UNIVERSITY OF COPENHAGEN
DEPARTMENT OF COMPUTER SCIENCE

Ph.D. Thesis

Ruixiang Cui

Evaluating Language Models: From Multilingual Compositional Semantic Parsing to Human-level Complex Reasoning

Advisor: Daniel Hershcovich, Anders Søgaard

Assessment Committee: Desmond Elliott, Ellie Pavlick, Adina Williams

This thesis has been submitted to the Ph.D. School of The Faculty of Science, University of Copenhagen on July 31st, 2024.

Eu tô te explicando
Pra te confundir
Eu tô te confundindo
Pra te esclarecer
Tô iluminado
Pra poder cegar
Tô ficando cego
Pra poder guiar
Suavemente pra poder rasgar
Olho fechado pra te ver melhor
Com alegria pra poder chorar
Desesperado pra ter paciência
Carinhoso pra poder ferir
Lentamente pra não atrasar
Atrás da vida pra poder morrer
Eu tô me despedindo pra poder voltar
- *Tom Zé, Tô*

Abstract

The pursuit of creating machines that can think and communicate like humans has led to the birth of artificial intelligence. Natural language processing is at the heart of this effort, a subfield concerning developing models that can process and generate human language. Recent advances in large language models (LLMs), trained on massive text corpora with neural architectures, have led to remarkable conversational capabilities. However, as these models grow in size and complexity, evaluating their ability to generalize across a wide range of languages, tasks, and linguistic phenomena becomes increasingly essential. While classic language model benchmarks have helped compare and develop models, many of them have already reached a point of saturation, with near-perfect scores achieved. This thesis explores various aspects of language model evaluation, reflecting the evolving landscape of model development and evaluation in the era of LLMs. It investigates the strengths and weaknesses of current language models, particularly their ability to generalize compositionally, a fundamental aspect of human language that allows for creating novel expressions by combining known building blocks.

The main contributions of this thesis are as follows: (1) We develop a method to migrate a question-answering dataset from one knowledge base to another and extend it to diverse languages and domains. The resulting benchmark, Multilingual Compositional Wikidata Questions (MCWQ), reveals language models lack the ability of cross-lingual compositional generalization. (2) We identified generalized quantifiers, i.e., words like "all", "some", and "most", as a significant challenge for understanding natural language and developed a benchmark to test this specific reasoning ability. (3) We investigate how language models reason with the word "respectively" in various learning settings and demonstrate the challenge they face in generalizing to the long tails of linguistic constructions (4) We introduce AGIEval, a bilingual benchmark comprising high-standard official exams, which uncovers the limitations of LLMs in human-level reasoning tasks. By pushing the boundaries of how we evaluate LLMs, this thesis provides valuable insights into their strengths and weaknesses. Ultimately, we argue that true language understanding requires more than good performance on existing tests. It requires the ability to generalize and adapt to new challenges, just like humans do.

Resumé

Jagten på at skabe maskiner, der kan tænke og kommunikere som mennesker, har ført til fødslen af kunstig intelligens. Naturlig sprogbehandling er kernen i denne indsats, et underfelt, der beskæftiger sig med at udvikle modeller, der kan behandle og generere menneskeligt sprog. Nylige fremskridt inden for store sprogmodeller (LLMs), trænet på massive tekstkorpora med neurale arkitekturer, har ført til bemærkelsesværdige samtaleevner. Men efterhånden som disse modeller vokser i størrelse og kompleksitet, bliver det stadig vigtigere at evaluere deres evne til at generalisere på tværs af en bred vifte af sprog, opgaver og sproglige fænomener. Mens klassiske sprogmodel benchmarks har hjulpet med at sammenligne og udvikle modeller, har mange af dem allerede nået et mætningspunkt, hvor næsten perfekte scores er opnået. Denne afhandling udforsker forskellige aspekter af sprogmodel evaluering, der afspejler det udviklende landskab af modeludvikling og evaluering i LLM-æraen. Den undersøger styrkerne og begrænsningerne ved nuværende sprogmodeller, især deres evne til at generalisere kompositionelt. Det er et grundlæggende aspekt af menneskeligt sprog, der muliggør skabelsen af nye udtryk ved at kombinere kendte byggesten.

De vigtigste bidrag fra denne afhandling er som følger: (1) Vi udvikler en metode til at migrere et spørgsmål-svar dataset fra en vidensbase til en anden og udvider det til forskellige sprog og domæner. Den resulterende benchmark, Multilingual Compositional Wikidata Questions (MCWQ), afslører begrænsningerne af sprogmodeller i tvær-sproglig kompositionel generalisering. (2) Vi identificerede generaliserede kvantifikatorer, dvs. ord som "alle", "nogle" og "de fleste", som en betydelig udfordring for forståelsen af naturligt sprog og udviklede en benchmark til at teste denne specifikke ræsonneringsevne. (3) Vi undersøger, hvordan sprogmodeller ræsonnerer med ordet "henholdsvis" i forskellige læringsindstillinger og demonstrerer den udfordring, de står over for i at generalisere til de lange haler af sproglige konstruktioner. (4) Vi introducerer AGIEval, en tosproget benchmark bestående af højtstående officielle eksamener, som afslører begrænsningerne af LLMs i menneskelige ræsonneringsopgaver. Ved at skubbe grænserne for, hvordan vi evaluerer LLMs, giver denne afhandling værdifulde indsigter i deres styrker og svagheder. I sidste ende argumenterer vi for, at ægte sprogforståelse kræver mere end god præstation på eksisterende tests. Det kræver evnen til at generalisere og tilpasse sig nye udfordringer, ligesom mennesker gør.

Acknowledgements

I am deeply in awe of how far I have come. My progress would not have been half as complete without the friends and mentors I have made along the way. The four-year journey of my PhD has been a kaleidoscope of experiences – colourful, adventurous, rewarding, and ever-changing. It has been one of intellectual growth and self-discovery, teaching me to embrace the unknown and find joy in the pursuit of knowledge.

First and foremost, I must express my heartfelt gratitude to my dear colleagues and friends from the CoAStaL group at the University of Copenhagen: Mostafa Abdou, Ali Al-Laith, Rahul Aralikatte, Marcel Bollmann, Lasse Borgholt, Emanuele Bugliarello, Yong Cao, Ilias Chalkidis, Miryam de Lhoneux, Simon Flachs, Nicolas Garneau, Victor Petrén Bach Hansen, Mareike Hartmann, Guimin Hu, Yova Kementchedjhieva, Antonia Kramolegkou, Seolhwa Lee, Heather Lent, Jiaang Li, Wenyan Li, Jonas Lotz, Katerina Margatina, Lukas Nielsen, Tommaso Pasini, Qiwei Peng, Rita Ramnos, Vinit Ravishankar, Davide Rigoni, Phillip Rust, Danae Sánchez Villegas, Alice Schiavone, Monorama Swain, Anna Katrine van Zee, Yifei Yuan, Sheng Zhang, and Li Zhou. You brilliant minds have created such a cozy space to study, work, and live in—a hygge environment unlike any other. CoAStaL has been my most important home in Denmark and will forever remain so in my heart.

I am immensely grateful to my two exceptional supervisors, Daniel Hershcovich and Anders Søgaard. You are the best supervisors I could have ever asked for. Daniel, you have always been there to listen to my ideas, provide constructive feedback, and offer encouragement when I needed it most. Your kindness, generosity, and patience have been invaluable to me. And I know that the lessons I have learned from you will stay with me throughout my career. Anders, your insights and willingness to think outside the box have been a constant source of inspiration for me. I am grateful for the countless thought-provoking discussions we have had, which have shaped my research and my growth as a scientist. Thank you, Daniel and Anders.

I am also thankful to my mentors during my internship at Microsoft, Yaobo Lang and Nan Duan. You opened new doors for me in academia, especially during a pivotal time for NLP. Your support and ideas solidified my work. I would also like to thank Wanjun Zhong, Yiduo Guo, and Shang Lu for our fruitful collaborations.

I am grateful to Christopher Potts for hosting me at the Stanford NLP group. The experience was eye-opening and introduced me into the realm of

mechanistic interpretability. I also appreciate Zhengxuan Wu for hosting me at Stanford. Your hospitality and insight turned my visit into an unforgettable chapter of my academic adventure.

I couldn't have made it without my friends, despite the distance between us, wherever we are in the world. To my long-time friends, Boyan, Chang, Lujie, Mengyu, Michael, Othmane, Yue, Xue, and Walter, your presence has filled my life with light and warmth. A special thanks goes to my friends in Copenhagen, Dennis, Dustin, Eric, Karolina and Nadav, for the fun times and memories we shared together.

I am honoured to have three wonderful assessment committee members, Desmond Elliott, Ellie Pavlick, and Adina Williams. Your work has been a huge inspiration during my PhD. It is a privilege to discuss my research with you and receive your valuable feedback. Thank you for evaluating this thesis.

Finally, I want to express my deepest gratitude to my parents, Zhimin and Haiyin, for your unconditional love and support, which allowed me to chase my dream in a faraway land. Your faith in me and the values you taught have been a source of strength, helping me navigate in turbulent times.

Many years later, as I look back on my PhD journey, I will remember the countless moments of discovery and support that shaped my life.

Contents

List of Publications

This is an article-based thesis. The articles are identical in content as they appear here and in the original publications, except for minor changes such as the correction of typos and reformatting of tables and figures. The following articles are included as chapters in the thesis, listed in the order of their appearance within the document (* denotes equal first-author contribution):

1. Ruixiang Cui, Rahul Aralikatte, Heather Lent, and Daniel Hershcovich. 2022a. Compositional generalization in multilingual semantic parsing over Wikidata. *Transactions of the Association for Computational Linguistics*, 10:937–955.

2. Ruixiang Cui, Daniel Hershcovich, and Anders Søgaard. 2022b. Generalized quantifiers as a source of error in multilingual NLU benchmarks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4875–4893, Seattle, United States. Association for Computational Linguistics.

3. Ruixiang Cui, Seolhwa Lee, Daniel Hershcovich, and Anders Søgaard. 2023. What does the failure to reason with "respectively" in zero/few-shot settings tell us about language models? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8786–8800, Toronto, Canada. Association for Computational Linguistics.

4. Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. AGIEval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*,

pages 2299–2314, Mexico City, Mexico. Association for Computational Linguistics

Below is a list of articles, in chronological order, that I co-authored during the course of my Ph.D., but not included as part of this thesis:

1. Ruixiang Cui and Daniel Hershcovich. 2021a. Great service! fine-grained parsing of implicit arguments. In *Proceedings of the 17th International Conference on Parsing Technologies (IWPT 2021)*, pages 65–77, Online. Association for Computational Linguistics

2. Ruixiang Cui and Daniel Hershcovich. 2021b. Meaning representation of numeric fused-heads in ucca. In *Proceedings of the 1st Workshop on Understanding Implicit and Underspecified Language*, Online. Association for Computational Linguistics

3. Stephanie Brandl, Ruixiang Cui, and Anders Søgaard. 2022. 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. Association for Computational Linguistics

4. Nikolaus Schrack, Ruixiang Cui, Hugo López, and Daniel Hershcovich. 2022. Can AMR assist legal and logical reasoning? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1555–1568, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics

5. 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. 2022. 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. Association for Computational Linguistics

6. Yong Cao, Yova Kementchedjhieva, Ruixiang Cui, Antonia Karamolegkou, Li Zhou, Megan Dare, Lucia Donatelli, and Daniel Hershcovich. 2024.

Cultural adaptation of recipes. *Transactions of the Association for Computational Linguistics*, 12:80–99

7. Antonia Karamolegkou, Phillip Rust, Yong Cao, Ruixiang Cui, Anders Søgaard, and Daniel Hershcovich. 2024. Vision-language models under cultural and inclusive considerations. In *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*, Bangkok, Thailand. Association for Computational Linguistics

List of Figures

List of Tables

Chapter 1

Introduction

Few fields have captured the imagination as profoundly as artificial intelligence in human history. In 1950, Alan Turing proposed the Turing test, a thought experiment to decide whether a machine exhibits human-like intelligence (Turing, 1950). It inspired people to imagine a future where machines can act like humans. At the Dartmouth Workshop in 1956, AI was coined as a research discipline "to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it" (McCarthy et al., 1956). AI has been driving computer science development for more than half a century, from mechanical inventions and mathematical theories to modern computing theory and technology (Russell and Norvig, 2016). Moreover, AI has transcended its field to become a profound artistic inspiration, birthing countless novels, films, and other creative works (Boden, 2016; Elgammal et al., 2017). Humans have sparked thinking and discussion about the future of AI in society.

Language, a fundamental element in our life, allows us to express our thoughts and experiences to others (Pinker, 2003; Tomasello, 2010). It is our primary means of storing and sharing knowledge (Clark, 2006; Vygotsky, 2012). The Turing test is essentially a language test; therefore, natural language processing (NLP) is one of the most critical areas in AI. It focuses on developing models that can communicate like humans. The fundamental goal of NLP researchers is to equip computers to process, understand, and generate text (Jurafsky and Martin, 2000). This goal requires addressing the inherent complexity of human language, including but not limited to its syntactic structure, semantic ambiguity, and contextual dependencies (Manning and Schutze, 1999).

Historically, two mainstream approaches have emerged in modelling languages: symbolic and connectionist. Symbolic approaches are rooted in formal linguistics, arguing that language can be effectively modelled through explicit rules and symbolic representations, aiming to encapsulate the grammatical structure and semantic relations inherent in language (Chomsky, 1957; Woods, 1970; Winograd, 1971; Schank and Abelson, 1978) In contrast, connectionist approaches are inspired by the neural structure of the human brain and use artificial neural networks to learn statistical patterns and representations from large amounts of language data (Seidenberg and McClelland, 1989; Elman, 1990; Bengio et al., 2000). Despite the early efforts by connectionists, the complexity of human language remains an obstacle to developing computational models that can generate fluent and coherent text due to historical

limitations in computing power and hardware.

However, the renewed focus on deep learning radically shifted the AI research paradigm (Krizhevsky et al., 2012). The Transformer architecture (Vaswani et al., 2017) inspired by the human attention mechanism has made large-scale model training feasible. Transformer-based models demonstrate excellent performance on assorted benchmarks (Devlin et al., 2019). Unlike early efforts in solving tasks individually, they are general language models (LMs) that can tackle multiple language tasks in a unified approach. The 2022 release of ChatGPT (OpenAI, 2022), with its exceptional conversational abilities, has captivated global attention. Its coherent and context-aware dialogue capacity has opened new possibilities for integrating AI across diverse fields, suggesting a future of seamless human-computer interaction (Bubeck et al., 2023).

The emergence of large language models (LLMs) undoubtedly marks an important milestone in AI and NLP. However, as these models evolve in sophistication and find real-world applications, their impact on human life is becoming increasingly profound and challenging to predict. Proactive and rigorous evaluation becomes increasingly essential. Researchers must actively evaluate the performance, reliability, and security of these systems (Sculley et al., 2015; Mitchell et al., 2019).

During development, the evaluation of language models focuses on metrics such as perplexity, which measures how well a model predicts the next word in a sequence based on the preceding context (Jelinek et al., 1977). While perplexity provides a helpful benchmark for comparing different models, it does not fully capture the nuances of human language understanding and generation (Ethayarajh and Jurafsky, 2020; Sai et al., 2022). As language models have evolved to tackle more complex tasks, such as question answering, text summarization, and dialogue generation, the need for more carefully designed, specialized evaluations is apparent to account for different capabilities and usage scenarios of language models (Chang et al., 2024).

The introduction of standardized benchmarks, such as the Penn Treebank (Marcus et al., 1993) and the CoNLL-2003 shared task (Tjong Kim Sang and De Meulder, 2003), has been significant in language model evaluation. These benchmarks have become widely adopted in the NLP community and made a lasting impact on LM development.

Recently, there has been a growing interest in evaluating neural networks and language models on their compositional generalization abilities (Fodor and Pylyshyn, 1988). Compositional generalization refers to the ability to

understand and generate novel combinations from known elements (Lake and Baroni, 2018). It is a vital aspect of human language, as it allows us to understand and produce infinite valid sentences using a finite set of linguistic building blocks (Chomsky, 1965). Researchers have developed tasks to evaluate the compositional generalization abilities of models (Lake and Baroni, 2018), but these tasks are primarily English-centric.

In fact, the majority of LM benchmarks are in English (Søgaard, 2022). Moreover, LMs nowadays are pre-trained on English-centric datasets, such as Common Crawl (Kreutzer et al., 2022). There is a significant performance gap between high-resource and low-resource languages. Developing fair models across linguistic groups is crucial (Joshi et al., 2020b). A critical step in this direction is to create benchmarks that assess cross-lingual compositional generalization abilities. In this dissertation, we introduced a dataset termed Multilingual Compositional Wikidata Questions (MCWQ), a benchmark for measuring compositional generalization in semantic parsing (SP) for knowledge base question answering (KBQA) over Wikidata in four typologically diverse languages. We discover that pretrained multilingual LMs can achieve similar within-language generalization performances but struggle to transfer and generalize compositionally across languages.

Linguists and philosophers are among the first to discuss the compositionality nature of languages (Fodor and Pylyshyn, 1988; Partee et al., 1995). Linguists analyze languages by describing the structure of languages at various levels, such as syntax (sentence structure), semantics (meaning), and pragmatics (language use in context) (Chomsky, 1965; Lyons, 1977; Levinson, 1983). The theories have been leveraged for building early NLP models like rule-based automated translation systems (Nagao, 1984; Nirenburg, 1987) or dialogue agents (Weizenbaum, 1966; Bobrow et al., 1977). Given the swift advancements in LLMs, linguistics has shifted from its central role in NLP research. However, linguistic knowledge remains valuable and essential for understanding language formalisms in LLMs and analogizing their language capabilities compared to humans (Bender and Koller, 2020).

Whether LMs display human-like compositionality in language generation is a critical question. In this thesis, we investigate two linguistic phenomena, generalized quantifiers (GQs; Barwise and Cooper, 1981) and respective readings (Gawron and Kehler, 2004), as a gateway to the question. Quantifier words, such as “some”, “all”, and “most”, are abstract, important and vary across languages. Linguists proposed generalized quantifier theory to account for its complexity in a logical approach (Mostowski, 1957; Lindström,

1966). We analyze how GQs are represented in multilingual natural language understanding (NLU) benchmarks and how their occurrence at test time contributes to errors by LMs. We find that GQs systematically cause performance drops for LMs even though they are exposed to GQs extensively during training. Respective reading is another interesting linguistic phenomenon, which requires various levels of reasoning abilities, such as syntactic-semantic and commonsense-world knowledge. We create challenge datasets related to respective readings and investigate whether LMs can compositionally generalize from simple to complex linguistic constructions and from synthetic to naturally occurring datasets.

Traditional benchmarks have guided LM development a long way. They are typically created bottom-up, targeting specific domains or abilities. However, they are often not directly associated with real-world scenarios. Humans' advanced reasoning abilities in complex tasks demonstrate our capacity to apply learned knowledge in new contexts. An ideal LLM should be able to perform human-level complex reasoning. In this thesis, we introduce AGIEval, a benchmark derived from 20 official, public, and high-standard admission and qualification exams intended for general human test-takers, such as general college admission tests, math competitions, and lawyer qualification exams. AGIEval targets multiple dimensions of model capabilities and underscores the importance of evaluating LLMs in the context of human-level tasks. Our findings reveal these models' understanding, knowledge utilization, reasoning, and calculation limitations.

1.1 Background

1.1.1 Benchmarking Language Models

Language model benchmarking has become essential in NLP system developments. As LMs progress rapidly, it is important to have benchmarks that assess their capabilities across different tasks, languages, and domains. One of the primary goals of LM benchmarking is to establish a standard of evaluation metrics and datasets that researchers can use to compare models objectively. This standardization ensures consistency and robustness in LM evaluation, regardless of the model architectures or training methodologies. Benchmarks can help researchers discover limitations and prioritize goals in developing language models. LM evaluation targets various language capabilities, from

natural language understanding inference to reasoning and generation.

The General Language Understanding Evaluation benchmark (GLUE; Wang et al., 2018) is one of the most recognized benchmarks. It comprises nine existing tasks, including single-sentence classification, natural language inference and similarity detection to provide a holistic view of model capabilities. GLUE has drawn the attention of researchers, which has consequently pushed LM's progress. However, new models have surpassed human-level performance in just over a year. The following-up benchmark, SuperGLUE (Wang et al., 2019), was introduced with increased challenging tasks for LMs. The benchmark requires more complex reasoning abilities, e.g., coreference resolution, causal reasoning, and multi-hop question answering. SuperGLUE has pushed the boundaries of more advanced language model development (Raffel et al., 2020; Brown et al., 2020).

Multilingual evaluation is another essential part of LM benchmarking. As LMs are being applied in various countries and regions, developing models that can process and generate multilingual texts becomes crucial. Benchmarks including XNLI (Conneau et al., 2018), XQuAD (Artetxe et al., 2020b), and XTREME (Hu et al., 2020) have been proposed to assess their multilingual and cross-lingual transfer capabilities.

Model developments rely on benchmarking. This requires creating more datasets targeting LLMs' increasingly sophisticated abilities. For example, the recently introduced Beyond the Imitation Game benchmark (BIG-bench; Srivastava et al., 2023) assesses the performance of LLMs on 204 tasks that require complex reasoning abilities and common-sense understanding. The Massive Multitask Language Understanding benchmark (MMLU; Hendrycks et al., 2020) has been one of the most popular benchmarks for LLM assessment and development in recent years. It spans 57 academic subjects, including mathematics, philosophy, law, and medicine.

1.1.2 Semantic Parsing and Compositionality

Vast amounts of data are stored in structured or semi-structured knowledge bases in today's world. Therefore, developing technologies that can access them efficiently is important. Semantic parsing maps natural language utterances to formal meaning representations, such as logical forms or structured queries. (Kamath and Das, 2019). However, a key challenge in semantic parsing is compositionality, which is the ability to understand the meaning of complex expressions by combining the meanings of their parts (Lake and

Baroni, 2018). Humans exhibit strong compositional reasoning abilities, understanding novel combinations of previously learned concepts. In contrast, neural models often struggle with compositional generalization, failing to correctly interpret new compositions of familiar elements.

Benchmarking efforts in semantic parsing have aimed to test compositional generalization in LLMs. For example, the SCAN task (Lake and Baroni, 2018) requires models to translate commands presented in simplified natural language into a sequence of actions. An ideal model should be able to interpret new commands like "jump twice and walk thrice" when only exposed to primitive commands, i.e., single elements like "jump", "twice" and "walk", during training. By using advanced neural architectures such as hierarchical poset decoding (Guo et al., 2020), span representation (Herzig and Berant, 2021) and graph encoding (Gai et al., 2021), researchers have achieved perfect scores on the dataset. However, more realistic semantic parsing benchmarks with larger vocabularies still pose significant challenges (Drozdov et al., 2023).

COGS (Kim and Linzen, 2020) is a semantic parsing dataset from English sentences to logical forms. The training and evaluation splits differ in their properties, which can only be addressed by compositional generalization; these include new combinations of familiar syntactic structures or new combinations of familiar words and familiar structures. The Compositional Freebase Questions (CFQ; Keysers et al., 2020) is another semantic parsing dataset targeting the evaluation of compositional generalization. The task is to parse English questions into SPARQL queries executable on Freebase. The instances in the dataset can be broken down into atoms and compounds so that the train-test split can satisfy specific compositionality challenges. Nevertheless, none of the existing works study the multilingual and cross-lingual aspects of compositional generalization in neural networks and language models.

1.1.3 Generalized Quantifier

Quantifier words have been a subject of extensive study in both logic and linguistics for over a century (Frege, 1879; Westerståhl, 1989; Peters and Westerståhl, 2006). Quantifier words express relationships between sets of entities, events, and properties. Generalized quantifiers (GQs) are a key concept in formal semantics that allow for a more precise and expressive representation of natural language quantification. GQs generalize the universal ($\forall$) and existential ($\exists$) quantifiers of first-order logic (Mostowski, 1957; Lindström, 1966; Montague, 1973; Bach et al., 1995; Keenan and Paperno, 2012).

Formally, a generalized quantifier is a relation between two sets. For example, in the sentence "Most students took the exam", "most" is the generalized quantifier, "students" is the restrictor, and "took the exam" is the scope. The meaning of this sentence can be represented as: "Most(x.Student(x), x.TookExam(x))". This logical form expresses that the set of individuals who are students and took the exam is larger than the set of individuals who are students and did not take the exam.

While the study of generalized quantifiers has its roots in English and other European languages, these linguistic constructions are found across a wide range of typologically diverse languages. Cross-linguistic research on generalized quantifiers has revealed universal patterns and language-specific variations in how quantification is expressed and interpreted (Fintel and Matthewson, 2008). However, languages also exhibit diversity in their quantifier systems. For example, some languages, like English, have a rich set of lexical quantifiers, while Mandarin Chinese relies more heavily on numeral classifiers and adverbial expressions to convey quantification (Cheng, 2009).

Evaluating LMs' understanding of GQs across languages poses additional challenges. Models need to capture the core semantic properties of quantifiers and the language-specific ways in which they are expressed and interpreted. Multilingual benchmarks like XNLI include some examples testing quantifier reasoning, but more targeted cross-linguistic evaluation is needed.

1.1.4 Respective Readings

Respective readings are interpretations that arise in specific linguistic constructions, particularly those involving coordination and plural entities. In a respective reading, elements of one set of entities are mapped to elements of another set one-to-one, often based on their order or some other salient correspondence (Gawron and Kehler, 2004; Chaves, 2012). For example, in the sentence "Niels Bohr and Kurt Cobain were born in Copenhagen and Seattle, respectively", the first conjunct in the subject (Niels Bohr) corresponds to the first conjunct in the object (born in Copenhagen). The second conjunct in the subject (Kurt Cobain) corresponds to the second conjunct in the object (born in Seattle). Respective readings pose challenges for compositional semantics because they require mechanisms for establishing and interpreting mappings between sets of entities that go beyond standard notions of predicate-argument structure and scope (Dalrymple and Kehler, 1995; Okada, 1999).

In NLP, respective readings are relevant for tasks that require deep natural

language understanding, such as question answering and natural language inference. Existing benchmarks do not explicitly target respective constructions but may include some relevant examples. Developing more targeted datasets and probing techniques for respective readings could help shed light on how LLMs capture this type of semantic mapping and provide insights into the generalizability of LMs to complex linguistic constructions.

1.2 Contributions

This dissertation contributes to the field by providing diverse perspectives on language model evaluation through the lens of generalizability across languages, tasks and linguistic constructions. We aim to provide a more comprehensive understanding of current models' progress, limitations, and potential. We highlight the generalization limitations of LMs, demonstrate the continued relevance of linguistic theories in understanding and evaluating LLMs, and emphasize the need for designing benchmarks that are closely aligned with humans.

Multilingual Compositional Generalization In Chapter 2, we propose a method to automatically migrate a KBQA dataset to another knowledge base (KB) and extend it to diverse languages and domains. Using this method, we create a multilingual, parallel dataset of question-query pairs grounded in Wikidata. We call the dataset Multilingual Compositional Wikidata Questions (MCWQ) and use it to analyze the compositional generalization of neural networks and language models in four languages: Hebrew, Kannada, Chinese and English. In our monolingual experiments, we find similar performances for within-language generalization. However, in zero-shot cross-lingual experiments, we show that compositional generalization from English to other languages fails even with pretrained multilingual encoders.

Generalized Quantifier Reasoning In Chapter 3, we examined the challenges NLU models face with generalized quantifiers (GQs). Specifically, we examined GQs in multilingual NLU tasks and proposed a new Natural Language Inference (NLI) dataset for GQ reasoning. We found that GQs cause significant performance drop for NLU models, particularly when combined with other logical operators or when multiple GQs co-occur, despite their frequent occurrence in pertaining data. We propose a new dataset called

GQNLI to facilitate researchers in studying and improving model performance on generalized quantifier reasoning.

Compositionality over Linguistic Constructions: a Case Study of Respective Readings In Chapter 4, we investigated LMs' generalizability of linguistic phenomena through the lens of respective readings. We proposed two datasets, WikiResNLI (a controlled synthetic dataset) and NatResNLI (a naturally occurring dataset), to probe their ability in zero-shot and few-shot settings. We show that LMs can generalize effectively in both few-shot fine-tuning and in-context learning when the word "respectively" is present. However, more training instances are required when the reading is implicitly evoked. It takes significantly more instances to generalize to naturally occurring datasets than humans. In both studies, we demonstrate that LMs lack generalizability to linguistic constructions unseen in their training data.

Benchmarking Human-level Complex Reasoning In Chapter 5, we introduced AGIEval, a benchmark designed to assess the general abilities of LLMs, emphasizing human-level reasoning and real-world relevance. The benchmark comprises high-standard official exams in both English and Chinese, including general college admission tests (SAT, Gaokao, GRE), law school admission exams (LSAT), lawyer qualification tests, graduate management admission tests (GMAT), civil service examinations, and math competitions (AMC, AIME). Our evaluation indicates impressive performance on human-centric exams by LLMs. However, our qualitative analysis reveals they exhibit limitations in understanding, knowledge, reasoning, and calculation.

Chapter 2

Compositional Generalization in Multilingual Semantic Parsing over Wikidata

Abstract

Semantic parsing (SP) allows humans to leverage vast knowledge resources through natural interaction. However, parsers are mostly designed for and evaluated on English resources, such as CFQ (Keysers et al., 2020), the current standard benchmark based on English data generated from grammar rules and oriented towards Freebase, an outdated knowledge base. We propose a method for creating a multilingual, parallel dataset of question-query pairs, grounded in Wikidata. We introduce such a dataset, which we call Multilingual Compositional Wikidata Questions (MCWQ), and use it to analyze the compositional generalization of semantic parsers in Hebrew, Kannada, Chinese and English. While within-language generalization is comparable across languages, experiments on zero-shot cross-lingual transfer demonstrate that cross-lingual compositional generalization fails, even with state-of-the-art pretrained multilingual encoders. Furthermore, our methodology, dataset and results will facilitate future research on SP in more realistic and diverse settings than has been possible with existing resources.

Lang.	Question
En	Did Lohengrin 's male actor marry Margarete Joswig
He	האם ה שחקן ה גברי של לוהנגרין התחתן עם מרגרט יוסוויג
Kn	ಲೋಹೆಂಗ್ರಿನ್ ಅವರ ಪುರುಷ ನಟ ವಿವಾಹವಾದರು ಮಾರ್ಗರೇಟ್ ಜೋಸ್ವಿಗ್
Zh	Lohengrin 的 男 演员 嫁给 了 Margarete Joswig 吗

SPARQL Query:

```
ASK WHERE { ?x0 wdt:P453 wd:Q50807639 . ?x0 wdt:P21 wd:Q6581097 . ?x0 wdt:P26
wd:Q1560129 . FILTER ( ?x0 != wd:Q1560129 )}
```

Figure 2.1: An example from the MCWQ dataset. The question in every language corresponds to the same Wikidata SPARQL query, which, upon execution, returns the answer (which is positive in this case).

2.1 Introduction

Semantic parsers grounded in knowledge bases (KBs) enable knowledge base question answering (KBQA) for complex questions. Many semantic parsers are grounded in KBs such as Freebase (Bollacker et al., 2008), DBpedia (Lehmann et al., 2015) and Wikidata (Pellissier Tanon et al., 2016), and models can learn to answer questions about unseen entities and properties (Herzig and Berant, 2017; Cheng and Lapata, 2018; Shen et al., 2019; Sas et al., 2020). An important desired ability is compositional generalization—the ability to generalize to unseen *combinations* of known components (Oren et al., 2020; Kim and Linzen, 2020).

One of the most widely used datasets for measuring compositional generalization in KBQA is CFQ (Compositional Freebase Questions; Keysers et al., 2020), which was generated using grammar rules, and is based on Freebase, an outdated and unmaintained English-only KB. While the need to expand language technology to many languages is widely acknowledged (Joshi et al., 2020b), the lack of a benchmark for compositional generalization in multilingual semantic parsing (SP) hinders KBQA in languages other than English. Furthermore, progress in both SP and KB necessitates that benchmarks can be reused and adapted for future methods.

Wikidata is a multilingual KB, with entity and property labels in a multitude of languages. It has grown continuously over the years and is an important complement to Wikipedia. Much effort has been made to migrate Freebase data to Wikidata (Pellissier Tanon et al., 2016; Diefenbach et al.,

CFQ field	Content
`questionWithBrackets`	Did ['Murder' Legendre]'s male actor marry [Lillian Lugosi]
`questionPatternModEntities`	Did M0 's male actor marry M2
`questionWithMids`	Did `m.0h4y854` 's male actor marry `m.0hpnx3b`
`sparql`	`SELECT count(*) WHERE { ?x0 ns:film.actor.film/ns:film.performance. character ns:m.0h4y854 . ?x0 ns:people.person.gender ns:m.05zppz . ?x0 ns:people.person.spouse_s/ns:fictional_universe. marriage_of_ fictional_characters.spouses ns:m.0hpnx3b . FILTER ( ?x0 != ns:m.0hpnx3b )}`
`sparqlPatternModEntities`	`SELECT count(*) WHERE { ?x0 ns:film.actor.film/ns:film.performance. character M0 . ?x0 ns:people.person.gender ns:m.05zppz . ?x0 ns:people.person.spouse_s /ns:fictional_universe.marriage_of_fictional_characters.spouses M2 . FILTER ( ?x0 != M2 )}`

Table 2.1: Selected fields in a CFQ entry. `questionWithBrackets` is the full English question with entities surrounded by brackets. `questionPatternModEntities` is the question with entites replaced by placeholders. In `questionWithMids`, the entity codes (Freebase machine IDs; MIDs) are given instead of their labels. `sparql` is the fully executable SPARQL query for the question, and in `sparqlPatternModEntities` the entity codes are replaced by placeholders.

2017; Hogan et al., 2021) but only in English. Investigating compositional generalization in cross-lingual SP requires a multilingual dataset, a gap we address in this work.

We leverage Wikidata and CFQ to create **M**ultilingual **C**ompositional **W**ikidata **Q**uestions (MCWQ), a new multilingual dataset of compositional questions grounded in Wikidata. Beyond the original English, an Indo-European language using the Latin script, we create parallel datasets of questions in Hebrew, Kannada and Chinese, which use different scripts and belong to different language families: Afroasiatic, Dravidian and Sino-Tibetan, respectively. Our dataset includes questions in the four languages and their associated SPARQL queries.

Our contributions are:

- a method to automatically migrate a KBQA dataset to another KB and extend it to diverse languages and domains,
- a benchmark for measuring compositional generalization in SP for KBQA over Wikidata in four typologically diverse languages,
- monolingual experiments with different SP architectures in each of the

four languages, demonstrating similar within-language generalization, and

- zero-shot cross-lingual experiments using pretrained multilingual encoders, showing that compositional generalization from English to the other languages fails.

Our code for generating the dataset and for the experiments, as well as the dataset itself and trained models, are publicly available on `https://github.com/coastalcph/seq2sparql`.

2.2 Limitations of CFQ

CFQ (Compositional Freebase Questions; Keysers et al., 2020) is a dataset for measuring compositional generalization in SP. It targets the task of parsing questions in English into SPARQL queries executable on the Freebase KB (Bollacker et al., 2008). CFQ contains questions as in Table 2.1, as well as the following English question (with entities surrounded by brackets):

> "Was [United Artists] founded by [Mr. Fix-it]'s star, founded by [D. W. Griffith], founded by [Mary Pickford], and founded by [The Star Boarder]'s star?"

