UNIVERSITY OF COPENHAGEN
DEPARTMENT OF COMPUTER SCIENCE

Ph.D. Thesis

Laura Cabello

Evaluating Social Biases, Fairness, and Applications of Language Models

Advisor: Anders Søgaard

Assessment Committee: Maria Maistro, Ivan Vulić, Sunipa Dev

This thesis has been submitted to the Ph.D. School of The Faculty of Science, University of Copenhagen on February 11th, 2025.

"It has to start somewhere,
it has to start sometime.
What better place than here,
what better time than now?"
Guerrilla Radio
—Rage Against the Machine

Abstract

Language is at the heart of human interaction, shaping how we express ourselves, convey emotions, organise and engage in society. Speaking, writing, and reading are integral to everyday life, making language a primary tool for communication. Language Models (LMs) are probabilistic, machine learning models that learn from natural language data to predict, generate, and communicate effectively in human-like ways. Unlike earlier machine learning models, LMs serve as general-purpose text interfaces, applicable across a wide range of tasks, from machine translation to question answering to toxicity detection. Given their ease of use and versatility, LMs have become central to many everyday applications, powering services people interact with daily, often without realizing it. As these systems increasingly mediate human interactions, understanding their potential risks is crucial to ensuring they deliver inclusive and equitable technology for all users.

Throughout this thesis, we explore LMs in depth, examining their capabilities, potential applications, and the risks they pose. In particular, the first part investigates the nuanced relationship between bias and fairness, highlighting their distinction as separate phenomena. It then provides a detailed evaluation of biases and fairness metrics as applied to language models across various contexts: First, we investigate whether multilingual LMs truly provide equal performance across languages,examining the responses given for a simple fill-in-the-gap task; Second, we acknowledge the subjectivity of rationales (explanations) and study whether the rationales provided by models exhibit bias and influence their perceived fairness. Here, we focus on explanations for classification tasks and compare users' and models' rationales, revealing that models' explanations often align more closely with opinions from certain social groups; Finally, we extend our analysis to vision-and-language models, tracking how gender bias emerges and evolves from pretrained models to downstream tasks.

The second part of this work takes a more direct user-centric approach by exploring two practical applications of LMs: offensive language detection and medical question answering. These chapters highlight the potential of LMs to improve real-world scenarios.

All things considered, this thesis has contributed to advancing our understanding of the capabilities and limitations of LMs, as well as to construct a more comprehensive narrative of their societal impact.

Resumé

Sprog er kernen i menneskelige interaktioner og former, hvordan vi udtrykker os, formidler følelser og organiserer og engagerer os i samfundet. At tale, skrive og læse er en integreret del af hverdagen, hvilket gør sproget til det primære kommunikationsredskab. Sprogmodeller (LM'er) er probabilistiske maskinlæringsmodeller, der lærer af data til at forudsige, generere og kommunikere effektivt på menneskelignende måder. I modsætning til tidligere maskinlæringsmodeller præsenteres LM'er som generelle tekst-baserede kontaktflader, der kan anvendes på tværs af en lang række opgaver; fra maskinoversættelse, til besvarelse af spørgsmål og til detektering af toksicitet. På grund af deres brugervenlighed og alsidighed er LM'er blevet centrale for mange hverdagsapplikationer og driver tjenester, som folk interagerer med dagligt, ofte uden at være klar over det. Da disse systemer i stigende grad formidler menneskelige interaktioner, er det afgørende at forstå deres potentielle risici for at sikre, at de leverer inkluderende og ligeværdig teknologi til alle brugere.

Gennem denne afhandling undersøger vi LM'er i dybden og afsøger deres muligheder, potentielle anvendelser og de risici, de udgør. Den første del af afhandlingen undersøger det nuancerede forhold mellem bias og fairness og fremhæver deres skelnen som separate fænomener. Derefter giver den en detaljeret evaluering af skævheder og retfærdighedsmålinger, som anvendes på sprogmodeller på tværs af forskellige sammenhænge: Først undersøger vi, om flersprogede LM'er leder til ens præstationer på tværs af sprog gennem en undersøgelse af de svar, der er givet for en simpel cloze-test; Dernæst anerkender vi rationalernes subjektivitet (forklaringer) og undersøger, om rationalerne fra modellerne udviser bias som påvirker deres opfattelse af retfærdighed. Her fokuserer vi på forklaringer på klassifikationsopgaver og sammenligner brugeres rationaler med modellernes og afslører, at modellernes forklaringer ofte stemmer mere overens med meninger fra bestemte sociale grupper; Endeligt udvider vi vores analyse til billede-og-sprog-modeller og sporer, hvordan kønsbias opstår og udvikler sig fra præ-trænede modeller til downstream evaluering.

Den anden del af afhandlingen tager en mere direkte brugercentreret tilgang ved at udforske to praktiske anvendelser af LM'er: detektering af stødende sprog og besvarelse af medicinske spørgsmål. Disse kapitler fremhæver LM'ers potentiale til at forbedre konkrete opgaver i den virkelige verden.

Alt i alt har denne afhandling bidraget til at fremme vores forståelse af LM'ers muligheder og begrænsninger samt til at konstruere en mere omfattende fortælling om deres samfundsmæssige indvirkning.

Acknowledgements

What an incredible journey this has been. Honestly, doing a PhD has been a wonderful experience for me. For that I have to thank first and foremost my research group, CoAStaL. Thank you all for the countless thought-provoking discussions, always encouraging curiosity and inspiring new perspectives. A special mention goes to Constanza, Rita, Stephanie and Yova. Thank you for keeping things fun and easygoing, I feel exceptionally lucky to have shared these years with you. And Seolhwa, thank you for a lovely friendship filled with warmth, laughter, foodie adventures and endless conversations.

Of course, none of this would have been possible without the support of my supervisor, Anders Søgaard. Anders, thank you for believing in me, your encouragement and endless optimism (and I really do mean *endless*). Thank you Desmond and Daniel for helping me to develop as a researcher too. I am also grateful to all my collaborators for giving me the opportunity to learn from them and grow together. In general, I would also like to thank everyone I have met during my PhD who has made my time in Copenhagen truly unforgettable. While I can't name each of you, please know that I deeply value and appreciate your role in this chapter of my life.

A big thank you to my family for always supporting me and providing me with a nurturing education. Eva, thank you for bringing that special spark and a touch of spice to life. And thank you, Borja, for being my rock through the highs and, especially, the lows.

List of Publications

This is an article-based thesis. The content of the articles remains largely the same as in their original publications, with only minor changes such as the correction of typos and the reformatting of tables and figures for consistency. The following articles are included as chapters in the thesis, in the order they appear:

Laura Cabello, Anna Katrine Jørgensen, and Anders Søgaard. On the independence of association bias and empirical fairness in language models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 370–378, New York, NY, USA, June 2023b. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594004. URL https://doi.org/10.1145/3593013.3594004

Laura Cabello Piqueras and Anders Søgaard. Are pretrained multilingual models equally fair across languages? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3597–3605, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.318

Terne Sasha Thorn Jakobsen, Laura Cabello, and Anders Søgaard. Being right for whose right reasons? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1033–1054, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.59. URL https://aclanthology.org/2023.acl-long.59

Laura Cabello, Emanuele Bugliarello, Stephanie Brandl, and Desmond Elliott. Evaluating Bias and Fairness in Gender-Neutral Pretrained Vision-and-Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8465–8483, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.525. URL https://aclanthology.org/2023.emnlp-main.525

Li Zhou, Laura Cabello, Yong Cao, and Daniel Hershcovich. Cross-cultural transfer learning for Chinese offensive language detection. In

Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP), pages 8–15, Dubrovnik, Croatia, May 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.c3nlp-1.2. URL https://aclanthology.org/2023.c3nlp-1.2

Laura Cabello, Carmen Martin-Turrero, Uchenna Akujuobi, Anders Søgaard, and Carlos Bobed. MEG: Medical Knowledge-Augmented Large Language Models for Question Answering. *Under review*, 2024

I was also involved in the following publications that are not included in this thesis:

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.482. URL https://aclanthology.org/2022.acl-long.482

Laura Cabello Piqueras, Constanza Fierro, Jonas F. Lotz, Phillip Rust, Joen Rommedahl, Jeppe Klok Due, Christian Igel, Desmond Elliott, Carsten B. Pedersen, Israfel Salazar, and Anders Søgaard. Date recognition in historical parish records. In *Frontiers in Handwriting Recognition*, pages 49–64, Cham, 2022. Springer International Publishing. ISBN 978-3-031-21648-0. URL https://link.springer.com/chapter/10.1007/978-3-031-21648-0_4

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.c3nlp-1.7. URL https://aclanthology.org/2023.c3nlp-1.7

Oliver Eberle, Ilias Chalkidis, Laura Cabello, and Stephanie Brandl. Rather a Nurse than a Physician - Contrastive Explanations under Investigation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6907–6920, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.427. URL https://aclanthology.org/2023.emnlp-main.427

Laura Cabello and Uchenna Akujuobi. It is Simple Sometimes: A Study On Improving Aspect-Based Sentiment Analysis Performance. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6597–6610, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.394. URL https://aclanthology.org/2024.findings-acl.394

Table of Contents

Chapter 1

Introduction

The idea of Artificial Intelligence (AI) began to captivate public imagination in the early 20th century. Science fiction movies introduced characters like the humanoid robot Maria in *Metropolis* or Gort in *The Day the Earth Stood Still*, exploring themes of mechanized intelligence and human-like machines. By the fifties, a generation of scientists, mathematicians, and philosophers had culturally assimilated the concept of AI[1], perhaps due to the influence of sci-fi cinema, and started the transition of these imaginative ideas into tangible realities. Alan Turing proposed a theoretical framework for constructing intelligent machines and evaluating their intelligence (Turing, 1950) in a formal attempt to answer the question "Can machines think?". Around the same time, Claude Shannon introduced the foundational principles of information theory (Shannon, 1948), establishing methods to quantify the predictability of symbols within a communication channel. These principles became instrumental in the development of statistical language modelling. Together, these groundbreaking contributions laid the theoretical foundation for decades of research in machine learning and continue to profoundly shape the evolution of modern AI systems[2].

Significant advancements in algorithm design, coupled with breakthroughs in hardware design —such as expanded computer storage and increased processing speeds—, have enabled the development of increasingly powerful AI systems. By October 2024, these breakthroughs culminated in

[1] sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/

[2] We agree with Basdevant et al. (2024)'s framework for AI systems, which encompasses infrastructure components such as storage, model elements such as datasets, code, and model weights, as well as product and user experience (UX) considerations.

global recognition when the Royal Swedish Academy of Sciences awarded the Nobel Prizes in Physics and Chemistry to work directly related to AI. The Nobel Prize in Physics was awarded to John J. Hopfield and Geoffrey E. Hinton for their transformative contributions that "enable machine learning with artificial neural networks" (Nobel Prize Outreach AB, 2024b). The Nobel Prize in Chemistry recognized Demis Hassabis and John M. Jumper for their pioneering work in creating an AI algorithm that solved a decades-long challenge in predicting protein structures, alongside the scientist David Baker, for his work in designing new proteins (Nobel Prize Outreach AB, 2024a). The historic announcement of these Nobel Prizes for AI-related work has sparked numerous discussions in mainstream media[3][4] about their implications in science, medicine, and society. At the same time, the laureates have also acknowledged important ethical aspects of AI that must be considered (Li and Gilbert, 2024). We align with the cautionary perspective and, throughout this thesis, systematically highlight the potential risks associated with the use of language models (LMs).

Closer to the subfield of Natural Language Processing (NLP), language models, once purely theoretical constructs, have evolved into transformative tools deeply integrated into modern society (Naveed et al., 2024). Millions of people are using general-purpose AI systems based on large language models (LLMs) as they have become the backbone of numerous real-world applications. They are now used to track public opinion on politicians and products (Nissim et al., 2020), to moderate online content (Winchcomb, 2019; Kumar et al., 2024), and to power conversational bots (OpenAI, 2024; Gemini Team Google, 2024). However, their widespread use also raises important questions about their safe development and impact on society. How do these systems navigate the complexities of language diversity and cultural contexts? Are they promoting social inequalities? And, if so, what are the broader implications of these inequalities?

This dissertation explores the multifaceted dimensions of language models. We study the behaviour of LMs to deepen our understanding of the biases they encode, their impact on users, and ultimately, how LMs can be effectively applied in real-world applications. Early research in NLP led to models based on deep neural networks that were specialized in specific tasks (LeCun et al., 2015), such us translation (Sutskever et al., 2014) or summarisation

[3] www.nytimes.com/2024/10/13/briefing/nobel-prize-artificial-intelligence.html
[4] www.economist.com/science-and-technology/2024/10/10/ai-wins-big-at-the-nobels

(Gambhir and Gupta, 2017). The models examined in this thesis are based on the more recent Transformer architecture (Vaswani et al., 2017b), and commonly follow a two-step learning strategy. Models are initially trained during the **pretraining** stage. The goal of pretraining an LM is to leverage large, unlabelled data sources to learn a general understanding of the world by predicting words or sentences in a way that captures the structure, semantics, and context of text. Pretraining requires substantial computational and financial resources to train increasingly large models on vast amounts of data. While academia has made significant technical contributions, the development of these foundation models is predominantly driven by industrial labs due to the high costs involved (Bommasani et al., 2022). As a result, the academic researchers have increasingly focused on probing the representations learned by these models to gain deeper insights into their capabilities and limitations. This phase provides the model with a broad foundation of linguistic knowledge that can be adapted for specific tasks later, in what is known as **fine-tuning**.

We engage in and advance the ongoing discussion on developing inclusive and equitable NLP technologies. Specifically, our research interest aims at understanding LM's representations from a "bias" point of view and their use across diverse social groups (Part I). We complement this study with an analysis of two popular use cases of LMs, namely content moderation and medical question answering (Part II), and a discussion of potential implications for ongoing and future research (Part III).

1.1 Definitions

As highlighted by Blodgett et al. (2020), Mehrabi et al. (2021), and Gallegos et al. (2024), NLP research often provides vague or inconsistent definitions of bias, or simply omits their explicit conceptualisations. To avoid common misconceptions, we outline the key definitions of bias and fairness used throughout this thesis.

A widely accepted definition of **bias** in NLP originates from Crawford's NeurIPS keynote in 2017, where she described it as a "skew that produces a type of harm" towards different social groups. Building on this definition, **social bias** serves as an umbrella term encompassing any form of disparate treatment between social groups stemming from historical and structural power imbalances (Gallegos et al., 2024). We can arguably say that early re-

search addressing NLP bias mostly focused on measuring social bias through word associations, *i.e.*, systematic differences in how words and phrases referring to demographic groups are represented in a model.

In this context, we define a **social group** as a segment of the population that shares a common identity trait —referred to as **sensitive** or **protected attribute**— which can be fixed, situational, or socially constructed. Protected attributes are typically demographic in nature, encompassing characteristics such as gender, race, ethnicity, age, socioeconomic status, and geographic location. For instance, gender and race are protected attributes well studied in NLP literature (Field et al., 2021; Stanczak and Augenstein, 2021) that we examine in Chapter 3, 4 and 5.

In general, inequalities —or disparate outcomes— between social groups are common across societies. As a result, data used to train models and inform decision-making processes often mirror the social biases of those who create or collect the data (Chang et al., 2019; Mehrabi et al., 2021); and, as Barocas and Selbst (2016) put it, an algorithm is only as good as the data it works with. To illustrate this phenomenon, we take arrest outcomes as an example. Arrest decisions depend on the discretion of patrol officers, who determine which individuals to investigate and whether specific behaviours qualify as crimes. Bunting et al. (2013) reported that between 2001 and 2010 in the United States, black and white people used marijuana at similar rates, yet black individuals were nearly four times more likely to be arrested for marijuana possession. By simply looking at police reports, one might mistakenly conclude that black people are more likely to commit this type of crime, when in reality, the detention rates reflected in these reports are themselves biased. Schramowski et al. (2022); Navigli et al. (2023) and Cabello et al. (2023a), amongst others, show that models trained on human-generated data can inherit, perpetuate and even exacerbate social biases leading to harmful consequences. Moreover, the use of biased data can perpetuate stereotypes (Nadeem et al., 2021) and create self-reinforcing cycles, such that, for instance, a proactive intensified policing in certain high-crime neighbourhoods can amplify disparities in rates of arrest across groups, which may in turn perpetuate the conditions that lead to increased crime in these areas (Behav, 2017).

The societal impact of language models has prompted researchers to investigate not only their biases but also the **fairness** of their applications (Barocas et al., 2019). The concept of fairness has long been a central theme in fields such as psychology, philosophy, and the social sciences (Goff, 1983;

Robeyns, 2009). John Rawls' philosophy is a prime example. In his seminal work, "A theory of Justice" (Rawls, 1971), Rawls argues that social institutions must be fair to all members of society, regardless of their background and circumstances. He outlines his framework for distributive justice as follows.

> *Social and economic inequalities are to be arranged so that they are both:*
>
> *a) to the greatest benefit of the least advantaged, consistent with the just savings principle, and*
>
> *b) attached to offices and positions open to all under conditions of fair equality of opportunity.*

Principle *a)*, often referred to as the Difference Principle, does not enforce strict quality but instead calls for maximizing the "benefit of the least advantaged". Principle *b)* refers to "equality of opportunity", which means that one action is preferred over another only if its true quality is superior. Despite the widespread adoption of Rawlsian fairness, it has also faced significant criticism across disciplines, especially the Difference Principle. Examples of such critiques include the focus on an abstract worst-off group rather than specific social groups or individuals (Altham, 1973); extreme risk aversion (Mueller et al., 1974)[5], and more recently, the existence of loopholes that may inadvertently exacerbate inequality (Jørgensen and Søgaard, 2023). In response to the significant number of critiques received, John Rawls presented a restatement (Rawls, 2001) adding further thought to his earlier theory of justice. This restatement as well as both his critics and supporters agree on the important idea of achieving **justice through fairness**.

Defining fairness in NLP is a challenging task that requires an understanding of its interdisciplinary nature and intricate nuances. Dwork et al. was amongst the first to introduce the fairness debate to computer science. Dwork et al. (2012) argued that computational systems which satisfy some definition of group fairness can still yield fairness related harms for people

[5]The Difference Principle corresponds to the maximin strategy in game theory. According to Mueller et al. (1974), Rawls assumes an extreme degree of risk aversion as he "ignores information concerning the alternative payoffs to all citizens except the worst-off under each strategy."

belonging to those groups. This argument has been discussed by many others since then (Joseph et al., 2016; Corbett-Davies et al., 2024), forming an ongoing debate about individual fairness and group fairness. In Part I, we focus on **group fairness**, which aims to ensure equal treatment of individuals regardless of their group membership—as opposed to **individual fairness**, which would focus on ensuring that similar individuals are treated similarly.

In general, fairness has multiple context-dependent, and sometimes even conflicting, theoretical understandings (Hutchinson and Mitchell, 2019; Miconi, 2017; Jacobs and Wallach, 2021). We deliberately avoid delving into the diversity of fairness definitions and generally adopt a pragmatic approach to investigating fairness, aligning with John Rawls' theory of distributive justice and stability. We define fairness as *equal performance across social groups*[6][7]. In this definition, we understand performance —*e.g.*, in terms of predictive accuracy— as a resource to be allocated.

While we offer a set of definitions here and in the relevant chapters, we acknowledge that the concepts of "bias" and "fairness" are inherently normative and subjective, which often depend on context and culture. We specifically define fairness and related notions within the context of each chapter.

1.2 Measures

Research on mitigating bias is often driven by fairness concerns. Researchers have, for example, measured social bias to ensure fairness or assume that bias causes unfairness. However, as we have emphasized above, bias and fairness are distinct concepts, and the metrics used to evaluate them should not be conflated. Part of the confusion likely stems from the lack of rigorous definitions and processes for detecting and measuring model biases (and fairness) (Blodgett et al., 2020; Oneto and Chiappa, 2020; Castelnovo et al., 2022c). This makes it challenging to understand what bias metrics actually measure and, ultimately, raises questions about whether observed differences in results are due to analytical errors or the design choices behind these metrics.

[6]This definition specifically frames the narrower concept of **empirical fairness**. Throughout the remainder of this thesis, we will primarily refer to the empirical aspect, even when using the term fairness alone.

[7]Consequently, **unfairness** is defined as *differences in performance across social groups that could potentially lead to unfair and discriminatory decision-making.*

Here, we summarise the shared aspects covered in Part I regarding bias and fairness measurements. The specific metrics and context are introduced in each chapter. This common background does not directly apply to Part II, as Part II focuses on the task performance of specific applications rather than unfair outcomes. Measurements used in Part II are based on predictive accuracy and F1-scores.

Measuring Bias Surveys of Blodgett et al. (2020); Mehrabi et al. (2021), and Czarnowska et al. (2021) contribute to the "bias and fairness" narrative, acknowledging the intertwined relationship between the two. Sheng et al. (2021) and Dev et al. (2022) discuss general trends on evaluating biases in language models; specifically, Dev et al. (2022) provides a taxonomy that enable direct comparisons between bias metrics and points out to what harms are being measured.

A common framework for evaluating the bias in an LM involves broadly classifying metrics as either **intrinsic** or **extrinsic**, depending on whether they assess biases during the language model's pre-training stage or within a specific downstream task. As written in Chapter 2, intrinsic metrics are typically applied to word representations and relate bias to the geometry of the embedding space, while extrinsic metrics quantify biases in model predictions for a specific task, *e.g.*, sentiment classification (Cabello et al., 2023b). However, the relationship between these type of biases remains unclear. Numerous studies (Goldfarb-Tarrant et al., 2021; Delobelle et al., 2021; Kaneko et al., 2022; Cao et al., 2022; Orgad et al., 2022) indicate that intrinsic bias in language models does not consistently align with bias measured extrinsically in downstream tasks or with empirical fairness (Shen et al., 2022; Cabello et al., 2023b).

In this thesis, we operationalise intrinsic bias through **representational bias** (Crawford, 2017), which occurs when a model captures associations between a protected attribute (*e.g.*, gender) and specific concepts (*e.g.*, job titles) within its embedding space (Beukeboom, 2014). This is, we aim at quantifying systematic differences in representations of demographic groups encoded in the embedding space of the models. Specifically, in Chapter 2 and Chapter 5 we measure gender bias towards occupations and contextual words, respectively. For instance, these measures allow us to answer questions like "Does this model show a bias toward referring to doctors as *male* or *female*?", "Does this model associate *cooking* with *women* more often than

with *men*?" When these questions are answered affirmatively, this skewed behaviour could lead to harmful consequences such as favouring the hiring of male doctors. In Chapter 5 we also look into bias amplification. **Bias amplification** occurs when a model exacerbates unwanted biases present in the training data. Unlike other biases, it cannot be solely attributed to the data since it can fluctuate significantly throughout training (Hall et al., 2022).

Measuring Fairness Fairness metrics can be classified into three main types: calibration-based, precision-based, and recall-based metrics. Kleinberg et al. (2016); Friedler et al. (2016) and Miconi (2017) show how pairs of fairness metrics can be mathematically incompatible—that is, one type of fairness may exclude another. This incompatibility occurs for any pair of metrics with different origins (*e.g.*, calibration-based vs. recall-based) unless the true base rates are the same for all groups or the classifier performs perfectly. Precision- and recall-based metrics are primarily used to evaluate LMs because they provide a reliable way to measure performance by comparing model outputs to a predefined standard.

Based on our definition of fairness given in § 1.1, we derive our metrics from differences in performance[8]. Performance differences across user demographics, either at individual or group level, is arguably the most common fairness metric (Barocas et al., 2019; Hutchinson and Mitchell, 2019). Assuming a fair algorithm would yield equally good (or bad) performance regardless of the user's group identity, **group disparity** measures how much the model deviates from this ideal behaviour and what group is favouring. The concept of *group disparity* broadly refers to any differences in performance between groups. In cases where more than two groups are defined, it is also crucial to examine the gap between the most advantaged (highest performing) and least advantaged (lowest performing) groups. This is also known as the **maximin difference**[9] in performance. The maximin difference captures the widely shared intuition that fairness is always in the service of the worst-off group (Rawls, 1971). Barocas et al. (2019); Mehrabi et al. (2021) and Czarnowska et al. (2021) offer excellent resources for discussing relevant metrics in different contexts and comparing available fairness metrics.

[8]The specific meaning of "performance" is defined individually in each chapter.
[9]Also referred to as "max-min" or "min-max".

1.3 Contributions

This dissertation contributes to the AI landscape as follows. Our goals are to highlight the importance of understanding the implications of social biases, gain insights into evaluating task performance from a fairness perspective, and adapt general-purpose models to specific scenarios. To do so, we put an emphasis on language models' societal impact (*e.g.*, inequity, misuse, and ethical considerations), technical principles (*e.g.*, model architectures, training procedures, data, evaluation) and their applications (*e.g.*, content moderation and medical question answering). Experiments are conducted on English data unless stated otherwise. We organise the work in two parts with distinct contributions.

Throughout **Part I**, we put forward the argument that language models are not only biased in various ways but also bring disparate user experiences.

In ***Chapter 2*** (Cabello et al., 2023b), we contribute to what was, at the time, a novel line of research demonstrating that bias in representations is not necessarily indicative of bias or fairness in downstream applications (Shen et al., 2022; Goldfarb-Tarrant et al., 2023). We clearly formulate the distinction between (association) bias and (empirical) fairness, and present a comprehensive study exploring the relationship between these two concepts from three points of view: theoretically, through a thought experiment; practically, using standard measurements; and socially, by reviewing social science literature. Finally, we propose the In-Group Affinity Assumption, which highlights the idea that a specific demographic group tend to use in-group terms more frequently—or in different ways—than other groups. For instance, this suggests that word co-occurrence metrics overlook use-mention distinctions, where harmful words might appear in the context of references to a social group without being used to directly target that group negatively. This was later also noticed by others (Gligoric et al., 2024; Gallegos et al., 2024). Our findings highlight potential limitations of embedding-based metrics and indicate that debiasing efforts in pretrained models, when guided by these metrics, cannot be guaranteed to effectively propagate downstream.

Chapter 3 (Cabello Piqueras and Søgaard, 2022) presents our study on group fairness of pretrained multilingual models, asking whether these models are equally fair across languages. For this purpose, we create a novel multilingual dataset of parallel cloze test examples—in English, Spanish,

French and German—equipped with annotator's demographic information (balanced with regard to gender and native tongue). We evaluate three popular multilingual models on the same cloze task given to participants, and compare their linguistic preferences. Our findings reveal different levels of group disparity in the models examined. The fact that all the languages in our study are Indo-European only reinforces the idea that if inequalities are observed in these closely related languages, they are likely to exist in other linguistic families as well.

In ***Chapter 4*** (Thorn Jakobsen et al., 2023), we focus on evaluating rationale-based explanations, and ask which demographics align with models' predictions best, and whose reasoning patterns align with the models' rationales best. To this end, we present a collection of three existing datasets augmented with demographic annotations (balanced across age and ethnicity), covering sentiment classification and common-sense reasoning tasks. Such data allow us to profile models by quantifying their alignment with rationales provided by different socio-demographic groups. Specifically, we examine label and rationale agreement across these groups and assess the extent to which the rationales of these groups align with those generated by Transformer-based models. This work makes a valuable contribution to the fairness literature: even if models are fair in terms of task performance, biases may still emerge when examining their reasoning process. This implies that undesired biases do not always manifest in task performance alone.

Chapter 5 (Cabello et al., 2023a) puts the focus on gender bias in multimodal models. The use of visual and grammatical gender information is essential for identifying the target of, for example, a question about an image. However, demographic factors should not disproportionately influence the outcome of vision-and-language (V&L) models, as this could reinforce harmful stereotypes (van Miltenburg, 2016; Bianchi et al., 2023). To address this issue, we investigate the relationship between intrinsic and extrinsic gender bias by quantifying bias amplification. Consistent with previous findings for language models, bias in a pretrained V&L model does not necessarily result in a comparable level of bias in downstream task performance. Additionally, we measure fairness through group disparity and demonstrate that it is not directly related to model bias either. Finally, we propose a simple method to improve fairness in V&L models: an extra pretraining epoch on gender-neutral data[10] reduces fine-tuning variance and group disparity

[10]Gender-neutral data refers to data in which gendered terms have been replaced with

across most models studied, without significantly affecting task performance.

Part II shifts toward a more applied narrative, where we directly examine the role of LMs in downstream tasks like offensive language detection and medical question answering.

In ***Chapter 6*** (Zhou et al., 2023b), we investigate the impact of transfer learning using data for offensive language detection in different languages and cultural backgrounds (English and Korean) to the target language (Chinese). We set up different training scenarios and demonstrate the presence of culture-specific biases in the source data that negatively impact the transferability of language models in this context. For instance, models often do not classify region-related comments as offensive, even though such language is typically considered toxic when used against someone in Chinese culture. While this distinction is clear to someone familiar with Chinese cultural norms, it is poorly represented in the model's training data, which disproportionately reflects American cultural values. A direct application for offensive language detection is online content moderation. In online content moderation, messages are processed by an NLP system that evaluates them as offensive or non-offensive, acceptable or unacceptable, and restricts the sharing of posts that fail to meet the acceptance criteria (Nobata et al., 2016). A key challenge is that offensive language can vary greatly depending on context and cultural backgrounds. Previous research highlighted the importance of cross-lingual transfer in offensive language detection (Stappen et al., 2020; Lamprinidis et al., 2021; Shi et al., 2022), but we are amongst the first to investigate the influence of cultural transfer learning for offensive language detection and, ultimately, content moderation.

Chapter 7 (Cabello et al., 2024) introduces a different type of NLP applications: medical question answering. NLP in question answering aims at building systems to answer questions about any topic in natural language. ChatGPT (OpenAI, 2024) is one of the most notorious examples that proves modern dialogue systems are able to write long coherent passages and generalize well to many domains. Yet, their human-like fluency can be deceiving. These models sometimes generate answers that, while plausible, are factually incorrect —a phenomenon known as AI hallucinations. This issue poses a significant risk, particularly in critical areas like healthcare, where misinformation can have serious consequences. To improve their reliability without

gender-neutral alternatives (*e.g.*, sister → sibling).

adding much to computational costs, researchers have experimented with retrieval-augmented systems (Lewis et al., 2020), or training from mixtures of corpora and knowledge bases (Pan et al., 2023, 2024). We take a step forward in this direction by designing a novel architecture for medical question answering that integrates language models and knowledge graphs. Our system creates an external storage of knowledge graph embeddings, retrieves relevant factual information from it, and feed it to an autoregressive large language model (LLM) to guide the response generation process. We experiment with three of the latest open-source instruction-tuned models, based on Mistral (Jiang et al., 2023) and Llama (Dubey et al., 2024). Our results prove the viability of augmenting LLMs with information from a large, curated knowledge graph, even surpassing strong LLM baselines like BioMistral Labrak et al. (2024) or MediTron Chen et al. (2023a), which have followed a costly continued pretraining of the base LLMs on curated biomedical data.

Part I

Social Biases & Fairness in Language Models

Chapter 2

On the Independence of Association Bias and Empirical Fairness in Language Models

Abstract

The societal impact of pre-trained language models has prompted researchers to probe them for strong associations between protected attributes and value-loaded terms, from slur to prestigious job titles. Such work is said to probe models for *bias or fairness*—or such probes "into representational biases" are said to be "motivated by fairness"—suggesting an intimate connection between bias and fairness. We provide conceptual clarity by distinguishing between association biases Caliskan et al. (2022) and empirical fairness Shen et al. (2022) and show the two can be independent. Our main contribution, however, is showing why this should *not* come as a surprise. To this end, we first provide a thought experiment, showing how association bias and empirical fairness can be completely orthogonal. Next, we provide empirical evidence that there is no correlation between bias metrics and fairness metrics across the most widely used language models. Finally, we survey the sociological and psychological literature and show how this literature provides ample support for expecting these metrics to be uncorrelated.

2.1 Introduction

The prevalence of unintended social biases in pre-trained language models (PLMs) is alarming, since they impact millions, if not billions of people every day. In recent years, more and more NLP researchers have studied such biases, making up an estimated 6.3% of the literature in 2022 Ruder et al. (2022). Much of this work has focused on what Crawford (2017) called *representational bias*, which manifests when portrayals of certain demographic groups are discriminatory. In NLP, representational bias often arises when associations between a protected attribute, *e.g.*, gender, and certain concepts, *e.g.*, job titles, are captured in the model space. Thus, to avoid ambiguity, we will refer to this type of bias as *association bias*, following Chaloner and Maldonado (2019).

Association bias is often confused with what is sometimes referred to as performance disparity Hashimoto et al. (2018) or *empirical fairness* Shen et al. (2022), *i.e.*, performance differences across end user demographics. Or mitigating association bias is assumed to improve empirical fairness Chen et al. (2020b); Friedrich et al. (2021); Cao et al. (2022); Dayanik and Padó (2020); Castelnovo et al. (2022b); Reddy et al. (2021). Note that most fairness metric focus on some form of equal performance and differ only in whether they focus on precision, recall or balancing the two Barocas et al. (2019). Empirical fairness refers to equal performance as measured by *de facto* standard metrics and is arguably the most common fairness metric Williamson and Menon (2019); Barocas et al. (2019).

In this paper, we will show that the two phenomena, association bias and empirical fairness, are often completely independent matters.[1] We devise a thought experiment (§2.3) to illustrate this, but also present a series of experiments (§2.4) to show that results obtained the way association bias is normally measured, do not correlate with results obtained the way empirical fairness is normally measured.

Our main contribution, however, is to show that this should not come as

[1]Note that the distinction between association bias and empirical fairness—between how expressions referring to demographic groups are encoded, and how these groups are treated as end users—is different from another distinction made in recent work Delobelle et al. (2021); Goldfarb-Tarrant et al. (2021); Kaneko et al. (2022) between intrinsic and extrinsic bias: Intrinsic bias, here, is what we call representational bias, whereas extrinsic bias refers to performance differences on sentences containing entities referring to different demographic groups.

a surprise. Research on mitigating association bias and empirical fairness is often motivated by fairness concerns, and bias and fairness are often considered near-synonymous terms in the research literature: Researchers have, for example, said that bias *causes* unfairness Chang et al. (2019); Friedrich et al. (2021); Castelnovo et al. (2022b). If this was the case, the independence of association bias and empirical fairness should come as a great surprise. However, the assumption that bias causes unfairness, is unwarranted, as we will see below, from a survey of relevant literature from the social sciences (§2.5). A causal link between association bias and empirical fairness would seem to require some sort of in-group affinity, *i.e.*, that groups use terms relating to their in-group peers more and in different ways than outsiders, like, for instance, Democrats on Twitter mention Trump and the Republican party more often than their Republican counterparts Duijnhoven (2018). This assumption, which we call the In-Group Affinity Assumption, seems intuitive, but without much support from the social sciences (§2.5).

Contributions In §2.2, we define association bias and empirical fairness and discuss related work. When we talk about association bias, we refer to systematic biases in how words and phrases referring to demographic groups are encoded. Figure 2.1 visualizes how models may exhibit biased associations because of sample biases, and may even amplify these. We define empirical fairness as equal performance across groups, because this is the most balanced and most widely applicable measure of fairness in NLP, except for specialized applications where equal base rates and calibration take priority over performance. We then move to study how association bias and empirical fairness relate. In §2.3, we show that theoretically, association bias and empirical fairness are completely independent. That is, mitigating association bias can hurt empirical fairness, and ensuring empirical fairness can introduce more bias. §2.4 shows there is no obvious correlation between results obtained from standard association bias measurements and results obtained from standard empirical fairness measurements of language models. Finally, §2.5 surveys the social science literature for explanations on why association bias and empirical fairness may be less related (or related in less obvious ways) than multiple works in the NLP literature have assumed up to this point. The finding that association bias and empirical fairness are independent in this three-way investigation, should help push research horizons and provide strong motivation for targeting empirical fairness directly,

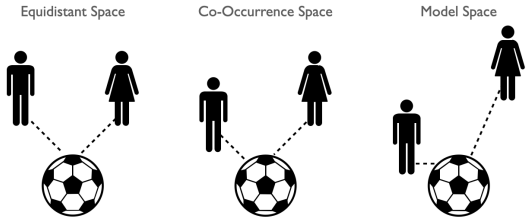


Figure 2.1: Association bias of group-related terms (*e.g.*, *woman* and *man*) can be quantified as degree of isomorphism relative to an empirical (**co-occurrence**) space or a normative, **equidistant** space. The graph illustrates how *man* may be more strongly associated with *soccer* in a **model**, less so empirically (the underlying data or real-world statistics), and not at all in an ideal world.

as well as for seeing association bias mitigation, not necessarily as a way of promoting fairness, but rather as a way of preventing poor inferences and generation of stereotypical text.

2.2 Definitions and Related Work

In the NLP literature, bias and fairness are often conflated, or it is argued that one follows from the other, *e.g.*, that we can ensure fairness by mitigating bias Chen et al. (2020b); Friedrich et al. (2021); Cao et al. (2022); Dayanik and Padó (2020); Castelnovo et al. (2022b); Reddy et al. (2021). In contrast, we will show that this is not always the case, and (association) bias and (empirical) fairness often are independent or at odds.

Bias Mitigating social biases in NLP models has become an important research goal Shah et al. (2020); Hutchinson et al. (2020); Romanov et al. (2019), but there is little consensus on how to evaluate such biases Blodgett et al. (2020); Stanczak and Augenstein (2021). We focus on association bias and show how, contrary to what seems to be popular belief, it is not unequivocally related to fairness; in fact, it is very often completely independent thereof.

Association bias in a model refers to systematic differences in how words and phrases referring to demographic groups are encoded. Classical tests

thereof include comparing the (cosine) distance of terms relating to protected attributes, *e.g.*, *woman* and *man*, or their vectors $\mathbf{v}_w$ and $\mathbf{v}_m$, to terms of particular interest, *e.g.*, slur Sap et al. (2019), sentiment Ali et al. (2022a), or job titles such as *doctor* ($\mathbf{w}_d$) Zhao et al. (2018). Early papers would quantify bias with respect to, say, gender, as cosine similarities ($\mathbf{cos}(\mathbf{v}_m, \mathbf{w}_d) - \mathbf{cos}(\mathbf{v}_w, \mathbf{w}_d)$) Caliskan et al. (2017); Bhatia (2017); Zhao et al. (2018); Brunet et al. (2019); Gonen and Goldberg (2019), and by seeing whether the nearest neighbor of $\mathbf{w}_d + \mathbf{v}_w - \mathbf{v}_m$ would be *nurse* or another job stereotypically associated with women Bolukbasi et al. (2016). In practice, NLP researchers have used tests such as the ones above for quantifying association bias Caliskan et al. (2017); Bhatia (2017); Zhao et al. (2018); Brunet et al. (2019); Gonen and Goldberg (2019). We will argue that such quantities are theoretically and, often practically, orthogonal to empirical fairness, which we define in terms of differences in performance estimates across demographics, *i.e.*, social groups Williamson and Menon (2019); Barocas et al. (2019); Shen et al. (2022), often defined by the cross-product of a subset of protected attributes such as gender, age, or race.

Fairness Fairness metrics come in multiple flavors, but are often divided in three: calibration-based, precision-based, and recall-based metrics. Miconi (2017), Friedler et al. (2016) and Kleinberg et al. (2016) show how pairs of fairness metrics can be mathematically incompatible, *i.e.*, one type of fairness can rule out another. In fact, incompatibility holds for all pairs of metrics such that the two metrics are of different flavor, *e.g.*, calibration-based and recall-based, unless the true base rates are identical across groups, or the classifier has perfect performance. Since the vast majority of NLP applications provide repetitive services, the quality of which can be measured against a gold standard, precision- and recall-based metrics are predominantly used in NLP. We follow several authors Hashimoto et al. (2018); Hansen and Søgaard (2021); Chalkidis et al. (2022) in using min-max differences in (the standard) performance (metric) as our go-to fairness metric. Relying on min-max difference captures the widely shared intuition that fairness is always in the service of the worst off group Rawls (1971). For a discussion of available fairness metrics, and in what contexts they are relevant, see Mehrabi et al. (2021) and Barocas et al. (2019). For a comparison of existing metrics used to quantify social biases in NLP, see Czarnowska et al. (2021).

Related Work Maity et al. (2020) study the effect of subpopulation shifts on performance disparities and show that these do not always relate in obvious ways. Goldfarb-Tarrant et al. (2021) study the correlation between what they refer to as "intrinsic and extrinsic measures of representational bias". Their intrinsic measures of representational bias amount to word association bias, but their extrinsic measures of representational bias are not empirical fairness measures. To see this, consider the coreference task used in Goldfarb-Tarrant et al. (2021). Goldfarb-Tarrant et al. (2021) correlate intrinsic gender bias measures (cosine distances in static word embedding spaces) with coreference performance on sentences with female and male referents. We argue that in this case, empirical fairness would be performance on sentences *written by* female and male authors.[2] Goldfarb-Tarrant et al. (2021) establish that there is no correlation between their two measures of representational bias. Their result superficially looks similar to and in agreement with ours, but is, in fact, unrelated. If anything, it shows that association bias has been assumed to correlate with many measures that it does not, in fact, correlate with. Cao et al. (2022) and Kaneko et al. (2022) studied the same problem as Goldfarb-Tarrant et al. (2021), but used contextualized token embeddings from PLMs rather than static word embeddings. They both found weak correlations between intrinsic and extrinsic evaluation measures. Again, we emphasize that these results do not contradict ours.

Shah et al. (2020) carefully avoid to discuss fairness, saying the fairness literature is outside the scope of their paper, but place outcome disparity (performance disparity) as a central motivation for social bias mitigation. They list four potential causes of outcome disparity: label bias, selection bias, bias amplification, and semantic (representation) bias. We show association (representation) bias and outcome disparity are theoretically, and also often practically, independent, questioning their fourth hypothesis. Moreover, we observe that outcome disparity can arise in the absence of *all* of the above four factors. Say a group exhibits more variance than others, *e.g.*, because of spelling variation in dyslexics. Even if dyslexics are represented proportionally or equally, they may still see worse performance with dyslexics than for non-dyslexics.

Finally, Shen et al. (2022) show how a different form of representational fairness, *i.e.*, whether protected author attributes can be detected from model

[2]Or, alternatively, sentences *read by* female and male authors. The latter is rarely studied, as it requires reader statistics, *e.g.*, from online media services.

representations, is also uncorrelated with empirical fairness. Together, our work and previous work Goldfarb-Tarrant et al. (2021); Cao et al. (2022); Kaneko et al. (2022); Shen et al. (2022) establish that four common bias-related measures — (i) association bias, (ii) performance on sentences with protected attribute terms (Goldfarb-Tarrant et al. (2021)'s extrinsic measure), (iii) decodability of protected attributes from representations, and (iv) empirical fairness are largely uncorrelated. Specifically, (i) is independent of (ii) and (iv), and (iii) is independent of (iv).

Our work is motivated by the large-spread assumption that association bias and empirical fairness are causally related Chen et al. (2020b); Friedrich et al. (2021); Cao et al. (2022); Dayanik and Padó (2020); Castelnovo et al. (2022b); Liu et al. (2020b); Qian et al. (2022); Sun et al. (2019); Ross et al. (2021); Bartl et al. (2020). Bartl et al. (2020), for example, aspire "to promoting fairness in NLP by exploring methods to measure and mitigate gender bias." Ross et al. (2021) say they "believe that by revealing biases, by providing tests for biases that are as focused as possible on the smallest units of systems, we can both assist the development of better models and allow the auditing of models to ascertain their fairness." Sun et al. (2019), argue that "biased predictions may discourage minorities from using those systems and having their data collected, thus worsening the disparity in the data sets", equating biased predictions with unfair predictions.

All three sets of authors see bias as the primary cause of fairness. Showing such causation is not a given, and that in fact, association bias and empirical fairness need not even correlate and are often orthogonal, is an important correction to this literature, with potential consequences for research methodology, applications of NLP in the social sciences, as well as AI ethics and regulation.

2.3 Association Bias and Empirical Fairness are Independent (in Theory)

In this section, we produce a thought experiment—a synthetic model—to illustrate how bias and fairness can in fact be completely independent of one another. We construct a synthetic ternary (positive/negative/neutral) sentiment analysis model with a small feature space, including words that refer to demographic subgroups of a population. These words, denoting various

groups, will be biased and associated with sentiment, because of biases in our training data. This assumption is also made in Ali et al. (2022b), for example. These associations lead to biased likelihood estimates and would, in the context of a linear model, lead to differences in the degree of isomorphism relative to the group-specific subgraphs. We will show, however, that the resulting biases are independent of the group fairness of the model, *i.e.*, to the min-max performance disparities across the same groups. Such a connection, if it exists, could be explained by an *in-group affinity*, which relies on the assumption that those biased terms are used by the in-group more frequently or in other ways than by other groups.

Say a population consists of members of groups $g_1, \ldots, g_4$, *e.g.*, defined according to their address as *north*, *east*, *west* and *south*. Everyone speaks the same language and expresses sentiment with a vocabulary of seven words: $w_{g_1}, \ldots, w_{g_4}, w_5, w_6, w_7$. Except w_6 (positive) and w_7 (neutral), all words express negative sentiment, including the words that refer to (or are associated with) other demographic subgroups (w_{g_i}), for instance, *northern*, *eastern*, *western* and *southern*. The subgroups use the terms with the following probabilities (Table 2.1):

	w_{g_1}	w_{g_2}	w_{g_3}	w_{g_4}	w_5	w_6	w_7
g_1	0.0	0.25	0.0	0.0	0.25	0.25	0.25
g_2	0.0	0.0	0.25	0.0	0.25	0.25	0.25
g_3	0.0	0.0	0.0	0.25	0.25	0.25	0.25
g_4	0.25	0.0	0.0	0.0	0.25	0.25	0.25

Table 2.1: Probability of a group g_i using the word w_j for expressing sentiment. Only w_6 (positive) and w_7 (neutral) express a non-negative sentiment.

