MiQA: A Benchmark for Inference on Metaphorical Questions

Iulia-Maria Comșa
Google Research, Zürich
`iuliacomsa@google.com`

Julian Martin Eisenschlos
Google Research, Zürich
`eisenjulian@google.com`

Srini Narayanan
Google Research, Zürich
`srinin@google.com`

Abstract

We propose a benchmark to assess the capability of large language models to reason with conventional metaphors. Our benchmark combines the previously isolated topics of metaphor detection and commonsense reasoning into a single task that requires a model to make inferences by accurately selecting between the literal and metaphorical register. We examine the performance of state-of-the-art pre-trained models on binary-choice tasks and find a large discrepancy between the performance of small and very large models, going from chance to near-human level. We also analyse the largest model in a generative setting and find that although human performance is approached, careful multiple-shot prompting is required.[1]

1 Introduction

Conceptual metaphor is an ubiquitous cognitive mechanism that allows us to structure and reason about abstract concepts by relating them to experiential domains (Lakoff and Johnson, 2003; Feldman, 2008). In language, metaphors allow human communication and reasoning about abstract ideas using concrete notions learned from sensorimotor, emotional, and other embodied experience (Thibodeau and Boroditsky, 2011): "a plan is *solid*"; "the economy *is stumbling*"; "I *see* what you mean".

To illustrate the role of metaphor in abstract reasoning, consider the following metaphorical statement: "the economy is *stumbling*". According to conceptual metaphor theory (CMT), we understand this statement through mental simulation, by connecting the abstract concept of economy to the imagined movement of a stumbling person. We use the same mental imagery to infer that the economy, just like a stumbling person, is "unstable" and "might fall". As another example, in the metaphorical statement "a proposal is *solid*", bringing to mind solid objects and their properties suggests that the proposal, just like a physical object, was "well-built" and "will not easily break". However, not all properties generalize: unlike a physical object, a proposal, which is an abstract entity, cannot be "thrown" or "bent"; unlike a stumbling person, the economy does not "wear shoes".

Large Language Models (LLMs) have achieved remarkable results on a variety of tasks. However, in contrast to humans, LLMs do not have access to commonsense and embodied experiences of the world (Bender and Koller, 2020). Although the data LLMs are trained on includes up to trillions of text tokens, it is unclear how much of this data allows them to capture human commonsense reasoning (Gordon and Van Durme, 2013; Becker et al., 2021). Conceptual metaphor theory suggests that embodied and implicit knowledge is required for the ability to utilize metaphors in commonsense reasoning about concrete and abstract ideas.

We propose a novel dataset, MiQA (Metaphorical Inference Questions and Answers), to assess the ability of a model to reason with conventional metaphors. The benchmark draws on the CMT research (Grady, 1997) to construct a representative set of primary metaphors, which are contrastively paired with literal statements. Our task requires a model to make a correct inference in simple situations without specifying whether the contextual register is literal or metaphorical, leveraging research on metaphor processing (Rai and Chakraverty, 2020), commonsense reasoning (Davis and Marcus, 2015), and natural language inference (Dagan et al., 2006; Bowman et al., 2015).

Our benchmark combines the previously isolated areas of metaphor detection and commonsense inference. Although there is considerable research on both of these areas separately, it is unclear whether such capabilities scale compositionally: LLMs could handle two separate tasks well, but not their combination (Keysers et al., 2020).

---

[1] The benchmark is available at https://github.com/google-research/language/tree/master/language/miqa.

"implies"-questions	*"implied-by"*-questions
"I see what you mean". Which of the following statements could that imply? (1) My eyes are working well **[incorrect]** (2) I understand you **[correct]**	"My eyes are working well" is implied by which of the following? (1) I see what you are pointing at **[correct]** (2) I see what you mean **[incorrect]**
"A plan is not solid". Which of the following statements could that imply? (1) A hammer could break it **[incorrect]** (2) We should not follow it **[correct]**	"A hammer could break it" is implied by which of the following? (1) A table is not solid **[correct]** (2) A plan is not solid **[incorrect]**
"My friend has a huge problem". Which of the following statements could that imply? (1) My friend needs space **[incorrect]** (2) My friend needs a solution **[correct]**	"My friend needs space" is implied by which of the following? (1) My friend has a huge dog **[correct]** (2) My friend has a huge problem **[incorrect]**

Table 1: Examples of MiQA items combined into two task types. Correct answers are labelled. The examples are built by pairing 150 sets of two premises (one metaphorical and one literal) and two respective implications.