Parsers trained on CFQ transform these questions into SPARQL queries, which can subsequently be executed against Freebase to answer the original questions (in this case, "Yes").

CFQ uses the Distribution-Based Compositionality Assessment (DBCA) method to generate multiple train-test splits with maximally divergent examples in terms of compounds, while maintaining a low divergence in terms of primitive elements (atoms). In these *maximum compound divergence* (MCD) splits, the test set is constrained to examples containing novel compounds, i.e., new ways of composing the atoms seen during training. For measuring compositional generalizations, named entities in the questions are anonymized so that models cannot simply learn the relationship between entities and properties. CFQ contains 239,357 English question-answer pairs, which encompass 49,320 question patterns and 34,921 SPARQL query patterns. Table 2.1 shows selected fields of an example in CFQ. In their experiments, Keysers et al. (2020) trained semantic parsers using several architectures on various train-test splits. They demonstrated strong negative correlation between

models' accuracy (correctness of the full generated SPARQL query) and compound divergence across a variety of system architectures - all models generalized poorly in the high-divergence settings, highlighting the need to improve compositional generalization in SP.

By the time CFQ was released, Freebase had already been shut down. On that account, to our knowledge, there is no existing SP dataset targeting compositional generalization that is grounded in a currently usable KB, which contains up-to-date information. We therefore migrate the dataset to such a KB, namely Wikidata, in §2.3.

Moreover, only few studies have evaluated semantic parsers' performance in a multilingual setting, due to the scarcity of multilingual KBQA datasets (Perevalov et al., 2022b). No comparable benchmark exists for languages other than English, and it is therefore not clear whether results are generalizable to other languages. Compositional generalization in typologically distant languages may pose completely different challenges, as these languages may have different ways to compose meaning (Evans and Levinson, 2009). We create such a multilingual dataset in §2.4, leveraging the multilinguality of Wikidata.

2.3 Migration to Wikidata

Wikidata is widely accepted as the replacement for Freebase. It is actively maintained and represents knowledge in a multitude of languages and domains, and also supports SPARQL. Migrating Freebase queries to Wikidata, however, is not trivial, as there is no established full mapping between the KBs' properties and entities. An obvious alternative to migration would be a replication of the original CFQ generation process but with Wikidata as the KB. Before delving into the details of the migration process, let us motivate the decision not to pursue that option: the grammar used to generate CFQ was not made available to others by Keysers et al. (2020) and is prohibitively too complex to reverse-engineer. Our migration process, on the other hand, is general and can similarly be applied for migrating other datasets from Freebase to Wikidata. Finally, many competitive models with specialized architecture have been developed for CFQ (Guo et al., 2020; Herzig et al., 2021; Gai et al., 2021). Our migrated dataset is formally similar and facilitates their evaluation and the development of new methods.

Lang.	MCWQ field	Content
En	`questionWithBrackets`	Did [Lohengrin] 's male actor marry [Margarete Joswig]
	`questionPatternModEntities`	Did M0 's male actor marry M2
He	`questionWithBrackets`	האם השחקן הגברי של [לוהנגרין] התחתן עם [מרגרט יוסוויג]
	`questionPatternModEntities`	האם השחקן הגברי של M0 התחתן עם M2
Kn	`questionWithBrackets`	[ಲೋಹೆಂಗ್ರಿನ್] ಅವರ ಪುರುಷ ನಟ ವಿವಾಹವಾದರು [ಮಾರ್ಗರೇಟ್ ಜೋಸ್ವಿಗ್]
	`questionPatternModEntities`	M0 ನ ಪುರುಷ ನಟ M2 ಅನ್ನು ಮದುವೆಯಾಗಿದ್ದಾರೆಯ
Zh	`questionWithBrackets`	[Lohengrin]的男演员嫁给了[Margarete Joswig]吗
	`questionPatternModEntities`	M0的男演员和M2结婚吗
	`sparql`	`ASK WHERE { ?x0 wdt:P453 wd:Q50807639 . ?x0 wdt:P21 wd:Q6581097 . ?x0 wdt:P26 wd:Q1560129 . FILTER ( ?x0 != wd:Q1560129 )}`
	`sparqlPatternModEntities`	`ASK WHERE { ?x0 wdt:P453 M0 . ?x0 wdt:P21 wd:Q6581097 . ?x0 wdt:P26 M2 . FILTER ( ?x0 != M2 )}`
	`recursionDepth`	20
	`expectedResponse`	True

Table 2.2: The MCWQ example from Figure 2.1. The English question is generated from the CFQ entry in Table 2.1 by the migration process described in §2.3.3, and the questions in the other languages are automatically translated (§2.4.1). The `questionWithBrackets`, `questionPatternModEntities`, `sparql` and `sparqlPatternModEntities` fields are analogous to the CFQ ones. `recursionDepth` (which quantifies the question complexity) and `expectedResponse` (which is the answer returned upon execution of the query) are copied from the CFQ entry.

2.3.1 Property Mapping

As can be seen in Table 2.1, the `WHERE` clause in a SPARQL query consists of a list of triples, where the second element in each triple is the property, e.g., `ns:people.person.gender`. CFQ uses 51 unique properties in its SPARQL queries, mostly belonging to the cinematography domain. These Freebase properties cannot be applied directly to Wikidata, which uses different property codes known as P-codes, e.g., `P21`. We therefore need to map the Freebase properties into Wikidata properties.

As a first step in the migration process, we check which Freebase properties used in CFQ have corresponding Wikidata properties. Using a publicly

available repository providing a partial mapping between the KBs,[1] we identify 22 out of the 51 Freebase properties in CFQ can be directly mapped to Wikidata properties.[2] The other 29 require further processing:

Fourteen properties are the reverse of other properties, which do not have Wikidata counterparts. For example, `ns:film.director.film` is the reverse of `ns:film.film.directed_by`, and only the latter has Wikidata mapping, `P57`. We resolve the problem by swapping the entities around the property.

The other 15 properties deal with judging whether an entity has a certain quality. In CFQ, `?x1 a ns:film.director` asks whether `?x1` is a director. Wikidata does not contain such unary properties. Therefore, we need to treat these CFQ properties as entities in Wikidata. For example, *director* is `wd:Q2526255`, so we paraphrase the query as `?x1 wdt:P106 wd:Q2526255`, asking whether `?x1`'s *occupation* (`P106`) is director. In addition, we substitute the *art director* property from CFQ with the *composer* property because the former has no equivalent in Wikidata. Finally, we filter out queries with reverse marks over properties, e.g., `?x0 ^ns:people.person.gender M0`, due to incompatibility with the question generation process (§2.3.2).

After filtering, we remain with 236,304 entries with only fully-mappable properties—98.7% of all entries in CFQ. We additionally make necessary SPARQL syntax modification for Wikidata.[3]

2.3.2 Entity Substitution

A large number of entities in Freebase are absent in Wikidata. For example, neither of the entities in Table 2.1 exist in Wikidata. Furthermore, unlike the case of properties, to our knowledge, there is no comprehensive or even partial mapping of Freebase entity IDs (i.e., Freebase machine IDs, MIDs, such as `s:m.05zppz`) to Wikidata entity IDs (i.e., Q-codes, such as `wd:Q6581097`). We replicate the grounding process carried out by Keysers et al. (2020), substituting entity placeholders with compatible entities codes by executing the queries against Wikidata:

1. Replacing entity placeholders with SPARQL **variables** (e.g., `?v0`), we

[1] `https://www.wikidata.org/wiki/Wikidata:WikiProject_Freebase/Mapping`

[2] While some Freebase properties have multiple corresponding Wikidata properties, we consider a property mappable as long as it has at least one mapping.

[3] CFQ uses `SELECT count(*) WHERE` to query yes/no questions, but this syntax is not supported by Wikidata. We replace it with `ASK WHERE`, intended for boolean queries.

obtain queries that return sets of compatible candidate entity assignments instead of simply an answer for a given assignment of entities.

2. We add constraints for the entities to be **distinct**, to avoid nonsensical redundancies (e.g., due to conjunction of identical clauses).

3. Special entities, representing **nationalities and genders**, are regarded as part of the question patterns in CFQ (and are not replaced with placeholders). Before running the queries, we thus replace all such entities with corresponding Wikidata Q-codes (instead of variables).

4. We **execute** the queries against the Wikidata query service[4] to get the satisfying assignments of entity combinations, with which we replace the placeholders in `sparqlPatternModEntities` fields.

5. Finally, we insert the Q-codes into the English **questions** in the `questionWithMids` field and the corresponding entity labels into the `questionWithBrackets` to obtain the English questions for our dataset.

Along this process, 52.5% of the queries have at least one satisfying assignment. The resulting question-query pairs constitute our English dataset. They maintain the SPARQL patterns in CFQ, but the queries are all executable on Wikidata.

We obtain 124,187 question-query pairs, of which 67,523 are yes/no questions and 56,664 are wh- questions. The expected responses of yes/no questions in this set are all "yes" due to our entity assignment process. To make MCWQ comparable to CFQ, which has both positive and negative answers, we sample alternative queries by replacing entities with ones from other queries whose preceding predicates are the same. Our negtive sampling results in 30,418 questions with "no" answers.

2.3.3 Migration Example

Consider the SPARQL pattern from Table 2.1:

```
SELECT count(*) WHERE { ?x0 ns:film.actor.film/ns:film
    .performance.character M0 . ?x0 ns:people.person.
    gender ns:m.05zppz . ?x0 ns:people.person.spouse_s/
```

[4] `https://query.wikidata.org/`

```
    ns:fictional_universe.
    marriage_of_fictional_characters.spouses M2 .
    FILTER ( ?x0 != M2 )}
```

We replace the properties and special entities (here the gender *male*: `ns:m.05zppz` → `wd:Q6581097`):

```
SELECT count(*) WHERE {?x0 wdt:P453 M0 .?x0 wdt:P21 wd
    :Q6581097 . ?x0 wdt:P26 M2 . FILTER ( ?x0 != M2 )}
```

Then we replace placeholders (e.g., `M0`) with variables and add constraints for getting only one assignment (which is enough for our purposes) with distinct entities. The resulting query is:

```
SELECT ?v0 ?v1 WHERE {?x0 wdt:P453 ?v0. ?x0 wdt:P21 wd
    :Q6581097. ?x0 wdt:P26 ?v1. FILTER ( ?x0 != ?v1 ).
    FILTER ( ?v0 != ?v1 )} LIMIT 1
```

We execute the query and get `wd:Q50807639` (Lohengrin) and `wd:Q1560129` (Margarete Joswig) as satisfying answers for `v0` and `v1` respectively. Note that these are different from the entities in the original question ('Murder' Legendre and Lillian Lugosi)—in general, there is no guarantee that the same entities from CFQ will be preserved in our dataset. Then we put back these answers into the query, and make necessary SPARQL syntax modification for Wikidata. The final query for this entry is:

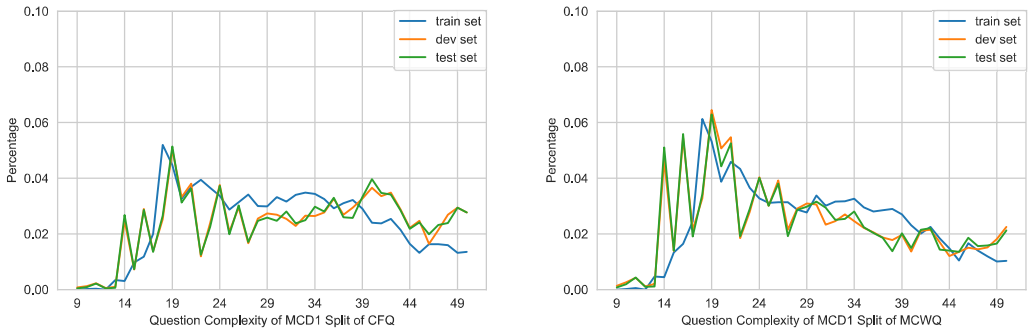
```
ASK WHERE {?x0 wdt:P453 wd:Q50807639. ?x0 wdt:P21 wd:
    Q6581097 . ?x0 wdt:P26 wd:Q1560129 . FILTER ( ?x0
    != wd:Q1560129 )}
```

As for the English question, we map the Freebase entities in the `questionWithMids` field with the labels of the obtained Wikidata entities. Therefore, the English question resulting from this process is:

> Did [Lohengrin] 's male actor marry [Margarete Joswig]?

2.3.4 Dataset Statistics

We compare the statistics of MCWQ with CFQ in Table 2.3. MCWQ has 29,312 unique question patterns (mod entities, verbs, etcs), i.e., 23.6% of questions cover all question patterns, compared to 20.6% in CFQ. Furthermore, MCWQ has 86,353 unique query patterns (mod entities), resulting in 69.5%



(a) CFQ (b) MCWQ

Figure 2.2: Complexity distribution of the MCD_1 split of CFQ (left) and MCWQ (right).

	CFQ	**MCWQ**
Unique questions	239,357	124,187
Questions patterns	49,320 (20.6%)	29,312 (23.6%)
Unique queries	228,149 (95.3%)	101,856 (82%)
Query patterns	123,262 (51.5%)	86,353 (69.5%)
Yes/no questions	130,571 (54.6%)	67,523 (54.4%)
Wh- questions	108,786 (45.5%)	56,664 (45.6%)

Table 2.3: Dataset statistics comparison for MCWQ and CFQ. Percentages are relative to all unique questions. Questions patterns refer to mod entities, verbs, etc. while query patterns refer to mod entities only.

of instances covering all SPARQL patterns, 18% higher than CFQ. Our dataset thus poses a greater challenge for compositional SP, and exhibits less redundancy in terms of duplicate query patterns. It is worth noting that less unique query percentage in MCWQ than CFQ results from the loss during swapping the entities in §2.3.1.

To be compositionally challenging, Keysers et al. (2020) generated the MCD splits to have high compound divergence while maintaining low atom divergence. As atoms in MCWQ are mapped from CFQ while leaving the compositional structure intact, we derive train-test splits of our dataset by inducing the train-test splits from CFQ on the corresponding subset of

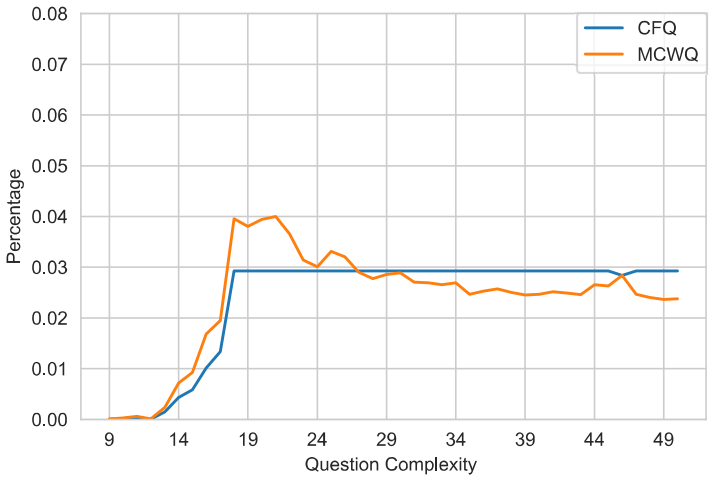


Figure 2.3: Complexity distribution of MCWQ, measured by recursion depth, compared to CFQ.

instances in our dataset.

The complexity of questions in CFQ is measured by recursion depth and reflects the number of rule applications used to generate a question, which encompasses grammar, knowledge, inference and resolution rules. While each question's complexity in MCWQ is the same as the corresponding CFQ question's, some cannot be migrated (see §2.3.1 and §2.3.2). To verify the compound divergence is not affected, we compare the question complexity distribution of the two datasets in one of the three compositional splits (MCD1) in Figure 2.2. The training, development and test sets of the split in CFQ and MCWQ follow a similar trend in general. The fluctuation in the complexity of questions in the MCWQ splits reflects the dataset's full distribution—see Figure 2.3.

Stemming from its entities and properties, CFQ questions are limited to the domain of movies. The entities in MCWQ, however, can in principle come from any domain, owing to our flexible entity replacing method. Though MCWQ's properties are still a subset of those used in CFQ, they are primarily in the movies domain. We also observe a few questions from literature, politics, and history in MCWQ.

2.4 Generating Multilingual Questions

To create a typologically diverse dataset, starting from our English dataset (an Indo-European language using the Latin script), we use machine translation to three other languages from different families (Afroasiatic, Dravidian and Sino-Tibetan), which use different scripts: Hebrew, Kannada and Chinese (§2.4.1). For a comparison to machine translation and a more realistic evaluation with regards to compositional SP, we manually translate a subset of the test sets of the three MCD splits (§2.4.2) and evaluate the machine translation quality (§2.4.3).

2.4.1 Generating Translations

Both question patterns and bracketed questions are translated separately with Google Cloud Translation[5] from English.[6] SPARQL queries remain unchanged, as both property and entity IDs are language-independent in Wikidata, which contains labels in different languages for each. Table 2.2 shows an example for a question in our dataset (which is generated from the same question as the CFQ instance from Table 2.1), as well as the resulting translations.

As an additional technical necessity, we add a question mark to the end of each question before translation (as the original dataset does not include question marks) and remove trailing question marks from the translated question before including it in our dataset. We find this step to be essential for translation quality.

2.4.2 Gold Test Set

CFQ and other datasets for evaluating compositional generalization (Lake and Baroni, 2018; Kim and Linzen, 2020) are generated from grammars. However, It has not been investigated how well models trained on them generalize to human questions. As a step towards that goal, we evaluate whether models trained with automatically generated and translated questions can generalize

[5] `https://cloud.google.com/translate`

[6] We attempted to translate bracketed questions and subsequently replace the bracketed entities with placeholders as question patterns. In preliminary experiments, we found that separate translation of question patterns is of higher translation quality. Therefore, we choose to translate question patterns and bracketed questions individually.

to high-quality human-translated questions. For that purpose, we obtain the intersection of the test sets of the MCD splits (1,860 entries), and sample two translated questions with yes/no questions and two with wh- questions for each complexity level (if available). This sample, termed *test-intersection-MT*, has 155 entries in total. The authors (one native speaker for each language) manually translate the English questions into Hebrew, Kannada and Chinese. We term the resulting dataset *test-intersection-gold*.

2.4.3 Translation Quality

We compute the BLEU (Papineni et al., 2002) scores of *test-intersection-MT* against *test-intersection-gold* using SacreBLEU (Post, 2018), resulting in 87.4, 76.6 and 82.8 for Hebrew, Kannada and Chinese, respectively. This indicates high quality of the machine translation outputs.

Additionally, one author for each language manually assesses translation quality for one sampled question from each complexity level from the full dataset (40 in total). We rate the translations on a scale of 1–5 for fluency and for meaning preservation, with 1 being poor, and 5 being optimal. Despite occasional translation issues, mostly attributed to lexical choice or morphological agreement, we confirm that the translations are of high quality. Across languages, over 80% of examples score 3 or higher in fluency and meaning preservation. The average meaning preservation scores for Hebrew, Kannada and Chinese are 4.4, 3.9 and 4.0. For fluency, they are 3.6, 3.9 and 4.4.

As a control, one of the authors (a native English speaker) evaluates English fluency for the same sample of 40 questions. Only 62% of patterns were rated 3 or above. While all English questions are grammatical, many suffer from poor fluency, tracing back to their automatic generation using rules. Some translations are rated higher in terms of fluency, mainly due to annotator leniency (focusing on disfluencies that might result from translation) and paraphrasing of unnatural constructions by the MT system (especially for lower complexities).

2.5 Experiments

While specialized architectures have been achieved state-of-the-art results on CFQ (Guo et al., 2020, 2021; Gai et al., 2021), these approaches are English- or Freebase-specific. We therefore experiment with sequence-to-sequence

	MCD_1				MCD_2				MCD_3				MCD_{mean}				Random			
Exact Match (%)	En	He	Kn	Zh	En	He	Kn	Zh	En	He	Kn	Zh	En	He	Kn	Zh	En	He	Kn	Zh
LSTM+Attention	38.2	29.3	27.1	26.1	6.3	5.6	9.9	7.5	13.6	11.5	15.7	15.1	19.4	15.5	17.6	16.2	96.6	80.8	88.7	86.8
E. Transformer	53.3	35	30.7	31	16.5	8.7	11.9	10.2	18.2	13	18.1	15.5	29.3	18.9	20.2	18.9	99	90.4	93.7	92.2
mBERT	49.5	38.7	34.4	35.6	13.4	11.4	12.3	15.1	17	18	18.1	19.4	26.6	22.7	21.6	23.4	98.7	**91**	95.1	**93.3**
T5-base+RIR	57.4	-	-	-	14.6	-	-	-	12.3	-	-	-	28.1	-	-	-	98.5	-	-	-
mT5-small+RIR	**77.6**	57.8	**55**	**52.8**	13	12.6	8.2	21.1	**24.3**	17.5	**31.4**	34.9	**38.3**	29.3	31.5	**36.3**	98.6	90	93.8	91.8
mT5-base+RIR	55.5	**59.5**	49.1	30.2	**27.7**	**16.6**	**16.6**	**23**	18.2	**23.4**	30.5	**35.6**	33.8	**33.2**	**32.1**	29.6	**99.1**	90.6	94.2	92.2

Table 2.4: Monolingual evaluation: exach match accuracies on MCWQ. MCD_{mean} is the mean accuracy of all three MCD splits. Random represents a random split of MCWQ. This is an upper bound on the performance shown only for comparison. As SPARQL BLEU scores are highly correlated with accuracies in this experiment, we only show the latter here.

(seq2seq) models, among which T5 (Raffel et al., 2020) has been shown to perform best on CFQ (Herzig et al., 2021). We evaluate these models for each lanugage separately (§2.5.1), and subsequently evaluate their cross-lingual compositional generalization (§2.5.2).

2.5.1 Monolingual Experiments

We evaluate six models' monolingual parsing performance on the three MCD splits and a random split of MCWQ. As done by Keysers et al. (2020), entities are masked during training, except those that are part of the question patterns (genders and nationalities).

We experiment with two seq2seq architectures on MCWQ for each language, with the same hyperparameters tuned by Keysers et al. (2020) on the CFQ random split: LSTM (Hochreiter and Schmidhuber, 1997) with attention mechanism (Bahdanau et al., 2015) and Evolved Transformer (So et al., 2019), both implemented using Tensor2Tensor (Vaswani et al., 2018). Separate models are trained and evaluated per language, with randomly initialized (not pretrained) encoders. We train a model for each of the three MCD splits plus a random split for each language.

We also experiment with pretrained language models (PLMs), to assess whether *multilingual* PLMs, mBERT (Devlin et al., 2019) and mT5 (Xue et al., 2020), are as effective for monolingual compositional generalization as an English-only PLM using the Transformers library (Wolf et al., 2020).

For mBERT, we fine-tune a `multi_cased_L-12_H-768_A-12` encoder and a randomly initialized decoder of the same architecture. We train for 100 epochs with patience of 25, batch size of 128, and learning rate of 5×10^{-5} with a linear decay.

For T5, we fine-tune `T5-base` on MCWQ English, and `mT5-small` and `mT5-base` on each language separately. We use the default hyperparameter settings except trying two learning rates, $5e^{-4}$ and $3e^{-5}$ (see results below). SPARQL queries are pre-processed using reversible intermediate representations (RIR), previously shown (Herzig et al., 2021) to facilitate compositional generalization for T5. We fine-tune all models for 50K steps.

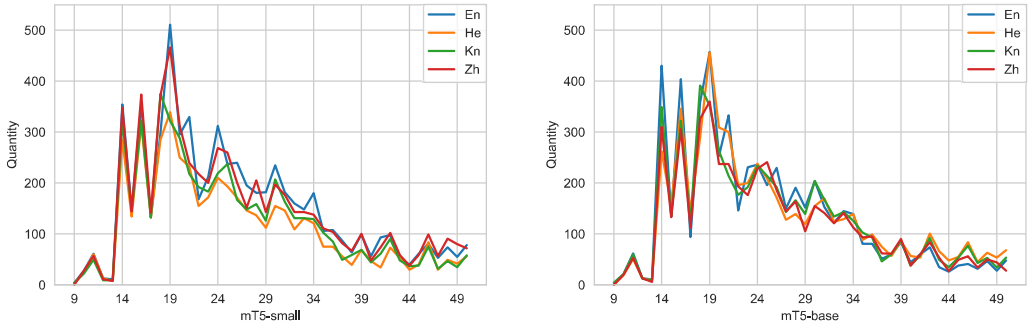
We use six Titan RTX GPUs for training, with batch size of 36 for `T5-base`, 24 for `mT5-small` and 12 for `mT5-base`. We use two random seeds for `T5-base`. It takes 384 hours to finish a round of `mT5-small` experiments, 120 hours for `T5-base` and 592 hours for `mT5-base`.

In addition to exact-match accuracy, we report the BLEU scores of the predictions computed with SacreBLEU, as a large portion of the generated queries is partially (but not fully) correct.

Results The results are shown in Table 2.4. While models generalize almost perfectly in the random split for all four languages, the MCD splits are much harder, with the highest mean accuracies of 38.3%, 33.2%, 32.1% and 36.3% for English, Hebrew, Kannada and Chinese, respectively. For comparison, on CFQ, `T5-base`+RIR has an accuracy of 60.8% on MCD_{mean} (Herzig et al., 2021). One reason for this decrease in performance is the smaller training data: the MCWQ dataset has 52.5% the size of CFQ. Furthermore, MCWQ has less redundancy than CFQ in terms of duplicate questions and SPARQL patterns, rendering models' potential strategy of simply memorizing patterns less effective.

Contrary to expectation, `mT5-base` does not outperform `mT5-small`. During training, we found `mT5-base` reached minimum loss early (after 1k steps). By changing the learning rate from the default $3e^{-5}$ to $5e^{-4}$, we seem to have overcome the local minimum. Training `mT5-small` with learning rate $5e^{-4}$ also renders better performance. Furthermore, the batch size we use for `mT5-base` may not be optimal, but we could not experiment with larger batch sizes due to resource limitations.

Comparing the performance across languages, `mT5-base` performs best on Hebrew and Kannada on average, while `mT5-small` has the best performance on English and Chinese. Due to resource limitations, we were not able to look deeper into the effect of hyperparameters or evaluate larger models. However, our experiments show that while multilingual compositional generalization is challenging for seq2seq semantic parsers, within-language generalization



(a) mT5-small (b) mT5-base

Figure 2.4: Two mT5 models' number of correct predictions summing over the three MCD splits in monolingual experiments, plotted by complexity level. Each line represents a language. While `mT5-small` generalizes better overall, `mT5-base` is better in lower complexities (which require less compositional generalization).

is comparable between languages. Nonetheless, English is always the easiest (at least marginally). A potential cause is that most semantic query languages were initially designed to represent and retrieve data stored in English databases, and thus have a bias towards English. Consequently, SPARQL syntax is closer to English than Hebrew, Kannada and Chinese. While translation errors might have an effect as well, we have seen in §2.4.3 that translation quality is high.

To investigate further, we plot the complexity distribution of true predictions (exactly matching the gold SPARQL) per language by the two best systems in Figure 2.4. We witness a near-linear performance decay from complexity level 19. We find that `mT5-base` is better than `mT5-small` on lower complexity despite the latter's superior overall performance. Interestingly, translated questions seem to make the parsers generalize better at higher complexity, as shown in the figure. For `mT5-small`, the three non-English models successfully parse more questions within the complexity range 46-50 than English, for `mT5-base` 44-50. As is discussed in §2.4.3, machine-translated questions tend to have higher fluency than English questions; we conjecture such a smoothing method helps the parser to understand and learn from higher complexity questions.

	MCD_{mean}				Random			
SPARQL BLEU	En	He	Kn	Zh	En	He	Kn	Zh
`mT5-small`+RIR	87.5	53.8	53.2	59	99.9	60.4	59.9	63.8
`mT5-base`+RIR	86.4	46.4	46	52.7	99.9	63.2	63.5	70.6
Exact Match (%)								
`mT5-small`+RIR	38.3	0.2	0.3	0.2	98.6	0.5	0.4	1.1
`mT5-base`+RIR	33.8	0.4	0.7	1.5	99.1	1.1	0.9	7.2

Table 2.5: Mean BLEU scores and exact match accuracies on the three MCD splits and on a random split in zero-shot cross-lingual transfer experiments on MCWQ. The grey texts represent the models' monolingual performance on English, given for reference (the exact match accuracies are copied from Table 2.4). The black texts indicate the zero-shot cross-lingual transfer performances on Hebrew, Kannada and Chinese of a model trained on English. While the scores for individual MCD splits are omitted for brevity, in all three MCD splits, the accuracies are below 1% (except on MCD_2 Chinese, being 4%).

2.5.2 Zero-shot Cross-lingual Parsing

Zero-shot cross-lingual SP has witnessed new advances with the development of PLMs (Shao et al., 2020; Sherborne and Lapata, 2022). Since translating datasets and training KBQA systems is expensive, it is beneficial to leverage multilingual PLMs, fine-tuned on English data, for generating SPARQL queries over Wikidata given natural language questions in different languages. While compositional generalization is difficult even in a monolingual setting, it is interesting to investigate whether multilingual PLMs can transfer in cross-lingual SP over Wikidata. Simple seq2seq T5/mT5 models perform reasonably well ($> 30\%$ accuracy) on monolingual SP on some splits (see §2.5.1). We investigate whether the learned multilingual representations of such models enable compositional generalization even without target language training. We use `mT5-small`+RIR and `mT5-base`+RIR, the best two models trained and evaluated on English from previous experiments, to predict on the other languages.

Results The results are shown in Table 2.5. Both BLEU and exact match accuracy of the predicted SPARQL queries drop drastically when the model is evaluated on Hebrew, Kannada and Chinese. `mT5-small`+RIR achieves 38.3% accuracy on MCD_{mean} English, but less than 0.3% in zero-shot parsing

	test-intersection-MT				*test-intersection-gold*			
SPARQL BLEU	En	He	Kn	Zh	En	He	Kn	Zh
`mT5-small`+RIR	86.1	82.5	78.9	85.1	-	81.8	77.7	86
`mT5-base`+RIR	85.5	83.7	81.8	83.2	-	83.8	80.9	83.8
Exact Match (%)								
`mT5-small`+RIR	45.6	35.7	32.7	38.5	-	35.9	28.2	39.8
`mT5-base`+RIR	40.4	41.9	40.2	38.7	-	41.1	34	38.9

Table 2.6: Mean BLEU scores and accuracies of monolingual models (§2.5.1) on *test-intersection-MT* and *test-intersection-gold*. The numbers are averaged over the accuracies of the predictions from the monolingual models trained on three MCD splits. Overall, there is no substantial difference between the performances on the two intersection sets, demonstrating the reliability of evaluating on machine translated data in this case.

on three non-English languages.

Even putting aside compositionality evaluation, as seen in the random split, the exact match accuracy in the zero-shot cross-lingual setting is still low. The relatively high BLEU scores can be attributed to the small overall vocabulary used in SPARQL queries. Interestingly, while `mT5-base`+RIR on MCD_{mean} English does not outperform `mT5-small`+RIR, it yields better performance in the zero-shot setting. For Hebrew, Kannada and Chinese, the accuracies are 0.2%, 0.4% and 1.3% higher. For `mT5-base`, Chinese is slightly easier than Kannada and Hebrew to parse in the zero-shot setting, outperforming 1.1% and 0.8%.

To conclude, zero-shot cross-lingual transfer from English to Hebrew, Kannada and Chinese fails to generate valid queries in MCWQ. A potential cause for such unsuccessful transfer is that all four languages in MCWQ belong to different language families and have low linguistic similarities. It remains to be investigated whether such cross-lingual transfer will be more effective on related languages, such as from English to German (Lin et al., 2019).

2.6 Analysis

2.6.1 Evaluation with Gold Translation

Most existing compositional generalization datasets focus on SP (Lake and Baroni, 2018; Kim and Linzen, 2020; Keysers et al., 2020). These datasets are composed either with artificial language or in English using grammar rules. With *test-intersection-gold* proposed in §2.4.2, we investigate whether models can generalize from a synthetic automatically translated dataset to a manually translated dataset.

We use the monolingual models trained on three MCD splits to parse *test-intersection-gold*. In Table 2.6, we present the mean BLEU scores and exact match accuracy of the predicted SPARQL queries. There is no substantial difference between the performances on the two intersection sets, except for Kannada, which has a 4% accuracy drop on average. These results testify that MCWQ has sufficiently high translation quality and that models trained with such synthetic data can be used to generalize to high-quality manually-translated questions.

2.6.2 Categorizing Errors

In an empirical analysis, we categorize typical prediction errors on *test-intersection-gold* and *test-intersection-MT* into six types: missing property, extra property, wrong property (where the two property sets have the same numbers of properties, but the elements do not match), missing entity, extra entity and wrong entity (again, same number of entities but different entity sets). We plot the mean number of errors per category, as well as the number of predictions with multiple errors, in Figure 2.5 for monolingual `mT5-small` models. Overall, model predictions tend to have more missing properties and entities than extra ones. Different languages, however, vary in error types. For example, on Hebrew, models make more missing property/entity errors than other languages; but on Kannada they make more extra property/entity errors than the others. About 70 out of the 155 examples contain multiple errors for all languages, with Kannada being slightly more.

Comparing errors on *test-intersection-gold* and *test-intersection-MT*, we find missing properties are more common in *gold* for all languages. For Hebrew and Kannada, extra properties and entities are also more common in *gold*. However, for Chinese, these and missing entities are less common in *gold*

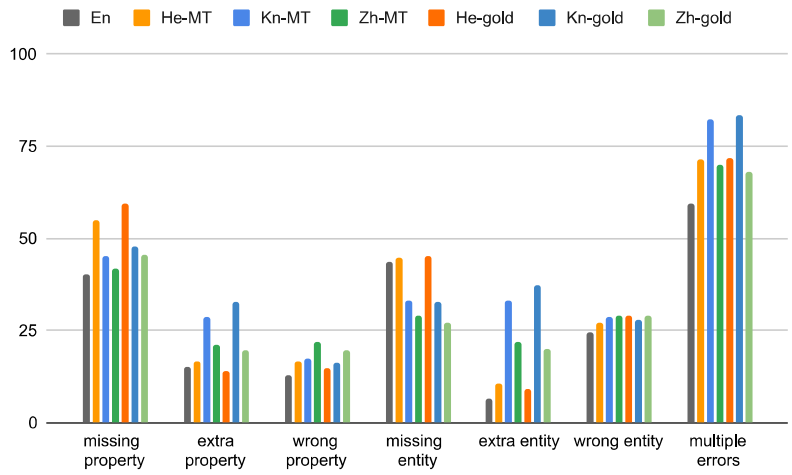


Figure 2.5: Number of errors per category in different SPARQL predictions on *test-intersection-MT* and *test-intersection-gold*, averaged across monolingual `mT5-small`+RIR models trained on the three MCD splits. The total number of items in each test set is 155.

compared to *MT*.

In Figure 2.6 we plot the error statistics for zero-shot cross-lingual transfer using `mT5-small` models. We can see there are drastically more error occurrences. For both missing and extra property/entity, the numbers are about double those from monolingual experiments. The number of wrong property/entity errors remain similar, due to the difficulty of even predicting a set of the correct size in this setting. For all three target languages, nearly all predictions contain multiple errors. The statistics indicate the variety and pervasiveness of errors.