This data exhibits four representational biases, *e.g.*, the association of g_1 with negative sentiment, the association of g_2 with negative sentiment, and so forth. If we have sufficient data, a simple model, *e.g.*, a Naive Bayes classifier trained on simple bag-of-words representations, should induce the maximum likelihood estimates (where '0' denotes negative, '1' positive and '2' neutral sentiment) showcased in Table 2.2.

Now, say we employ an existing debiasing approach and manage to debias the model with respect to its representation of group g_1 by setting $P(w_{g_1}|0) = P(w_{g_1}|1) = P(w_{g_1}|2)$, which, in this case, would equal zero. This would hurt performance on data from g_4 (bottom row), increasing the empirical risk on

	$P(w_{g_1}\|0)$	$P(w_{g_2}\|0)$	$P(w_{g_3}\|0)$	$P(w_{g_4}\|0)$	$P(w_5\|0)$	$P(w_6\|0)$	$P(w_7\|0)$
g_1	0.0	0.25	0.0	0.0	0.25	0.0	0.0
g_2	0.0	0.0	0.25	0.0	0.25	0.0	0.0
g_3	0.0	0.0	0.0	0.25	0.25	0.0	0.0
g_4	0.25	0.0	0.0	0.0	0.25	0.0	0.0

	$P(w_{g_1}\|1)$	$P(w_{g_2}\|1)$	$P(w_{g_3}\|1)$	$P(w_{g_4}\|1)$	$P(w_5\|1)$	$P(w_6\|1)$	$P(w_7\|1)$
g_1	0.0	0.0	0.0	0.0	0.0	0.25	0.0
g_2	0.0	0.0	0.0	0.0	0.0	0.25	0.0
g_3	0.0	0.0	0.0	0.0	0.0	0.25	0.0
g_4	0.0	0.0	0.0	0.0	0.0	0.25	0.0

	$P(w_{g_1}\|2)$	$P(w_{g_2}\|2)$	$P(w_{g_3}\|2)$	$P(w_{g_4}\|2)$	$P(w_5\|2)$	$P(w_6\|2)$	$P(w_7\|2)$
g_1	0.0	0.0	0.0	0.0	0.0	0.0	0.25
g_2	0.0	0.0	0.0	0.0	0.0	0.0	0.25
g_3	0.0	0.0	0.0	0.0	0.0	0.0	0.25
g_4	0.0	0.0	0.0	0.0	0.0	0.0	0.25

Table 2.2: Maximum likelihood estimates from a linear classifier on our synthetic data modelled in Table 2.1.

this sub-population, but more surprisingly, note that it would not help us on classifying the data from g_1. That is, an attempt to make the model fairer towards *north* by equalizing the use of the term *northern*, would result in increased unfairness towards members from *south*, who tend to use *northern* more often (and in a negative context). Removing bias in how terms referring to a group are represented, only improves performance on data from members from that group, if these members use such in-group terms in non-standard ways, *i.e.*, differently from everyone else. In the absence of this assumption, association bias and empirical fairness are orthogonal. We will refer to this assumption as the **In-Group Affinity Assumption**.

Note that while we make use of a linear model and likelihood estimates in our thought experiment, it would be very easy to translate this into a deep neural network and cosine distances instead. To see this, consider, for example, how any Naive Bayes model can be translated into a deep neural network, and how the differences in likelihood can, under such a translation, be translated into differences in cosine instances.

2.4 Association Bias and Empirical Fairness Scores are Uncorrelated (in Practice)

In this section, we study whether association bias and empirical fairness are correlated in practice,*i.e.*, when *actual* models are evaluated on *actual* data designed to probe bias and fairness. We apply well-established metrics for measuring the two. While bias and fairness can be studied with respect to any prospective attribute, the vast majority of NLP research has focused on (binary) gender Sun et al. (2019); Stanczak and Augenstein (2021). Binary gender is often correlated with terms referring to occupations, *e.g.*, the co-occurrence of *woman* and *man*—or *she* and *he*—in the context of *nurse* and *doctor*. For convenience, we rely on existing benchmarks and do the same. It is important to remember, however, that bias and fairness may arise across any groups in society, and that all those defined in terms of protected attributes, *e.g.*, race, religion, sexuality, or impairment, are legally irrelevant. As mentioned, the two—association bias and empirical fairness—are often conflated, or one is said to cause the other. This reflects an In-Group Affinity Assumption, saying that members of social groups refer to themselves more often or in different ways than other members of a linguistic community. If this were the case, mitigating biases would contribute positively to equal performance across groups.

The analysis of these experiments concludes our three-way investigation of the In-Group Affinity Assumption and the independence of bias and fairness. All three perspectives suggest that NLP research should not further assume an intimate connection between the two.

Bias To measure representational bias, we use three popular metrics, *i.e.*, the Log Probability Bias Score (LPBS) proposed by Kurita et al. (2019), as well as two variants of the Word Embedding Association Test (WEAT) Caliskan et al. (2017) for assessing bias in contextual word representations: the adaption proposed by Tan and Celis (2019) (henceforth, WEAT_T), and the alternative suggested by Lauscher et al. (2021) (henceforth, WEAT_L). All these metrics rely on association tests to compute the relationship between a set of related targets $\{t_1, t_2, \ldots\}$, *e.g.*, gender words, and attributes $\{a_1, a_2, \ldots\}$, *e.g.*, occupation words, through definitions of template sentences designed to convey no meaning beyond that of the terms inserted into them.

Kurita et al. (2019) use template sentences like T="[TARGET] is a [ATTRIBUTE]". The target word is masked, and the attribute word is a placeholder for a specific word denoting an occupation, *e.g.*, T_m="[MASK] is a *chef*". LPBS uses the prior probability of the target word (p_{prior}), *i.e.*, the probability of a target t_i being generated when both t_i and the attribute a_j are masked, as a normalizer, and computes the association as the relative increase in log probability:

$$a^{lpbs}_{t_i,a_j} = \log \frac{p([\text{MASK}] = t_i | T_m)}{p_{prior}} \tag{2.1}$$

The difference between the relative increased log probability scores for two targets is the LPBS measure of bias. For linear models, this correlates strongly with the ϵ-isometry of the target word subgraph relative to an equidistant space, if we make the centroid of the set of attribute vectors the reference point. For a non-linear language model, we can compute the ϵ-isometry of its linear approximation. Table 2.3 are for the targets "he" and "she". A t-test is used to evaluate the statistical significance of the metric, in which the means of a^{lpbs}_{he,a_j} and a^{lpbs}_{she,a_j} are compared. We draw 10^5 random permutations, meaning that the p-values observed will not be less than 10^{-5}.

Tan and Celis (2019) follow the methodology of May et al. (2019), who extended the WEAT metric to sentences (SEAT) inserting the word of interest in context templates such as T="This is _". Tan and Celis (2019) use the contextual embedding of the token of interest, instead of using the sentence encoding, to compute the cosine similarities (associations). Lauscher et al. (2021) follow Vulić et al. (2020) and average the pooled embeddings of the first four attention layers for the word of interest (t_i or a_j) in a template without context, e.g, "[CLS] t_i [SEP]". Both approaches report the *effect size* Caliskan et al. (2017), a normalized measure of how separated the association distributions of target and attributes are. The statistical significance of the associations is also computed with a permutation test as in Caliskan et al. (2017). Both approaches are an instance of computing the ϵ-isometry of the template sentence subgraphs in the cosine metric space. See Table 2.3 for empirical results.[3] We see that results are somewhat mixed, with LPBS and the two variants of WEAT often disagreeing which models are more biased. All the metrics are evaluated on the same list of sixty attributes —equally

[3]PLM names follow the same nomenclature as in the Hugging Face Transformers library. The pre-trained models can be downloaded at huggingface.co/models.

split into female and male stereotyped professions from the US bureau of labour—, provided in Delobelle et al. (2021).

	$LPBS$	$WEAT_L$	$WEAT_T$
bert-base-uncased	0.86*	1.01*	0.33
bert-base-cased	0.90*	1.00*	0.52
bert-large-uncased	0.20	0.83*	0.73*
bert-large-cased	-1.10*	0.60	0.83*
bert-base-multilingual-cased	-1.98*	0.36	0.12
distilbert-base-uncased	-0.46*	0.79*	0.58
albert-base-v2	-7.02*	0.72*	0.56
albert-large-v2	-1.58*	0.84*	0.61*
albert-xxlarge-v2	0.18	0.46	0.95*
roberta-base	-2.32*	0.51	0.36
roberta-large	-2.63*	0.24	0.82*
google/electra-small-generator	-0.20	0.71*	0.85*
google/electra-large-generator	-2.64*	0.73*	0.63*

Table 2.3: Three metrics of representational bias. Values are the average difference of associations between the target words "he"/"she", and a list of occupations as attributes. Larger values reflect a more severe bias. A positive value hints a skewed distribution towards males. A negative value hints a skewed distribution towards females. $*$: statistically significant at 0.01.

Fairness Our fairness evaluation is based on Zhang et al. (2021)'s work, who study how the predictions of various PLMs align with the linguistic preferences of different social groups. They directly compare masked word predictions to human cloze tests, quantifying how often a language model agrees with the members of a particular social group on what is the most likely word in contexts such as:

After waiting three hours, Cal whined and started to [MASK].

Zhang et al. (2021) use, as their fairness metric, the min-max difference in precision (ΔP@1) across groups defined by the cross-product of several protected attributes, including gender, age, race, and level of education. Since we are comparing with binary gender bias probes, we only consider fairness

across (binary) gender here. We sample members of each group (female and male) in a balanced way across subgroups, as defined by the other variables. This is equivalent to reporting the macro-average across subgroups for each group. ΔP@1 is thus the difference in performance between male and female groups, macro-averaged across subgroups in the cloze test data. We follow Zhang et al. (2021) in also reporting the difference in mean reciprocal rank as a second performance metric (ΔMRR). See the individual scores in Table 2.4.

	$\Delta P@1$	ΔMRR
bert-base-uncased	0.69	1.57
bert-base-cased	0.15	0.74
bert-large-uncased	0.91	1.34
bert-large-cased	-0.07	0.32
bert-base-multilingual-cased	0.89	0.54
distilbert-base-uncased	1.63	0.64
albert-base-v2	0.74	0.94
albert-large-v2	1.45	1.21
albert-xxlarge-v2	0.48	0.41
roberta-base	0.14	0.06
roberta-large	0.68	0.69
google/electra-small-generator	0.97	0.43
google/electra-large-generator	1.22	0.97

Table 2.4: Macro-averaged precision and mean reciprocal rank differences between male and female subgroups following experiments in Zhang et al. (2021). Values close to zero are preferred for a more equitable model.

Results show performance gaps between binary gender groups. Consequently, we would expect models exhibiting high degree of bias in Table 2.3 to be the least fair. However, this is not the case. Figure 2.2 displays the results for bias and fairness jointly, often highlighting the lack of correlation. Note that, ideally, all data-points should belong to the bottom-right quadrant.

Metrics are uncorrelated Now that we have our evaluation framework defined, let us analyze whether representational bias correlates with outcome

disparity. This amounts to studying the correlations between LPBS and WEAT metrics and the min-max P@1 difference across groups. We report the sign of the Pearson correlation coefficient to ease the interpretation of the (ideally) monotonic relationship[4] between each set of metrics in Figure 2.2.

Results are two-fold:

(i) The discrepancy across sub-graphs in Figure 2.2 aligns with results in May et al. (2019), Delobelle et al. (2021) and Cao et al. (2022), who all found different representational bias metrics to lead to mutually inconsistent results. WEAT_L and WEAT_T are related and show some agreement, but generally, results are wildly different across metrics.

(ii) More importantly, for our purposes, representational bias and fairness-as-equal-performance (quantified as min-max differences across performance scores for different groups) are, in fact, uncorrelated. Models with high bias values are the most fair according to our fairness metric, and vice versa. These cases are highlighted in red in Figure 2.2. For example, `roberta-base` (`rb`) is among the most biased models according to LPBS, but it exhibits the highest degree of fairness wrt. the MRR metric —and second highest wrt. P@1. The bigger PLM, `roberta-large` (`rl`) is slightly less biased according to LPBS, but it is generally less fair. Values from the WEAT metrics are, in this case, somewhat mixed.

Result (ii) is evidence *against* the In-Group Affinity Assumption and *for* the independence of bias and fairness. Looking at each model family—separated by horizontal lines in Table 2.3 and 2.4—model size does not systematically lead to larger or smaller bias scores, and it does not seem strongly correlated with any of the fairness metrics either.

In the following section, we survey research in the social sciences that also suggest the In-Group Affinity Assumption is mostly false, with one important caveat: Slur words have marked in-group usage. In most applications, this exception would be insufficient to drive a causal link between association bias and empirical fairness, because slur words are rare, and performance differences across social groups are pervasive.

[4]We deliberately omit the magnitude of the Pearson coefficient to emphasize the sign of the correlation. Ideally, bias and fairness metrics should have a negative *linear* dependence ($p < 0$).

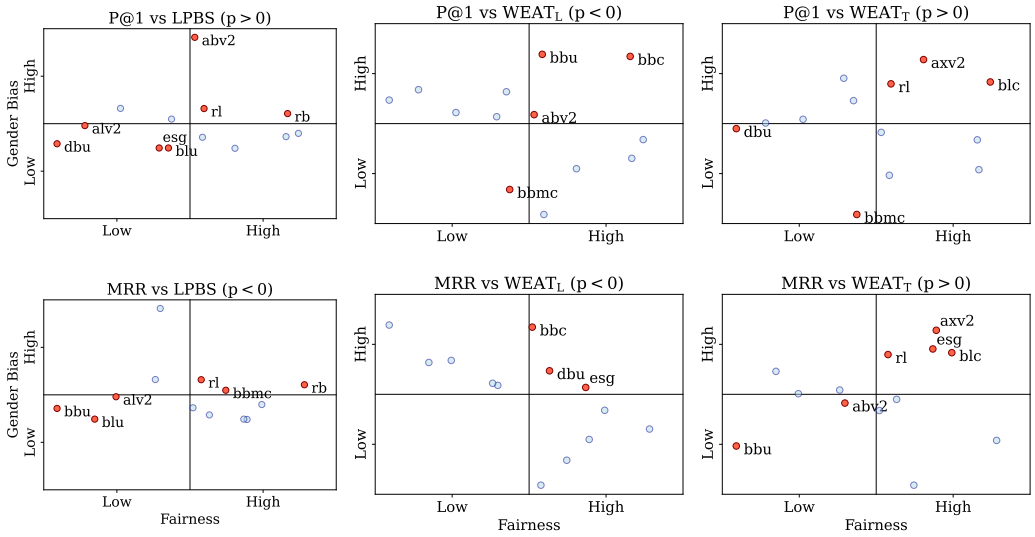


Figure 2.2: Scatter plots show the relationship between different representational bias metrics and fairness evaluation. The upper row displays results when evaluating fairness through precision at top-1 (P@1). The bottom row displays results when considering MRR to evaluate fairness. The division into quadrants is done according to average scores. Each point represents a language model, labelled with its initials. We see no support for a strong negative correlation between bias and fairness. Red points mark the clear counter-examples to such a negative correlation. Global trend for each plot is summarized with the sign of Pearson coefficient (p).

2.5 Association Bias and Empirical Fairness are Sometimes at Odds (in Humans)

The thought experiment in §2.3 shows that bias and fairness can in fact be completely independent or orthogonal. The experiments in §2.4 further showed that there is no direct correlation between the association bias in a model $\mathbf{M}$ toward social groups $g_1, \ldots, g_n$, and the performance disparity (fairness) of $\mathbf{M}$ across data from these groups $g_1, \ldots, g_n$.

In such cases, debiasing a model with respect to the representation of a certain group (*e.g.*, g_1) has no impact on the performance of the model for users from the group. The beneficiaries of such a debiasing procedure are, in other words, not necessarily the group the debiasing was intended to increase fairness for. The idea that debiasing word representations that are related to a particular group increases the fairness of the model for that group, relies on the assumption that those words are also used by the in-group more frequently or in other ways than by other groups. This assumption—which we called the In-Group Affinity Assumption—seems problematic, since there are plenty of examples in the literature of the opposite. In the following, we briefly review some examples that originate from the NLP literature; others from the social sciences.

We are often likely to talk more about members of other groups than our in-group peers. Li and Dickinson (2017), for example, find that some of the most indicative n-grams for detecting young female users on Chinese social media are the names of male pop stars. Correcting or debiasing the representations of these names would not improve model fairness on texts written by the male pop stars, but rather on texts written by young female users.

Morgan-Lopez et al. (2017) show that young (pre-college) children talk more about college on Twitter than adults in their college age. Wei and Santos Jr. (2020) analyze data from Twitter and Reddit and find that the most predictive n-grams for Israeli users include "Iraqis" and "Palestinians", while for Palestinian users "israeli military detention centres" and "Lieberman settler rabbis" (referring to the Israeli Defence Minister, Avigdor Lieberman) are among the most predictive n-grams.

Generally, political debates are often experienced as negative in both tone and nature. According to a 2019 Pew Research Center study, 85% of

Americans say that the political debate has become "more negative". [5] One explanation for the increase in negative sentiment in political discourse is increased attention to what members of other (political) groups do wrong compared to what the in-group peers do right. Supporting this explanation, Jensen et al. (2012) show, for example, that one of the most partisan phrases used by US Democrats in congressional texts was "great Republican Party".

Similarly, Duijnhoven (2018) finds that Democrats on Twitter mention Trump and the Republican party more often than their Republican counterparts. In analyzing the language of German political parties, Biessmann (2016) likewise finds that the left-wing party, Linke, has a high frequency of mentions of large corporations (*konzerne*) and policies that negatively impact the social welfare.

On Slur Some slurring terms (*e.g.*"dyke", "queer" and "bitch") have been reclaimed or reappropriated by the target group resulting in a semantic discrepancy dependent on the speaker's group membership Ritchie (2017); Henry et al. (2014). This results in what we term the In-Group Affinity Assumption, where the in-group's use of the term will differ significantly from that of the out-groups. Any debiasing of the term will have no significant impact on the performance for the in-groups, since the language model's representation of the term will reflect the majority use of the term, which will not be that of the in-group. However, since slurs are per definition defamatory terms, debiasing these terms will result in less insulting outputs in downstream tasks, and this may result in a higher perception of fairness for the target group.

2.6 Discussion and Conclusion

The independence of representational bias and fairness-as-equal-performance shown here, along with the falsification of the In-Group Affinity Assumption, runs counter to the NLP literature. Bias and fairness have been assumed to be intimately connected, and the In-Group Affinity Assumption has been implicit and unquestioned in much recent work. The results we present in this paper are, at the same time, in a sense not surprising. Or they *should* not be surprising. In many aspects of private and public life, we encounter

[5] pewresearch.org/politics/2019/06/19/public-highly-critical-of-state-of-political-discourse-in-the-u-s/

decisions or patterns where bias and fairness exist or fluctuate independently of each other, or in which they are negatively correlated. In affirmative action, for example, we tolerate and encourage a (more) biased decision-making process to achieve (higher) fairness. While positive discrimination is heavily debated Holzer and Neumark (2000); Barmes (2009); Noon (2010), it is a good example of a biased process intended to increase the level of fairness.

Methods for correctly assessing model biases remains an open research question. Current evaluation benchmarks give inconsistent results May et al. (2019); Delobelle et al. (2021); Cao et al. (2022). Moreover, as discussed in §2.2, evaluating model biases with metrics that only consider local geometries, such as cosine-based metrics, can be inadequate. The fairness metric literature is also full of controversies Miconi (2017); Friedler et al. (2016); Kleinberg et al. (2016); Hedden (2021), but there is a broad consensus that performance disparity or outcome disparity is a real challenge for responsible NLP research and development. This consensus is not only limited to NLP research, but also found in legal studies, machine ethics, and the social sciences. Our results have shown that regardless of these open problems in bias and fairness research, the assumption that bias and fairness are always negatively correlated, and that one is a cause of the other, is not always true. Despite being closely related, it is important to understand that biases exist everywhere, but might not be unequivocally harmful. And similarly, fairness issues may arise in non-biased scenarios.

Finally, it is worth noting that we should not solely focus on the correlation between protected attributes such as race or gender and the model's output, but rather ask the question if *they* are causing the outcome, and, whether the model is unfair to individuals in virtue of their membership in a certain group Hedden (2021).

Conclusion We reviewed part of the NLP literature showing how many researchers conflate bias and fairness, *i.e.*, representational bias and fairness-as-equal-performance, or argue that fixing one will solve the other. In an attempt to explain why this does not hold always true, we devised a thought experiment in §2.3: a synthetic model that illustrates how bias and fairness can be completely independent of one another. We introduced the In-Group Affinity Assumption to highlight the assumption that a particular demographic groups use in-group terms more frequently—or in different

ways—than other groups (non-standard). This, we argue, is a necessary assumption to drive a causal connection between bias and fairness, if it exists. In §2.5, we surveyed the social science literature and found evidence that often the opposite is the case, which substantiates our findings in §2.3 and §2.4. Our survey includes examples from the social sciences, as well as from NLP research, where bias and fairness are (locally) *negatively* correlated. This provides strong reason to be skeptical of the In-Group Affinity Assumption and shows that bias and fairness are often independent or orthogonal to each other.

In sum, we have shown the importance of studying bias and fairness independently of one another and cautioned against the In-Group Affinity Assumption. We think this, potentially, could lead to a valuable reorientation of the NLP literature, enabling researchers to study representational bias in more adequate ways, focusing on robustness and generation (to avoid bias reinforcement). This also highlights the different contributions of representational bias benchmarks and in-the-wild evaluation datasets with demographic information that can be used to evaluate performance disparities across groups. Bias and fairness seem to be separate issues, and we believe research should be done by disentangling the two.

2.7 Limitations

Our paper addresses the relationship between the two specific interpretations of bias and fairness, *i.e.*, representational bias and fairness-as-equality. These are, in our view, the most common and most important definitions of bias and fairness in the NLP literature, but they are not the only ones. We hope others will follow up with studies of how other definitions relate. Our experiments in §3 were limited to English benchmark datasets. We agree with Ruder et al. (2022) that the prevalence of bias and fairness studies using English data, is most unfortunate, and we are, in parallel, working to create multilingual benchmarks for bias and fairness studies.

Chapter 3

Are Pretrained Multilingual Models Equally Fair across Languages?

Abstract

Pretrained multilingual language models can help bridge the digital language divide, enabling high-quality NLP models for lower-resourced languages. Studies of multilingual models have so far focused on performance, consistency, and cross-lingual generalisation. However, with their wide-spread application in the wild and downstream societal impact, it is important to put multilingual models under the same scrutiny as monolingual models. This work investigates the group fairness of multilingual models, asking whether these models are equally fair across languages. To this end, we create a new four-way multilingual dataset of parallel cloze test examples (MozArt), equipped with demographic information (balanced with regard to gender and native tongue) about the test participants. We evaluate three multilingual models on MozArt —mBERT, XLM-R, and mT5— and show that across the four target languages, the three models exhibit different levels of group disparity, *e.g.*, exhibiting near-equal risk for Spanish, but high levels of disparity for German.

3.1 Introduction

Fill-in-the-gap cloze tests Taylor (1953) ask language learners to predict what words were removed from a text and it is a "procedure for measuring the effectiveness of communication". Today, language models are trained to do the same Devlin et al. (2019). This has the advantage that we can now use fill-in-the-gap cloze tests to directly compare the linguistic preferences of humans and language models, *e.g.*, to investigate task-independent sociolectal biases (group disparities) in language models Zhang et al. (2021). This paper presents a novel four-way parallel cloze dataset for English, French, German, and Spanish that enables apples-to-apples comparison across languages of group disparities in multilingual language models.[1]

Language models induced from historical data are prone to implicit biases Zhao et al. (2017b); Chang et al. (2019); Mehrabi et al. (2021), *e.g.*, as a result of the over-representation of male-dominated text sources such as Wikipedia and newswire Hovy and Søgaard (2015). This may lead to language models that are *unfair* to groups of users in the sense that they work better for some groups rather than others Zhang et al. (2021). Multilingual language models can be unfair to their training languages in similar ways Choudhury and Deshpande (2021); Wan (2022); Wang et al. (2022b), but this work goes beyond previous work in evaluating whether multilingual language models are *equally fair to demographic groups across languages*.

To this end, we create MozArt, a multilingual dataset of fill-in-the-gap sentences covering four languages (English, French, German and Spanish). The sentences reflect diastratic variation within each language and can be used to compare biases in pretrained language models (PLMs) across languages. We study the influence of four demographic groups, *i.e.*, the cross-product of our annotators' gender —male (*M*) or female (*F*)[2]— and first language —native (*N*) or non-native (*NN*)[3]—. Table 3.1 presents a summary of dataset characteristics.

[1]The language selection was given to us, because we rely on an existing word alignment dataset; see §2.

[2]None of our annotators identified as non-binary.

[3]See Schmitz (2016); Faez (2011) for discussion of the native/non-native speaker dichotomy. Participants were asked "What is your first language?" and "Which of the following languages are you fluent in?". We use *native* (*N*) for people whose first language coincides with the example sentences, and non-native (*NN*) otherwise, without any sociocultural implications.

	EN	ES	DE	FR
WordPiece (avg. #tokens)	19.7	22.0	23.6	23.1
SentencePiece (avg. #tokens)	22.3	22.9	24.9	25.3
#Sentences	100	100	100	100
#Annotations	600	600	600	600
#Annotators	60	60	60	60
Demographics	id_u, id_s, gender, age, nationality, first language, fluent languages, current country of residence, country of birth, time taken			

Table 3.1: MozArt details. The average number of tokens per sentence is reported using WordPiece and SentecePiece. The bottom row lists the demographic attributes shared; id_u refers to user id (anonymised) and id_s to sentence id.

3.2 Dataset

We introduce MozArt, a four-way multilingual cloze test dataset with annotator demographics. We sampled 100 sentence quadruples from each of the four languages (English, French, German, Spanish) in the corpus provided for the WMT 2006 Shared Task.[4] The data was extracted from the publicly available Europarl corpus Koehn (2005) and enhanced with word-level bitext alignments Koehn and Monz (2006). The word alignments are important for what follows. We manually verify that sentences make sense out of context and use the data to generate *comparable cloze examples*, *e.g.*:

en [MASK] that deplete the ozone layer
es [MASK] que agotan la capa de ozono
de [MASK], die zum Abbau der Ozonschicht führen
fr [MASK] appauvrissant la couche d'ozone

We only mask words which are (i) aligned by one-to-one alignments, and which are (ii) either nouns, verbs, adjectives or adverbs.[5] We mask one word

[4] www.statmt.org/wmt06/shared-task

[5] We use spaCy's part-of-speech tagger Honnibal and Montani (2017) to predict the syntactic categories of the input words.

in each sentence and verify that one-to-one alignments exist in all languages. Following Kleijn et al. (2019), we rely on part-of-speech information to avoid masking words that are *too* predictable, *e.g.*, auxiliary verbs or constituents of multi-word expressions, or words that are *un*-predictable, *e.g.*, proper names and technical terms.

Annotators were recruited using Prolific.[6] We applied eligibility criteria to balance our annotators across demographics. Participants were asked to report (on a voluntary basis) their demographic information regarding gender and languages spoken. Each eligible participant was presented with 10 cloze examples. We collected answers from 240 annotators, 60 per language batch, divided in four balanced demographic groups (gender $\times$ native language). We made sure that each sentence had at least six annotations. Annotation guidelines for each language were given in that language, to avoid bias and ensure a minimum of language understanding for non-native speakers. We manually filtered out spammers to ensure data quality.

The dataset is made publicly available at github.com/coastalcph/mozart under a CC-BY-4.0 license. We include all the demographic attributes of our annotators as per agreement with the annotators. The full list of protected attributes is found in Table 3.1. We hope MozArt will become a useful resource for the community, also for evaluating the fairness of language models across other attributes than gender and native language.

3.3 Experimental Setup

Models We evaluate three PLMs: mBERT Devlin et al. (2019), XLM-RoBERTa/XLM-R Conneau et al. (2020), and mT5 Xue et al. (2021).[7] All three models were trained with a masked language modelling objective. mBERT differs from XLM-R and mT5 in including a next sentence prediction objective Devlin et al. (2019). mT5 differs from mBERT and XLM-R in allowing for consecutive spans of input tokens to be masked Raffel et al. (2020). We adopt beam search decoding with early stopping and constrain the generation to single words. This enables better correlation of mT5's output with our group preferences. t-SNE plots are included in Appendix 3.8.2 to show how languages are distributed in the PLM vector spaces.

[6] prolific.co

[7] We use the base models available from huggingface.co/models. We report results using uncased mBERT, since it performed better on our data than its cased sibling.

Metrics We use several metrics to compare how the PLMs align with group preferences across languages. These include top-k precision P@k with k={1, 5}, mean reciprocal rank (MRR), and two classical univariate rank correlations: Spearman's ρ Spearman (1987) and Kendall's τ Kendall (1938).

Given a set of $|S|$ cloze sentences and a group of annotators, for each sentence s, we denote the list of answers, ranked by their frequency, as $\mathrm{W_s} = [w_1, w_2, ...]$, and the list of model's predictions as $\mathrm{C_s} = [c_1, c_2, ...]$, ranked by their model likelihood. Then, we report $\mathrm{P@k} = \mathbb{1}[c_i \in \mathrm{W_s}]$ with $i \in [1, k]$, where $\mathbb{1}[\cdot]$ is the indicator function. Precision is reported together with its standard deviation, to account for the group-wise disparity in both dimensions (social groups and language):

$$\sigma_{\mathrm{gd}} = \sqrt{\frac{\sum_{j=1}^{G}(\mathrm{P@k}_j - \overline{\mathrm{P@k}})^2}{\mathrm{G}}} \tag{3.1}$$

where $\overline{\mathrm{P@k}}$ is the mean value of all observations, and G the total number of groups across the dimension fixed each time *i.e.*, $\mathrm{G} = 4$ across social groups (*MN, FN, MNN, FNN*) and $\mathrm{G} = 4$ across languages (EN, ES, DE, FR). We also compute the mean-reciprocal rank (MRR) of the elements of $\mathrm{W_s}$ with respect to the top-n ($\mathrm{n} = 5$) elements of $\mathrm{C_s}$ ($\mathrm{C_s^n}$):

$$\mathrm{MRR} = \frac{1}{|S|}\sum_{s=1}^{|S|}\frac{1}{Rank_i^{C_s^n}} \tag{3.2}$$

Finally, we compute Spearman's ρ Spearman (1987) and Kendall's τ Kendall (1938) between $\mathrm{W_s}$ and $\mathrm{C_s^5}$. These metrics are generally more robust to outliers.

3.4 Results

Following previous work on examining fairness of document classification Huang et al. (2020); Dixon et al. (2018); Park et al. (2018); Garg et al. (2019), we focus on group-level performance differences (group disparity). We measure the group disparity as the variance in PLM's performance (P@k) across demographics (gender and native language). Table 3.2 shows better precision for native speakers in German and French (*MN*, *FN*) for P@1. In terms of group disparity, male non-natives (*MNN*) is the demographic

mBERT

P@1	EN	ES	DE	FR	
MN	13.3	12.7	11.3	10.7	12.0 (1.0)
FN	13.3	12.0	15.3	8.0	12.2 (2.7)
MNN	12.7	12.4	11.4	3.6	10.0 **(3.8)**
FNN	13.3	10.0	5.6	6.9	9.0 (3.0)
	13.2 (0.3)	11.8 (1.1)	10.8 **(3.5)**	7.3 (2.5)	$\overline{\text{P@1}}(\sigma_{gd})$

XLM-R

P@1	EN	ES	DE	FR	
MN	16.7	13.3	20.7	16.7	16.9 (2.6)
FN	16.0	15.3	24.0	17.3	18.2 **(3.5)**
MNN	15.3	13.5	15.0	11.4	13.8 (1.5)
FNN	20.0	14.7	13.1	12.7	15.1 (3.0)
	17.0 (1.8)	14.2 (0.8)	18.2 **(4.4)**	14.5 (2.6)	$\overline{\text{P@1}}(\sigma_{gd})$

mT5

P@1	EN	ES	DE	FR	
MN	2.0	4.7	8.7	5.3	5.2 **(2.4)**
FN	4.0	3.3	6.7	3.3	4.3 (1.4)
MNN	2.0	4.7	6.4	4.3	4.4 (1.6)
FNN	3.3	6.7	1.9	6.2	4.5 (2.0)
	2.8 (0.9)	4.8 (1.2)	5.8 **(2.5)**	4.8 (1.1)	$\overline{\text{P@1}}(\sigma_{gd})$

Table 3.2: Results on P@1 score across groups (rows) and languages (columns), average performance in each language ($\overline{\text{P@1}}$) and standard deviation for group disparity (σ_{gd}). Cells are coloured language-wise. Cells with a darker background are language-wise above the average. Worst group performance in terms of group disparity (highest variance) is highlighted in red.

exhibiting the highest disparity across languages in mBERT, while it is female natives (*FN*) in XLM-R and male natives (*MN*) in mT5. Language-wise, we see the largest group disparity with German in all three models. Here, we see 2.5–4.4 between-group differences, compared to, *e.g.*, 0.3–1.8 between-group differences for English. See Appendix 3.8.1 for results with P@5.

XLM-R consistently exhibits better overall performance on average, but higher between-group and between-language differences in terms of precision (σ_{gd}).

Figure 3.1 complements results from Table 3.2 with MRR scores. We observe a common trend that the models often underperform on non-native

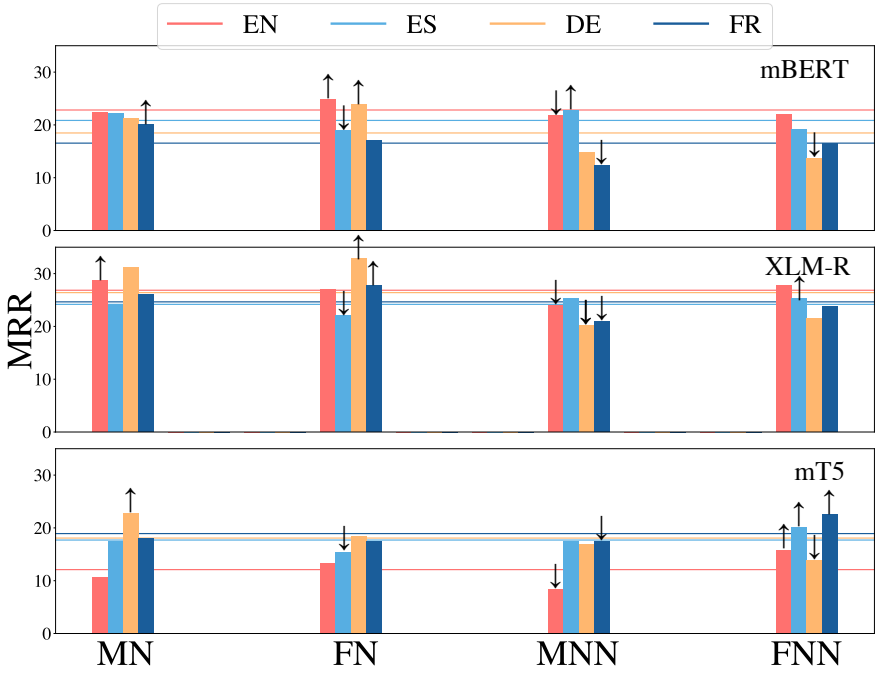


Figure 3.1: Average MRR (in percentage) per group in each language. Horizontal lines denote the average per language. Best-off (↑) and worst-off (↓) subgroups for each language are marked.

male speakers in all languages except for Spanish: Performance is (always) below the average, and they are the worst-off group (↓) in most of the cases. At the same time, predictions with mBERT and XLM-R seem to be biased towards native speakers because answers from *MN* and *FN* generally rank highest. Despite none of the models perform equally across groups, XLM-R shows a lower divergence across languages: Between-group differences are more than 50% smaller than with mBERT and mT5 when looking at the average MRR per language.

Table 3.3 gathers group level Spearman's ρ and average correlation per language. XLM-R predictions are more uniformly correlated across languages compared to mBERT, whose lexical preferences are better aligned in English and Spanish setups, and mT5, whose predictions correlate poorly with human cloze test answers. However, in line with previous results, the model exhibits bias towards male native speakers and *MNN* outlines as the worst performing group across languages, with a coefficient always below the average. Looking

mBERT

ρ	EN	ES	DE	FR
MN	0.33 (p=0.00)	0.23 (p=0.01)	-0.14 (p=0.09)	0.10 (p=0.21)
FN	0.27 (p=0.00)	0.07 (p=0.42)	-0.01 (p=0.89)	0.14 (p=0.08)
MNN	0.30 (p=0.00)	0.16 (p=0.03)	-0.10 (p=0.23)	0.08 (p=0.32)
FNN	0.37 (p=0.00)	0.16 (p=0.06)	0.03 (p=0.69)	0.08 (p=0.30)
Avg.	0.32 (p=0.00)	0.16 (p=0.00)	-0.05 (p=0.21)	0.10 (p=0.01)

XLM-R

ρ	EN	ES	DE	FR
MN	0.45 (p=0.00)	0.46 (p=0.00)	0.35 (p=0.00)	0.48 (p=0.00)
FN	0.30 (p=0.00)	0.35 (p=0.00)	0.45 (p=0.00)	0.33 (p=0.00)
MNN	0.30 (p=0.00)	0.38 (p=0.00)	0.22 (p=0.01)	0.32 (p=0.00)
FNN	0.40 (p=0.00)	0.48 (p=0.00)	0.11 (p=0.16)	0.36 (p=0.00)
Avg.	0.36 (p=0.00)	0.41 (p=0.00)	0.28 (p=0.00)	0.37 (p=0.00)

mT5

ρ	EN	ES	DE	FR
MN	0.01 (p=0.89)	0.14 (p=0.08)	0.14 (p=0.08)	0.25 (p=0.00)
FN	-0.12 (p=0.13)	0.13 (p=0.12)	0.00 (p=0.99)	0.14 (p=0.08)
MNN	-0.10 (p=0.22)	0.12 (p=0.11)	0.03 (p=0.74)	0.11 (p=0.18)
FNN	-0.07 (p=0.41)	0.28 (p=0.00)	0.04 (p=0.58)	0.11 (p=0.16)
Avg.	-0.07 (p=0.07)	0.17 (p=0.00)	0.05 (p=0.23)	0.15 (p=0.00)

Table 3.3: Correlation between groups of annotators (*MN, FN, MNN, FNN*) and models' predictions, classified by language. The degree of correlation is measured with Spearman's ρ coefficient ($\rho \in [-1, 1]$). Cells are coloured language-wise. Cells with a darker background show a stronger correlation compared to the average in each language. Samples highlighted in red fail to reject the null hypothesis, meaning that their difference is not statistically significant ($p > 0.05$).

into the dimension of languages, German is the least aligned with human's answers in all models. Kendall's τ yields similar results. See Appendix 3.8.1 for details.

It is worth mentioning that our study does not aim to compare models' performance, but rather to motivate a discussion about the between-group and between-language differences within each model. The general low precision of mT5 outputs compared to human answers is likely due to the nature of the task itself. Because mT5 was trained with a span-mask denoising objective, it tends to complete the masked-out span with more than one token. When constraining generation to output one token, we are conditioning its default behaviour. Better correlation could be achieved by fine-tuning the

model on completing cloze tests.

(Dis)agreement amongst annotators on the same language gives a measure of the difficulty of the task. French and German present a higher variability in the responses (with a vocabulary of 442 and 443 words respectively), compared to English (374 words), and Spanish (427 words), which reflects in a lower correlation with models' predictions.

3.5 Related Work

Multilingual PLMs have been analyzed in many ways: Researchers have, for example, looked at performance differences across languages Singh et al. (2019); Wu and Dredze (2020), looked at their organization of language types Rama et al. (2020), used similarity analysis to probe their representations Kudugunta et al. (2019), and investigated how learned self-attention in the Transformer blocks affects different languages Ravishankar et al. (2021).

Previous work on fairness of multilingual models has, to the best of our knowledge, focused exclusively on task-specific models, rather than PLMs: Huang et al. (2020) evaluate the fairness of multilingual hate speech detection models, and several researchers have explored gender bias in multilingual models Zhao et al. (2020); González et al. (2020). Dayanik and Padó (2021) consider the effects of adversarial debiasing in multilingual models.

Cloze tests were previously used in Zhang et al. (2021) to evaluate the fairness of English (monolingual) language models. In psycholinguistics, cloze tests have been performed with different age groups Hintz et al. (2020) and native language Stringer and Iverson (2020), but these datasets have, to the best of our knowledge, not been used to evaluate language models.

3.6 Conclusion

In this paper, we present MozArt, a new multilingual dataset of parallel cloze examples with annotations from balanced demographics. This dataset is, to the best of our knowledge, the first to enable apples-to-apples comparison of group disparity of multilingual PLMs across languages. The dataset includes several demographic attributes, but we present preliminary experiments with gender and native language. We show that mBERT, XLM-R and mT5 are not equally fair across languages. For example, group disparities are much

higher for German (and French) than for English and Spanish. This shows the importance of evaluating fairness across languages instead of stipulating from results for a single language. We further show that cloze test answers of female native speakers tend to rank highest in both predictive PLMs. We followed best practices for mitigating the dangers of crowdsourcing Karpinska et al. (2021); Kleijn et al. (2019) (see §2) and hope MozArt will be widely adopted and, over time, generate more results for other languages, PLMs and demographic attributes.

3.7 Limitations

As described in the paper, MozArt builds on top of another dataset, which is only available in four languages. The original dataset with its manual word alignments provided a unique opportunity to build MozArt in a way in which we could account for context, across languages. This of course limits our work to the languages provided. We acknowledge how multilingual studies of Indo-European languages may not generalize to languages outside this language family, and hope we or others will be able to contribute resources for a more diverse set of languages in the future.

Ethics Statement

The dataset released contains publicly available content from the proceedings of the European Parliament. Our work is based on sensitive information provided by the participants that took on our study in Prolific. The protected attributes collected are self-reported on a voluntary basis, and participants gave their consent to share them. In addition to the specific attributes analyzed in our study, which served as prescreening filters, Prolific also provides baseline data for all studies with the consent of participants to share it with researchers. For these base attributes, there might be gaps in the data because it is optional for participants to provide this information. These attributes are filled as *null* in the dataset. We performed a pilot study to determine the amount of time a task would take on average. The participants were paid based on time worked, and were given the option to opt out at any time of the study. Participants who revoked consent at any stage are not included in our study nor in the data released.

3.8 Appendix

3.8.1 Additional results

In this section, we provide additional analysis results of the PLM's performance on MozArt. We report precision at 5 (P@5), which corresponds to the number of relevant answers amongst the top 5 candidates. It provides a more flexible metric for measuring model alignments with open-ended text answers, but fails to take into account the exact position within the top-k. Considering the top-5, the bias towards native speakers is diminished especially in English and Spanish, despite being *MNN* and *FNN* the worst groups —in terms of group disparity— in mBERT and XLM-R respectively. At the same time, the group disparities are exacerbated as shown in Table 3.4.

Table 3.5 complements results on correlation of the alignment of group responses. It shows Kendall's τ coefficient. Conclusions remain almost the same as studied with Spearman's coefficient, albeit non-native subgroups in Spanish are more correlated in mBERT.

mBERT

P@5	**EN**	**ES**	**DE**	**FR**	
MN	30.7	26.7	22.0	24.0	25.9 (3.3)
FN	32.0	18.7	24.7	22.0	24.4 (4.9)
MNN	34.0	25.9	12.1	15.0	21.8 **(8.7)**
FNN	32.7	25.3	16.3	16.3	22.7 (6.9)
	32.3 (1.2)	24.2 (3.1)	18.8 **(4.9)**	19.3 (3.8)	$\overline{\text{P@5}}(\sigma_{gd})$

XLM-R

P@5	**EN**	**ES**	**DE**	**FR**	
MN	39.3	30.7	34.7	32.7	34.4 (3.2)
FN	30.7	25.3	38.0	35.3	32.3 (4.8)
MNN	30.7	29.4	22.1	25.4	26.9 (3.4)
FNN	36.7	34.0	19.4	26.9	29.3 **(6.7)**
	34.3 (3.8)	29.8 (3.1)	28.5 **(7.9)**	30.3 (4.1)	$\overline{\text{P@5}}(\sigma_{gd})$

mT5

P@5	**EN**	**ES**	**DE**	**FR**	
MN	10.0	12.7	16.0	11.3	12.5 (2.2)
FN	11.3	10.0	16.7	18.0	14.0 **(3.4)**
MNN	6.0	11.8	9.3	10.7	9.5 (2.2)
FNN	13.3	16.0	8.7	15.0	13.3 (2.8)
	10.2 (2.7)	12.6 (2.2)	12.7 **(3.7)**	13.8 (3.0)	$\overline{\text{P@5}}(\sigma_{gd})$

Table 3.4: Results on P@5 score across groups and languages, average performance in each language ($\overline{\text{P@5}}$) and standard deviation for group disparity (σ_{gd}). Cells are coloured language-wise. Cells with a darker background are language-wise above the average. Worst group performance in terms of group disparity (highest variance) is highlighted in red.

mBERT

τ	EN	ES	DE	FR
MN	0.27 (p=0.00)	0.19 (p=0.00)	-0.09 (p=0.15)	0.09 (p=0.16)
FN	0.23 (p=0.00)	0.07 (p=0.24)	0.01 (p=0.89)	0.13 (p=0.04)
MNN	0.25 (p=0.00)	0.15 (p=0.01)	-0.06 (p=0.32)	0.07 (p=0.28)
FNN	0.29 (p=0.00)	0.14 (p=0.01)	0.03 (p=0.57)	0.06 (p=0.27)
Avg.	0.26 (p=0.00)	0.14 (p=0.00)	-0.03 (p=0.41)	0.09 (p=0.01)

XLM-R

τ	EN	ES	DE	FR
MN	0.40 (p=0.00)	0.43 (p=0.00)	0.32 (p=0.00)	0.45 (p=0.00)
FN	0.26 (p=0.00)	0.33 (p=0.00)	0.43 (p=0.00)	0.31 (p=0.00)
MNN	0.26 (p=0.00)	0.35 (p=0.00)	0.20 (p=0.01)	0.29 (p=0.00)
FNN	0.35 (p=0.00)	0.45 (p=0.00)	0.10 (p=0.15)	0.34 (p=0.00)
Avg.	0.32 (p=0.00)	0.39 (p=0.00)	0.25 (p=0.00)	0.34 (p=0.00)

mT5

τ	EN	ES	DE	FR
MN	0.02 (p=0.79)	0.13 (p=0.06)	0.13 (p=0.06)	0.21 (p=0.00)
FN	-0.09 (p=0.16)	0.11 (p=0.11)	0.00 (p=0.98)	0.12 (p=0.08)
MNN	-0.08 (p=0.21)	0.10 (p=0.10)	0.03 (p=0.69)	0.10 (p=0.17)
FNN	-0.04 (p=0.51)	0.25 (p=0.00)	0.03 (p=0.61)	0.10 (p=0.15)
Avg.	-0.07 (p=0.07)	0.15 (p=0.00)	0.05 (p=0.18)	0.13 (p=0.00)

Table 3.5: Correlation between groups of annotators (*MN, FN, MNN, FNN*) and models' predictions, classified by language. The degree of correlation is measured with Kendall's τ coefficient ($\tau \in [-1, 1]$). Cells are coloured language-wise. Cells with a darker background show a stronger correlation compared to the average in each language. Samples highlighted in red fail to reject the null hypothesis, meaning that their difference is not statistically significant ($p > 0.05$).

