A big ant or a small elephant: metaphor interpretation on large language models

Luiz Matos
Universidade Federal Fluminense
lfmatosmelo@id.uff.br

Aline Paes
Universidade Federal Fluminense
alinepaes@ic.uff.br

Abstract

Figurative language is a fundamental aspect of human communication, making it crucial for any computational system that aims to mimic human interaction to be able to understand and use it. Previous works have indicated that pretrained language models (PLMs) can process figurative language across different layers of the model. This paper evaluates how well large language models (LLMs) can interpret metaphors, a specific form of figurative language. We focus on smaller LLMs, particularly using Llama2-7B for our complete set of results, while also including GPT-3 for comparison. Our early findings suggest a variation in how well different models can handle metaphors.

1 Introduction and motivation

Figurative languages (FLs) are meaningful but not literally true expressions, according to the Merriam-Webster dictionary[1]. They enrich human communication by including unusual expressions that creatively evoke emotion and imagery (Roberts and Kreuz, 1994). One of the most creative ways of introducing FL into a conversation is with metaphors. In a metaphor, a word or phrase denoting literally one thing is used in place of another to suggest a similarity or analogy between them. Previous works have proposed several ways to represent, understand, and analyze metaphors (Lakoff and Johnson, 1980; Group, 2007; Steen, 2008). One of the most well-investigated frameworks is the conceptual metaphors (CMs) (Lakoff and Johnson, 1980), which defines metaphors as mappings between source and target domains. Figure 1[2][3] shows two different uses of the verb *flew*. The first one is metaphorical, juxtaposing source (*"organism that flies"*) and target (*"speed"*) domains to describe the subject's speed in terms of an ability to fly. Otherwise, the second sentence is literal, lacking contrast between the word's usage context and its common sense meaning.

Dealing with the complexity of metaphors is challenging for PLMs. Jang et al. (2023) finds that models perform worse for metaphor detection compared to other FL types even if obvious class cues exist. On the other side, Aghazadeh et al. (2022) show that BERT (Devlin et al., 2018) and variants encode metaphor knowledge, enabling transfer learning for similarly annotated datasets. Also, Wachowiak and Gromann (2023) show that GPT-3 achieves 65.15% accuracy in CM source domain inference according to manual inspection of predictions' alignment, but suffers from domain hallucinations in some cases.

This paper investigates the extent and limitations of LLMs' metaphorical knowledge by exploring CMs. To this end, three tasks are selected: 1) metaphor classification, 2) inference of CM lexicons, and 3) inference of CM domains. Notably, we target at investigating the abilities of not-that-large LMs on those tasks eliciting the smallest Llama version (Llama2-7B) (Touvron et al., 2023). To investigate even in a limited extent the role of the size to the tasks we compare with GPT-3 (Brown et al., 2020), reporting preliminary results below. Code can be found on GitHub[4].

2 Datasets and tasks

2.1 Datasets

The datasets we rely on this work are metaphor interpretation datasets such as TroFi (Birke and Sarkar, 2006), VUA Verbs, VUA POS (Steen et al.,

[1] https://www.merriam-webster.com/grammar/figurative-language

[2] https://players.fcbarcelona.com/en/player/763-ronaldinho-ronaldo-assis-moreira, credits to Miguel Ruiz

[3] https://unsplash.com/pt-br/fotografias/bando-de-passaro-amarelo-voando-EGcfyDiUv58, credits to Gareth Davies

[4] https://github.com/lfmatosm/metaphor-interpretation-on-llms

"Ronaldinho flew over the field to score another goal."
"The birds flew over the fields."

Figure 1: Non-literal and literal uses of the verb *flew*.

2010), Metaphor List and LCC's English version (Mohler et al., 2016). TroFi includes metaphorical and non-metaphorical usage of English verbs. VUA includes metaphorical language annotations following the MIPVU procedure (Steen et al., 2010). Metaphor List includes CM sentences and was collected by Wachowiak and Gromann (2023). Finally, LCC includes CM domain and lexicon annotations on news/web data.

TroFi and VUA Verbs/POS follow the same preprocessing schema as Aghazadeh et al. (2022) but reducing VUA Verbs/POS to respectively $\frac{1}{4}$ and $\frac{1}{6}$ of their sizes, while Metaphor List follow the procedure of Wachowiak and Gromann (2023). Only sentences with level 3 metaphoricity - the maximum defined in the dataset - were considered for LCC, and infrequent source/target domain combinations were dropped. Following Aghazadeh et al. (2022), the final set was divided into train, test and dev splits with respective ratios of 0.7, 0.2 and 0.1. The first three datasets consist of sentence and metaphor presence indicator pairs, while the last two comprise sentences and CM information. Table 1 shows the number of instances and examples in the datasets.