Our contributions are the following:

- We propose MiQA, a benchmark for commonsense inference with conventional metaphors;

- We show a large discrepancy between the performance of small and large models in a binary-choice MiQA task, from chance to human-level accuracy;

- We use a generative MiQA task to corroborate the performance of the largest model in an open-ended setting, showing that although human-level performance is approached, careful multiple-shot prompting is required.

2 Related Work

Metaphor has received renewed attention in natural language processing, but most tasks have been restricted to detection (e.g. Leong et al., 2018, 2020; Choi et al., 2021) on large annotated corpora (Steen et al., 2010; Beigman Klebanov et al., 2016). Human-level performance has not been reached by LLMs, but the progress is promising and an active area of research. However, these tasks may excessively rely on context-dependent word meanings (Neidlein et al., 2020) and do not measure the ability to reason with metaphor.

Metaphor paraphrasing is another active area of research. BIG-bench (Srivastava et al., 2022), a collaborative multi-task benchmark intended to test a variety of capabilities of LLMs, includes four tasks related to metaphor. While these tasks contain novel metaphors, they still do not assess the ability to employ metaphoric knowledge in reasoning.

Recently, Chakrabarty et al. (2022) built a dataset for multiple-choice story continuations involving similes and idioms extracted from books. Subsequently, Liu et al. (2022) proposed a metaphor interpretation task that requires models to choose the correct out of two interpretations of a simile. In contrast, our task combines metaphor interpretation with commonsense inference, uses a more systematic data source, and has an additional adversarial character, as it requires the selection between two semantically-close items instead of items with opposite meanings.

3 Dataset

3.1 Motivation

Most existing studies of metaphor have primarily started from corpus-based methods, using frequency and other corpus-based metrics to detect or classify metaphors. This process leads to a primary focus on corpus distributions and makes it hard to compare studies across different corpora. Furthermore, it ignores the central tenet of metaphor theory that foundational metaphors are grounded in non-linguistic and experiential domains which may be assumed as a background and thus underrepresented in corpora.

To address this, we chose to use a foundational ontology of primary conceptual metaphors (Grady, 1997) based on CMT (Lakoff and Johnson, 2003; Feldman, 2008). Our choice of primary metaphors has multiple desirable properties. Primary metaphors are a good starting point for the investigation of more complex, compositional metaphor. The metaphors in our dataset are developmentally early in child experience with primary scenes and language. Moreover, the chosen metaphors are embodied, in that the source domain is often observable and sensory-motor (size, warmth, height) while the targets are less observ-

able and often subjective or abstract (importance, affection, quantity). The metaphors we chose form a basis set of mappings that can create, through composition, more complex mappings, such as the Event Structure Metaphor, that maps movement and manipulation to actions. Our approach ensures that the distributions of metaphor categories in the task is balanced and hence reflective of the capability of large models to use primary metaphors as building blocks of reasoning.

3.2 Construction

We constructed a novel dataset consisting of 150 items. The items were manually created by the authors based on the work of Grady (1997), which lists 100 primary metaphors that are conventional, developmentally early, and form a basis set for composing complex mappings.

Each item in the dataset is a tuple consisting of four sentences in English: a literal premise (L_p), a premise containing a conventional metaphor (M_p), an implication of the literal premise (L_c), and an implication of the metaphorical premise (M_c).

The tuples are paired so that a mistaken literal interpretation of M_p can falsely suggest that L_c is implied. For example, a wrong inference would be: "*I see what you mean*" implies that "*I am using my eyes*". The false implication $M_p \rightarrow L_c$ thus serves as an adversarial element that probes whether the model correctly registers the metaphorical context.

For each primary metaphor proposed by Grady (1997), we manually created 1-2 pairs of items in the form described above, where M_p is an example of the metaphor, while L_p relates to the source domain of the metaphor only.

To create the final benchmark for LLMs, we used these items to generate two types of adversarial questions. The first type ("*implies*"-questions) requests the model to select the most likely inference given a metaphorical statement. Answering correctly requires the model to not be tricked by a possible literal interpretation of the metaphorical premise. The second type ("*implied-by*"-questions) requests the model to select the most likely premise that a literal statement is implied by. Answering correctly requires the model to not be tricked by a possible metaphorical interpretation of the literal conclusion. See Table 1 for examples.

We combined these items to obtain a benchmark consisting of 300 questions, of which half are "*implies*"-questions and half are "*implied-by*"-

questions. This pairing of tasks ensures that the model does not achieve a better score if biased towards assigning a higher likelihood to either literal or metaphorical continuations of a statement.