3.8.2 t-SNE

To give a brief overview of the semantic multilinguality encoded in the pretrained models, we run several representations with t-SNE. Figure 3.2 and Figure 3.3 represent the top-1000 predictions in a t-SNE plot for mBERT and XLM-R respectively. The same sentence is queried to the model in four languages and, accordingly, to annotators:

en	We want to [MASK] innovation .
es	Queremos [MASK] la innovación .
de	Wir wollen zur Innovation [MASK] .
fr	Nous voulons [MASK] l'innovation .

Highest scored predictions are highlighted with a ($\star$). Annotator's answers that fell into the top-1000 predictions are denoted with a black edge. In line with results in Choenni and Shutova (2020), we observe that languages are mostly projected in separate sub-spaces instead of yielding a neutral representation, even though they share a common space (vocabulary).

Similarly, Singh et al. (2019) shown a trend towards dissimilarity between representations for semantically similar inputs in different languages, in deeper layers of an uncased mBERT. Serve Figure 3.4 as an example, where the same word "gases" was answered in different languages but is represented in different subspaces. Figure 3.5 shows a similar behaviour in XLM-R. The sentences queried are:

en	[MASK] that deplete the ozone layer
es	[MASK] que agotan la capa de ozono
de	[MASK], die zum Abbau der Ozonschicht führen
fr	[MASK] appauvrissant la couche d'ozone

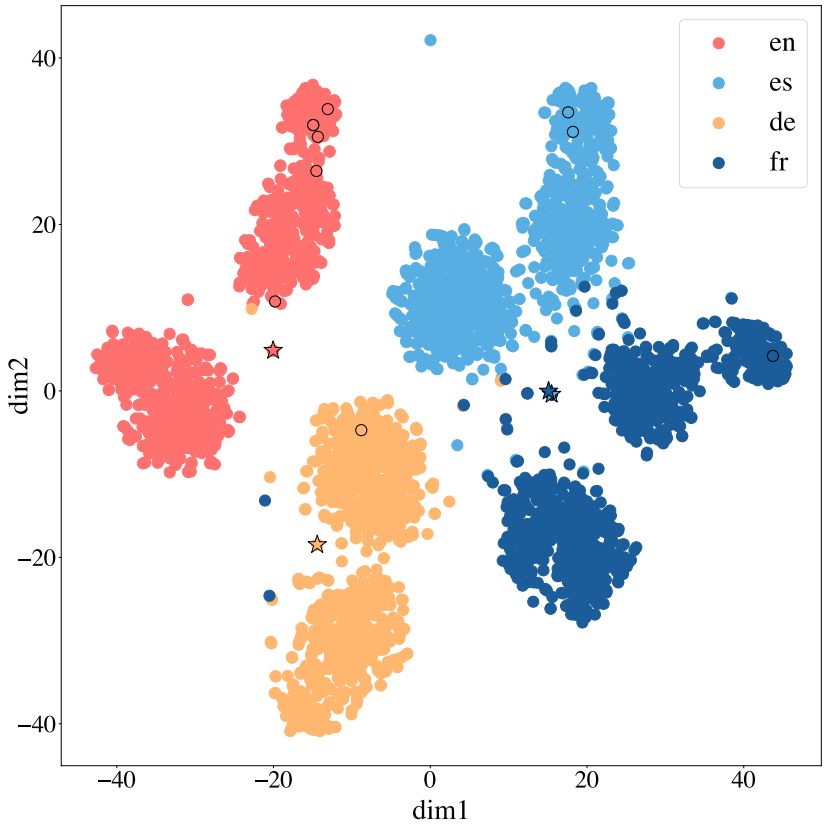


Figure 3.2: t-SNE representation from the last layer of mBERT for the top-1000 predictions for the parallel sentences in the list above ("We want to [MASK] innovation ." in English). Highest scored prediction is starred; annotator's answers are denoted by a dot with black edge. Legend shows language-color mapping.

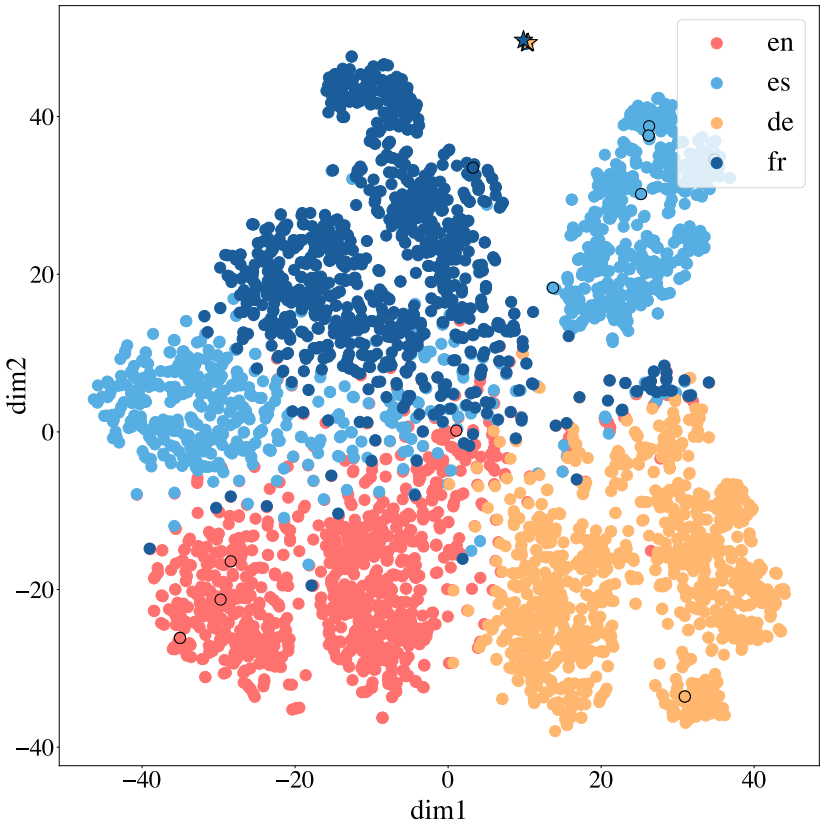


Figure 3.3: t-SNE representation from the last layer of XLM-R for the top-1000 predictions for the parallel sentences in the list above ("We want to [MASK] innovation ." in English). Highest scored prediction is starred; annotator's answers are denoted by a dot with black edge. Legend shows language-color mapping.

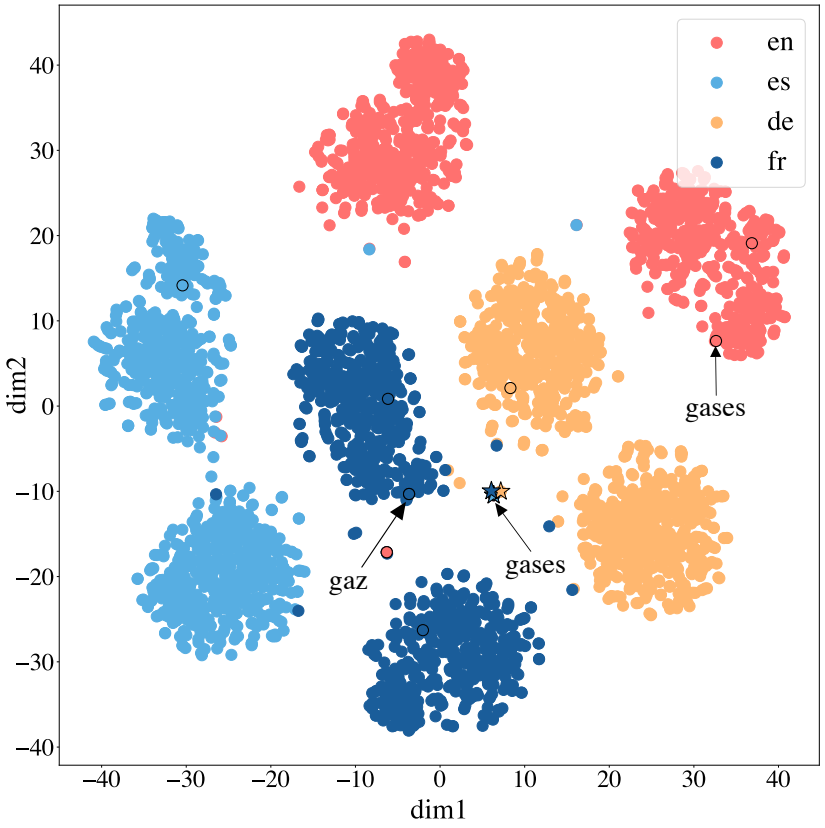


Figure 3.4: t-SNE representation from the last layer of mBERT for the top-1000 predictions for the parallel sentences in the list above ("[MASK] that deplete the ozone layer" in English). The word "gases" is pointed out in each language (en: gases, es: gases, fr: gaz), as it was a recurrent answer from different annotators. Highest scored prediction in each language is starred; annotator's answers are denoted by a dot with black edge. Legend shows language-color mapping.

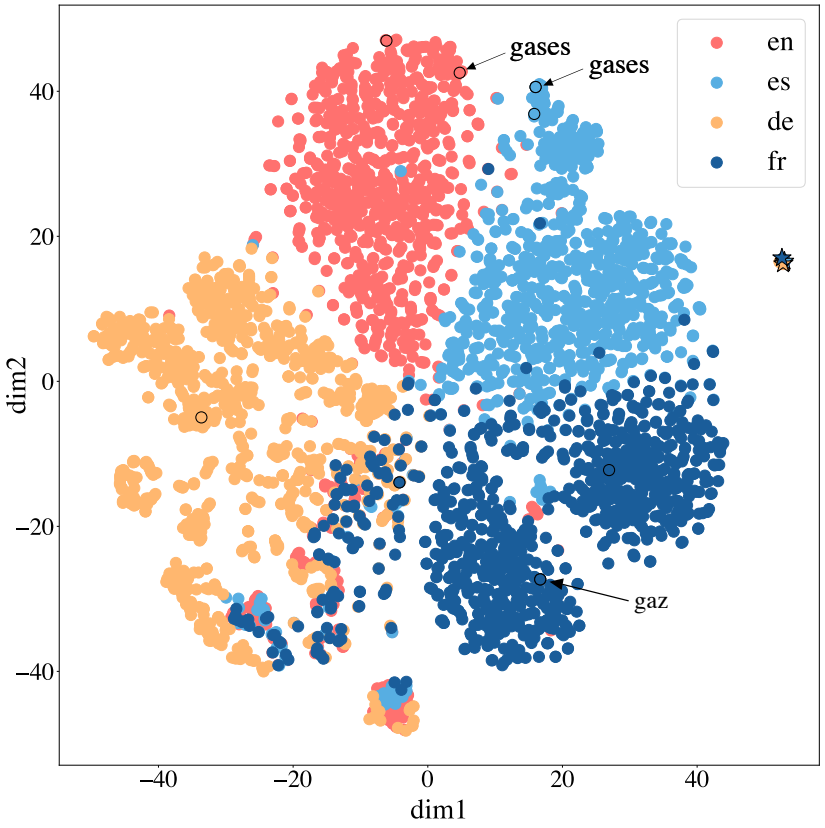


Figure 3.5: t-SNE representation from the last layer of XLM-R for the top-1000 predictions for the parallel sentences in the list above ("[MASK] that deplete the ozone layer" in English). The word "gases" is pointed out in each language (en: gases, es: gases, fr: gaz), as it was a recurrent answer from different annotators. Highest scored prediction in each language is starred; annotator's answers are denoted by a dot with black edge. Legend shows language-color mapping.

Chapter 4

Being Right for Whose Right Reasons?

Abstract

Explainability methods are used to benchmark the extent to which model predictions align with human rationales*i.e.*, are "right for the right reasons". Previous work has failed to acknowledge, however, that what counts as a rationale is sometimes subjective. This paper presents what we think is a first of its kind, a collection of human rationale annotations augmented with the annotators demographic information. We cover three datasets spanning sentiment analysis and common-sense reasoning, and six demographic groups (balanced across age and ethnicity). Such data enables us to ask both what demographics our predictions align with and whose reasoning patterns our models' rationales align with. We find systematic inter-group annotator disagreement and show how 16 Transformer-based models align better with rationales provided by certain demographic groups: We find that models are biased towards aligning best with older and/or white annotators. We zoom in on the effects of model size and model distillation, finding –contrary to our expectations– negative correlations between model size and rationale agreement as well as no evidence that either model size or model distillation improves fairness.

4.1 Introduction

Transparency of NLP models is essential for enhancing protection of user rights and improving model performance. A common avenue for providing such insight into the workings of otherwise opaque models come from explainability methods Páez (2019); Zednik and Boelsen (2022); Baum et al. (2022); Beisbart and Räz (2022); Hacker and Passoth (2022). Explanations for model decisions, also called *rationales*, are extracted to detect when models rely on spurious correlations, *i.e.*, are right for the wrong reasons McCoy et al. (2019), or to analyze if they exhibit human-like inferential semantics Piantadosi and Hill (2022); Ray Choudhury et al. (2022). Furthermore, model rationales are used to evaluate how well models' behaviors align with humans, by comparing them to human-annotated rationales, constructed by having annotators mark *evidence* in support of an instance's label DeYoung et al. (2020). Human rationales are, in turn, used in training to improve models by guiding them towards what features they should (or should not) rely on Mathew et al. (2021); Rajani et al. (2019).

While genuine disagreement in labels is by now a well-studied phenomenon Beigman Klebanov and Beigman (2009); Plank et al. (2014); Plank (2022), little attention has been paid to disagreement in rationales. Since there is evidence that human rationales in ordinary decision-making differ across demographics Stanovich and West (2000), we cannot, it seems, blindly assume that what counts as a rationale for one group of people, e.g,. young men, also counts as a rationale for another group of people, *e.g.*, elderly women. This dimension has not been explored in fairness research either. Could it be that some models that exhibit performance parity, condition on factors that align with the rationales of some groups, but not others?

Contributions We present a collection of three existing datasets with demographics-augmented annotations to enable profiling of models, *i.e.*, quantifying their alignment[1] with rationales provided by different socio-demographic groups. Such profiling enables us to ask *whose* right reasons models are being right for. Our annotations span two NLP tasks, namely *sentiment classification* and *common-sense reasoning*, across three datasets and six demographic groups, defined by age {Young, Old} and ethnicity

[1]We use the terms "agreement" and "alignment" interchangeably.

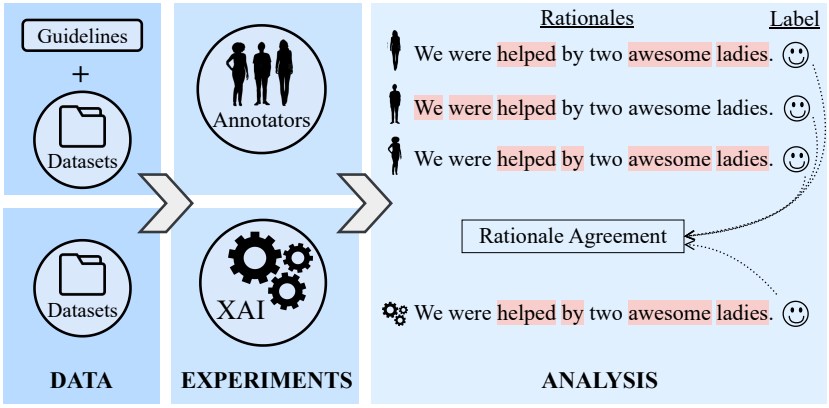


Figure 4.1: Experimental setup for a sentiment analysis task. For a given instance, annotators are asked to choose a label and mark supporting evidence for their choice. For instances with full label agreement, we compare alignment of rationales (group-group alignment). We do the same to measure group-model alignment through attention- and gradient-based explainability methods.

{Black/African American, White/Caucasian, Latino/Hispanic}. We investigate label and rationale agreement across groups and evaluate to what extent groups' rationales align with 16 Transformer-based models' rationales, which are computed through attention- and gradient-based methods. We observe that models generally align best with older and/or white annotators. While larger models have slightly better prediction performance, model size does not correlate positively with neither rationale alignment nor fairness. Our work constitutes multi-dimensional research in off-the-beaten-track regions of the NLP research manifold Ruder et al. (2022). We make the annotations publicly available. [2] [3]

4.2 Fairness and Rationales

Fairness generally concerns the distribution of resources, often across society as a whole. In NLP, the main resource is system performance. Others include computational resources, processing speed and user friendliness, but

[2] github.com/terne/Being_Right_for_Whose_Right_Reasons
[3] huggingface.co/datasets/coastalcph/fair-rationales

performance is king. AI fairness is an attempt to regulate the distribution of performance across subgroups, where these are defined by the product of legally protected attributes Williamson and Menon (2019).

NLP researchers have uniformly adopted American philosopher John Rawls' definition of fairness Larson (2017); Vig et al. (2020); Ethayarajh and Jurafsky (2020); Li et al. (2021b); Chalkidis et al. (2022), defining fairness as performance parity, except where it worsens the conditions of the least advantaged. Several dozen metrics have been proposed, based on Rawls' definition (Castelnovo et al., 2022a), some of which are argued to be inconsistent or based on mutually exclusive normative values Friedler et al. (2021); Castelnovo et al. (2022a). Verma and Rubin (2018) grouped these metrics into metrics based only on predicted outcome, *e.g.*, statistical parity, and metrics based on both predicted and actual outcome, *e.g.*, performance parity and accuracy equality. Corbett-Davies et al. (2024) argue that metrics such as predictive parity and accuracy equality do not track fairness in case of infra-marginality, *i.e.*, when the error distributions of two subgroups are different. For a better understanding of the consequences of infra-marginality we refer to Biswas et al. (2019) and Sharma et al. (2020). Generally, there is some consensus that fairness in NLP is often best evaluated in terms of performance parity using standard performance metrics Williamson and Menon (2019); Koh et al. (2020); Chalkidis et al. (2022); Ruder et al. (2022). We do the same and evaluate fairness in group-model rationale agreement quantifying performance differences (understanding performance as degree of rationale agreement) across end user demographics. In doing so, we are embodying group fairness values: that individuals should be treated equally regardless of their protected attributes, *i.e.*, group belonging.

Fairness and explainability are often intertwined in the literature due to the assumption that transparency, through explainability methods, makes it possible to identify which models are right for the right reasons or, on the contrary, right by relying on spurious, potentially harmful, patterns Langer et al. (2021); Balkir et al. (2022). This study tightens the connection between fairness and explainability, investigating whether model rationales align better with those of some groups rather than others. If so, this would indicate that models can be more robust for some groups rather than others, even in the face of performance parity on dedicated evaluation data. That is: We ask whether models are equally right for the right reasons (with the promise of generalization) across demographic groups.

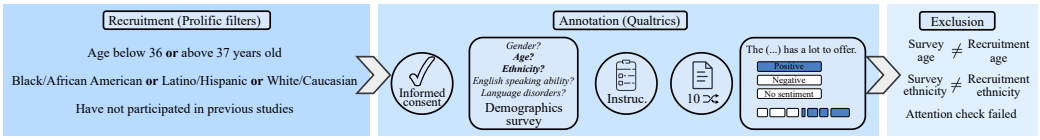


Figure 4.2: Overview of the annotation collection process from annotator recruitment criteria, to the annotation itself, and finally annotator exclusion criteria. Separately for each dataset, annotators are recruited via Prolific using specific filters for age, ethnicity and participation status. Recruits are directed to a Qualtrics survey containing, in consecutive order, a consent form, a short demographics survey, instructions for the annotation task and then approx. 10 randomly selected instances of which annotators provide both labels and rationales for. After annotation, some annotators' responses are excluded from our analysis due to certain mismatches in responses. The annotation process is detailed further in section 4.3.1 and we show the instructions and task examples in appendix A.

4.3 Data

We augment a subset of data from three publicly available datasets spanning two tasks: DynaSent Potts et al. (2020) and SST Socher et al. (2013)[4], for sentiment classification and CoS-E Talmor et al. (2019); Rajani et al. (2019) for common-sense reasoning.[5] For each dataset, we crowd-source annotations for a subset of the data. We instruct annotators to select a label and provide their rationale for their choice by highlighting supporting words in the given sentence or question. Table 4.1 shows statistics of the annotations collected. Annotation guidelines are explained in § 4.3.1 (and included in full in Appendix 4.7.1) and recruitment procedures are explained in § 4.3.2.

4.3.1 Annotation Process

We summarize the process of collecting annotations in Figure 4.2, where we depict a three-step process: recruitment, annotation and exclusion. In this section, we start by describing the second step – annotation – and explain *what* is annotated and *how* it is annotated. We describe our recruitment and exclusion criteria in the following section, 4.3.2.

[4]We work with its binary version, SST-2.

[5]We use the simplified version of CoS-E released by DeYoung et al. (2020).

	Annotators		Annotations
	×Group	Total	Total
DynaSent	48	288	2,880
SST-2	26	156	1,578
CoS-E	50	300	3,000
Total	**124**	**744**	**7,458**
Before excl.*	-	929	9,310

Table 4.1: Summary of the annotated data, showing, for each dataset, the amount of annotators within the six demographic groups, the total amount of annotators and the amount of annotations after workers have annotated approx. 10 instances each. Reported numbers are after exclusions as described in § 4.3.2. *We publicly share all annotated data which includes annotators that were excluded from our analyses.

Annotators are directed to a Qualtrics[6] survey and presented with *i)* a consent form, *ii)* a short survey on demographics, *iii)* instructions for their annotation task and lastly, *iv)* a randomly selected set of $n \approx 10$ instances to annotate, out of a subset of size N. As a result of this procedure, each group, for each dataset, is represented by approximately N/n annotators. Data points are annotated for both classification labels and extractive rationales, *i.e.*, input words that motivate the classification.

Existing rationale datasets are typically constructed by giving annotators "gold standard" labels, and having them provide rationales for these labels. Instead, we let annotators provide rationales for labels they choose themselves. This lets them engage in the decision process, but it also acknowledges that annotators with different backgrounds may disagree on classification decisions. Explaining other people's choices is error-prone Barasz and Kim (2022), and we do not want to bias the rationale annotations by providing labels that align better with the intuitions of some demographics than with those of others. For the sentiment analysis datasets, we discard neutral instances because rationale annotation for neutral instances is ill-defined. Yet, we still allow annotators to evaluate a sentence as neutral, since we do not want to force our annotators to provide rationales for positive and negative sentiment that they do not see.

[6] www.qualtrics.com

DynaSent We re-annotate $N = 480$ instances six times (for six demographic groups), comprising 240 instances labeled as positive, and 240 instances labeled as negative in the DynaSent Round 2 test set (see Potts et al. (2020)). This amounts to 2,880 annotations, in total. Our sentiment *label* annotation follows the instructions of Potts et al. (2020). To annotate *rationales*, we formulate the task as marking "supporting evidence" for the label, following how the task is defined by DeYoung et al. (2020). Specifically, we ask annotators to mark all the words, in the sentence, they think shows evidence for their chosen label.

SST-2 We re-annotate $N = 263$ instances six times (for six demographic groups), which are all the positive and negative instances from the Zuco dataset of Hollenstein et al. (2018)[7], comprising a mixture of train, validation and test set instances from SST-2, which we remove from the original data before training the models. Instructions for sentiment annotations build on the instructions by Potts et al., combined with a few examples from Zaidan et al. (2007). The instructions for annotating rationales are the same as for DynaSent.

CoS-E We re-annotate $N = 500$ instances from the test set six times (for six demographic groups) and ask annotators to firstly select the answer to the question that they find most correct and sensible, and then mark words that justifies that answer. Following Chiang and Lee (2022), we specify the rationale task with a wording that should guide annotators to make short, precise rationale annotations:

> "For each word in the question, if you think that removing it will decrease your confidence toward your chosen label, please mark it."

4.3.2 Annotator Population

We recruited annotators via Prolific based on two main criteria, age and ethnicity, previously identified as related to unfair performance differences of

[7]The Zuco data contains eye-tracking data for 400 instances from SST. By annotating some of these with rationales, we add an extra layer of information for future research. Note that there is a typo in Hollenstein et al. (2018). There is 263 positive and negative instances (not 277).

NLP systems Hovy and Søgaard (2015); Jørgensen et al. (2016); Sap et al. (2019); Zhang et al. (2021).

Recruitment In our study, there is a trade-off between collecting annotations for a diverse set of data instances (number of tasks and sentences) and for a diverse set of annotators (balanced by demographic attributes), while keeping the study affordable and payment fair. Hence, when we want to study differences between individuals with different ethnic backgrounds, we can only study a subset of possible ethnic identities (of which there are many categories and diverging definitions). We balanced the number of annotators across *three* ethnic groups — Black/African American (B), Latino/Hispanic (L) and White/Caucasian (W) — and *two* age groups —below 36 (young, Y) and above 37 (old, O), excluding both — whose cross-product results in six sub-groups: {BO, BY, LO, LY, WO, WY}. We leave a two-year gap between the age groups in order to not compare individuals with very similar ages. Furthermore, the age thresholds are inspired by related studies of age differences in NLP-tasks and common practices in distinguishing groups with an age gap Johannsen et al. (2015); Hovy and Søgaard (2015) and around the middle ages Zhang et al. (2021). Our threshold also serves to guarantee sufficient proportions of available crowdworkers in each group. Our ethnicity definition follows that of Prolific, which features in a question workers have previously responded to and hence are recruited by, defining ethnicity as:

> "[a] feeling of belonging and attachment to a distinct group of a larger population that shares their ancestry, colour, language or religion"

While we do not require all annotators to be fluent in English, we instead ask about their English-speaking abilities in the demographics survey and find that 75% of the participants speak English "very well" and only 1% "not well", and the remaining "well".

Exclusions Annotators who participated in annotating one task were excluded from participating in others. *After* annotation, we manually check whether a participant's answers to our short demographics survey correspond to their recruitment criteria. We found many discrepancies between recruitment ethnicity and reported ethnicity, especially for Latino/Hispanic individuals, who often report to identify as White/Caucasian. This highlights the

difficulty of studying ethnicities as distinct, separate groups, as it is common to identify with more than one ethnicity[8]. Hence, the mismatches are not necessarily errors. For our experiments, we decided to exclude participants with such mismatches and recruit new participants to replace their responses (see Appendix 4.7.2 for further details). A smaller amount of participants were excluded due to mismatch in reported age or due to failing a simple attention check. We release annotations both with and without the instances excluded from our analyses. The final data after pre-processing consist of one annotation per instance for each of the six groups, *i.e.*, six annotations per instance in total. Annotators annotated (approximately) 10 instances each. *All* participants were paid equally.

4.4 Experiments

We first conduct an analysis of *group-group* label agreement (i.e., comparing human annotator groups with each other, measuring human agreement on the sentiment and answer labels) and rationale agreement (measuring human agreement on rationale annotations) to characterize inter-group differences. We then move to *group-model* agreement (comparing the labels and rationales of our annotator groups to model predictions and model rationales) and ask: Do models' explanations align better with certain demographic groups compared to others? In our analysis, we further focus on how rationale agreement and fairness behave depending on model size and model distillation.

We probe 16 Transformer-based models[9]. To ease readability, we will use abbreviations following their original naming when depicting models' performance[10].
We fine-tune the models individually on each dataset (see Figure 4.3). SST-2

[8]General Social Survey as well as US Census allow respondents to report multiple ethnicities for this reason. See, *e.g.*, a GSS 2001 report commenting on multi-ethnicity: shorturl.at/BCP49.

[9]All pretrained models can be downloaded at huggingface.co/models.

[10]{`abv2`: albert-base-v2, `alv2`: albert-large-v2, `mlm-l6`: MiniLM-L6-H384-uncased, `mlm-l12`: MiniLM-L12-H384-uncased, `axlv2`: albert-xlarge-v2, `dbu`: distilbert-base-uncased, `dr`: distilroberta-base, `bbu`: bert-base-uncased, `rb`: roberta-base, `mrb`: muppet-roberta-base, `dv3b`: deberta-v3-base, `axxlv2`: albert-xxlarge-v2, `blu`: bert-large-uncased, `rl`: roberta-large, `mrl`: muppet-roberta-large, `dv3l`: microsoft/deberta-v3-large}

and CoS-E simplified[11] are modeled as binary classification tasks; DynaSent is modeled as a ternary (positive/negative/neutral) sentiment analysis task. We exclude all annotated instances from the training splits; for CoS-E, we downsample the negative examples to balance both classes in the training split. After fine-tuning for 3 epochs, we select the checkpoint with the highest validation accuracy to run on our test (annotated) splits and apply two explainability methods to obtain input-based explanations, *i.e.*, rationales, for the predictions made.

We measure label agreement with appropriate variants of F_1 (SST-2 binary-F_1; DynaSent macro-F_1; CoS-E mean of binary-F_1 towards the negative and the positive class). CoS-E simplified represents a slightly different task (see footnote 11) from what the annotators were presented to solve (a multi-class question-answering task). To correctly measure label agreement, we evaluate whether a model predicts "True" for the question-answer pair with the answer selected by the annotator. Therefore, to avoid misleading F_1 scores if, for example, a model predominantly predict True, we report the mean of the F_1 towards each class. We explain below how we measure rationale agreement.

Explainability methods We analyze models' predictions through two families of post-hoc, attribution-based[12] explainability methods: Attention Rollout (AR) Abnar and Zuidema (2020) and Layer-wise Relevance Propagation (LRP) Bach et al. (2015), a gradient-based method. Ali et al. (2022c) compare these methods, showing how their predicted rationales are frequently uncorrelated. Both AR and LRP thus provide token level rationales for a given input, but while AR approximates the relative importance of input tokens by accumulating attention, LRP does so by backpropagating "relevance" from the output layer to the input, leading to sparser attribution scores. We rely on the rules proposed in Ali et al. (2022c), an extension of the original LRP method Bach et al. (2015); Arras et al. (2017) for Transformers, aiming to uphold the conservation property of LRP in Transformers as well. This extension relies on an "implementation trick", whereby the magnitude of any

[11] CoS-E simplified represents each of the original questions into five question-answer pairs, one per potential answer, and label them as True (the right question-answer pair) or False.

[12] The methods are applied at inference time and provide explanations *locally*, *i.e.*, for each individual instance, indicating the relative importance of each input token through a score distribution.

output remains intact during backpropagation of the gradients of the model.

Comparing rationales Attention-based and gradient-based methods do not provide categorical relevance of the input tokens, but a vector S_i with continuous values for each input sentence i. We translate S_i into a binary vector S_i^b following the procedure from Wang et al. (2022c) for each group. We define the top-k^{gd} tokens as rationales, where k^{gd} is the product of the current sentence length (tokens) and the average rationale length ratio (RLR) of a group g within a dataset d. On average, RLR for SST-2 are shorter (29.6%) compared to DynaSent (31.9%) and CoS-E (33.0%) (see Appendix 4.7.2 for specific values). Models' outputs are also preprocessed to normalize different tokenizations and to match the input format given to annotators.

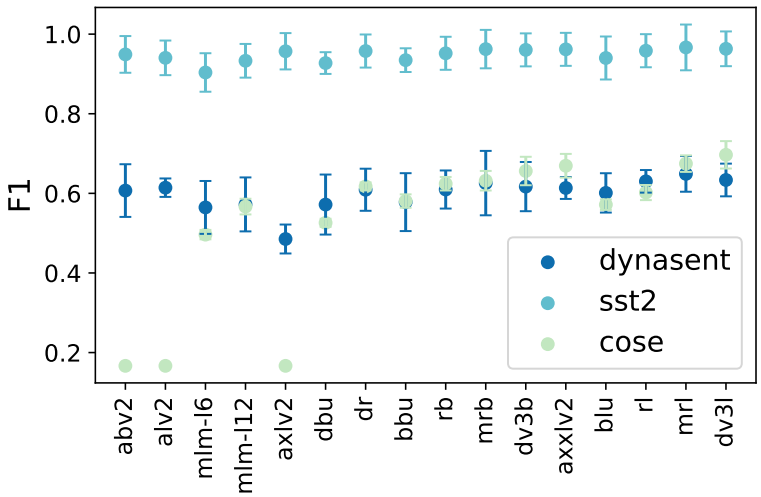


Figure 4.3: Group-model label agreement over our annotated data, measured by F1-score. Error bars show variance between the best and worst performing groups. Models are ordered by size from smallest to largest from left to right.

After aligning explanations from models and annotators in the same space, we can compare them. We employ two metrics specifically designed to evaluate discrete rationales: token-level F_1 (token-F_1) (Equation 4.1) DeYoung et al. (2020); Wang et al. (2022c), and Intersection-Over-Union F_1 (IOU-F_1) (Equation 4.3) as presented in DeYoung et al. (2020). These metrics are flexible enough to overcome the strictness of exact matching.[13]

[13]Formally,

4.5 Results and Discussion

Figure 4.3 shows group-model label agreement over our annotated data.[14] Error bars show the variability between best and worst performing groups. CoS-E exhibits the lowest variability, indicating less variability in label agreement between groups.

When annotators disagree on the label of an instance, it is to be expected that their rationales will subsequently be different. Therefore, to compare group-group (§ 4.5.1) and group-model (§ 4.5.2) rationales more fairly, we focus on the subset of instances where all groups are in agreement about the label, *i.e.*, instances with full label agreement. This amounts to 209, 152 and 161 instances for DynaSent, SST-2 and CoS-E, respectively.

$$\text{token-}F_1 = \frac{1}{N}\sum_{i=1}^{N} 2 \times \frac{P_i \times R_i}{P_i + R_i} \tag{4.1}$$

where P_i and R_i are the precision and recall for the i^{th} instance, computed by considering the overlapped tokens between models' and annotators' rationales. To measure Intersection-Over-Union, we define the categorical vector given by the annotators for each sample as A_i. Thereby,

$$\text{IOU}_i = \frac{|S_i^b \cap A_i|}{|S_i^b \cup A_i|} \tag{4.2}$$

and

$$\text{IOU-}F_1 = \frac{1}{N}\sum_{i=1}^{N} \begin{cases} 1 & \text{if IOU}_i \geq 0.5 \\ 0 & \text{otherwise.} \end{cases} \tag{4.3}$$

These metrics account for *plausibility* DeYoung et al. (2020) of the models' rationales, *i.e.*, the degree to which they are agreeable to humans, as well as the extent to which models are "right for the right reasons" McCoy et al. (2019). Since we are interested in comparing rationale alignment between groups and between groups and models, measuring plausability is our go-to. Other research Jacovi and Goldberg (2020); Setzu et al. (2021) focus on properties like *faithfulness*, which reflect a model's true decision process, *i.e.*, whether the provided rationale influenced the corresponding decision, generally measured through perturbation experiments.

[14] See Figure 4.15 in Appendix 4.7.3 for a detailed representation of group-model label agreement.

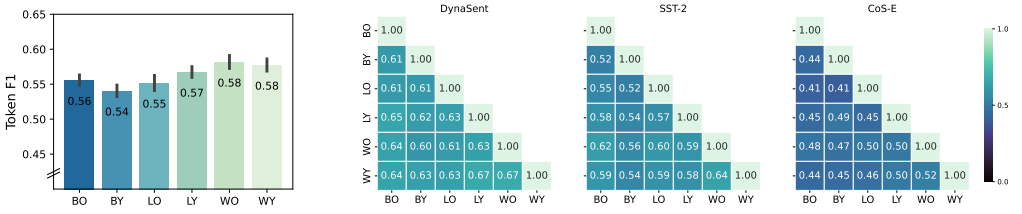


Figure 4.4: Group-group **rationale agreement** for instances with full label agreement. Agreement is measured by token-level binary F_1. On the left side, average and std (error bar) token-F_1 for 20 random combinations of paired group rationales over all datasets. On the right, each group-group agreement for each dataset. We observe lower agreement for BY except in CoS-E. WO tends to agree more with other groups, especially in CoS-E.

4.5.1 Analysis of Group-Group Agreement

We first want to quantify how different the rationales of one group are to those of others, and more generally to a random population. We compare each groups' set of rationales to a random paired set of rationales, where the rationale of each instance is randomly picked from one of the five other groups. Figure 4.4 shows the overall agreement score, average token-F_1 across datasets, and its standard deviation from 20 random seeds, *i.e.*, 20 random combinations of paired rationales. We observe that rationales of White annotators (WO, WY) are on average more similar to others while the average difference with the rationales of minority groups like, for example, Black Young (BY), is greater.

We then compute the level of rationale agreement (token-F_1) between all groups (heatmaps on Figure 4.4) and observe that, in general, differences in group-group rationale agreement are consistent across datasets (tasks): Black Youngs (BY) have lower alignment with others, especially in sentiment analysis tasks. While the definition of rationales for DynaSent seems to be easier (higher values of agreement), it seems to be harder (lower values of agreement) for CoS-E, even when the label is agreed upon. We hypothesize this is due to the complexity of the CoS-E task itself, which also leads to more lengthy rationales, as reflected by the average RLR reported on § 4.4, probably in the absence of a clear motivation for the selected answer.

The definition of what is *common-sense* varies across cultures and it is related to a person's background Hershcovich et al. (2022), which makes CoS-E

a highly subjective task[15]. Take for example the question "Where would you find people standing in a line outside?" with these potential answers: "bus depot", "end of line", "opera", "neighbor's house" and "meeting". Even if there is agreement on the *correct* choice as "bus depot", the rationale behind it could easily differ amongst people, *i.e.*, it could be due to "people standing", or the fact that they are standing in "a line outside", or all together.

4.5.2 Analysis of Group-Model Agreement

Now that we have analyzed group-group agreement, we measure the alignment between groups' rationales and models' rationales. We analyze predictions from 16 Transformer-based models and employ AR and LRP to extract model rationales. Methods for comparing rationales and measuring group-model agreement are explained in Section 4.4.

Socio-demographic fairness Figure 4.5 shows a systematic pattern of model rationales aligning better with the rationales of older annotators in each ethnic group (BO, LO, WO) on the sentiment datasets. The only exception is White Young (WY) annotators in SST-2, whose median token-F_1 is higher than their older counterpart. We argue this is due, in part, to the data source of the tasks themselves. While DynaSent constitutes an ensemble of diverse customer reviews, SST is based on movie review excerpts from Rotten Tomatoes with a more informal language, popular amongst younger users. Findings from Johannsen et al. (2015) and Hovy and Søgaard (2015) indicate that there exist grammatical differences between age groups. Johannsen et al. (2015) further showed several age and gender-specific syntactic patterns that hold even across languages. This would explain not only the noticeable group-group differences when marking supporting evidence (lexical structures) for their answers, but also the agreement disparity reflected by models fine-tuned on potentially age-biased data.

Results are consistent with previous findings of Zhang et al. (2021), who show a variety of language models aligning better with older, white annotators, and worse with minority groups, in word prediction tasks. We observe that group-model rationale agreement does not correlate with group-model class agreement, *i.e.*, when a model performs well for a particular group, it does not necessarily entail that its rationales, or learned patterns, align.

[15]This is specially notorious on the query type *people*.

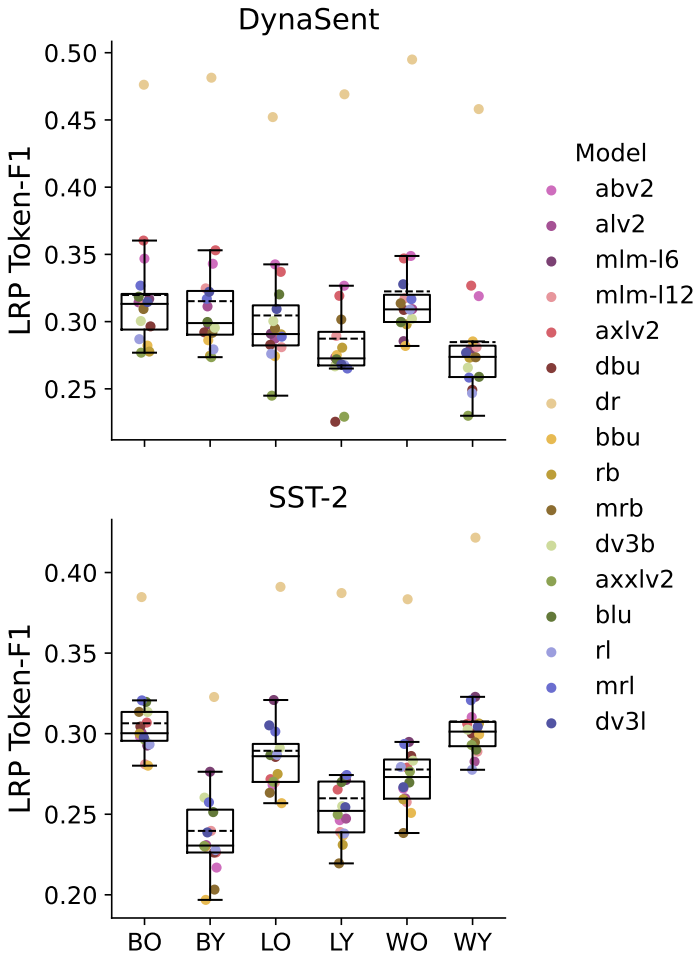


Figure 4.5: Box-plots of group-model rationale alignment for the two sentiment datasets measured with token-F_1. Model rationales are extracted with LRP. Each dot represents a model's token-F_1 score for the respective group. We see that for each ethnic group, model rationales align better with rationales of older annotators, except for White Young (WY) annotators of SST-2. DistilRoBERTa (`dr`) is an outlier, consistently showing the best scores in both datasets across groups.

Group-model rationale agreement evaluated with Attention Rollout and CoS-E are shown in Figure 4.16 in Appendix 4.7.3, along with results using the complementary metric (IOU-F_1). The patterns derived from them are in line with those in Figure 4.5: AR shows similar behaviours as LRP, but leads

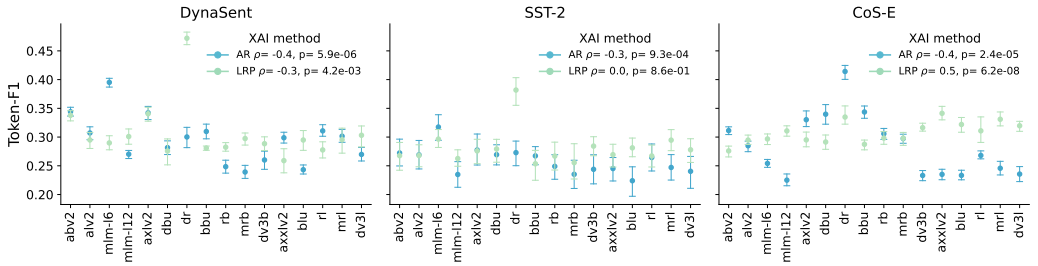


Figure 4.6: Group-model rationale alignment (token-F_1). Error bars show the distance between the groups with the highest and lowest scores. On the X-axis, models are ordered from smallest to largest. We show Spearman correlation coefficients, ρ, between token-F_1 scores (the concatenation of all groups' scores) and model sizes (in Million parameters), finding token-F_1 to be negatively correlated with model size in most cases.

to larger variation between models. However CoS-E, which, as explained, is a very different task, does not seem to exhibit big group differences. This is also noticeable in Figure 4.6, where error bars show the distance between groups with the highest and lowest level of agreement in every model.

The role of model size In general, larger language models seem to perform better on NLP tasks. In our setting, Figure 4.3 shows a positive trend with model size: larger models achieve, in general, higher performance. Could it be the case that larger language models also show higher rationale agreement? And, are they consequently more fair? We evaluate fairness in terms of performance parity: min-max difference between the group with the lowest and highest token-F_1 (per model). Relying on min-max difference captures the widely shared intuition that fairness is always in the service of the worst off group Rawls (1971).

Contrary to our expectations, Figure 4.6 shows how token-F_1 scores actually *decrease* with model size – with CoS-E model rationales from LRP being the only exception to the trend. We report Spearman correlation values for each dataset and explainability method: The negative correlation between token-F_1 and model size is significant in all three datasets with AR, but only in DynaSent with LRP. The positive correlation in CoS-E with LRP rationales is also significant.

