Log probability scores provide a closer match to human plausibility judgments than prompt-based evaluations

Anna A. Ivanova*[1], Aalok Sathe*[2], Benjamin Lipkin*[2], Evelina Fedorenko[2], Jacob Andreas[3]
a.ivanova@gatech.edu {asathe,lipkinb,evelina9,jda}@mit.edu
[1]School of Psychology, Georgia Tech
[2]Department of Brain and Cognitive Sciences, MIT [3]Computer Science and Artificial Intelligence Lab, MIT

Background

- Traditional approach for assessing LLM knowledge: LogProbs

$$P(S \mid C) = \sum_{t=1}^{n} \log P(w_t \mid C, S_{<t})$$

- LogProbs capture many aspects of LLMs' commonsense world knowledge, including knowledge of object properties (Misra et al, 2023) and common events (Kauf, Ivanova et al, 2023), but are sensitive to other factors too.
- New(er) approach: Prompting

```
Rate the likelihood of S given C…
```

- Advantages of Prompting: user-friendly, sensitive to specific tasks
- However: Hu & Levy (2023) showed that Prompting underestimates linguistic knowledge in LLMs relative to LogProbs.
- **WE ASK: How do LogProbs vs. Prompting compare when assessing world knowledge in LLMs using a context-sensitive sentence plausibility task?**

Approach

We compare LLM and human performance on two prompt-based tasks - Choice and Likert - and additionally evaluate LLM performance with the classical LogProbs approach.

Controlled stimuli probing commonsense social relations knowledge (see details below).

We use naïve prompting: identical instructions to those of humans.

For prompting, we constrain outputs to 1-2 (Choice) or 1-5 (Likert); this approach works comparably to free generation.

Choice

Contexts:
1. "Aalok likes Ben."
2. "Aalok hates Ben."
Scenario:
"Aalok and Ben are friends."
Enter the number corresponding to the context that makes more sense. Your response must be either "1" or "2".

Likert

"Aalok likes Ben. Aalok and Ben are friends."
Rate the scenario using a number from 1 to 5, with 1 meaning "makes no sense", and 5 meaning "makes perfect sense".

LogProbs

"Aalok likes Ben. Aalok and Ben are friends."

Results

1. LogProbs is a better metric of LLM knowledge than naive Prompting.
2. Human but *not* LLMs' performance is robust to task variations.

1

	GPT2_XL	MPT_7B	MPT_7B-chat	MPT_30B	MPT_30B-chat
Choice	0.53	0.49	0.50	0.49	0.51
Likert	0.50	0.50	0.51	0.51	0.64
LogProbs	**0.72**	**0.79**	**0.82**	**0.82**	**0.83**