4 Human Evaluation

We estimated the human performance on the binary-choice task using the responses of 15 human adult volunteers with English as first or second language. The participants were told that the aim of the research was to gather a set of commonsense responses and compare them to LLMs responses. No additional information about the task was given.

5 Large Language Models Evaluation

We evaluated the performance of two pre-trained LLMs: PaLM with 8B, 62B, 540B parameters (Chowdhery et al., 2022) and GPT-3 Ada, Babbage, Curie and DaVinci (Brown et al., 2020). The parameter counts of the GPT-3 models are not publicly available, but have been estimated at 350M, 1.3B, 6.7B, and 175B respectively [2].

The main purpose of this study is to assess the capabilities of LLMs on the MiQA benchmark. For comparison, we also verify the capabilities of pre-trained fine-tuned smaller language models on our benchmark. We follow Liu et al. (2022) in using encoder-only models trained on the natural language inference datasets SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018) for zero-shot evaluation. We opt for this approach because MiQA is designed as a small dataset suitable as a benchmark and not for fine-tuning. We test the state-of-the-art encoder-only model DeBERTaV3 (He et al., 2021) in sizes small, medium and large. These models have 44M, 86M and 304M parameters respectively, and their weights are available online[3]. The models take a premise-implication pair and produce a probability distribution over three classes: "entailment", "contradiction", and "undetermined". We report the results for the best-performing score, in this case $1 - P(contradiction)$, which outranks $P(entailment)$.

5.1 Binary-Choice Tasks

We first assessed the models by prompting them with the question types illustrated in Table 1 in 0, 1 and 5-shot settings. Few-shot prompts were

---

Prompt: "I am sticking to my plans". Could that imply that "I am using an adhesive"?	
Scored as correct: No, it means that you are not changing your plans.	
Prompt: "'I have to move ahead with my project". Could that imply that "a trolley could help"?	
Scored as wrong: I think it's a reference to the trolley problem.	
Prompt: "My future looks bright". Could that imply that "I could use sunglasses"?	
Scored as ambiguous: I think it's a bit of a stretch.	

Table 2: Examples of prompts from the MiQA dataset and manual scoring of answers generated by PaLM-540b.

obtained by prefixing with randomly selected questions followed by their correct answers. To score the results, we obtained the log likelihood of each of the two choices as candidate continuations to the given prompt question. A response was scored as correct if the log likelihood of the correct choice was larger than that of the incorrect choice.

Prompts can greatly influence LLM predictions (Lu et al., 2022; Cao et al., 2021). As expected, we observed variability with changing prompts. To mitigate this, we tried multiple prompts, as detailed in Appendix A. For each model, we selected the prompt that performed best with 0-shot, and subsequently used this prompt to obtain and report its results in few-shot settings. Additionally, we used two baseline prompts (an empty prompt with no choices, and a prompt containing an unrelated question), which can indicate if the models simply learn to select either metaphorical or literal statements independently of the prompt in few-shot settings.

5.2 Generative Task

In addition to the binary-choice tasks, we also tested the largest model, PaLM-540b, in a generative setting. We prompted this model with questions of the form: "M_p. Could that imply that L_c?". This capitalises on the adversarial false implication $M_p \rightarrow L_c$ described in section 3.

We obtained completions to the 150 questions of this form generated from the MiQA dataset. Answers were manually and independently scored by every author. Every author scored at least two thirds of all responses and the scores were averaged. Scoring consisted of labelling the first paragraph of an answer as "correct", "wrong" or "ambiguous". To compute the accuracy over the generative task, "correct" responses were scored as 1 and "ambiguous" responses were scored as 0.5. Agreement between raters was medium, with intraclass correlation (Shrout and Fleiss, 1979) $ICC(2, k)$ at 0.56. Examples of scored answers are shown in Table 2.

As before, we evaluated the model in 0, 1 and 5-shot settings. From the 0-shot setting, we selected 32 answers produced by the model that all authors independently scored as "correct", and the same number of answers of the form $M_p \rightarrow M_c$ and $L_p \rightarrow L_c$. We randomly selected 1 or 5 of these answers and their corresponding questions to prefix each prompt question in 1- and 5-shot settings.

6 Results

The full results are shown in Table 3.

Firstly, for the binary-choice tasks, there was a considerable gap between small and large LLMs. While the smaller models performed at or close to chance level, the largest models achieved very good performance even with 0 shots, and approached human-level performance with few shots. We note that the "*implied-by*" task was overall more difficult than the "*implies*" task for both humans and LLMs.