When we zoom in on the min-max Token-F_1 gaps (error bars on Fig-

	token-F_1 (↑)	IOU-F_1 (↑)	min-max token-F_1 (↓)	min-max IOU-F_1 (↓)
minilm-l6-h384-unc.	**.31**	**.28**	**.045**	**.068**
minilm-l12-h384-unc.	.27	.21	**.045**	.083
distilbert-base-unc.	**.29**	**.24**	.064	.100
distilroberta-base	**.36**	**.36**	.065	**.069**
Avg. (16 models)	.29	.24	.054	.081

Table 4.2: Group-model alignment for four distilled models. Bottom row shows average scores across all 16 models considered in this paper. Values in **bold** are better than the average (lower if ↓, higher if ↑). While rationale alignment (token-F_1 and IOU-F_1) seem to be better for distilled models, only `minilm-l6-h384-uncased` is also fairer than the average (in terms of min-max difference) with both metrics.

ure 4.6)[16], we find that performance gaps are uncorrelated with model size. Therefore, there is no evidence that larger models are more fair, *i.e.*, rationale alignment does not become more equal for demographic groups. In the context of toxicity classification, work from Baldini et al. (2021) also hints that size is not well correlated with fairness of models.

Do distilled models align better? Knowledge distillation has been proven to be effective in model compression while maintaining model performance Gou et al. (2021). But can it also be effective in improving NLP fairness? Xu and Hu (2022) find a consistent pattern of toxicity and bias reduction after model distillation. Chai et al. (2022) show promising results when approaching fairness without demographics through knowledge distillation. Tan et al. (2018) discuss the benefits of applying knowledge distillation to leverage model interpretability. Motivated by these findings, we take results from LRP to look closer into group-model rationale agreement for distilled models, which we show in Table 4.2. We find overall higher rationale agreement for distilled models. However, there is no evidence that distilled models are also more fair: Only `minilm-l6-h384-uncased` has a smaller performance gap between the best and worst-off group for both metrics compared to the average.

[16]See Figure 4.17 in Appendix 4.7.3.2 for a plot of the gaps themselves.

4.6 Conclusion

In this paper, we present a new collection of three existing datasets with demographics-augmented annotations, balanced across age and ethnicity. By having annotators choose the right label and marking supporting evidence for their choice, we find that what counts as a rationale differs depending on peoples' socio-demographic backgrounds.

Through a series of experiments with 16 popular model architectures and two families of explainability methods, we show that model rationales align better with older individuals, especially on sentiment classification. We look closer at model size and the influence of distilled pretraining: despite the fact that larger models perform better in general NLP tasks, we find negative correlations between model size and rationale agreement. Furthermore, from the point of view of performance parity, we find no evidence that increasing model size improves fairness. Likewise, distilled models do not seem to be more fair in terms of rationale agreement, however they do present overall higher scores.

This work indicates the presence of undesired biases that *do not necessarily surface in task performance*. We believe this provides an important addendum to the fairness literature: Even if models are fair in terms of predictive performance, they may still exhibit biases that can only be revealed by considering model rationales. If models are equally right, but only right for the right reasons in the eyes of some groups rather than others, they will likely be less robust for the latter groups.

Limitations

Our analysis is limited to non-autoregressive Transformer-based models, fine-tuned with the same set of hyperparameters. Hyperparameter optimization would undoubtedly lead to better performance for some models, but we fine-tuned each model with standard hyperparameter values for solving sentiment analysis tasks DeYoung et al. (2020) to reduce resource consumption. This should not affect the conclusions drawn from our experiments.

Comparing human rationales and rationales extracted with interpretability methods such as Attention Rollout and LRP is not straightforward. Overall agreement scores depend on how model rationales are converted into categorical values (top-k^{gd}). See Jørgensen et al. (2022) for discussion.

Acknowledgments

Many thanks to Stephanie Brandl, David Dreyer Lassen, Frederik Hjort, Emily Pitler and David Jurgens for their insightful comments. This work was supported by the Novo Nordisk Foundation.

Ethics Statement

Broader impact Although explainability and fairness are broadly viewed as intertwined subjects, very little work has studied the two concepts together (Feng and Boyd-Graber, 2019; González et al., 2021; Ruder et al., 2022). This study is a first of its kind to examine fairness issues of explainability methods and to publish human rationales with diverse socio-demographic information. We hope this work will impact the NLP research community towards more data-aware and multi-dimensional investigations of models and methods, and towards further studies of biases in NLP.

Personal and sensitive data This study deals with personal and sensitive information. The responses are anonymous and cannot be used to identify any individual.

Informed consent The participants were informed of the study's overall aim, the procedure and confidentiality of their responses. With this information, the participants consented to the use and sharing of their responses.

Potential risks We do not anticipate any risks of participation in the study, yet we do note a recent awareness of poor working conditions among crowdworkers for AI data labeling in some countries (Williams et al., 2022). The recruitment platform Prolific, used in this study, is targeted towards research (rather than AI development) and has stricter rules on participant screening and minimum wages (Palan and Schitter, 2017), compared to other popular platforms, which we hope reduce the risk of such poor working conditions.

Remuneration The participants were paid an average of 7.1£/hour (≈ 8.8$/hour).

Intended use The collected annotations and demographic information will be publicly available to be used for research purposes only.

4.7 Appendix

4.7.1 Annotation guidelines and task examples

On the next pages, we firstly show the annotation instructions given to annotators within the Qualtrics surveys. Full exports of the surveys are available in our GitHub repository.[17]

We created instructions specific for each dataset (DynaSent, SST-2, and CoS-E), leaning on prior work of annotating labels and rationales for these and similar datasets Potts et al. (2020); Zaidan et al. (2007); DeYoung et al. (2020), as described in the paper, section 4.3.1.

Figures 4.7–4.8, 4.9–4.10, and 4.11–4.12 show the instructions for DynaSent, SST-2 and CoS-E, respectively, and Figure 4.13 shows an example of how an instance for the sentiment task and the common-sense reasoning task is annotated, i.e. how it looked from the perspective of the crowdworkers.

Annotating rationales for the common-sense reasoning task is somewhat more complex than annotating rationales for sentiment: while we can ask annotators to mark "evidence" for a sentiment label —often resulting in marking words that are positively or negatively loaded— we cannot as simply ask for "evidence" for a common-sense reasoning answer without risking some confusion. Take, for instance, the question "Where do you find the most amount of leafs?" with the answer being "Forest", as shown in Figure 4.11. Here, the term "evidence" might be misunderstood as actual evidence for why there would be more leafs in the forest compared to a field —evidence which cannot be found within the question itself. We therefore re-phrase the rationale annotation instructions for CoS-E, following an example from Chiang and Lee (2022), and ask, "For each word in the question, if you think that removing it will decrease your confidence toward your chosen label, please mark it." Furthermore, the subset of the CoS-E dataset, that we re-annotate, consists of the more "difficult" split of the CommonsenseQA dataset Talmor et al. (2019); DeYoung et al. (2020). To make the task as clear as possible to the annotators, we explain, in the instructions, that the question and answer-options have been created by other crowdworkers who

[17] github.com/terne/Being_Right_for_Whose_Right_Reasons.

were instructed to create questions that could be "easily answered by humans without context, by the use of common-sense knowledge", as is described by Talmor et al. (2019).

Dataset	N	Complete label agreement			
		Pos	Neg	Neutral	Total
DynaSent	480	105	102	2	209
SST	263	79	73	0	152
CoS-E	500	-	-	-	161

Table 4.3: Number of instances, in our (re-)annotated data, where all annotator groups agreed upon the instance's label.

4.7.2 Annotations Overview

Table 4.3 shows the number of instances in the data subsets, we work with, and the number of instances where all our annotator groups agreed on the label and that are therefore used for rationale-agreement analyses. Table 4.4 gives further information on the distribution of annotators, across groups and datasets, as well as ratios of rationale lengths to input lengths.

4.7.3 Supplementary Figures

For completeness, we provide supplementary figures for all the metrics and datasets analyzed in the paper.

4.7.3.1 Label Agreement

Heatmaps in Figure 4.14 show the level of group-group label agreement across datasets. Similar to what is shown in Figure 4.4, BY consistently exhibit lower level of agreement.

Box-plots in Figure 4.15 represent group-model label agreement. Each dot represents the F1-score of each model. While for Cos-E the models generally exhibit lower variability across groups, the level of agreement is also lower (as shown in Figure 4.3).

Dataset		BO	BY	LO	LY	WO	WY	Total/Avg.
DynaSent	Annot.	51	56	61	73	54	51	346
	Annot.*	48 (58%F)	48 (67%F)	48 (44%F)	48 (40%F)	48 (56%F)	48 (48%F)	288
	RLR	33.7	32.5	31.5	29.8	34.7	29.1	31.9
SST	Annot.	28	27	53	43	27	29	207
	Annot.*	26 (69%F)	26 (58%F)	26 (38%F)	26 (31%F)	26 (38%F)	26 (69%F)	156
	RLR	32.1	25.1	30.7	27.8	29.1	32.7	29.6
CoS-E	Annot.	52	56	74	85	54	55	376
	Annot.*	50 (60%F)	50 (60%F)	50 (40%F)	50 (48%F)	50 (48%F)	50 (40%F)	300
	RLR	31.9	32.9	34.1	32.2	33.3	33.6	33.0

Table 4.4: Overview of our annotated data. Rows display statistics per dataset. Columns refer to each demographic group: Black/African American old (BO) and young (BY), Latino/Hispanic old (LO) and young (LY), White/Caucasian old (WO) and young (WY). Last column show the total quantity of each feature over all groups. Row-wise within each dataset: 'Annot.' and 'N' reflect the total number of annotators and instances, respectively. Annot.* refers to the number of annotators left after pre-processing (see exclusion criteria in Section 4.3.2). Number shown between brackets refers to the percentage of female annotators. RLR represents the ratio of rationale length to its input length (percentage).

4.7.3.2 Rationale Alignment

Figure 4.16 is the extended version of Figure 4.5, showing the group-model rationale agreement for each dataset, each explainability method and with two metrics for measuring agreement, token-F_1 and IOU-F_1.

The bar charts in Figure 4.17 shows, per model and dataset, the distance between the group with the lowest and highest agreement with the model (by token-F_1), which we refer to as the "min-max token-F_1 gaps" in section 4.5.2. We include this plot because it serves to better illustrate the gaps themselves, and how they are uncorrelated with model size, compared to what Figure 4.6 in the paper can convey.

Instructions
Please read these instructions carefully.

You will be shown 10 sentences from reviews of products and services. For each, your task is to choose from one of our three labels:

Positive: The sentence conveys information about the author's positive evaluative sentiment.

Negative: The sentence conveys information about the author's negative evaluative sentiment.

No sentiment: The sentence does not convey anything about the author's positive or negative sentiment.

Here are some examples of the labels:

Sentence: This is an under-appreciated little gem of a movie.
(This is Positive because it expresses a positive overall opinion.)

Sentence: I asked for my steak medium-rare, and they delivered it perfectly!
(This is Positive because it puts a positive spin on an aspect of the author's experience.)

Sentence: The screen on this device is a little too bright.
(This is Negative because it negatively evaluates an aspect of the product.)

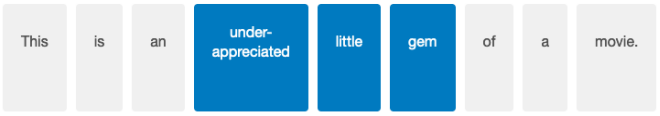
Sentence: The book is 972 pages long.
(This is No sentiment because it describes a factual matter with not evaluative component.)

Sentence: The entrees are delicious, but the service is so bad that it's not worth going.
(This is Negative because the negative statement outweighs the positive one.)

Sentence: The acting is great! The soundtrack is run-of-the mill, but the action more than makes up for it.
(This is Positive because the positive statements outweighs the negative.)

Figure 4.7: DynaSent annotation instructions (part 1/2).

We further ask you to specify what snippets of text, in the sentence, you think acts as supporting evidence for your chosen label. The sentence will be shown to you as illustrated below, and your task is to mark (by clicking on them) all the words you think shows evidence for the sentiment label you chose.



Be aware that some sentences might be too long to fit on your screen. You therefore have to remember to scroll in order to see all the words that can be marked as evidence.

Click the forward button below when you are ready to start the task.

Figure 4.8: DynaSent annotation instructions (part 2/2).

Instructions
Please read these instructions carefully.

You will be shown approximately 10 sentences from reviews of movies. For each, your task is to choose from one of our three labels:

Positive: The sentence conveys information about the author's positive evaluative sentiment.
Negative: The sentence conveys information about the author's negative evaluative sentiment.
No sentiment: The sentence does not convey anything about the author's positive or negative sentiment.

Here are some examples of the labels:

Sentence: This is an under-appreciated little gem of a movie.
(This is Positive because it expresses a positive overall opinion.)

Sentence: he is one of the most exciting martial artists on the big screen, continuing to perform his own stunts and dazzling audiences with his flashy kicks and punches.
(This is Positive because it positively evaluates an aspect of the movie.)

Sentence: The acting is great! The soundtrack is run-of-the-mill, but the action more than makes up for it.
(This is Positive because the positive statements outweigh the negative.)

Sentence: The story is interesting but the movie is so badly put together that even the most casual viewer may notice the miserable pacing and stray plot threads.
(This is Negative because the negative statement outweighs the positive one.)

Sentence: A woman in peril. A confrontation. An explosion. The end. Yawn. Yawn. Yawn.
(This is Negative because it puts a negative spin on the author's experience.)

Figure 4.9: SST-2 annotation instructions (part 1/2).

Sentence: don't go see this movie.
(This is Negative because it recommends against seeing the movie, reflecting a negative evaluation.)

Sentence: it is directed by Steven Spielberg.
(This is No sentiment because it describes a factual matter with no evaluative component.)

Sentence: I saw it in the local theater with my best friend.
(This is No sentiment because it does not say anything about the movie.)

We further ask you to specify what snippets of text, in the sentence, you think acts as supporting evidence for your chosen label. The sentence will be shown to you as illustrated below, and your task is to mark (by clicking on them) all the words you think shows evidence for the sentiment label you chose.

Be aware that some sentences might be too long to fit on your screen. In that case you have to scroll in order to see all the words that can be marked as evidence.

Click the forward button below when you are ready to start the task.

Figure 4.10: SST-2 annotation instructions (part 2/2).

Instructions
(Please read these instructions carefully.)

You will be shown 10 multiple-choice questions. All questions and their answer-options have been created by other crowdworkers, who where instructed to create questions that can be fairly easily answered by humans without context, by the use of common-sense knowledge.

Your task is to firstly select the answer you think is most correct and sensible. We call this the label of the question. Secondly, we ask you to mark relevant words in the question that justifies your choice. Specifically, for each word in the question, if you think that removing it will decrease your confidence toward your chosen label, you should mark it.

In the image below, you see an example of how the task will be presented to you. To the question "Where do you find the most amount of leafs?", the option "Forest" is selected as the correct answer and four words have been marked as justification.

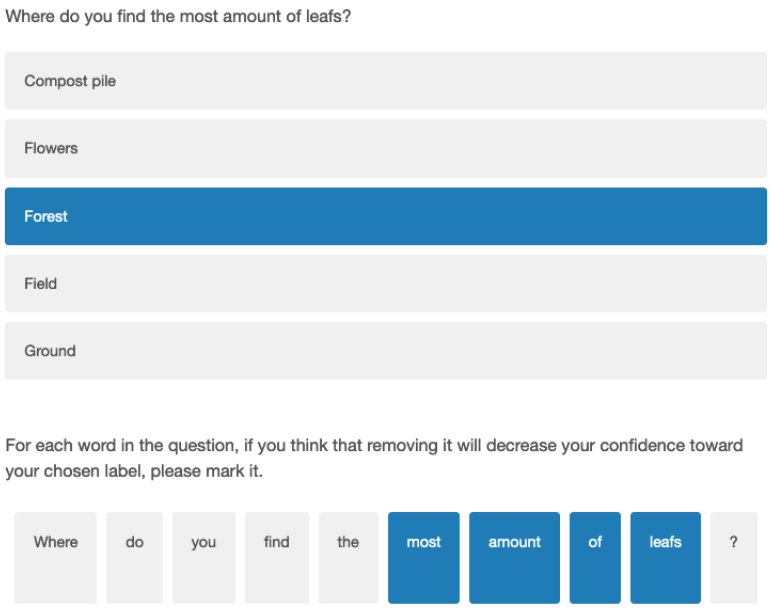
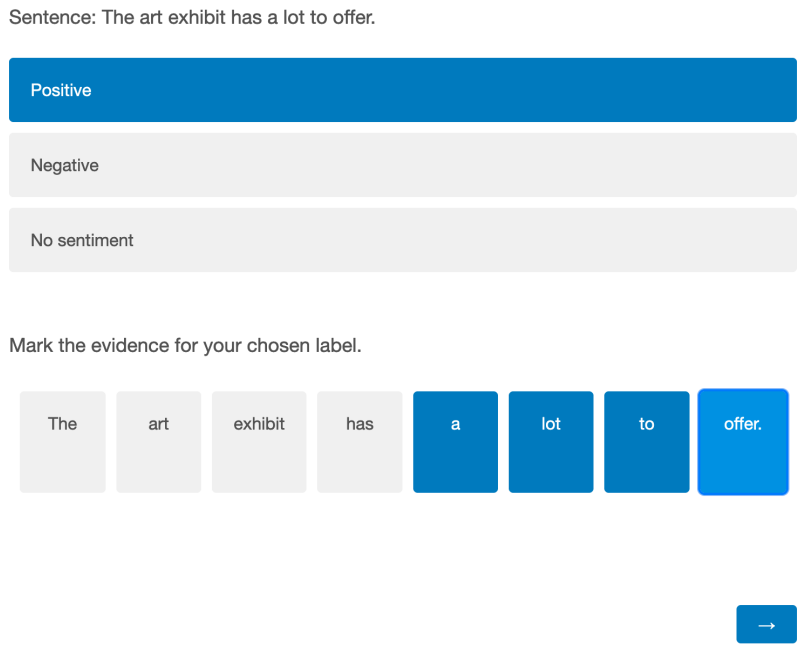


Figure 4.11: CoS-E annotation instructions (part 1/2).

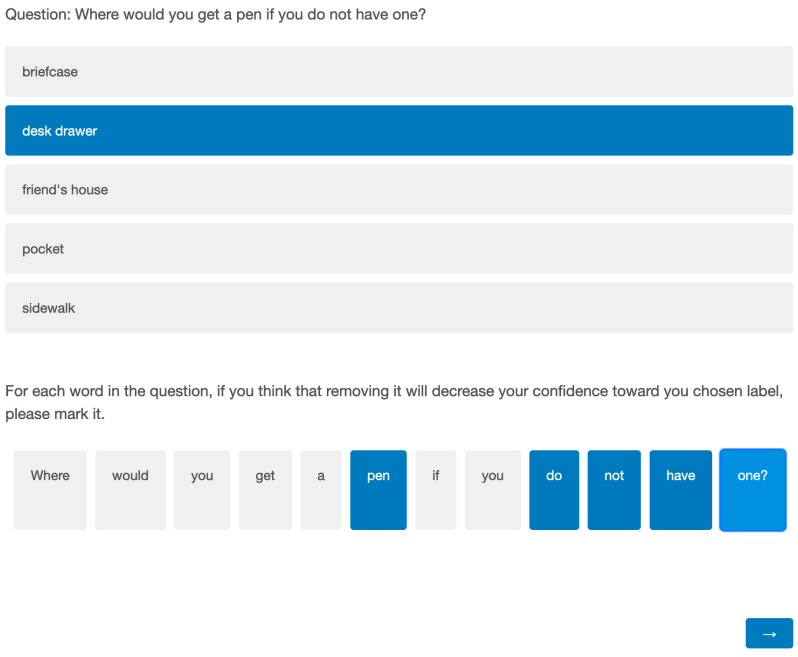
When marking words, be aware that some questions might be longer and not fit perfectly on your screen. In that case you have to scroll in order to see all the words that can be marked. Also, the texts may have misspellings, typos and wrongly put spaces before punctuation – pay no attention to this.

Click the forward button below when you are ready to start the task.

Figure 4.12: CoS-E annotation instructions (part 2/2).



(a) Sentiment annotation example.



(b) Common-sense reasoning annotation example.

Figure 4.13: Screenshots of the annotation tasks as they are viewed in Qualtrics surveys.

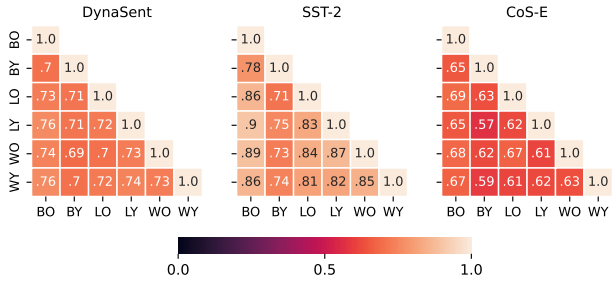


Figure 4.14: Group-group label agreement (F1-scores).

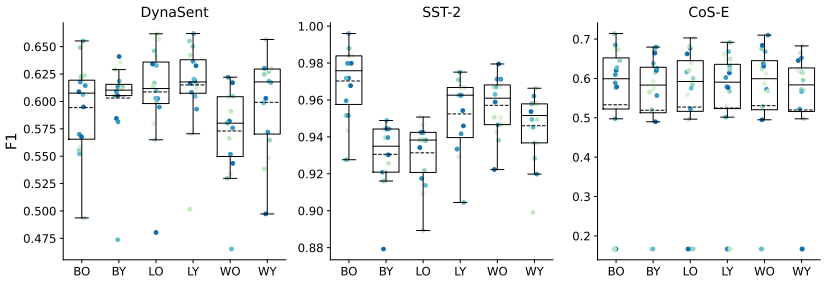
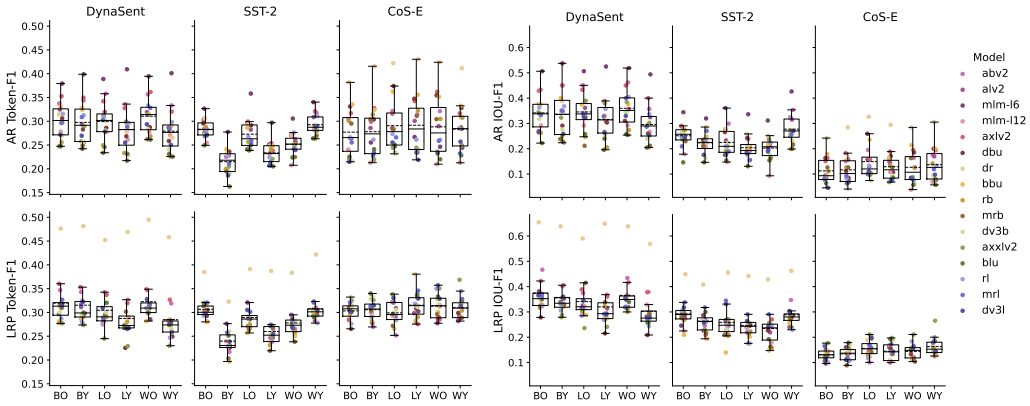


Figure 4.15: Group-model label agreement (F1-scores).



(a) Token-F1 scores (b) IOU-F1 scores

Figure 4.16: Box-plots of group-model rationale agreement for the each dataset measured with Token-F1 (left) and IOU-F1 (right). Model rationales are extracted with Attention Rollout (top row) and LRP (bottom row). Each dot represents a model's agreement with the respective group.

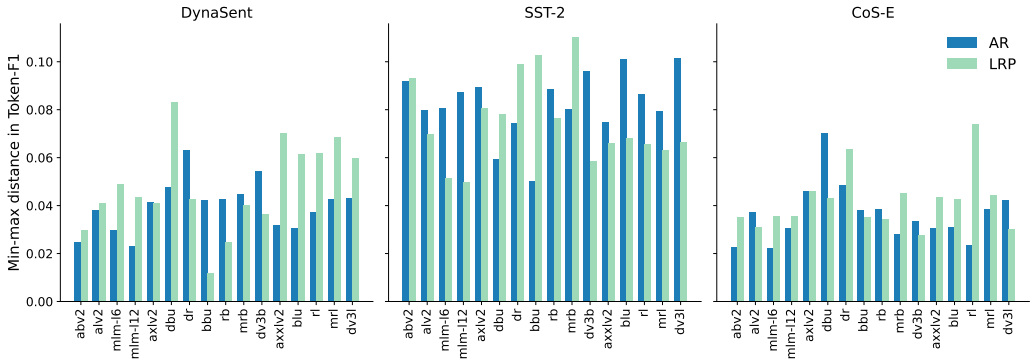


Figure 4.17: Per-model difference between the group with the lowest (min) and highest (max) model-group agreement measured with token-F1. Models on the x-axis are sorted by model size. The min-max captures a measure of fairness, with a smaller difference entailing more equal model-group rationale alignments. We find that the differences are uncorrelated with model size (in Million parameters), as is visible in this plot.

Chapter 5

Evaluating Bias and Fairness in Gender-Neutral Pretrained Vision-and-Language Models

Abstract

Pretrained machine learning models are known to perpetuate and even amplify existing biases in data, which can result in unfair outcomes that ultimately impact user experience. Therefore, it is crucial to understand the mechanisms behind those prejudicial biases to ensure that model performance does not result in discriminatory behaviour toward certain groups or populations. In this work, we define gender bias as our case study. We quantify bias amplification in pretraining and after fine-tuning on three families of vision-and-language models. We investigate the connection, if any, between the two learning stages, and evaluate how bias amplification reflects on model performance. Overall, we find that bias amplification in pretraining and after fine-tuning are independent. We then examine the effect of continued pretraining on gender-neutral data, finding that this reduces group disparities, *i.e.*, promotes fairness, on VQAv2 and retrieval tasks without significantly compromising task performance.

5.1 Introduction

As shown by Mitchell (1980) and Montañez et al. (2019), inductive biases are essential for learning algorithms to outperform random guessing. These task-specific biases allow algorithms to generalize beyond training data but, necessarily, they should not be conflated with prejudicial or unwanted biases. Unwanted bias, such as bias against demographic groups, can be found in many applications, from computer vision systems to natural language processing (NLP). Vision-and-language (V&L) models lie at the intersection of these areas, where one of the key challenges is deploying robust models to perform high-level reasoning based on the multimodal context instead of exploiting biases in data (Zhao et al., 2017a).

Multiple studies Lee et al. (2021); Hirota et al. (2022a); Zhou et al. (2022) have shown that V&L models leverage co-occurrences between objects and their context to make predictions, and thus are susceptible to unwanted biases. However, these authors do not explore the broad landscape of V&L models and focus on biases in common visual datasets Wang et al. (2022a); Hirota et al. (2022b), only on pretrained models Zhou et al. (2022) or only focus on one application, *e.g.*, image captioning Burns et al. (2018); Hirota et al. (2022a) or semantic segmentation Lee et al. (2021).

In this work, we investigate to what extent the unwanted bias in a V&L model is caused by the *pretraining data*. To answer this question, we focus on one important aspect of bias encoded in V&L models, namely *bias amplification*. Bias amplification occurs when a model exacerbates unwanted biases from the training data and, unlike other forms of bias, it is not solely attributed to the data, yet it can vary greatly during training Hall et al. (2022).

We explore bias amplification in two encoder-only V&L models: LXMERT Tan and Bansal (2019) and ALBEF Li et al. (2021a), and the encoder-decoder model BLIP Li et al. (2022). Specifically, we quantitatively and qualitatively analyse the relationship between the bias encoded in pretrained models, and after fine-tuning on downstream tasks including visual question answering, visual reasoning and image–text retrieval.

While bias can be studied with respect to any protected attribute, the majority of NLP research has focused on (binary) gender Sun et al. (2019); Stanczak and Augenstein (2021); Shrestha and Das (2022). We also use gender bias as our case study but different to previous work, we advocate for the

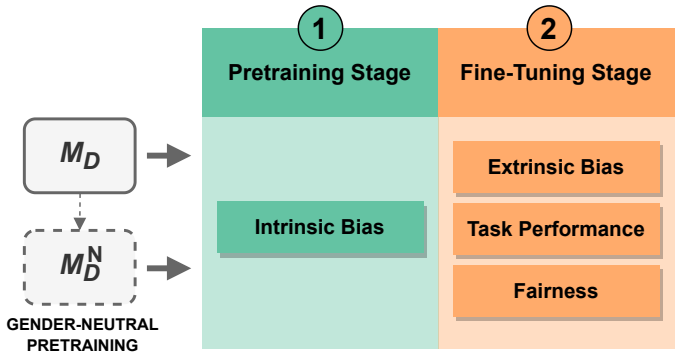


Figure 5.1: A V&L model pretrained on data D (M_D) is further pretrained on gender-neutral multimodal data D^N, resulting in a gender neutral V&L model (M_D^N). Both models can then be used in a two-phase analysis: 1) bias amplification is measured on the intrinsic bias of pretrained models, and 2) bias amplification, task performance and fairness are evaluated on the extrinsic performance of fine-tuned models.

inclusion of gender-neutral terms Dev et al. (2021) and consider three gender categories based on visual appearance: male, female and gender-neutral (*e.g.*, PERSON). The use of both visual and grammatical gender information across V&L tasks is needed for identifying the target of, for example, a question. But the demographics of the subject should not solely influence the outcome of the model. Otherwise, the model may reinforce harmful stereotypes resulting in negative consequences for certain group identities van Miltenburg (2016).

Motivated by this argument, we investigate the effect of shifting the projection of gender-marking to a gender-neutral space by continued pretraining on gender-neutral multimodal data–a form of domain adaptation Gururangan et al. (2020)–and how it reflects on task performance after fine-tuning. Figure 5.1 depicts an overview of our full workflow.

Contributions We examine whether bias amplification measured on pretrained V&L models (intrinsic bias) relates to bias amplification measured on downstream tasks (extrinsic bias). We show that a biased pretrained model might not translate into biased performance on a downstream task to a similar degree. Likewise, we measure model fairness through group disparity and show that it is not unequivocally related to bias in a model. Further-

more, we empirically present a simple, viable approach to promote fairness in V&L models: performing an extra epoch of pretraining on unbiased (gender-neutral) data reduces fine-tuning variance and group disparity on VQAv2 and retrieval tasks on the majority of models studied, without significantly compromising task performance.

We make our code publicly available to ensure reproducibility and foster future research.[1]

5.2 Related Work

Bias in language In general, bias can be defined as "undue prejudice" Crawford (2017). Studies targeting language models Kurita et al. (2019); Zhao et al. (2019) have shown that biases encoded in pretrained models (*intrinsic bias*) can be transferred to downstream applications (*extrinsic bias*), but the relationship between these biases is unclear.[2] There are several studies Goldfarb-Tarrant et al. (2021); Delobelle et al. (2021); Kaneko et al. (2022); Cao et al. (2022); Orgad et al. (2022), showing that intrinsic bias in language models does not consistently correlate with bias measured extrinsically on a downstream task or, similarly, with empirical fairness Shen et al. (2022); Cabello et al. (2023b). Contrarily, Jin et al. (2021b) observed that the effects of intrinsic bias mitigation are indeed transferable in fine-tuning language models. To the best of our knowledge, we are the first to investigate if the same holds for V&L models.

Bias in vision & language Prior research observed the presence of gender disparities in visual datasets like COCO Bhargava and Forsyth (2019); Zhao et al. (2021); Tang et al. (2021) and Flickr30k van Miltenburg (2016). Recent studies also revealed the presence of unwanted correlations in V&L models. Prejudicial biases found in V&L models are not only attributed to one domain, *i.e.*, vision or language, but they are compound Wang et al. (2019), and this should be studied together. Srinivasan and Bisk (2021);

[1] github.com/coastalcph/gender-neutral-vl

[2] As first suggested by Goldfarb-Tarrant et al. (2021), we can broadly categorize bias into *intrinsic* and *extrinsic*. Therefore, *intrinsic metrics* are applied directly to word representations and relate bias to the geometry of the embedding space, whereas *extrinsic metrics* evaluate bias in downstream tasks.

Hirota et al. (2022a) and Zhou et al. (2022) show that different model architectures exhibit gender biases, often preferring to reinforce a stereotype over faithfully describing the visual scene. Bianchi et al. (2023) show the presence of stereotypes in image generation models and discuss the challenges of the compounding nature of language–vision biases. Another line of work addresses visual *contextual* bias Choi et al. (2012); Zhu et al. (2018); Singh et al. (2020) and study a common failure of recognition models: an object fails to be recognized without its co-occurring context. So far, little work has investigated bias amplification in pretrained V&L models. Our study is among the first to cast some light on the gender bias encoded in pretrained V&L models and evaluate how it translates to downstream performance.

Gender-neutral language Zhao et al. (2019) examine the effect of learning gender-neutral embeddings during training of static word embeddings like GloVe Pennington et al. (2014). Sun et al. (2021) and Vanmassenhove et al. (2021) present rule-based and neural rewriting approaches to generate gender-neutral alternatives in English texts. Brandl et al. (2022) find that upstream perplexity substantially increases and downstream task performance severely drops for some tasks when gender-neutral language is used in English, Danish and Swedish. Amend et al. (2021) show that the substitution of gendered for gender-neutral terms on image captioning models poses a viable approach for reducing gender bias. In our work, we go one step beyond and investigate the effect of continued pretraining V&L models on in-domain data where gendered terms have been replaced by their gender-neutral counterparts (*e.g.*, *sister* → *sibling*).

5.3 Problem Formulation

We characterize the gender bias encoded in V&L models in a two-phase analysis:

i) Intrinsic bias: First, we investigate the bias encoded after the V&L pretraining phase.

ii) Extrinsic bias and task performance: Second, we fine-tune the models on common downstream tasks to further investigate how bias affects model performance.

These investigations will be performed using a set of original, pretrained models M_D, and models that have been further pretrained on gender-neutral data M_D^{N} in order to mitigate any biases learned during pretraining (§5.4.4). We hypothesize that this bias mitigation technique will decrease both intrinsic and extrinsic biases encoded in the models.

Data Our analysis relies on data where the gender of the main actor of the image is known. This is, to some degree, annotated in the crowdsourced text, *e.g.*, image captions or questions.[3] Following Zhao et al. (2017a) and Burns et al. (2018), images are labelled as 'Male' if the majority of its captions include a word from a set of male-related tokens (*e.g.*, BOY), and no caption includes a word from the set of female-related tokens (*e.g.*, GIRL); and vice-versa for 'Female'. Images are labelled as 'Neutral' if most of the subjects are listed as gender-neutral (*e.g.*, PERSON), or if there is no majority gender mention in the texts. Finally, images are discarded from the analysis when the text mentions both male and female entities, or there are no people mentioned. This process can be applied to both pretraining data and downstream task data. See Appendix 5.8.1 for the complete word list.

5.4 Measuring Bias in V&L Models

5.4.1 Intrinsic Bias

When we measure the intrinsic bias of a model, we are interested in whether there are systematic differences in how phrases referring to demographic groups are encoded Beukeboom (2014). We can measure the intrinsic bias using the model's language modelling task, where the tokens related to grammatical gender are masked.[4]

Let M_D be a V&L model pretrained on corpora D. The masked words related to grammatical gender are categorised on $N = 3$ disjoint demographic

[3] Zhao et al. (2021) annotated samples from the COCO dataset Lin et al. (2014) with the perceived attributes (gender and skin-tone) of the people in the images. However, their gender labels agree on 66.3% of the images compared to caption-derived annotations. To be consistent across all datasets used in our project, we will not use their human-collected annotations for analysing gender bias on COCO.

[4] We define *gender* correlations as our case study of representational bias, but note that our methodology can be extended to analyse bias with regard to any protected attribute(s).

groups $A = \{\text{Male}, \text{Female}, \text{Neutral}\}$ based on reported visual appearance in the image. The gender associated with an image is considered as the ground truth (see previous section for more details). Let g_i for $i \in [1, N]$ be the categorical random variable corresponding to the presence of the group i. We investigate the **gender–context distribution**: the co-occurrence between attributes $A_i = \{a_1, \ldots, a_{|A_i|}\}$, *e.g.*, gender terms, for a demographic group g_i, and contextual words $T = \{t_1, \ldots, t_T\}$, *e.g.*, objects that appear in a given text. This results in a co-occurrence matrix $C_{a,t}^{g_i}$ that captures how often pairs of attribute–context words occur in a defined context S, *e.g.*, an image caption in a corpus $\mathcal{C}$. Formally, for every demographic group g_i, over the A_i attributes and T objects, and all possible contexts in corpus $\mathcal{C}$

$$C_{a,t}^{g_i} = \sum_{S \in \mathcal{C}} \sum_{j=1}^{|A_i|} \sum_{k=1}^{|T|} S(a_j, t_k) \quad \text{with } i \in [1, N], \tag{5.1}$$

where $S(a_j, t_k) = 1$ if the attribute and object co-occur, zero otherwise. Based on $C_{a,t}^{g_i}$, standard statistical metrics like precision, recall and F1 can be computed. In addition, we will quantify the bias amplification in a given model M_D to better understand the degree of bias exacerbated by the model. We use the metric presented by Wang and Russakovsky (2021), which is described in more detail in the next section.

5.4.2 Bias Amplification

We use the BiasAmp metric introduced by Wang and Russakovsky (2021), as it accounts for varying base rates of group membership and naturally decouples the direction of bias amplification: While $\text{BiasAmp}_{T \to A}$ measures the bias amplification due to the *task* influencing the protected *attribute* prediction,[5] $\text{BiasAmp}_{A \to T}$ measures the bias amplification due to the *protected attribute* influencing the task prediction. We give a concise treatise of $\text{BiasAmp}_{A \to T}$ here, and refer to Wang and Russakovsky (2021) for further details.

In our setup, the set of attributes $a \in A$ is given by $A = \{\text{Male}, \text{Female}, \text{Neutral}\}$, and the set of tasks (or objects) $t \in T$ are the most frequent nouns

[5] We do not consider gender prediction as a task per se, as gender –or any other sensitive attribute– prediction entangles a complex categorization and a moral debate Keyes (2018); Larson (2017). Instead, we use a MLM task as proxy and ask the model to predict the subject of a sentence given its context.

co-occurring with gendered terms in the training sets (see Appendix 5.8.1 for details). Denote by $P(T_t = 1)$ the probability that an example in the dataset belongs to class t. And, similarly, $P(\hat{T}_t = 1)$ the probability that an example in the dataset is labelled as class t by the model. Wang and Russakovsky (2021) introduce two terms to disambiguate the direction of bias amplification. The first term, Δ_{at}, quantifies the difference between the bias in the training data and the bias in model predictions.

The second term, y_{at}, identifies the direction of correlation of A_a with T_t; that is, y_{at} alters the sign of the Δ_{at} to correct for the fact that the bias can have two directions. Thereby,

$$\text{BiasAmp}_{A \to T} = \frac{1}{|A||T|} \sum_{\substack{a \in A \\ t \in T}} y_{at}\Delta_{at} - (1 - y_{at})\Delta_{at} \tag{5.2}$$

$\text{BiasAmp}_{A \to T}$ will be *positive if the model predictions amplify the prevalence of a class label* $t \in T$ *between groups* $a \in A$ *in the dataset*. For instance, bias is amplified if A_a = MALE images are more likely to appear in the presence of a T_t = SKATEBOARD in the model predictions, compared to the prior distribution from the dataset. In contrast, a negative value indicates that model predictions diminish the bias present in the dataset. A value of 0 implies that the model does not amplify the bias present in the dataset. Note that this does not imply that the model predictions are unbiased.

5.4.3 Extrinsic Bias & Fairness

The second phase of our analysis measures extrinsic bias amplification: downstream performance and fairness (group disparity). A given model is fine-tuned on downstream tasks that require different reasoning skills based on the image context. We evaluate model performance with respect to the three demographic groups defined in A and compare results in search of the more equitable system.

5.4.4 Gender-neutral Domain Adaptation

Motivated by the fact that models are known to acquire unwanted biases during pretraining Hall et al. (2022), we also investigate what happens if a model M_D is further pretrained for one additional epoch on gender-neutral

data, with the goal of creating a more gender-neutral model M_D^{N}. We hypothesize that this may be sufficient to reduce the biases encoded in the original model. Given a dataset D, a new dataset D^N is created by substituting gender-related tokens in the text for gender-neutral tokens. The substitution is based on a hand-crafted lexicon,[6] *e.g.*, *woman* or *man* may be substituted to *person*.[7] The new model M_D^{N} is used for both the intrinsic and extrinsic bias evaluations.

5.5 Experimental setup

5.5.1 Models

We take the LXMERT architecture Tan and Bansal (2019) as a popular representative of V&L models, and build our controlled analysis on `VOLTA` Bugliarello et al. (2021). `VOLTA` is an implementation framework that provides a fair setup for comparing V&L models pretrained under the same conditions, which enables us to compare the influence of diverse training data on representational bias. In this case, $\mathrm{LXMERT}_{\mathrm{180K}}$ refers to the original checkpoint and $\mathrm{LXMERT}_{\mathrm{3M}}$ to the model trained on CC3M Bugliarello et al. (2021). We also study ALBEF in two sizes and BLIP. Table 5.1 lists the models included in our analysis.

5.5.2 Gender-neutral Data

As a natural extension to study representational gender bias, we want to evaluate to what extent gender-neutral data helps to mitigate gender bias. Amend et al. (2021) showed that gender-neutral training might be a viable approach for reducing gender bias in image captioning models. We study its effect in more generic pretrained V&L models.

The gender-neutral pretraining data is the result of substituting terms with grammatical gender for gender-neutral equivalents, *e.g.*, "A woman walking her dog" translates into "A person walking their dog." To this end,

[6] See Appendix 5.8.1

[7] Note that when the pretraining data D is composed of multiple corpora, we argue that domain adaptation to a non-biased space should be performed only on *clean* data, and, therefore, $|D^N| \leq |D|$.

Model (M_D)	Gender-neutral model (M_D^{N})
$\text{LXMERT}_{\text{180K}}$	$\text{LXMERT}^{\text{N}}_{\text{180K}}$
$\text{LXMERT}_{\text{3M}}$	$\text{LXMERT}^{\text{N}}_{\text{3M}}$
ALBEF_{4M}	$\text{ALBEF}^{\text{N-COCO}}_{\text{4M}}$, $\text{ALBEF}^{\text{N-CC3M}}_{\text{4M}}$
$\text{ALBEF}_{\text{14M}}$	$\text{ALBEF}^{\text{N-COCO}}_{\text{14M}}$, $\text{ALBEF}^{\text{N-CC3M}}_{\text{14M}}$
$\text{BLIP}_{\text{129M}}$	$\text{BLIP}^{\text{N}}_{\text{129M}}$

Table 5.1: Summary of the models. The subscript in the model name indicates the number of images in the pretraining set. *All* gender-neutral models are pretrained with in-domain data ($\text{LXMERT}^{\text{N}}_{\text{180K}}$ and $\text{BLIP}^{\text{N}}_{\text{129M}}$ on COCO; $\text{LXMERT}^{\text{N}}_{\text{3M}}$ on CC3M). For models with more than one gender-neutral version, the superscript indicates the dataset used for gender-neutral pretraining.

we create a list of gender entities[8] by merging previous hand-curated lexicons used in a similar context to ours, provided by Antoniak and Mimno (2021).[9]

Starting from a pretrained checkpoint, we perform an extra epoch of pretraining. The training is done based on a linear function that increases the probability for a model to learn from gender-neutral captions. The starting rate is p=0.15 and, as the training progresses, the probability of getting a gender-neutral caption increases to p=1.0 at the last step. Note that as the probability of getting a gender-neutral caption increases, the learning rate decreases. This methodology supports our intuition that starting with a gender-neutral corpus would be too drastic for the model to adapt to, and instead cause catastrophic forgetting.

Finally, we continue pretraining the original model checkpoints for an extra epoch *without* the gender-neutral alternative (*i.e.*, p=0.0). The evaluation on this new checkpoint will help us to draw conclusions on longer training, as well as ensure the correct implementation of our setup.

[8]See Appendix 5.8.1 for the complete list.

[9]We deliberately omit tokens like 'actor' from the list if the female (or male) equivalent is not always used (people do not always use the word 'actress' when referring to a female character). We also discard 'male' and 'female' as we suspect that they are more often used on non-human entities.

5.5.3 Evaluation Tasks

For evaluation of downstream tasks, we report task performance and analyse group disparities. Bias amplification is reported on the validation splits.

MLM We follow standard practice for assessing gender bias in V&L models Zhao et al. (2017a); Burns et al. (2018); Wang et al. (2019); Tang et al. (2021); Srinivasan and Bisk (2021); Agarwal et al. (2021); Cho et al. (2022) and expose representational bias in a masked language modelling (MLM) task. The words masked are gendered terms given by the same lexicon used in §5.5.2. Personal pronouns (if any) are also masked to avoid leaking gender information into the model representation. For example, "A woman walking her dog" would be masked as "A [MASK] walking [MASK] dog". The image associated with each sentence is also input to the model, in a setup that reflects the pretraining conditions.

We investigate the intrinsic bias of the models as detailed in §5.3, *i.e.*, we look at the co-occurrence of context words (*e.g.*, car, ball) with particular word choices from the model (*e.g.*, gender words like woman, child). Previous work Sedoc and Ungar (2019); Antoniak and Mimno (2021); Delobelle et al. (2021) showcases how the measure of bias can be heavily influenced by the choice of target seed words. To avoid misleading results from low frequency words, we define the set of target words to be the 100 most frequent common nouns that co-occur with the gender entities in the corresponding training data. Table 5.2 provides a summary of gender distribution.

To evaluate intrinsic bias, we do not look at the exact word prediction but instead consider two options to annotate the gender of the predicted word. First, we can extract and sum the probabilities of *all* male, female and gender-neutral tokens within our set to select the most probable gender entity. However, given that the distributions of tokens follows Zipf's Law, the probability mass computed for each gender group is nearly equal, yielding inconclusive results. Therefore, we use the gender category of the most probable token. Then, the bias present in model predictions is measured with the statistical and bias amplification metrics presented in §5.4.2.