2.6.3 Other Observations

We also find that comparatively, parsers perform well on short questions on all four languages. This is expected as the compositionality of these questions is inherently low. On languages other than English, the models perform well when the translations are faithful. On occasions when they are less faithful or fluent but still generate correct queries, we hypothesize that translation acts as data regularizers, especially at higher complexities, as demonstrated

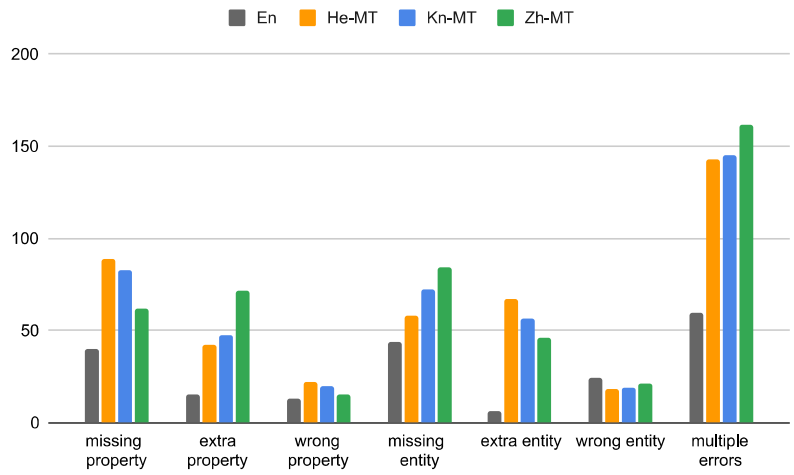


Figure 2.6: Number of errors per category in different zero-shot cross-lingual SPARQL predictions on *test-intersection-MT*, averaged across `mT5-small`+RIR models trained on the three MCD splits in English. Additionally, mean error counts on the English set are given for comparison. The total number of items in each test set is 155.

in Figure 2.4.

Among wrong entity errors, the most common cause across languages is the shuffling of entity placeholders. In the example shown in Figure 2.7, we see that the model generates `M1 wdt:P57 M2` instead of `M0 wdt:P57 M2`, which indicates incorrect predicate-argument structure interpretation.

2.7 Related Work

Compositional Generalization Compositional generalization has witnessed great developments in recent years. SCAN (Lake and Baroni, 2018), a synthetic dataset consisting of natural language and command pairs, is an early dataset designed to systematically evaluate neural networks' generalization ability. CFQ and COGS are two more realistic benchmarks following SCAN. There are various approaches developed to enhance compositional generalization, for example, by using hierarchical poset decoding (Guo et al., 2020), combining relevant queries (Das et al., 2021) using span representation (Herzig and Berant, 2021) and graph encoding (Gai et al., 2021). In addition

Question	Was M0 written by and directed by M1 , M2 , and M3
Gold	ASK WHERE { M0 wdt:P57 M1 . M0 wdt:P57 M2 . M0 wdt:P57 M3 . M0 wdt:P58 M1 . M0 wdt:P58 M2 . M0 wdt:P58 M3 }
Inferred	ASK WHERE { M0 wdt:P57 M1 . M1 wdt:P57 M2 . M0 wdt:P58 M3 }

Figure 2.7: Example for an error reflecting incorrect predicate-argument structure. `wdt:P57` is *director* and `wdt:P58` is *screenwriter*. Incorrect triples are shown in red and missed triples in blue.

to pure language, the evaluation of compositional generalization has been expanded to image captioning and situated language understanding (Nikolaus et al., 2019; Ruis et al., 2020). Multilingual and cross-lingual compositional generalization is an important and challenging field to which our paper aims to bring researchers' attention.

Knowledge Base Question Answering Comparing to machine reading comprehension (Rajpurkar et al., 2016; Joshi et al., 2017; Shao et al., 2018; Dua et al., 2019; d'Hoffschmidt et al., 2020), KBQA is less diverse in terms of datasets. Datasets such as WebQuestions (Berant et al., 2013), SimpleQuestions (Bordes et al., 2015), ComplexWebQuestions (Talmor and Berant, 2018), FreebaseQA (Jiang et al., 2019), GrailQA (Gu et al., 2021),CFQ and *CFQ (Tsarkov et al., 2021) were proposed on Freebase, a now-discontinued KB. SimpleQuestions2Wikidata (Diefenbach et al., 2017) and ComplexSequentialQuestions (Saha et al., 2018) are based on Wikidata, but like most others, they are monolingual English datasets. Related to our work is RuBQ (Korablinov and Braslavski, 2020; Rybin et al., 2021), an English-Russian dataset for KBQA over Wikidata. While the dataset is bilingual, it uses crowdsourced questions and is not designed for compositionality analysis. Recently, Thorne et al. (2021) proposed WIKINLDB, a Wikidata-based English KBQA dataset, focusing on scalability rather than compositionality. Other related datasets include QALM (Kaffee et al., 2019), a dataset for multilingual question answering over a set of different popular knowledge graphs, intended to help determine the multilinguality of those knowledge graphs. Similarly, QALD-9 (Ngomo, 2018) and QALD-9-plus (Perevalov et al., 2022a) support the development of multilingual question answering systems, tied to DBpedia and Wikidata, respectively. The goal of both datasets is to expand QA systems to more languages rather than improving compositionality. KQA Pro (Cao et al., 2022), a concurrent work to us, is an English KBQA dataset over

Wikidata with a focus on compositional reasoning.

Wikidata has been leveraged across many NLP tasks such as coreference resolution (Aralikatte et al., 2019), frame-semantic parsing (Sas et al., 2020), entity linking (Kannan Ravi et al., 2021) and named entity recognition (Nie et al., 2021). As for KBQA, the full potential of Wikidata is yet to be explored.

Multilingual and Cross-lingual Modelling Benchmarks such as XGLUE (Liang et al., 2020) and XTREME (Hu et al., 2020) focus on multilingual classification and generation tasks. Cross-lingual learning has been studied across multiple fields, such as sentiment analysis (Abdalla and Hirst, 2017), document classification (Dong and de Melo, 2019), POS tagging (Kim et al., 2017) and syntactic parsing (Rasooli and Collins, 2017). In recent years, multilingual PLMs have been a primary tool for extending NLP applications to low-resource languages, as these models ameliorate the need to train individual models for each language, for which less data may be available. Several works have attempted to explore the limitations of such models in terms of practical usability for low-resource languages (Wu and Dredze, 2020), and also the underlying elements that make cross-lingual transfer learning viable (Dufter and Schütze, 2020). Beyond these PLMs, other works focus on improving cross-lingual learning by making particular changes to the encoder-decoder architecture, such as adding adapters to attune to specific information (Artetxe et al., 2020b; Pfeiffer et al., 2020).

For cross-lingual SP, Sherborne and Lapata (2022) explored zero-shot SP by aligning latent representations. Zero-shot cross-lingual SP has also been studied in dialogue modelling (Nicosia et al., 2021). Yang et al. (2021) present augmentation methods for Discourse Representation Theory (Liu et al., 2021c). Oepen et al. (2020) explore cross-framework and cross-lingual SP for meaning representations. To the best of our knowledge, our work is the first on studying cross-lingual transfer learning in KBQA.

2.8 Limitations

MCWQ is based on CFQ, a rule-base generated dataset, and hence the inherited unnaturalness in question-query pairs of high complexity. Secondly, we use machine translation to make MCWQ multilingual. Although this is the dominant approach for generating multilingual datasets (Ruder et al.,

2021) and we have provided evidences that MCWQ has reasonable translation accuracy and fluency with human evaluation and comparative experiments in §2.4.3 and §2.5.1, machine translation would nevertheless create substandard translation artifacts (Artetxe et al., 2020a). One alternative way is to write rules for template translation. The amount of work can possibly be reduced by refering to a recent work (Goodwin et al., 2021) in which English rules are provided for syntactic dependency parsing on CFQ's question fields.

Furthermore, the assumption that an English KB is a "canonical" conceptualization is unjustified, as speakers of other languages may know and care about other entities and relationships (Liu et al., 2021a; Hershcovich et al., 2022). Therefore, future work must create multilingual SP datasets by sourcing questions from native speakers rather than translating them.

2.9 Conclusion

The field of KBQA has been saturated with work on English, due to both the inherent challenges of translating datasets and the reliance on English-only DBs. In this work, we presented a method for migrating the existing CFQ dataset to Wikidata and created a challenging multilingual dataset, MCWQ, targeting compositional generalization in multilingual and cross-lingual SP. In our experiments, we observe that pretrained multilingual language models struggle to transfer and generalize compositionally across languages. Our dataset will facilitate building robust multilingual semantic parsers by serving as a benchmark for evaluation of cross-lingual compositional generalization.

Chapter 3

Generalized Quantifiers as a Source of Error in Multilingual NLU Benchmarks

Abstract

Logical approaches to representing language have developed and evaluated computational models of quantifier words since the 19th century, but today's NLU models still struggle to capture their semantics. We rely on Generalized Quantifier Theory for language-independent representations of the semantics of quantifier words, to quantify their contribution to the errors of NLU models. We find that quantifiers are pervasive in NLU benchmarks, and their occurrence at test time is associated with performance drops. Multilingual models also exhibit unsatisfying quantifier reasoning abilities, but not necessarily worse for non-English languages. To facilitate directly-targeted probing, we present an adversarial generalized quantifier NLI task (GQNLI) and show that pre-trained language models have a clear lack of robustness in generalized quantifier reasoning.

3.1 Introduction

Quantifier words—such as *each* or *most* or *more than three*—have been extensively studied, both in logic and in linguistics (Westerståhl, 1989; Peters and Westerståhl, 2006), going all the way back to Frege (1879). In this paper, we examine the extent to which they present a challenge to modern NLU systems. Our analysis is motivated by three observations:

Quantifier words are abstract Unlike nouns, verbs and adjectives, quantifier words do not have referents out in the world. Rather, quantifier words specify relationships between sets of entities, events and properties. To provide intuitions about the semantics of quantifier words, and to be able to refer to quantifiers in a language-independent way, we rely on the notion of generalized quantifiers (Mostowski, 1957), as described in §2.

Quantifier words vary across languages Quantifier word inventories differ across languages. Often what is considered rough translation equivalents also differ in syntax, fine-grained semantics or pragmatics. Stateva et al. (2019) show, e.g., that perceptions of the numerical bounds of existential quantifiers differ across speakers of English, French, Slovenian, and German. Other papers showing discrepancies between quantifier systems include comparisons of Salish to English (Matthewson, 2001), Adyghe to English (Nikolaeva, 2012), or of Dutch, Hebrew and Bengali (Gil, 1982). The cross-linguistic differences in how generalized quantifiers are expressed motivates a cross-lingual error

QA_English	CONTEXT: *A piece of paper was later found on which he had written his last statements in **two** languages, Latin and German. Only **one** statement was in Latin and the rest in German.* QUESTION: *In what language were **most** statements written?* ANSWER: `German` PREDICTED ANSWER: `Latin and German`
NLI_Spanish	PREMISE: ***Más de tres** personas resultaron heridas en un accidente de **dos** vehículos el lunes por la noche. (translation: **More than three** people were injured in a **two**-vehicle crash Monday evening.)* HYPOTHESIS: *Había **4** personas involucradas. (translation: There were **4** people involved.* LABEL: `Neutral` PREDICTED LABEL: `Entailment`

Table 3.1: Examples of quantifiers (marked in bold texts) in NLP tasks, with RoBERTa's prediction for QA and XLM-R's prediction for NLI after fine-tuning.

analysis, since quantifiers may contribute more to error when processing some languages rather than others.

Quantifier words are important Quantifier words are extremely important for tasks that require inference, including natural language inference, question answering, fact-checking, etc. Datasets have, for example, been developed for numerical reasoning in English (Dua et al., 2019). Several researchers have identified quantifier words as important sources of errors for natural language processing systems (Joshi et al., 2020a); see Table 3.1 for examples of such errors. Unfortunately, most efforts have concentrated on subsets of quantifier words and on English.

Contributions We analyze how quantifiers are represented in NLU benchmarks, and how their occurrence at test time contributes to errors by neural language models (LMs). We derive a linguistically motivated 11-way categorization set for generalized quantifiers and look into their distribution in three steps: (a) monolingual NLI; (b) cross-lingual NLI; (c) cross-lingual question answering. We also propose GQNLI[1], an adversarial generalized quantifier NLI challenge dataset. Our work shows that (i) generalized quantifiers are pervasive and cause overall performance drops in NLU benchmarks; (ii) the contribution of quantifier words to system error varies across languages; and (iii) generalized quantifiers are particularly difficult for LMs in interaction with negation and subsumption.

[1] `https://github.com/ruixiangcui/GQNLI`

Generalized Quantifiers	**Logical Denotation**	**Example**
some(A)(B) = 1	$A \cap B \neq \varnothing$	This process is known to increase security in several ways.
all(A)(B) = 1	$A \subseteq B$	Everyone agreed the food was terrible.
more than k the(A)(B) = 1	$\|A \cap B\| > k$	They do let them go more than twice a week.
less than k the(A)(B) = 1	$\|A \cap B\| < k$	San Augustin Acolman has less than 1,000 residents.
k (A)(B) = 1	$\|A \cap B\| = k$	Please donate 100 million to the School of Nursing.
between p and k the(A)(B) = 1	$p < \|A \cap B\| < k$	The USA added ten states to its nation between 1800 and 1850.
the **p/k** (A)(B) = 1	$\|A \cap B\| = p \cdot (\|A\|/k)$	Captain Blood has 20/20 vision.
the **k%** (A)(B) = 1	$\|A \cap B\| = k \cdot (\|A\|/100)$	The lending fund is always guaranteed 9% interest.
most (A)(B) = 1	$\|A \cap B\| > \|A \backslash B\|$	Most ZIP Codes cover roughly ten thousand addresses.
few (A)(B) = 1	$\|A \cap B\| < \|A \backslash B\|$	Only a few teenagers were still listening to Rock 'n' Roll.
each other (A)(B) = 1	$\forall a \in (A \cap B) \exists b \in (A \cap B)(a \neq b)$	All of these trails are located within the a one hour drive of each other.

Table 3.2: The categorization set of quantifiers for task analysis. The first six are Aristotelian/counting quantifiers and the following four are proportional quantifiers. The last one is a Ramsey quantifier (Schmerl and Simpson, 1982). For each quantifier, its logical denotation is listed in the second column. The third conlumn contains English examples with quantifiers taken from XNLI.

3.2 Background

Generalized quantifiers (GQs) are developed upon first-order predicate logic, denoting relations between sets (Mostowski, 1957). Given a universe E, a quantifier Q would be treated as a mapping Q_E from the Cartesian product of powersets $\mathcal{P}(E) \times \mathcal{P}(E)$ to the set $\{false, true\}$ or, as a binary relation on subsets of E (Dvořák and Holčapek, 2015). GQs are generalizations of the $\forall$,$\exists$ quantifiers from first-order predicate logic (Mostowski, 1957; Lindström, 1966; Montague, 1973; Bach et al., 1995; Keenan and Paperno, 2012). A generalized quantifier is, abstractly, a relation between sets. Generalized quantifier theory, while developed by logicians, is used by formal linguists to analyze the meaning of quantifier words in combination with referential expressions (Barwise and Cooper, 1981; Higginbotham and May, 1981).

Most human languages contain ways of expressing generalized quantifiers, and their semantics exhibit striking similarities across languages (Matthewson, 2004; Fintel and Matthewson, 2008; Steinert-Threlkeld, 2019). At the same time, generalized quantifiers can be instantiated very differently across languages due to pragmatic considerations (Grice, 1989) or cognitive economy and cost-benefit optimisation in the exchange of information (Levinson et al., 2000; Steinert-Threlkeld, 2021; Uegaki, 2022). Quantifier words also exhibit syntactic differences, e.g., with some languages having specialized words to express quantity, while others rely on metaphorical usage of common nouns (Katsos et al., 2012). In English, *most* is a determiner, but Spanish and French express the same concept through common nouns, *la mayoría* and *la majorité*. The relative stability of the core semantics of quantifiers makes a cross-linguistic comparison possible, but the syntactic and pragmatic variation associated with the expression of generalized quantifiers poses a challenge for multilingual NLU. We consult quantifier taxonomy studies (Keenan and Westerståhl, 1997; Peters and Westerståhl, 2006; Szymanik and Thorne, 2015; Szymanik, 2016) and derive a categorization set for quantifier analysis in NLU benchmarks. In Table 3.2, we list the 11-way quantifier categorization set and their logical denotation based on set theory.

While other foci of formal linguistics have attracted the attention of NLP researchers—including coreference (Ogrodniczuk et al., 2019, 2020), negation (Hossain et al., 2020; Hartmann et al., 2021), and consistency (Li et al., 2019; Ribeiro et al., 2019; Asai and Hajishirzi, 2020; Geva et al., 2022)—there has been little work on generalized quantifiers as a source of error in NLU, let alone in multilingual NLU. It remains an open problem whether LMs represent the semantics of quantifiers words adequately, or if they provide a basis for resolving scopal ambiguities.[2]

3.3 NLU Benchmarks

We conduct an error analysis focusing on the role of generalized quantifiers in two NLU tasks, Natural Language Inference (NLI) and Question Answering (QA), which generally require understanding of quantifiers. For each type

[2]Note that generalized quantifiers are not always *explicit* in discourse. The sentence *inadequate sleep causes obesity* should be interpreted as *Most of those who do not sleep adequately, gain weight* (Zadeh, 1983). Such implicit quantifiers related to pragmatic variation are important for language understanding, but will be ignored in this work.

of task, both monolingual and cross-lingual evaluation are conducted. We focus on generalized quantifiers in the *hypotheses* in NLI examples—and on generalized quantifiers in the *question* fields in question answering. To

Quantifier	English						Cross-lingual
	MNLI_m	MNLI_mm	SNLI	ANLI_R1	ANLI_R2	ANLI_R3	XNLI
some	171	132	191	5	1	17	115
all	255	239	65	15	8	29	166
$> k$	14	23	8	10	16	14	16
$< k$	3	3	0	6	7	5	1
k	266	269	988	55	62	48	159
between	2	3	0	3	2	0	1
p/k	1	5	1	1	1	0	2
$k\%$	10	7	0	0	0	1	5
most	35	39	1	0	2	1	9
few	14	15	11	0	0	6	11
each other	4	3	35	0	0	2	5
Total	775	738	1300	95	99	124	499
Frequency	7.9%	7.5%	13.2%	9.5%	9.9%	12.4%	10.0%

Table 3.3: Quantifier distribution in four NLI tasks, among which three are monolingual English and one is cross-lingual. The table show statistics of the test set, if not available, dev set, of the target task. All but the last rows show the occurrence time of the type of quantifier in the first column. The last row represents the distribution rate of any quantifier in the dataset.

this end, we identify quantifiers by the lemma and the universal dependency relation (Nivre et al., 2020) of a quantifier after preprocessing the sentences using *Stanza* (Qi et al., 2020). Take the sentence "The Yiddish culture has survived for more than a thousand years.", we annotate it as "The/*det* Yiddish/*amod* culture/*nsubj* have/*aux* survive/*root* for/*case* more/*advmod* than/*fixed* a/*det* thousand/*nummod* year/*obl* ./*punct*". By matching the regex pattern of the quantifier "more than k", in this case *"((more|great)\/advmod than\/(fixed|case)|at\/case least\/nmod) .+\/nummod .+\/(nsubj|obj|obl)"*, we approximate the surface form of the type "more than k".Through matching quantifier patterns, we are able to find entries in which quantifiers are

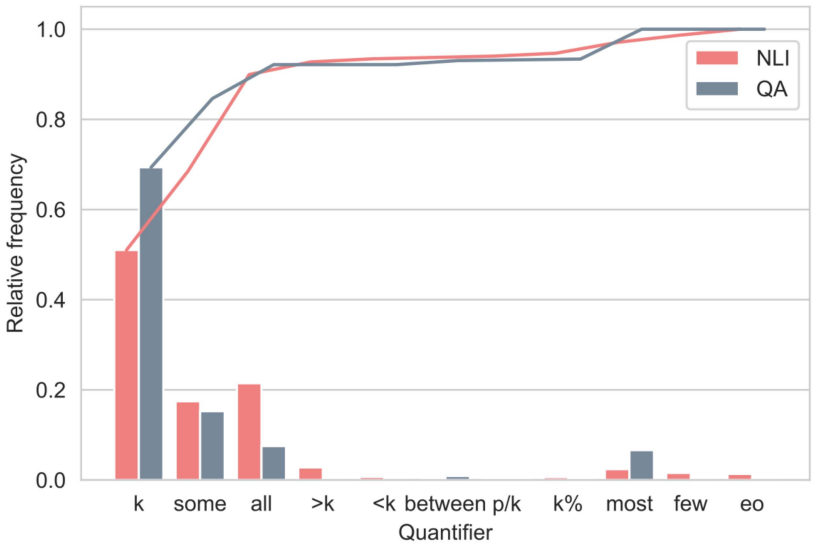


Figure 3.1: Relative distribution of quantifiers in NLI and QA tasks ranked by semantic complexity. The bars show the relative frequency of such quantifier and the lines indicate the cumulative frequency for a task.

instantiated. See Appendix 3.A for the list of regex patterns we write to identify GQs. In Table 3.3 and Table 3.6, we present the statistics of the quantifier distributions in NLI and QA tasks, respectively. As can be seen, quantifiers are indeed widespread in NLU tasks, accounting for roughly 10% in NLI tasks and 5% in QA tasks. We will further discuss the statistics and experiments in the following section.

3.4 Quantifiers in English NLI Benchmarks

NLI is commonly framed as a three-way classification task with labels *entailment*, *contradiction* and *neutral* (Bowman et al., 2015a). While SOTA models exhibit low error rates on NLI benchmarks, it is unclear when they succeed or fail in their underlying reasoning. We are interested in whether generalized quantifers challenge modern NLI models. In our error analysis, we initially focus on three English NLI datasets, MultiNLI (MNLI; Williams et al., 2018), SNLI (Bowman et al., 2015a) and ANLI (Nie et al., 2020) as testbeds.

Quantifier	BERT							RoBERTa						
	M_m	M_mm	SNLI	A_R1	A_R2	A_R3	*weig.*	M_m	M_mm	SNLI	A_R1	A_R2	A_R3	*weig.*
some	**82.5**	**84.1**	**86.9**	100	**0**	47.1	**83.4**	**83**	**84.8**	**86.9**	100	100	**41.1**	**83.7**
all	**85.9**	88.3	**89.2**	**46.7**	**37.5**	**34.5**	**83.2**	**85.9**	92.1	**90.8**	**66.7**	**37.5**	**34.5**	**85.3**
$> k$	**85.7**	100	**87.5**	70	**43.8**	**42.9**	**73**	**85.7**	91.3	**87.5**	80	**37.5**	**28.5**	**68.2**
$< k$	100	100		**33.3**	57.1	80	**66.7**	100	100		83.3	85.7	100	91.7
k	87.2	**81.8**	92.4	**43.6**	**43.5**	**33.3**	84.8	88.3	**88.8**	92.9	**56.3**	61.3	**43.8**	87.8
between	100	100		66.7	50		**80**	100	**66.7**		**66.7**	50		**70**
p/k	100	**60**	100	100	100		**77.8**	100	**80**	100	100	**0**		**77.8**
$k\%$	90	100				100	94.4	**70**	**85.7**				**0**	**72.2**
most	**74.3**	**79.5**	**0**		50	**0**	**74.4**	**77**	**87.2**	100		59	**0**	**80.9**
few	**78.6**	**73.3**	**90.9**			**33.3**	**73.9**	**85.7**	**80**	**90.9**			**33.3**	**78.3**
each other	**75**	100	**85.7**			50	**84.1**	**50**	100	**88.6**			50	**84.1**
all GQs	**85**	**84.8**	**91.2**	**50.5**	**44.4**	**39**	**83.3**	**85.4**	**88.8**	**91.7**	**65.3**	56.5	**40.3**	**85.5**
full	86.5	86.1	91.3	58.6	48	43.2	84.4	89.5	89.4	92.3	71.7	49.6	49	87.3

Table 3.4: BERT and RoBERTa performance on NLI tasks. The *weig.* column represents the percentage of all true predictions in six subtasks over total instances. The penultimate row stands for the overall performance when quantifiers exist in a dataset. The last row reports the overall performance in a dataset. Number marked in bold signifies a lower score than the overall performance.

Table 3.3 presents statistics of quantifier distribution in these datasets, where we observe that, across, about 10% of all hypotheses contain quantifier words, indicating the pervasiveness of quantification. We also plot the frequency of quantifiers in NLI in Figure 3.1 and find the quantifier word distribution follows Zipf's law (Zipf, 1949). Note the top three most common quantifiers account for more than 90% of all.

Experiments and Results In order to investigate whether NLU systems can solve quantifiers in NLI, we experiment with two pretrained LMs: BERT[3] (Devlin et al., 2019) and RoBERTa[4] (Liu et al., 2019). We use the codebase by Nie et al. (2020). The training data combines SNLI, MNLI, FEVER-NLI (Nie et al., 2019) and ANLI.

In Table 3.4, we report the test set performance on SNLI and ANLI, and the dev set performance on MLNI *matched* and *mismatched* sections. We can observe that SOTA models suffer from performance drops across almost all quantification phenomena in every task. When it comes to performance

[3] `wwm_cased_L-24_H-1024_A-16`

[4] `roberta-large`

over all quantifiers, the improvement from RoBERTa to BERT (2.2%) is less prominent than that over full datasets (2.9%), suggesting RoBERTa is particularly challenged.

Taking a closer look at error by category, proportional quantifiers seem harder to solve than Aristotelian/counting quantifiers. Except for *k%*, all proportional quantifiers—*p/k*, *most*, and *few*—are about 10% lower than the five counting quantifiers (except *less than k*) with BERT; and about 5% lower with RoBERTa. RoBERTa is not generally superior to BERT; e.g., for *k%*, BERT outperforms it by 22%. We show a pairwise analysis of how GQs affect performance when they appear in both the premises and hypotheses in the Appendix 3.B. Generally, our results attest to the difficulty of resolving GQs in NLI benchmarks.

3.5 Quantifiers in Cross-lingual NLU Benchmarks

Quantifiers are acquired in similar orders across languages (Katsos et al., 2016), although languages express quantifiers in different ways. For example, there are eight different universal quantifiers with different level of distributivity in Malagasy (Matthewson, 2008). This poses challenges to training multilingual LMs and transfer learning. We are interested in whether quantifiers are universally and evenly challenging for all languages.

Quantifiers in Cross-lingual NLI We choose XNLI (Conneau et al., 2018), a manual translation of the development and test set of MNLI into 15 languages, for this multilingual error analysis. We should clarify that for XNLI, the authors annotate entailment labels for the English data only and apply them to the other languages. We do not assume label changes due to translation in this study, but it is worth investigate in the future. We choose five languages belonging to different language families, namely Arabic, Chinese, German, Spanish and Vietnamese as targets. The last column in Table 3.3 shows the numbers of quantifiers in XNLI. The distribution rate is 10%. Note that the universal quantifier is the most common quantifier in XNLI.

Quantifier	mBERT							XLM						
	en	zh	es	ar	vi	de	*weig.*	en	zh	es	ar	vi	de	*weig.*
some	85.2	**69.6**	80	**63.5**	**67.8**	74.8	73.4	85.2	70.3	79.1	71.3	73.9	**69.6**	**69.6**
all	**80.1**	**65.7**	**72.8**	69.3	**63.9**	74.1	**70.9**	**82.5**	**62.7**	74.1	67.5	71.7	73.5	**72**
$> k$	87.5	**50**	**68.8**	**43.8**	**56.2**	**62.5**	**61.6**	**81.2**	**62.5**	**56.2**	**62.5**	**50**	75	75
$< k$	100	100	100	100	100	100	100	100	100	100	100	100	100	100
k	86.2	**69.1**	80.5	71.7	76.7	82.4	77.7	83	66.7	78.6	71.7	74.2	81.1	75.8
between	100	100	100	100	100	100	100	100	100	100	100	100	100	100
p/k	100	**50**	100	100	100	100	91.7	100	**0**	100	100	**50**	**50**	**66.7**
$k\%$	100	100	80	100	100	100	96.7	**80**	80	80	100	100	80	86.7
most	**55.6**	**55.6**	**66.7**	66.7	**33.3**	66.7	**57.4**	**55.6**	**33.3**	**66.7**	**55.6**	**44.4**	77.8	**55.6**
few	**72.7**	**54.5**	**72.7**	**63.6**	**45.5**	72.7	**63.6**	**63.6**	**36.4**	**54.5**	**63.6**	**54.5**	72.7	**57.5**
each other	**60**	**60**	**60**	**60**	80	80	**66.7**	**80**	**20**	**60**	**20**	**40**	**60**	**46.7**
all GQs	83	**67.1**	76.7	68.1	**68.3**	76.9	73.3	**82.4**	**64.2**	**75.7**	69.3	**71.4**	74.8	73
full	82.6	88.9	74.7	65.6	70.7	71.4	72.4	83.1	64.8	76.3	66.9	71.6	71.3	72.3

Table 3.5: Results of mBERT and XLM performance on XNLI tasks decomposed by quantifier categories.

We fine-tune mBERT[5] (Devlin et al., 2019) and XLM[6] (Lample and Conneau, 2019) on the MNLI training set and evaluate them on XNLI. We report the results in Table 3.5. We find that performance varies across languages. For Chinese and Vietnamese, we see significant drops in performance for examples with GQs, whereas for Arabic and German, we see improvements. The results *per* quantifier are more homogeneous, however.

Similar to our results for English, we can see that the lowest accuracies in XNLI are with proportional quantifiers, such as *most* and *few*. But the gap in non-English languages is wider for these two categories, especially for Chinese, the difference reaches 30%. Other hard quantifiers include *all*, $> k$, $< k$, and *each other*.

Quantifiers in Cross-lingual QA Cross-lingual question answering (XQA) is another important NLU task that evaluates the cross-lingual transferability of LMs. We evaluate the effect of quantifiers on system errors across two XQA datasets, namely XQuAD (Artetxe et al., 2020b) and MLQA (Lewis et al., 2020). As demonstrated in Figure 3.1, quantifier word distributions in XQA tasks also follow Zipf's law, as in NLI tasks, but k is more frequent (perhaps because of a traditional emphasis on numerical reasoning), and we see less variance across languages. This is probably because question

[5] `multi_cased_L-12_H-768_A-12`

[6] `xlm-mlm-100-1280`

Quantifier	MLQA						XQuAD
	en	zh	es	ar	vi	de	*all_lang*
some	66	39	41	44	37	33	12
all	31	14	26	21	19	16	7
$< k$	1	0	0	0	1	0	0
k	322	168	166	195	204	149	32
between	4	2	2	2	3	0	3
p/k	1	1	1	0	0	0	0
$k\%$	1	1	0	1	0	0	0
most	27	19	11	30	17	9	5
Total	453	244	247	293	281	207	59
Frequency	3.9%	4.7%	4.7%	5.4%	5.1%	4.5%	5.0%

Table 3.6: Quantifier distribution in two multilingual QA tasks, MLQA and XQuAD. We choose six common languages apprearing in both tasks to facilitate comparisons. XQuAD is strictly parellel while MLQA is not, hence only the latter has statistics by languges. Categories that no entry exists are omitted.

answering is targeting quantification less directly. To evaluate cross-lingual QA performance on GQs, we fine-tune mBERT and XLM-R[7] (Conneau et al., 2020) using Hu et al. (2020)'s architecture. We present results for mBERT in Table 3.7; for XLM-R results, please refer to Appendix 3.D.

Just as with XNLI, LMs suffer from performance drops across all languages for almost all GQ phenomena with significant, cross-lingual variation. The most distinguished is that Exact Match (EM) suffers from a greater deterioration than F1 scores for all languages. For example, the weighted EM difference for mBERT on MLQA is 2.9% while the weighted F1 is 1%. As one example in Table 3.1, we observe that the plausible answers selected by models, while being incorrect, result in a sharper decrease of EMs comparing to F1s. Questions containing GQs also tend to have less verbal answers comparing to those without GQs, and therefore require higher precision.

Regarding cross-lingual comparisons, Chinese and Arabic are the two languages that do not have lower performance over GQs compared to the performance over the complete dataset. Despite the overall trends, subtle differences from XNLI performance still exist. For example, XLM-R is worse

[7] `xlm-roberta-large`

	XQuAD													MLQA											
Quantifier	en		zh		es		ar		vi		de		*weighted*	en		zh		es		ar		vi		de		*weighted*	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
some	75	84.2	50	**55.5**	58.3	76.1	50	**50**	16.6	42.4	**33.3**	**43.8**	47.2	58.7	**59**	80	**28.2**	**52.1**	**34.1**	**59.2**	36.3	54.9	**5.4**	**24**	**33.3**	58.4	**32.7**	**54.8**
all	**28.5**	**62.2**	**14.2**	**35.2**	**28.5**	82	42.8	**52.3**	**14.2**	**29.4**	**28.5**	**56**	**26.1**	**52.9**	67.7	**79.8**	14.2	**46.4**	**38.4**	**62.8**	33.3	57.9	**10.5**	30.1	**31.2**	**51.6**	**32.6**	**54.8**
$< k$															**0**	**0**							0	**13.3**			**0**	**6.7**
k	78.1	90.1	68.7	80.4	56.2	**72.1**	**40.6**	64.3	**12.5**	35.7	56.2	77.1	52.1	70	74.9	**79.4**	47	63.4	**41.5**	**65.9**	**27.6**	50.3	**6.3**	**23.7**	**38.2**	**53**	39.3	**56**
between	100	100	**33.3**	**72.2**	66.6	93.3	100	100	**0**	**19**	**0**	**56.5**	50	73.5	**50**	88.5	50	83.3	**0**	**26.6**	**0**	68.7	**0**	**26.6**			**20**	58.7
p/k															100	100	**0**	**0**	**0**	**0**							**33.3**	**33.3**
$k\%$															100	100	**0**	**26.6**			**0**	**23.7**					**33.3**	**50.1**
most	**40**	**53.3**	**40**	**40**	**0**	**10**	**0**	**26.6**	**0**	**0**	**20**	**49.3**	**16.7**	**29.9**	**55.5**	**76**	47.3	62.1	**45.4**	**61.7**	30	**46.8**	**5.8**	**15.7**	**33.3**	**40.7**	**36.2**	**50.3**
all GQs	**70**	**83.2**	55	66.7	**50**	**70.3**	41.6	58.2	**11.6**	32.5	**43.3**	**65**	**45.3**	62.7	**63.5**	**79.2**	41.8	60.3	**39.6**	**63.7**	**29.3**	51.3	**6.4**	**23.6**	**36.1**	**53.2**	**36.1**	**55.2**
full	71.8	83.7	48	59.1	56	74.5	40.8	57.9	13.9	32.4	50.7	67.2	46.9	62.5	67.2	80.6	37.5	57.9	47.3	66	30	48.4	11.2	28	40.8	56	39	56.2

Table 3.7: Results of mBERT performance on XQA tasks decomposed by quantifier categories.

than mBERT on quantifier reasoning on XQuAD Chinese, especially at proportional quantifiers, but this is not the case on MLQA Chinese.

3.6 GQNLI

We have seen how quantifiers present challenges to NLI and QA models. Using an approach similar to ANLI (Nie et al., 2020) and DynaBench (Kiela et al., 2021), we use model difficulty (RoBERTa's) as a heuristic to select hard examples for a challenge dataset that can hopefully be used to evaluate any future progress on this. We propose GQNLI, a generalized quantifier NLI challenge dataset, consisting of 30 premises and 300 hypotheses. The average sentence lengths of hypothesis and premises are 15.97 and 7.35, respectively. Both numbers are comparable to those of MNLI, but lower than ANLI's (Williams et al., 2022). It should be noted that GQNLI is designed for evaluating future models; obviously not for benchmarking RoBERTa.