Table 1: Data splits with avg. sentence length.

Dataset	**Train**	**Test**	**Dev**	**Length**
TroFi	3838	1096	548	28
VUA Verb	2294	656	328	27
VUA POS	3506	1002	502	29
Metaphor List	132	244	120	7
LCC (en)	1206	345	172	25

2.2 Tasks

Classification and CM domain inference tasks mostly follow the setting defined by Aghazadeh et al. (2022) and Wachowiak and Gromann (2023). For classification, our prompt-based approach includes a sentence and a question if it is metaphoric, expecting a binary answer. For CM domain inference, our prompt consists of the definition of a CM, a sentence, and a domain, asking for the remaining one. Additionally, CM source and target lexeme inference are proposed here. They consist in choosing lexemes associated with the CM from the given metaphorical sentence. Prompts for each task can be seen in the appendix.

3 Methods

3.1 Experiments

Llama2-7B was the focus of the experiments with GPT-3 used as-is to determine the ideal performance to be achieved by the former. The default Llama2-7B model without fine-tuning was our baseline. Llama's Transformers implementation (Wolf et al., 2019) is used, with fine-tuning (FT) with QLoRA (Dettmers et al., 2023), combining low rank adapters (LoRA) (Hu et al., 2021) and 4-bits quantization. The LoRA parameters used during fine-tuning were $\alpha = 16$, dropout $d = 0.1$ and $|dim_{LoRA}| = 64$. Some of the parameters chosen for training were $lr = 2e - 4$, $wd = 1e - 3$. A fitted model for each pair $\{dataset, task\}$ was created. Google Colab Pro's NVIDIA T4 GPU with 16GB RAM was used, with costs including GPT-3 API amounting for a total of US$62 at the time of this writing.

Temperature was set to $t = 0$ during inference to allow reproducibility. Few-shot examples were tried on the dev set to obtain the ideal number of examples to provide for both models; best-performing ones were selected as prefix for test set inference. For $n \in [2, 12]$, n samples were concatenated into a single textual prompt, including expected labels, with the unanswered example for model prediction at the end. This approach is used for test set evaluation before and after FT, but not for FT itself. Due

Table 2: Classification results on each dataset. Problematic results are indicated with an asterisk.

Model	Dataset											
	TroFi				VUA Verb				VUA POS			
	f1	prec	rec	acc	f1	prec	rec	acc	f1	prec	rec	acc
GPT-3	0.59	**0.57**	0.61	**0.58**	0.59	0.57	0.61	0.57	0.56	0.53	**0.60**	0.54
Llama2-7B	0.61	0.55	0.70	0.56	**0.60**	0.57	0.62	0.58	0.55	0.52	0.58	0.52
+ FT	**0.67***	0.50*	**1.00***	0.50*	**0.65**	**0.64**	**0.67**	**0.65**	**0.58**	**0.59**	0.57	**0.59**

Table 3: Source and target domain/lexeme inference results on each dataset.

Model	Dataset					
	Metaphor List		LCC (en)		LCC (en)	
	SD	TD	SD	TD	SL	TL
	$\cos\theta \pm \sigma$	$\cos\theta \pm \sigma$	$\cos\theta \pm \sigma$	$\cos\theta \pm \sigma$	$\cos\theta \pm \sigma$	$\cos\theta \pm \sigma$
GPT-3	0.51 ± 0.21	**0.60** ± 0.24	0.65 ± 0.26	**0.84** ± 0.27	**0.84** ± 0.27	**0.88** ± 0.26
Llama2-7B	0.49 ± 0.12	0.55 ± 0.14	0.55 ± 0.13	0.64 ± 0.13	0.58 ± 0.18	0.63 ± 0.17
+ FT	**0.52** ± 0.11	0.58 ± 0.15	**0.70** ± 0.17	0.71 ± 0.13	0.69 ± 0.13	0.66 ± 0.18

to variability in metrics results depending on the number of few-shot samples provided, the selected number was chosen on a $\{model, dataset, task\}$ basis, that is, an optimal global number of examples could not be defined. Cosine similarity with fastText (Bojanowski et al., 2016; Joulin et al., 2016) between inference and gold label was used to evaluate CM tasks, replacing manual inspection and avoiding out-of-vocabulary tokens.