Visual Question Answering VQA Antol et al. (2015) requires the model to predict an answer given an image and a question. LXMERT formulates VQA as a multi-answer classification task, and ALBEF and BLIP treat it as a language generation task. We evaluate models on the VQAv2 Goyal

	COCO	CC3M	VQAv2	GQA	NLVR2	F30K
	Image	Image	Question	Question	Sentence	Image
Male	725	901	20000	8265	91	345
Female	363	945	9498	4860	99	207
Neutral	1187	1095	18549	4442	377	336
Total	2275	2941	48047	17567	567	889

Table 5.2: Gender distribution across validation splits in each dataset. Note that for COCO, this refers to the minival split in Tan and Bansal (2019). COCO and F30K have five captions per image. Gender was inferred from image captions for COCO, CC3M and F30K. Gender was inferred from questions in VQAv2, GQA and from the sentence given in NLVR2.

et al. (2017) and GQA datasets Hudson and Manning (2019), and report performance as VQA-Score and accuracy, respectively.

Bias amplification is measured on the subset of question–answer pairs targeting people. Gender is inferred from the question, considering all the gender entities presented in Appendix 5.8.1. We filter any answer category whose answer does not occur with gender entities at least 50 times in the training set. Finally, numerical and yes/no question-answer pairs are also removed leaving a total of 165 answer categories in VQAv2 and 214 in GQA.

Natural Language for Visual Reasoning NLVR2 Suhr et al. (2019) requires the model to predict whether a text describes a pair of images. The notion of bias amplification considered in this project would require us to manually annotate the gender from all the images to be able to extract gender-context patterns from the training data. For this reason, we only evaluate the group disparity in NLVR2 through differences in performance, reported as accuracy.

Image–Text Retrieval This retrieval task contains two subtasks: text-to-image retrieval (IR), where we query the model with a caption to retrieve an image, and image-to-text retrieval (TR), where we use an image to retrieve a suitable caption. We report Recall@1 on the Flickr30K Plummer et al. (2015) benchmark. Bias amplification is measured on the subset of data targeting people. In IR, we query the model with captions that include a

word from the set of male-related or female-related tokens and compare to the gender annotated in the image retrieved. In TR, we query the model with images annotated as 'Male' or 'Female' and compare to the gendered terms in the caption retrieved. Captions with gender-neutral terms are treated as a separate case to assess how often the models retrieve images from each group, yet the image retrieved could be potentially valid for any gender case. In both subtasks, we consider that the model does not amplify gender bias when the image or caption retrieved has a gender-neutral subject.

5.6 Results

5.6.1 Intrinsic Bias

We evaluate intrinsic bias in encoder-only models. Considering that bias varies as a function of the bias in a dataset, amongst other variables Hall et al. (2022), we define our experiments with LXMERT variants as our *control setup*: the same model architecture is trained with the same hyperparameters on disjoint corpora yielding two versions of the model, LXMERT_{180K} and LXMERT_{3M}.

Gender-neutral pretraining mitigates gendered outputs Figure 5.2 shows results for LXMERT_{180K} models; complete results are in Appendix 5.8.3. A model is penalised when it predicts a token from the opposite gender, but we consider a gender-neutral term as a valid output.[10] The models pretrained with gender-neutral data, have near perfect F1 performance as they learnt to predict gender-neutral tokens when their standard counterparts, LXMERT_{180K} and LXMERT_{3M}, had low confidence on the most probable token.[11] We presume these are images where the visual appearance of the main subject is unclear. Interestingly, the trade-off between precision and recall has opposite directions for Female and Male groups *vs* Neutral in LXMERT_{180K} and LXMERT_{3M}: the models tend to

[10] Predicting a gender-neutral term shows that the model understands the depicted visual concept at the generic level.

[11] The models do not *forget* to predict gender-related tokens. LXMERT^{N}_{180K} predicts ∼37% of the time a word from the set of neutral-related tokens (compared to ∼20% in LXMERT_{180K}).

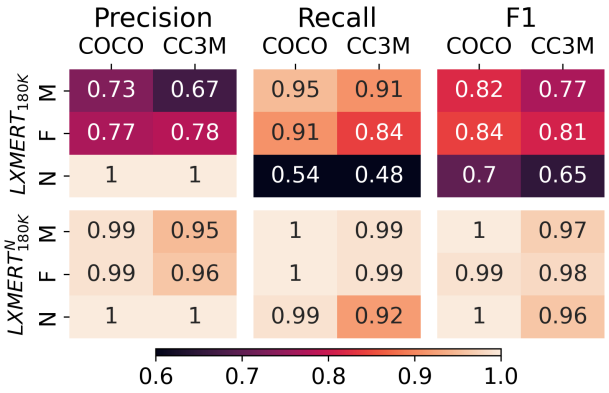


Figure 5.2: Statistical analysis of gender bias in MLM with gendered terms masked. Predicting a token from the gender-neutral set is always considered correct (Precision=1). Models report higher recall scores for Male (M) and Female (F) groups, showcasing the completeness of positive predictions; it is the opposite for Neutral (N) tokens.

output female- and male- tokens more often than neutral-related, even when the subject in the image was annotated as gender neutral (low recall).

Pretrained models reflect training data biases Table 5.3 shows the aggregated bias amplification measured in encoder-only model variants. Our bias mitigation strategy has the same consistent behaviour across LXMERT models and evaluation data (COCO or CC3M): models tend to reflect the same degree of bias present in the data ($\text{BiasAmp}_{T \to A}$ closer to zero). $\text{ALBEF}^{\text{N-COCO}}_{\text{14M}}$ and $\text{ALBEF}^{\text{N-CC3M}}_{\text{14M}}$ models benefit from pretraining on gender-neutral data differently, as both decrease the overall bias amplification. Wang and Russakovsky (2021) caution against solely reporting the aggregated bias amplification value, as it could obscure attribute-task pairs that exhibit strong bias amplification. We report it here as a relative metric to compare the overall amplified bias between the models, and should not be considered in its own. See Appendix 5.8.3 for results broken down by gender.

We also investigated the equivalent to $\text{LXMERT}^{\text{N}}_{\text{3M}}$, but pretrained on gender-neutral data for a reduced number of steps to match those in $\text{LXMERT}^{\text{N}}_{\text{180K}}$. We verified that more pretraining steps on gender-neutral data equates to a reduced bias amplification in absolute terms.

	LXMERT_{180K}	LXMERT^{N}_{180K}	LXMERT_{3M}	LXMERT^{N}_{3M}	ALBEF_{14M}	$\text{ALBEF}^{\text{N-COCO}}_{14M}$	$\text{ALBEF}^{\text{N-CC3M}}_{14M}$
COCO	-.0359	-.0008	-.0617	-.0014	-.0742	-.0517	-.0792
CC3M	-.0346	-.0062	-.0007	-.0002	-.0182	-.0367	-.0570

Table 5.3: $\text{BiasAmp}_{T\rightarrow A}$ averaged over attributes (gender entities) and tasks (top-100 nouns) for LXMERT and ALBEF_{14M} models. Light and dark backgrounds indicate bias amplification measured in-domain and out-of-domain data respectively. Negative values indicate an overall decrease of the bias in model's predictions.

5.6.2 Extrinsic Bias & Fairness

Trade-offs in task performance Downstream performance on the test sets is shown in Table 5.4. LXMERT_{180K} may require more pretraining steps to converge, as we verify that the performance improvement observed in LXMERT^{N}_{180K} is mainly due to the extra pretraining steps regardless of gender-neutral data. Our strategy for mitigating gender bias on pretrained models generally leads to lower task performance on NLVR2 and image retrieval, revealing a *trade-off between bias mitigation and task performance*. The same trade-off has been observed in language models He et al. (2022); Chen et al. (2023b). However, gender-neutral models report similar or even superior performance on question answering and text retrieval tasks compared to their original versions.

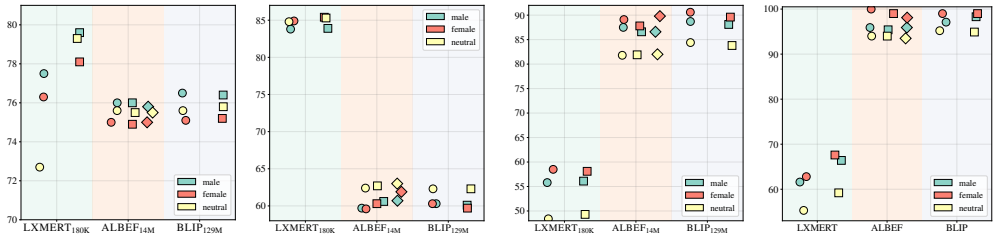
Gender-neutral models consistently reduce group disparity Group performance is depicted in Figure 5.3 for a subset of models and tasks. Table 5.7 in Appendix 5.8.4 shows the complete results. We observe that group disparity is consistently reduced on VQAv2 and retrieval tasks. An exception are LXMERT models, which show a minor, undesirable increase in group disparity on VQAv2, GQA and text retrieval tasks. For instance, in question-answering tasks with LXMERT, we observe a reduction in the min-max gap of 4.5 (LXMERT^{N}_{180K}) points in VQAv2, while the min-max gap increase in GQA is *only* of 0.4 points. Note that Tan and Bansal (2019) pretrained LXMERT_{180K} on GQA train and *validation* data, which results in a very high performance (∼85.0 for all groups) on the GQA validation set. We speculate that the gains in performance equality across groups could be due to a shift of the final word representations to a more equidistant vector space between gendered terms and their context. That is, the conditional probability

	VQAv2	**GQA**	**NLVR2**	**F30K**	
	test-dev	test-dev	test-P	test IR	test TR
LXMERT_{180K}	70.3	**59.4**	**74.5**	53.0	61.1
LXMERT^{N}_{180K}	**71.6**	59.3	**74.5**	**53.9**	**66.2**
LXMERT_{3M}	67.2	55.4	**71.5**	**54.4**	**59.5**
LXMERT^{N}_{3M}	**68.1**	**56.0**	70.0	50.2	57.4
ALBEF_{4M}	**72.9**	**56.6**	**79.3**	**82.6**	93.3
$\text{ALBEF}^{\text{N-COCO}}_{4M}$	**72.9**	56.3	77.1	82.5	94.0
$\text{ALBEF}^{\text{N-CC3M}}_{4M}$	**72.9**	**56.6**	78.4	82.4	**94.2**
ALBEF_{14M}	**74.4**	**58.4**	**82.4**	**85.9**	95.1
$\text{ALBEF}^{\text{N-COCO}}_{14M}$	74.1	57.3	52.3[12]	85.5	**95.4**
$\text{ALBEF}^{\text{N-CC3M}}_{14M}$	74.1	58.1	81.0	85.1	95.2
BLIP_{129M}	**75.3**	58.1	**79.7**	**87.5**	**96.7**
BLIP^{N}_{129M}	75.2	**58.3**	79.3	86.9	96.2

Table 5.4: Test results for a model M_D and its gender-neutral version M_D^{N}. We report VQA-accuracy in VQAv2, accuracy in GQA and NLVR2, and Recall@1 in F30K. Results for original models computed by us.

distribution of a gendered term given its context is smoother across different demographic groups. We leave exploration of this for future work. In recent work, Feng et al. (2023) continued pretraining language models on partisan corpora and observed that these models *do* acquire (political) bias from said corpora. In our case, the continued pretraining could make the M_D^{N} models more robust regarding gendered terms.

Gender-neutral training reduces fine-tuning variance Dodge et al. (2020) and Bugliarello et al. (2021) analysed the impact of random seeds in fine-tuning. We do this analysis on our control setup and observe that gender-neutral variants of LXMERT consistently report lower variance in performance on all tasks, except for NLVR2. We, however, observe a strong variance in the fine-tuning process for NLVR2 due to the random weight initialisation of the classification layer. See Appendix 5.8.5 for specific results



(a) VQAv2 (b) GQA (c) F30K - IR (d) F30K - TR

Figure 5.3: Validation-set results of selected models (○: $\text{LXMERT}_{180\text{K}}$, $\text{ALBEF}_{14\text{M}}$ and $\text{BLIP}_{129\text{M}}$) and their gender-neutral version (□: $\text{LXMERT}^{\text{N}}_{180\text{K}}$, $\text{ALBEF}^{\text{N-COCO}}_{14\text{M}}$ and $\text{BLIP}^{\text{N}}_{129\text{M}}$, ◊: $\text{ALBEF}^{\text{N-CC3M}}_{14\text{M}}$). We report VQA-accuracy in VQAv2, accuracy in GQA, and Recall@1 in F30K by gender group: male (M), female (F), and neutral (N).

across 6 runs.

Intrinsic & extrinsic bias are independent We estimate bias amplification in VQA tasks by evaluating the fluctuations in models' predictions when they differ from the correct answer. Otherwise, the models are said to not amplify the bias from the data. We find that *all* model variants – M_D and M^{N}_D – reduce the gender bias across tasks. However, contrary to what we observed in pretrained models (Table 5.3), there is no evidence that the gender-neutral pretraining influenced positively (nor negatively) the extrinsic bias of the models: it depends on the model, downstream task and gender group (see Appendix 5.8.5 for results on $\text{BiasAmp}_{A\to T}$ fine-tuning variance). Figure 5.4 displays $\text{BiasAmp}_{A\to T}$ broken-down by gender category measured on GQA for a subset of models. Whereas the degree of bias amplification is fairly consistent between a model M_D and M^{N}_D in VQAv2 (see Appendix 5.8.4), there is higher variance in GQA: $\text{ALBEF}^{\text{N-COCO}}_{14\text{M}}$ reduces the bias amplification compared to $\text{ALBEF}_{14\text{M}}$, but we observe the opposite effect on $\text{BLIP}^{\text{N-COCO}}$.

In retrieval tasks, we look into models' behavior when querying them with neutral instances. Regardless of the degree of intrinsic bias in the model, models exhibit the same trend: in IR, all models mostly retrieve images

[12]This result is inexplicably low, despite fifteen attempts at fine-tuning with different random seeds. We saw similar instabilities when fine-tuning the released LXMERT models, but we found seeds that gave above-chance accuracy.

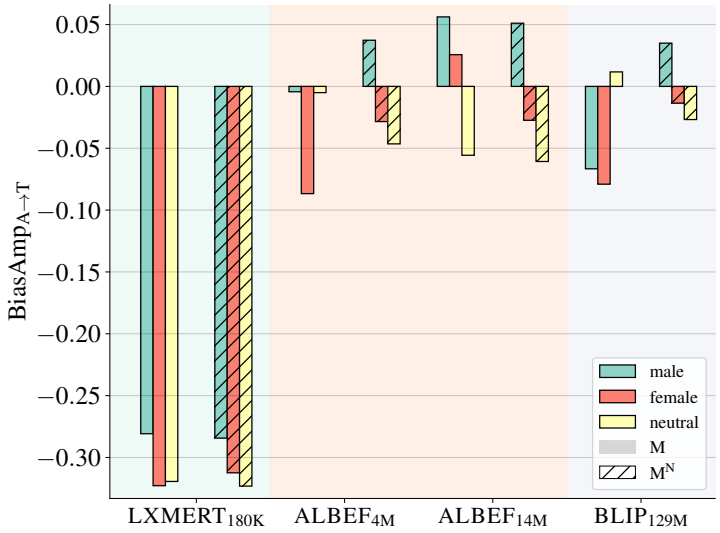


Figure 5.4: Bias amplification measured on question-answering (GQA) broken down by gender group. M^{N} are gender-neutral pretrained on COCO.

labeled as 'Neutral', but twice as much 'Male' images as 'Female'. We find similar results for TR, *i.e.*, query images whose main actor is defined as Neutral, but, in this scenario, only half of the captions retrieved relate to people. See Appendix 5.8.4 for detailed results.

5.7 Conclusion

This paper presented a comprehensive analysis of gender bias amplification and fairness of encoder-only and encoder-decoder V&L models. The intrinsic bias analysis shows consistent results – in terms of bias mitigation – in models trained on gender-neutral data, even if these models reflect biases present in data instead of diminishing them (as we observed with LXMERT). In line with previous findings in language models Goldfarb-Tarrant et al. (2021); Kaneko et al. (2022); Orgad et al. (2022), intrinsic bias in V&L models does not necessarily transfer to extrinsic bias on downstream tasks. Similarly, we find that the bias in a model and its empirical fairness –group disparity on task performance– are in fact independent matters, which is in line with the NLP literature Shen et al. (2022); Cabello et al. (2023b). Intrinsic bias can potentially reinforce harmful biases, but these may not impact the treatment

of groups (or individuals) on downstream tasks. We believe that bias and fairness should always be carefully evaluated as separate matters. One of they key findings of our work is that the extra pretraining steps on gender-neutral data are beneficial to reduce the group disparity in every model architecture tested on VQAv2, and in the majority of models for both retrieval tasks. Crucially, there is no penalty to pay for this fair outcome: the overall task performance of gender-neutral models is similar or better than their original versions.

Limitations

The framework to characterize gender bias in V&L presented in this study is general and extensible to analyse other forms of bias in multimodal models. We consider three base architectures to settle on the implementation. However, our work would benefit from analyzing a wider range of models. Studying the effects of gender-neutral pretraining on V&L models with a frozen language model, such as ClipCap (Mokady et al., 2021) and BLIP-2 (Li et al., 2023), is left as future work.

Due to computational limitations, we restricted most of our analysis to single runs. We perform a first analysis across multiple random seeds for LXMERT models in Appendix 5.8.5. There, we notice that gender-neutral models seem to have lower variance after fine-tuning. Yet, the cross-seed performance of a given model can fluctuate considerably for some tasks (*e.g.*, NLVR2), corroborating previous findings from Bugliarello et al. (2021). Likewise, bias amplification, along with other fairness metrics like group disparity, often fluctuates across runs. We report bias amplification variance in fine-tuning of LXMERT models, but the absence of confidence intervals for all models and tasks –due to the same reason stated above– should be considered. We hope to motivate future work to address this issue.

Moreover, despite the existence of multilingual multimodal datasets (Elliott et al., 2016; Liu et al., 2021; Bugliarello et al., 2022, *inter-alia*), our experimental setup is limited to English datasets and models. Studies of (gender) bias using only English data are not complete and might yield inaccurate conclusions, albeit overcoming the structural pervasiveness of gender specifications in grammatical gender languages such us German or Spanish is not trivial Gabriel et al. (2018). Likewise, our work considers a single dimension of social bias (gender). Further research on analyzing social biases

on V&L models should account for intersectionality: how different social dimensions, *e.g.*, gender and race, can intersect and compound in ways that can potentially impact model performance on most disfavoured groups, *e.g.*, Black Women as discussed in Crenshaw (1989).

Ethics Statement

The models and datasets used in this study are publicly available, and we strictly follow the ethical implications of previous research related to the data sources. Our work is based on sensitive information such as gender, based on reported visual appearance in the image captions. We would like to emphasize that we are not categorizing biological sex or gender identity, but rather using the given image captions as proxies to the outward gender appearance.

Acknowledgments

We are grateful to Benjamin Rotendahl and Rita Ramos for initial discussions about data and evaluation. We also thank members of CoAStaL and LAMP groups for their valuable feedback. Laura Cabello is funded by the Novo Nordisk Foundation (grant NNF 20SA0066568). Emanuele Bugliarello is supported by the funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 801199. Stephanie Brandl is funded by the European Union under the Grant Agreement no. 10106555, FairER. Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or European Research Executive Agency (REA). Neither the European Union nor REA can be held responsible for them. This work was supported by a research grant (VIL53122) from VILLUM FONDEN.

5.8 Appendix

5.8.1 Seed words

Gender terms

- **Female:** aunt, bride, businesswoman, daughter, daughters, fiancee, fiancée, gal, gals, girl, girlfriend, girls, grandmother, her, herself, lady, landlady, mama, mom, mother, queen, she, sister, sisters, spokeswoman, wife, woman, women, womens.
- **Male:** boy, boyfriend, boys, brother, brothers, businessman, dad, dude, dudes, father, fiance, fiancé, gentleman, grandfather, groom, guy, he, him, himself, his, husband, king, landlord, man, men, mens, papa, son, sons, spokesman, uncle.
- **Neutral:** businessperson, child, childs, grandparent, kid, kids, landlord, monarch, newlywed, parent, partner, pbling, people, person, sibling, siblings, someone, spokesperson, spouse, their, them, themself, they.

Gender-neutral mappings Using the gender terms listed above, we generate mappings from male and female to neutral terms: see Table 5.5 for details. These mappings are used to continue pre-training on gender-neutral (debiased) data as explained in §5.5.2.

Objects List of top-100 most frequent nouns co-occurring with gender terms in the training split in COCO Lin et al. (2014) and Conceptual Captions (CC3M) Sharma et al. (2018).

- **COCO:** tennis, group, street, baseball, table, dog, front, ball, player, field, snow, game, beach, horse, skateboard, umbrella, water, phone, kite, hand, top, board, ski, couple, motorcycle, food, elephant, People, picture, pizza, surfboard, room, shirt, bench, wave, frisbee, court, park, air, cake, bed, laptop, train, cell, racket, bat, bus, kitchen, plate, glass, ocean, side, grass, giraffe, building, city, skier, road, car, suit, trick, cat, tie, tree, bike, photo, boat, hat, slope, baby, area, sign, chair, sidewalk, computer, hill, head, surfer, mountain, video, skateboarder,

soccer, truck, banana, couch, camera, skate, crowd, lot, snowboard, background, wine, bear, day, back, luggage, cow, living, fence, ramp.

- **CC3M:** player, team, actor, football, game, artist, hand, day, match, background, dress, beach, car, photo, dog, event, street, home, ball, wedding, family, city, film, time, tree, award, goal, hair, front, night, water, baby, business, illustration, politician, sport, show, way, portrait, face, book, premiere, fan, room, head, friend, year, athlete, park, house, fashion, soccer, character, flower, country, style, field, side, party, festival, picture, stage, rock, eye, couple, world, shirt, vector, camera, pop, tv, ceremony, hat, glass, snow, horse, school, road, phone, arm, art, window, crowd, sea, table, part, boat, suit, basketball, model, top, birthday, star, student, view, tennis, smile, wall, celebrity, baseball.

5.8.2 Models

In this section, we provide an overview on the models we use in our evaluation. We refer to their original work for more details.

LXMERT Tan and Bansal (2019) is a cross-modal architecture pretrained to learn vision-and-language representations. It consists of three Transformer Vaswani et al. (2017a) encoders, where visual and language inputs are encoded separately in two independent stacks of Transformer layers before feeding them into the cross-modality encoder. The cross-modality encoder uses bi-directional cross attention to exchange information and align the entities across the two modalities. LXMERT is trained with four objectives: masked language modelling (MLM), masked object prediction, image–text matching (ITM) and image question answering.

Similar to LXMERT, **ALBEF** Li et al. (2021a) is a dual-stream encoder Bugliarello et al. (2021) that first learns separate visual and textual embeddings using Transformer-based image and text encoders; and then fuses them in a cross-modal Transformer using image–text contrastive loss (ITC), which enables a more grounded vision and language representation learning. The model is pretrained with two other objectives: masked language modelling (MLM) and image–text matching (ITM) on the multimodal encoder. Unlike LXMERT, ALBEF does not rely on image features extracted from an off-the-shelf object detector, but directly feeds the raw image into a Vision Transformer (Dosovitskiy et al., 2021)

Male	**Female**	**Neutral**
boy	girl	child
boyfriend	girlfriend	partner
boys	girls	kids
brother	sister	sibling
brothers	sisters	siblings
businessman	businesswoman	businessperson
dad	mom	parent
dude	gal	person
dudes	gals	people
father	mother	parent
fiance	fiancee	partner
fiancé	fiancée	partner
gentleman	lady	person
grandfather	grandmother	grandparent
groom	bride	newlywed
guy	gal	person
he	she	they
him	her	them
himself	herself	themself
his	her	their
husband	wife	spouse
king	queen	monarch
landlord	landlady	landlord
man	woman	person
men	women	someone
mens	womens	people
papa	mama	parent
son	daughter	kid
sons	daughters	childs
spokesman	spokeswoman	spokesperson
uncle	aunt	pbling

Table 5.5: Gender-neutral mappings used for continual pre-training in gender-neutral data as described in §5.5.2.

BLIP Li et al. (2022) is a versatile model based on a multimodal mixture of encoder–decoder network, that can be applied to a wide range of downstream tasks. The authors introduce a novel boostrapping method to generate synthetic captions and remove noisy pairs from large-scale web data. Unlike LXMERT and ALBEF, BLIP is trained with an autoregressive language modelling objective that allows the generation of coherent captions given an image. The model is also pretrained using the unimodal image–text contrastive loss (ITC) and the cross-modal image–text matching (ITM) loss used by ALBEF.

5.8.3 Bias in Pretrained Models

Intrinsic bias Figure 5.5 complements Figure 5.2 from the main paper showing statistical results measured on the intrinsic bias analysis in our *control setup*.

MLM experiment broken down by gender Table 5.6 provides a more granular look at which gender groups are actually amplifying/decreasing the bias in the pretrained models.

COCO

	Male	Female	Neutral
$\text{LXMERT}_{\text{180K}}$	-.0295	-.0048	-.0733
$\text{LXMERT}^{\text{N}}_{\text{180K}}$	-.0004	-.0008	-.0014
$\text{LXMERT}_{\text{3M}}$	-.0577	-.0230	-.1062
$\text{LXMERT}^{\text{N}}_{\text{3M}}$	-.0014	+.0001	-.0028
$\text{LXMERT}^{\text{N-SC}}_{\text{180K}}$	-.0082	-.0009	-.0109
ALBEF_{4M}	-.1006	-.0517	-.1083
$\text{ALBEF}^{\text{N-COCO}}_{\text{4M}}$	-.0748	-.1293	-.1529
$\text{ALBEF}^{\text{N-CC3M}}_{\text{4M}}$	-.0754	-.0337	-.1073
$\text{ALBEF}_{\text{14M}}$	-.0418	-.1146	-.0663
$\text{ALBEF}^{\text{N-COCO}}_{\text{14M}}$	-.0559	-.0169	-.0824
$\text{ALBEF}^{\text{N-CC3M}}_{\text{14M}}$	-.0556	-.0983	-.0837

CC3M

	Male	Female	Neutral
$\text{LXMERT}_{\text{180K}}$	-.0281	-.0276	-.0482
$\text{LXMERT}^{\text{N}}_{\text{180K}}$	+.0008	-.0081	-.0113
$\text{LXMERT}_{\text{3M}}$	-.0043	-.0030	+.0055
$\text{LXMERT}^{\text{N}}_{\text{3M}}$	+.0002	+.0004	-.0011
$\text{LXMERT}^{\text{N-SC}}_{\text{180K}}$	-.0011	+.0003	-.0012
ALBEF_{4M}	-.0473	-.0569	-.0422
$\text{ALBEF}^{\text{N-COCO}}_{\text{4M}}$	-.0329	-.0514	-.0152
$\text{ALBEF}^{\text{N-CC3M}}_{\text{4M}}$	-.0295	-.0497	-.0313
$\text{ALBEF}_{\text{14M}}$	+.0159	-.0642	-.0062
$\text{ALBEF}^{\text{N-COCO}}_{\text{14M}}$	-.0290	-.0250	-.0561
$\text{ALBEF}^{\text{N-CC3M}}_{\text{14M}}$	-.0535	-.0641	-.0534

Table 5.6: $\text{BiasAmp}_{T\rightarrow A}$ (BA.) per gender group, averaged over tasks (top-100 nouns) for LXMERT and ALBEF models, evaluated on validation splits on COCO (top) and CC3M (bottom). Light and dark backgrounds indicate bias amplification measured within in-domain and out-of-domain data respectively. A model amplifies the bias in the dataset if the value is positive. A negative value indicates an overall decrease of the bias in model's predictions.

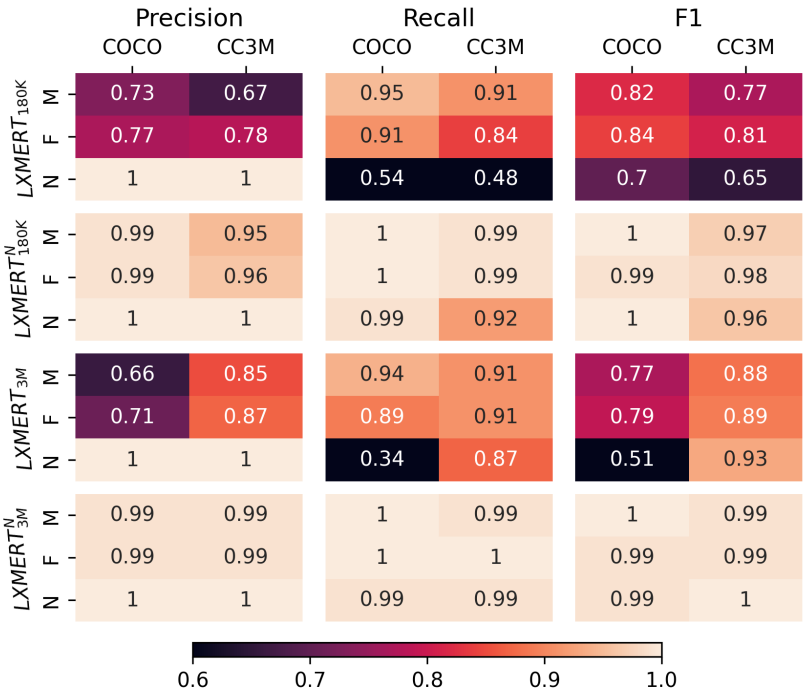


Figure 5.5: Statistical analysis of gender bias found through masked language modelling with gendered terms masked. Prediction of a token from the gender-neutral set is always considered correct (Precision=1). Models report higher recall scores for Male (M) and Female (F) groups, showcasing the completeness of positive predictions, whereas it is the opposite for Neutral-related (N) tokens.

5.8.4 Bias & Fairness in Downstream Tasks

Extrinsic Bias The following graphs complement results shown in § 5.6.2 for bias amplification measured on downstream tasks: Figure 5.6 shows results on GQA; Figure 5.7 shows results on VQAv2; Figure 5.8 and Figure 5.9 show the bias revealed on image–text retrieval tasks when querying the models with a gender-neutral caption (or image), respectively.

Task performance & Fairness We present granular results on task performance in validation in Table 5.7 and group disparity, defined as the min-max difference between group performance (Δ).

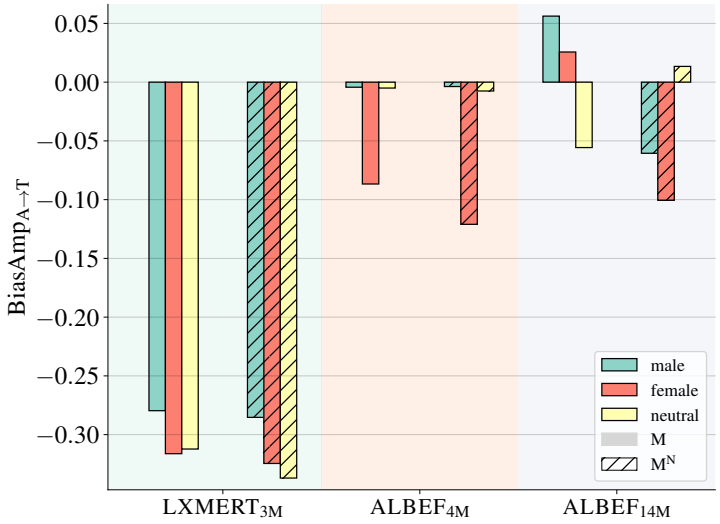
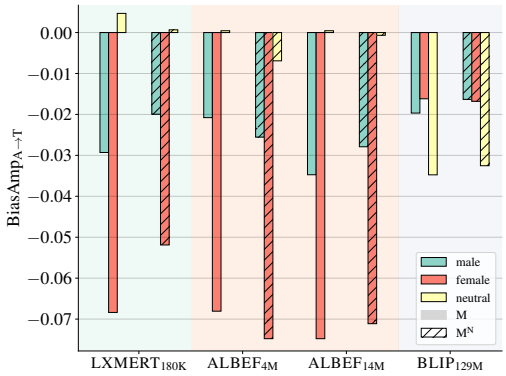
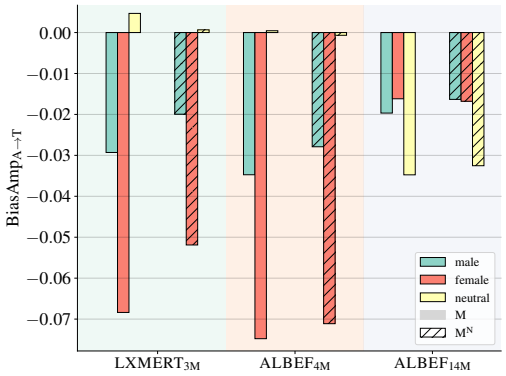


Figure 5.6: Bias amplification measured on question-answering (GQA) broken down by gender group. M^N are gender-neutral pretrained on CC3M.



(a) M^N are gender-neutral models pretrained on COCO.

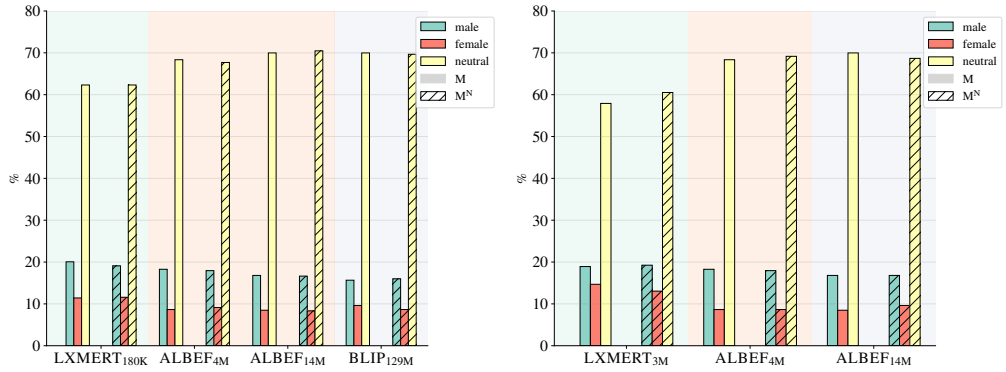


(b) M^N are gender-neutral models pretrained on COCO.

Figure 5.7: Bias amplification measured on question-answering (VQAv2) broken down by gender group.

5.8.5 Variance in fine-tuning

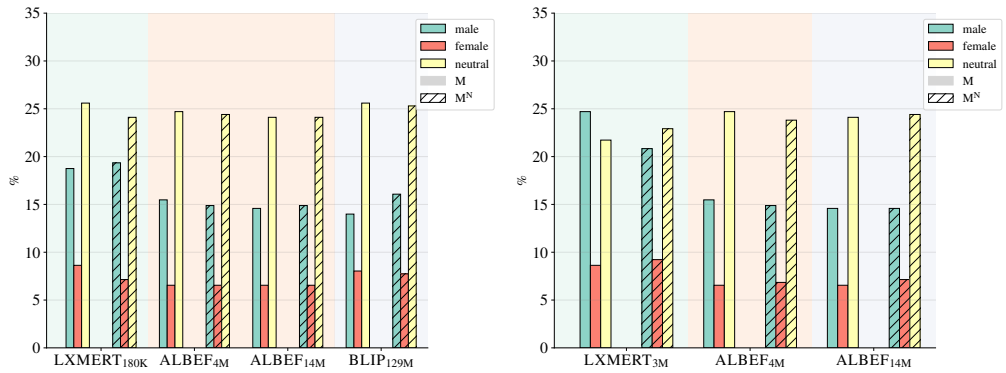
Table 5.8 shows the mean and standard deviation in bias amplification when fine-tuning LXMERT models with different random seeds. The variance



(a) IR - M^{N} are gender-neutral models pretrained on COCO.

(b) IR - M^{N} are gender-neutral models pretrained on CC3M.

Figure 5.8: Extrinsic bias measured on text-to-image retrieval (IR) on Flickr30K. Bias is measured as the percentage of images retrieved from each group when querying the models with a gender-neutral caption.



(a) TR - M^{N} are gender-neutral models pretrained on COCO.

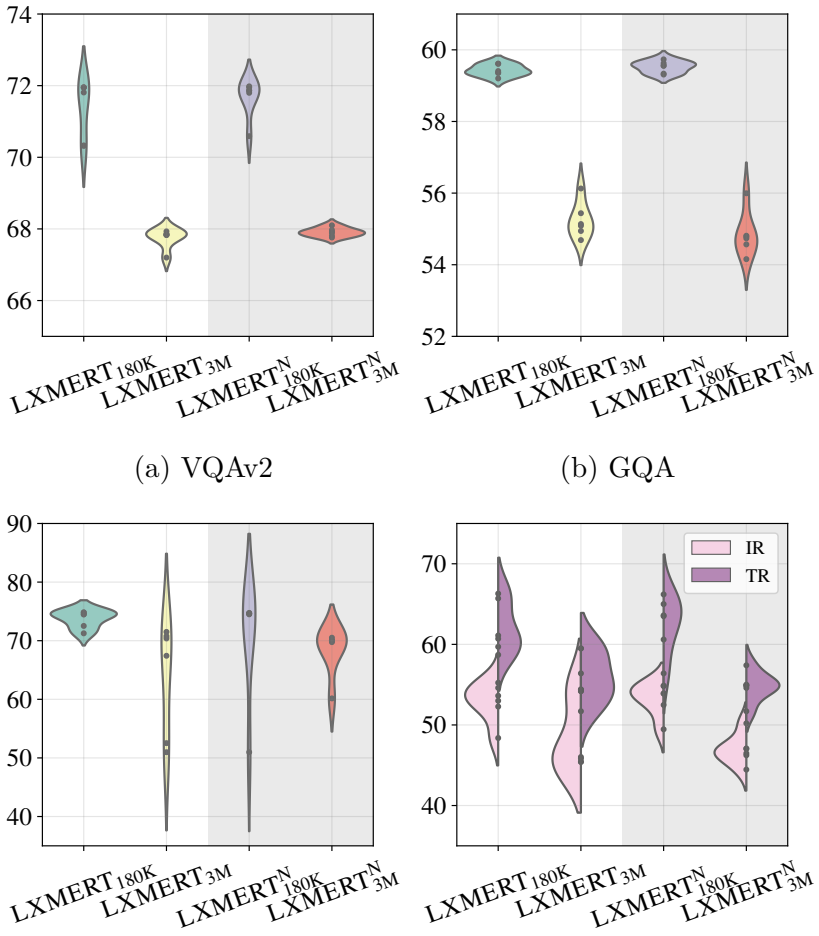
(b) TR - M^{N} are gender-neutral models pretrained on CC3M.

Figure 5.9: Extrinsic bias measured on image-to-text retrieval (TR) (c)-(d) on Flickr30K. Bias is measured as the percentage of captions retrieved from each group when querying the models with a gender-neutral image.

is due to random initialization. In line with what we observed in §5.6.2, there is no clear trend when comparing a model M with it's gender-neutral pretraining counterpart, M_D^{N}.

Figure 5.10 shows violin plots of the distribution of results when fine-

tuning LXMERT models with different random seeds. The variance is due to random initialization. Gender-neutral models reveal lower standard deviation across tasks. This finding reveals one of the benefits to perform extra steps of pretraining on gender-neutral data: to reduce variance in downstream performance. This observation aligns with the NLP literature showing that biases in a model are independent from model performance Cabello et al. (2023b).



(a) VQAv2 (b) GQA

(c) NLVR2 (d) F30K

Figure 5.10: Fine-tuning variance of LXMERT models across tasks. On the left with white background, original models (M_D). On the right with darker background, models after gender-neutral pretraining (M_D^{N}). Each model is fine-tuned 6 times on each task. The dots represent the experimental observations. We report average VQA-accuracy in VQAv2, accuracy in GQA and NLVR2, and recall@1 in F30k.

		VQAv2		GQA		NLVR2		F30K			
		Acc.	Δ(↓)	Acc.	Δ(↓)	Acc.	Δ(↓)	r@1 IR	Δ(↓)	r@1 TR	Δ(↓)
LXMERT_{180K}	M	77.5		83.8		81.9		55.8		61.6	
	F	76.3	4.8	84.9	**1.1**	75.0	6.9	58.5	10.1	62.8	**7.5**
	N	72.7		84.8		81.1		48.4		55.3	
LXMERT^{N}_{180K}	M	79.6		83.9		79.3		56.1		66.4	
	F	78.1	**1.5**	85.4	1.5	74.2	**6.0**	58.1	**8.8**	67.6	8.4
	N	79.3		85.3		80.2		49.3		59.2	
LXMERT_{3M}	M	68.4		63.7		72.3		56.0		63.0	
	F	66.5	6.0	65.6	**1.9**	64.0	14.8	59.4	8.2	63.7	**8.7**
	N	62.4		64.5		78.8		51.2		55.0	
LXMERT^{N}_{3M}	M	70.3		64.6		79.3		51.4		56.2	
	F	67.9	**2.4**	66.8	2.3	67.7	**11.6**	52.3	**5.0**	61.4	9.0
	N	70.1		64.5		78.8		47.3		52.4	
ALBEF_{4M}	M	75.1		60.0		87.9		83.1		94.5	
	F	73.7	1.4	61.5	2.5	76.8	11.1	87.9	10.1	98.1	7.0
	N	74.3		62.5		79.6		77.8		91.1	
$\text{ALBEF}^{\text{N-COCO}}_{4M}$	M	75.0		60.5		85.7		83.7		94.2	
	F	73.6	1.4	60.7	**1.6**	75.8	9.9	87.2	10.1	96.6	**5.2**
	N	74.0		62.1		79.8		77.1		91.4	
$\text{ALBEF}^{\text{N-CC3M}}_{4M}$	M	75.0		61.0		84.6		82.2		94.8	
	F	73.7	**1.3**	60.7	2.0	74.8	**9.8**	87.1	**9.2**	97.6	7.7
	N	74.5		62.7		79.3		77.9		89.9	
ALBEF_{14M}	M	76.0		59.7		86.8		87.5		95.9	
	F	75.0	1.0	59.6	2.8	79.8	**7.0**	89.1	7.3	100.0	6.0
	N	75.6		62.4		81.2		81.8		94.0	
$\text{ALBEF}^{\text{N-COCO}}_{14M}$	M	76.0		60.6		60.4		86.6		95.4	
	F	74.9	1.1	60.3	2.4	52.5	7.9	87.8	**5.9**	99.0	5.0
	N	75.5		62.7		57.6		81.9		94.0	
$\text{ALBEF}^{\text{N-CC3M}}_{14M}$	M	75.8		60.7		87.9		86.6		95.9	
	F	75.0	**0.8**	61.9	**2.3**	77.8	10.1	89.8	7.8	98.1	**4.6**
	N	75.5		63.0		81.7		82.0		93.5	
BLIP_{129M}	M	76.5		60.3		82.4		88.7		97.1	
	F	75.1	1.4	60.3	**2.0**	77.8	**6.0**	90.6	6.2	99.0	**3.8**
	N	75.6		62.3		83.8		84.4		95.2	
BLIP^{N}_{129M}	M	76.4		60.1		84.6		88.1		98.3	
	F	75.2	**1.2**	59.7	2.6	73.7	10.9	89.6	**5.8**	99.0	4.1
	N	75.8		62.3		80.9		83.8		94.9	

Table 5.7: Validation results per group: male (M), female (F), and neutral (N). We report VQA-accuracy in VQAv2, accuracy in GQA and NLVR2, recall@1 in F30k and group disparity (Δ) across tasks. Lower Δ is better.

		VQAv2 mean± std	GQA mean ± std
$\text{LXMERT}_{\text{180K}}$	male	-.0311±.0057	-.0192±.0071
	female	-.0497±.0053	-.0477±.0088
	neutral	+.0020±.0030	-.0252±.0089
$\text{LXMERT}^{\text{N}}_{\text{180K}}$	male	-.0301±.0031	-.0227±.0036
	female	-.0538±.0034	-.0528±.0106
	neutral	+.0007±.0014	-.0245±.0042
$\text{LXMERT}_{\text{3M}}$	male	-.0169±.0054	-.0254±.0135
	female	-.0667±.0041	-.0935±.0110
	neutral	-.0157±.0056	-.0269±.0069
$\text{LXMERT}^{\text{N}}_{\text{3M}}$	male	-.0164±.0038	-.0188±.0060
	female	-.0634±.0046	-.0971±.0131
	neutral	-.0183±.0035	-.0194±.0109

Table 5.8: $\text{BiasAmp}_{A \to T}$ fine-tuning variance of LXMERT models across question answering tasks. Each model is fine-tuned 6 times on each task. We report average VQA-accuracy in VQAv2 and average accuracy in GQA, together with its standard deviation.

Part II

Applications of Language Models

Chapter 6

Cross-Cultural Transfer Learning for Chinese Offensive Language Detection

Abstract

Detecting offensive language is a challenging task. Generalizing across different cultures and languages becomes even more challenging: besides lexical, syntactic and semantic differences, pragmatic aspects such as cultural norms and sensitivities, which are particularly relevant in this context, vary greatly. In this paper, we target Chinese offensive language detection and aim to investigate the impact of transfer learning using offensive language detection data from different cultural backgrounds, specifically Korean and English. We find that culture-specific biases in what is considered offensive negatively impact the transferability of language models (LMs) and that LMs trained on diverse cultural data are sensitive to different features in Chinese offensive language detection. In a few-shot learning scenario, however, our study shows promising prospects for non-English offensive language detection with limited resources. Our findings highlight the importance of cross-cultural transfer learning in improving offensive language detection and promoting inclusive digital spaces.