Dataset Creation Firstly, we manually create 100 premise-hypothesis pairs, in which various types of GQs appear. For each premise and hypothesis, the number of GQs varies from one to three. To choose the premises, we randomly sampled 100 premises with GQs from SNLI and ANLI test sets, respectively, and selected 10 premises in total, that we consider are semantically adequate for adding GQs and making simple hypotheses.

To construct the hypotheses, we rely on RoBERTa fine-tuned on MNLI and manually select examples about which the model is unsure or incorrect. To focus on GQs, we keep the challenge examples otherwise simple (Ribeiro et al., 2020), and avoid lexical variations in the hypotheses. Hard examples were found to be characterized by (i) mixing generalized quantifiers with other logical operators, such as subsumption or negation, and (ii) combining

Quantifier		some	all	$> k$	$< k$	k	between	p/k	$k\%$	most	few	each other	*Overall*
# Occurrence		27	51	51	33	170	21	24	45	18	9	36	485
Model	**Training Data**	**% Performance**											
BERT	S,M,F,ANLI	40.7	41.2	33.3	30.3	30.6	14.3	37.5	22.2	61.1	22.2	41.7	30
ELECTRA	S,M,F,ANLI	37.0	17.6	54.9	27.3	38.2	14.3	62.5	31.1	61.1	0.0	16.7	38.0
SBERT	S,M,F,ANLI	66.7	43.1	47.1	24.2	32.4	14.3	25.0	31.1	**77.8**	**66.7**	36.1	39.3
RoBERTa	MNLI	55.6	25.5	17.6	27.3	24.7	23.8	45.8	17.8	33.3	33.3	11.1	28.2
	S,M,F,ANLI	63.0	41.2	41.2	27.3	34.1	28.6	**75.0**	33.3	50.0	33.3	38.9	39.3
ALBERT	S,M,F,ANLI	70.4	45.1	35.3	33.3	36.5	19.0	37.5	37.8	50.0	11.1	36.1	41.7
BART	MNLI	40.7	21.6	**60.8**	36.4	**50.6**	**66.7**	37.5	46.7	27.8	33.3	22.2	41.3
	S,M,F,ANLI	59.3	51.0	35.3	30.3	35.3	19.0	66.7	20.0	50.0	**66.7**	47.2	42.7
DeBERTa-v3	MNLI	48.1	37.3	33.3	33.3	35.9	33.3	41.7	33.3	33.3	33.3	41.7	34.7
	M,F,ANLI	**81.5**	54.9	49.0	33.3	44.7	28.6	50.0	**48.9**	66.7	55.6	44.4	**48.0**
	M,F,Ling,DocNLI	77.8	**70.6**	49.0	**54.5**	44.7	4.8	33.3	42.2	50.0	**66.7**	**58.3**	45.0

Table 3.8: GQNLI statstics and seven types of models' performance with different combinations of training data. The second row shows the occurrence time of the type of GQ in GQNLI. The following rows show models' performance on the dataset. We tested most competitive models fine-tuned for NLI available on Hugging Face. All but ALBERT (`xxlarge`) and DeBERTa-v3 (`base`) are size `large`. S, M, F, Ling, A, DocNLI refer to SNLI, MNLI, Fever-NLI, LingNLI (Parrish et al., 2021), ANLI and DocNLI (Yin et al., 2021), respectively. Numbers in bold represent the highest accuracy in one category. Due to space limitation we provide the link to each model in the Appendix 3.H.

multiple different generalized quantifiers. We discuss these observations in Section 3.7.

Two of the authors annotated the examples. The inter-annotator agreement (Fleiss' kappa) was 0.895, substantially higher than ANLI's (0.672–0.740). It is worth noting that the level of semantic or pragmatic interpretation difference of GQs is reflected in the measurement.

We augmented the examples by substituting non-quantifier words (e.g., replacing "dogs" with "cats") while keeping the labels, to exclude the effect of specific lexical items. The resulting labels are uniformly distributed. Table 3.8 presents GQNLI statistics. Since the dataset is curated to probe the ability to reason with quantifiers, the distribution of generalized quantifiers does not follow Zipf's law; see §3.4. A list of GQNLI examples per category is shown in Appendix 3.E.

Experiments and Results We evaluate seven types of models on GQNLI, fine-tuned with different combinations of NLI datasets. As data creation only relied on RoBERTa and MNLI, nothing prevents that models with different architectures and training data will perform well. They do not, however. The results are shown in Table 3.8.

We see that all models have great difficulty with GQNLI. With more training data, models improve, but the best performance is 48%, less than 15 points above chance level. In general, the counting quantifiers, especially the existential and universal quantifiers, are easier than proportional quantifiers. Particularly, most models struggle with *less than k* and *between*. This is in some contrast with the NLU tasks studied above, where these quantifiers were among the easiest.

We also observe unstable GQ reasoning ability in simple word substitution cases. For instance, it happens for DeBERTa fine-tuned with M, F, Ling, DocNLI that it predicted correctly the contradiction relation between "There are six children standing on top of a yellow mountain. Two thirds wear red tops and one third wear green." and "Between 80% and 90% children do not wear red tops.", but incorrectly when "red" is substituted with "beige" and "green" with "cyan". We are yet to study what kind of cues lead to the instability. Our experiments suggest a lack of testing proportionality reasoning and robustness in existing benchmarks.

3.7 Discussion

Negation The interaction between negation words and quantifiers increases semantic complexity (Partee, 1970; Horn, 2010). We investigate whether this holds for NLI tasks, using negation cue detection to find all cases where a negation word and a quantifier appear in the hypotheses.

We break down the performances on negation of the seven models in Appendix 3.F. As indicated, LMs overall have polarized results for negation cases comparing to the entire dataset. We can see a majority of the models even predicted opposite labels for some GQ categories, with 0% accuracy. BART is no longer the second best model, replaced by RoBERTa. The improvement by training with more data is overall consistent for reasoning over GQs with negation.

For a cross-lingual investigation of the interaction of GQs and negation, we find that in XNLI, the number of cases combining both phenomena is

insufficient: we identified four such cases, involving only the quantifiers "all" and "more than." For English, mBERT predicted two cases successfully. For Chinese, German, Vietnamese and Arabic, one is correct. For Spanish, all are wrongly predicted.

It is evident that NLU models suffer from reasoning difficulties in certain cases when negation interacts with GQs, especially in cross-lingual evaluation. In future work, we are interested in expanding GQNLI to more instances and more languages to facilitate qualitative investigations.

Subsumption In generalized term subsumption languages (TSLs; Yen, 1991; Ali and Shapiro, 1993), a term a subsumes another term b if and only if the extension of a is a superset of the extension of b . Rather than surface number comparison, subsumption reasoning requires knowledge of the relations between supersets and subsets. For example, to decide whether "There are six dogs. Three brown dogs, a black dog and a white dog run along the green grass" entails "One dog sits", LMs should be aware that "six dogs" is a superset of the extension of the "brown dogs", "black dog" and "white dog". Another example in GQNLI is to infer whether "There are twelve singers on a stage, less than half from Argentina and one from Cape Verde" entails "Several singers do not come from Chile".

We annotate 63 cases out of the first 100 in GQNLI requiring subsumption reasoning. We show the statistics and results regarding subsumption in Appendix 3.G. It can be seen that more training data leads to higher accuracies. Especially, DeBERTa fine-tuned with DocNLI, which unifies the two classes "neutral" and "contradict" into a new class "not entail", has a significant improvement on subsumption cases with neutral label. The training bias give an advantage to the model on the subsumption subset, half cases of which are labelled neutral. But such bias has a negative effect on non-subsumption cases; the accuracy drops by 20.2% comparing to the model without training with DocNLI. It is worth investigating whether DocNLI is truly helping subsumption reasoning in future work. Subsumption is a key concept in the study of knowledge representation (Woods, 1991), but is neglected in current NLP research. The fact that LMs struggle to perform subsumption reasoning asserts the necessity to explicit tackle the problem.

3.8 Related Work

We examine the sensitivity of NLU models to generalized quantifiers. These models are designed to induce correlations from large volumes of data, not to reason symbolically with logical quantifiers. Such models have, nevertheless, been probed for logical knowledge.

Mul and Zuidema (2019), for example, show neural networks encode fragments of first-order logic and exhibit zero-shot generalization ability. Evans et al. (2018) present a neural architecture that improves performance on propositional logical inference. Bowman et al. (2015b) also suggest neural networks learn semantic representations for logical inference in natural languages. However, on the same task, Veldhoen and Zuidema (2017) find neural networks fail to do so on a more stringent test. Geiger et al. (2019) also show that neural networks fail to exhibit robust logical inference. Srivastava et al. (2018) use semantic parsers to encode quantifiers and improve zero-shot learning in classification tasks. Haruta et al. (2020) present a system that computes logical inference over GQs and see improvements on two specialized datasets, FraCaS (Cooper et al., 1994) and MED (Yanaka et al., 2019). None of these papers explicitly discussed generalized quantifiers, and all were limited to studying the ability of neural networks to capture the logical semantics of English.

Many studies have instead focused on LMs' ability to capture negation (Gururangan et al., 2018; Naik et al., 2018; Hossain et al., 2020; Ettinger, 2020; Hartmann et al., 2021) or coreference (Ye et al., 2020; Varkel and Globerson, 2020; Abdou et al., 2020). Others have focused on LMs' ability to reason with numbers (Johnson et al., 2020). DROP (Dua et al., 2019), for example, is a question answering dataset designed specifically to probe LMs' ability to count, add and subtract for answering factoid questions. Models have also been tailored for numerical reasoning (Geva et al., 2020; Zhang et al., 2020). Cobbe et al. (2021) proposes to use a verification task during pretraining of LMs to improve their ability to solve math word problems. Others have studied monotonicity inference (Hu et al., 2019; Yanaka et al., 2019, 2020), and Fang and Lou (2021) recently focused on the two quantifier words *part* and *whole* in an error analysis for named entity recognition.

Many NLU benchmarks contain quantifier words, but their influence on performance has not been studied systematically. One exception to this is that generalized quantifiers have been used to generate adversarial examples in the context of numerical reasoning (Naik et al., 2018; Nie et al., 2020). TaxiNLI

(Joshi et al., 2020a), which categorizes 15 types of reasoning abilities, is a dataset drawn from MNLI. In their taxonomy, the Quantifier category only refers to universal and existential quantifiers, *not* to generalized quantifiers, and ditto for Kim et al. (2019). All of the above focused on English, but in an extension to TaxiNLI, K et al. (2021) incorporated quantifiers into the Logic class and found a large cross-lingual transfer gap on LMs.

3.9 Conclusion

Quantifiers lie in the intersection of logic, linguistics and NLP research. It is essential for NLU systems to learn quantifier reasoning. We examined generalized quantifiers in multilingual NLU tasks with regards to their expressiveness and logical reasoning requirement. Our survey and experiments indicate quantifiers are neglected to a degree and cause significant performance drops for neural LMs. To better understand LMs' reasoning abilities, we release GQNLI, a novel generalized quantifier NLI challenge dataset. With the pervasiveness of generalized quantifiers, we stress that more efforts are necessary to investigate: (1) when and why models systematically fail when quantifiers interact with other operators; (2) how to improve cross-lingual transferability of quantifiers; (3) how we can exploit the theoretical results about generalized quantifiers from logic and linguistic studies, so as to improve the logical inference ability of neural LMs.

Acknowledgements

We would like to thank Miryam de Lhoneux, Constanza Fierro, Desmond Elliott and the anonymous reviewers for their valuable feedback.

Appendix

3.A Regular Expressions for Generalized Quantifiers

Table 3.9 lists the regex we use to parse generalized quntifiers in sentences augmented with universal dependency tags. The approach does not find all the generalized quantifiers exhuastively but rather approximates the common distributions.

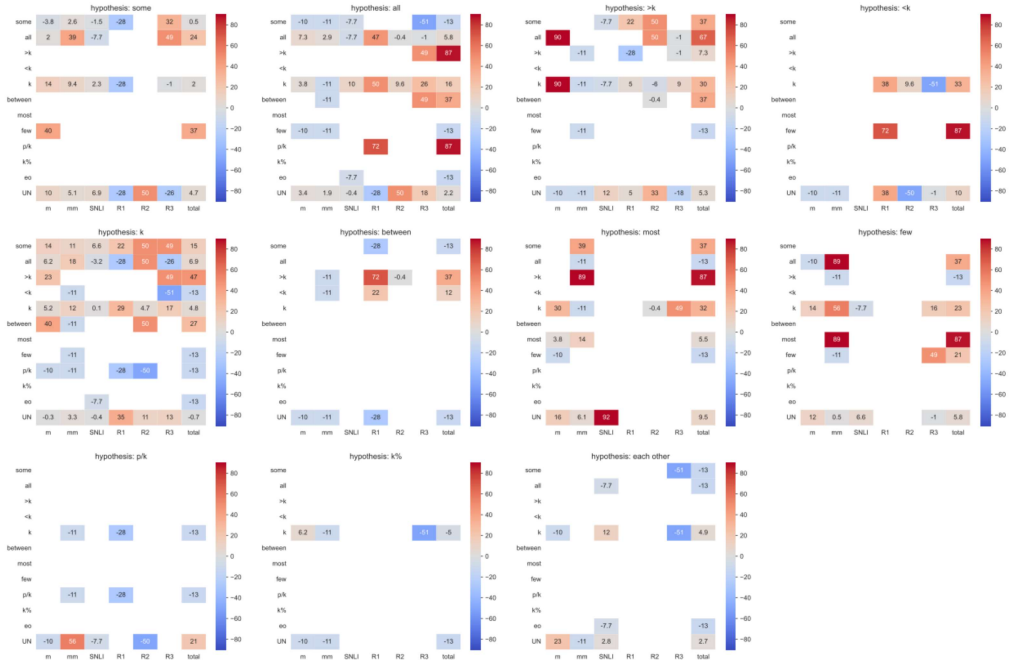


Figure 3.2: Fine-grained analysis of RoBERTa performance on 6 English NLI subtasks. Each heatmap represents hypotheses with a type of quantifier. The rows stand for premises with the quantifier of that label. The numbers are calculated as the accuracy over the whole dataset minus the fine-grained accuracy given a specific premise and hypothesis (the higher the number, the worse the performance). For each heatmap, the last column represents the accuracy gap weighted by all 6 tasks. "UN" stands for an entry where no explicit quantifier is identified.

Generalized Quantifiers	**Regular Expressions**
some(A)(B) = 1	(some|several|much|many)\/det .*\/(nsubj|obj|obl)|(some|several|much|many)\/nsubj|(some|several|much|many)\/amod \w+\/nsubj:pass
all(A)(B) = 1	(every|all|each)\/det .*\/(nsubj|obj|obl)|all\/det:predet .*\/(nsubj|obj|obl) |everything|everyone|everybody
more than k the(A)(B) = 1	((more|great)\/advmod than\/(fixed|case)|at\/case least\/nmod) .+\/nummod .+\/(nsubj|obj|obl)
less than k the(A)(B) = 1	((few|less)\/advmod than\/(fixed|case)|at\/case most\/amod) .+\/nummod .+\/(nsubj|obj|obl)
k (A)(B) = 1	\w+\/nummod .+\/(nsubj|obj|obl)
between p and k the(A)(B) = 1	between\/case \w+\/(nummod|nsubj|obj|obl) and\/cc \w+\/conj|between\/case .+\/(nummod|nsubj|obj|obl) %\/obl
the **p/k** (A)(B) = 1	\d+\/\d+\/(nummod|nsubj|obj|obl)|half\/nummod|third\/(nsubj|obj|obl)|fourth\/(nsubj|obj|obl)|fifth\/(nsubj|obj|obl)
the **k%** (A)(B = 1)	\d+\/nummod %\/(nsubj|obj|obl)
most (A)(B) = 1	most\/amod \w+\/(nsubj|obj|obl)|most\/nsubj:pass of\/case .+\/nmod
few (A)(B) = 1	few\/amod \w+\/(nsubj|obj|obl)|few\/nsubj:pass of\/case .+\/nmod
each other (A)(B) = 1	each\/det other\/(nsubj|obj|obl)

Table 3.9: Regular Expressions for generalized quantifiers.

3.B Pairwise Observation

While the analysis in Section 3.4 is based on quantifiers in hypotheses, next we consider the interaction of quantifiers in hypotheses and quantifiers in premises. To this end, we calculate the difference between overall performance and performance for premise-hypothesis pairs of GQs. In Figure 3.2, we visualize the results as heatmaps (see Table 3.10 for exact numbers of occurences and accuracies). Surprisingly, whenever quantifiers appear in both the premise and the hypothesis, LMs largely fail to predict the entailment. Percentage quantifiers, supposed to be semantically more complex than counting quantifiers, are not *de facto* harder in NLI. We studied all 27 cases of percentage quantifiers in the English NLI datasets, and found that in most cases, percentage quantifiers occurrences are *identical* across premises and hypotheses, i.e., triggering little or no inference. The other two proportional quantifiers, *most* and *few*, are hard for LMs to resolve, e.g., in some quantifier pairs, models yield 0% accuracy. Although *each other* is supposed to be hardest to resolve due to the complex semantics of reciprocals (Szymanik and Thorne, 2015), it is not reflected in NLI tasks as such. The reason is similar to percentage quantifiers, while annotators intend to alter counting quantifiers when writing hypotheses, reciprocality is seldomly considered a linguistic ability that needs testing for NLU systems. And the annotation for Ramsey quantifier is simply a knockoff, making reciprocal relation identification unwarranted through shallow correlations.

3.C Fine-grained NLI Analysis

3.D XQA Result: mBERT and XLM-R

Table 3.11 compares the results of mBERT and XLM-R on two XQA tasks, XQuAD and MLQA.

3.E GQNLI Examples

Table 3.12 list one example per category in GQNLI.

3.F GQNLI Negation Cases

We present the results of seven models' performance on cases with negation cues in GQNLI in Table 3.13.

3.G GQNLI Subsumption Cases

See Table 3.14 for models 'performance on cases requiring subsumption reasoning in GQNLI. We also break down subsumption results by entailment labels into two categories: neutral and non-neutral.

3.H GQNLI Experiment Details

We reused the fine-tuned BERT and RobERTa in Section 3.4. The other fine-tuned LMs are from Hugging Face. We list the models and thier links in Table 3.15.

Hypothesis	Premise	MNLI_m_dev #occurrence	%Acc	MNLI_mm_dev #occurrence	%Acc	SNLI_test #occurrence	%Acc	ANLI_R1_test #occurrence	%Acc	ANLI_R2_test #occurrence	%Acc	ANLI_R3_test #occurrence	%Acc	Total #occurrence	%Acc	#correctpred
some	some	15	93.3	38	86.8	16	93.8	1	100			6	16.7	106	86.8	92
	all	8	87.5	8	50	3	100					3	0	22	63.6	14
	>k															
	<k															
	k	12	75	10	80	40	90	4	100	0		2	50	68	85.3	58
	between															
	most															
	few	2	50											2	50	1
	p/k															
	k%															
	eachother															
	"unmatched"	110	79.1	83	84.3	137	85.4	1	100	1	0	8	75	340	82.6	281
all	some	11	100	12	100	2	100					1	100	26	100	26
	all	73	82.2	74	86.5	3	100	4	25	2	50	6	50	162	81.5	132
	>k											1	0	1	0	0
	<k															
	k	28	85.7	19	100	22	81.8	9	22.2	5	40	13	23.1	96	70.8	68
	between			1	100							1	0	2	50	1
	most															
	few	4	100	2	100									6	100	6
	p/k							1	0					1	0	
	k%															
	eachother	0		0		1	100	0		0		0		1	100	1
	"unmatched"	151	86.1	144	87.5	41	92.7	5	100	2	0	13	30.8	356	85.1	303
>k	some					1	100	2	50	1	0			4	50	2
	all	1	0							2	0	2	50	5	20	1
	>k			2	100			1	100			2	50	5	80	4
	<k															
	k	1	0	3	100	2	100	3	66.7	9	55.6	10	40	28	57.1	16
	between									2	50			2	50	1
	most															
	few			1	100									1	100	1
	p/k															
	k%															
	eachother															
	"unmatched"	12	100	18	100	5	80	6	66.7	6	16.7	3	66.7	50	82	41
<k	some															
	all															
	>k															
	<k															
	k							3	33.3	5	40	3	100	11	54.5	6
	between															
	most															
	few							1	0					1	0	0
	p/k															
	k%															
	eachother															
	"unmatched"	3	100	3	100			3	33.3	2	100	2	50	13	76.9	10
k	some	8	75	14	78.6	28	85.7	2	50	2	0	4	0	58	72.4	42
	all	12	83.3	14	71.4	22	95.5	1	100	3	0	4	75	56	80.4	45
	>k	3	66.7									2	0	5	40	2
	<k			2	100							1	100	3	100	3
	k	140	84.3	121	76.9	593	92.2	42	42.9	49	44.9	37	32.4	982	82.5	810
	between	2	50	2	100					1	0			5	60	3
	most															
	few			1	100									1	100	1
	p/k	1	100	1	100			1	100	1	100			4	100	4
	k%															
	eachother					7	100							7	100	7
	"unmatched"	118	89.8	137	86.1	383	92.7	11	36.4	13	38.5	11	36.4	673	88	592
between	some							1	100					1	100	1
	all															
	k			1	100			1	0	2	50			4	50	2
	between			2	100			2	50					4	75	3
	most															
	few															
	>p/k:more/greaterthanp/k															
	<p/k:fewer/lessthanp/k															
	p/k															
	k%															
	eachother															
	"unmatched"	2	100	1	100			1	100					4	100	4
most	some			2	50									2	50	1
	all			2	100									2	100	2
	>k			1	0									1	0	0
	<k															
	k	5	60	1	100					2	50	1	0	9	55.6	5
	between															
	most	7	85.7	4	75									11	81.8	9
	few	1	100											1	100	1
	p/k															
	k%															
	eachother															
	"unmatched"	23	73.9	30	83.3	1	0							54	77.8	42
few	some															
	all	1	100	1	0							0		2	50	1
	>k			1	100							0		1	100	1
	<k											0		0		0
	k	4	75	3	33.3	4	100					3	33.3	14	64.3	9
	between															
	most	0		1	0	0								1	0	0
	few	0		2	100	0						1	0	3	66.7	2
	p/k															
	k%															
	eachother															
	"unmatched"	9	77.8	9	88.9	7	85.7					2	50	27	81.5	22
p/k	some															
	all															
	>k															
	<k															
	k			2	100	0		1	100	0		0		3	100	3
	between															
	most															
	few															
	p/k			2	100			1	100					3	100	3
	k%															
	eachother															
	"unmatched"	1	100	3	33.3	1	100	0		1	100	0		6	66.7	4
k%	some															
	all															
	>k															
	<k															
	k	6	83.3	6	100							1	100	13	92.3	12
	between															
	most															
	few															
	p/k															
	k%															
	eachother															
	"unmatched"	4	100	1	100									5	100	5
eachother	some											1	100	1	100	1
	all					3	100							3	100	3
	>k															
	<k															
	k	1	100			15	80					1	100	17	82.4	14
	between															
	most															
	few															
	p/k															
	k%															
	eachother					1	100							1	100	1
	"unmatched"	3	66.7	3	100	19	89.5							26	84.6	22

Table 3.10: Statistics of pairwise analysis in Monolingual NLI Benchmarks

Quant.	mBERT														XLM-R											
	en		zh		es		ar		vi		de		*weighted*		en		zh		es		ar		vi		de		*weighted*	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
XQuAD																										
some	75	84.2	50	55.5	58.3	76.1	50	**50**	16.6	42.4	**33.3**	**43.8**	47.2	58.7	**66.7**	**76.1**	**41.6**	**51.3**	**50**	**71.5**	66.7	73.6	66.7	**76.9**	66.7	80.6	59.7	71.7
all	**28.5**	**62.2**	**14.2**	**35.2**	**28.5**	82	42.8	**52.3**	**14.2**	**29.4**	**28.5**	**56**	**26.1**	**52.9**	**57.1**	91.8	**14.2**	**21.4**	**57.1**	**78.6**	**42.8**	**54.9**	85.7	85.7	**57.1**	79.3	**52.3**	**68.6**
$> k$																												
$< k$																												
k	78.1	90.1	68.7	80.4	56.2	**72.1**	**40.6**	64.3	**12.5**	35.7	56.2	77.1	52.1	70	75	87.4	53.1	58.8	**46.8**	**77.4**	65.6	86.3	62.5	85.4	62.5	86.9	60.9	80.4
between	100	100	**33.3**	**72.2**	66.6	93.3	100	100	**0**	**19**	**0**	**56.5**	50	73.5	100	100	66.7	66.7	**33.3**	**60**	100	100	100	100	**33.3**	**55.5**	72.2	80.4
p/k																												
$k\%$																												
most	**40**	**53.3**	**40**	**40**	**0**	**10**	**0**	**26.6**	**0**	**0**	**20**	**49.3**	**16.7**	**29.9**	**40**	**48**	**20**	**33.3**	**40**	**50**	**0**	**26.6**	**0**	**0**	**20**	**49.3**	**20**	**34.5**
few																												
each other																												
all GQs	**70**	**83.2**	55	66.7	**50**	**70.3**	41.6	58.2	**11.6**	32.5	**43.3**	**65**	**45.3**	62.7	**70**	**83.6**	**43.3**	**50.2**	**48.3**	**73.6**	60	76	68.3	83.6	**58.3**	80.3	**58**	74.6
comp.	71.8	83.7	48	59.1	56	74.5	40.8	57.9	13.9	32.4	50.7	67.2	46.9	62.5	74.5	86	43	52.8	61	80	53.3	71.7	58.1	78	61.1	77.1	58.5	74.3
MLQA																												
some	**59**	80	**28.2**	**52.1**	**34.1**	**59.2**	36.3	54.9	**5.4**	**24**	**33.3**	58.4	**32.7**	**54.8**	**69.6**	86.1	**33.3**	**60.6**	**41.4**	**70**	43.1	62.9	**43.2**	78	**45.4**	**61.1**	**46**	**69.8**
all	67.7	**79.8**	**14.2**	**46.4**	**38.4**	**62.8**	33.3	57.9	**10.5**	30.1	**31.2**	**51.6**	**32.6**	**54.8**	77.4	90.6	**35.7**	70	**42.3**	**66.4**	**38**	**60**	57.8	79.8	**37.5**	**51**	**48.1**	**69.6**
$> k$																												
$< k$	0	0							0	13.3			0	6.7	0	40							0	20			0	30
k	**74.9**	79.4	47	63.4	**41.5**	**65.9**	**27.6**	50.3	**6.3**	**23.7**	**38.2**	**53**	39.3	**56**	**69.2**	**82.1**	45.2	66.2	**48.7**	73.3	43	64.9	**48.5**	**71.9**	**46.3**	**62.1**	**50.2**	70.1
between	**50**	88.5	50	83.3	**0**	**26.6**	**0**	68.7	**0**	**26.6**			**20**	58.7	**50**	88.5	50	**50**	**50**	**65.3**	**0**	**54.6**	**0**	77.4			**30**	**67.2**
p/k	100	100	**0**	**0**	**0**	**0**							**33.3**	**33.3**	100	100	100	100	100	100							100	100
$k\%$	100	100	**0**	**26.6**			**0**	**23.7**					**33.3**	**50.1**	100	100	0	**26.6**			**0**	71.4					**33.3**	**66**
most	**55.5**	**7**	47.3	62.1	**45.4**	**61.7**	30	**46.8**	**5.8**	**15.7**	**33.3**	**40.7**	**36.2**	**50.3**	**59.2**	**76**	47.3	69.5	**45.4**	**59.5**	**40**	63.2	**47**	75.7	**22.2**	**31.7**	**43.5**	**62.6**
few																												
each other																												
all GQs	**63.5**	**79.2**	41.8	60.3	**39.6**	**63.7**	**29.3**	51.3	**6.4**	**23.6**	**36.1**	**53.2**	**36.1**	**55.2**	**69**	**83**	43	65.6	**46.9**	**71.5**	**41.9**	64.1	**47.6**	73.2	**44.4**	**59.8**	**48.8**	**69.5**
comp.	67.2	80.6	37.5	57.9	47.3	66	30	48.4	11.2	28	40.8	56	39	56.2	70.4	83.3	38.7	62.5	54.1	72.2	42.5	62.9	50.5	72.3	52.2	67.3	51.4	70.1

Table 3.11: Results of mBERT and XLM-R performance on XQA tasks decomposed by quantifier categories.

Quantifier	**Premise**	**Hypothesis**	**Label**
some	"There are six dogs. Three brown dogs, a black dog and a white dog run along the green grass."	"Some dogs sit."	Neutral
all	"In 2021, there are 490 million people in Africa living in extreme poverty, or 36% of the total population."	"Not all people in Africa live in extreme poverty."	Entailment
$> k$	"Two young men in blue stand over a stove and look at the camera while another young man in red stands behind them."	"At least two men wear red."	Contradiction
$< k$	"More than five guys chased two girls in the classroom."	"No less than four guys chased two girls in the classroom."	Entailment
k	"There are twelve singers on a stage, less than half from Argentina and one from Cape Verde."	"Two singers come from Argentina."	Neutral
between	"Only half out of six cleaners are sweeping up animal faeces from the street during a parade."	"Between four and five cleaners are sweeping up animal faeces."	Contradiction
p/k	"More than 50% but less than 65% of Americans worry about global warming."	"Two thirds of Americans worry about global warming."	Contradiction
$k\%$	"More than five guys chased two girls in the classroom."	"100% of the guys chased two girls in the classroom."	Neutral
most	"Two young men in blue stand over a stove and look at the camera while another young man in red stands behind them."	"Most men wear blue."	Entailment
few	"More than 50% but less than 65% of Americans worry about global warming."	"A few people from America do not worry about global warming."	Entailment
each other	"There are 100 villagers and 100 townsmen. Most villagers and most townsmen hate each other."	"All villagers and all townsmen hate each other."	Neutral

Table 3.12: GQNLI examples.

Quantifier		some	all	$> k$	$< k$	k	between	p/k	$k\%$	most	few	each other	*Overall*
# Occurrence with negation cues		9	6	6	9	18	3	6	6	6	9	3	81
Model	**Training Data**	**% Performance**											
BERT	S,M,F,ANLI	0	66.7	100	33.3	50	0	50	0	50	22.2	33.3	39.2
ELECTRA	S,M,F,ANLI	33.3	50.0	100.0	33.3	50.0	0.0	50.0	0.0	66.7	0.0	0.0	43.1
SBERT	S,M,F,ANLI	55.6	50.0	66.7	11.1	27.8	0.0	50.0	0.0	100.0	66.7	0.0	54.9
RoBERTa	MNLI	33.3	16.7	0	33.3	27.8	66.7	33.3	33.3	50	33.3	33.3	31.4
	S,M,F,ANLI	66.7	83.3	100.0	33.3	66.7	100.0	50.0	50.0	50.0	33.3	66.7	58.8
ALBERT	S,M,F,ANLI	88.9	50.0	66.7	33.3	55.6	100.0	0.0	50.0	50.0	11.1	0.0	49.0
BART	MNLI	33.3	0.0	50.0	66.7	66.7	100.0	0.0	100.0	0.0	33.3	0.0	35.3
	S,M,F,ANLI	66.7	50.0	100.0	33.3	50.0	0.0	50.0	0.0	50.0	66.7	100.0	52.9
DeBERTa-v3	MNLI	33.3	0.0	50.0	33.3	50.0	100.0	66.7	50.0	0.0	33.3	0.0	37.3
	M,F,ANLI	55.6	66.7	100.0	33.3	66.7	100.0	50.0	50.0	100.0	55.6	33.3	66.7
	M,F,Ling,DocNLI	33.3	100.0	100.0	0.0	33.3	0.0	83.3	0.0	50.0	66.7	100.0	51.0

Table 3.13: Models' performance on instances with negation cues in GQNLI.

Type		Subsumption (neutral)	Subsumption (non-neutral)	Subsumption (total)	Non-subsumption
# Occurrence		90	99	189	111
Model	**Training Data**	**% Performance**			
BERT	S,M,F,ANLI	22.2	24.2	23.3	41.4
ELECTRA	S,M,F,ANLI	3.3	52.5	29.1	53.2
SBERT	S,M,F,ANLI	68.9	35.4	51.3	18.9
RoBERTa	MNLI	27.8	18.2	22.8	37.8
	S,M,F,ANLI	21.1	33.3	27.5	59.5
ALBERT	S,M,F,ANLI	33.3	38.4	36.0	49.5
BART	MNLI	36.7	46.5	41.8	40.5
	S,M,F,ANLI	44.4	23.2	33.3	58.6
DeBERTa-v3	MNLI	45.6	26.3	35.4	33.3
	M,F,ANLI	52.2	37.4	44.4	54.1
	M,F,Ling,DocNLI	86.7	17.2	50.3	36.0

Table 3.14: Models' performance on instances requiring subsumption reasoning.

Model	**Training Data**	**Model's link**
ELECTRA	S,M,F,ANLI	`https://huggingface.co/ynie/electra-large-discriminator-snli_mnli_fever_anli_R1_R2_R3-nli`
SBERT	S,M,F,ANLI	`https://huggingface.co/usc-isi/sbert-roberta-large-anli-mnli-snli`
BART	MNLI	`https://huggingface.co/facebook/bart-large-mnli`
	S,M,F,ANLI	`https://huggingface.co/ynie/bart-large-snli_mnli_fever_anli_R1_R2_R3-nli`
ALBERT	S,M,F,ANLI	`https://huggingface.co/ynie/albert-xxlarge-v2-snli_mnli_fever_anli_R1_R2_R3-nli`
DeBERTa-v3	MNLI	`https://huggingface.co/MoritzLaurer/DeBERTa-v3-base-mnli`
	M,F,ANLI	`https://huggingface.co/MoritzLaurer/DeBERTa-v3-base-mnli-fever-anli`
	M,F,Ling,DocNLI	`https://huggingface.co/MoritzLaurer/DeBERTa-v3-base-mnli-fever-docnli-ling-2c`

Table 3.15: Links to the models we use to test on GQNLI.

Chapter 4

What does the Failure to Reason with "Respectively" in Zero/Few-Shot Settings Tell Us about Language Models?

Abstract

Humans can effortlessly understand the coordinate structure of sentences such as "Niels Bohr and Kurt Cobain were born in Copenhagen and Seattle, *respectively*". In the context of natural language inference (NLI), we examine how language models (LMs) reason with respective readings (Gawron and Kehler, 2004) from two perspectives: syntactic-semantic and commonsense-world knowledge. We propose a controlled synthetic dataset WikiResNLI and a naturally occurring dataset NatResNLI to encompass various explicit and implicit realizations of "respectively". We show that fine-tuned NLI models struggle with understanding such readings without explicit supervision. While few-shot learning is easy in the presence of explicit cues, longer training is required when the reading is evoked implicitly, leaving models to rely on common sense inferences. Furthermore, our fine-grained analysis indicates models fail to generalize across different constructions. To conclude, we demonstrate that LMs still lag behind humans in generalizing to the long tail of linguistic constructions.

Bohr and Cobain were born in Copenhagen and Seattle respectively .

Bohr and Cobain were born in Copenhagen and Seattle .

Bohr was born in Copenhagen .
Cobain was born in Seattle .