Secondly, the chance-level performance on the baseline prompts suggests that the increase in performance in few-shot settings was not due to the model learning to select either metaphorical or literal statements independently of the prompt. On the other hand, the strong performance of the DeBERTaV3 models suggests a high level of transfer from the NLI datasets to MiQA, although there is a still a considerable gap to human performance.

Finally, the generative results on PaLM-540b estimated the model performance in an open-ended setting. Similarly to the binary task, the model performed considerably better with 5 shots compared to 0 shots, approaching human performance. However, the gap between human and model performance for the generative task was greater compared with the gap for the binary-choice task.

Overall, the results demonstrate that LLMs can correctly select between the metaphorical and literal contextual registers to perform inferences with conventional metaphors, but there is still a considerable gap between human and LLM performance

		Accuracy		
		"implies" questions		
Model	**Shots**	**0**	**1**	**5**
PaLM-8b		54.0	51.3	51.0
PaLM-62b		53.7	62.0	66.3
PaLM-540b		**89.7**	**97.0**	**96.3**
GPT-3-Ada		52.3	49.0	51.3
GPT-3-Babbage		50.7	51.0	51.3
GPT-3-Curie		50.7	57.3	55.7
GPT-3-DaVinci		**89.3**	**97.7**	**98.7**
DeBERTaV3-NLI-small		78.0		
DeBERTaV3-NLI-base		**82.7**		
DeBERTaV3-NLI-large		80.0		
Baseline prompts				
no choices		48.7	49.3	52.3
no questions		50.0	56.0	58.0
Human		**99.6**		
		"implied-by" questions		
Model	**Shots**	**0**	**1**	**5**
PaLM-8b		51.0	57.7	55.3
PaLM-62b		53.0	58.0	65.3
PaLM-540b		**71.3**	**84.7**	**92.3**
GPT-3-Ada		53.0	53.0	46.3
GPT-3-Babbage		50.7	50.3	52.0
GPT-3-Curie		55.0	53.7	50.0
GPT-3-DaVinci		**77.3**	**88.0**	**95.7**
DeBERTaV3-NLI-small		74.0		
DeBERTaV3-NLI-base		70.7		
DeBERTaV3-NLI-large		**76.7**		
Baseline prompts				
no choices		49.3	53.0	58.7
no questions		50.0	43.0	47.0
Human		**96.4**		
Chance		50.0		
Generative task		**71.7**	**73.1**	**88.9**

Table 3: Results obtained on the MiQA tasks by pre-trained LLMs in few-shot settings. Small models perform close to chance level, while large models perform close to human level. We include two prompting baselines for PaLM-540b, whose performance close to chance level shows that few-shot performance is not due to metaphor detection only. We also include the accuracy for the generative task that asks PaLM-540 to answer open-ended "$M_p \rightarrow L_c$?" questions, scored as descried in Table 2. Multiple-shot prompting is required to approach human-level performance.

in 0-shot settings.

7 Limitations

Our work used foundational metaphors from CMT to test basic metaphoric reasoning in LLMs. We will expand this benchmark using additional and more complex sources of conceptual metaphor (e.g. Narayanan and Hong, 2013). Future work will assess LLMs on novel non-conventional mappings.

Although we mitigated for prompt sensitivity by using multiple prompts, the result interpretation should allow for small accuracy variations. Further, in the binary-choice tasks we compare the LLM results with a human baseline, but we do not provide a baseline for the generative task. This is near perfect for humans, but a more systematic baseline can be created to quantify the exact headroom on this task. Finally, while the task holistically measures the performance of LLMs on a complex task, it is difficult to disentangle the component effects (metaphor detection, reasoning, response generation) in the overall accuracy.

8 Conclusion

We have proposed a novel compositional benchmark based on conceptual metaphor theory to assess the capacity of LLMs to make inferences with metaphors. Successful performance on this task requires metaphor detection and commonsense inference. Using a metaphor theory-based approach allows us to systematically explore capabilities and limitations of LLMs. This is the first in a planned series of increasingly complex metaphor inference datasets.

Three main findings emerged from our proposed task. Firstly, there is a vast difference between the performance of small and large LLMs, with the former performing at chance level and the latter approaching human level in few-prompt setting. This observation is informative in the context of previous results showing that some, but not all, tasks observe a qualitative performance jump with model size and scale: for example, this is the case for reasoning about goal-step relationships between events and ordering events, but not for navigation and mathematical induction tasks (Chowdhery et al., 2022). This result invites more research into the question of how and whether the performance of smaller models can be improved. Secondly, this reflects a true ability of LLMs to reason with conventional metaphor, and not simply to detect it. Whether this ability extends to novel metaphor is ongoing work. Finally, the performance of large LLMs approaches that of humans in binary-choice and generative tasks, but careful multiple-shot prompting is required.