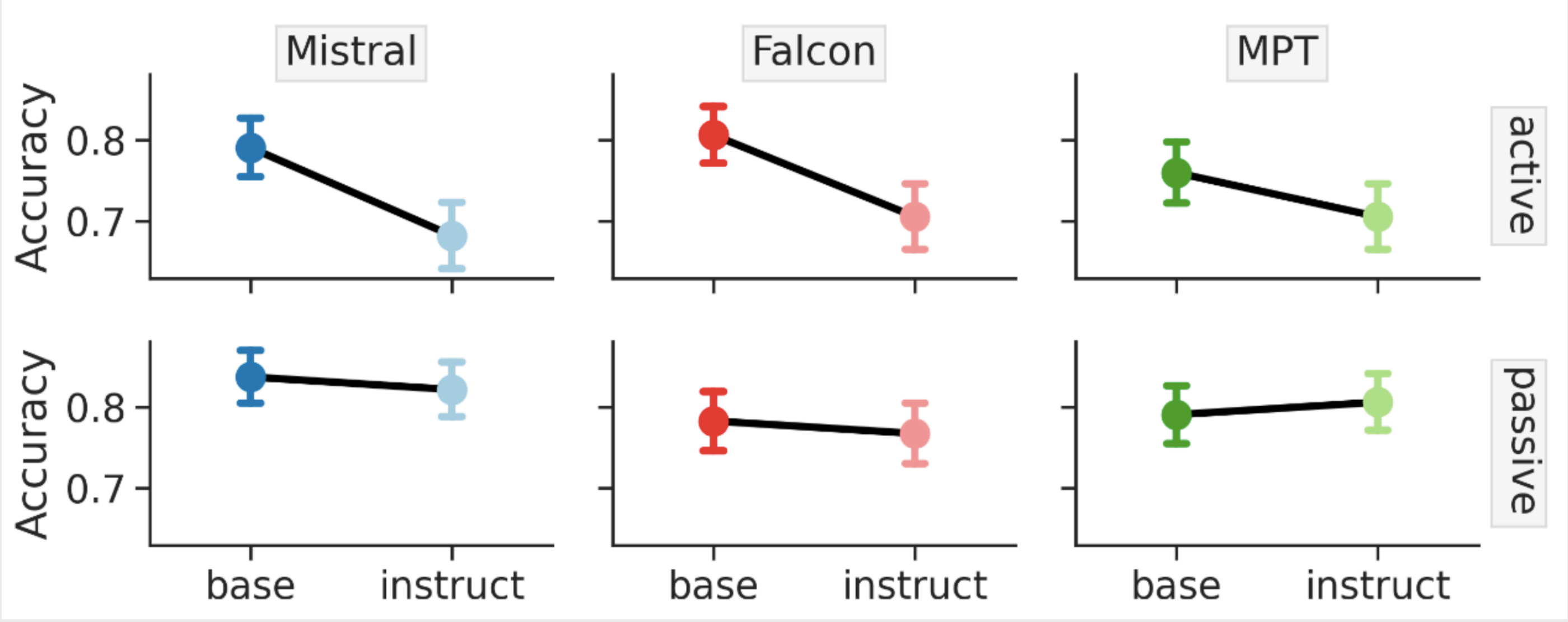
2

	Choice-Likert consistency
Human	0.96
GPT2_XL	0.54
MPT_7B	0.83
MPT_7B-chat	0.63
MPT_30B	0.71
MPT_7B-chat	0.74