Figure 4.1: An example of explicit (top, evoked by "respectively") and implicit (middle, with no overt marker) respecitve readings. Humans can infer that both sentences have the same "cross-serial" meaning (bottom) by relying on commonsense knowledge (that a person is only born in one location) and world knowledge (that Copenhagen and Seattle are mutually exclusive).

4.1 Introduction

Transformer-based language models (LMs) (Devlin et al., 2019; Raffel et al., 2020; Brown et al., 2020) induce useful representations for a wide range of natural language understanding (NLU) tasks, including natural language inference (NLI; Wang et al., 2018; Hu et al., 2020), especially in in zero-shot or few-shot settings. To what extent this usefulness results from memorization, generalization or the ability of LMs to draw common sense inferences remains an open question.

To approach it, the linguistic phenomenon of respective readings (Gawron and Kehler, 2004) serves as an excellent probe. This phenomenon has so far been underexplored in NLP, even though it has been studied extensively in linguistic semantics (McCawley, 1968; Pullum and Gazdar, 1982; Dalrymple and Kehler, 1995; Eggert, 2000). In English, "respectively" is a rare word[1] used to establish a one-to-one mapping between two sets of participants and to distribute predicates over sets (Okada, 1999). For example, in Figure 4.1, the first conjunct in the subject corresponds to the first conjunct in the object and the second conjunct in the subject corresponds to the second conjunct in the object. The respective relation is bijective and respects the relative order of the elements of two different coordinate expressions; it is, in other words, cross-serial. "Respectively" can have different syntactic or semantic properties depending on the context, e.g., as a conjunction or adverb.

[1]In terms of frequency, in the British National Corpus, "respectively" is ranked 13,606th among 18,089 words, and 233rd among 429 adverbs (Leech et al., 2014).

In this paper, we investigate how LMs reason with respective readings. We propose two datasets, WikiResNLI (a controlled synthetic dataset) and NatResNLI (a naturally occurring dataset) to cover various explicit and implicit realizations of "respectively". Our research questions are:

1. Can NLI models reason with "respectively" constructions in zero-shot settings?
2. Can LMs generalize from explicit to implicit respective readings?
3. Can LMs generalize from synthetic to natural respective readings?
4. What cues do LMs leverage for prediction?

We experiment with state-of-the-art LMs and analyze the results to gain insights into the limitations of current models and potential directions for future research. We show that LMs are able to generalize effectively in a few-shot learning scenario when the word "respectively" is present. However, when the reading is evoked implicitly, a greater number of training instances are necessary. LMs require significantly more instances to generalize to naturally occurring datasets than humans. In conclusion, our study demonstrates that LMs continue to exhibit a deficit in generalizability to infrequent linguistic constructions with limited coverage in their training data.

4.2 Respective Readings

Respective readings are closely related to several types of readings instantiated by plurals and mass terms: distributive readings, collective readings and cumulative readings (Champollion, 2015).

Distributive readings. These usually refer to the application of a predicate to the subsets of a set or group. As for sentence 1(a), it is equivalent to "John smiled and Mary smiled". The reading is available because of the nature of the predicate is *atomic* (Winter, 2002), similar instances including "sing" and "sleep". Distributive reading can be enforced with overt distributive markers, i.e., "every" and "each" (Scha, 1984). In example 1(b), we enforce the reading by adding "each" at the end of the sentence so as to rule out the reading "John and Mary earn 200 dollars together".

1. (a) **Distributive reading:** John and Mary smiled.
   (b) **Distributive reading with an enforced marker:** John and Mary earn 200 dollars *each*.

Collective readings. These are the opposite of distributive readings in that the predicates apply to the whole plural entity instead of individuals. The quantifiers "all" and "most" instead of "every" and "each" are usually compatible with collective readings as in example 2(b) (Dowty et al., 1987).

2. (a) **Collective reading:** The men gathered.
   (b) **Collective reading with overt marker:** *All* of the men gathered.

Cumulative readings. These involves two entities but in a symmetric non-scopal relation as in the canonical example 3 (Scha, 1984). The sentence can be paraphrased into "There are three boys and two girls, each of the three boy saw at least one of the two girls, and each of the two girls was seen by at least one of the three boys.". It is discussed sometimes with weak reciprocity (Langendoen, 1978).

3. **Cumulative reading:** Three boys saw two girls.

Respective readings. These are thought to be a special case of cumulative readings in which a bijective relation holds between the two (or more) sets of entities that enter into the cumulative relation (Chaves, 2012). For example 4(a), the pair (Emiliano Zapata, Morelos) and the pair (Gerhart Münch, Michoacán) are grouped under the *died in* relation. Respective reading can also arise without the adverb *respectively*, and the absence is even sometimes preferred. As in example 4(b), the binomial expression "husband and wife" is so strong that the adverb "respectively" is unwarranted.

4. (a) **Respective reading with overt marker:** Emiliano Zapata and Gerhart Münch. died in Morelos and Michoacán, *respectively*.
   (b) **Respective reading without overt marker:** John and Mary are husband and wife.

4.3 An NLI Benchmark for "Respectively"

Understanding the coordinate structures in respective readings is effortless for humans, but it remains a question whether LMs, after being pre-trained on billions of tokens and fined-tuned on thousands of NLI instances, can reliably process them.

To probe LMs' behaviour in the presence of respective readings, we construct two English NLI datasets: WikiResNLI, a synthetic dataset based on an analogy corpus, and NatResNLI, a dataset sourced and created from natural occurrences. We release both datasets on Github[2] and describe the detailed creation steps below.

4.3.1 Synthetic Dataset: WikiResNLI

To generate a controlled synthetic challenge set for reasoning with respective readings, we exploit a useful relationship between coordination constructions and *analogies*. Analogy is concerned with similarities between observable properties and causal similarities.

Analogy dataset. Garneau et al. (2021) proposed WiQueen, a multilingual analogy dataset consisting of 78,000 analogies extracted from Wikidata. A subset of 9,000 instances is annotated where all four entities are *unique*. These are the analogies in which all relations are informative (Newman-Griffis et al., 2017). See Table 4.1 for an example. Their experiment showed that pretrained LMs can predict 29% of analogous entities in a zero-shot setting and 41% after training. This indicates that analogical knowledge already exists in pretrained models and can be enhanced by training.

Generating premises with "respectively". Given four analogical entities $\langle w_1, w_2, w_3, w_4 \rangle$ and the predicate p, we form a natural language premise consisting of the analogical information in a respective reading setting of 5(a) after adapting p for phrasing and conjugation. Such a premise is unambiguous and equivalent to 5(b), where the predication is distributed over the two pairs of entities. 5(a) is marked by an explicit respective reading indicator. As an implicit respective reading case, 5(c) has the same meaning as 5(b) but there is no explicit respective operator. In such implicit cases, the predicate

[2] `https://github.com/ruixiangcui/WikiResNLI_NatResNLI`

	Denotation	**Natural Language Example**
Premise:	w_1 and w_3 p w_2 and w_4, respectively.	*Emiliano Zapata* *and* *Gerhart Münch* *died in* *Morelos* *and* *Michoacán*, *respectively*
Hypotheses:		
Entailment (1), 1S1O	w_1 p w_2.	*Emiliano Zapata* *died in* *Morelos*.
Entailment (2), 1S1O	w_3 p w_4.	*Gerhart Münch* *died in* *Michoacán*.
Contradiction (1), 1S1O	w_1 p w_4.	*Emiliano Zapata* *died in* *Michoacán*.
Contradiction (2), 1S1O	w_3 p w_2.	*Gerhart Münch* *died in* *Morelos*.
Contradiction (3), 1S2O	w_1 p w_2 and w_4.	*Emiliano Zapata* *died in* *Morelos* *and* *Michoacán*.
Contradiction (4), 1S2O	w_3 p w_2 and w_4.	*Gerhart Münch* *died in* *Morelos* *and* *Michoacán*.
Contradiction (5), 2S1O	w_1 and w_3 p w_2.	*Emiliano Zapata* *and* *Gerhart Münch* *died in* *Morelos*.
Contradiction (6), 2S1O	w_1 and w_3 p w_4.	*Emiliano Zapata* *and* *Gerhart Münch* *died in* *Michoacán*.

Table 4.1: Example analogy in the spirit of Garneau et al. (2021). Both entity pairs (w_1, w_2; w_3, w_4) share the p relation. Object entities are *unique* in that given an entity pair and a subject, the fourth is uniquely determined. We generate eight hypotheses for each premise: 1S1O refers to one subject and one object, 1S2O refers to one subject and two objects and 2S1O refers to two subjects and one object.

p is usually mutually exclusive in that each subject can have only one object. For example, in Sentence 6(a) a person can only die in one place but not two places. Non-mutually exclusive predicates are disqualified for an implicit respective reading since they causes ambiguity, as in Sentence 6(b).

5. (a) w_1 and w_3 p w_2 and w_4, respectively.
   (b) w_1 p w_2 and w_3 p w_4.

(c) w_1 and w_3 p w_2 and w_4.

6. (a) Emiliano Zapata and Gerhart Münch died in Morelos and Michoacán.
   (b) John and Mary ate a falafel and a tortilla.

Generating hypotheses. We subsequently generate hypotheses and pair them with the generated explicit and implicit premises. In Table 4.1, we show the rules to write entailment or contradiction hypotheses given a premise created from the analogical entities and properties.

Statistics. The resulting dataset, which we call WikiResNLI$_{\text{EXPLICIT}}$, contains 2,317 premises with different analogical entities, each of which has two entailment hypotheses and six contradiction hypotheses, resulting in 18,536 premise-hypothesis pairs in total. The dataset has 139 different predicates derived from Wikidata properties. For the development set, we randomly sample 13 predicates from the 126 predicates left and trimmed them if the number of premises for each predicate exceeds 100. We have 1,312 premise-hypothesis pairs for the development set. The rest is used as the training set, with 1,577 premises and 12,616 premise-hypothesis pairs.

Generating premises with implicit "respectively". We aim to test whether LMs can reason with respective readings and generalize from explicit construction to instances without overt markers. For this purpose, we derive an implicit dataset from WikiResNLI$_{\text{EXPLICIT}}$ by simply removing the word "respectively" from the premises. We call this dataset WikiResNLI$_{\text{IMPLICIT}}$. In this process, we need to pay special attention to the fact that ambiguity usually occurs in the 1S2O setting when the predicates allow conjunction of objects; given the sentence 6(b), it is ambiguous whether the hypothesis "John ate a falafel and a tortilla" is entailed. To form a high-quality test set for WikiResNLI$_{\text{IMPLICIT}}$, we first need to exclude the ambiguous contradiction hypotheses. Therefore, two of the authors manually annotate the 139 predicates for whether they allow a single subject predicating conjunction of two objects. In total, 13 predicates are annotated by both authors as unambiguous. Subsequently, we keep only the premises with these predicates from the complete WikiResNLI, and for each predicate, we cap it if the number of premises exceeds 100. Eventually, we are left with 451 premises for the 13 predicates. The 3,608 premise-hypotheses pairs are used as the test set.

Human Reference	Entailment	Neutral	Contradiction
Entailment	93.4	2.1	4.5
Contradiction	5.9	4.1	90

Table 4.2: NatResNLI human annotated label distribution in percentages for each assigned reference label. Humans mostly agree with the pre-assigned reference labels (demonstrated in Table 4.1), but not always.

4.3.2 Naturally-occurring Dataset: NatResNLI

While the synthetic dataset is well-controlled, it does not necessarily cover the natural usage of "respectively". To address this, we also collect a dataset of naturally-occuring usages.

Collecting premises. As data resources for "respectively" in publicly available naturally-occuring data, we leverage two online dictionaries[3] and a writing advice blog,[4] which provide English examples containing specific words in real-world examples. We curate the sentences that included "respectively" and further filter some of them to avoid context ambiguity. In total, 76 sentences remain as the premise set.

Generating hypotheses. Two of the authors manually write hypotheses based on the fine-grained categorization of Table 4.1 for each collected premise. Given that the labels are pre-assumed, and to determine whether these inference relations align with humans, we employ crowd workers to verify them. See the annotation details in Appendix 4.A.

Statistics. The resulting dataset, which we call NatResNLI, consists of 76 premises and 608 hypotheses. The average sentence lengths of NatResNLI's premise and hypothesis are 20.1 and 10.1, respectively. Sentences have 2.32 conjnucts in average, with 4 as the maximum.

[3] `https://sentence.yourdictionary.com/respectively` and `https://www.dictionary.com/browse/respectively`

[4] `https://crosstalk.cell.com/blog/how-to-use-respectively-respectfully`

Variety. NatResNLI's sentences have more complicated linguistic constuctions than WikiResNLI, such as relative clauses, e.g., sentence 7(a), implicit coreferences in sentence 7(b), and inverted sentences in sentence 7(c).

7. (a) The annual value of the Hulse endowment is between £800 and £900, of which eight-tenths go to the professor of divinity and one-tenth to the prize and lectureship, respectively.
   (b) In 1910 the export of palm kernels was 6,141 tons, of palm oil 2,160 tons; in 1916 the figures were 22,391 tons and 3,852 tons respectively.
   (c) Above this, approached by a stair, are the Lesche and the theatre, occupying respectively the north-east and northwest corner of the precinct.

Inter-annotator Agreement. The inter-annotator agreement (Fleiss' kappa; Fleiss, 1971) of the workers for NatResNLI is 0.65, lower than ANLI's (0.67–0.74) and SNLI's (0.70). This can be attributed to that we have five annotators rather than the commonly chosen three annotators, as a larger number of annotators can sometimes lead to more diverse interpretations and disagreements, potentially lowering the inter-annotator agreement.

Verification of pre-assigned labels. In Table 4.2, we calculate the average agreement percentage of human annotation with reference labels, showing that humans do not always agree with them. Investigating the examples where the majority votes are distinct from the pre-assigned labels, we find nine instances distributed over four premises. For the sentence in 8(a), humans actually correct the label as the respective reading here does not cause a mutually exclusive effect. For sentence 8(b), humans show more caution towards sentence ambiguity caused by unknown world knowledge of Kilia and Dniester's locations, and hence the neutral label.

8. (a) **Premise**: The annual value of imports and exports exceeds seven and nine million sterling respectively. **Hypothesis**: The annual value of imports and exports exceeds seven million sterling. **Pre-assigned Label**: contradiction. **Majority Vote**: entailment
   (b) **Premise**: In that year a Turkish fleet captured the strongholds of Kilia and Akkerman, commanding respectively the mouths of the Danube and Dniester. **Hypothesis**: In that year a Turkish fleet captured the stronghold of Kilia, commanding the mouths of the Danube and Dniester. **Pre-asigned Label**: contradiction. **Majority Vote**: neutral

Model	**Training data**	entailment	contradiction			*overall*
			1S1O	1S2O	2S1O	
RoBERTa	MNLI	99.7	47.9	0.3	4.5	38.1
	S,M,F,ANLI	100	55.1	0.1	1	39.1
ALBERT	S,M,F,ANLI	99.8	31.6	3.1	3.9	34.6
DeBERTa-v3	MNLI	99.4	36.1	0.6	3.9	35
	M,F,ANLI	98.8	40.2	3.9	10.8	38.4
	M,F,Ling, WANLI	100	77.8	36.7	59.4	68.5

Table 4.3: Zero-shot performance on the $\text{WikiResNLI}_{\text{EXPLICIT}}$ test set.

Considering human annotations as ground truth, we discard the pre-assigned labels and adopt the majority votes as the final labels for NatResNLI.

4.4 Experiments

We begin our experiments with the datasets by addressing our first research question:

Theorem 4.1. *Can NLI models reason with the coordinate structure in "respectively" construction in zero-shot settings?*

Given the popularity of NLI as a classification task to test LMs' ability of language understanding, many works have proposed new models achieving state-of-the-art results on datasets such as SNLI (Bowman et al., 2015a), MultiNLI (MNLI; Williams et al., 2018) and ANLI (Nie et al., 2020). On the GLUE leaderboard,[5] the state-of-the-art models have surpassed 90% and 95% accuracy on MNLI and QNLI which are deemed as solved challenges. ANLI has been one of the most challenging tasks in recent years, and the latest models such as DeBERTa-v3-large (He et al., 2021; Moritz et al., 2022) and PaLM 540B (Huang et al., 2022) can achieve 64% and 67.9%, respectively. While many works use ANLI as a medium to exhibit the models' growing reasoning ability, few of them analyze in depth in which case it fails and at which stage it gets to learn certain linguistic abilities.

We report the zero-shot performances of three LMs fine-tuned with different combinations of NLI corpora. The models include RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2019) and DeBERTa with fine-tuning data of

[5] `https://gluebenchmark.com/leaderboard`

Model	**Training data**	entailment	contradiction			*overall*
			1S1O	1S2O	2S1O	
RoBERTa	MNLI	97.1	26.4	0.4	8.6	33.1
	S,M,F,ANLI	99.9	23.5	0.3	3.4	31.8
ALBERT	S,M,F,ANLI	100	14	0.1	0.8	28.7
DeBERTa-v3	MNLI	99.3	25.5	1.4	5.2	32.9
	M,F,ANLI	96.9	26.6	5.7	16.9	36.5
	M,F,Ling, WANLI	100	59.3	2.4	13.2	43.7

Table 4.4: Zero-shot performance on the $\text{WikiResNLI}_{\text{IMPLICIT}}$ test set.

MNLI, SNLI, ANLI, FEVER-NLI (Nie et al., 2019), LingNLI (Parrish et al., 2021) and WANLI (Liu et al., 2022).

The experiment results on $\text{WikiResNLI}_{\text{EXPLICIT}}$ and $\text{WikiResNLI}_{\text{IMPLICIT}}$ are presented in Table 4.3 and Table 4.4, respectively.

As can be seen in Table 4.3, models cannot fully correctly reason with respective readings. The best model, DeBERTa, only achieves 35% accuracy if fine-tuned with MNLI, and will reach 68.5% if fine-tuned with almost all NLI training datasets mentioned above. It gains a large increase in the 1S1O setting by 41.7%. However, the accuracy on 1S2O is still at a chance level, and the 2S1O setting performance is only approaching around 60%, leaving room for improvement.

The performance on $\text{WikiResNLI}_{\text{IMPLICIT}}$ is even worse, as indicated in Table 4.4. Similarly, DeBERTa is again the best performance model on the dataset, with an accuracy of 43.7% if fine-tuned with all NLI corpora. The accuracy is just 10% above the chance level, and it completely fails in the 1S2O and 2S1O settings.

Results on both datasets show that when training with more data, models improve on respective readings. However, the question of what leads to improvement remains. We examine how many times explicit respective readings appear in the training and testing datasets of MNLI, SNLI Fever-NLI and ANLI. We find that the adverb "respectively" occurs 177 and 12 times in the MNLI training and dev sets, 15 and 0 times in the SNLI training and test sets, 1,064 and 64 times in the Fever-NLI training and test sets, and 216 and 5 times in the combined ANLI training and dev sets. We randomly sampled a subset of each dataset and manually check whether they tackle reasoning over coordination structure. We find that in most cases, "respectively" works simply as a context word and has little to do with the

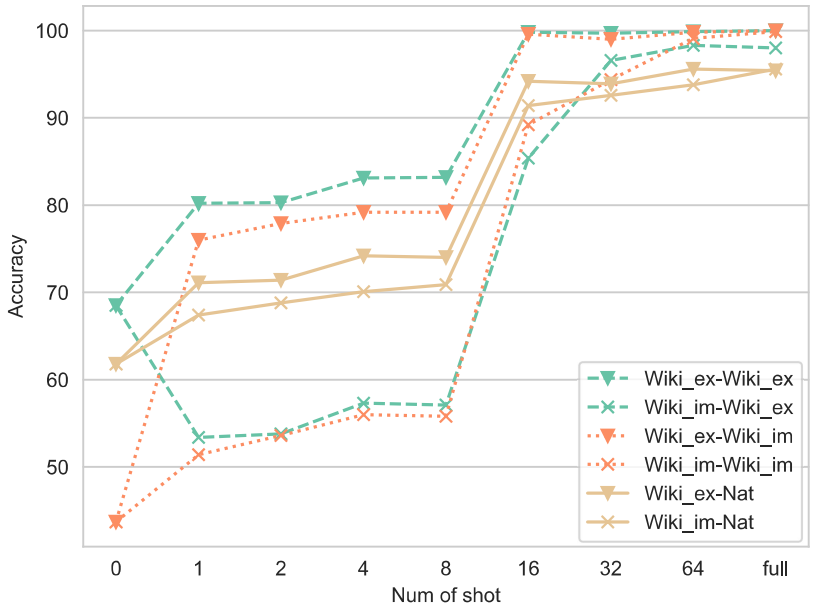


Figure 4.2: Overall performance of DeBERTa WikiResNLI$_{\text{EXPLICIT}}$, WikiResNLI $_{\text{IMPLICIT}}$ and NatResNLI from zero-shot to fully supervised. *Wiki_ex-Wiki_ex* refers to training with WikiResNLI$_{\text{EXPLICIT}}$ instances and evaluating on WikiResNLI$_{\text{EXPLICIT}}$ test set. Similarly, *Wiki_im-Nat* refers to training with WikiResNLI$_{\text{IMPLICIT}}$ and testing on NatResNLI.

actual inference relations. Thus it is still not clear whether it is simply the exposure to the explicit cues (the word "respectively") or some instances with implicit coordinate structures that result in the performance improvement. We thus ask the following three research questions and experiment with few-shot learning.

Theorem 4.2. *Can LMs Generalize from Explicit to Implicit Respective Readings?*

Instances of WikiResNLI have the coordinate structures of an equal number of conjuncts, and linguists have argued that such semantic relations are reflected in the syntactic relations (Goodall, 1987; Moltmann, 1992). It is essentially semantic but also relies on pragmatically available information of the truth conditions. Respective readings in fact also commonly omit explicit lexical indicators but remain available and preferred as 2(a) (Gawron and

# Shots **Type**	1	2	4	8	16	32	64	Full
All	8	16	32	64	128	256	512	12,616
Basic	4	8	16	32	64	128	256	6,308

Table 4.5: Number of training instances for each number of shots. A "shot" contains multiple training instances since we always take a premise along with all of its generated hypotheses—8 in the general case and 4 in the basic case.

Kehler, 2004). We are therefore interested in whether LMs can learn the semantic-pragmatic meaning of respective reading sentences rather than only making use of lexical and syntactic cues.

We fine-tune the DeBERTa model previously fine-tuned with M, F, Ling and WANLI with different numbers of $\text{WikiResNLI}_{\text{EXPLICIT}}$ examples without a dev set, since we do not want to bias the model towards our datasets hence hurting performance on the other NLI tasks.

We fine-tune the model with $\text{WikiResNLI}_{\text{EXPLICIT}}$ and $\text{WikiResNLI}_{\text{IMPLICIT}}$ separately and report the overall accuracy on both dataset in Figure 4.2. Training with $\text{WikiResNLI}_{\text{EXPLICIT}}$ contributes to a steady performance increase on both $\text{WikiResNLI}_{\text{EXPLICIT}}$ and $\text{WikiResNLI}_{\text{IMPLICIT}}$. Especially, 1-shot learning enhances the performance clearly, with a 10% increase for in-domain evaluation, and a remarkable 30% increase for explicit to implicit generalization. The improvements are small from 1-shot to 8-shot. Only at 16-shot, both $\text{WikiResNLI}_{\text{EXPLICIT}}$ in-domain learning and transferring to $\text{WikiResNLI}_{\text{IMPLICIT}}$ reach 100% accuracy. This shows the possibility to learn respective readings, despite the need to see relevant instances 128 times (see Table 4.5).

Interestingly, in-domain few-shot learning of $\text{WikiResNIL}_{\text{IMPLICIT}}$ witnesses a relatively cold start. The accuracy does not increase above 60% until 16 shots. Generalization from implicit respective reading to explicit reading is surprisingly not reaching 100% accuracy even after full supervision. We are keen to investigate what types of instances are difficult to learn for explicit to implicit respective reading generalization. In Figure 4.3, we break down $\text{WikiResNLI}_{\text{IMPLICIT}}$ with contradiction labels by categories (1S1O, 1S2O and 2S1O) and plot the accuracy against number of shot.

As can be seen, the performance on explicit readings is always better than on implicit readings across all three contradiction types. Among them,

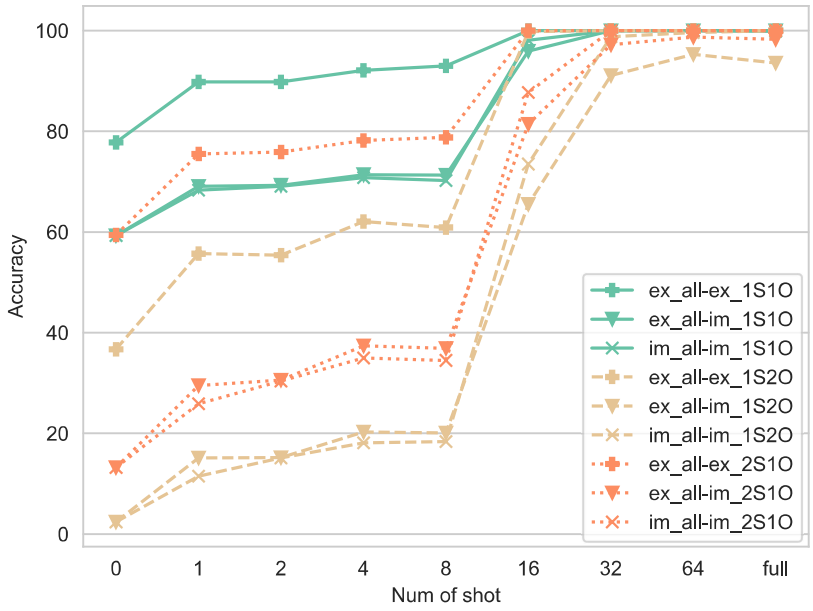


Figure 4.3: DeBERTa's Performances on $\text{WikiResNLI}_{\text{IMPLICIT}}$ after fine-tuning on $\text{WikiResNLI}_{\text{EXPLICIT}}$ or $\text{WikiResNLI}_{\text{IMPLICIT}}$. The result is broken down by contradiction fine-grained set.

1S2O and 2S1O instances are the most difficult. Their accuracies are below 40% and 20%, respectively before 16 shots. And only until 32 shots do both types reach above 95% accuracy. Unlike in-domain learning, 1S2O never gets perfectly solved.

Theorem 4.3. *Can LMs Generalize from Synthetic to Natural Respective Readings?*

WikiResNLI is a synthetic dataset, and it remains unclear whether models can reason with respective readings in realistic settings if we generate enough synthetic data and feed it to models. With NatResNLI, we are able to investigate LM's respective reading reasoning generalizability from synthetic to natural data and its alignment with humans.

We evaluate the models fine-tuned with $\text{WikiResNLI}_{\text{EXPLICIT}}$ on NatResNLI and plot the performance in Figure 4.4. We can observe that scores on NatResNLI are almost always lower than on WikiResNLI due to domain drift. Particularly, 1S2O and 2S1O are 10% and 20% lower in zero-shot settings.

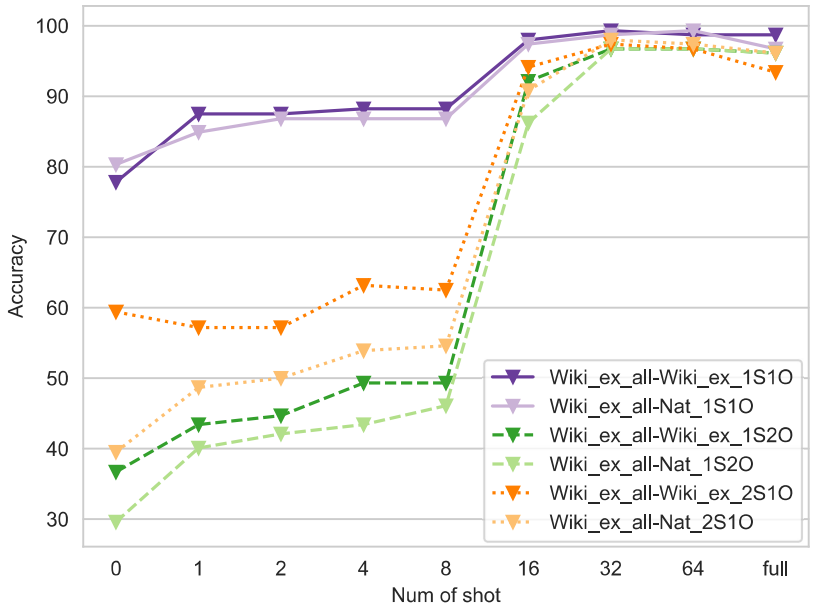


Figure 4.4: Performance of DeBERTa on NatResNLI after being fine-tuned on $\text{WikiResNLI}_{\text{EXPLICIT}}$. To facilitate comparison, we mark performances on $\text{WikiResNLI}_{\text{EXPLICIT}}$ in darker colours.

1S2O manage to reach on-par performance with WikiResNLI after 16 shots, while 2S1O after 32 shots.

Interestingly, the models are able to surpass 95% after 32 shots, while pre-assigned labels only have 90% match (see Table 4.2). Although we are comparing a rule-based method with 32-shot (256 examples) training, we can conclude that models are able to align with humans for respective reading reasoning. In addition, we notice that for 1S2O and 2S1O generalization, the complex linguistic structures discussed in Section 4.3.2 do have a high impact in the low-number few-shot learning, but the difficulty diminished as more training data are used.

Theorem 4.4. *What Cues do LMs Rely on?*

So far we have discussed LMs' ability to generalize on the syntactic-semantic level, from explicit to implicit and from synthetic to natural in respective readings. But it is yet to be determined whether the model is simply adopting the lexical-syntactic heuristics for prediction and whether it

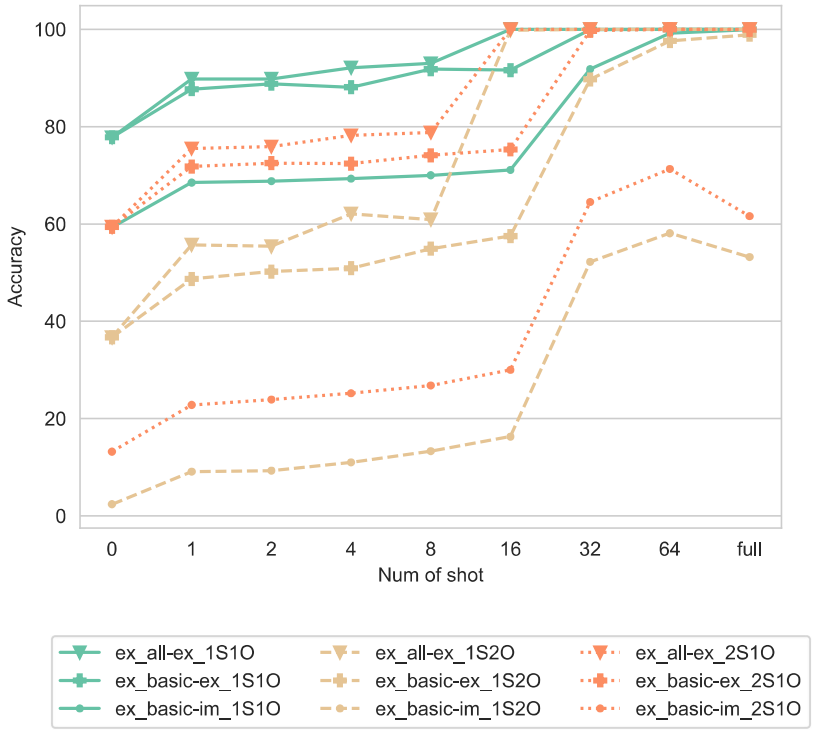


Figure 4.5: Performance of DeBERTa on WikiResNLI$_{\text{EXPLICIT}}$ and WikiResNLI $_{\text{IMPLICIT}}$ after being fine-tuned only with the basic types (entailment and 1S1O contradiction) of WikiResNLI$_{\text{EXPLICIT}}$.

leverages common sense and world knowledge. If models can reason over basic hypothesis structures (1S1O entailment and 1S1O contradiction), it would be expected they are aware that the one-to-one relation correspondences should exclude 1S2O and 2S1O propositions due to common sense and world knowledge. Although there are cases such as 8(a) where one object entity includes the other in NatResNLI, all cases of the WikiResNLI test set disallow the situation due to the mutually exclusive properties.

Therefore, we fine-tuned the DeBERTa models with only WikiResNLI$_{\text{EXPLICIT}}$ instances of basic structures and evaluated their performances on both WikiResNLI$_{\text{EXPLICIT}}$ and WikiResNLI$_{\text{IMPLICIT}}$ 1S2O and 2S1O. The results can be seen in Figure 4.5. We can observe that the generalization from basic structures to unseen structures is indeed difficult: while training with all structures and evaluating will all structures achieve perfect scores on 1S2O and 2S1O of

$\text{WikiResNLI}_{\text{EXPLICIT}}$ at 16 shots, training with basic structures are only 58% and 75% accuracies. It is worth noting that all fine-tuning instances have either entailment or contradiction labels, and therefore a random-guessing baseline would be 50% instead of 33.3%.

The generalization performances from explicit respective readings with basic structures to implicit 1S2O and 2S1O are more disappointing. At 16 shots, the accuracies are only 18% and 30%, respectively, well below the chance level. Even full supervision can only achieve around 60% accuracy for both structures. The results indicate that the models do not effectively learn the abstract respective reading relations due to not understanding the commonsense and world knowledge.

We look into the intersection errors of 32-shot, 64-shot and fully-supervised models which are fine-tuned on $\text{WikiResNLI}_{\text{EXPLICIT}}$ and are evaluated on $\text{WikiResNIL}_{\text{IMPLICIT}}$. There are 358 1S2O and 248 2S1O instances that are consistently mistaken by the models. The top-5 frequent properties are: twinned administrative bodies, took place, are capitals of, buried in, and family names. Knowledge about relative location 9(a) and knowledge about humans 9(b) thus seem to play an important role in reasoning with implicit respective readings.

9. (a) **Premise**: Battle of Tours and Battle of Verdun took place in Poitiers and Verdun. **Hypothesis**: Battle of Tours took place in Poitiers and Verdun. **WikiResNLI Label**: contradiction. **Prediction**: entailment

   (b) **Premise**: Theresa of León and Maria Solé Cuñat died in Galicia and Catalonia. **Hypothesis**: Theresa of León and Maria Solé Cuñat died in Galicia. **WikiResNLI Label**: contradiction. **Prediction**: entailment

Impact on other NLI tasks. We evaluate all models fine-tuned with WikiResNLI above on other NLI tasks, i.e, MNLI-m and ANLI-R3, to check whether fine-tuning on such a label-imbalanced dataset hurts performance. Interestingly, full supervision with $\text{WikiResNIL}_{\text{IMPLICIT}}$ of basic structures results in new state-of-the-art performance for DeBERTa. On MNLI-m, the score improves from 90.8% to 91.4%; and on ANLI-R3, the performance rises from 63.6% to 64.1%.