Acknowledgements

References

Maria Becker, Siting Liang, and Anette Frank. 2021. Reconstructing implicit knowledge with language models. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 11–24, Online. Association for Computational Linguistics.

Beata Beigman Klebanov, Chee Wee Leong, E. Dario Gutierrez, Ekaterina Shutova, and Michael Flor. 2016. Semantic classifications for detection of verb metaphors. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 101–106, Berlin, Germany. Association for Computational Linguistics.

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.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. 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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. 2021. Knowledgeable or educated guess? revisiting language models as knowledge bases. 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 1860–1874, Online. Association for Computational Linguistics.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022. It's not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1763–1773, Online. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling language modeling with pathways. arXiv:2204.02311. Version 5.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM*, 58(9):92–103.

Jerome Feldman. 2008. *From molecule to metaphor: A neural theory of language*. MIT press.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 25–30.

Joseph Edward Grady. 1997. *Foundations of meaning: Primary metaphors and primary scenes*. University of California, Berkeley.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. arXiv:2111.09543. Version 2.

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 *8th International Conference on Learning Representations (ICLR) 2020, Addis Ababa, Ethiopia*.

George Lakoff and Mark Johnson. 2003. *Metaphors we live by*. University of Chicago press.

Chee Wee (Ben) Leong, Beata Beigman Klebanov, Chris Hamill, Egon Stemle, Rutuja Ubale, and Xianyang Chen. 2020. A report on the 2020 VUA and TOEFL metaphor detection shared task. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 18–29, Online. Association for Computational Linguistics.

Chee Wee (Ben) Leong, Beata Beigman Klebanov, and Ekaterina Shutova. 2018. A report on the 2018 VUA metaphor detection shared task. In *Proceedings of the Workshop on Figurative Language Processing*, pages 56–66, New Orleans, Louisiana. Association for Computational Linguistics.

Emmy Liu, Chen Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. arXiv:2204.12632. Version 2.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Srini Narayanan and Jisup Hong. 2013. A multilingual functional repository for cognitive linguists. In *International Cognitive Linguistics Conference (ICLC)*.

Arthur Neidlein, Philip Wiesenbach, and Katja Markert. 2020. An analysis of language models for metaphor recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3722–3736, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sunny Rai and Shampa Chakraverty. 2020. A survey on computational metaphor processing. *ACM Comput. Surv.*, 53(2).

Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420–428.

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 Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew 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, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris 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, 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, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, 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 Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, 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 Kocoń, Jana Thompson, 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 Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory 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 Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. 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, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, 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 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, Ramón Risco Delgado, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, 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 S. 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, 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 T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Timothy Telleen-Lawton, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay 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. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv:2206.04615. Version 2.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, Tina Krennmayr, and Tryntje Pasma. 2010. *A method for linguistic metaphor identification. From MIP to MIPVU.* Number 14 in Converging Evidence in Language and Communication Research. John Benjamins.

Paul H. Thibodeau and Lera Boroditsky. 2011. Metaphors we think with: The role of metaphor in reasoning. *PLOS ONE*, 6(2):1–11.

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.

A Appendix

The following prompts were used to assess model performance on the binary-choice MiQA tasks with "*implies*"-questions:

1. "M_p". Which of the following two statements could that imply? L_c or M_c?
   (chosen for PaLM-8b, PaLM-16b, GPT3-Ada, GPT3-Babbage, GPT3-Curie)

2. "M_p". Which of the following two statements could that imply? (1) L_c (2) M_c

3. Q: "M_p". Which of the following two statements could that imply? (1) L_c (2) M_c A:
   (chosen for PaLM-540b)

4. Question: "M_p". Which of the following two statements could that imply? (1) L_c (2) M_c Answer: It could imply
   (chosen for GPT3-DaVinci)

Similar prompts were used with "*implied-by*"-questions.

The following prompts were used as a baseline. In a 0-shot setting, these will produce an accuracy related to the independent log likelihood of each candidate answer. In few-shot settings, these will produce a better performance if the model learns to act as a metaphor detector, independently of any statement connecting the two answers:

1. "" (empty prompt)

2. Pick between the following statements: (1) L_c (2) M_c
   (random choice)

We verified that the results were similar when the random choice baseline was altered to be more similar to the best-scoring prompt for the largest model (PaLM-540b).

To minimise the influence of the answer order on the scores of the model, we presented each question twice, swapping the order of the choices the second time. This has the effect of producing a better mean approximation than presenting each question once with randomised order of choices.