Comparing Plausibility Estimates in Base and Instruction-Tuned Large Language Models (Abstract)

Carina Kauf
Massachusetts Institute of Technology
ckauf@mit.edu

Emmanuele Chersoni
The Hong Kong Polytechnic University
emmanuele.chersoni@polyu.edu.hk

Alessandro Lenci
University of Pisa
alessandro.lenci@unipi.it

Evelina Fedorenko
Massachusetts Institute of Technology
evelina9@mit.edu

Anna A. Ivanova
Georgia Tech University
a.ivanova@gatech.edu

1 Introduction

The success of large language models (LLMs) on diverse linguistic tasks (e.g., Devlin et al., 2019; Liu et al., 2019; Brown et al., 2020; Achiam et al., 2023) fueled an increase in their popularity, and in the research aiming at assessing their capabilities. An important domain to test is LLMs' world knowledge: language training data contains vast amounts of information about the world, including both explicit factual data and distributional knowledge, inferrable via text co-occurrence patterns (Elazar et al., 2022; Kang and Choi, 2023).

We focus on a specific way to assess general the world knowledge of LLMs: estimates of semantic plausibility. Plausible sentences conform with world knowledge whereas implausible sentences violate it; thus, the ability to distinguish plausible and implausible sentences is an indicator of world knowledge. Among the methods for evaluating the linguistic knowledge of LLMs, minimal sentence pair comparisons of log-likelihoods have been widely-adopted, as they allow for an unsupervised evaluation of what the model has absorbed just with pretraining (Futrell et al., 2019; Warstadt et al., 2020; Hu et al., 2020; Aina and Linzen, 2021; Pedinotti et al., 2021; Sinha et al., 2022; Kauf et al., 2023; Misra et al., 2024).

Since recently the focus in NLP shifted towards LLMs that have been fine-tuned to follow instructions (Chung et al., 2022; Touvron et al., 2023; Almazrouei et al., 2023; Jiang et al., 2023), which are designed to interact with users via *prompts*, prompting emerged as a way to directly query LLMs for the knowledge they encode (Li et al., 2022; Blevins et al., 2023; Hu and Levy, 2023).

What is the effect of the instruction tuning process on a model's knowledge of semantic plausibility? And as prompting does not suffer from the confounders of log-likelihoods (e.g. frequency, word length etc.), could it turn out to be a better method for extracting plausibility knowledge? To address such questions, we use both log probabilities and prompting with base and instruction LLMs to rate the sentences of two datasets, and compare the predictions with human-elicited ratings.

2 Experiments

2.1 Datasets

We used two sentence sets adapted from previous studies: a schematic illustration of the items in each of the datasets can be seen in Table 1.

EventsAdapt (Fedorenko et al., 2020) is composed of 391 items, each of which includes (i) a plausible active sentence that describes a transitive event in the past tense, (ii) the implausible version of the same sentence, constructed by swapping the noun phrases, as well as passive voice alternatives. The items fall into one of two categories: a) animate-inanimate items (AI), where the swap of the noun phrases leads to impossible sentences; and b) animate-animate ones (AA), where role-reversed sentences have milder plausibility violations. Given these differences, we model the two subsets independently.

DTFit (Vassallo et al., 2018) contains 395 items, each of which includes (i) a plausible active sentence that describes a transitive event in the past tense, where an animate agent is interacting with an inanimate patient that is typical for the agent; (ii) or less plausible version of the same sentence with a less typical patient. Typicality values, in this case, depend on the interaction of the patient with both the agent and the verb.

For each set, human plausibility ratings have

Dataset	Plausible?	Possible?	Voice	Example	Source
EventsAdapt (AA, unlikely)	Yes	Yes	Active	The nanny tutored the boy.	Fedorenko et al. (2020)
			Passive	The boy was tutored by the nanny.	
	No	Yes	Active	The boy tutored the nanny.	
			Passive	The nanny was tutored by the boy.	
EventsAdapt (AI, impossible)	Yes	Yes	Active	The teacher bought the laptop.	
			Passive	The laptop was bought by the teacher.	
	No	No	Active	The laptop bought the teacher.	
			Passive	The teacher was bought by the laptop.	
DTFit (AI, unlikely)	Yes	Yes	Active	The actor won the award.	Vassallo et al. (2018)
	No	Yes	Active	The actor won the battle.	

Table 1: Example stimuli from the datasets used in Experiment 1. Names in parentheses indicate event participant animacy (AI = animate agent, inanimate patient; AA = animate agent, animate patient) and the plausibility type of the implausible sentences in the dataset (impossible vs. unlikely).

been collected. We averaged human ratings to obtain a single score for each sentence, and assigned a hit every time the plausible version of the sentence was scored higher than the implausible one.

2.2 Model Plausibility Judgments

We used the Base and the Instruct 7B version of three autoregressive LLMs: `Mistral` (Jiang et al., 2023), `Falcon` (Almazrouei et al., 2023), and `MPT` (MosaicML NLP Team, 2023), and we included `GPT2-XL` (Radford et al., 2019) (1.5B parameters) as a baseline.

We evaluated the models using (i) *LL score*, and (ii) several zero-shot prompting methods. The *LL score* is simply the sum of the log-likelihoods of each token w_i in a sentence.For (ii), we test several prompts designed to explicitly query the LLMs' knowledge of plausibility, using either the same or similar instructions to the task that humans solved on the original datasets.

For each item, we compared the scores of the plausible and the implausible sentence conditions, and assigned a hit every time the plausible version gets a higher score. We considered, as a model's *accuracy*, the ratio of the dataset items in which plausible sentences received a higher probability score.