Warning: *This paper contains content that may be offensive or upsetting.*

6.1 Introduction

The proliferation of offensive language and hate speech in online platforms, especially on social media, has significantly increased in recent years Zampieri et al. (2019, 2020); Gao et al. (2020). There is a fine line between offensive language and hate speech as few universal definitions exist Davidson et al. (2017). Therefore, hate speech can be classified as a subtype of offensive language. In this paper, we do not differentiate them in detail, and instead, refer to the task of offensive language detection (OLD).

Despite numerous breakthroughs in the development of NLP methods for OLD Liu et al. (2022); Rusert et al. (2022), some significant obstacles remain unsolved Vidgen et al. (2019), including the shortage of data resources for research purposes and bias in human annotation. Since most of the available approaches and resources for OLD are designed for English Arango Monnar et al. (2022), the resulting trained models operate within a monocultural background that caters to English speakers.[1] However, Schmidt and Wiegand (2017) believe that OLD has strong cultural implications, unlike other NLP tasks, because an utterance's offensiveness can vary based on an individual's cultural background.

People with different backgrounds react to inputs differently and communicate differently, so their tolerance for the presence of offensive terms, *e.g.*, slur, may differ, as well as what is altogether considered offensive Jay and Janschewitz (2008). Cultural differences have been explored in humor perception Jiang et al. (2019), swearing reception Pavesi and Zamora (2022), translation in semantic inconsistencies Sperber et al. (1994) and honorifics expression Song (2015); Liu and Kobayashi (2022). Even in less obvious cases, however, they bear meaningful significance on how to pose and solve NLP tasks, as cultures differ with respect to style, values, common ground and topics of interest Hershcovich et al. (2022).

Therefore, we argue that there is a need for addressing cross-cultural aspects in offensive language detection. Although culture is intricate and challenging to define clearly, language still remains as one of the most straightforward manifestation of culture. While recent work Ringel et al. (2019);

[1]Importantly, "culture" is multifaceted and complex. When referring to English speakers, we assume that there are general unique features that characterize them, but of course there is enormous diversity within speakers of the same language. As a first step towards the analysis of cross-cultural OLD, we restrict ourselves to the level of language categories.

Ranasinghe and Zampieri (2021) has demonstrated the effectiveness of cross-lingual transfer learning in the text classification and offensive Language (hate speech) detection, they don't consider the impact of cultural background differences (*e.g.*, Eastern and Western culture). In this paper, we take a step forward in this direction and explore the influence of offensive content from diverse cultural background on OLD, focusing on evaluation in Chinese.

Our contributions are as follows: 1) We explore the impact of transfer learning using offensive language data from different cultural backgrounds on Chinese offensive language detection (§6.3). 2) We find cultural differences in offensive language are expressed in the text topics, and that LMs are sensitive to these differences, learning culture-specific biases that negatively impact their transfer ability (§6.4). 3) We find that in the few-shot scenario, even with very limited Chinese examples, the model quickly adapts to the target culture.

6.2 Related work

Offensive language detection Although most of the research on OLD has focused on English Fortuna and Nunes (2018), there exist datasets in multiple languages: Chinese Deng et al. (2022), Korean Jeong et al. (2022), Danish Sigurbergsson and Derczynski (2020), Bengali Das et al. (2022), and Nepali Niraula et al. (2021), to name a few. However, language models commonly rely on prior distributions from training data, that reflects a discourse that is temporally and culturally situated Ghosh et al. (2021). In a comprehensive analysis of geographically-related content and its influence on performance disparities of offensive language detection models, Lwowski et al. (2022) find that current models do not generalize across locations. Sap et al. (2022) call for contextualizing offensive (toxicity) labels in social variables as determining what is toxic is subjective, and annotator beliefs can be reflected in the data collected.

Cross-lingual transfer learning Cross-lingual transfer appears as a potential solution to the issue of language-specific resource scarcity Lamprinidis et al. (2021). Nozza (2021) demonstrates the limits of cross-lingual zero-shot transfer for hate speech detection in English, Italian and Spanish. The benefits of few-shot learning is evident in works from Stappen et al. (2020) and

Röttger et al. (2022), who confirmed the effectiveness of few-shot learning for the task of hate speech detection in under-resourced languages. Ringel et al. (2019) harness cross-cultural differences for English formality and sarcasm detection based on German and Japanese, respectively. Litvak et al. (2022) show that, in the context of OLD, knowledge transfer is not bidirectional and efficient transfer learning holds from Arabic to Hebrew in terms of recall.

6.3 Method

6.3.1 Datasets

To explore the influence of different cultural backgrounds on Chinese OLD, the most straightforward approach is to adopt OLD datasets whose context and annotation process reflect diverse cultural backgrounds. We first select COLD Deng et al. (2022), a Chinese benchmark dataset covering the topics of racial, gender, and regional bias as our test dataset. We then select two other datasets that will be used in different training scenarios (see § 6.3.2): KOLD Jeong et al. (2022), a Korean dataset suited for OLD covering topics such as race, gender, political affiliation and religion; and HatEn, the English subset of HatEval Basile et al. (2019) composed of tweets which tends to capture a Western cultural background. Table 6.1 reports the statistics of the three datasets and the topic distributions of COLD. Notably, the three languages come from three different language families, making linguistic similarities between them less likely to be a factor in effective transfer learning between the datasets.

6.3.2 Learning settings

We explore different learning settings by utilizing **intra-cultural** and **cross-cultural** training sets during fine-tuning. For the intra-cultural setting, we only use COLD as the training set, which ensures cultural consistency in the training and testing process. In the cross-cultural setting, we further set up two ways: 1) *zero-shot*: only use KOLD or HatEn as the training set, which makes the fine-tuning process of LMs come from completely different cultural backgrounds; 2) *mix-training few-shot*: mix COLD with another language (KOLD or HatEn) as the final training set, which introduces cultural interference and makes the acquisition of the target culture more challenging.

Dataset	Language	Train	Dev	Test
COLD	Chinese	25726 (12723:13003=0.98)	6431 (3211:3220=1.00)	5323 (2107:3216=0.66)
KOLD	Korean	24257 (12190:12067=1.01)	8086 (4076:4010=1.02)	8086 (4044:4022=1.01)
HatEn	English	9000 (3782:5217=0.72)	1000 (427:573=0.75)	3000 (2343:657=3.57)
Region		8449	2104	2087
Gender		6579	1657	1551
Race		10698	2670	1685

Table 6.1: Datasets statistics (**top**) and topic distributions of COLD (**bottom**). Particularly, statistics of offensive and non-offensive data and the ratio between them are indicated in **parentheses**.

For convenience, we use $\mathcal{D}[X]$ to represent the detector with X as training set. Since the datasets are in different languages, we apply multilingual LMs in these experiments.

Translated data setting As an additional control experiment, to avoid the difference from the language itself, we also translate COLD and KOLD into English with *googletrans*[2] and conduct experiments with *English* PLMs under the same settings.

6.4 Experiments

Implementation In our experiments, we only evaluate on COLD and try different training settings with COLD, KOLD and HatEn. In particular, because the data volume of HatEn is relatively small, we use all of its data as the training set. The actual training set of three datasets has offensive data to non-offensive data ratios of 0.98, 1.01, and 1.02 (refer to Table 6.1). In the cross-cultural zero-shot setting, we also randomly sample 13,000 examples[3] from the Korean training set to ensure the consistency of the training data

[2] https://pypi.org/project/googletrans/

[3] The ratio of offensive data to non-offensive data is 0.96.

Model	Train Set	Test F1	Test ACC
mBERT$_{\texttt{base}}$	COLD	77.90±0.25	80.86±0.26
	CO+KO	78.23±0.05*	81.16±0.19
	CO+HE	78.19±0.18*	81.07±0.10
	KOLD	49.27±4.04**	67.85±0.70**
	KOLD†	50.34±3.49**	69.47±0.71**
	HatEn	35.96±3.95**	63.54±0.54**
XLM-R$_{\texttt{base}}$	COLD	78.77±0.27	81.51±0.20
	CO+KO	78.90±0.10	81.78±0.15*
	CO+HE	78.96±0.15	81.66±0.18
	KOLD	58.13±1.78**	72.14±0.67**
	KOLD†	60.86±1.44**	72.93±0.37**
	HatEn	29.84±2.07**	63.36±0.90**
XLM-R$_{\texttt{large}}$	COLD	79.09±0.24	81.87±0.16
	CO+KO	79.76±0.19**	82.45±0.19**
	CO+HE	79.43±0.22*	82.16±0.26**
	KOLD	63.48±1.63**	74.45±0.34**
	KOLD†	61.71±2.37**	74.09±0.80**
	HatEn	28.94±2.50**	63.76±0.40**

Table 6.2: Overall results on COLD test set. † marks KOLD training set is the same size as HatEn. CO, KO and HE are short for COLD, KOLD and HatEn respectively. By conducting Paired Student's t-test, $*$ = differs significantly from intra-cultural at $p < 0.05$, $**$ = significant difference at $p < 0.01$.

sizes with HatEn. For the multilingual LMs, we choose $\texttt{mBERT}_{\texttt{base}}$ Devlin et al. (2019), $\texttt{XLM-R}_{\texttt{base}}$ and $\texttt{XLM-R}_{\texttt{large}}$ Conneau et al. (2020). In the translated data setting, we apply the English models $\texttt{BERT}_{\texttt{base}}$ Devlin et al. (2019), $\texttt{RoBERTa}_{\texttt{base}}$ and $\texttt{RoBERTa}_{\texttt{large}}$ Liu et al. (2019).

Our models are optimized with a learning rate of $5e-5$. We fine-tune each model for 100 epochs using early-stopping with a patience of 5, and run 5 times with different random seeds for each setting.

Overall results The experimental results on COLD test set are shown in Table 6.2.[4] Compared to the intra-cultural setting, we find that: 1) In the cross-cultural few-shot scenario, the performance differences between

[4]We only report the test set score, because only the test set of COLD is annotated manually, and the training and dev sets are labeled semi-automatically.

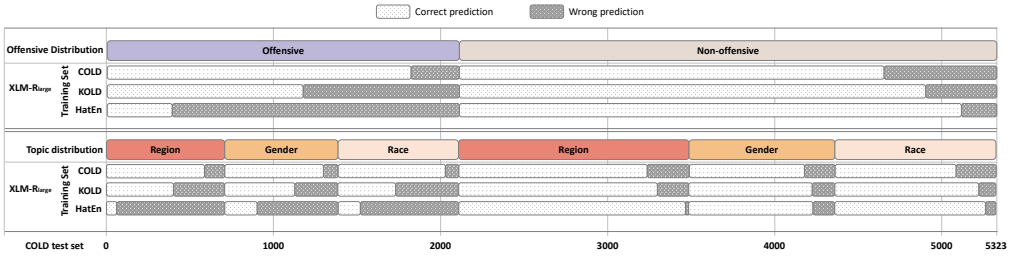


Figure 6.1: A fine-grained view of the distribution of offensive detection results based on $\texttt{XLM-R}_{\texttt{large}}$. For reference, the colored part represent the distribution of related data in COLD test set. The model learns culture-specific biases—*e.g.*, when training on English, it tends not to classify region-related text as offensive.

$\mathcal{D}\,[\text{COLD}]$ and $\mathcal{D}\,[\text{CO}+\text{KO}]$, $\mathcal{D}\,[\text{COLD}]$ and $\mathcal{D}\,[\text{CO}+\text{HE}]$ are both very small (less than one point at the maximum), which implies that with sufficient knowledge of the Chinese target culture, the intervention of other cultures does not diminish the ability to detect Chinese offensive language, but has a slight contribution. 2) In the cross-cultural zero-shot scenario, the detection ability of $\mathcal{D}\,[\text{KOLD}]$ and $\mathcal{D}\,[\text{HatEn}]$ get worse. In particular, the former is slightly better than the latter. This implies that it is easier to detect Chinese offensive language in Korean cultural background compared to a Western cultural background.

To better understand the detection ability of Chinese offensive language with different cultural backgrounds, we look closer at offensive detection results for the intra-cultural and cross-cultural zero-shot settings. Figure 6.1 shows the distribution of the data and the predictions from our best performing model $\texttt{XLM-R}_{\texttt{large}}$. First, $\mathcal{D}\,[\text{COLD}]$, which is in the same cultural background as the test set, has the best ability to detect offense. $\mathcal{D}\,[\text{HatEn}]$ is the worst detector, with less than 50% accuracy for offensive data. Because of this, it can be highly accurate in non-offensive data. This is why $\mathcal{D}\,[\text{HatEn}]$ gets a spurious high accuracy on the test set but a very low F1 score (Table 6.2). However, it is noteworthy that the HatEn-trained model requires more severe language to be labeled as offensive,[5] so some instances that should be classified as offensive, may not be considered hate speech and

[5]This could be a reason to treat Hate Speech Detection as a separate task, contrary to our simplified view here.

Model	Train Set	Test F1	Test ACC
$\text{BERT}_{\texttt{base}}$	COLD	77.59±0.41	80.67±0.37
	CO+KO	77.86±0.19*	80.90±0.20
	CO+HE	77.50±0.17*	80.47±0.18
	KOLD	61.84±1.46**	71.26±0.34**
	KOLD†	61.64±1.06**	71.21±0.27**
	HatEn	21.20±1.36**	61.53±0.21**
$\text{RoBERTa}_{\texttt{base}}$	COLD	77.89±0.46	81.01±0.40
	CO+KO	78.25±0.40	81.35±0.37*
	CO+HE	78.08±0.34	81.12±0.25
	KOLD	63.85±1.12**	73.60±0.43**
	KOLD†	63.47±0.84**	73.21±0.25**
	HatEn	26.09±2.82**	62.81±0.36**
$\text{RoBERTa}_{\texttt{large}}$	COLD	78.22±0.40	81.24±0.33
	CO+KO	78.74±0.21**	81.70±0.15**
	CO+HE	78.24±0.30*	81.17±0.25**
	KOLD	65.56±1.16**	73.70±0.49**
	KOLD†	64.39±1.60**	73.71±0.37**
	HatEn	26.69±1.38**	63.20±0.44**

Table 6.3: The experimental results on the COLD test set, with all training and testing data translated to English. † marks KOLD training set is the same size as HatEn. By conducting Paired Student's t-test, $*$ = differs significantly from intra-cultural at $p < 0.05$, $**$ = significant difference at $p < 0.01$.

will not be classified as such. Moreover, for specific-topic offensive language detection, the performance of each detector is also different, with $\mathcal{D}$ [HatEn] performing the worst in the regional topic.

Translated results For the experiments of the translated version of the Chinese and Korean datasets into English. The experimental results are shown in Table 6.3, showing similar trends to the results in Table 6.2. This demonstrates that the results hold for cross-cultural transfer and are not simply due to linguistic similarities.

Few-shot learning While the diverse cultural backgrounds of Korean and English may not enable precise detection of Chinese offensive language in a zero-shot scenario, it is not detrimental when integrated into the target cul-

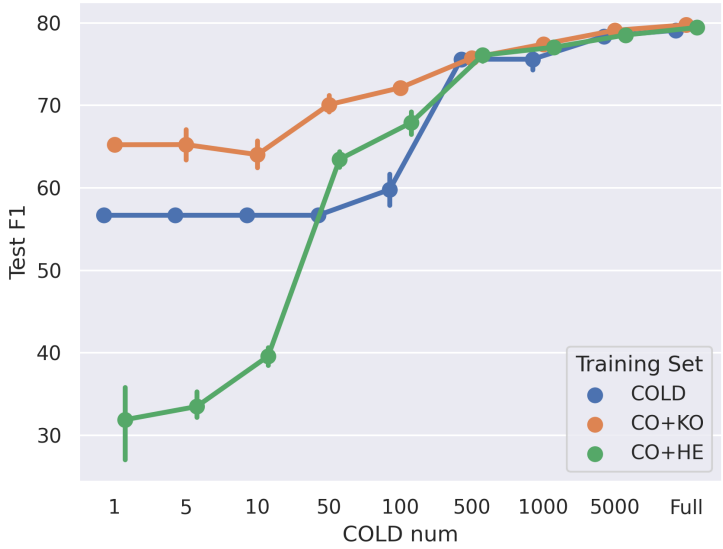


Figure 6.2: The experimental results (F1) in few-shot setting based on `XLM-R`$_{\texttt{large}}$, evaluated on the COLD (Chinese) test set. Performance improves rapidly with training examples from the target culture. Pre-training on KOLD (Korean) provides a better starting point, while pre-training on HatEn (English) is detrimental.

ture in a few-shot scenario. Therefore, when mixing heterogeneous cultural background knowledge, is it necessary to provide sufficient target cultural background knowledge? To investigate this problem, we conduct an analytical experiment under a few-shot setting by incorporating different scales of COLD data into the training set. Figure 6.2 displays experimental results indicating that the correlation between the ability to detect offensive language and target cultural knowledge follows a pattern similar to that of an increasing logarithmic function. This implies that offensive language detection performance improves rapidly with limited target cultural knowledge acquisition, but gradually slows down as the amount of target knowledge increases. Specifically, when the training focuses on COLD within the range of 1 to 50, $\mathcal{D}$[COLD] possesses limited knowledge of the training concentration, and its detection capability stems primarily from the pre-training model itself. At this stage, HatEn has a clearly negative effect, while KOLD

Chinese	Korean	Labels
黑人反对歧视黑人有啥错? *What is wrong with blacks against discrimination against blacks?*	흑인 대통령도 나온 미국, 이제 인종차별은 사라졌다? *America with a black president, now racism has disappeared?*	0 / 1
中国哪有那么容易搞到毒品? *How can it be so easy to get drugs in China.*	중국에서 범은 잡히면 뭐가 잘릴까.. *What will happen if a criminal is caught in China?*	0 / 1

Table 6.4: Cases with reversed labels through semantic vector retrieval were listed, suggesting the existence of cultural differences across languages. Non-offensive and offensive cases are labeled as 0 and 1.

has a positive effect. Within the range of 50 to 500, both HatEn and KOLD have an obvious positive effect, while for COLD data scales greater than 500, the effect is still present but less pronounced. These findings offer promising opportunities for low-resource offensive language detection systems.

Case study. To provide an intuitive explanation of cultural differences, we use semantic similarity retrieval Reimers and Gurevych (2019) to find the most similar cases from KOLD to COLD with the similarity threshold set to 0.7. As depicted in Table 6.4, sentences with similar topics and semantics (*e.g.*racial discrimination, politics) hold different labels among languages, suggesting the presence of cultural distinctions in offensive language detection and highlighting the significant obstacles for few-shot learning. Thus, we emphasize the necessity of greater cultural adaptation models that can integrate diverse cultural knowledge.

6.5 Conclusion

Our study highlights the challenges of detecting offensive language across different cultures and languages. We show that transfer learning using data from diverse cultural backgrounds have different negative effects on the transferability of language models due to culture-specific biases. However, our findings also indicate promising prospects for improving offensive language detection in promoting inclusive digital spaces, particularly in a few-shot

learning scenario. We call for more research on cross-cultural offensive language detection, which is important to deploy effective moderation strategies for social media platforms, improving cross-cultural communication, and reducing harmful online behavior.

Limitations

Our study explores the impact of transfer learning on offensive language detection using data from different cultural backgrounds. However, treating HatEn as representative of "Western cultural background" is too vague, as it ignores the cultural differences between American and British cultures. Moreover, "culture" is multifaceted and complex, and there is enormous diversity among speakers of the same language. To focus on language categories, we limit our analysis to a first step towards cross-cultural offensive language detection.

Ethics Statement

The datasets used in this study are publicly available, and we strictly follow the ethical implications of previous research related to the data sources. It is important to note that the content of these datasets does not represent our opinions or views.

Acknowledgments

Thanks to the anonymous reviewers for their helpful feedback. The authors gratefully acknowledge financial support from China Scholarship Council. (CSC No. 202206070002 and No. 202206160052).

Chapter 7

MEG: Medical Knowledge-Augmented Large Language Models for Question Answering

Abstract

Question answering is a natural language understanding task that involves reasoning over both explicit context and unstated, relevant domain knowledge. Large language models (LLMs), which underpin most contemporary question answering systems, struggle to induce how concepts relate in specialized domains such as medicine. Existing medical LLMs are also costly to train. In this work, we present MEG, a parameter-efficient approach for medical knowledge-augmented LLMs. MEG uses a lightweight mapping network to integrate graph embeddings into the LLM, enabling it to leverage external knowledge in a cost-effective way. We evaluate our method on four popular medical multiple-choice datasets and show that LLMs greatly benefit from the factual grounding provided by knowledge graph embeddings. MEG attains an average of +10.2% accuracy over the Mistral-Instruct baseline, and +6.7% over specialized models like BioMistral. We also show results based on Llama-3. Finally, we show that MEG's performance remains robust to the choice of graph encoder.

7.1 Introduction

Large language models (LLMs) induce knowledge from vast text corpora. Through self-supervised learning, these models capture deeply contextualized representations of input tokens that enable them to generalize to new tasks with remarkable performance. This, as well as their ability to write long coherent passages, has made LLMs incredibly popular, despite their considerable inference costs (Cheng et al., 2023) and their concerning carbon footprint (Strubell et al., 2019). Moreover, current LLMs face significant challenges with handling complex reasoning and ensuring trustworthiness Liu et al. (2023); Huang et al. (2024) and factual consistency Maynez et al. (2020); Zhou et al. (2023a); Tam et al. (2023); Hager et al. (2024), essential to critical fields like healthcare. While LLMs are poised to revolutionize our medical system, already performing well on medical licensing exams Jin et al. (2021a); Pal et al. (2022); Singhal et al. (2023a); Brin et al. (2023) and other tasks Nazario-Johnson et al. (2023); Van Veen et al. (2023); Tu et al. (2023); Carl et al. (2024), there is still much room for improvement.

To improve reliability and reduce computational costs, researchers have experimented with training from mixtures of corpora and knowledge bases (Pan et al., 2023, 2024). Knowledge Graphs (KGs), such as the Unified Medical Language System (UMLS) Bodenreider (2004), are structured knowledge bases that explicitly store rich factual knowledge. KGs are good at capturing the nuances of complex data and can provide complementary information to LLMs, especially useful for tasks requiring structured understanding. The potential of knowledge-augmented LLMs[1] outlines an interesting research paradigm that can alleviate current challenges of LLMs, and reduce the need of training ever-larger models Hooker (2024). However, how to effectively model interactions between LLMs and KGs remains an open question.

Recent efforts have focused on self-supervised methods for jointly training graph neural networks and pretrained language models Yang et al. (2021); Chien et al. (2022); Brannon et al. (2024). Others Yasunaga et al. (2022); Tang et al. (2024); Plenz and Frank (2024), propose new model architectures to leverage the two modalities, graph and text, during pretraining. These

[1]In this work, we define a knowledge-augmented LLM as an LLM enhanced with KG embeddings (KGEs). KGEs are dense vector representations of graph entities Ju et al. (2024). Therefore, we also refer to knowledge-augmented LLMs as KGE-augmented LLMs throughout the paper.

methods learn deep interactions over text and graph, but they require carefully curated pretraining data, are mainly studied for graph-oriented tasks Yang et al. (2021); Chien et al. (2022); Tang et al. (2024), or are yet to be adapted to a generative framework Yasunaga et al. (2022); Plenz and Frank (2024).

In this work, we introduce MEG, a parameter-efficient approach to MEdical knowledGe-augmented LLMs for question answering (QA). We design a lightweight mapping network to unidirectionally translate KG embeddings into the LLM's vector space. This enables the LLM to interpret the new input embeddings, which, in turn, further conditions its response generation. We use Mistral-Instruct (7B) Jiang et al. (2023) as our base LLM and report results with our best setup: a KG encoder based on GraphSAGE Hamilton et al. (2017) combined with a simple MLP as mapping network. We also provide results with the recently released Llama-3-Instruct (8B) Dubey et al. (2024) as base LLM.

In sum, our **contributions** are as follows: *i)* We introduce MEG, a novel approach to knowledge-augmented LLMs based on KGEs. *ii)* We conduct extensive evaluation on the four popular multiple-choice QA datasets from the MultiMedQA Singhal et al. (2023a) clinical benchmark, and demonstrate the effectiveness of integrating pretrained KGEs into LLMs for medical question answering. Specifically, MEG surpasses strong LLM baselines like BioMistral-7B Labrak et al. (2024) or MediTron-7B Chen et al. (2023a), which have followed a costly continued pretraining of the base LLMs on curated biomedical data. *iii)* We provide insights into the inner workings of MEG, examining the contributions of each module and comparing embedding spaces. We intuitively explain the shifts in the LLM's representations that drive MEG's stronger performance. *iv)* We publicly release the code, trained KGEs and model checkpoints at github.com/lautel/MEG.

7.2 Related Work

Medical Language Models Current state-of-the-art (SOTA) in medical QA benchmarks like MedQA Jin et al. (2021a), PubMedQA Jin et al. (2019) or MedMCQA Pal et al. (2022) belongs to close-sourced models of unknown size like Med-Gemini Saab et al. (2024), Med-PaLM2 Singhal et al. (2023b) or GPT-4 Nori et al. (2023). Popular open-source LLMs in biomedicine include MedAlpaca Han et al. (2023) and PMC-LLaMA Wu et al. (2023)

based on Llama Touvron et al. (2023a), MediTron Chen et al. (2023a) based on Llama-2 Touvron et al. (2023b), or BioMistral Labrak et al. (2024) based on Mistral-Instruct Jiang et al. (2023). These models continue pretraining the base general-purpose models on curated medical corpora. More recently, Kim et al. (2024) present the Meerkat models trained with chain-of-thought Wei et al. (2024) synthetic data. Meerkat-7B outperforms the previous best 7B models across several medical benchmarks. However, it takes eight 80G A100 GPUs and 1.5 days to complete training. In contrast, our approach is the first to leverage pretrained medical KGEs and can be trained on four A10G GPUs within a few hours (see § 7.5 for details).

Knowledge-Augmented Language Models Bringing together LLMs and KGs is an active line of research that has gained increasing attention from both academia and industry (Pan et al., 2023, 2024). Among numerous efforts in this area, Zhang et al. (2019); Yasunaga et al. (2022); Tang et al. (2024); Zhu et al. (2023), to name a few, propose different methods for combining text and graphs during pretraining. Parallel to these lines of work, Sarmah et al. (2024); Edge et al. (2024); Hu et al. (2024); Mavromatis and Karypis (2024) approach the integration of LLMs and KGs through retrieval-augmented generation (RAG) Lewis et al. (2020). However, the deployment of such knowledge-augmented LLMs for medical QA remains understudied. Our work fills this gap and presents a novel approach to medical knowledge-augmented LLMs based on KGEs. We note that MEG may resemble a sort of RAG system, where an LLM leverages knowledge from an external database of KGEs. In this case, the grounding module would act as retrieval module, fetching appropriate KGEs that ground text information in KG entities as part of a prompt.

7.3 Problem Formulation

We augment an LLM with KG embeddings to answer medical questions drawing on factual knowledge from the KG. We rely on a large KG in the target domain, namely UMLS Bodenreider (2004). Our proposed approach, MEG, consists of four key components:

i) A **KG encoder** to represent knowledge graph entities in a continuous vector space, while preserving their semantic meaning.

ii) An **instruction-tuned language decoder** capable of generating textual answers.

iii) A **mapping function** $f_{\mathbf{k}}$ that transforms the output of the KG encoder into a representation that can be used by the language decoder. f_{k} is parameterized by a neural network. Thus, we interchangeably use the term mapping network.

iv) A **KG grounding module** that detects textual entities and grounds them in graph entities.

Figure 7.1 depicts the full pipeline of MEG. We carefully investigate the design of these components and how they interact with each other (§ 7.4.1). To attain the best accuracy on downstream tasks, we conduct a two-phase training (§ 7.4.2).

Definitions. A generic dataset for multiple-choice question answering (QA) consists of examples with a context paragraph, a question and a candidate answer set, all expressed in text. Given a QA example, each prompt W is the concatenation of context, question and candidate answer set. We denote the sequence of tokens (words) in W as $\{w_1, \ldots, w_S\}$, where S is the maximum sequence length. We denote the sequence of tokens (vectors) in the language model embedding space as $W_e = \{w_{e1}, \ldots, w_{eS}\}$.

We define a knowledge graph (KG) as a directed graph $G = (V, E)$, where V is the set of entity nodes, and $E \subseteq V \times R \times V$ is the set of edges (triples) that connect nodes in V, with R being the set of relation types. Each triple (s, p, o) in a KG represents a knowledge fact, such as (Headache, is_a, Cephalgia). A KGE e is a mathematical representation that maps each entity $v \in V$ and each relation $r \in R$ of a directed knowledge graph G to low-dimensional vectors in $\mathbb{R}^g$, preserving the semantic relationships within the graph.

Finally, we define a KGE-augmented language model to be a function $f_{\mathrm{l}}(W_e \oplus f_{\mathrm{k}}(X))$[2] with $f_{\mathrm{l}} \in \mathbb{R}^l$, where $f_{\mathrm{k}}(X)$ is a set of KGEs, $\{e_1, \ldots, e_N\}$ with $e_i \in \mathbb{R}^g$, that has been mapped to the LLM's space using a learned mapping function $f_k : \mathbb{R}^g \to \mathbb{R}^l$. The language model f_{l} concatenates these representations to the token word embeddings W_e to perform downstream tasks in the fine-tuning steps. A language model is a special case of a KGE-augmented language model with no KGE (N=0).

[2]Formally, its domain is the set of sequences of elements $x_i \in \mathbb{R}^l$.

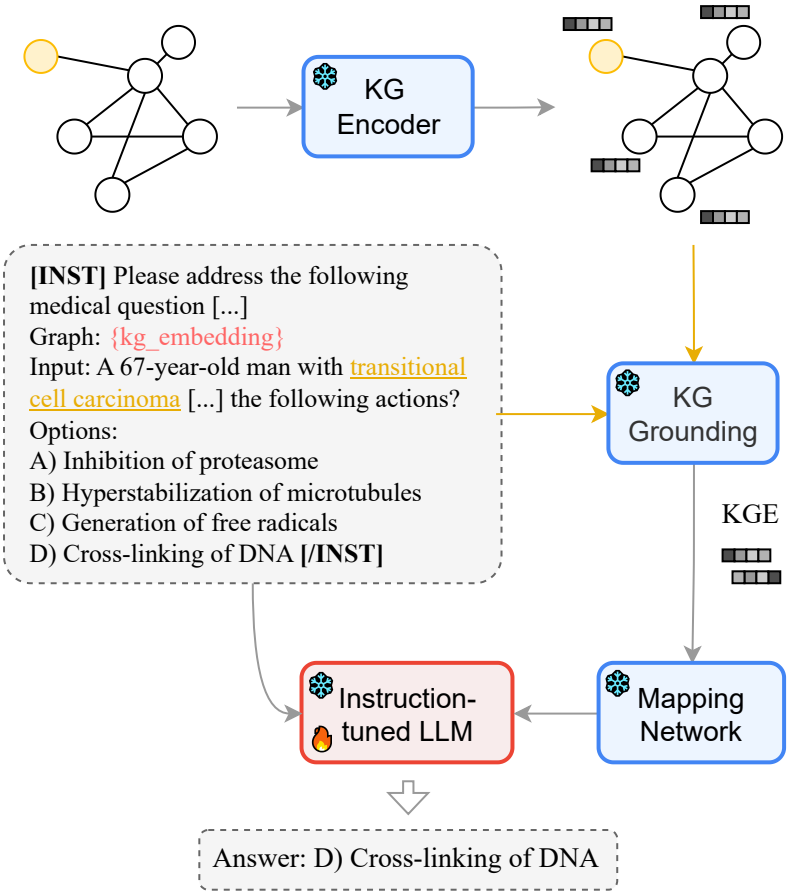


Figure 7.1: MEG leverages a pretrained KG encoder and an LLM. During an initial phase of training, MEG learns a mapping network to convert relevant graph features (KGEs) retrieved by the grounding module into token embeddings. During downstream fine-tuning, only the LLM's weights are updated, keeping the LLM's embedding layer and mapping network frozen. At inference, the LLM takes the text and the mapped KGEs as input and generates a response.

7.4 Method

7.4.1 MEG

MEG combines a pretrained KG encoder and a pretrained LLM by means of an intermediate mapping network (see Figure 7.1). The KG encoder, which

is trained separately on a large medical KG[3], provides graph embeddings that are directly fed to the mapping network. Then, the LLM uses the text content and mapped KG embeddings as input to generate an answer.

Knowledge Graph Encoder The KG encoder is trained up-front over the selected graph to generate KGEs. We choose GraphSAGE Hamilton et al. (2017) as our preferred KG encoder. In § 7.6.1 we present an ablation study with random-walk-based, energy-based translational, and message-passing encoders.

Mapping Network The mapping function f_{k} transforms a sequence of graph features from the KG encoder into a sequence that can be consumed by the LLM. We parameterize f_{k} as an MLP with four hidden layers of size $d_h = 128$. In particular, a set of graph embeddings is transformed from $d_g = 256$ to $d_l = 4096$ after a series of non-linear transformations through the hidden layers of f_{k}. We denote the initial embedding sets as $X = \{x_i\}_{i=1}^{N}$, $Y = \{y_j\}_{j=1}^{N}$, $x_i \in \mathbb{R}^{d_g}, y_j \in \mathbb{R}^{d_l}$, being x_i the KGEs, y_j the averaged token embeddings of the entity in the LLM, and N the total number of graph embeddings (entities). We further denote the set of mapped embeddings as $f_{\text{k}}(X) := \{f_{\text{k}}(x_i)\}_{i=1}^{n}$.

The goal is to learn the mapping f_{k} that transforms X to the LLM's vector space, while preserving its semantic meaning and structural information. Rather than minimizing the sum of squared differences between $f_{\text{k}}(X)$ and Y, we aim at positioning each x_i in the neighborhood of its counterpart in Y. Pursuing an exact matching of space distributions, such as through a Procrustes transformation Schönemann (1966); Gower (1975), would disregard the structural knowledge encoded in X.

To achieve this, we design an architecture similar to Xu et al. (2018) with two mappings $f_{\text{k}} : X \to Y$ and $g_{\text{k}} : Y \to X$, as illustrated in Figure 7.2. We construct an instruction dataset with labels from UMLS's entities to teach the LLM to interpret the transformed graph embedding $f_{\text{k}}(x_i)$. Figure 7.2 shows an example of an instruction, where the placeholder `{kg_embedding}` is replaced by $f_{\text{k}}(x_i)$. We train the full network jointly with the LLM[4].

[3]Specifically, we use **UMLS** Bodenreider (2004), a widely-used KG in biomedicine with ∼300K nodes (entities) and one million edges in total.

[4]We conducted experiments by training the mapping network and LLM separately on UMLS. However, this approach resulted in worse performance on the downstream tasks

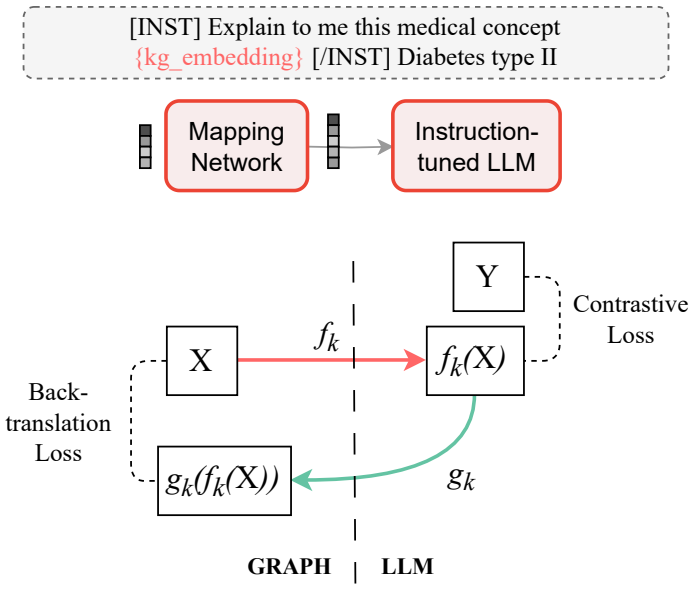


Figure 7.2: f_k and g_k are embedding transfer functions. f_k takes a set of KGEs X (*i.e.*, node entities) as input, and outputs a mapping of X to the LLM's vector space. Y is the set of averaged token embeddings of entities in the LLM space. During training, g_k prevents degenerated transformation of graph embeddings. The dashed lines indicate the input for the objective losses.

Our loss function consists of three parts: a standard next-token prediction objective (**cross-entropy loss** $\mathcal{L}_{ce}$), and a sum of a contrastive loss and a back-translation loss to optimize the mapping network. Specifically,

- Given a batch X_b including a positive pair of examples x_i and x_j, a contrastive objective Hadsell et al. (2006) is a function whose value is low when x_i is similar to x_j and dissimilar to all others, which are considered negative pairs for x_i. We employ a popular contrastive self-supervised learning objective Sohn (2016); van den Oord et al. (2019); He et al. (2019), dubbed as **NT-Xent loss** by Chen et al. (2020a). NT-Xent uses dot product as similarity measure, and computes a normalized

tested.

temperature-scaled cross-entropy loss for a positive pair as follows,

$$\ell_{i,j} = -\log \frac{\exp(x_i \cdot x_j/\tau)}{\sum_{k=1,k\neq i}^{B} \exp(x_i \cdot x_k/\tau)}, \tag{7.1}$$

where B is the batch size and τ is the temperature. We set the hyper-parameter $\tau = 1.0$ [5]. The final loss $\mathcal{L}_{\text{c}}$ is computed across all positive pairs in a batch, summed across all batches. Intuitively, the contrastive loss serves as an unsupervised objective function for training the network to bring similar entities closer together in Y and push dissimilar ones apart.

- We also employ a **back-translation loss** for preventing degenerated transformation. We enforce that the graph embedding after the forward and the backward transformation should not diverge much from its original direction. Following Xu et al. (2018), we choose the back-translation loss based on cosine similarity. Note that our primary goal is to optimize the forward mapping $f_{\text{k}} : X \to Y$. Thus, we do not control for back-translation in the reversed path, $g_{\text{k}} : Y \to X$,

$$\mathcal{L}_{\text{bt}}(f_{\text{k}}, g_{\text{k}}) = \sum_{i} (1 - \cos(x_i, g_{\text{k}}(f_{\text{k}}(x_i)))) \tag{7.2}$$

Thereby, when training the mapping network jointly with the LLM, we minimize the following objective function:

$$\mathcal{L} = \alpha\mathcal{L}_{\text{c}} + \beta\mathcal{L}_{\text{bt}} + \mathcal{L}_{\text{ce}}, \tag{7.3}$$

where α and β are scalar hyperparameters to weight each objective in the transformation process. Our network design achieves a good trade-off between expressivity and parameter count, totaling 1.22M parameters. After the mapping is learnt, we freeze the network's weights and the LLM's embedding layer during the fine-tuning to downstream tasks (Phase II of training). The backwards transfer network g_{k} is disconnected and only f_{k} is used to do the mapping.

[5] We also evaluate $\tau = 0.5$ as in Chen et al. (2020a) and $\tau = 0.07$ as in He et al. (2019). We choose the final value of $\tau = 1.0$ based on accuracy attained on a zero-shot setting on the validation split in MedQA.

Grounding Module The grounding module takes textual data W as input and links entity mentions in W to their corresponding nodes in the knowledge graph G, generating a set of KG embeddings for the LLM to use. A critical step in this process is medical entity disambiguation Vretinaris et al. (2021); Lu et al. (2024), which involves detecting named entities in W and linking them to their unique counterparts in G. Since an entity can be referred to in multiple ways, for instance "heart attack" and "myocardial infarction", this step standardizes variations by linking them to a unique identifier in G (following with the previous example, both mentions correspond to Concept Unique Identifier (CUI) "C0155626" in UMLS). This grounding ensures retrieval of relevant information for each example. We use the entity linker presented in Neumann et al. (2019)[6], which covers 99% of the concepts mentioned in the MedMentions dataset Mohan and Li (2019) and 86% of the concepts mentioned in the MedQA dataset Jin et al. (2021a). These two datasets provide ground truth UMLS annotations.

7.4.2 Training

We aim to train MEG to achieve competent results on medical question answering benchmarks while minimizing computational cost. To do this, we conceive a two-phase training strategy with a minimal part of the model's parameters updated.

Phase I: Embedding Transfer Learning We first learn the optimal transformation $f_{\text{k}} : X \rightarrow Y$ so that the mapped KG embeddings retain relevant information from the KG and can be effectively used by the LLM. As explained in § 7.4.1, we create an instruction dataset to guide the LLM in learning the relationship between its original representation of medical entities and their mapped graph embeddings. The train set contains 297,927 examples, following the same template shown in Figure 7.2 for every entity label in UMLS.[7] We train for one epoch jointly the mapping network and

[6]We use the last version of scispaCy (v2.5.0), which supports linking to UMLS and has near 3M unique concepts.

[7]We investigate whether data augmentation at this stage could lead to more accurate results in downstream tasks. We augment the initial ∼300K examples by creating new instructions with multiple entities, *e.g.*"Explain to me these medical concepts: […]", to better match the setting from downstream tasks, which often include several entities per sample. This process doubles the dataset size, with an augmentation that normally dis-

the LLM to minimize the objective from Eq. 7.3.

Phase II: Downstream task Given a medical multiple choice QA dataset, we fine-tune MEG to answer the input question based on the textual content and information leveraged from the mapped KG embeddings. We format the input prompts W as follows. For each example in the dataset, we concatenate the context (if any), question and candidate answer set following the pseudo-code shown in Figure 7.3. The placeholder `{kg_embedding}` is replaced with N KGEs transformed by the mapping network[8]. We assume f_{k} is learned in phase I, so we keep the mapping network weights frozen and disconnect the backward network g_{k}, which merely served to regulate the learning of f_{k} and prevent degenerated transformation. Similarly, the LLM's embedding layer is also frozen.

7.5 Experimental Details

Data Following previous research on medical LLMs, we evaluate MEG on four well-known medical benchmarks that require extensive background knowledge. The first one is MedQA-USMLE (**MedQA**) Jin et al. (2021a), which consists of 10,178 train questions and 1,273 test questions, formatted with four choices each. The content was originally curated by experts from the US Medical License Exam. The second benchmark, **PubMedQA** Jin et al. (2019), was collected from PubMed abstracts and includes 1,000 expert labeled question-answer pairs. The task is to produce a yes/no/maybe answer based on the question and an abstract as context. As previously done by others Singhal et al. (2023a); Chen et al. (2023a); Labrak et al. (2024), we

tributes the number of entities per instruction between 2 and 10. Results are within ±0.2 accuracy in MedQA compared to training without the augmented data. Due to the extra computational costs and minor (if any) gains, we did not explore this option further.

[8]In both training phases, we investigate the effect of injecting the mapped KGEs at the last layer of the LLM instead of after the embedding layer. These early experiments revealed little to no impact on zero-shot downstream accuracy, but slightly worsen the fine-tuned accuracy as measured on the validation set of MedQA with three random initialization seeds. This finding suggests that the LLM benefits from attending the external KGEs during fine-tuning, enabling more contextualized representations of these embeddings.

[INST] Please address the following medical question based on the Input text and any useful information you may find in the given concepts from a medical graph.
Input: context question
Options:
{% for option in options %}
{{letter}}) {{text}}
{% endfor %}
Answer with the best option directly. Ignore irrelevant information.
Graph: {{kg_embeddings}} [/INST]
Answer: {{correct_option}}

Figure 7.3: Template used to generate instructions for all QA datasets. The context is optional, depending on the dataset. At inference time, the text after [/INST] is generated by the language model.

use 500 random[9] samples for evaluation. The remaining 500 samples, though limited in size, serve as our only source of training data. We exclude the 211k artificially labeled yes/no samples provided by Jin et al. (2019) to avoid bias towards these two options. The third benchmark, **MedMCQA** Pal et al. (2022), contains 179,722[10] train questions from Indian medical entrance exams. Due to the unavailability of answer keys for the test set, we follow others Wu et al. (2023); Tu et al. (2023); Labrak et al. (2024) and report results on the validation set (4,183 questions). Lastly, **MMLU-Medical** Singhal et al. (2023a) includes 1,089 questions, each with four options, across six medical and biology-related categories drawn from Hendrycks et al. (2021). Since this dataset only provides test data, we evaluate the generalization performance of MEG fine-tuned on MedMCQA as in Chen et al. (2023a). Thus, results on MMLU-Medical report out-of-distribution inference.

Training Details We initialize the KG node embeddings with token embeddings from SapBERT Liu et al. (2020a). SapBERT leverages contextualized embeddings from a pretrained BERT-based language model for biomed-

[9]We split the data following a similar distribution of answers between train and test splits.
[10]We detect 3,100 duplicate questions in the train split, which we remove.

ical KGs like UMLS. This initialization leads to improved performance compared to random embedding initialization. We train GraphSAGE with same hyperparameters as in Hamilton et al. (2017).

During **phase I** of training, described in § 7.4.2, we randomly initialize the mapping network and load the pretrained weights of the LLM. We fully train the mapping network and perform low-rank adaptation (LoRA, Hu et al. (2022)) fine-tuning on every linear layer of the LLM, while the remaining parameters are frozen. This parameter-efficient tuning approach allows to learn the equivalent of 2% of the model's parameters. Our full architecture results on 216M trainable parameters for MEG-Mistral. After fine-tuning, we merge the LLM's updated parameters with the base model. Since the input prompt has a fixed size (see Figure 7.2), we use a reduced sequence length (124) to optimize computational efficiency. We train for one epoch with gradient accumulation over 8 steps to achieve an effective batch size of 128. We employ a cosine learning rate scheduler with learning rate of $1e-5$, warmup ratio of 3% and no weight decay. We use mixed precision (bfloat16) and FlashAttention2 Dao (2023) to optimize memory usage and speed up computations on the LLM. Training takes 4h on 4 NVIDIA A10G GPUS using DeepSpeed[11] for distributed training.