Experiments on LLaMA, FLAN-T5 and GPT-JT Significant advancements in large generative LMs have been achieved in the realm of general natural language understanding. These improvements can be attributed

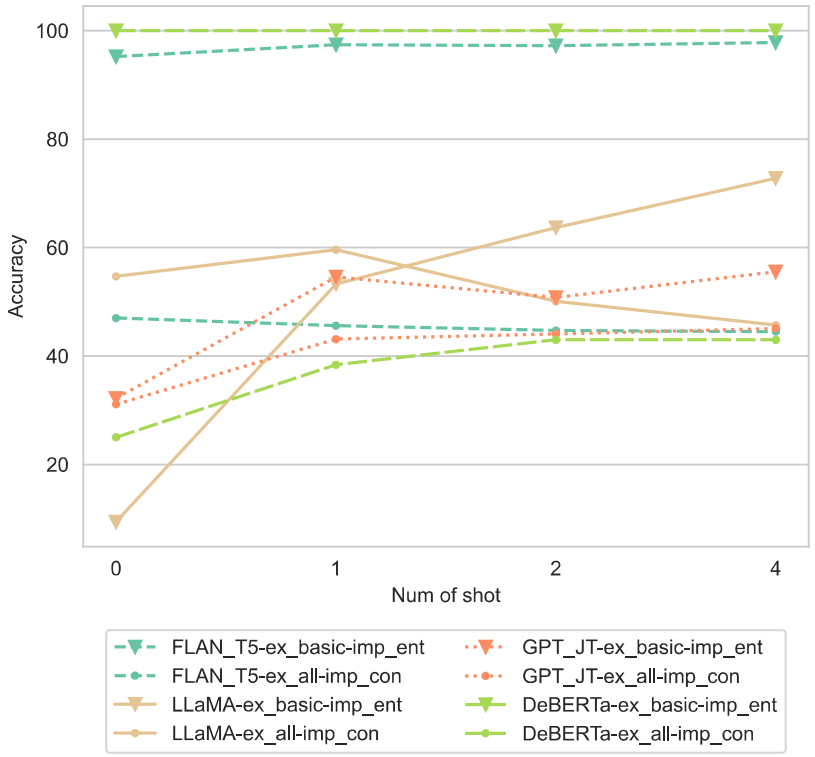


Figure 4.6: LLaMA, FLAN-T5, GPT-JT and DeBERTa's performances on $\text{WikiResNLI}_{\text{IMPLICIT}}$ after in-context learning of $\text{WikiResNLI}_{\text{EXPLICIT}}$. The last suffix *ent* of a legend means the performance on entailment pairs and *con* on contradiction pairs.

to enhanced training strategies, such as incorporating code and human instructions into pretraining/fine-tuning data and RLHF (Christiano et al., 2017; OpenAI, 2023). We assess the zero-shot and in-context learning abilities of three open-source generative models, that is, LLaMA-7B (Touvron et al., 2023), FLAN-T5-XL (Chung et al., 2022) and GPT-JT-6B (Wang and Komatsuzaki, 2021; Together, 2022). In this study, our focus is on two representative scenarios, namely generalizing from explicit to implicit readings and generalizing from synthetic to natural readings. We adopt the template *{premise} Question: Does this imply that {hypothesis}?* as it attains top-tier results for NLI tasks (Webson and Pavlick, 2022).

Figure 4.6 illustrates the explicit to implicit generalization results. No-

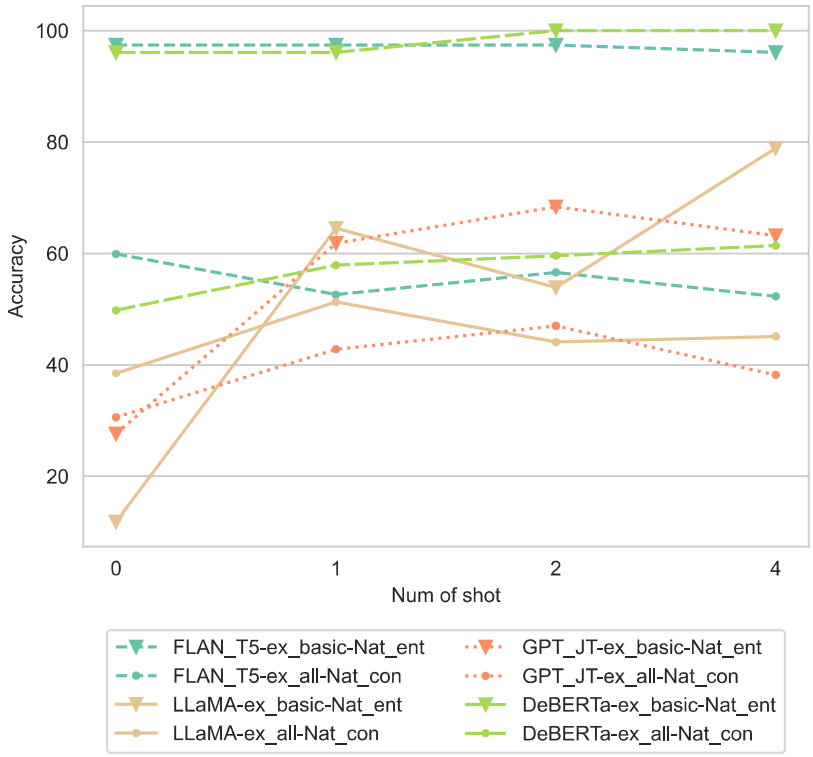


Figure 4.7: LLaMA, FLAN-T5, GPT-JT and DeBERTa's performances on NatResNLI after in-context learning of $\text{WikiResNLI}_{\text{EXPLICIT}}$.

tably, FLAN-T5 achieved a near-perfect score on zero-shot entailment pairs, comparable to the fine-tuned DeBERTa. However, GPT-JT, despite being instruction-tuned on NLI datasets, performed at a mere chance level on entailment pairs, while LLaMA scored below 10% accuracy. In terms of contradiction instances, all three models scored below 60% accuracy, with in-context learning offering limited improvement at the 4-shot level. Specifically, FLAN-T5's performance decreased after in-context learning.

For the generalization from WikiResNLI to NatResNLI, in Figure 4.7, we observed similar trends as in the previous experiments. FLAN-T5 outperformed the other models on entailment instances, and LLaMA demonstrated significant improvement within a few shots. However, for contradiction pairs, all models experienced only a modest increase in accuracy from 1 to 4 shots, with the highest accuracy remaining below 60%.

To conclude, while large generative models have made significant strides in natural language understanding, they still face substantial challenges in reasoning with respective readings, highlighting the need for further research and development in the long tail of linguistic constructions.

4.5 Related Work

Logical relations between two sentences are a core aspect of language understanding (Frege, 1879; Heijenoort, 1967; Blackburn et al., 2006). To facilitate large-scale model evaluation, NLP researchers have developed manually labelled NLI corpora, typically for 2/3-way classification (Dagan et al., 2013; Bowman et al., 2015a; Williams et al., 2018). In recent years, researchers start to analyze the characteristics of these datasets, such as annotation artefacts (Gururangan et al., 2018), syntactic heuristics (McCoy et al., 2019) and adversarial collection process (Williams et al., 2022).

In computational linguistics, distributive predication has been analyzed through means of distributivity operators (Massey, 1976; Link et al., 1983; Roberts, 1987; Lasersohn, 1998). And linguists have been working on extending first-order logical forms to include distributive and collective readings (Martin, 1981; Alshawi and van Eijck, 1989). Scha and Stallard (1988) present a recursive translation rule scheme to account for multi-level plurals. Aone (1991) proposed a reasoner consisting of domain-dependent constraints and domain-independent axioms for collective and distributive ambiguity. Shaw and McKeown (2000) described a simplified quantifier system to minimize distributive and collective ambiguities.

Respective readings have not yet been studied in modern NLP. Relevant works include plural understanding, which has been studied as a coreference resolution task (Jain et al., 2004; Zhou and Choi, 2018; Yu et al., 2020b). Manshadi et al. (2011) proposed quantifier scope annotation in which plurals are annotated with distributive and collective readings at the constraint level. Yu et al. (2020a) show that LMs are better at reflexive anaphora tasks with distributive than collective constructions.

4.6 Conclusions

The "respectively" construction is simple yet entails multiple levels of reasoning skills, including syntactic-semantic and commonsense-world knowledge. It is crucial that when an out-of-the-box model cannot reason over it, it should be able to learn with as few examples as possible. We proposed two datasets, WikiResNLI (a controlled synthetic dataset) and NatResNLI (a naturally occurring dataset) to probe their ability to do so in zero-shot and few-shot settings. We find that explicit reasoning is easier to learn than implicit reasoning, and LMs fail to generalize when common sense inference is needed. We confirm that diverse and complex training data are necessary to achieve human-level performance.

4.7 Limitation

Linguistic studies have shown that respective readings are not necessary to have two coordinate structures in the same sentence (Dalrymple and Kehler, 1995). Both WikiResNLI and NatResNLI have only one sentence in the premise and do not exhaust all possible and complicated realizations of respective readings. However, we are able to discuss and investigate LMs' generalizability with "respectively" with three constructions, i.e., 1S1O, 1S2O and 2S1O.

Our experiments are English-specific and are limited to LMs that can be run with an academic budget. However, our conclusions about the generalizability towards respective readings should be viewed as language-agnostic given there are linguistic constructions under-discussed in many other languages and it is worth researchers' attention to study them.

4.8 Acknowledgments

We would like to thank the members of the CoAStaL NLP group and the anonymous reviewers for their helpful suggestions.

Appendix

4.A Annotation Details

We employ Amazon Mechanical Turk workers. A qualified worker is one who has completed more than 10,000 HITs and has an approval rate greater than 99%. We set the location to the United States as there was no option to choose language proficiency. They are shown only three examples with entailment, neutral and contradiction labels before annotation. For each premise-hypothesis pair, five workers were asked to annotate the entailment relation (entailment, neutral or contradiction) following the guidelines of Nie et al. (2020). The worker gains a reward of 12 cents. Based on the workers' feedback, our hourly rate ranges between 16 to 27 US dollars, which is above the federal or Californian hourly wage. In total, 170 annotators participated in the step of label annotation of the hypotheses written by the authors. The number of HITs (annotation) per worker ranges from 5 to 200 based on their wishes. We assure to have 5 annotations per each premise-hypothesis pair.

Chapter 5

AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models

Abstract

Assessing foundation models' abilities for human-level tasks is crucial for Artificial General Intelligence (AGI) development. Traditional benchmarks, which rely on artificial datasets, may not accurately represent these capabilities. In this paper, we introduce AGIEval, a novel bilingual benchmark designed to assess foundation models in the context of human-centric standardized exams, such as college entrance exams, law school admission tests, math competitions, and lawyer qualification tests. We evaluate several state-of-the-art foundation models on our benchmark. Impressively, we show that GPT-4 exceeds the average human performance in SAT, LSAT, and math contests, with 95% accuracy on SAT Math and 92.5% on the Chinese college entrance English exam. This demonstrates the exceptional performance of contemporary foundation models. In contrast, we also find that GPT-4 is less proficient in tasks requiring complex reasoning or specific domain knowledge. Our comprehensive analyses of model capabilities (understanding, knowledge, reasoning, and calculation) reveal their strengths and limitations, providing valuable insights into future directions for enhancing general capabilities. By concentrating on tasks pertinent to human cognition and decision-making, our benchmark delivers a meaningful and robust evaluation of foundation

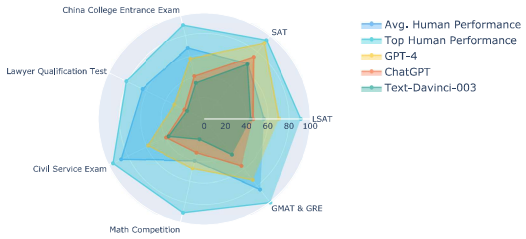


Figure 5.1: The performance of LLMs (text-davinci-003, ChatGPT, and GPT-4) was evaluated on several human-centric exams under zero-shot learning with a Chain-of-Thought (CoT) prompting setting. Human performance (avg.) refers to the average performance of all test takers, while human performance (top) refers to the performance of the top 1% of test takers. Compared to the averaged human performance, GPT-4 achieves better scores on the SAT, LSAT, and math competitions.

models' performance in real-world scenarios[1].

5.1 Introduction

Recently, large foundation models, such as the large language models (LLMs) ChatGPT(OpenAI, 2022) and GPT-4 (OpenAI, 2023), exhibited remarkable versatility and adaptability, with plethora of applications spanning various domains as a decision-making assistant, from processing daily events to assisting in specialized fields such as law and finance. With these advancements, AI systems are inching closer to achieving Artificial General Intelligence (AGI). As these AI systems continue to evolve and become more integrated into our daily lives, it is essential to effectively assess their general abilities in handling human-centric tasks, identify potential shortcomings, and ensure that they can handle complex, human-centric tasks effectively. Moreover, evaluating their reasoning abilities is also crucial to ensure their reliability and trustworthiness across diverse settings.

Traditional benchmarks for evaluating foundation models often fall short in providing an accurate assessment of their general abilities in handling human-level tasks. This is primarily due to the use of artificial datasets and a lack of emphasis on real-world tasks that require human-like cognitive

[1]The data, code, and all model outputs are released in `https://github.com/ruixiangcui/AGIEval`

capabilities. Moreover, these benchmarks often focus on tasks that do not truly represent the complexities and nuances of real-world human cognition and decision-making, leading to a skewed evaluation of models' capabilities and limiting their ability to provide meaningful insights into the models' real-world applicability. Consequently, there is a growing need for a more human-centric benchmark that allows for a robust evaluation of foundation model in the context of tasks that are relevant to human reasoning and problem-solving.

We introduce a human-centric benchmark, AGIEval, specifically designed to evaluate the general abilities of foundation models in tasks pertinent to human-level problem-solving. This benchmark is derived from official, public, and high-standard admission and qualification exams intended for general human test-takers, such as general college admission tests (e.g., Chinese College Entrance Exam (Gaokao) and American SAT), law school admission tests, math competitions, lawyer qualification tests, and national civil service exams. These exams are taken by a diverse range of individuals seeking entry into higher education institutions or new career paths, with millions participating annually (e.g., 12 million for the Chinese Gaokao and 1.7 million for the American SAT). As a result, these exams establish officially recognized standards for assessing human-level capabilities. Additionally, the benchmark covers bilingual tasks in both Chinese and English, allowing for a more comprehensive evaluation. By concentrating on these tasks, our benchmark provides a more meaningful and comprehensive evaluation of large language model performance in scenarios directly relevant to human decision-making.

We employ 20 human-centric tasks across a wide variety of subjects in our benchmark to assess the performance of cutting-edge foundation models, encompassing close-source models, i.e., text-davinci-003, ChatGPT and GPT-4, and an open-source model, Vicuna (Chiang et al., 2023). Our experiments explore their performance under various settings, including few-shot learning, zero-shot learning, and chain-of-thought prompting techniques. We compare the performance of these models with human performance, as illustrated in Fig. 5.1. Remarkably, the results reveal that GPT-4 outperforms the average human performance on LSAT, SAT, and math competitions under the zero-shot chain-of-thought (CoT) setting, demonstrating its capability on human-centric tasks. However, there remains a gap between GPT-4 and the top human performance, indicating opportunities for future improvement. We also discover that these models struggle with tasks requiring complex reasoning (e.g., LSAT-analytical reasoning and physics) or specific domain knowledge,

such as law and chemistry. Moreover, our comprehensive qualitative analyses of the four dimensions of model capabilities (i.e., *understanding, knowledge, reasoning, and calculation*) delve into their respective strengths and limitations, providing valuable insights into their general capabilities. This multi-faceted approach enables us to examine the models' single-task behavior and identify general patterns, ultimately contributing to a more robust understanding of these state-of-the-art models and their potential applications in tackling human-level tasks.

5.2 Background and Related Work

Large Foundation Model: Recently, large foundation models, like LLMs (e.g., GPT-3 (Brown et al., 2020), GPT-4 (OpenAI, 2023), OPT (Zhang et al., 2022a) and FLAN-T5 (Chung et al., 2022)) have successfully demonstrated unprecedented performance in a wide range of natural language tasks. The success of these models can be attributed to advances in deep learning techniques, architectural improvements, and the availability of massive amounts of data for training. The most recent cutting-edge language models, such as ChatGPT(OpenAI, 2022) and GPT-4 (OpenAI, 2023), have continued to demonstrate substantial adaptability to a diverse array of tasks and domains and have served as a daily decision-making assistant for human beings. However, despite their impressive performance on various benchmarks, concerns have been raised about the reasoning abilities, trustfulness and real-world applicability of these models (Marcus and Davis, 2019).

Evaluation of Language Models: Constructing benchmarks is a reliable way to establish evaluation standards and monitor model performance. Numerous benchmarks (Thorne et al., 2018; Rajpurkar et al., 2016) have been proposed and widely adopted for evaluating single-task performance, such as SQuAD (Rajpurkar et al., 2016) for assessing answer extraction ability and SNLI (Bowman et al., 2015a) for evaluating natural language inference capability. The emergence of general language models (LMs) like BERT (Devlin et al., 2019) has made it increasingly essential to develop more comprehensive benchmarks to assess the general capabilities of these LMs. GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) are popular benchmarks that evaluate language model performance across diverse NLP tasks. GLUE series benchmarks have significantly influenced language model development,

encouraging researchers to enhance their models' generalization capabilities. The LAMBADA language modeling task (Paperno et al., 2016) assesses language models' ability to capture long-range dependencies in text. SentEval (Conneau and Kiela, 2018) and DecaNLP (McCann et al., 2018) also set benchmarks for evaluating models' general capabilities. ToxiGen (Hartvigsen et al., 2022) and BOLD (Dhamala et al., 2021) evaluate the bias in language models. Despite their broad applicability, these benchmarks mainly consist of artificially curated datasets designed to evaluate specific machine skills, rather than real-world problems aimed at assessing human behaviors. Consequently, these benchmarks primarily focus on simpler textual understanding rather than complex reasoning abilities aligned with real-world applicability. MMLU (Hendrycks et al., 2020) addresses this issue by collecting questions from online sources covering a diverse set of subjects (e.g., history, humanities) that humans learn, pushing towards human-centric evaluation. Our work differs from MMLU in two main ways: (1) We derive our benchmark from high-standard human-centric exams like college admissions tests, ensuring a robust, standardized evaluation, unlike MMLU which lacks explicit sourcing details. (2) AGIEval is bilingual (English and Chinese), broadening the assessment scope across languages and cultures, whereas MMLU is solely English-based. The official technical report of GPT-4 (OpenAI, 2023) also underscored the importance of evaluating models' behaviors on human exams and analyzed GPT-4's performance on several such exams. However, the relevant benchmarks in these reports and the corresponding model outputs are not publicly available, and the evaluation metric is also not transparent. These factors limit further research to follow up their evaluation.

5.3 Human-Centric Benchmark

5.3.1 Design Principles

Emphasis on human-level cognitive tasks: Our human-centric benchmark is designed to mimic human cognition and problem-solving, aiming for a comprehensive evaluation of foundation models. We use a diverse set of public, official exams, such as college admission tests, law tests, and national civil service exams. These exams, taken by millions seeking further education or careers, provide standards for assessing human-level capabilities, making our benchmark directly relevant to human cognition and decision-making.

Relevance to real-world scenarios: The second design principle is emphasizing tasks relevant to real-world situations. By utilizing high-standard admission and qualification exams, we capture the complexity and practicality of challenges in various fields. This not only measures model performance against human cognition, but also their applicability in real-life scenarios, fostering AI development that is reliable, practical, and capable of solving diverse real-world problems.

Exams	#Participants	Language	Tasks	Subject	#Instance	#Avg. Token
Gaokao	12M	Chinese	GK-geography	Geography	199	144
			GK-biology	Biology	210	141
			GK-history	History	243	116
			GK-chemistry	Chemistry	207	113
			GK-physics	Physics	200	124
			GK-En	English	306	356
			GK-Ch	Chinese	246	935
			GK-Math-QA	Math	351	68
			GK-Math-Cloze	Math	118	60
SAT	1.7M	English	SAT-En.	English	206	656
			SAT-Math	Math	220	54
Lawyer Qualification Test	820K	Chinese	JEC-QA-KD	Law	1000	146
			JEC-QA-CA	Law	1000	213
Law School Admission Test (LSAT)	170K	English	LSAT-AR	Law-Analytics	230	154
			LSAT-LR	Law-Logic	510	178
			LSAT-RC	Law-Reading	260	581
Civil Service Examination	2M	English	LogiQA-en	Logic	651	144
	2M	Chinese	LogiQA-ch	Logic	651	242
GRE	340K	English	AQuA-RAT	Math	254	77
GMAT	150K	English				
AMC	300K	English	MATH	Math	1000	40
AIME	3000	English				

Table 5.1: Exams included in AGIEval. We highlight the number of human participants taking these exams annually (column "# Participants"). We also report the number of instances and average token number in AGIEval.

5.3.2 Exam Selection

Our human-centric benchmark features various standardized exams, each serving unique assessment roles. Some exams are participated by millions of human test-takers annually. For example, 12 millions of students participate in Gaokao. Statistics of annual human participants are reported in Table 5.1. **Dataset collection is introduced in Appendix 5.B**. The following categories of human-centric exams are included in our benchmark:

General College Entrance Exams: Including the GRE, SAT, and Gaokao, these exams assess critical thinking, problem-solving, and analytical skills for entry into higher education. We selected tasks from eight subjects in the Gaokao and mathematical questions from the GRE and SAT. These exams are designed to assess the general aptitude and subject-specific knowledge of humans.

Law School Admission Test: LSAT measures reasoning and analytical skills of prospective law students. These tests include sections on logical reasoning, reading comprehension, and analytical reasoning, aiding us in evaluating language models' legal reasoning abilities and ability to analyze complex information and draw accurate conclusions.

Lawyer Qualification Test: Including the bar exam, these tests assess legal knowledge, analytical skills, and ethical understanding. Questions from Chinese lawyer qualification tests are included. By incorporating lawyer qualification tests in our benchmark, we can evaluate language models' performance in the context of professional legal expertise and ethical judgment.

Graduate Management Admission Test (GMAT): The GMAT is a standardized exam designed to assess the analytical, quantitative, verbal, and integrated reasoning skills of prospective graduate business school students. It assess LLMs' potential to assist in decision-making and problem-solving in management scenarios.

High School Math Competitions: Math competitions like **American Mathematics Competitions (AMC)** and the **American Invitational Mathematics Examination (AIME)** test mathematical abilities, creativity, and problem-solving skills, helping to evaluate models' proficiency in tackling complex mathematical problems.

Chinese Civil Service Examination: This exam assesses a range of competencies for prospective civil servants. These exams evaluate a range of competencies, such as general knowledge, reasoning abilities, language skills, and subject-specific expertise, allowing us to gauge models' performance in public administration contexts.

5.4 Evaluation of Foundation Models

5.4.1 Model Selection

In this section, we evaluate the performance of various state-of-the-art language models on our benchmark dataset. **(1) GPT-4:** The fourth iteration of the GPT series, GPT-4 is a large-scale, generative pre-trained transformer with enhanced performance and a broad knowledge base. It exhibits human-level performance in numerous scenarios, including factuality, steerability, and adherence to guardrails. **(2) ChatGPT:** An OpenAI-developed conversational model, ChatGPT is trained on extensive instruction data and fine-tuned using reinforcement learning with human feedback, enabling contextually relevant responses. **(3) text-davinci-003:** As an intermediate version between GPT-3 and GPT-4, GPT-3.5 offers improved performance, providing a comparative perspective. We specifically evaluate the text-davinci-003 variant. **(4) Vicuna-13B (Chiang et al., 2023):** It is an open-source LLM, trained on user-shared conversations from ShareGPT by fine-tuning LLaMA. It achieves over 90% of the quality of OpenAI's ChatGPT.

5.4.2 Experimental Setup

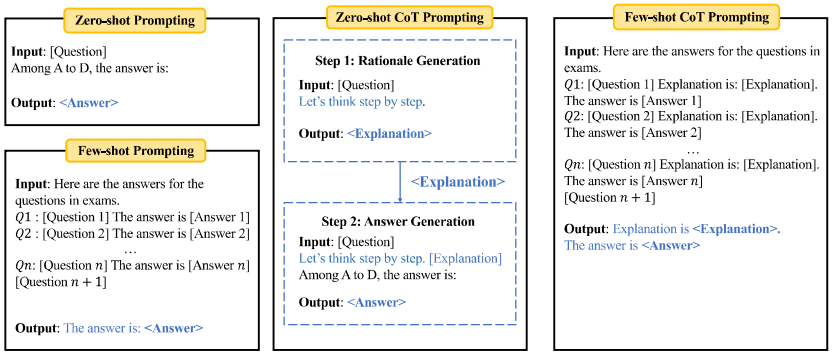


Figure 5.2: Prompting examples of different settings.

To gauge the adaptability of LLMs, we conduct two types of evaluations: zero-shot and few-shot. We further implement a "Chain-of-Thought (CoT)" reasoning evaluation. Fig. 5.2 describes the concrete prompting examples for zero-shot testing, few-shot testing and chain-of-thought prompting.

5.4.2.1 Zero-shot and Few-shot Evaluation

In the zero-shot setting, models were evaluated on the questions without being provided examples of the specific tasks. This scenario tests the models' innate ability to reason and solve problems without explicit training. In the few-shot setting, models were given a small number of examples (e.g., 5) from the same task before being evaluated on the test samples. This evaluation setup tests the models' ability to quickly adapt from limited examples.

5.4.2.2 Chain-of-Thought (CoT) Reasoning

We employ the Chain-of-Thought (CoT) prompting method (Wei et al., 2022) to assess models' reasoning capabilities. CoT enables large language models to break down a complex question to a series of decomposed reasoning steps. As shown in Fig. 5.2, CoT involves two steps: Firstly, with prompt "*[question] Let's think step by step:* "(Zhang et al., 2022b), the model generates an explanation for a given question, which evaluates its comprehension and problem-solving strategy identification. Secondly, the model provides an answer based on its explanation, testing its ability to generate a solution using its self-derived reasoning, mirroring human problem-solving processes. In the few-shot CoT setting, the explanation and answer are generated simultaneously.

5.4.2.3 Evaluation Metrics

We use both quantitative and qualitative evaluation metrics. Quantitative metrics included accuracy for multi-choice questions and use Exact Match (EM) for fill-in-blank questions. We also perform qualitative evaluations, which involved human evaluators assessing the models' responses in terms of semantic understanding capability, knowledge utilization, and reasoning and calculation.

5.4.3 Main Results

The results of closed-source models are reported in Table 5.2, while the results of the open-source model are reported in Table 5.3. We also report average and top human performance on each task. From the results, we highlight the following findings.

Task/Model	Human		Zero-Shot			Zero-Shot CoT			Few-Shot			Few-Shot CoT		
	Avg.	Top	TD	CG	G4	TD	CG	G4	TD	CG	G4	TD	CG	G4
AQuA-RAT	85	100	29.9	31.9	40.6	42.1	55.9	73.2	30.3	31.1	50.8	47.2	60.6	74.0
MATH	40	90	11.9	26.4	35.7	19.1	31.9	47.7	10.3	14.8	15.1	15.1	30.1	25.3
LogiQA (English)	86	95	22.7	35.0	49.3	36.9	39.9	57.8	43.5	43.5	63.9	37.5	38.9	62.7
LogiQA (Chinese)	88	96	40.3	41.0	58.8	36.7	38.9	57.5	43.2	46.2	65.0	40.0	38.6	61.9
JEC-QA-KD	71	78	21.9	21.1	33.4	18.4	21.2	31.9	22.4	27.6	41.3	23.6	23.4	40.4
JEC-QA-CA	58	85	21.0	22.0	31.1	16.7	19.6	29.8	22.2	25.1	37.4	16.1	20.0	34.7
LSAT-AR	56	91	21.7	24.4	35.2	23.9	22.6	34.4	22.6	25.7	33.9	22.6	25.2	31.7
LSAT-LR	56	91	47.5	52.6	80.6	50.0	52.6	80.6	60.4	59.2	85.9	51.2	52.2	84.5
LSAT-RC	56	91	64.7	65.4	85.9	57.6	62.1	85.1	70.6	67.7	87.7	64.3	57.6	87.7
SAT-Math	66	94	35.5	42.7	64.6	54.6	70.9	95.0	44.6	40.9	71.8	55.5	65.0	89.6
SAT-English	66	94	74.8	81.1	88.8	75.7	77.7	85.9	84.0	81.1	88.8	76.7	78.2	85.9
GK-Cn	65	85	43.9	39.0	53.3	35.4	33.7	44.7	25.6	41.5	61.4	29.3	37.8	51.6
GK-En	69	91	81.4	84.9	91.9	83.0	84.3	92.5	86.9	86.3	93.8	80.7	84.6	93.1
GK-geography	65	85	53.3	59.8	76.9	48.7	55.8	72.4	59.8	63.8	75.9	52.3	61.8	76.4
GK-history	64	85	47.3	59.7	77.4	37.0	50.2	76.5	49.0	57.6	77.8	51.9	58.4	78.2
GK-biology	68	89	40.5	52.9	75.7	30.0	42.4	71.9	44.3	52.4	80.0	32.9	50.0	72.9
GK-chemistry	66	86	27.1	38.7	51.7	24.6	33.8	52.2	32.4	44.0	54.6	35.8	33.8	54.1
GK-physics	71	94	22.0	33.0	39.0	18.5	29.5	45.5	31.0	33.5	43.5	27.5	36.5	54.5
GK-Math-QA	73	96	28.2	36.5	47.0	28.8	33.3	50.7	27.6	31.3	39.9	33.1	31.6	49.0
GK-Math-Cloze	73	96	17.0	7.6	16.1	4.2	5.1	15.3	5.9	5.9	11.0	5.93	8.5	16.1
Average	67	91	38.1	42.9	56.4	37.4	43.2	58.4	41.2	44.4	59.2	40.4	45	61.3

Table 5.2: Performance of close-source LLMs on 20 tasks under **zero-shot**, **zero-shot CoT**, **few-shot** and **few-shot CoT** settings. We also report human performance on each task. For LSAT, Gaokao and SAT, we report average (50%) and top (1%) human performance. The Text-Davinci-003 is abbreviated as TD, ChatGPT is abbreviated as CG, and GPT-4 is abbreviated as G4.

Task/Model	Computation		LogiQA		JEC-QA		LSAT			SAT		GK							
	AQaA	MATH	En.	Cn.	KD	CA	AR	LR	RC	Math	En.	Cn	En	Geo.	His.	Bio.	Che.	Phy.	M.-QA	M.-Cloze
Vicuna (ZS)	26.4	6.8	18.4	23.5	14.3	12.4	22.2	25.5	30.5	24.6	50.5	25.6	50.7	24.6	28.9	20.5	26.6	15	22.5	2.5
Vicuna (ZS-CoT)	22.1	6.6	30.3	27.1	14.9	15.2	20.9	36.1	44.2	35.5	57.8	23.6	67	28.6	34.9	24.3	23.2	17	21.7	1.7

Table 5.3: Performance of Vicuna-13B under zero-shot and zero-shot CoT setting. Task names are abbreviated.

(1) Superior Performance of GPT-4: On average, GPT-4 significantly outperforms its counterparts (e.g., ChatGPT) across all settings. Impressively, GPT-4 achieves 93.8% accuracy on Gaokao-English and 95% accuracy on SAT-MATH, demonstrating its superior capabilities.

(2) ChatGPT v.s. TD-003: ChatGPT excels over text-davinci-003 in tasks requiring extensive knowledge like geography, biology, chemistry, physics, and mathematics, implying a stronger knowledge base of ChatGPT. In tasks

emphasizing simple comprehension and logical reasoning, like English and LSAT tasks, both models perform comparably, indicating their proficiency in language understanding and logical reasoning.

(3) Challenge of Complex Tasks: All models face difficulties with complex tasks, such as those in MATH or LSAT-AR, revealing limitations in handling advanced reasoning. This presents future research opportunities to bolster models' reasoning abilities.

(4) Few-shot Learning vs. Zero-shot Learning: Few-shot learning marginally outperforms zero-shot learning, suggesting that LLMs' zero-shot capabilities are nearing their few-shot performance. This development, a marked improvement from the original GPT-3 (few-shot performance of GPT-3 is significantly better), may stem from enhanced human-alignment and instruction tuning in recent models. This progress demonstrates the effectiveness of recent advancements in LLM tuning, which allows them to better understand the meaning and context of tasks even in zero-shot settings. As shown in Fig. 5.3, Vicuna, despite excelling on OpenLLM leaderboard (Beeching et al., 2023) and its claimed comparable ability with ChatGPT, falls short on AGIEval, highlighting the valuable challenges AGIEval presents to open-source models.

5.4.4 Analyses of Chain-of-thought Prompting

As reported in Table 5.2, the CoT prompting demonstrates its potential by improving performance. However, the performance gains from CoT are not consistently observed across all tasks. Our analysis leads to the following findings:

(1) Performance Variability: CoT mainly enhances performance in English math and logic reasoning tasks but degrades performance in others, implying inconsistent effects on different tasks, which may be a consequence of the generated misleading reasoning processes. It's vital to understand what drives these variations to uniformly optimize CoT for diverse tasks.

(2) Backbone Dependency: CoT's efficacy is linked to the base model. GPT-4, for instance, generates more illustrative reasoning processes, improving CoT performance. This underscores the importance of model compatibility with CoT.

(3) Language Sensitivity: CoT performance varies with language. For LogiQA, CoT improves English tests but decreases Chinese ones. Similar findings are observed in mathematical tests, where performance increase on

English math tests (MATH, AQuA) but decrease on Chinese math exam in Gaokao. This suggests CoT's sensitivity to language differences, necessitating further optimization across languages to ensure its consistent and generalizable reasoning capabilities.

In conclusion, CoT's effectiveness is relevant to task, model capability, and language. These factors need careful consideration when employing CoT or developing future models.

5.4.5 Qualitative Analyses of Model Capabilities

We conduct a qualitative analysis of ChatGPT's outputs under a zero-shot CoT setting, with 100 erroneously answered instances for each task, to assess its alignment with human capabilities. We enlist human annotators with expert knowledge, such as Ph.D. students and professional researchers, to evaluate the model outputs (i.e., explanations and answers) along the following four dimensions and report average scores for tasks. **(1) Understanding**: Assessing whether the model comprehends the context and questions. **(2) Knowledge**: Evaluating the model's ability to recall relevant knowledge or formula for problem-solving. **(3) Reasoning**: Determining the model's ability to reason accurately. **(4) Calculation**: Evaluating the model's correctness in mathematical calculations.

Each instance is scored 1 for correct skill application and 0 otherwise. Certain tasks like LSAT and English reading tasks, primarily emphasize understanding not requiring external knowledge or calculations, were excluded from respective skill analyses. This detailed evaluation provides insights into the models' strengths and weaknesses, guiding future improvements of LLMs. Annotators also provided insights into the models' behavior patterns. We summarize the overall trend in the paper and give **detailed analyses about strength and weaknesses in Appendix 5.D**.

5.4.5.1 Overall Trend of Model Capabilities

The average scores on tasks for the four dimensions of capabilities are shown in Fig. 5.3. As shown From the qualitative analysis, we summarize the following observations:

Understanding: The model generally performs well in understanding. For most tasks, it can accurately interpret the meaning of questions, demonstrating its ability to comprehend context.

Knowledge: In the knowledge dimension, the model demonstrates proficiency in identifying correct knowledge or formulas for tasks. However, it encounters difficulties in recalling specific domain knowledge, such as law, biology, and physics. This observation emphasizes the significance of integrating more domain-specific knowledge into the model, potentially through the utilization of specialized domain-specific knowledge bases or knowledge-enhanced pre-training techniques.

Reasoning: Among the four dimensions, the model's reasoning capability appears to be relatively worse. For tasks necessitating complex, multi-step reasoning (e.g., LSAT-AR, LogiQA, and GK-Physics), the model struggles to accurately execute multi-step reasoning process. This underlines the importance of research concentrating on augmenting the model's reasoning capabilities, potentially through the exploration of prompting methods or training strategies that encourage complex reasoning and problem-solving skills.