3 Findings

In our experiments, we found that LL scores are the most effective estimates of semantic plausibility across model architectures, performing consistently above chance on all the sentence subsets, although we observed that on the more challenging *EventsAdapt (AA, unlikely)* subset (i.e. the one not including any animacy distinction between agent and patient), the performance of all models drops significantly (see Figure 1).

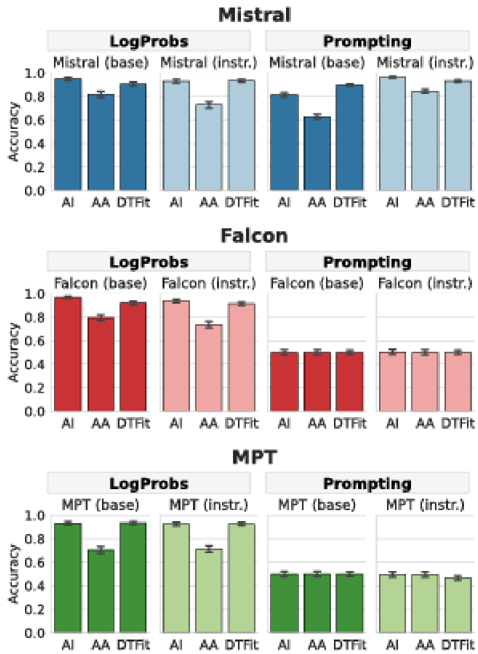

Figure 1: Results of sentence plausibility judgment performance across models and datasets, using LL scores vs. prompting (scores under best prompt settings).

On the other hand, prompting approaches were often a hit-and-miss, with Mistral-7B being the only LLM being consistently above chance, at least in the best prompt settings.

Finally, we found that instruction models perform similar or slightly worse than the corresponding base models, mainly due to a weak performance in estimating plausibility with active voice sentences. The result is in line with other recent findings: although instruction tuning seems to improve LLM alignment with brain representations, it does not always help for alignment at the behavioral level (Kuribayashi et al., 2024).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 Technical Teport. *arXiv preprint arXiv:2303.08774*.

Laura Aina and Tal Linzen. 2021. The Language Model Understood the Prompt was Ambiguous: Probing Syntactic Uncertainty through Generation. In *Proceedings of the EMNLP BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks*.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The Falcon Series of Open Language Models. *arXiv preprint arXiv:2311.16867*.

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2023. Prompting Language Models for Linguistic Structure. In *Proceedings of ACL*.

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.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling Instruction-finetuned Language Models. *arXiv preprint arXiv:2210.11416*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Amir Feder, Abhilasha Ravichander, Marius Mosbach, Yonatan Belinkov, Hinrich Schütze, and Yoav Goldberg. 2022. Measuring Causal Effects of Data Statistics on Language Model's Factual Predictions. *arXiv preprint arXiv:2207.14251*.

Evelina Fedorenko, Idan Asher Blank, Matthew Siegelman, and Zachary Mineroff. 2020. Lack of Selectivity for Syntax Relative to Word Meanings Throughout the Language Network. *Cognition*, 203:104348.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural Language Models as Psycholinguistic Subjects: Representations of Syntactic State. In *Proceedings of NAACL*.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger P Levy. 2020. A Systematic Assessment of Syntactic Generalization in Neural Language Models. In *Proceedings of ACL*.

Jennifer Hu and Roger Levy. 2023. Prompting Is Not a Substitute for Probability Measurements in Large Language Models. In *Proceedings of EMNLP*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Cheongwoong Kang and Jaesik Choi. 2023. Impact of Co-occurrence on Factual Knowledge of Large Language Models. In *Findings of EMNLP*.

Carina Kauf, Anna A Ivanova, Giulia Rambelli, Emmanuele Chersoni, Jingyuan Selena She, Zawad Chowdhury, Evelina Fedorenko, and Alessandro Lenci. 2023. Event Knowledge in Large Language Models: The Gap Between the Impossible and the Unlikely. *Cognitive Science*, 47(11):e13386.

Tatsuki Kuribayashi, Yohei Oseki, and Timothy Baldwin. 2024. Psychometric Predictive Power of Large Language Models. In *Findings of NAACL*.

Jiaoda Li, Ryan Cotterell, and Mrinmaya Sachan. 2022. Probing via Prompting. In *Proceedings of NAACL*.

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*.

Kanishka Misra, Allyson Ettinger, and Kyle Mahowald. 2024. Experimental Contexts Can Facilitate Robust Semantic Property Inference in Language Models, but Inconsistently. *arXiv preprint arXiv:2401.06640*.

MosaicML NLP Team. 2023. Introducing MPT-7B: A New Standard for Open-Source, Commercially Usable LLMs. *www.mosaicml.com/blog/mpt-7b*.

Paolo Pedinotti, Giulia Rambelli, Emmanuele Chersoni, Enrico Santus, Alessandro Lenci, and Philippe Blache. 2021. Did the Cat Drink the Coffee? Challenging Transformers with Generalized Event Knowledge. In *Proceedings of *SEM*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models Are Unsupervised Multitask Learners. *OpenAI Blog*, 1(8):9.

Koustuv Sinha, Jon Gauthier, Aaron Mueller, Kanishka Misra, Keren Fuentes, Roger Levy, and Adina Williams. 2022. Language Model Acceptability Judgements Are Not Always Robust to Context. *arXiv preprint arXiv:2212.08979*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Grave Edouard, and Guillaume Lample. 2023. Llama: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*.

Paolo Vassallo, Emmanuele Chersoni, Enrico Santus, Alessandro Lenci, and Philippe Blache. 2018. Event Knowledge in Sentence Processing: A New Dataset for the Evaluation of Argument Typicality. In *Proceedings of the LREC Workshop on Linguistic and Neuro-Cognitive Resources (LiNCR)*.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. BLiMP: The Benchmark of Linguistic Minimal Pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.