After phase I, the models are evaluated in a zero-shot setup on downstream tasks, already able to use information from graph embeddings. To fine-tune on a specific task, we perform **phase II** of training. To allow batching, the number of KGEs injected to the LLM is fixed across samples[12]. If the grounding module retrieves more KGEs, we randomly select N. Otherwise, we add zero-padding. We assume the mapping network is learned in phase I, so we freeze its weights and disconnect the backward network g_{k}. We also freeze the LLM's embedding layer. This approach reduces the computational complexity and speeds up fine-tuning with a total of 83M trainable parameters in both MEG-Mistral and MEG-Llama. Fine-tuning is done for 3 epochs, with a sequence length of 400 (500 for PubMedQA), learning rate of $1e-4$ and effective batch size of 32. The remaining hyperparameters have the same value as in phase I. We did not optimize hyperparameters for Llama-3-Instruct.

[11] https://www.deepspeed.ai/

[12] The average number of ground entities per instance varies across datasets according to the median number of ground entities. We set $N = 20$ in MedQA, PubMedQA and "professional medicine" in MMLU-Medical; $N = 3$ in MedMCQA and $N = 2$ in the remaining categories from MMLU-Medical.

7.6 Results

We evaluate accuracy on four medical multiple-choice question datasets in three variants of MEG: MEG-Mistral1 and MEG-Mistral3, based on Mistral-7B-Instruct-v0.1 and -v0.3, respectively; and MEG-Llama, based on Llama-3-Instruct. We report average accuracy and standard deviation across three random seeds. Our results in Tables 7.1 and 7.2 reveal consistent average improvement across datasets compared to baselines.

	Model	Acc
ZS	Mistral-Instruct-v0.1†	$42.3_{\pm 0.3}$
	BioMistral†	$44.4_{\pm 0.2}$
	Mistral-Instruct-v0.1 w/ graph	$40.4_{\pm 0.4}$
FT	Mistral-Instruct-v0.1†	$42.0_{\pm 0.2}$
	BioMistral†	$50.6_{\pm 0.3}$
	Mistral-Instruct-v0.1 w/ graph	$52.7_{\pm 0.2}$

Table 7.1: Ablation study on the utility of the explicit information encoded in knowledge graph triples. We report accuracy on MedQA. ZS stands for "zero-shot"; FT stands for "fine-tuning". †Results from Labrak et al. (2024).

In-prompt graph triples provide useful information We investigate whether the inclusion of KG information can positively influence the LLM's answers. To establish a primary baseline, we take Mistral-Instruct-v0.1 and MedQA as a running example. For each question, we select a maximum of 10 named entities s and randomly retrieve 2 graph neighbors o for each, resulting in a maximum of 20 graph triples (s, p, o). We include them as part of the prompt, in natural language[13]. Table 7.1 shows a degradation in zero-shot accuracy when including triples to the prompt. This can be due to the random selection of the final triples (to fit in the context length), since the semantic information varies among them significantly. This limitation speaks in favor of using KGEs to condense representation of entities to a single embedding. Also, as Hager et al. (2024) point out, LLMs' face difficulties

[13]We append the triples at the end of the instruction in JSONL-style, *i.e.*, [{s1,p1,o1}, {s2,p2,o2}, ...].

	MedQA	PubMedQA	MedMCQA	MMLU-Medical	Avg
Human (pass)	60.0	60.0			
Human (expert)	87.0	78.0	90.0		
`Models based on Llama`					
MedAlpaca (7B)†	$40.1_{\pm 0.4}$	$73.6_{\pm 0.3}$	$37.0_{\pm 0.3}$	$55.1_{\pm 1.1}$	51.4
MEDITRON-7B*‡	$52.0_{\pm -}$	$74.4_{\pm -}$	$59.2_{\pm -}$	$54.2_{\pm -}$	60.0
Meerkat-8B*‡	**74.2**$_{\pm -}$	-	**62.7**$_{\pm -}$	**75.2**$_{\pm -}$	**70.7**
MEG-Llama (8B)	$66.0_{\pm 0.2}$	**78.0**$_{\pm 0.3}$	$\underline{60.6}_{\pm 0.3}$	$\underline{74.9}_{\pm 0.7}$	$\underline{69.9}$
`Models based on Mistral-Instruct 7B`					
Mistral-Instruct-v0.1†	$42.0_{\pm 0.2}$	$73.8^{\triangledown}_{\pm 0.4}$	$46.1_{\pm 0.1}$	$59.1_{\pm 1.0}$	55.3
BioMistral†	$50.6_{\pm 0.3}$	$77.5_{\pm 0.1}$	$48.1_{\pm 0.2}$	$59.1_{\pm 1.3}$	58.8
BioMistral DARE†	$51.1_{\pm 0.3}$	$\underline{77.7}_{\pm 0.1}$	$48.7_{\pm 0.1}$	$61.9_{\pm 1.2}$	59.9
Meerkat-7B ‡	$\underline{70.3}_{\pm -}$	-	$\underline{60.6}_{\pm -}$	$70.5_{\pm -}$	67.1
MEG-Mistral1	$54.6_{\pm 0.2}$	$74.6_{\pm 0.6}$	$56.4_{\pm 0.4}$	$60.3_{\pm 0.9}$	61.5
MEG-Mistral3	$60.8_{\pm 0.2}$	$74.4_{\pm 0.5}$	$58.4_{\pm 0.6}$	$68.2_{\pm 0.4}$	65.5

MMLU-Medical						
Clinical K.	Genetics	Anatomy	P. Medicine	C. Biology	C. Medicine
`Models based on Llama`						
$53.1_{\pm 0.9}$	$58.0_{\pm 2.2}$	$54.1_{\pm 1.6}$	$58.8_{\pm 0.3}$	$58.1_{\pm 1.3}$	$48.6_{\pm 0.5}$
$57.2_{\pm -}$	$64.6_{\pm -}$	$49.3_{\pm -}$	$55.4_{\pm -}$	$53.8_{\pm -}$	$44.8_{\pm -}$
$74.3_{\pm -}$	$76.7_{\pm -}$	$74.8_{\pm -}$	$75.3_{\pm -}$	$76.1_{\pm -}$	$74.3_{\pm -}$
$72.3_{\pm 0.5}$	$83.0_{\pm 1.5}$	$64.5_{\pm 0.7}$	$79.4_{\pm 0.3}$	$80.6_{\pm 0.4}$	$69.4_{\pm 0.9}$
`Models based on Mistral-Instruct 7B`						
$62.9_{\pm 0.2}$	$57.0_{\pm 0.8}$	$55.6_{\pm 1.0}$	$59.4_{\pm 0.6}$	$62.5_{\pm 1.0}$	$57.2_{\pm 2.1}$
$59.9_{\pm 1.2}$	$64.0_{\pm 1.6}$	$56.5_{\pm 1.8}$	$60.4_{\pm 0.5}$	$59.0_{\pm 1.5}$	$54.7_{\pm 1.0}$
$62.3_{\pm 1.3}$	$67.0_{\pm 1.6}$	$55.8_{\pm 0.9}$	$61.4_{\pm 0.3}$	$66.9_{\pm 2.3}$	$58.0_{\pm 0.5}$
$71.6_{\pm -}$	$74.8_{\pm -}$	$63.2_{\pm -}$	$77.3_{\pm -}$	$70.8_{\pm -}$	$65.2_{\pm -}$
$58.1_{\pm 0.8}$	$68.7_{\pm 0.2}$	$54.4_{\pm 0.5}$	$62.9_{\pm 0.9}$	$61.1_{\pm 2.2}$	$56.6_{\pm 1.0}$
$64.9_{\pm 0.2}$	$69.6_{\pm 0.8}$	$63.0_{\pm 1.0}$	$72.8_{\pm 0.4}$	$73.6_{\pm 0.0}$	$65.2_{\pm 0.2}$

Table 7.2: Main results on four medical multiple-choice question answering benchmarks (left) and fine-grained results on MML-Medical (right). We report accuracy (↑) and standard deviation (↓), when available, of other 7B and 8B medical open-source models. Avg stands for average across datasets. †Results reproduced by Labrak et al. (2024). ▽Reproduced by us with the same data splits used in this work. ‡Results from the original papers.

in interpreting large amounts of information. However, fine-tuning the model with triples boosts accuracy, even surpassing BioMistral, a model adapted from Mistral-Instruct-v0.1 through continued pretraining on curated biomedical data. This baseline highlights the value of graph data for the LLM, but it still does not fully leverage the structural and semantic information provided by the KG.

KGE-augmented LLMs show accuracy gains To fully exploit the KG's rich structural information, we replace the text triples by node embeddings. This approach also compacts the KG's information into a much shorter input sequence. Table 7.2 shows performance on the four medical multiple-choice question answering datasets (left), and a fine-grained evaluation on MMLU-Medical subjects (right). Across all datasets, MEG-Mistral consistently outperforms the baselines' accuracy, and the merged model DARE, the best BioMistral model in Labrak et al. (2024). The only exception is PubMedQA, where BioMistral surpasses MEG. Their higher accuracy may result from using the artificially labeled training set. Instead, we rather train with the small subset of manually labeled samples to avoid biasing the model towards yes/no answers (see § 7.5). The high accuracy on MMLU-Medical indicates that MEG retains good generalization capabilities. Current SOTA for 7B models, Meerkat-7B, proves the effectiveness of training on chain-of-

thought (CoT) in-domain synthetic data. Future work includes exploring CoT instruction tuning in our phase I of training, exploiting information from graph triples instead of relying only on entity labels.

7.6.1 Ablation study

In this section, we evaluate how the choice of graph encoder and mapping network architecture impact MEG-Mistral1's performance on a downstream task (case study on MedQA).

On the impact of the graph encoder We train encoders based on random-walk (RDF2Vec Ristoski and Paulheim (2016)), energy (DistMult Yang et al. (2015)) and message-passing (GraphSAGE Hamilton et al. (2017) and eGraphSAGE, an edge-type-aware variant inspired by Hu et al. (2020)'s adaptation). Along with their impact in MEG's performance, we include a link classification task as a proxy to evaluate their capabilities. Since these encoders are fundamentally distinct, they capture diverse graph properties, as reflected in classification accuracy in Figure 7.4, plain (orange) bars. eGraphSAGE stands out with a considerably higher score (73.9), as it naturally integrates edge-type information during training.

However, higher accuracy in a graph-oriented task such as link classification, does not lead to better performance in a language-oriented downstream task in MEG. When we integrate these KGEs in MEG-Mistral1 and evaluate zero-shot and fine-tune settings on MedQA (Figure 7.4, stripped bars), eGraphSAGE's substantial advantage in link classification does not carry over to MEG-Mistral1, as evidenced by the smaller performance gap across encoders (ranging from 52.1 to 54.2). This suggests that the role of the KGEs in our setup aligns with our intuition: they guide the answer generation by activating the LLM's semantic region that leads to the correct answer. The difference is more notable in a zero-shot setting, where RDF2Vec produces the highest rate of not valid answers (NA). DistMult's lower NA rate indicates it may better align with the LLM's embedding space.

On the impact of the mapping network To assess the impact of the mapping network architecture we replace our 4-layer, 128-dimensional MLP (MLP 4×128) with the alternative designs from Table 7.3. Our final choice, MLP 4×128, outperforms all others in the fine-tuning setting while maintain-

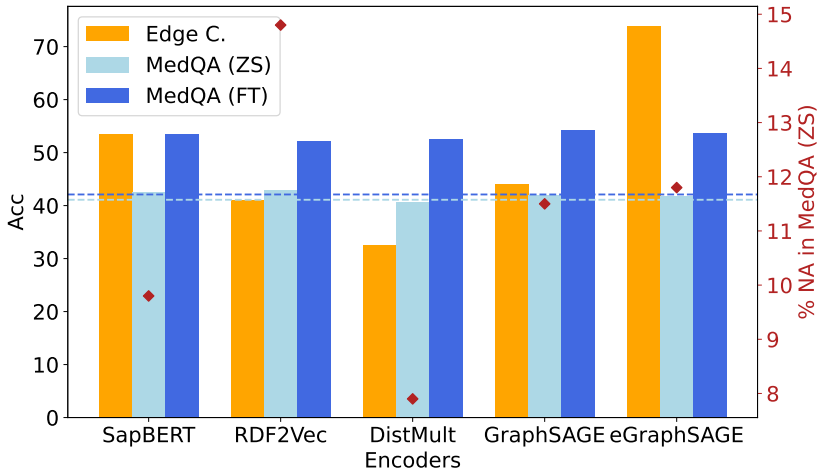


Figure 7.4: Ablation study on KG encoder choice. Plain bars show edge classification accuracy over UMLS; stripped bars show MEG-Mistral1's zero-shot (//) and fine-tuned (\\) accuracy on MedQA; the dashed line represents accuracy with random embeddings; red dots mark the ratio of not valid answers (NA) in the zero-shot setting.

	Parameters (M)	**MedQA (ZS)**	**MedQA (FT)**
MLP 2×512	4.98	41.6	53.5
MLP 2×384	3.74	41.5	53.3
Transf. 4×128	2.18	39.9	53.2
MLP 4×128	1.22	41.6	54.2
MLP 3×128	1.18	42.7	52.7
MLP 2×128	1.15	38.2	51.3

Table 7.3: Comparison of MEG-Mistral1's accuracy on MedQA with different mapping networks. ZS: zero-shot; FT: fine-tuned.

ing a small size. The transformer's lower score highlights the unnecessary overhead of attention layers, as the mapping network's task of embedding transformation does not benefit from attending the input sequence.

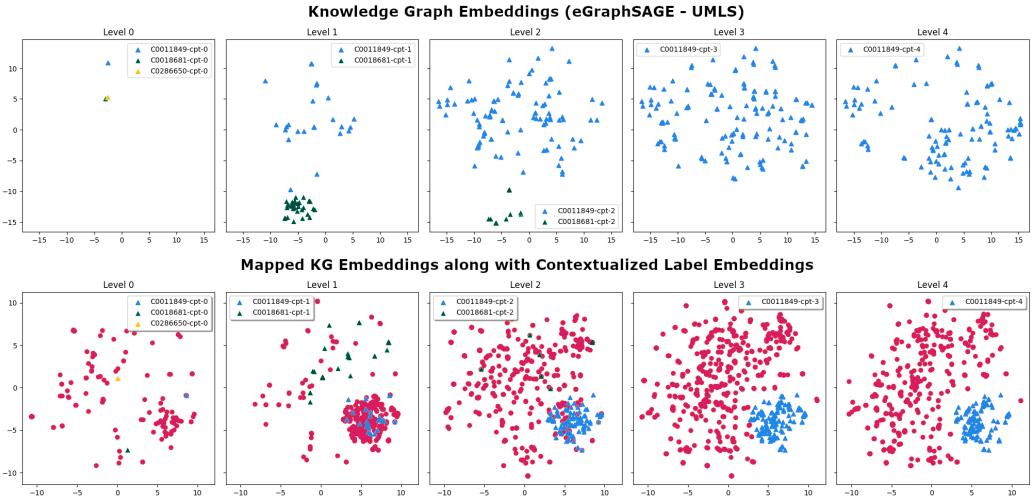


Figure 7.5: t-SNE visualization of the embeddings: before and after the mapping network. After mapping, the relative KGEs' structure along hierarchy levels is preserved, albeit slightly rotated (*e.g.*, in Level 4, see the diagonal gap which hints the orientation of the blobs) and with reversed sparsity. Note the clustering effect over contextualized label embeddings: the mapped KGEs draw them to a specific region.

7.6.2 Qualitative Analysis

This section provides insights into the representation spaces before and after the mapping network, comparing them with the LLM's vector space.

Visualizing the embeddings To track the mapped KGEs, we select three UMLS concepts (entities) representing different semantic and specificity levels within the graph's hierarchy, measured by the number of descendants (IS_A and SUBCLASS_OF relations): "Diabetes Mellitus" (CUI: C0011849, a broad disease category), "Headache" (CUI: C0018681, a symptom), and "Atorvastatin Calcium" (CUI: C0286650, a specific pharmacological substance).

Figure 7.5 depicts t-SNE[14] plots of the concepts (Level 0) and their hierarchies (Level 1 to 4) in the KGE space (top row) and the mapped KGE

[14]We have also witnessed the same overall behavior using UMAP and MDS to visualize the embeddings.

along with the concept's contextualized embeddings in the LLM space (bottom row). Such contextualized embeddings are the LLM's token embeddings generated for the labels of the concepts when verbalizing their KG's triples, *i.e.*, they represent the concepts in all the contexts given by the KG.

Examining the distribution of concepts in the upper and lower rows, we observe two effects of the mapping. First, the relative structure of KGEs is preserved after mapping to the LLM's space, albeit slightly rotated (see Level 1 and Level 4) and with reversed sparsity (high-density groups become less dense, while low-density groups tighten; yet the global between-group structure is the same)[15]. Second, the clustering effect of the contextualized label embeddings (see Level 0): while the labels (circles in the bottom row) are scattered, the mapped embeddings cover such spaces. As we descend in the graph's hierarchy, the contextualized labels become increasingly scattered, yet the mapped KGEs, while remaining within the same region, stay more cohesive. Note that these contextualized labels are distinct vectors for the same label across different contexts. Indeed, their sparsity suggests that they do not completely capture the conceptual dimension of the entity they represent. The mapping seems to activate the LLM's embedding region associated with each concept, providing semantic information the model uses to condition its answer generation.

Probing MEG for acquired knowledge We informally probe MEG to estimate its acquired medical knowledge. We handcraft a set of open-ended questions to query Mistral-Instruct before and after the phase I training on UMLS (§ 7.4.2), and MEG. As an example, we show an excerpt (irrelevant information to the question is omitted for brevity) of models' answers to the question *"What are the active and inactive compounds present in Atorvastatin?"*[16] and highlight the key differences. MEG uses the KGE for Atorvastatin and provides a more knowledgeable response.

Mistral-Instruct (baseline): …The active compound in Atorvastatin is Atorvastatin itself, while the inactive compounds are Atorvastatin calcium salt and Atorvastatin magnesium salt.

Mistral-Instruct after phase I: …The active compound present in Atorvastatin is Atorvastatin calcium. …is responsible for the drug's therapeutic

[15]Given that KGE's and LLM's embeddings represent two spaces with different dimensionalities, we applied t-SNE separately to each; thus, a difference in scale is expected.

[16]This question targets Atorvastatin (C0286650 in UMLS).

effect. There are no inactive compounds present in Atorvastatin.
MEG-Mistral1: …the active compound Atorvastatin calcium.The inactive compounds present in Atorvastatin include excipients such as microcrystalline cellulose, croscarmellose sodium, magnesium stearate, and sodium lauryl sulfate …

7.7 Analysis and Discussion

Our approach is efficient not only because it trains a small fraction of the LLM's parameters but also because it seamlessly handles out-of-vocabulary terms. A new term's KGE can be initialized, for instance, by averaging its one-hop neighbors in the KG, making MEG both lightweight and adaptable to new vocabulary. The efficacy of this method should be evaluated in future work.

Garikipati et al. (2024) demonstrate that prompt engineering can outperform fine-tuning in medical QA for open-source LLMs. However, our focus was to investigate the viability of integrating knowledge from KG embeddings into LLMs rather than optimizing for peak downstream performance. Our experiments show that supervised fine-tuning of KGE-augmented LLMs yields more accurate answers than other specialized baselines. Chain-of-thought tuning, as shown by Kim et al. (2024), is another promising step forward to improve MEG's accuracy. MEG improves response generation by injecting KGEs in a single generation step. This suggests that MEG could also benefit from chain-of-thought tuning, as each of the reasoning steps would increase precision of the model's response.

Besides, the sensitivity of LLMs to the information order in multiple-choice questions, also known as positional bias, is well-documented Pezeshkpour and Hruschka (2023); Zheng et al. (2024). More specific to biomedicine, Liévin et al. (2024) and Hager et al. (2024) show how variations in sequence can significantly impact diagnostic accuracy of medical-aligned language models. However, the robustness to changes in information order remain understudied in most medical model evaluations. Recent studies Wang et al. (2024b,a) find that greedy decoding combined with text answer evaluation gives more consistent answers compared to first-token evaluation, particularly for instruction-tuned LLMs. Thus, in an attempt to alleviate this issue, we inspect the text answer generated by the model instead of ranking the candidate answers by the log probability of its first token

prediction. However, whether KGEs further help the LLMs in mitigating positional biases —as well as other Lyell and Coiera (2017); Moor et al. (2023); Ness et al. (2024) biases— needs to be explored in future work.

7.8 Conclusion

We introduce MEG, a novel medical knowledge-augmented LLM based on KGEs for question answering tasks. To the best of our knowledge, we are the first to inject pretrained KGEs into an LLM via a lightweight mapping network, enabling the model to interpret structural graph information from the medical domain. We present a comprehensive evaluation on four medical multiple-choice question benchmarks, revealing that LLMs can highly benefit from the factual information encoded in KG embeddings. Our results suggest that integrating KGEs with LLMs offers a promising path towards specialized language models.

Acknowledgments

We are grateful to CoAStaL members for their comments on earlier versions of this project. Special thanks to Rita Ramos, Constanza Fierro and Jonas F. Lotz for their insightful feedback and to Nicolas Garneau for his support during initial experiments. Laura Cabello was supported by the Novo Nordisk Foundation (grant NNF 20SA0066568). The work of Carlos Bobed was supported by the I+D+i project PID2020-113903RB-I00 (funded by MCIN/AEI/ 10.13039/501100011033) and the project T42_23R (funded by Gobierno de Aragón, Spain).

Part III

Conclusion

Chapter 8

Discussion

The primary goal of this dissertation was to deepen our understanding of language model representations. The included publications have collectively advanced research on probing language models for social bias, fairness, and task-oriented adaptability. The published works in Part I explored model evaluation beyond traditional performance metrics, and delved into the nuanced interplay between bias and fairness, as well as their independence. Part II complemented our comprehension of NLP technologies by exploring cross-cultural transfer learning and enhancing LLMs' ability to answer specialized domain questions using knowledge graphs.

Open Problems As discussed throughout various chapters of this thesis, the contested nature of fairness and its multidisciplinary implications make it impractical to adopt a single universal definition. This conceptual ambiguity has been, and continues to be, a significant challenge in both theoretical and empirical analysis, complicating the comparison of measurements and the reading of reliable conclusions. We have shown that concepts are inconsistently applied across studies which, in turn, undermines the coherence of fairness research and gives the false hope of progress within this field.

Although the performance of language models in isolation may not fully reflect the impact of AI systems in everyday use (Reuel et al., 2024), a thorough understanding of their potential risks and challenges is crucial for the safe deployment of more complex AI systems. As Eckhouse et al. (2019) put it in the context of risk assessment models, "without assurances about the foundational layers, the fairness of the top layers is irrelevant". This is, the just behaviour observed in the higher, more visible parts of a system depends

on the integrity and trustworthiness of the underlying components. In this regard, we identify a non-exhaustive list of open problems and directions for future research on language modelling.

a. Deploying effective evaluation benchmarks that are dynamic (continuous adaptation), relevant (LLMs are evaluated in an scenario as close as possible to their real-world application), and prioritize inclusivity (representing diverse cultures and user demographics).

b. Establishing reliable approaches to anticipate the broader societal impact of language models and, ultimately, AI systems. While comprehending AI's societal impact is a central aim of much research, it remains a complex challenge that spans both social and technical dimensions.

c. Optimizing for equal performance rather than average performance. Research should strive for more inclusive and equitable models, thereby prioritizing to optimize for performance of the least advantage. This approach has the potential to address and mitigate performance disparities embedded in current systems, promoting the deployment of fairer and more equitable AI.

d. Assessing intersectional biases. Evaluating intersectional biases involves understanding how multiple attributes interact to influence the manifestation of bias in language, such as examining how gender and race together impact the way individuals are represented. For instance, a language model might consistently associate certain professions with women of one racial group while neglecting others, reflecting compounded stereotypes.

e. Monitoring the long-term effects of technology usage disparities across social groups. As AI becomes increasingly integrated into our daily lives, it is crucial to gain a deeper understanding of its role in either exacerbating or alleviating social challenges.

f. Promoting cultural awareness. Efforts to deploy AI systems, such as content moderators or dialogue bots, should ensure both consistent behaviour across diverse cultures and adaptability to specific cultural contexts.

g. Adopting hybrid approaches. To mitigate the need for training ever-larger models and address current limitations of LLMs—including issues like hallucinations, high environmental costs, and lack of explainability—future research should prioritize leveraging the advantages of hybrid strategies. This requires going beyond simply scaling up LLMs and focusing on improving systems that integrate external knowledge sources, such as vector databases or knowledge graphs, to develop more efficient and reliable solutions.

As we stand at the forefront of the Fourth Industrial Revolution, the past decades have brought remarkable technological advancements that are reshaping the way we interact with the world. From how we communicate with each other to how we solve complex problems, technology is now a cornerstone of our daily lives. However, with these incredible opportunities come significant responsibilities. The journey of AI is far from complete—it has only just begun. There is much work ahead to ensure its safe deployment and responsible adoption in society, paving the way for innovations that not only enhance our capabilities but also uphold our shared values and ethical principles.

Bibliography

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.385. URL https://aclanthology.org/2020.acl-main.385.

Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. Evaluating clip: Towards characterization of broader capabilities and downstream implications, 2021. URL https://arxiv.org/abs/2108.02818.

Ameen Ali, Thomas Schnake, Oliver Eberle, Grégoire Montavon, Klaus-Robert Müller, and Lior Wolf. XAI for transformers: Better explanations through conservative propagation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 435–451. PMLR, 2022a. URL https://proceedings.mlr.press/v162/ali22a.html.

Ameen Ali, Thomas Schnake, Oliver Eberle, Grégoire Montavon, Klaus-Robert Müller, and Lior Wolf. Xai for transformers: Better explanations through conservative propagation, 2022b. URL https://arxiv.org/abs/2202.07304.

Ameen Ali, Thomas Schnake, Oliver Eberle, Grégoire Montavon, Klaus-Robert Müller, and Lior Wolf. Xai for transformers: Better explanations through conservative propagation, 2022c. URL https://arxiv.org/abs/2202.07304.

J. E. J. Altham. Rawls's difference principle. *Philosophy*, 48(183):75–78, 1973. doi: 10.1017/S0031819100060447.

Jack J Amend, Albatool Wazzan, and Richard Souvenir. Evaluating gender-neutral training data for automated image captioning. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 1226–1235, 2021. doi: 10.1109/BigData52589.2021.9671774.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2425 – 2433, 2015. URL https://doi.org/10.1109/ICCV.2015.279.

Maria Antoniak and David Mimno. Bad seeds: Evaluating lexical methods for bias measurement. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1889–1904, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.148. URL https://aclanthology.org/2021.acl-long.148.

Ayme Arango Monnar, Jorge Perez, Barbara Poblete, Magdalena Saldaña, and Valentina Proust. Resources for multilingual hate speech detection. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 122–130, Seattle, Washington (Hybrid), July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.woah-1.12. URL https://aclanthology.org/2022.woah-1.12.

Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. Explaining recurrent neural network predictions in sentiment analysis. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 159–168, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5221. URL https://aclanthology.org/W17-5221.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS*

ONE, 10(7):1–46, 07 2015. doi: 10.1371/journal.pone.0130140. URL https://doi.org/10.1371/journal.pone.0130140.

Ioana Baldini, Dennis Wei, Karthikeyan Natesan Ramamurthy, Mikhail Yurochkin, and Moninder Singh. Your fairness may vary: Pretrained language model fairness in toxic text classification, 2021. URL https://arxiv.org/abs/2108.01250.

Esma Balkir, Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen Fraser. Challenges in applying explainability methods to improve the fairness of NLP models. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*, pages 80–92, Seattle, U.S.A., July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.trustnlp-1.8. URL https://aclanthology.org/2022.trustnlp-1.8.

Kate Barasz and Tami Kim. Choice perception: Making sense (and nonsense) of others' decisions. *Current opinion in psychology*, 43:176–181, 2022. ISSN 2352-250X.

Lizzie Barmes. Equality law and experimentation: The positive action challenge. *The Cambridge Law Journal*, 68(3):623–654, 2009. ISSN 00081973, 14692139. URL http://www.jstor.org/stable/40388838.

Solon Barocas and Andrew D. Selbst. Big Data's Disparate Impact. *California Law Review*, 104:671–732, 2016. doi: 10.2139/ssrn.2477899. URL https://ssrn.com/abstract=2477899.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

Marion Bartl, Malvina Nissim, and Albert Gatt. Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 1–16, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.gebnlp-1.1.

Adrien Basdevant, Camille François, Victor Storchan, Kevin Bankston, Ayah Bdeir, Brian Behlendorf, Merouane Debbah, Sayash Kapoor, Yann LeCun, Mark Surman, Helen King-Turvey, Nathan Lambert, Stefano Maffulli, Nik

Marda, Govind Shivkumar, and Justine Tunney. Towards a framework for openness in foundation models: Proceedings from the columbia convening on openness in artificial intelligence, 2024. URL https://arxiv.org/abs/2405.15802.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2007. URL https://aclanthology.org/S19-2007.

Kevin Baum, Susanne Mantel, Timo Speith, and Eva Schmidt. From responsibility to reason-giving explainable artificial intelligence. *Philosophy and Technology*, 35(1):1–30, 2022.

Nat Hum Behav. Fair play in research and policy. *Nature Human Behaviour*, 1(10):693–693, 2017. ISSN 2397-3374. doi: 10.1038/s41562-017-0231-1. URL https://doi.org/10.1038/s41562-017-0231-1.

Beata Beigman Klebanov and Eyal Beigman. Squibs: From annotator agreement to noise models. *Computational Linguistics*, 35(4):495–503, December 2009. doi: 10.1162/coli.2009.35.4.35402. URL https://aclanthology.org/J09-4005.

Claus Beisbart and Tim Räz. Philosophy of science at sea: Clarifying the interpretability of machine learning. *Philosophy Compass*, 17(6):e12830, 2022.

Camiel J. Beukeboom. *Mechanisms of linguistic bias: How words reflect and maintain stereotypic expectancies.*, pages 313–330. Psychology Press, 2014.

Shruti Bhargava and David Forsyth. Exposing and correcting the gender bias in image captioning datasets and models, 2019. URL https://arxiv.org/abs/1912.00578.

Sudeep Bhatia. The semantic representation of prejudice and stereotypes. *Cognition*, 164:46–60, 2017. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2017.03.016. URL https://www.sciencedirect.com/science/article/pii/S0010027717300872.

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1493–1504, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594095. URL https://doi.org/10.1145/3593013.3594095.

Felix Biessmann. Automating political bias prediction. arXiv preprint, 2016. arXiv:1608.02195.

Arpita Biswas, Siddharth Barman, Amit Deshpande, and Amit Sharma. Quantifying infra-marginality and its trade-off with group fairness. *CoRR*, abs/1909.00982, 2019. URL http://arxiv.org/abs/1909.00982.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.485. URL https://aclanthology.org/2020.acl-main.485.

Olivier Bodenreider. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–D270, 2004. ISSN 0305-1048. doi: 10.1093/nar/gkh061. URL https://doi.org/10.1093/nar/gkh061.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel,

Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2022. URL https://arxiv.org/abs/2108.07258.

Stephanie Brandl, Ruixiang Cui, and Anders Søgaard. How conservative are language models? adapting to the introduction of gender-neutral pronouns. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3624–3630, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.265. URL https://aclanthology.org/2022.naacl-main.265.

William Brannon, Wonjune Kang, Suyash Fulay, Hang Jiang, Brandon Roy, Deb Roy, and Jad Kabbara. ConGraT: Self-supervised contrastive pretraining for joint graph and text embeddings. In Dmitry Ustalov, Yanjun Gao, Alexander Panchenko, Elena Tutubalina, Irina Nikishina, Arti Ramesh, Andrey Sakhovskiy, Ricardo Usbeck, Gerald Penn, and Marco Valentino, editors, *Proceedings of TextGraphs-17: Graph-based*

Methods for Natural Language Processing, pages 19–39, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.textgraphs-1.2.

Dana Brin, Vera Sorin, Akhil Vaid, Ali Soroush, Benjamin S. Glicksberg, Alexander W. Charney, Girish Nadkarni, and Eyal Klang. Comparing ChatGPT and GPT-4 performance in USMLE soft skill assessments. *Scientific Reports*, 13:16492, 2023. doi: 10.1038/s41598-023-43436-9. URL https://doi.org/10.1038/s41598-023-43436-9.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. Understanding the origins of bias in word embeddings. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 803–811. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/brunet19a.html.

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language BERTs. *Transactions of the Association for Computational Linguistics*, 9:978–994, 2021. doi: 10.1162/tacl_a_00408. URL https://aclanthology.org/2021.tacl-1.58.

Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulić. IGLUE: A benchmark for transfer learning across modalities, tasks, and languages. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2370–2392. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/bugliarello22a.html.

William Bunting, Lynda Garcia, and Ezekiel Edwards. The war on marijuana in black and white. American Civil Liberties Union, June 2013. Available at SSRN: https://ssrn.com/abstract=2819708.

Kaylee Burns, Lisa Anne Hendricks, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models, 2018. URL https://arxiv.org/abs/1803.09797.

Laura Cabello and Uchenna Akujuobi. It is Simple Sometimes: A Study On Improving Aspect-Based Sentiment Analysis Performance. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6597–6610, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.394. URL https://aclanthology.org/2024.findings-acl.394.

Laura Cabello, Emanuele Bugliarello, Stephanie Brandl, and Desmond Elliott. Evaluating Bias and Fairness in Gender-Neutral Pretrained Vision-and-Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8465–8483, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.525. URL https://aclanthology.org/2023.emnlp-main.525.

Laura Cabello, Anna Katrine Jørgensen, and Anders Søgaard. On the independence of association bias and empirical fairness in language models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 370–378, New York, NY, USA, June 2023b. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594004. URL https://doi.org/10.1145/3593013.3594004.

Laura Cabello, Carmen Martin-Turrero, Uchenna Akujuobi, Anders Søgaard, and Carlos Bobed. MEG: Medical Knowledge-Augmented Large Language Models for Question Answering. *Under review*, 2024.

Laura Cabello Piqueras and Anders Søgaard. Are pretrained multilingual models equally fair across languages? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3597–3605, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.318.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017. doi: 10.1126/science.aal4230. URL https://www.science.org/doi/abs/10.1126/science.aal4230.

Aylin Caliskan, Pimparkar Parth Ajay, Tessa Charlesworth, Robert Wolfe, and Mahzarin R. Banaji. Gender bias in word embeddings: A comprehensive analysis of frequency, syntax, and semantics. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, page 156–170, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392471. doi: 10.1145/3514094.3534162. URL https://doi.org/10.1145/3514094.3534162.

Yang Trista Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 561–570, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.62. URL https://aclanthology.org/2022.acl-short.62.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.c3nlp-1.7. URL https://aclanthology.org/2023.c3nlp-1.7.

Nicolas Carl, Franziska Schramm, Sarah Haggenmüller, Jakob Nikolas Kather, Martin J. Hetz, Christoph Wies, Maurice Stephan Michel, Frederik Wessels, and Titus J. Brinker. Large language model use in clinical oncology. *npj Precision Oncology*, 8:240, 2024. doi: 10.1038/s41698-024-00733-4. URL https://doi.org/10.1038/s41698-024-00733-4.

Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Penco, and Andrea Cosentini. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12, 03 2022a. doi: 10.1038/s41598-022-07939-1.

Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12(1), March 2022b. doi: 10.1038/s41598-022-07939-1. URL https://doi.org/10.1038/s41598-022-07939-1.

Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12:4209, 2022c. doi: 10.1038/s41598-022-07939-1. URL https://doi.org/10.1038/s41598-022-07939-1.

Junyi Chai, Taeuk Jang, and Xiaoqian Wang. Fairness without demographics through knowledge distillation. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=8gjwWnN5pfy.

Ilias Chalkidis, Tommaso Pasini, Sheng Zhang, Letizia Tomada, Sebastian Schwemer, and Anders Søgaard. FairLex: A multilingual benchmark for evaluating fairness in legal text processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4389–4406, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.301. URL https://aclanthology.org/2022.acl-long.301.

Kaytlin Chaloner and Alfredo Maldonado. Measuring gender bias in word embeddings across domains and discovering new gender bias word categories. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 25–32, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3804. URL https://aclanthology.org/W19-3804.

Kai-Wei Chang, Vinodkumar Prabhakaran, and Vicente Ordonez. Bias and fairness in natural language processing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts*, Hong Kong, China, November 2019. Association for Computational Linguistics. URL https://aclanthology.org/D19-2004.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020a. URL https://arxiv.org/abs/2002.05709.

Wei-Fan Chen, Khalid Al Khatib, Henning Wachsmuth, and Benno Stein. Analyzing political bias and unfairness in news articles at different levels of granularity. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 149–154, Online, November 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.nlpcss-1.16. URL https://aclanthology.org/2020.nlpcss-1.16.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. MEDITRON-70B: Scaling medical pretraining for large language models, 2023a. URL https://arxiv.org/abs/2311.16079.

Zhenpeng Chen, Jie M. Zhang, Federica Sarro, and Mark Harman. A comprehensive empirical study of bias mitigation methods for machine learning classifiers, 2023b.

Zhoujun Cheng, Jungo Kasai, and Tao Yu. Batch prompting: Efficient inference with large language model APIs. In Mingxuan Wang and Imed Zitouni, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 792–810, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-industry.74. URL https://aclanthology.org/2023.emnlp-industry.74.

Cheng-Han Chiang and Hung-yi Lee. Re-examining human annotations for interpretable nlp, 2022. URL https://arxiv.org/abs/2204.04580.

Eli Chien, Wei-Cheng Chang, Cho-Jui Hsieh, Hsiang-Fu Yu, Jiong Zhang, Olgica Milenkovic, and Inderjit S Dhillon. Node feature extraction by self-supervised multi-scale neighborhood prediction. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=KJggliHbs8.

Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers, 2022. URL https://arxiv.org/abs/2202.04053.

Rochelle Choenni and Ekaterina Shutova. What does it mean to be language-agnostic? probing multilingual sentence encoders for typological properties, 2020.

Myung Jin Choi, Antonio Torralba, and Alan S. Willsky. Context models and out-of-context objects. *Pattern Recognition Letters*, 33 (7):853–862, 2012. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2011.12.004. URL https://www.sciencedirect.com/science/article/pii/S0167865511004235. Special Issue on Awards from ICPR 2010.

Monojit Choudhury and Amit Deshpande. How linguistically fair are multilingual pre-trained language models? In *AAAI-21*. AAAI, AAAI, February 2021. URL https://www.microsoft.com/en-us/research/publication/how-linguistically-fair-are-multilingual-pre-trained-language-models/.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.

Sam Corbett-Davies, Johann D. Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. The measure and mismeasure of fairness. *J. Mach. Learn. Res.*, 24(1), March 2024. ISSN 1532-4435.

Kate Crawford. The trouble with bias. In *Conference on Neural Information Processing Systems, invited speaker*, 2017.

Kimberle Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *The University of Chicago Legal Forum*, 140:139–167, 1989.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. Quantifying Social Biases in NLP: A Generalization and Empirical Comparison of Extrinsic Fairness Metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267, 11 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00425. URL https://doi.org/10.1162/tacl_a_00425.

Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning, 2023. URL https://arxiv.org/abs/2307.08691.

Mithun Das, Somnath Banerjee, Punyajoy Saha, and Animesh Mukherjee. Hate speech and offensive language detection in Bengali. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 286–296, Online only, November 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.aacl-main.23.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515, May 2017. doi: 10.1609/icwsm.v11i1.14955. URL https://ojs.aaai.org/index.php/ICWSM/article/view/14955.

Erenay Dayanik and Sebastian Padó. Masking actor information leads to fairer political claims detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4385–4391, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.404. URL https://aclanthology.org/2020.acl-main.404.

Erenay Dayanik and Sebastian Padó. Disentangling document topic and author gender in multiple languages: Lessons for adversarial debiasing. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–61, Online, April 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wassa-1.6.

Pieter Delobelle, Ewoenam Kwaku Tokpo, Toon Calders, and Bettina Berendt. Measuring fairness with biased rulers: A survey on quantifying biases in pretrained language models. *CoRR*, abs/2112.07447, 2021. URL https://arxiv.org/abs/2112.07447.

Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. COLD: A benchmark for Chinese offensive language detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11580–11599, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi:

10.18653/v1/2022.emnlp-main.796. URL https://aclanthology.org/2022.emnlp-main.796.

Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang. Harms of gender exclusivity and challenges in non-binary representation in language technologies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1994, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.150. URL https://aclanthology.org/2021.emnlp-main.150.

Sunipa Dev, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. On measures of biases and harms in NLP. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 246–267, Online only, November 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-aacl.24.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.408. URL https://aclanthology.org/2020.acl-main.408.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360128. doi: 10.1145/3278721.3278729. URL https://doi.org/10.1145/3278721.3278729.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping, 2020. URL https://arxiv.org/abs/2002.06305.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar

Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana

Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang,

Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Coen Van Duijnhoven. Predicting political preference through content- and stylistic text features and distant labeling. Master's thesis, Tilburg University, 2018.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, page 214–226, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450311151. doi: 10.1145/2090236.2090255. URL https://doi.org/10.1145/2090236.2090255.

Oliver Eberle, Ilias Chalkidis, Laura Cabello, and Stephanie Brandl. Rather a Nurse than a Physician - Contrastive Explanations under Investigation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6907–6920, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.427. URL https://aclanthology.org/2023.emnlp-main.427.

Lauren Eckhouse, Kristian Lum, Cynthia Conti-Cook, and Julie Ciccolini. Layers of bias: A unified approach for understanding problems with risk assessment. *Criminal Justice and Behavior*, 46(2):185–209,

2019. doi: 10.1177/0093854818811379. URL https://doi.org/10.1177/0093854818811379.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization, 2024. URL https://arxiv.org/abs/2404.16130.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-3210. URL https://aclanthology.org/W16-3210.

Kawin Ethayarajh and Dan Jurafsky. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.393. URL https://aclanthology.org/2020.emnlp-main.393.

Farahnaz Faez. Reconceptualizing the native/nonnative speaker dichotomy. *Journal of Language, Identity & Education*, 10(4):231–249, 2011. doi: 10.1080/15348458.2011.598127. URL https://doi.org/10.1080/15348458.2011.598127.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.656. URL https://aclanthology.org/2023.acl-long.656.

Shi Feng and Jordan Boyd-Graber. What can ai do for me? evaluating machine learning interpretations in cooperative play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, page 229–239, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362726. doi: 10.1145/3301275.3302265. URL https://doi.org/10.1145/3301275.3302265.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. A survey of race, racism, and anti-racism in NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1905–1925, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.149. URL https://aclanthology.org/2021.acl-long.149.

Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4), jul 2018. ISSN 0360-0300. doi: 10.1145/3232676. URL https://doi.org/10.1145/3232676.

Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im)possibility of fairness, 2016. URL https://arxiv.org/abs/1609.07236.

Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. *Commun. ACM*, 64(4):136–143, mar 2021. ISSN 0001-0782. doi: 10.1145/3433949. URL https://doi.org/10.1145/3433949.

Niklas Friedrich, Anne Lauscher, Simone Paolo Ponzetto, and Goran Glavaš. DebIE: A platform for implicit and explicit debiasing of word embedding spaces. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 91–98, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-demos.11. URL https://aclanthology.org/2021.eacl-demos.11.

Ute Gabriel, Pascal Mark Gygax, and Elisabeth A. Kuhn. Neutralising linguistic sexism: Promising but cumbersome? *Group Processes & Intergroup Relations*, 21:844 – 858, 2018.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770*, 2024.

Mahak Gambhir and Vishal Gupta. Recent automatic text summarization techniques: a survey. *Artif. Intell. Rev.*, 47(1):1–66, January 2017. ISSN 0269-2821. doi: 10.1007/s10462-016-9475-9. URL https://doi.org/10.1007/s10462-016-9475-9.

Zhiwei Gao, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. Offensive language detection on video live streaming chat. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1936–1940, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.175. URL https://aclanthology.org/2020.coling-main.175.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H. Chi, and Alex Beutel. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 219–226, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450363242. doi: 10.1145/3306618.3317950. URL https://doi.org/10.1145/3306618.3317950.

Anurag Garikipati, Jenish Maharjan, Navan Preet Singh, Leo Cyrus, Mayank Sharma, Madalina Ciobanu, Gina Barnes, Qingqing Mao, and Ritankar Das. OpenmedLM: Prompt engineering can out-perform fine-tuning in medical question-answering with open-source large language models. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024. URL https://openreview.net/forum?id=WIHH0iOOUt.

Gemini Team Google. Gemini: A family of highly capable multimodal models, 2024. URL https://arxiv.org/abs/2312.11805.

Sayan Ghosh, Dylan Baker, David Jurgens, and Vinodkumar Prabhakaran. Detecting cross-geographic biases in toxicity modeling on social media, 2021. URL https://arxiv.org/abs/2104.06999.

Kristina Gligoric, Myra Cheng, Lucia Zheng, Esin Durmus, and Dan Jurafsky. NLP systems that can't tell use from mention censor counterspeech, but teaching the distinction helps. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5942–5959, Mexico City, Mexico, June 2024. Association for

Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.331. URL https://aclanthology.org/2024.naacl-long.331.

Edwin L. Goff. Justice as fairness: The practice of social science in a rawlsian model. *Social Research*, 50(1):81–97, 1983. ISSN 0037783X. URL http://www.jstor.org/stable/40958869.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.150. URL https://aclanthology.org/2021.acl-long.150.

Seraphina Goldfarb-Tarrant, Adam Lopez, Roi Blanco, and Diego Marcheggiani. Bias beyond English: Counterfactual tests for bias in sentiment analysis in four languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4458–4468, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.272. URL https://aclanthology.org/2023.findings-acl.272.

Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1061. URL https://aclanthology.org/N19-1061.