Calculation: The model's calculation ability is weaker than their understanding capacity and displays variability across different subjects. They perform better in math exams, but face challenges in chemistry and biology exams, which often require variable substitution involving chemical elements. This suggests that enhancing the calculation and combinatorial abstraction and calculation ability of the model, particularly in subject areas with specialized notations or customized symbol substitutions, is a crucial challenge for further improvement.

5.4.6 Data Contamination Issue

The issues surrounding data contamination and future web scrapes on training data for LLMs are noteworthy. Most of current benchmarks and datasets up to date suffer from these vulnerabilities. To exam the situation of contamination, we provided timestamp for the 4 new Gaokao datasets and we can evaluate on the latest tests (later than 2022) released later than the training data timestamp of ChatGPT and GPT-4. Hereinafter, from AGIEval, we provide results comparing the GPT-4 zero-shot performance on six Gaokao subjects with and without risk of data contamination (Chinese, English, and History have not been included in this analysis due to the constrained size of the exams for these subjects). The uncontaminated dataset comprises entries released in 2022, which postdates the GPT-4 training data's timestamp (September 2021). The results are reported on Table 5.4. Evidently, we observe that

barring the Mathematics subjects, the performance experiences a minor drop in the absence of contamination, yet remains proximate to the performances on the complete datasets. This finding substantiates that while AGIEval still retains its value as a useful and effective human-centric benchmark for evaluating the abilities of foundation models against complex human-oriented tasks.

	#test	Full acc.	Un. acc.
Gaokao-geo.	37	76.9%	73%
Gaokao-bio.	58	75.7%	77.6%
Gaokao-chem.	64	51.7%	42.2%
Gaokao-phy.	20	40%	40%

Table 5.4: Analysis on data contamination risk on AGIEval. The uncontaminated set (performance on the last column) includes examples released later than the time stamp of training data of ChatGPT and GPT-4.

5.5 Conclusion

We introduce AGIEval, a novel benchmark specifically designed to assess the general capabilities of large foundation models with respect to human-level cognition. The benchmark comprises high-quality official admission tests, qualification exams, and advanced competitions tailored for human participants, including law school admission tests and college entrance examinations. These assessments establish officially recognized standards for gauging human capabilities, making them well-suited for evaluating foundation models in the context of human-centric tasks. Additionally, AGIEval incorporates bilingual tasks in both Chinese and English, offering a more comprehensive assessment of model behavior. We have carried out an extensive evaluation of three cutting-edge large foundation models: text-davinci-003, ChatGPT, and GPT-4, using AGIEval. Remarkably, GPT-4 surpasses average human performance on LSAT, SAT, and math competition, attaining a 95% accuracy rate on the SAT Math test and a 92.5% accuracy on the Gaokao English test, demonstrating the impressive performance of contemporary foundation models. Despite their significant achievements, our in-depth manual analyses also reveal the limitations of these large language models in terms of

understanding, knowledge utilization, reasoning and calculation. Guided by these findings, we explore potential future research avenues in this domain. By assessing these foundation models on human-centric tasks and probing their capabilities more deeply, we strive to foster the development of models that are more closely aligned with human cognition.

5.6 Limitation

Until the time we finished this work, state-of-the-art foundation models, such as text-davinci-003, ChatGPT, and GPT-4, only have publicly available APIs for language-only tasks. Therefore, we release the language-only version of AGIEval and focus on evaluating a wider range of large language models in the present paper. In the future, we will study on the multi-modal test set.

Appendix

5.A Discussion about Future Directions

In light of the findings and limitations identified in our analysis, we point out several potential future directions for the development of large foundation models. These directions aim to address the weaknesses observed and further improve the models' capabilities in various human-centric tasks.

Inclusion of External Knowledge and Formulas: Enriching the models with external knowledge sources, like formulas and domain-specific knowledge can help enhance their performance in mathematical and knowledge-intensive tasks. Specifically, developing models that can effectively handle domain-specific tasks, such as those in law, biology, or physics, requires the integration of specialized knowledge bases and expertise into the model, and enables the model to adapt to different verticals more effectively. This could involve integrating structured knowledge repositories, mathematical and scientific concepts into the models with pre-training or knowledge-enhanced prompting methods, allowing them to access and apply relevant information more efficiently.

Strict Complex Logical Reasoning: Improving the models' capacity for strict complex logical reasoning is crucial for their performance in a wide range of human-centric tasks. This could involve the creation of new datasets

that emphasize complex reasoning, as well as incorporating APIs and external symbolic compilers that can execute strict logical or mathematical deduction, and use the execution results to further facilitate logical analysis and reasoning verification.

Multi-lingual Reasoning Capabilities Generalization: As mentioned in Sec. 5.4.4, the reasoning capabilities of models are variant across different language, where the reasoning ability is relatively better for rich-resourced language like English. Enhancing the models' multi-lingual reasoning capabilities is essential for their applicability in a diverse range of real-world scenarios. Therefore, future directions can put more focus on enhancing the multilingual generalization of the reasoning capability of foundation models.

Multi-modal Evaluation: Expanding the evaluation framework to include multi-modal tasks can provide a more comprehensive assessment of the models' capabilities. This could involve incorporating visual, auditory, or interactive tasks that require the models to process and reason with multiple types of input simultaneously and generate multi-modal outputs for comprehensive real-world applications. In future work, we will focus on the multi-modal version of AGIEval.

Better Automatic Evaluation Metrics for Human-centric Tasks: Developing more robust and meaningful automatic evaluation metrics is crucial for the objective assessment of large language models' performance. Future research should focus on devising metrics that can accurately capture the models' understanding, knowledge, and reasoning abilities while taking into account the nuances and complexities of real-world tasks.

Robustness of Reasoning Capability: Improving the robustness of the models' reasoning capabilities is essential for ensuring their consistency and reliability across various contexts. This can be achieved by exploring techniques that enhance the models' ability to maintain consistent reasoning performance, even when faced with changes in the surrounding context or variations in the input data.

By addressing these future directions, foundation models can be further developed and refined to exhibit more advanced capabilities that align closely with human cognition, ultimately enabling them to tackle a broader range of complex, human-centric tasks with greater accuracy and reliability.

5.B Dataset Collection

As previously mentioned, our human-centric benchmark comprises questions from a diverse range of official and high-quality exams, originally designed for human test-takers. These exams include general college admission tests (GRE, Gaokao, SAT), entrance exams for specific majors (such as LSAT and GMAT), high school math competitions (AMC and AIME), as well as the national civil service examination and lawyer qualification test in China.

Since evaluating model performance on subjective questions is challenging without human expert scoring, we believe such questions are unsuitable for inclusion in this benchmark for consistent assessment. To ensure a robust and standardized evaluation metric, we have removed all subjective questions, retaining only objective ones, such as multiple-choice and fill-in-the-blank questions.

With regard to data collection, we gather Gaokao[2] and SAT questions[3] from publicly available online sources, along with their corresponding solutions or explanations. Throughout our data collection phase, we encountered various challenges. Consider the instance of Gaokao: our approach encompassed not only discerning reliable sources while respecting copyright regulations, but also the annotation and removal of examples with multi-modal components, elimination of duplications, identification of items unsuitable for the QA format, as well as reformatting and connecting passages and questions. Furthermore, we invite professional human experts to manually check the correctness of latex formula in each question and answer, to ensure the correctness and robustness of QA pairs.

For the LSAT, we utilize data from Wang et al. (2022) and Zhong et al. (2022), which encompasses three tasks (logical reasoning, reading comprehension, and analytical reasoning) from the LSAT administered between 1991 and 2016. For Chinese civil service examinations, we repurpose data from LogiQA (Liu et al., 2021b), a dataset built on various types of logical reasoning questions collected from the National Civil Servants Examination of China. It is worth noting that LogiQA consists of bilingual questions (English and Chinese), where the English version is a translated version of the original Chinese version.

[2]Gaokao questions are collected from officially announced exam questions and answers like `http://www.hbccks.cn/html/gkgzzt/ggsjjda/`.

[3]`https://satsuite.collegeboard.org/sat/practice-preparation/practice-tests/paper`

For high school math competitions, we employ data from the MATH dataset (Hendrycks et al.), comprising questions from AMC and AIME. Furthermore, we incorporate GRE and GMAT questions from AQaA-RAT (Ling et al., 2017), which emphasizes algebraic word problems. In the case of the Chinese Civil Service Examination, we reuse instances from JEC-QA (Zhong et al., 2020), a large-scale dataset derived from the National Judicial Examination of China. We down-sample the two types of JEC-QA and MATH to 1,000 instances each.

As a result, we construct a benchmark consisting of 8,062 questions for evaluation. Detailed data statistics are presented in Table 5.1. It is worth noting that our benchmark is bilingual, encompassing both **English and Chinese tests**. This design enables the evaluation of a broader scope of model capabilities, reflecting their performance and adaptability across different languages. A few data examples in Gaokao are shown in Fig. 5.4, and an example in SAT and the corresponding Chain-of-Thought reasoning process generated by GPT-4 is shown in Fig. 5.5.

5.C Implementation Details

5.C.1 API Details

All experiments were conducted using the respective language models' API provided by Azure OpenAI Service[4]. The Azure OpenAI services offer two types of APIs: completion and chat completion. The completion API generates text based on prompts, while the chat completion API generates the next AI response based on the conversation history and new human input. For text-davinci-003 and few-shot ChatGPT, we use the completion API, and for zero-shot ChatGPT and GPT-4, we use the chat completion API. Notably, only the chat completion API is available for GPT-4 at present. We use a temperature of zero to generate output using greedy search and set the maximum number of tokens for generation to 2048. Additionally, we set the frequency penalty to zero and top p to 1, which are the default values for these APIs.

The Chat Completion API exhibits distinct properties in comparison to the Completion API. In a zero-shot context, the Chat Completion API

[4] `https://azure.microsoft.com/en-us/products/cognitive-services/openai-service`

has the potential to autonomously generate reasoning steps, eliminating the necessity for prompt engineering and potentially enhancing performance. For few-shot scenarios, it is imperative to adapt the few-shot examples into conversational history, as recommended in the Azure guidelines. The inquiry is transformed into a user input, while the AI's response is composed of a chain-of-thought explanation and answer. However, we have observed that the models, particularly ChatGPT, encounter difficulties in adhering to the pattern using the Chat Completion API. Consequently, we employ the Completion API to conduct few-shot experiments with ChatGPT, which is analogous to text-davinci-003, in order to gain a deeper understanding of the disparities between text-davinci-003 and ChatGPT. If a completion API for GPT-4 become accessible in the future, we will revise and update the few-shot outcomes accordingly.

5.C.2 Few-shot Examples Construction:

For AQuA-RAT, LogiQA and LSAT, we randomly sample five examples of medium sentence length of the test set from the provided training set. Similarly, for Gaokao and SAT, we randomly select five examples of medium sentence length from the dataset that was initially collected and exclude them from the test set. For JEC-QA, given that the test set is not publicly available, we take the first 1,000 examples from the training set as the test set and again sample five examples of medium sentence length from the rest. For MATH, we use the same instances as in the appendices of Lewkowycz et al. (2022).

To generate explanations for few-shot CoT experiments, for AQuA-RAT and MATH, we use the existing rationales from these datasets. For Gaokao and SAT, we collected expert annotations. For LogiQA, JEC-QA and LSAT, we use ChatGPT to generate explanations given the questions and the answers. We release all CoT demonstrations in the Github repository.

5.D Qualitative Analysis Details

By closely examining the models' output explanations and analyzing their behavior patterns, we identify several strengths that highlight the capabilities of these models in handling various aspects of problem-solving. The models demonstrate remarkable performance in the following areas:

Good Understanding: The models excel in accurately comprehending the semantic meaning of context and questions. They effectively discern nuances, interpret complex questions, and parse intricate sentences, showcasing their strong natural language understanding skills. This capability enables them to grasp the core concepts of a problem and lays the foundation for subsequent reasoning and problem-solving steps.

Proficiency in Simple Reasoning and Deduction: The models are adept at handling tasks that require simple reasoning and deduction. They can draw straightforward conclusions, identify logical connections, and perform basic inference, which is crucial for addressing a wide variety of problems. Their ability to effectively perform simple reasoning tasks is an essential component of their overall problem-solving skillset.

For example, the model can comprehend "could be true except" is equals to "cannot be true". Also, taking a question in the LSAT-AR task as an example, it requires the model to place 8 books to a bookcase with three shelves following conditions: *"each shelf should have at least 2 books and more books should be placed on the bottom shelf than on the top shelf"*. The model successfully deduced that "*there are at least 3 books on the bottom shelf and at most 2 books on the top shelf.*"

Grasping General Reasoning Process: The models demonstrate an ability to understand and generate the general idea of reasoning processes. They can identify the main components of a problem, recognize the structure of a solution, and outline a high-level reasoning strategy. This capability allows them to generate meaningful explanations and provides a starting point for more detailed reasoning and problem-solving tasks.

These strengths indicate that the models have made significant progress in aligning with human problem-solving capabilities. However, there is still room for improvement, especially in complex reasoning tasks and domain-specific knowledge, as discussed in the subsequent section on weaknesses.

5.D.1 Weaknesses

Despite the significant strengths displayed by the models, there are certain limitations that need to be addressed to improve their overall performance. We outline these weaknesses based on the analysis of the models' output explanations:

Understanding:

- *Difficulty with Variable Substitution:* The models struggle to understand questions that require variable substitution, often failing to recognize the need for this operation and how it should be applied to solve the problem. This limitation can hinder their ability to tackle a wide range of mathematical and logical tasks. For instance, the model frequently struggles to answer chemistry questions that involve substituting a variable in a chemical equation with a chemical element and analyzing its properties.

- *Challenges with Complex Math Concepts and Symbols:* The models find it difficult to comprehend complex mathematical concepts and interpret the meaning of symbols, particularly when multiple symbols are involved. This weakness limits their ability to effectively address advanced mathematical problems.

- *Confusion with Similar Concepts:* The models can easily be confused by similar concepts or terms, sometimes leading to incorrect or misleading reasoning. For example, in the physics exam, the model is confused by the difference between vertical speed and horizontal speed of moving object. This issue underscores the need for better disambiguation and concept understanding techniques in future model iterations.

- *Difficulty in Handling Long Contexts:* The models are prone to being disrupted by long contexts, leading to a decline in their comprehension and reasoning abilities. Improving the models' capacity to maintain focus and process extensive information is essential for enhancing their performance in real-world scenarios.

Knowledge:

- *Insufficiency in Commonsense and Domain-Specific Knowledge:* The models occasionally demonstrate a lack of commonsense or domain-specific knowledge, which hinders their ability to generate plausible explanations and provide accurate answers. This limitation underscores the importance of incorporating diverse knowledge sources into the training data and exploring techniques that can more effectively integrate and access this information within the models. Moreover, it emphasizes the necessity to broaden the models' exposure to a wider array of subjects and fields, ensuring a more comprehensive understanding of various domains.

For instance, given the conditions "*if Julio and Kevin both lead morning sessions, we know that Kevin and Rebecca must lead sessions that meet on the same day*," the model incorrectly deduces that "*Therefore, Rebecca must also lead a morning session.*" This indicates a lack of commonsense knowledge about the relationship between *morning* and *day*, leading to an erroneous explanation. Additionally, the model generally performs poorly on tasks requiring specific domain knowledge, such as law and chemistry.

- *Difficulty Identifying Correct Formulas:* The models occasionally struggle to recall and apply the appropriate formulas necessary to solve particular problems, especially in tasks that demand specialized knowledge or expertise. This shortcoming suggests that there is potential for improvement in the models' knowledge retrieval mechanisms and their ability to recognize the relevance of specific formulas to a given problem. Developing strategies to enhance the models' proficiency in identifying and applying correct formulas will be essential for improving their performance in tasks requiring a deep understanding of domain-specific concepts and techniques.

Addressing these weaknesses in knowledge will contribute to the development of more robust and versatile large language models, better equipped to tackle a broader range of human-centric tasks and exhibit a more comprehensive understanding of various domains.

Reasoning:

- *Challenges in Strict Logical Deduction:* The models frequently encounter difficulties when attempting to perform strict logical deduction accurately. Common issues include ignoring premise conditions, misconstruing sufficient and necessary conditions, or making errors in logical chaining. These types of errors are commonly observed in manual analyses.

  For instance, given a condition, "*If Myers is on the team, neither Ortega nor Paine can be*", and a solution, "*Ortega, Paine, Thomson, and Zayre are on the team*", the model incorrectly states that this solution is wrong because "*Paine and Ortega are on the team*", neglecting to first satisfy the premise condition "*If Myers is on the team*". Furthermore, the model

demonstrates a misunderstanding of the difference between sufficient and necessary conditions in its explanation of another question and states: "*If Kayne is assigned to an ambassadorship, then so is Jaramillo. This constraint is essentially the same as the given constraint that if Jaramillo is assigned to one of the ambassadorships, then so is Kayne*".

To address these limitations, it is essential to improve the models' abilities to recognize and apply logical rules and refine their understanding of logical structures.

- *Difficulty with Counterfactual Reasoning:* The models consistently struggle with counterfactual reasoning tasks. They have difficulty generating alternative scenarios, evaluating hypothetical outcomes, or exploring potential consequences based on varying assumptions. For instance, the models frequently make incorrect judgments for counterfactual questions in the LSAT-AR task: "*Which one of the following, if substituted for the constraint that [Constraint A], would have the same effect in determining the assignment?*" Enhancing the models' capabilities in handling counterfactual reasoning tasks is vital for developing a more comprehensive problem-solving skillset.

- *Struggles in Multi-hop Complex Reasoning:* The models have difficulty accurately executing multi-hop complex reasoning tasks, often displaying inconsistent logic, omitting inference steps, or producing flawed reasoning chains. To address a broader range of complex problems, it is crucial to improve the models' abilities to systematically navigate and process multi-step reasoning tasks.

- *Establishing Incorrect Conclusions and Contradictory Reasoning:* The models occasionally set an incorrect conclusion first and then generate contradictory reasoning based on that faulty foundation. This behavior emphasizes the need for improved reasoning verification and error correction techniques in the models' problem-solving processes.

- *Concealed Substitution of Concepts:* The models sometimes covertly substitute one concept with another similar one, leading to inaccurate or misleading reasoning. For example, in a biology exam, the model replaces the concept of "*isotopically labeled amino acids*" with "*isotopically labeled tRNA (a tool for transporting amino acids)*", resulting

in erroneous reasoning. This issue underscores the importance of better concept disambiguation and reasoning coherence in future model iterations.

- *Difficulty in Identifying Solutions:* The models occasionally struggle to discover feasible solutions for specific problems, possibly due to limitations in their knowledge, reasoning capabilities, or problem-solving strategies. Addressing this shortcoming involves refining the models' ability to explore, evaluate, and select appropriate solutions based on the given problem context.

- *Vulnerability to Contextual Disturbance:* The reasoning ability of large language models is often easily disrupted by changes in the surrounding context. When the context is modified, the models may produce different deductions for the same condition, suggesting that the robustness of their reasoning ability is not yet sufficient. This observation emphasizes the need to develop models that can maintain consistent reasoning performance, even in the presence of varying contextual information, ensuring more reliable and stable problem-solving capabilities.

Calculation: The model is prone to making calculation errors, particularly when dealing with complex variable substitutions. This may be attributed to the inherent limitations of the model's computation process in handling mathematical operations, as well as its difficulty in parsing intricate relationships between variables. Consequently, the model may struggle to maintain accuracy and precision when attempting to solve problems involving advanced algebraic manipulations or multi-step calculations. To address this issue, future iterations of the model should focus on enhancing its mathematical reasoning capabilities and improving its ability to recognize and apply relevant mathematical rules. This could involve incorporating specialized modules or mechanisms specifically designed to handle complex calculations, variable substitutions, and numerical problem-solving tasks. By refining the model's ability to accurately process and solve intricate mathematical problems, we can expand its applicability across a broader range of disciplines and domains, ensuring a more comprehensive and robust problem-solving skillset.

By addressing these reasoning weaknesses, future large language models can be developed with more robust problem-solving capabilities, enabling

them to effectively tackle a broader range of human-centric tasks and exhibit more sophisticated reasoning skills that align closely with human cognition.

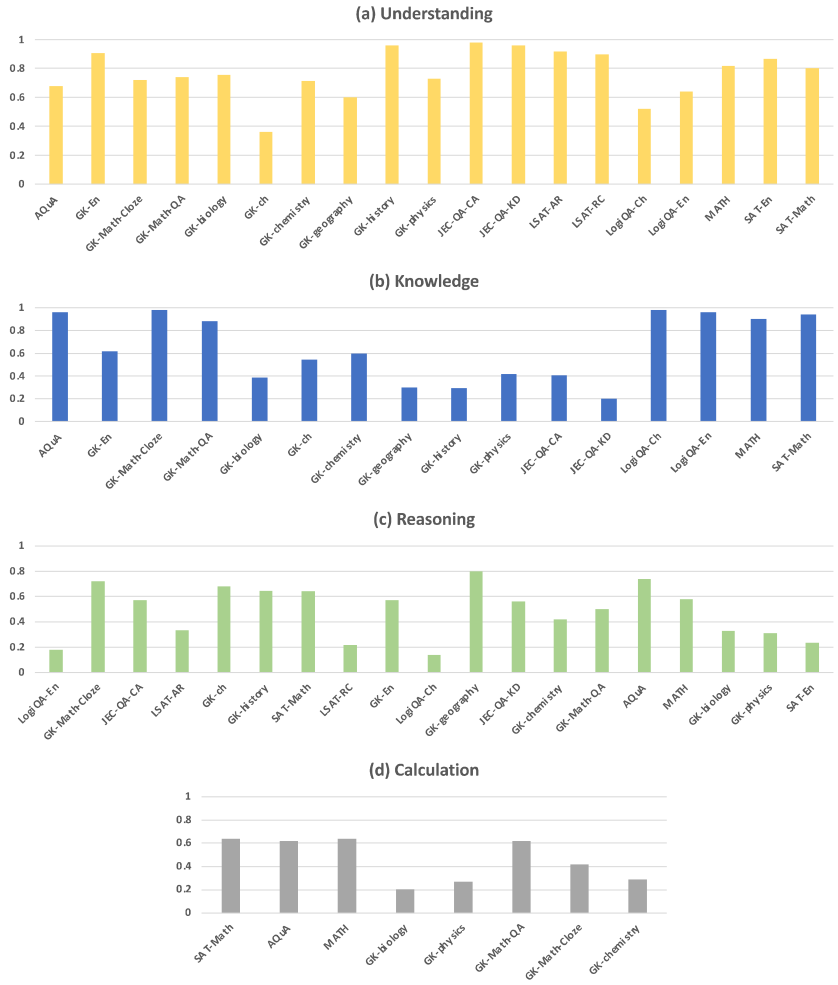


Figure 5.3: Qualitative assessment of inaccurately answered questions by the model focuses on four dimensions of capabilities: understanding, knowledge acquisition, reasoning and calculation.

Example in Gaokao-MathQA

Question:设 O 为坐标原点, 直线 $x=a$ 与双曲线 $C:\frac{x^2}{a^2}-\frac{y^2}{b^2}=1(a>0,b>0)$ 的两条渐近线分别交于 D,E 两点, 若 $\triangle ODE$ 的面积为 8 , 则 C 的焦距的最小值为 ( )?
(Let O be the origin of the coordinate system, and let the line $x=a$ intersect the two asymptotes of the hyperbola $C:\frac{x^2}{a^2}-\frac{y^2}{b^2}=1(a>0,b>0)$ at points D and E. If the area of triangle $\triangle ODE$ is 8, what is the minimum value of the focal length of C?)
Options: (A)4, (B)8, (C)16, (D)32
Answer: (B)

Example in Gaokao-Biology

Question:人体下丘脑具有内分泌功能, 也是一些调节中枢的所在部位。下列有关下丘脑的叙述, 错误的是选项 (The hypothalamus in the human body has endocrine functions and is also the location of some regulatory centers. Which of the following statements about the hypothalamus is incorrect?)
Options:
(A)下丘脑能感受细胞外液渗透压的变化 (The hypothalamus can sense changes in the osmotic pressure of extracellular fluid)
(B)下丘脑能分泌抗利尿激素和促甲状腺激素 (The hypothalamus can secrete antidiuretic hormone and thyroid-stimulating hormone)
(C)下丘脑参与水盐平衡的调节: 下丘脑有水平衡调节中枢 (The hypothalamus is involved in the regulation of water-salt balance; the hypothalamus has a water balance regulation center)
(D)下丘脑能感受体温的变化; 下丘脑有体温调节中枢 (The hypothalamus can sense changes in body temperature; the hypothalamus has a body temperature regulation center)
Answer: (B)

Figure 5.4: Data examples in Gaokao.

Example in SAT-Math

Question: $\begin{matrix}3x+4y=-23\\2y-x=-19\end{matrix}$ What is the solution (x,y) to the system of equations above?
Options: (A). (-5,-2) (B). (3,-8) (C). (4,-6) (D). (9,-6)
Answer: (B)

Chain-of-Thought Reasoning by GPT-4

Step 1: Rewrite the second equation to isolate x.
Given: $\begin{matrix}3x+4y=-23\\2y-x=-19\end{matrix}$ Rewrite the second equation as $x=2y+19$
Step 2: Substitute the expression for x from the second equation into the first equation.
Substitute x in the first equation: $3(2y+19)+4y=-23$
Step 3: Solve for y.
Distribute the 3: $6y+57+4y=-23$
Combine like terms: $10y+57=-23$
Subtract 57 from both sides: $10y=-80$
Divide by 10: $y=-8$
Step 4: Substitute the value of y back into the expression for x. $x=2(-8)+19$
Step 5: Solve for x. $x=3$
Step 6: Write the solution as an ordered pair (x,y)
The solution to the system of equations is (3, -8), which corresponds to answer choice (B).

Figure 5.5: Data example in SAT and corresponding CoT reasoning process.

Chapter 6

Conclusion

This thesis explores various facets of language model evaluation, reflecting the evolving landscape of both LLMs and their assessment. From multilingual compositional generalization, our research progresses towards linguistic theory-motivated probing and benchmarking against human-level general abilities.

A unifying focus throughout this work is the exploration of generalizability in neural networks. Chapter 2 examined the multilingual dimensions of generalizability, investigating how well models trained on entities in multiple languages transfer knowledge across them. Chapter 3 leveraged the theoretical framework of generalized quantifiers and revealed systematic errors in LM performance on natural language inference tasks involving quantifier reasoning. This highlighted the limitations of current models in generalizing to unseen quantifier constructions and reasoning over them. Chapter 4 further investigated linguistic phenomena by dissecting the challenges LMs face in comprehending respective readings. This chapter identified specific failure cases and explored training strategies to enhance the models' ability to generalize this understanding. Finally, Chapter 5 culminated in an investigation of how LLMs, immersed with knowledge from the vast World Wide Web, fare when tasked with reasoning challenges at a human level.

In conclusion, this thesis provides valuable insights into the strengths and limitations of contemporary language models, particularly concerning their generalizability. We designed datasets and probe model capabilities across diverse linguistic phenomena and reasoning tasks. This research contributes to developing more robust and reliable AI models and, ultimately, superintelligences.

Multilingual Compositional Generalization In pursuing a future without language barriers, researchers hypothesize that models should be able to communicate and solve problems equally for every language. However, the availability of training resources varies significantly across languages, with English being the most commonly studied and accepted as the universal language for research communication. Consequently, it is essential to develop models capable of effectively transferring learned knowledge from one language to another, treating all individuals equally regardless of their spoken language. Previous works have focused on English tasks in compositional generalization in neural networks. For example, SCAN is constructed for English to synthetic translation, while CFQ is based on English to SPARQL parsing of an outdated knowledge base.

In Chapter 2, we present pioneering research on the compositional generalization ability of neural networks and language models in the context of cross-lingual knowledge transfer. Specifically, we investigate the semantic parsing task of text to SPARQL. We propose a method to migrate the CFQ dataset to Wikidata and extend it to three typologically distant languages and domains. The resulting dataset, MCWQ, enables researchers to perform analyses and improve the compositionality of models. Our monolingual experiments suggest that current models can achieve a degree of compositional generalization but fail to perform zero-shot transfer to other languages.

Generalized Quantifier Probing Generalized quantifiers (GQs) are ubiquitous in human language and lie at the intersection of logic, linguistics, and NLP research. While AI researchers have recently focused more on the mathematical reasoning abilities of LLMs, they have largely overlooked the phenomenon of quantifiers in natural languages. In Chapter 3, we derive a categorization set for quantifier analysis based on quantifier taxonomy studies. Our findings reveal that GQs account for 10% of the corpora in natural language inference and question-answering tasks, and they systematically contribute to model failure. We analyze the multilingual performance of models and highlight discrepancies in quantifier reasoning across different languages. In the end, we introduce GQNLI, the first benchmark in modern NLP research designed to study the GQ reasoning abilities of LMs. Our results demonstrate that GQ reasoning remains challenging for models, particularly with proportional quantifiers compared to existential or universal quantifiers. We also observe that models lack robustness in GFQ reasoning

when subjected to simple ablation studies. The challenge becomes even more significant when GQs interact with other logical operators. These findings underscore the need for further research and development of models that can effectively handle the complexities of GQ reasoning.

Compositional Learning of Linguistic Constructions: Respective Readings Researchers have attempted to investigate specific aspects of linguistic reasoning in LMs over the years. For instance, negation has been extensively studied in the NLP community. However, countless linguistic constructions, particularly those with long tails, cannot receive such thorough examination. In Chapter 4, we focus on the respective reading construction as a representative example to study compositional learning of linguistic constructions. Despite its simplicity, it requires multiple levels of reasoning skills, including syntactic-semantic and commonsense-world knowledge. We examine the ability of LMs to generalize from simple to complex linguistic constructions in zero-shot and few-shot settings. Our findings indicate that LMs struggle to comprehend respective readings without explicit supervision and learn more effectively with explicit reasoning than implicit reasoning. Nevertheless, they can achieve perfect generalization with a relatively large number of training examples. This confirms that diverse and complex training data are vital for attaining human-level language skills.

Benchmarking Human-level Complex Reasoning As LLMs advance, traditional benchmarks such as SQuAD and GLUE have become increasingly saturated, with models achieving near-perfect performance. This underscores the need for more challenging, human-centric benchmarks that reflect real-world tasks. In Chapter 5, we introduce AGIEval, a benchmark specifically designed to assess the general cognitive abilities of LLMs in tasks that mirror human problem-solving and cognition. Our benchmark comprises 20 high-quality tasks, including general college admission tests, law school admission tests, math competitions, lawyer qualification tests, and national civil service exams, all designed for human test-takers. AGIEval covers bilingual tasks in both Chinese and English. Our evaluation of state-of-the-art models, including GPT-4, reveals that while they excel in certain areas, they struggle with tasks that require more complex reasoning or domain-specific knowledge. Through qualitative analyses across four key aspects - understanding, knowledge, reasoning, and calculation - we shed light on the strengths and weaknesses of

LLMs, providing valuable insights to inform future development directions.

6.1 Future Directions

Benchmarking has made significant progress in recent years. As the capabilities of language models continue to expand in various dimensions, researchers must stay ahead of the curve and anticipate future challenges. In compositional generalization studies, existing English-based benchmarks such as CFQ have been perfectly solved (Drozdov et al., 2023). However, limited effort has been dedicated to enhancing the cross-lingual generalizability of LLMs (Zhang et al., 2023). It is important to investigate the influence of neural architectures and training data on the cross-lingual generalization abilities of LLMs. This line of research can lead to the development of more robust and inclusive models that perform well across a diverse range of languages.

Regarding generalized quantifiers, there is a pressing need for more extensive and more systematically designed benchmarks to facilitate comprehensive evaluation. Recent works have proposed promising directions to address this gap. For instance, Madusanka et al. (2023) construct a large-scale GQ dataset using algorithms defined purely logically. This approach offers a principled and scalable way to generate diverse and challenging test cases. On the other hand, Li et al. (2023) crowd-source a dataset of human-annotated generalized quantifiers in Wikipedia sentences, capturing the nuances and complexities of real-world language use. Despite approaching the research question from different perspectives, both works converge on the conclusion that there is significant room for improvement in the reasoning capabilities of LLMs when it comes to generalized quantifiers. Another important direction is to investigate the relations between quantifier reasoning and world knowledge, especially multimodal knowledge. Therefore, developing multimodal quantifier datasets and evaluating multimodal models is the next straightforward step.

To achieve a more comprehensive understanding of LLMs' linguistic abilities, it is necessary to foster collaboration between LLM trainers and linguists. This interdisciplinary approach will enable a rigorous evaluation and analysis of how LLMs learn, generalize, and process language at a level comparable to human language understanding. We can gain deeper insights into the strengths and limitations of these models, identifying the linguistic phenomena that LLMs capture effectively and the areas where they struggle or exhibit biases. To effectively assess the linguistic reasoning abilities of language models,

particularly for low-resource languages, it is crucial to design more challenging and comprehensive benchmarks. One valuable resource that can be leveraged is the International Linguistics Olympiad (IOL), an annual competition that tackles self-sufficient linguistics problems in theoretical, mathematical, and applied linguistics. Researchers can develop benchmarks that are challenging and grounded in real-world linguistic phenomena by drawing inspiration from the IOL problem sets.

As LLMs advance rapidly, it is important to consider three key aspects when creating comprehensive and effective benchmarks. Firstly, benchmarks should be dynamic, allowing for continuous updates and adaptation to reflect the language's evolving nature and usage. This can be achieved by incorporating real-time data sources and crowdsourced annotations. Secondly, benchmarks should cover a range of everyday life domains, such as agent-based interactions, medical consultations, and legal discourse, to ensure that LLMs are evaluated on their ability to understand and generate text relevant to real-world applications. Finally, benchmarks should prioritize inclusivity, encompassing the breadth and depth of multilingual and multicultural aspects, including diverse linguistic, cultural, and socio-economic perspectives, to guarantee that LLMs serve the needs of a global and diverse user base.

Bibliography

Mohamed Abdalla and Graeme Hirst. 2017. Cross-lingual sentiment analysis without (good) translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 506–515, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Mostafa Abdou, Vinit Ravishankar, Maria Barrett, Yonatan Belinkov, Desmond Elliott, and Anders Søgaard. 2020. The sensitivity of language models and humans to Winograd schema perturbations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7590–7604, Online. Association for Computational Linguistics.

Syed S Ali and Stuart C Shapiro. 1993. Natural language processing using a propositional semantic network with structured variables. *Minds and machines*, 3(4):421–451.

Hiyan Alshawi and Jan van Eijck. 1989. Logical forms in the core language engine. In *27th Annual Meeting of the Association for Computational Linguistics*, pages 25–32, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Chinatsu Aone. 1991. Resolution of collective-distributive ambiguity using model-based reasoning. In *29th Annual Meeting of the Association for Computational Linguistics*, pages 1–8, Berkeley, California, USA. Association for Computational Linguistics.

Rahul Aralikatte, Heather Lent, Ana Valeria Gonzalez, Daniel Hershcovich, Chen Qiu, Anders Sandholm, Michael Ringaard, and Anders Søgaard. 2019. Rewarding coreference resolvers for being consistent with world knowledge. 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 1229–1235, Hong Kong, China. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020a. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020b. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Akari Asai and Hannaneh Hajishirzi. 2020. Logic-guided data augmentation and regularization for consistent question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5642–5650, Online. Association for Computational Linguistics.