See also

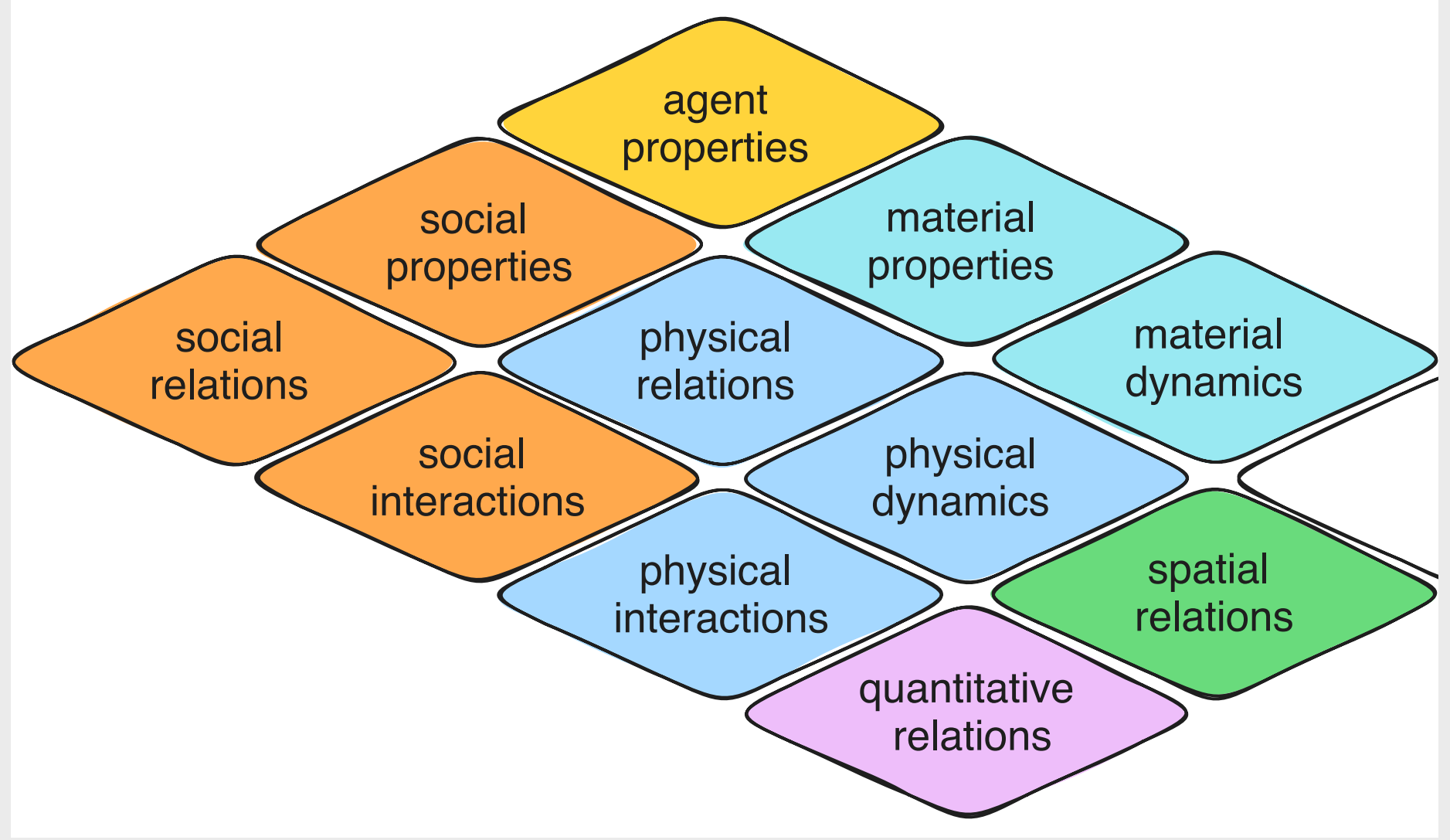
A. In the Kauf et al (2024) preprint, we replicate and extend result #1 on additional datasets and models, both in context-free and context-sensitive settings.