3.2 Preliminary Results

Tables 2 and 3 describe results in the test set. For classification, FT Llama2-7B performed better in most metrics, with a higher performance in VUA Verb and VUA POS, reinforcing Aghazadeh et al. (2022) as both datasets share annotation criteria, explaining similar results. However, its results in TroFi indicate misclassification of samples into a single class. Further investigation is needed to understand what caused such behavior.

For domain inference, target was easier to infer, as source requires more effort, being not trivially derivable from the sentence. Overall, FT Llama2-7B was better than its baseline counterpart. For CM lexemes, GPT-3 achieved the best results overall across tasks. This result may indicate the task's ease in comparison with domain inference. Though accuracy is not reported, it was noted during evaluation that FT Llama2-7B inferences often included gold label followed by an excerpt of the input prompt context prefix.

4 Conclusion

This work aimed to analyze the extent of metaphor knowledge in models such as Llama2-7B and GPT-3. Preliminary results show that this knowledge do exist, but vary between models, with each one being better on distinct tasks. FT improved Llama2-7B's classification performance, and even without it, both models achieved results slightly above chance. As such, a smaller task-tuned model can be as competitive as large ones. Regarding inference, Llama2-7B frequently hallucinates responses according to prefix prompts, which indicates an improvement area. Additionally, performance in TroFi was poor, with results pending investigation.

Besides fixing prompts and model's behavior across datasets for more precise results, this study creates other possibilities. We are mapping available metaphor knowledge inside the model architecture as ongoing work. Additionally, the exploration of smaller and linguistically aware LM architectures, alongside external knowledge bases, can increase the interpretability and explainability of such models.

Acknowledgments

This research was financed by CNPq (National Council for Scientific and Technological Development), grant 307088/2023-5, FAPERJ - *Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro*, processes SEI-260003/002930/2024, SEI-260003/000614/2023.

We are also deeply thankful for the anonymous reviewers' comments and suggestions that contributed with the improvement of this work in its final version.

References

Ehsan Aghazadeh, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. 2022. Metaphors in pre-trained language models: Probing and generalization across datasets and languages. *arXiv preprint arXiv:2203.14139*.

Julia Birke and Anoop Sarkar. 2006. A clustering approach for nearly unsupervised recognition of nonliteral language. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 329–336.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Pragglejaz Group. 2007. Mip: A method for identifying metaphorically used words in discourse. *Metaphor and symbol*, 22(1):1–39.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Hyewon Jang, Qi Yu, and Diego Frassinelli. 2023. Figurative language processing: A linguistically informed feature analysis of the behavior of language models and humans. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9816–9832.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

George Lakoff and Mark Johnson. 1980. Metaphors we live by. *University of Chicago, Chicago, IL.*

Michael Mohler, Mary Brunson, Bryan Rink, and Marc Tomlinson. 2016. Introducing the lcc metaphor datasets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4221–4227.

Richard M Roberts and Roger J Kreuz. 1994. Why do people use figurative language? *Psychological science*, 5(3):159–163.

Gerard Steen. 2008. The paradox of metaphor: Why we need a three-dimensional model of metaphor. *Metaphor and Symbol*, 23(4):213–241.

Gerard Steen, Aletta G Dorst, J Berenike Herrmann, Anna Kaal, Tina Krennmayr, Trijntje Pasma, et al. 2010. A method for linguistic metaphor identification. *Amsterdam: Benjamins.*

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lennart Wachowiak and Dagmar Gromann. 2023. Does gpt-3 grasp metaphors? identifying metaphor mappings with generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1018–1032.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

A Task prompt examples

Table 4: Prompt and completion examples for each task.

Task	Datasets	Labels	Prompt	Inference
Classification	TroFi VUA Verb VUA POS	yes/no	"**Sentence**: For off-duty shifts, the Air Force is starting to build concrete dugouts for about 80 persons, where one takes off the chem suit and rests on a cot **Question**: Is the sentence metaphoric? **Answer**:"	yes
SD/TD Inference	Metaphor List LCC (en)	-	"**Context**: In linguistics, conceptual metaphors consists of understanding a given concept in terms of another **Task**: Extract the source domain from the sentence **Sentence**: I’ve lost all hope of a solution. **Target (or source) domain**: hope (or possessions) **Answer**:"	possessions (or hope)
SL/TL Inference	LCC (en)	-	"**Context**: In linguistics, conceptual metaphors consists of understanding a given concept in terms of another **Task**: Extract the source lexeme from the sentence **Sentence**: 8th June 2014 & 01:26 PM I think my gun safe must be a ""fertile zone"" for gun breeding. **Target (or source) lexeme**: gun (or breeding) **Answer**:"	breeding (or gun)