Ana Valeria González, Maria Barrett, Rasmus Hvingelby, Kellie Webster, and Anders Søgaard. Type B reflexivization as an unambiguous testbed for multilingual multi-task gender bias. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2637–2648, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.209. URL https://aclanthology.org/2020.emnlp-main.209.

Ana Valeria González, Gagan Bansal, Angela Fan, Yashar Mehdad, Robin Jia, and Srinivasan Iyer. Do explanations help users detect errors in open-domain QA? an evaluation of spoken vs. visual explanations. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1103–1116, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.95. URL https://aclanthology.org/2021.findings-acl.95.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, mar 2021. doi: 10.1007/s11263-021-01453-z. URL https://doi.org/10.1007/s11263-021-01453-z.

J. C. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, March 1975. ISSN 1860-0980. doi: 10.1007/BF02291478. URL https://doi.org/10.1007/BF02291478.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.740. URL https://aclanthology.org/2020.acl-main.740.

Philipp Hacker and Jan-Hendrik Passoth. *Varieties of AI Explanations Under the Law. From the GDPR to the AIA, and Beyond*, pages 343–373. Springer International Publishing, Cham, 2022. ISBN 978-3-031-04083-2. doi: 10.1007/978-3-031-04083-2_17. URL https://doi.org/10.1007/978-3-031-04083-2_17.

R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742, 2006. doi: 10.1109/CVPR.2006.100.

Patrick Hager, Fabian Jungmann, Richard Holland, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, 30:2613–2622, 2024. doi: 10.1038/s41591-024-03097-1.

Melissa Hall, Laurens van der Maaten, Laura Gustafson, Maxwell Jones, and Aaron Adcock. A systematic study of bias amplification, 2022. URL https://arxiv.org/abs/2201.11706.

Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/5dd9db5e033da9c6fb5ba83c7a7ebea9-Paper.pdf.

Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K. Bressem. MedAlpaca – An open-source collection of medical conversational ai models and training data, 2023. URL https://arxiv.org/abs/2304.08247.

Victor Petrén Bach Hansen and Anders Søgaard. Is the lottery fair? evaluating winning tickets across demographics. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3214–3224, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.284. URL https://aclanthology.org/2021.findings-acl.284.

Tatsunori B. Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1934–1943. PMLR, 2018. URL http://proceedings.mlr.press/v80/hashimoto18a.html.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*

(CVPR), pages 9726–9735, 2019. URL https://api.semanticscholar.org/CorpusID:207930212.

Zexue He, Yu Wang, Julian McAuley, and Bodhisattwa Prasad Majumder. Controlling bias exposure for fair interpretable predictions. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5854–5866, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.431. URL https://aclanthology.org/2022.findings-emnlp.431.

Brian Hedden. On statistical criteria of algorithmic fairness. *Philosophy and Public Affairs*, 49(2):209–231, 2021. doi: 10.1111/papa.12189.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

P.J. Henry, Sarah E. Butler, and Mark J. Brandt. The influence of target group status on the perception of the offensiveness of group-based slurs. *Journal of Experimental Social Psychology*, 53:185–192, 2014. ISSN 0022-1031. doi: https://doi.org/10.1016/j.jesp.2014.03.012. URL https://www.sciencedirect.com/science/article/pii/S0022103114000390.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.482. URL https://aclanthology.org/2022.acl-long.482.

F Hintz, M Dijkhuis, V van Hoff, JM McQueen, and AS Meyer. A behavioural dataset for studying individual differences in language skills. *Scientific Data*, 7(1), 2020.

Yusuke Hirota, Yuta Nakashima, and Noa Garcia. Quantifying societal bias amplification in image captioning. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6 2022a. URL https://facctconference.org/static/pdfs_2022/facct22-102.pdf. 10 pages.

Yusuke Hirota, Yuta Nakashima, and Noa Garcia. Gender and racial bias in visual question answering datasets. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1280–1292, New York, NY, USA, 2022b. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533184. URL https://doi.org/10.1145/3531146.3533184.

Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. *Scientific Data*, 5, 2018.

Harry Holzer and David Neumark. Assessing affirmative action. *Journal of Economic Literature*, 38(3):483–568, September 2000. doi: 10.1257/jel.38.3.483. URL https://www.aeaweb.org/articles?id=10.1257/jel.38.3.483.

Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

Sara Hooker. On the limitations of compute thresholds as a governance strategy, 2024. URL https://arxiv.org/abs/2407.05694.

Dirk Hovy and Anders Søgaard. Tagging performance correlates with author age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 483–488, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-2079. URL https://aclanthology.org/P15-2079.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJlWWJSFDH.

Yuntong Hu, Zhihan Lei, Zheng Zhang, Bo Pan, Chen Ling, and Liang Zhao. Grag: Graph retrieval-augmented generation, 2024. URL https://arxiv.org/abs/2405.16506.

Xiaolei Huang, Linzi Xing, Franck Dernoncourt, and Michael J. Paul. Multilingual Twitter corpus and baselines for evaluating demographic bias in hate speech recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1440–1448, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.180.

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Hanchi Sun, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bertie Vidgen, Bhavya Kailkhura, Caiming Xiong, Chaowei Xiao, Chunyuan Li, Eric P. Xing, Furong Huang, Hao Liu, Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao, Manolis Kellis, Marinka Zitnik, Meng Jiang, Mohit Bansal, James Zou, Jian Pei, Jian Liu, Jianfeng Gao, Jiawei Han, Jieyu Zhao, Jiliang Tang, Jindong Wang, Joaquin Vanschoren, John Mitchell, Kai Shu, Kaidi Xu, Kai-Wei Chang, Lifang He, Lifu Huang, Michael Backes, Neil Zhenqiang Gong, Philip S. Yu, Pin-Yu Chen, Quanquan Gu, Ran Xu, Rex Ying, Shuiwang Ji, Suman Jana, Tianlong Chen, Tianming Liu, Tianyi Zhou, William Yang Wang, Xiang Li, Xiangliang Zhang, Xiao Wang, Xing Xie, Xun Chen, Xuyu Wang, Yan Liu, Yanfang Ye, Yinzhi Cao, Yong Chen, and Yue Zhao. Position: TrustLLM: Trustworthiness in large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 20166–20270. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/huang24x.html.

Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In

Conference on Computer Vision and Pattern Recognition (CVPR), pages 6700–6709, 2019. URL https://arxiv.org/abs/1902.09506.

Ben Hutchinson and Margaret Mitchell. 50 years of test (un)fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 49–58, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287600. URL https://doi.org/10.1145/3287560.3287600.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.487. URL https://aclanthology.org/2020.acl-main.487.

Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 375–385, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445901. URL https://doi.org/10.1145/3442188.3445901.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.386. URL https://aclanthology.org/2020.acl-main.386.

Timothy Jay and Kristin Janschewitz. The pragmatics of swearing. *Journal of Politeness Research 4*, 4(2):267–288, 2008. doi: doi:10.1515/JPLR.2008.013. URL https://doi.org/10.1515/JPLR.2008.013.

Jacob Jensen, Ethan Kaplan, Suresh Naidu, and Laurence Wilse-Samson. Political Polarization and the Dynamics of Political Language: Evidence from 130 Years of Partisan Speech. *Brookings Papers on Economic Activity*, 43(2 (Fall)):1–81, 2012. URL https://ideas.repec.org/a/bin/bpeajo/v43y2012i2012-02p1-81.html.

Younghoon Jeong, Juhyun Oh, Jongwon Lee, Jaimeen Ahn, Jihyung Moon, Sungjoon Park, and Alice Oh. KOLD: Korean offensive language dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10818–10833, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.744. URL https://aclanthology.org/2022.emnlp-main.744.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B, 2023. URL https://arxiv.org/abs/2310.06825.

Tonglin Jiang, Hao Li, and Yubo Hou. Cultural differences in humor perception, usage, and implications. *Frontiers in Psychology*, 10, 2019. ISSN 1664-1078. doi: 10.3389/fpsyg.2019.00123. URL https://www.frontiersin.org/articles/10.3389/fpsyg.2019.00123.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14), 2021a. ISSN 2076-3417. doi: 10.3390/app11146421. URL https://www.mdpi.com/2076-3417/11/14/6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1259. URL https://aclanthology.org/D19-1259.

Xisen Jin, Francesco Barbieri, Brendan Kennedy, Aida Mostafazadeh Davani, Leonardo Neves, and Xiang Ren. On transferability of bias mitigation effects in language model fine-tuning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3770–3783, Online, June

2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.296. URL https://aclanthology.org/2021.naacl-main.296.

Anders Johannsen, Dirk Hovy, and Anders Søgaard. Cross-lingual syntactic variation over age and gender. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 103–112, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.18653/v1/K15-1011. URL https://aclanthology.org/K15-1011.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. Learning a POS tagger for AAVE-like language. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1115–1120, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1130. URL https://aclanthology.org/N16-1130.

Anna Katrine Jørgensen and Anders Søgaard. Rawlsian ai fairness loopholes. *AI and Ethics*, 3(4):1185–1192, 2023. ISSN 2730-5961. doi: 10.1007/s43681-022-00226-9. URL https://doi.org/10.1007/s43681-022-00226-9.

Rasmus Jørgensen, Fiammetta Caccavale, Christian Igel, and Anders Søgaard. Are multilingual sentiment models equally right for the right reasons? In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 131–141, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.blackboxnlp-1.11. URL https://aclanthology.org/2022.blackboxnlp-1.11.

Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/eb163727917cbba1eea208541a643e74-Paper.pdf.

Wei Ju, Zheng Fang, Yiyang Gu, Zequn Liu, Qingqing Long, Ziyue Qiao, Yifang Qin, Jianhao Shen, Fang Sun, Zhiping Xiao, Junwei Yang, Jingyang Yuan, Yusheng Zhao, Yifan Wang, Xiao Luo, and Ming Zhang. A comprehensive survey on deep graph representation learning. *Neural Networks*, 173:106207, 2024. ISSN 0893-6080. doi: https://doi.org/10.

1016/j.neunet.2024.106207. URL https://www.sciencedirect.com/science/article/pii/S089360802400131X.

Masahiro Kaneko, Danushka Bollegala, and Naoaki Okazaki. Debiasing isn't enough! – on the effectiveness of debiasing MLMs and their social biases in downstream tasks. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1299–1310, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.111.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. The perils of using Mechanical Turk to evaluate open-ended text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265–1285, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.97. URL https://aclanthology.org/2021.emnlp-main.97.

M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2): 81–93, June 1938. ISSN 0006-3444. doi: 10.1093/biomet/30.1-2.81. URL https://doi.org/10.1093/biomet/30.1-2.81.

Os Keyes. The misgendering machines: Trans/hci implications of automatic gender recognition. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW), nov 2018. doi: 10.1145/3274357. URL https://doi.org/10.1145/3274357.

Hyunjae Kim, Hyeon Hwang, Jiwoo Lee, Sihyeon Park, Dain Kim, Taewhoo Lee, Chanwoong Yoon, Jiwoong Sohn, Donghee Choi, and Jaewoo Kang. Small language models learn enhanced reasoning skills from medical textbooks, 2024. URL https://arxiv.org/abs/2404.00376.

Suzanne Kleijn, Henk Pander Maat, and Ted Sanders. Cloze testing for comprehension assessment: The hytec-cloze. *Language Testing*, 36(4): 553–572, 2019. doi: 10.1177/0265532219840382. URL https://doi.org/10.1177/0265532219840382.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores, 2016. URL https://arxiv.org/abs/1609.05807.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*,

pages 79–86, Phuket, Thailand, September 13-15 2005. URL https://aclanthology.org/2005.mtsummit-papers.11.

Philipp Koehn and Christof Monz. Manual and automatic evaluation of machine translation between European languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City, June 2006. Association for Computational Linguistics. URL https://aclanthology.org/W06-3114.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts, 2020.

Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. Investigating multilingual NMT representations at scale. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1565–1575, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1167. URL https://aclanthology.org/D19-1167.

Deepak Kumar, Yousef Anees AbuHashem, and Zakir Durumeric. Watch your language: Investigating content moderation with large language models. *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1):865–878, May 2024. doi: 10.1609/icwsm.v18i1.31358. URL https://ojs.aaai.org/index.php/ICWSM/article/view/31358.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3823. URL https://aclanthology.org/W19-3823.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. BioMistral: A collection of open-source pretrained large language models for medical domains. In Lun-Wei

Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics ACL 2024*, pages 5848–5864, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-acl.348.

Sotiris Lamprinidis, Federico Bianchi, Daniel Hardt, and Dirk Hovy. Universal joy a data set and results for classifying emotions across languages. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 62–75, Online, April 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wassa-1.7.

Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. What do we want from explainable artificial intelligence (xai)? - A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artif. Intell.*, 296:103473, 2021. doi: 10.1016/j.artint.2021.103473. URL https://doi.org/10.1016/j.artint.2021.103473.

Brian Larson. Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain, April 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-1601. URL https://aclanthology.org/W17-1601.

Anne Lauscher, Tobias Lueken, and Goran Glavaš. Sustainable modular debiasing of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4782–4797, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.411. URL https://aclanthology.org/2021.findings-emnlp.411.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. ISSN 1476-4687. doi: 10.1038/nature14539. URL https://doi.org/10.1038/nature14539.

Seungho Lee, Minhyun Lee, Jongwuk Lee, and Hyunjung Shim. Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5491–5501, 2021.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

B. Li and S. Gilbert. Artificial intelligence awarded two nobel prizes for innovations that will shape the future of medicine. *npj Digital Medicine*, 7:336, 2024. doi: 10.1038/s41746-024-01345-9. URL https://doi.org/10.1038/s41746-024-01345-9.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021a. URL https://openreview.net/forum?id=OJLaKwiXSbx.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2023. URL https://arxiv.org/abs/2301.12597.

Mike Li, Hongseok Namkoong, and Shangzhou Xia. Evaluating model performance under worst-case subpopulations. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 17325–17334, Vancouver, CA, 2021b. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper/2021/file/908075ea2c025c335f4865f7db427062-Paper.pdf.

Wen Li and Markus Dickinson. Gender prediction for chinese social media data. In *Proceedings of Recent Advances in Natural Language Processing*, 2017.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

Marina Litvak, Natalia Vanetik, Chaya Liebeskind, Omar Hmdia, and Rizek Abu Madeghem. Offensive language detection in Hebrew: can other languages help? In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3715–3723, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.396.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. Self-alignment pretraining for biomedical entity representations. In *North American Chapter of the Association for Computational Linguistics*, 2020a.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.818. URL https://aclanthology.org/2021.emnlp-main.818.

Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. Does gender matter? towards fairness in dialogue systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4403–4416, Barcelona, Spain (Online), December 2020b. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.390. URL https://aclanthology.org/2020.coling-main.390.

Jiexi Liu, Dehan Kong, Longtao Huang, Dinghui Mao, and Hui Xue. Multiple instance learning for offensive language detection. In *Findings of the As-*

sociation for Computational Linguistics: EMNLP 2022, pages 7387–7396, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.546. URL https://aclanthology.org/2022.findings-emnlp.546.

Muxuan Liu and Ichiro Kobayashi. Construction and validation of a Japanese honorific corpus based on systemic functional linguistics. In *Proceedings of the Workshop on Dataset Creation for Lower-Resourced Languages within the 13th Language Resources and Evaluation Conference*, pages 19–26, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.dclrl-1.3.

Yang Liu, Yuanshun Yao, Jean-François Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hanguang Li. Trustworthy LLMs: a survey and guideline for evaluating large language models' alignment. *ArXiv*, abs/2308.05374, 2023. URL https://api.semanticscholar.org/CorpusID:260775522.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*, 2019.

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. Can large language models reason about medical questions? *Patterns*, 5(3):100943, 2024. ISSN 2666-3899. doi: https://doi.org/10.1016/j.patter.2024.100943. URL https://www.sciencedirect.com/science/article/pii/S2666389924000424.

Wenpeng Lu, Guobiao Zhang, Xueping Peng, Hongjiao Guan, and Shoujin Wang. Medical entity disambiguation with medical mention relation and fine-grained entity knowledge. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11148–11158, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.972.

Brandon Lwowski, Paul Rad, and Anthony Rios. Measuring geographic performance disparities of offensive language classifiers. In *Proceedings of*

the 29th International Conference on Computational Linguistics, pages 6600–6616, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.574.

D. Lyell and E. Coiera. Automation bias and verification complexity: a systematic review. *Journal of the American Medical Informatics Association*, 24(2):423–431, Mar 2017. doi: 10.1093/jamia/ocw105.

Subha Maity, Debarghya Mukherjee, Mikhail Yurochkin, and Yuekai Sun. Does enforcing fairness mitigate biases caused by subpopulation shift?, 2020. URL https://arxiv.org/abs/2011.03173.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence 35(17)*, pages 14867–14875, 2021.

Costas Mavromatis and George Karypis. Gnn-rag: Graph neural retrieval for large language model reasoning, 2024. URL https://arxiv.org/abs/2405.20139.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1063. URL https://aclanthology.org/N19-1063.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.173. URL https://aclanthology.org/2020.acl-main.173.

Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computa-*

tional Linguistics, pages 3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1334. URL https://aclanthology.org/P19-1334.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), jul 2021. ISSN 0360-0300. doi: 10.1145/3457607. URL https://doi.org/10.1145/3457607.

Thomas Miconi. The impossibility of "fairness": a generalized impossibility result for decisions. *arXiv: Applications*, 2017. doi: 10.48550/ARXIV.1707.01195.

Tom Michael Mitchell. The need for biases in learning generalizations. In *Rutgers CS tech report CBM-TR-117*, 1980.

Sunil Mohan and Donghui Li. MedMentions: A large biomedical corpus annotated with umls concepts, 2019. URL https://arxiv.org/abs/1902.09476.

Ron Mokady, Amir Hertz, and Amit H Bermano. ClipCap: CLIP prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. URL https://arxiv.org/abs/2111.09734.

George D. Montañez, Jonathan Hayase, Julius Lauw, Dominique Macias, Akshay Trikha, and Julia Vendemiatti. The futility of bias-free learning and search. In Jixue Liu and James Bailey, editors, *AI 2019: Advances in Artificial Intelligence*, pages 277–288, Cham, 2019. Springer International Publishing. ISBN 978-3-030-35288-2.

Michael Moor, Oishi Banerjee, Zein S.H. Abad, et al. Foundation models for generalist medical artificial intelligence. *Nature*, 616:259–265, 2023. doi: 10.1038/s41586-023-05881-4.

Antonio Alexander Morgan-Lopez, Annice E Kim, Robert F. Chew, and Paul Ruddle. Predicting age groups of twitter users based on language and metadata features. *PLoS ONE*, 12, 2017.

Dennis C. Mueller, Robert D. Tollison, and Thomas D. Willett. The Utilitarian Contract: A Generalization of Rawls' Theory of Justice. *Theory and Decision*, 4(3-4):345–367, 1974. doi: 10.1007/bf00136654.

Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.416. URL https://aclanthology.org/2021.acl-long.416.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models, 2024. URL https://arxiv.org/abs/2307.06435.

Roberto Navigli, Simone Conia, and Björn Ross. Biases in Large Language Models: Origins, Inventory, and Discussion. *J. Data and Information Quality*, 15(2), June 2023. ISSN 1936-1955. doi: 10.1145/3597307. URL https://doi.org/10.1145/3597307.

Lleayem Nazario-Johnson, Hossam A. Zaki, and Glenn A. Tung. Use of large language models to predict neuroimaging. *Journal of the American College of Radiology*, 20(10):1004–1009, 2023. ISSN 1546-1440. doi: https://doi.org/10.1016/j.jacr.2023.06.008. URL https://www.sciencedirect.com/science/article/pii/S1546144023004830.

Robert Osazuwa Ness, Katie Matton, Hayden Helm, Sheng Zhang, Junaid Bajwa, Carey E. Priebe, and Eric Horvitz. Medfuzz: Exploring the robustness of large language models in medical question answering, 2024. URL https://arxiv.org/abs/2406.06573.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5034. URL https://aclanthology.org/W19-5034.

Nobal B. Niraula, Saurab Dulal, and Diwa Koirala. Offensive language detection in Nepali social media. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 67–75, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.woah-1.7. URL https://aclanthology.org/2021.woah-1.7.

Malvina Nissim, Viviana Patti, Barbara Plank, and Esin Durmus, editors. *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.peoples-1.0.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 145–153, Republic and Canton of Geneva, CHE, 2016. International World Wide Web Conferences Steering Committee. ISBN 9781450341431. doi: 10.1145/2872427.2883062. URL https://doi.org/10.1145/2872427.2883062.

Nobel Prize Outreach AB. Press release: The nobel prize in chemistry 2024. https://www.nobelprize.org/prizes/chemistry/2024/press-release/, 2024a. URL https://www.nobelprize.org/prizes/chemistry/2024/press-release/. Accessed: 2024-11-27.

Nobel Prize Outreach AB. Press release: The nobel prize in physics 2024. https://www.nobelprize.org/prizes/physics/2024/press-release/, 2024b. URL https://www.nobelprize.org/prizes/physics/2024/press-release/. Accessed: 2024-11-27.

Mike Noon. The shackled runner: time to rethink positive discrimination? *Work, Employment and Society*, 24(4):728–739, 2010. doi: 10.1177/0950017010380648. URL https://doi.org/10.1177/0950017010380648.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. Can generalist foundation models outcompete special-purpose tuning? case study in medicine, 2023. URL https://arxiv.org/abs/2311.16452.

Debora Nozza. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914, Online, August 2021. Association for Computational Linguistics.

doi: 10.18653/v1/2021.acl-short.114. URL https://aclanthology.org/2021.acl-short.114.

Luca Oneto and Silvia Chiappa. Fairness in machine learning. In *Recent Trends in Learning From Data*, pages 155–196. Springer, 2020.

OpenAI. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Hadas Orgad, Seraphina Goldfarb-Tarrant, and Yonatan Belinkov. How gender debiasing affects internal model representations, and why it matters. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2602–2628, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.188. URL https://aclanthology.org/2022.naacl-main.188.

Andrés Páez. The pragmatic turn in explainable artificial intelligence (xai). *Minds and Machines*, 29(3):441–459, 2019.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann, editors, *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR, 07–08 Apr 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

Stefan Palan and Christian Schitter. Prolific.ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27, 2017.

Jeff Z. Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini, Russa Biswas, Gerard de Melo, Angela Bonifati, Edlira Vakaj, Mauro Dragoni, and Damien Graux. Large Language Models and Knowledge Graphs: Opportunities and Challenges. *Transactions on Graph Data and Knowledge*, 1(1):2:1–2:38, 2023. doi: 10.4230/TGDK.1.1.2. URL https://drops.dagstuhl.de/entities/document/10.4230/TGDK.1.1.2.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599, 2024. doi: 10.1109/TKDE.2024.3352100.

Ji Ho Park, Jamin Shin, and Pascale Fung. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1302. URL https://aclanthology.org/D18-1302.

Maria Pavesi and Pablo Zamora. The reception of swearing in film dubbing: a cross-cultural case study. *Perspectives*, 30(3):382–398, 2022. doi: 10.1080/0907676X.2021.1913199. URL https://doi.org/10.1080/0907676X.2021.1913199.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions, 2023. URL https://arxiv.org/abs/2308.11483.

Steven T. Piantadosi and Felix Hill. Meaning without reference in large language models, 2022. URL https://arxiv.org/abs/2208.02957.

Laura Cabello Piqueras, Constanza Fierro, Jonas F. Lotz, Phillip Rust, Joen Rommedahl, Jeppe Klok Due, Christian Igel, Desmond Elliott, Carsten B. Pedersen, Israfel Salazar, and Anders Søgaard. Date recognition in historical parish records. In *Frontiers in Handwriting Recognition*, pages 49–64, Cham, 2022. Springer International Publishing. ISBN 978-3-031-21648-0. URL https://link.springer.com/chapter/10.1007/978-3-031-21648-0_4.

Barbara Plank. The 'problem' of human label variation: On ground truth in data, modeling and evaluation. *ArXiv*, abs/2211.02570, 2022.

Barbara Plank, Dirk Hovy, and Anders Søgaard. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. doi: 10.3115/v1/E14-1078. URL https://aclanthology.org/E14-1078.

Moritz Plenz and Anette Frank. Graph language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4477–4494, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.245.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2641–2649, 2015. doi: 10.1109/ICCV.2015.303.

Christopher Potts, Zhengxuan Wu, Atticus Geiger, and Douwe Kiela. DynaSent: A dynamic benchmark for sentiment analysis. *arXiv preprint arXiv:2012.15349*, 2020. URL https://arxiv.org/abs/2012.15349.

Rebecca Qian, Candace Ross, Jude Fernandes, Eric Smith, Douwe Kiela, and Adina Williams. Perturbation augmentation for fairer nlp, 2022. URL https://arxiv.org/abs/2205.12586.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), January 2020. ISSN 1532-4435.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 2019 Conference of the Association for Computational Linguistics (ACL2019)*, 2019. URL https://arxiv.org/abs/1906.02361.

Taraka Rama, Lisa Beinborn, and Steffen Eger. Probing multilingual BERT for genetic and typological signals. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1214–1228, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.105. URL https://aclanthology.org/2020.coling-main.105.

Tharindu Ranasinghe and Marcos Zampieri. Multilingual offensive language identification for low-resource languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(1), nov 2021. ISSN 2375-4699. doi: 10.1145/3457610. URL https://doi.org/10.1145/3457610.

Vinit Ravishankar, Artur Kulmizev, Mostafa Abdou, Anders Søgaard, and Joakim Nivre. Attention can reflect syntactic structure (if you let it). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3031–3045, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.264. URL https://aclanthology.org/2021.eacl-main.264.

John Rawls. *A Theory of Justice*. Belknap Press of Harvard University Press, Cambridge, Massachussets, 1 edition, 1971. ISBN 0-674-88014-5.

John Rawls. *Justice as Fairness: A Restatement*. Harvard University Press, 2001. ISBN 9780674005105. URL http://www.jstor.org/stable/j.ctv31xf5v0.

Sagnik Ray Choudhury, Anna Rogers, and Isabelle Augenstein. Machine reading, fast and slow: When do models "understand" language? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 78–93, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.8.

Charan Reddy, Deepak Sharma, Soroush Mehri, Adriana Romero Soriano, Samira Shabanian, and Sina Honari. Benchmarking bias mitigation algorithms in representation learning through fairness metrics. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1,

2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/2723d092b63885e0d7c260cc007e8b9d-Paper-round1.pdf.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410.

Anka Reuel, Ben Bucknall, Stephen Casper, Tim Fist, Lisa Soder, Onni Aarne, Lewis Hammond, Lujain Ibrahim, Alan Chan, Peter Wills, Markus Anderljung, Ben Garfinkel, Lennart Heim, Andrew Trask, Gabriel Mukobi, Rylan Schaeffer, Mauricio Baker, Sara Hooker, Irene Solaiman, Alexandra Sasha Luccioni, Nitarshan Rajkumar, Nicolas Moës, Jeffrey Ladish, Neel Guha, Jessica Newman, Yoshua Bengio, Tobin South, Alex Pentland, Sanmi Koyejo, Mykel J. Kochenderfer, and Robert Trager. Open problems in technical ai governance, 2024. URL https://arxiv.org/abs/2407.14981.

Dor Ringel, Gal Lavee, Ido Guy, and Kira Radinsky. Cross-cultural transfer learning for text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3873–3883, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1400. URL https://aclanthology.org/D19-1400.

Petar Ristoski and Heiko Paulheim. Rdf2vec: Rdf graph embeddings for data mining. In Paul Groth, Elena Simperl, Alasdair Gray, Marta Sabou, Markus Krötzsch, Freddy Lecue, Fabian Flöck, and Yolanda Gil, editors, *The Semantic Web – ISWC 2016*, pages 498–514, Cham, 2016. Springer International Publishing.

Katherine Ritchie. Social identity, indexicality, and the appropriation of slurs. *Croatian Journal of Philosophy*, 17(2):155–180, 2017.

Ingrid Robeyns. Justice as fairness and the capability approach. *Arguments for a Better World. Essays for Amartya Sen's*, 75:397–413, 2009.

Alexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Tauman Kalai. What's in a name? reducing bias in bios without access to protected attributes, 2019. URL https://arxiv.org/abs/1904.05233.

Candace Ross, Boris Katz, and Andrei Barbu. Measuring social biases in grounded vision and language embeddings. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 998–1008, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.78. URL https://aclanthology.org/2021.naacl-main.78.

Paul Röttger, Debora Nozza, Federico Bianchi, and Dirk Hovy. Data-efficient strategies for expanding hate speech detection into under-resourced languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5674–5691, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.383. URL https://aclanthology.org/2022.emnlp-main.383.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. Square one bias in NLP: Towards a multi-dimensional exploration of the research manifold. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2340–2354, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.184. URL https://aclanthology.org/2022.findings-acl.184.

Jonathan Rusert, Zubair Shafiq, and Padmini Srinivasan. On the robustness of offensive language classifiers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7424–7438, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.513. URL https://aclanthology.org/2022.acl-long.513.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambrano Chaves, Szu-Yeu Hu, Mike Schaekermann, Aishwarya Kamath, Yong Cheng, David G. T. Barrett, Cathy Cheung, Basil Mustafa,

Anil Palepu, Daniel McDuff, Le Hou, Tomer Golany, Luyang Liu, Jean baptiste Alayrac, Neil Houlsby, Nenad Tomasev, Jan Freyberg, Charles Lau, Jonas Kemp, Jeremy Lai, Shekoofeh Azizi, Kimberly Kanada, SiWai Man, Kavita Kulkarni, Ruoxi Sun, Siamak Shakeri, Luheng He, Ben Caine, Albert Webson, Natasha Latysheva, Melvin Johnson, Philip Mansfield, Jian Lu, Ehud Rivlin, Jesper Anderson, Bradley Green, Renee Wong, Jonathan Krause, Jonathon Shlens, Ewa Dominowska, S. M. Ali Eslami, Katherine Chou, Claire Cui, Oriol Vinyals, Koray Kavukcuoglu, James Manyika, Jeff Dean, Demis Hassabis, Yossi Matias, Dale Webster, Joelle Barral, Greg Corrado, Christopher Semturs, S. Sara Mahdavi, Juraj Gottweis, Alan Karthikesalingam, and Vivek Natarajan. Capabilities of gemini models in medicine, 2024. URL https://arxiv.org/abs/2404.18416.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1163. URL https://aclanthology.org/P19-1163.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.431. URL https://aclanthology.org/2022.naacl-main.431.

Bhaskarjit Sarmah, Benika Hall, Rohan Rao, Sunil Patel, Stefano Pasquali, and Dhagash Mehta. Hybridrag: Integrating knowledge graphs and vector retrieval augmented generation for efficient information extraction, 2024. URL https://arxiv.org/abs/2408.04948.

Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain, April 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-1101. URL https://aclanthology.org/W17-1101.

John Schmitz. On the native/nonnative speaker notion and world englishes: Debating with k. rajagopalan. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, 32:597–611, December 2016. doi: 10.1590/0102-445083626175745488.

Peter H Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A. Rothkopf, and Kristian Kersting. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4:258–268, 2022. doi: 10.1038/s42256-022-00458-8. URL https://doi.org/10.1038/s42256-022-00458-8.

João Sedoc and Lyle Ungar. The role of protected class word lists in bias identification of contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 55–61, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3808. URL https://aclanthology.org/W19-3808.

Mattia Setzu, Riccardo Guidotti, Anna Monreale, Franco Turini, Dino Pedreschi, and Fosca Giannotti. Glocalx - from local to global explanations of black box ai models. *Artificial Intelligence*, 294:103457, 2021. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2021.103457. URL https://www.sciencedirect.com/science/article/pii/S0004370221000084.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.468. URL https://aclanthology.org/2020.acl-main.468.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 1948. URL http://plan9.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf.

Amit Sharma, Arpita Biswas, and Siddharth Barman. Inframarginality audit of group-fairness. Symposium on the Foundations of Responsible

Computing (FORC), June 2020. URL https://www.microsoft.com/en-us/research/publication/inframarginality-audit-of-group-fairness/.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018.

Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. Does representational fairness imply empirical fairness? In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 81–95, Online only, November 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-aacl.8.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. Societal biases in language generation: Progress and challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4275–4293, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.330. URL https://aclanthology.org/2021.acl-long.330.

Xiayang Shi, Xinyi Liu, Chun Xu, Yuanyuan Huang, Fang Chen, and Shaolin Zhu. Cross-lingual offensive speech identification with transfer learning for low-resource languages. *Computers and Electrical Engineering*, 101:108005, 2022. ISSN 0045-7906. doi: https://doi.org/10.1016/j.compeleceng.2022.108005. URL https://www.sciencedirect.com/science/article/pii/S0045790622002725.

Sunny Shrestha and Sanchari Das. Exploring gender biases in ml and ai academic research through systematic literature review. *Frontiers in Artificial Intelligence*, 5, 2022. ISSN 2624-8212. doi: 10.3389/frai.2022.976838. URL https://www.frontiersin.org/articles/10.3389/frai.2022.976838.

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. Offensive language and hate speech detection for Danish. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3498–3508, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.430.

Jasdeep Singh, Bryan McCann, Richard Socher, and Caiming Xiong. BERT is not an interlingua and the bias of tokenization. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 47–55, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-6106. URL https://aclanthology.org/D19-6106.

Krishna Kumar Singh, Dhruv Kumar Mahajan, Kristen Grauman, Yong Jae Lee, Matt Feiszli, and Deepti Ghadiyaram. Don't judge an object by its context: Learning to overcome contextual bias. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11067–11075, 2020.

Karan Singhal, Shekoofeh Azizi, Tao Tu, et al. Large language models encode clinical knowledge. *Nature*, 620(7974):172–180, 2023a. doi: 10.1038/s41586-023-06291-2. URL https://doi.org/10.1038/s41586-023-06291-2.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Towards expert-level medical question answering with large language models, 2023b. URL https://arxiv.org/abs/2305.09617.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/6b180037abbebea991d8b1232f8a8ca9-Paper.pdf.

Sanghoun Song. Representing honorifics via individual constraints. In *Proceedings of the Grammar Engineering Across Frameworks (GEAF) 2015 Workshop*, pages 57–64, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3308. URL https://aclanthology.org/W15-3308.

Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 100(3/4):441–471, 1987. ISSN 00029556. URL http://www.jstor.org/stable/1422689.

Ami D. Sperber, Robert F. Devellis, and Brian Boehlecke. Cross-cultural translation: Methodology and validation. *Journal of Cross-Cultural Psychology*, 25(4):501–524, 1994. doi: 10.1177/0022022194254006. URL https://doi.org/10.1177/0022022194254006.

Tejas Srinivasan and Yonatan Bisk. Worst of both worlds: Biases compound in pre-trained vision-and-language models, 2021. URL https://arxiv.org/abs/2104.08666.

Karolina Stanczak and Isabelle Augenstein. A survey on gender bias in natural language processing, 2021. URL https://arxiv.org/abs/2112.14168.

K. E. Stanovich and R. F. West. Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, 23: 645–665, 2000.

Lukas Stappen, Fabian Brunn, and Björn Schuller. Cross-lingual zero- and few-shot hate speech detection utilising frozen transformer language models and axel. *ArXiv*, abs/2004.13850, 2020.

Louise Stringer and Paul Iverson. Non-native speech recognition sentences: A new materials set for non-native speech perception research. *Behavior Research Methods*, 52(2), 2020.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1355. URL https://aclanthology.org/P19-1355.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1644. URL https://aclanthology.org/P19-1644.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1159. URL https://aclanthology.org/P19-1159.

Tony Sun, Kellie Webster, Apu Shah, William Yang Wang, and Melvin Johnson. They, them, theirs: Rewriting with gender-neutral english, 2021.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA, 2014. MIT Press.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421.

Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. Evaluating the factual consistency of large language models through news summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5220–5255, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.322. URL https://aclanthology.org/2023.findings-acl.322.

Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1514. URL https://aclanthology.org/D19-1514.

Sarah Tan, Rich Caruana, Giles Hooker, and Yin Lou. Distill-and-compare: Auditing black-box models using transparent model distillation. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 303–310, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360128. doi: 10.1145/3278721.3278725. URL https://doi.org/10.1145/3278721.3278725.

Yi Chern Tan and L. Elisa Celis. Assessing social and intersectional biases in contextualized word representations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/201d546992726352471cfea6b0df0a48-Paper.pdf.

Jiabin Tang, Yuhao Yang, Wei Wei, Lei Shi, Lixin Su, Suqi Cheng, Dawei Yin, and Chao Huang. GraphGPT: Graph instruction tuning for large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 491–500, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657775. URL https://doi.org/10.1145/3626772.3657775.

Ruixiang Tang, Mengnan Du, Yuening Li, Zirui Liu, Na Zou, and Xia Hu. Mitigating gender bias in captioning systems. In *Proceedings of the Web Conference 2021*, WWW '21, page 633–645, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383127. doi: 10.1145/3442381.3449950. URL https://doi.org/10.1145/3442381.3449950.

Wilson L. Taylor. Cloze procedure: A new tool for measuring readability. *Journalism & Mass Communication Quarterly*, 30:415 – 433, 1953.

Terne Sasha Thorn Jakobsen, Laura Cabello, and Anders Søgaard. Being right for whose right reasons? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1033–1054, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.59. URL https://aclanthology.org/2023.acl-long.59.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models, 2023a. URL https://arxiv.org/abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023b. URL https://arxiv.org/abs/2307.09288.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David Fleet, Philip Mansfield, Sushant Prakash, Renee Wong, Sunny Virmani, Christopher Semturs, S Sara Mahdavi, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Karan Singhal, Pete Florence, Alan Karthikesalingam,

and Vivek Natarajan. Towards generalist biomedical ai, 2023. URL https://arxiv.org/abs/2307.14334.

Alan Turing. Computing machinery and intelligence. *Mind*, 59(October): 433–60, 1950. doi: 10.1093/mind/lix.236.433.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. URL https://arxiv.org/abs/1807.03748.

Emiel van Miltenburg. Stereotyping and bias in the flickr30k dataset. In Jens Edlund, Dirk Heylen, and Patrizia Paggio, editors, *Proceedings of Multimodal Corpora*, pages 1–4, 2016.

Dave Van Veen, Cara Van Uden, Maayane Attias, Anuj Pareek, Christian Bluethgen, Malgorzata Polacin, Wah Chiu, Jean-Benoit Delbrouck, Juan Zambrano Chaves, Curtis Langlotz, Akshay Chaudhari, and John Pauly. RadAdapt: Radiology report summarization via lightweight domain adaptation of large language models. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 449–460, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.bionlp-1.42. URL https://aclanthology.org/2023.bionlp-1.42.

Eva Vanmassenhove, Chris Emmery, and Dimitar Shterionov. NeuTral Rewriter: A rule-based and neural approach to automatic rewriting into gender neutral alternatives. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8940–8948, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.704. URL https://aclanthology.org/2021.emnlp-main.704.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017a. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc., 2017b. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18, page 1–7, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450357463. doi: 10.1145/3194770.3194776. URL https://doi.org/10.1145/3194770.3194776.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3509. URL https://aclanthology.org/W19-3509.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401, Vancouver, CA, 2020. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Alina Vretinaris, Chuan Lei, Vasilis Efthymiou, Xiao Qin, and Fatma Özcan. Medical entity disambiguation using graph neural networks. In *Proceedings of the 2021 International Conference on Management of Data*, SIGMOD '21, page 2310–2318, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383431. doi: 10.1145/3448016.3457328. URL https://doi.org/10.1145/3448016.3457328.

Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, Roi Reichart, and Anna Korhonen. Multi-SimLex: A Large-Scale Evaluation of

Multilingual and Crosslingual Lexical Semantic Similarity. *Computational Linguistics*, 46(4):847–897, 02 2020. ISSN 0891-2017. doi: 10.1162/coli_a_00391. URL https://doi.org/10.1162/coli_a_00391.

Ada Wan. Fairness in representation for multilingual NLP: Insights from controlled experiments on conditional language modeling. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=-llS6TiOew.

Angelina Wang and Olga Russakovsky. Directional bias amplification. In *ICML*, 2021.

Angelina Wang, Alexander Liu, Ryan Zhang, Anat Kleiman, Leslie Kim, Dora Zhao, Iroha Shirai, Arvind Narayanan, and Olga Russakovsky. Revise: A tool for measuring and mitigating bias in visual datasets. *Int. J. Comput. Vision*, 130(7):1790–1810, jul 2022a. ISSN 0920-5691. doi: 10.1007/s11263-022-01625-5. URL https://doi.org/10.1007/s11263-022-01625-5.

Jialu Wang, Yang Liu, and Xin Wang. Assessing multilingual fairness in pre-trained multimodal representations. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2681–2695, Dublin, Ireland, May 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.211. URL https://aclanthology.org/2022.findings-acl.211.

Lijie Wang, Yaozong Shen, Shu ping Peng, Shuai Zhang, Xinyan Xiao, Hao Liu, Hongxuan Tang, Ying Chen, Hua Wu, and Haifeng Wang. A fine-grained interpretability evaluation benchmark for neural nlp. *ArXiv*, abs/2205.11097, 2022c.

Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *ICCV*, 2019.

Xinpeng Wang, Chengzhi Hu, Bolei Ma, Paul Rottger, and Barbara Plank. Look at the text: Instruction-tuned language models are more robust multiple choice selectors than you think. In *First Conference on Language Modeling*, 2024a. URL https://openreview.net/forum?id=qHdSA85GyZ.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. "My Answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics ACL 2024*, pages 7407–7416, Bangkok, Thailand and virtual meeting, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.441. URL https://aclanthology.org/2024.findings-acl.441.

Jason Wei and Eugene Santos Jr. Narrative origin classification of Israeli-Palestinian conflict texts. In *The Thirty-Third International FLAIRS Conference*, 2020.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NeurIPS '22, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9781713871088.

Adrienne Williams, Milagros Miceli, and Timnit Gebru. The exploited labor behind artificial intelligence. *Noema*, 2022. URL https://www.noemamag.com/the-exploited-labor-behind-artificial-intelligence/.

Robert Williamson and Aditya Menon. Fairness risk measures. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6786–6797. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/williamson19a.html.

T. Winchcomb. Use of ai in online content moderation. Technical report, Ofcom, July 2019. Last updated: 16 March 2023.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. PMC-LLaMA: Towards building open-source language models for medicine, 2023. URL https://arxiv.org/abs/2304.14454.

Shijie Wu and Mark Dredze. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online, July 2020. Association for

Computational Linguistics. doi: 10.18653/v1/2020.repl4nlp-1.16. URL https://aclanthology.org/2020.repl4nlp-1.16.

Guangxuan Xu and Qingyuan Hu. Can model compression improve nlp fairness, 2022. URL https://arxiv.org/abs/2201.08542.

Ruochen Xu, Yiming Yang, Naoki Otani, and Yuexin Wu. Unsupervised cross-lingual transfer of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2465–2474, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1268. URL https://aclanthology.org/D18-1268.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6575.

Junhan Yang, Zheng Liu, Shitao Xiao, Chaozhuo Li, Defu Lian, Sanjay Agrawal, Amit S, Guangzhong Sun, and Xing Xie. Graphformers: GNN-nested transformers for representation learning on textual graph. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=yILzFBjR0Y.

Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D Manning, Percy S Liang, and Jure Leskovec. Deep bidirectional language-knowledge graph pretraining. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*,

volume 35, pages 37309–37323. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/f224f056694bcfe465c5d84579785761-Paper-Conference.pdf.

Omar Zaidan, Jason Eisner, and Christine Piatko. Using "annotator rationales" to improve machine learning for text categorization. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 260–267, Rochester, New York, April 2007. Association for Computational Linguistics. URL https://aclanthology.org/N07-1033.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2010. URL https://aclanthology.org/S19-2010.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online), December 2020. International Committee for Computational Linguistics. doi: 10.18653/v1/2020.semeval-1.188. URL https://aclanthology.org/2020.semeval-1.188.

Carlos Zednik and Hannes Boelsen. Scientific exploration and explainable artificial intelligence. *Minds Mach.*, 32(1):219–239, mar 2022.

Sheng Zhang, Xin Zhang, Weiming Zhang, and Anders Søgaard. Sociolectal analysis of pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4581–4588, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.375. URL https://aclanthology.org/2021.emnlp-main.375.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. ERNIE: Enhanced language representation with informative enti-

ties. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1139. URL https://aclanthology.org/P19-1139.

Dora Zhao, Angelina Wang, and Olga Russakovsky. Understanding and evaluating racial biases in image captioning. In *International Conference on Computer Vision (ICCV)*, 2021.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark, September 2017a. Association for Computational Linguistics. doi: 10.18653/v1/D17-1323. URL https://aclanthology.org/D17-1323.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints, 2017b.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1521. URL https://aclanthology.org/D18-1521.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1064. URL https://aclanthology.org/N19-1064.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2896–2907, Online, July

2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.260. URL https://aclanthology.org/2020.acl-main.260.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=shr9PXz7T0.

Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023a. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581318. URL https://doi.org/10.1145/3544548.3581318.

Kankan Zhou, Eason Lai, and Jing Jiang. VLStereoSet: A study of stereotypical bias in pre-trained vision-language models. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 527–538, Online only, November 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.aacl-main.40.

Li Zhou, Laura Cabello, Yong Cao, and Daniel Hershcovich. Cross-cultural transfer learning for Chinese offensive language detection. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 8–15, Dubrovnik, Croatia, May 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.c3nlp-1.2. URL https://aclanthology.org/2023.c3nlp-1.2.

Hongyin Zhu, Hao Peng, Zhiheng Lyu, Lei Hou, Juanzi Li, and Jinghui Xiao. Pre-training language model incorporating domain-specific heterogeneous knowledge into a unified representation. *Expert Systems with Applications*, 215:119369, 2023. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2022.119369. URL https://www.sciencedirect.com/science/article/pii/S0957417422023879.

Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2:

More deformable, better results. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9300–9308, 2018.