In some cases, RLHF fine-tuning decreases LogProbs performance

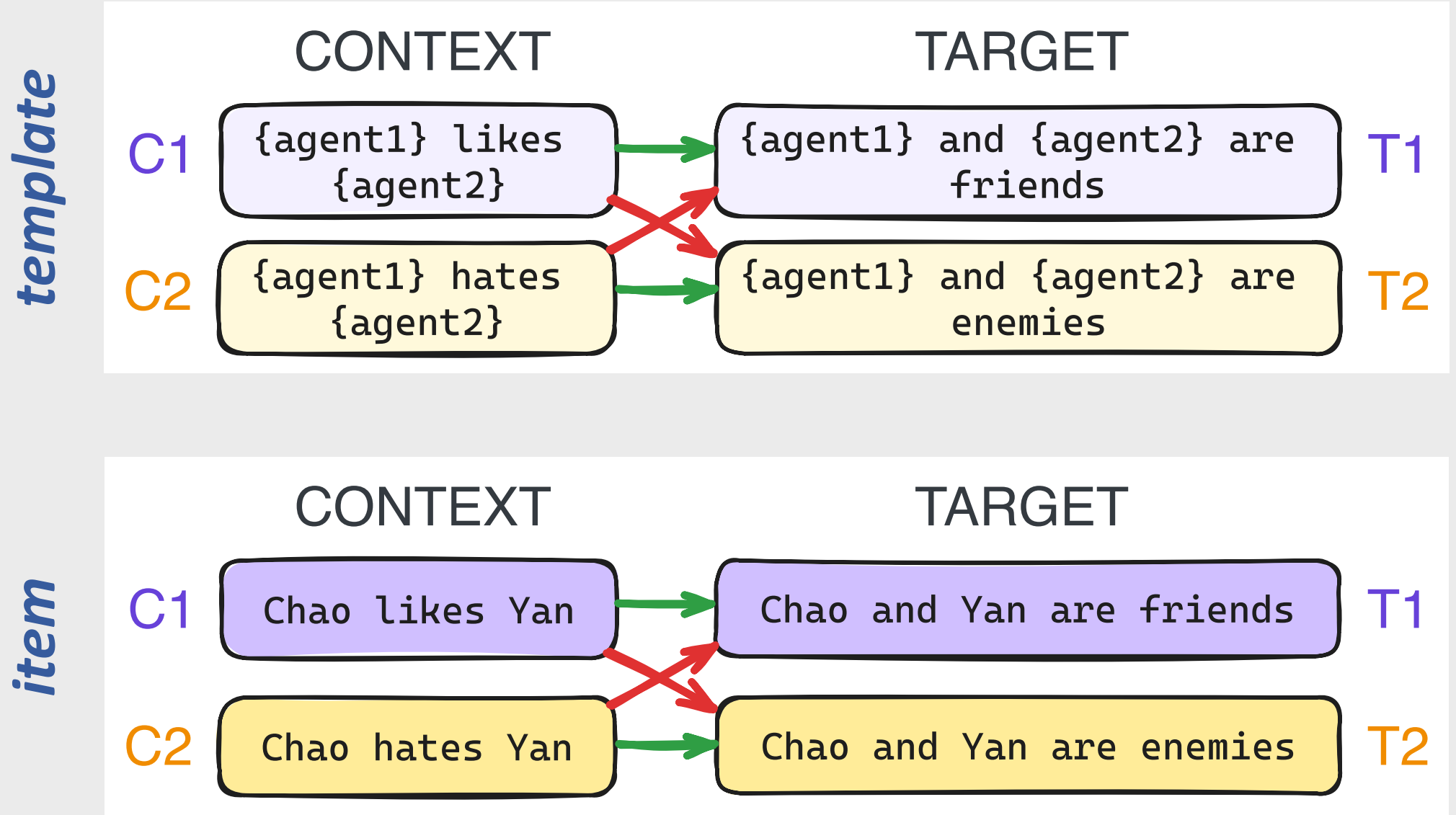


B. This work is part of a broader effort by Ivanova, Lipkin, Sathe et al (in prep) to build a cognitively inspired commonsense benchmark, Elements of World Knowledge (EWoK).

many domains



controlled, curated, generalizable stimuli



Conclusion

- LogProbs are an easy, straightforward way to quickly estimate commonsense world knowledge in LLMs.
- Naive Prompting with instructions identical to humans results in bad performance even when the knowledge is there.
- Tailored prompting will result in better performance but requires model-specific tweaks. Could LogProbs serve as a quick estimate of how successful/easy prompt-engineering can be?

Acknowledgments & References

This work was funded by the Language Mission of MIT Quest for Intelligence.

Hu, J., & Levy, R. (2023). Prompting is not a substitute for probability measurements in large language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 5040–5060).

Kauf, C., Ivanova, A. A., Rambelli, G., Chersoni, E., She, J. S., Chowdhury, Z., Fedorenko, E., & Lenci, A. (2023). Event knowledge in large language models: the gap between the impossible and the unlikely. Cognitive Science, 47(11), e13386.

Kauf, C., Chersoni, E., Lenci, A., Fedorenko, E., & Ivanova, A. A. (2024). Comparing Plausibility Estimates in Base and Instruction-Tuned Large Language Models. *arXiv preprint arXiv:2403.14859*.

Misra, K., Rayz, J., & Ettinger, A. (2023). COMPS: Conceptual Minimal Pair Sentences for testing Robust Property Knowledge and its Inheritance in Pre-trained Language Models. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics (pp. 2928–2949).