Emmon W. Bach, Eloise Jelinek, Angelika Kratzer, and Barbara H. Partee. 1995. Quantification in natural languages.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Jon Barwise and Robin Cooper. 1981. Generalized quantifiers and natural language. In *Philosophy, language, and artificial intelligence*, pages 241–301. Springer.

Edward Beeching, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open llm leaderboard. `https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard`.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. *Advances in neural information processing systems*, 13.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics.

Patrick Blackburn, Johan van Benthem, and Frank Wolter. 2006. *Handbook of Modal Logic*. Elsevier.

Daniel G Bobrow, Ronald M Kaplan, Martin Kay, Donald A Norman, Henry Thompson, and Terry Winograd. 1977. Gus, a frame-driven dialog system. *Artificial intelligence*, 8(2):155–173.

Margaret A Boden. 2016. *AI: Its nature and future*. Oxford University Press.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, page 1247–1250, New York, NY, USA. Association for Computing Machinery.

Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale simple question answering with memory networks. *arXiv preprint arXiv:1506.02075*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015a. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Samuel R. Bowman, Christopher Potts, and Christopher D. Manning. 2015b. Recursive neural networks can learn logical semantics. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, pages 12–21, Beijing, China. Association for Computational Linguistics.

Stephanie Brandl, Ruixiang Cui, and Anders Søgaard. 2022. 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. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Shulin Cao, Jiaxin Shi, Liangming Pan, Lunyiu Nie, Yutong Xiang, Lei Hou, Juanzi Li, Bin He, and Hanwang Zhang. 2022. KQA pro: A dataset with explicit compositional programs for complex question answering over knowledge base. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6101–6119, Dublin, Ireland. Association for Computational Linguistics.

Yong Cao, Yova Kementchedjhieva, Ruixiang Cui, Antonia Karamolegkou, Li Zhou, Megan Dare, Lucia Donatelli, and Daniel Hershcovich. 2024. Cultural adaptation of recipes. *Transactions of the Association for Computational Linguistics*, 12:80–99.

Lucas Champollion. 2015. Ten Men and Women Got Married Today: Noun Coordination and the Intersective Theory of Conjunction. *Journal of Semantics*, 33(3):561–622.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A

survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Rui Pedro Chaves. 2012. Conjunction, cumulation and respectively readings. *Journal of Linguistics*, 48:297 – 344.

Jianpeng Cheng and Mirella Lapata. 2018. Weakly-supervised neural semantic parsing with a generative ranker. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 356–367, Brussels, Belgium. Association for Computational Linguistics.

Lisa Lai-Shen Cheng. 2009. On every type of quantificational expression in chinese. *Quantification, definiteness, and nominalization*, 5375.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Noam Chomsky. 1957. *Syntactic structures*. Mouton de Gruyter.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. 11. MIT press.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416.

Andy Clark. 2006. Language, embodiment, and the cognitive niche. *Trends in cognitive sciences*, 10(8):370–374.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *ArXiv*, abs/2110.14168.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Robin Cooper, Richard Crouch, Jan Van Eijck, Chris Fox, Josef Van Genabith, Jan Jaspers, Hans Kamp, Manfred Pinkal, Massimo Poesio, Stephen Pulman, et al. 1994. Fracas: A framework for computational semantics. *Deliverable D6*.

Ruixiang Cui, Rahul Aralikatte, Heather Lent, and Daniel Hershcovich. 2022a. Compositional generalization in multilingual semantic parsing over Wikidata. *Transactions of the Association for Computational Linguistics*, 10:937–955.

Ruixiang Cui and Daniel Hershcovich. 2021a. Great service! fine-grained parsing of implicit arguments. In *Proceedings of the 17th International Conference on Parsing Technologies (IWPT 2021)*, pages 65–77, Online. Association for Computational Linguistics.

Ruixiang Cui and Daniel Hershcovich. 2021b. Meaning representation of numeric fused-heads in ucca. In *Proceedings of the 1st Workshop on Understanding Implicit and Underspecified Language*, Online. Association for Computational Linguistics.

Ruixiang Cui, Daniel Hershcovich, and Anders Søgaard. 2022b. Generalized quantifiers as a source of error in multilingual NLU benchmarks. In

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4875–4893, Seattle, United States. Association for Computational Linguistics.

Ruixiang Cui, Seolhwa Lee, Daniel Hershcovich, and Anders Søgaard. 2023. What does the failure to reason with "respectively" in zero/few-shot settings tell us about language models? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8786–8800, Toronto, Canada. Association for Computational Linguistics.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–220.

Mary Dalrymple and Andrew Kehler. 1995. On the constraints imposed by respectively. *Linguistic Inquiry*, pages 531–536.

Rajarshi Das, Manzil Zaheer, Dung Thai, Ameya Godbole, Ethan Perez, Jay Yoon Lee, Lizhen Tan, Lazaros Polymenakos, and Andrew McCallum. 2021. Case-based reasoning for natural language queries over knowledge bases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9594–9611, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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. Association for Computational Linguistics.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 862–872.

Martin d'Hoffschmidt, Wacim Belblidia, Quentin Heinrich, Tom Brendlé, and Maxime Vidal. 2020. FQuAD: French question answering dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1193–1208, Online. Association for Computational Linguistics.

Dennis Diefenbach, Thomas Pellissier Tanon, K. Singh, and P. Maret. 2017. Question answering benchmarks for Wikidata. In *International Semantic Web Conference*.

Xin Dong and Gerard de Melo. 2019. A robust self-learning framework for cross-lingual 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 6306–6310, Hong Kong, China. Association for Computational Linguistics.

David Dowty et al. 1987. Collective predicates, distributive predicates, and all. In *Proceedings of the 3rd ESCOL*, pages 97–115. Ohio State University Ohio.

Andrew Drozdov, Nathanael Schärli, Ekin Akyürek, Nathan Scales, Xinying Song, Xinyun Chen, Olivier Bousquet, and Denny Zhou. 2023. Compositional semantic parsing with large language models. In *The Eleventh International Conference on Learning Representations*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. 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 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Philipp Dufter and Hinrich Schütze. 2020. Identifying elements essential for BERT's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437, Online. Association for Computational Linguistics.

Antonín Dvořák and Michal Holčapek. 2015. Type 1 , 1 fuzzy quantifiers determined by fuzzy measures on residuated lattices. part iii. extension, conservativity and extensionality. *Fuzzy Sets Syst.*, 271(C):133–155.

Randall Eggert. 2000. Grammaticality and context with respect to and... and or... respectively. In *Proceedings of the Thirty-Sixth Annual Meeting of the Chicago Linguistics Society (CLS-36)*.

Ahmed Elgammal, Bingchen Liu, Mohamed Elhoseiny, and Marian Mazzone. 2017. Can: Creative adversarial networks, generating "art" by learning about styles and deviating from style norms.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

Kawin Ethayarajh and Dan Jurafsky. 2020. 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. Association for Computational Linguistics.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Nicholas Evans and Stephen C Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and brain sciences*, 32(5):429–448.

Richard Evans, David Saxton, David Amos, Pushmeet Kohli, and Edward Grefenstette. 2018. Can neural networks understand logical entailment? *arXiv preprint arXiv:1802.08535*.

Lei Fang and Jian-Guang Lou. 2021. Part & whole extraction: Towards a deep understanding of quantitative facts for percentages in text. *ArXiv*, abs/2110.13505.

Kai Fintel and Lisa Matthewson. 2008. Universals in semantics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Jerry A Fodor and Zenon W Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71.

Gottlob Frege. 1879. *Begriffsschrift: Eine der Arithmetischen Nachgebildete Formelsprache des Reinen Denkens*. Halle a.d.S.: Louis Nebert.

Yu Gai, Paras Jain, Wendi Zhang, Joseph Gonzalez, Dawn Song, and Ion Stoica. 2021. Grounded graph decoding improves compositional generalization in question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1829–1838, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nicolas Garneau, Mareike Hartmann, Anders Sandholm, Sebastian Ruder, Ivan Vulic, and Anders Søgaard. 2021. Analogy training multilingual encoders. In *AAAI*.

Jean Mark Gawron and Andrew Kehler. 2004. The semantics of respective readings, conjunction, and filler-gap dependencies. *Linguistics and Philosophy*, 27:169–207.

Atticus Geiger, Ignacio Cases, Lauri Karttunen, and Christopher Potts. 2019. Posing fair generalization tasks for natural language inference. 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 4485–4495, Hong Kong, China. Association for Computational Linguistics.

Mor Geva, Ankit Gupta, and Jonathan Berant. 2020. Injecting numerical reasoning skills into language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 946–958, Online. Association for Computational Linguistics.

Mor Geva, Tomer Wolfson, and Jonathan Berant. 2022. Break, Perturb, Build: Automatic Perturbation of Reasoning Paths Through Question Decomposition. *Transactions of the Association for Computational Linguistics*, 10:111–126.

David Gil. 1982. Quantifier scope, linguistic variation, and natural language semantics. *Linguistics and Philosophy*, 5(4):421–472.

Grant Goodall. 1987. *Parallel Structures in Syntax: Coordination, Causatives, and Restructuring*. Cambridge University Press.

Emily Goodwin, Siva Reddy, Timothy J. O'Donnell, and Dzmitry Bahdanau. 2021. Compositional generalization in dependency parsing.

Paul Grice. 1989. *Studies in the Way of Words*. Harvard University Press.

Yu Gu, Sue Kase, Michelle Vanni, Brian Sadler, Percy Liang, Xifeng Yan, and Yu Su. 2021. Beyond iid: three levels of generalization for question answering on knowledge bases. In *Proceedings of the Web Conference 2021*, pages 3477–3488.

Yinuo Guo, Zeqi Lin, Jian-Guang Lou, and Dongmei Zhang. 2020. Hierarchical poset decoding for compositional generalization in language. *Advances in Neural Information Processing Systems*, 33:6913–6924.

Yinuo Guo, Hualei Zhu, Zeqi Lin, Bei Chen, Jian-Guang Lou, and Dongmei Zhang. 2021. Revisiting iterative back-translation from the perspective of compositional generalization. In *AAAI'21*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Mareike Hartmann, Miryam de Lhoneux, Daniel Hershcovich, Yova Kementchedjhieva, Lukas Nielsen, Chen Qiu, and Anders Søgaard. 2021. A multilingual benchmark for probing negation-awareness with minimal pairs. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 244–257, Online. Association for Computational Linguistics.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*.

Izumi Haruta, Koji Mineshima, and Daisuke Bekki. 2020. Logical inferences with comparatives and generalized quantifiers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 263–270, Online. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Jean van Heijenoort. 1967. Logic as calculus and logic as language. In *Proceedings of the Boston Colloquium for the Philosophy of Science 1964/1966*, pages 440–446. Springer.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *Sort*, 2(4):0–6.

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. 2022. 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. Association for Computational Linguistics.

Jonathan Herzig and Jonathan Berant. 2017. Neural semantic parsing over multiple knowledge-bases. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 623–628, Vancouver, Canada. Association for Computational Linguistics.

Jonathan Herzig and Jonathan Berant. 2021. Span-based semantic parsing for compositional generalization. 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 908–921, Online. Association for Computational Linguistics.

Jonathan Herzig, Peter Shaw, Ming-Wei Chang, Kelvin Guu, Panupong Pasupat, and Yuan Zhang. 2021. Unlocking compositional generalization in pre-trained models using intermediate representations. *arXiv preprint arXiv:2104.07478*.

James Higginbotham and Robert May. 1981. Questions, quantifiers and crossing.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, José Emilio Labra Gayo, S. Kirrane, Sebastian Neumaier, Axel Polleres, R. Navigli, Axel-Cyrille Ngonga Ngomo, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. 2021. Knowledge graphs. *Communications of the ACM*, 64:96 – 104.

Laurence R. Horn. 2010. The expression of negation.

Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. 2020. An analysis of natural language inference benchmarks through the lens of negation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9106–9118, Online. Association for Computational Linguistics.

Hai Hu, Qi Chen, and Larry Moss. 2019. Natural language inference with monotonicity. In *Proceedings of the 13th International Conference on Computational Semantics - Short Papers*, pages 8–15, Gothenburg, Sweden. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve. *ArXiv*, abs/2210.11610.

Prateek Jain, Manav Ratan Mital, Sumit Kumar, Amitabha Mukerjee, and Achla M. Raina. 2004. Anaphora resolution in multi-person dialogues. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 47–50, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

Kelvin Jiang, Dekun Wu, and Hui Jiang. 2019. FreebaseQA: A new factoid QA data set matching trivia-style question-answer pairs with Freebase. 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 318–323, Minneapolis, Minnesota. Association for Computational Linguistics.

Devin Johnson, Denise Mak, Andrew Barker, and Lexi Loessberg-Zahl. 2020. Probing for multilingual numerical understanding in transformer-based language models. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 184–192, Online. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Pratik Joshi, Somak Aditya, Aalok Sathe, and Monojit Choudhury. 2020a. TaxiNLI: Taking a ride up the NLU hill. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 41–55, Online. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020b. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Dan Jurafsky and James H. Martin. 2000. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, 2nd edition. In *Prentice Hall series in artificial intelligence*.

Karthikeyan K, Aalok Sathe, Somak Aditya, and Monojit Choudhury. 2021. Analyzing the effects of reasoning types on cross-lingual transfer performance. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 86–95, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lucie-Aimée Kaffee, Kemele M. Endris, Elena Simperl, and Maria-Esther Vidal. 2019. Ranking knowledge graphs by capturing knowledge about languages and labels. In *Proceedings of the 10th International Conference on Knowledge Capture, K-CAP 2019, Marina Del Rey, CA, USA, November 19-21, 2019*. ACM.

Aishwarya Kamath and Rajarshi Das. 2019. A survey on semantic parsing. In *Automated Knowledge Base Construction (AKBC)*.

Manoj Prabhakar Kannan Ravi, Kuldeep Singh, Isaiah Onando Mulang', Saeedeh Shekarpour, Johannes Hoffart, and Jens Lehmann. 2021. CHOLAN: A modular approach for neural entity linking on Wikipedia and Wikidata. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 504–514, Online. Association for Computational Linguistics.

Antonia Karamolegkou, Phillip Rust, Yong Cao, Ruixiang Cui, Anders Søgaard, and Daniel Hershcovich. 2024. Vision-language models under cultural and inclusive considerations. In *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*, Bangkok, Thailand. Association for Computational Linguistics.

Napoleon Katsos, Chris Cummins, Maria-José Ezeizabarrena, Anna Gavarró, Jelena Kuvač Kraljević, Gordana Hrzica, Kleanthes K Grohmann, Athina Skordi, Kristine Jensen De López, Lone Sundahl, et al. 2016. Cross-linguistic patterns in the acquisition of quantifiers. *Proceedings of the National Academy of Sciences*, 113:9244 – 9249.

Napoleon Katsos, Maria-Jose Ezeizabarrena, Anna Gavarro, Jelena Kuvac Kraljevic, Gordana Hrzica, Kleanthes K Grohmann, Athina Skordi, Kristine Jensen de Lopez, Lone Sundahl, Angeliek Van Hout, et al. 2012. The acquisition of quantification across languages: some predictions.

Edward L. Keenan and Denis Paperno. 2012. Handbook of quantifiers in natural language.

Edward L. Keenan and Dag Westerståhl. 1997. Generalized quantifiers in linguistics and logic. In *Handbook of Logic and Language*.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz

Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. 2020. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations*.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. 2017. Cross-lingual transfer learning for POS tagging without cross-lingual resources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2832–2838, Copenhagen, Denmark. Association for Computational Linguistics.

Najoung Kim and Tal Linzen. 2020. COGS: A compositional generalization challenge based on semantic interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105, Online. Association for Computational Linguistics.

Najoung Kim, Roma Patel, Adam Poliak, Patrick Xia, Alex Wang, Tom McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick. 2019. Probing what different NLP tasks teach machines about function word comprehension. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 235–249, Minneapolis, Minnesota. Association for Computational Linguistics.

Vladislav Korablinov and Pavel Braslavski. 2020. RuBQ: A Russian dataset for question answering over Wikidata. In *International Semantic Web Conference*.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin,

Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Brenden M. Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *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 2879–2888. PMLR.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. In *NeurIPS*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *ArXiv*, abs/1909.11942.

D Terence Langendoen. 1978. The logic of reciprocity. *Linguistic inquiry*, 9(2):177–197.

Peter Lasersohn. 1998. Generalized distributivity operators. *Linguistics and philosophy*, pages 83–93.

Geoffrey Leech, Paul Rayson, et al. 2014. *Word frequencies in written and spoken English: Based on the British National Corpus*. Routledge.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2015. DBpedia - a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2):167–195.

Stephen C Levinson. 1983. *Pragmatics*. Cambridge university press.

Stephen C Levinson, C Stephen, and Stephen C Levinson. 2000. *Presumptive meanings: The theory of generalized conversational implicature*. MIT press.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. Solving quantitative reasoning problems with language models.

Tao Li, Vivek Gupta, Maitrey Mehta, and Vivek Srikumar. 2019. A logic-driven framework for consistency of neural models. 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 3924–3935, Hong Kong, China. Association for Computational Linguistics.

Yiyuan Li, Rakesh Menon, Sayan Ghosh, and Shashank Srivastava. 2023. Pragmatic reasoning unlocks quantifier semantics for foundation models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 573–591, Singapore. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark

dataset for cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Per Lindström. 1966. First order predicate logic with generalized quantifiers. *Theoria*, 32(3):186–195.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 158–167.

Godehard Link et al. 1983. The logical analysis of plurals and mass terms: A lattice-theoretical approach. *Formal semantics: The essential readings*, 127:147.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021a. 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. Association for Computational Linguistics.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2021b. Logiqa: a challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth*

International Conference on International Joint Conferences on Artificial Intelligence, pages 3622–3628.

Jiangming Liu, Shay B. Cohen, Mirella Lapata, and Johan Bos. 2021c. Universal Discourse Representation Structure Parsing. *Computational Linguistics*, 47(2):445–476.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

John Lyons. 1977. *Semantics: Volume 2*, volume 2. Cambridge university press.

Tharindu Madusanka, Iqra Zahid, Hao Li, Ian Pratt-Hartmann, and Riza Batista-Navarro. 2023. Not all quantifiers are equal: Probing transformer-based language models' understanding of generalised quantifiers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8680–8692, Singapore. Association for Computational Linguistics.

Christopher Manning and Hinrich Schutze. 1999. *Foundations of statistical natural language processing*. MIT press.

Mehdi Manshadi, James Allen, and Mary Swift. 2011. A corpus of scope-disambiguated English text. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 141–146, Portland, Oregon, USA. Association for Computational Linguistics.

Gary Marcus and Ernest Davis. 2019. *Rebooting AI: Building artificial intelligence we can trust*. Vintage.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

William A. Martin. 1981. Formal roles, co-descriptors, and the representation of quantified English expressions. *American Journal of Computational Linguistics*, 7(3):137–148.

Gerald J Massey. 1976. Tom, dick, and harry, and all the king's men. *American Philosophical Quarterly*, 13(2):89–107.

Lisa Matthewson. 2001. Quantification and the nature of crosslinguistic variation. *Natural Language Semantics*, 9(2):145–189.

Lisa Matthewson. 2004. On the methodology of semantic fieldwork. *International journal of American linguistics*, 70(4):369–415.

Lisa Matthewson. 2008. Quantification: A cross-linguistic perspective.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.

John McCarthy, Nathaniel Rochester, and Claude Shannon. 1956. Dartmouth workshop.

James D. McCawley. 1968. The role of semantics in a grammar. In Emmon Bach and R. Harms, editors, *Universals in Linguistic Theory*, pages 124–169. Holt, Rinehart, and Winston.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229.

Friederike Moltmann. 1992. *Coordination and Comparatives*. Ph.D. thesis, MIT.

Richard Montague. 1973. The proper treatment of quantification in ordinary english. In *Approaches to natural language*, pages 221–242. Springer.

Laurer Moritz, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2022. Less annotating, more classifying – addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert - nli.

Andrzej Mostowski. 1957. On a generalization of quantifiers. *Fundamenta Mathematicae*, 44(1):12–36.

Mathijs Mul and Willem Zuidema. 2019. Siamese recurrent networks learn first-order logic reasoning and exhibit zero-shot compositional generalization.

Makoto Nagao. 1984. A framework of a mechanical translation between japanese and english by analogy principle. *Artificial and human intelligence*, pages 351–354.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Denis Newman-Griffis, Albert Lai, and Eric Fosler-Lussier. 2017. Insights into analogy completion from the biomedical domain. In *BioNLP 2017*, pages 19–28, Vancouver, Canada,. Association for Computational Linguistics.

Ngonga Ngomo. 2018. 9th challenge on question answering over linked data (qald-9). *language*, 7(1):58–64.

Massimo Nicosia, Zhongdi Qu, and Yasemin Altun. 2021. Translate & Fill: Improving zero-shot multilingual semantic parsing with synthetic data. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3272–3284, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Binling Nie, Ruixue Ding, Pengjun Xie, Fei Huang, Chen Qian, and Luo Si. 2021. Knowledge-aware named entity recognition with alleviating heterogeneity. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association*

for Computational Linguistics, pages 4885–4901, Online. Association for Computational Linguistics.

Liudmila Nikolaeva. 2012. *Quantifiers in Adyghe*, pages 21–82. Springer Netherlands, Dordrecht.

Mitja Nikolaus, Mostafa Abdou, Matthew Lamm, Rahul Aralikatte, and Desmond Elliott. 2019. Compositional generalization in image captioning. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 87–98, Hong Kong, China. Association for Computational Linguistics.

Sergei Nirenburg. 1987. *Machine translation: Theoretical and methodological issues*. Cambridge University Press.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Stephan Oepen, Omri Abend, Lasha Abzianidze, Johan Bos, Jan Hajic, Daniel Hershcovich, Bin Li, Tim O'Gorman, Nianwen Xue, and Daniel Zeman. 2020. MRP 2020: The second shared task on cross-framework and cross-lingual meaning representation parsing. In *Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing*, pages 1–22, Online. Association for Computational Linguistics.

Maciej Ogrodniczuk, Vincent Ng, Yulia Grishina, and Sameer Pradhan, editors. 2020. *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*. Association for Computational Linguistics, Barcelona, Spain (online).

Maciej Ogrodniczuk, Sameer Pradhan, Yulia Grishina, and Vincent Ng, editors. 2019. *Proceedings of the Second Workshop on Computational Models of Reference, Anaphora and Coreference*. Association for Computational Linguistics, Minneapolis, USA.

Sadayuki Okada. 1999. On the function and distribution of the modifiers respective and respectively. *Linguistics*, 37(5):871–903.

OpenAI. 2022. Chatgpt. `https://chat.openai.com/chat`.

OpenAI. 2023. Gpt-4 technical report.

Inbar Oren, Jonathan Herzig, Nitish Gupta, Matt Gardner, and Jonathan Berant. 2020. Improving compositional generalization in semantic parsing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2482–2495, Online. Association for Computational Linguistics.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The lambada dataset: Word prediction requiring a broad discourse context. *arXiv preprint arXiv:1606.06031*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alicia Parrish, William Huang, Omar Agha, Soo-Hwan Lee, Nikita Nangia, Alexia Warstadt, Karmanya Aggarwal, Emily Allaway, Tal Linzen, and Samuel R. Bowman. 2021. Does putting a linguist in the loop improve NLU data collection? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4886–4901, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Barbara Partee et al. 1995. Lexical semantics and compositionality. *An invitation to cognitive science: Language*, 1:311–360.

Barbara Hall Partee. 1970. Negation, conjunction, and quantifiers: Syntax vs. semantics. *Foundations of Language*, 6(2):153–165.

Thomas Pellissier Tanon, Denny Vrandečić, Sebastian Schaffert, Thomas Steiner, and Lydia Pintscher. 2016. From Freebase to Wikidata: The great migration. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 1419–1428.

Aleksandr Perevalov, Dennis Diefenbach, Ricardo Usbeck, and Andreas Both. 2022a. QALD-9-plus: A multilingual dataset for question answering over

DBpedia and Wikidata translated by native speakers. In *2022 IEEE 16th International Conference on Semantic Computing (ICSC)*. IEEE.

Aleksandr Perevalov, Axel-Cyrille Ngonga Ngomo, and Andreas Both. 2022b. Enhancing the accessibility of knowledge graph question answering systems through multilingualization. In *2022 IEEE 16th International Conference on Semantic Computing (ICSC)*, pages 251–256.

Stanley Peters and Dag Westerståhl. 2006. *Quantifiers in language and logic*. Oxford University Press.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Steven Pinker. 2003. *The language instinct: How the mind creates language*. Penguin uK.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Geoffrey K Pullum and Gerald Gazdar. 1982. Natural languages and context-free languages. *Linguistics and Philosophy*, 4(4):471–504.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language*

Processing, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Mohammad Sadegh Rasooli and Michael Collins. 2017. Cross-lingual syntactic transfer with limited resources. *Transactions of the Association for Computational Linguistics*, 5:279–293.

Marco Tulio Ribeiro, Carlos Guestrin, and Sameer Singh. 2019. Are red roses red? evaluating consistency of question-answering models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6174–6184, Florence, Italy. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Craige Roberts. 1987. *Modal subordination, anaphora, and distributivity.* University of Massachusetts Amherst.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Laura Ruis, Jacob Andreas, Marco Baroni, Diane Bouchacourt, and Brenden M Lake. 2020. A benchmark for systematic generalization in grounded language understanding. *Advances in neural information processing systems*, 33:19861–19872.

Stuart J Russell and Peter Norvig. 2016. *Artificial intelligence: a modern approach.* Pearson.

Ivan Rybin, Vladislav Korablinov, Pavel Efimov, and Pavel Braslavski. 2021. RuBQ 2.0: An innovated Russian question answering dataset. In *Eighteenth Extended Semantic Web Conference - Resources Track.*

Amrita Saha, Vardaan Pahuja, Mitesh M. Khapra, Karthik Sankaranarayanan, and A. P. S. Chandar. 2018. Complex sequential question answering: Towards learning to converse over linked question answer pairs with a knowledge graph. In *AAAI*.

Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys (CSUR)*, 55(2):1–39.

Cezar Sas, Meriem Beloucif, and Anders Søgaard. 2020. WikiBank: Using Wikidata to improve multilingual frame-semantic parsing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4183–4189, Marseille, France. European Language Resources Association.

Remko Scha. 1984. Distributive, collective and cumulative quantification. *Truth, Interpretation, and Information, GRASS*, 2:131–158.

Remko Scha and David Stallard. 1988. Multi-level plurals and distributivity. In *26th Annual Meeting of the Association for Computational Linguistics*, pages 17–24, Buffalo, New York, USA. Association for Computational Linguistics.

Roger C. Schank and Robert P. Abelson. 1978. Scripts, plans, goals and understanding: an inquiry into human knowledge structures.

James H. Schmerl and Stephen G. Simpson. 1982. On the role of ramsey quantifiers in first order arithmetic. *J. Symb. Log.*, 47:423–435.

Nikolaus Schrack, Ruixiang Cui, Hugo López, and Daniel Hershcovich. 2022. Can AMR assist legal and logical reasoning? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1555–1568, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. 2015. Hidden technical debt in machine learning systems. *Advances in neural information processing systems*, 28.

Mark S Seidenberg and James L McClelland. 1989. A distributed, developmental model of word recognition and naming. *Psychological review*, 96(4):523.

Bo Shao, Yeyun Gong, Weizhen Qi, Nan Duan, and Xiaola Lin. 2020. Multi-level alignment pretraining for multi-lingual semantic parsing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3246–3256, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Chih Chieh Shao, Trois Liu, Yuting Lai, Yiying Tseng, and Sam Tsai. 2018. Drcd: a chinese machine reading comprehension dataset.

James Shaw and Kathleen McKeown. 2000. Generating referring quantified expressions. In *INLG'2000 Proceedings of the First International Conference on Natural Language Generation*, pages 100–107, Mitzpe Ramon, Israel. Association for Computational Linguistics.

Tao Shen, Xiubo Geng, Tao Qin, Daya Guo, Duyu Tang, Nan Duan, Guodong Long, and Daxin Jiang. 2019. Multi-task learning for conversational question answering over a large-scale knowledge base. 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 2442–2451, Hong Kong, China. Association for Computational Linguistics.

Tom Sherborne and Mirella Lapata. 2022. Zero-shot cross-lingual semantic parsing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4134–4153, Dublin, Ireland. Association for Computational Linguistics.

David So, Quoc Le, and Chen Liang. 2019. The evolved transformer. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5877–5886. PMLR.

Anders Søgaard. 2022. Should we ban English NLP for a year? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5254–5260, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat

Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Johan Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, Cesar Ferri, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Christopher Waites, Christian Voigt, Christopher D Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, C. Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Xinyue Wang, Gonzalo Jaimovitch-Lopez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Francis Anthony Shevlin, Hinrich Schuetze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B Simon, James Koppel, James Zheng, James Zou, Jan Kocon, Jana Thompson, Janelle Wingfield,

Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh Dhole, Kevin Gimpel, Kevin Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros-Colón, Luke Metz, Lütfi Kerem Senel, Maarten Bosma, Maarten Sap, Maartje Ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramirez-Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael Andrew Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Andrew Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter W Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Russ Salakhutdinov, Ryan Andrew Chi, Seungjae Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh,

Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel Stern Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Shammie Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven Piantadosi, Stuart Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsunori Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, vinay uday prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Shashank Srivastava, Igor Labutov, and Tom Mitchell. 2018. Zero-shot learning of classifiers from natural language quantification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 306–316, Melbourne, Australia. Association for Computational Linguistics.

Penka Stateva, Arthur Stepanov, Viviane Déprez, Ludivine Emma Dupuy, and Anne Colette Reboul. 2019. Cross-linguistic variation in the meaning of quantifiers: Implications for pragmatic enrichment. *Frontiers in Psychology*, 10:957.

Shane Steinert-Threlkeld. 2019. Learnability and semantic universals. *Semantics and Pragmatics*, 12:4.

Shane Steinert-Threlkeld. 2021. Quantifiers in natural language: Efficient communication and degrees of semantic universals. *Entropy*, 23(10).

Jakub Szymanik. 2016. *Cognitive Processing of Quantifiers*, pages 51–83. Springer International Publishing, Cham.

Jakub Szymanik and Camilo Thorne. 2015. Semantic complexity of quantifiers and their distribution in corpora. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 64–69, London, UK. Association for Computational Linguistics.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 641–651, New Orleans, Louisiana. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

James Thorne, Majid Yazdani, Marzieh Saeidi, Fabrizio Silvestri, Sebastian Riedel, and Alon Halevy. 2021. Database reasoning over text. 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 3091–3104, Online. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Together. 2022. Releasing GPT-JT powered by open-source AI. https://www.together.xyz/blog/releasing-v1-of-gpt-jt-powered-by-open-source-ai.

Michael Tomasello. 2010. *Origins of human communication*. MIT press.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aur'elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Dmitry Tsarkov, Tibor Tihon, Nathan Scales, Nikola Momchev, Danila Sinopalnikov, and Nathanael Schärli. 2021. *-CFQ: Analyzing the scalability of machine learning on a compositional task. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Alan M. Turing. 1950. Computing machinery and intelligence. *Mind*, LIX:433–460.

Wataru Uegaki. 2022. The Informativeness/Complexity Trade-Off in the Domain of Boolean Connectives. *Linguistic Inquiry*, pages 1–39.

Yuval Varkel and Amir Globerson. 2020. Pre-training mention representations in coreference models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8534–8540, Online. Association for Computational Linguistics.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2tensor for neural machine translation. *CoRR*, abs/1803.07416.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Sara Veldhoen and Willem Zuidema. 2017. Can neural networks learn logical reasoning? *CLASP Papers in Computational Linguistics*, page 34.

Lev S Vygotsky. 2012. *Thought and language*. MIT press.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. `https://github.com/kingoflolz/mesh-transformer-jax`.

Siyuan Wang, Zhongkun Liu, Wanjun Zhong, Ming Zhou, Zhongyu Wei, Zhumin Chen, and Nan Duan. 2022. From lsat: The progress and challenges of complex reasoning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2201–2216.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Dag Westerståhl. 1989. Quantifiers in formal and natural languages. In *Handbook of philosophical logic*, pages 1–131. Springer.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Adina Williams, Tristan Thrush, and Douwe Kiela. 2022. ANLIzing the adversarial natural language inference dataset. In *Proceedings of the Society for Computation in Linguistics 2022*, pages 23–54, online. Association for Computational Linguistics.

Terry Winograd. 1971. Procedures as a representation for data in a computer program for understanding natural language.

Yoad Winter. 2002. *Flexibility principles in Boolean semantics: The interpretation of coordination, plurality, and scope in natural language*. MIT press.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

William A Woods. 1970. Transition network grammars for natural language analysis. *Communications of the ACM*, 13(10):591–606.

William A. Woods. 1991. Understanding subsumption and taxonomy: A framework for progress. In *Principles of Semantic Networks*.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, and Kentaro Inui. 2020. Do neural models learn systematicity of monotonicity inference in natural language? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6105–6117, Online. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019. Can neural networks understand monotonicity reasoning? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 31–40, Florence, Italy. Association for Computational Linguistics.

Jingfeng Yang, Federico Fancellu, Bonnie Webber, and Diyi Yang. 2021. Frustratingly simple but surprisingly strong: Using language-independent features for zero-shot cross-lingual semantic parsing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5848–5856, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Deming Ye, Yankai Lin, Jiaju Du, Zhenghao Liu, Peng Li, Maosong Sun, and Zhiyuan Liu. 2020. Coreferential Reasoning Learning for Language Representation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7170–7186, Online. Association for Computational Linguistics.

John Yen. 1991. Generalizing term subsumption languages to fuzzy logic. In *IJCAI*.

Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. DocNLI: A large-scale dataset for document-level natural language inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4913–4922, Online. Association for Computational Linguistics.

Charles Yu, Ryan Sie, Nicolas Tedeschi, and Leon Bergen. 2020a. Word frequency does not predict grammatical knowledge in language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4040–4054, Online. Association for Computational Linguistics.

Juntao Yu, Nafise Sadat Moosavi, Silviu Paun, and Massimo Poesio. 2020b. Free the plural: Unrestricted split-antecedent anaphora resolution. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6113–6125, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Lotfi A Zadeh. 1983. A computational approach to fuzzy quantifiers in natural languages. In *Computational linguistics*, pages 149–184. Elsevier.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022a. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Xikun Zhang, Deepak Ramachandran, Ian Tenney, Yanai Elazar, and Dan Roth. 2020. Do language embeddings capture scales? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4889–4896, Online. Association for Computational Linguistics.

Yusen Zhang, Jun Wang, Zhiguo Wang, and Rui Zhang. 2023. XSemPLR: Cross-lingual semantic parsing in multiple natural languages and meaning representations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15918–15947, Toronto, Canada. Association for Computational Linguistics.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022b. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Jec-qa: A legal-domain question answering dataset. In *Proceedings of AAAI*.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. AGIEval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314, Mexico City, Mexico. Association for Computational Linguistics.

Wanjun Zhong, Siyuan Wang, Duyu Tang, Zenan Xu, Daya Guo, Yining Chen, Jiahai Wang, Jian Yin, Ming Zhou, and Nan Duan. 2022. Analytical

reasoning of text. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2306–2319, Seattle, United States. Association for Computational Linguistics.

Ethan Zhou and Jinho D. Choi. 2018. They exist! introducing plural mentions to coreference resolution and entity linking. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 24–34, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

George Kingsley Zipf. 1949. Human behavior and the principle of least effort.