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

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

Aalok Sathe*
MIT
asathe@mit.edu

Benjamin Lipkin*
MIT
lipkinb@mit.edu

Evelina Fedorenko
MIT
evelina9@mit.edu

Jacob Andreas
MIT
jda@mit.edu

Abstract

Recent large language model (LLM) advances enable users to interact with LLMs as with another person, by inputting queries and receiving responses in the form of freeform text. Yet such prompt-based interactions might fail to fully leverage LLMs' internal knowledge. We demonstrate a case where naive prompt evaluations fail to capture human judgments on a simple plausibility evaluation task, whereas the traditional scoring method based on extracting conditional log probability scores matches human judgments more closely.

1 Introduction

The explosive popularity of large language models (LLMs) has made it critical to provide reliable assessments of their abilities. Here, we present work that is part of a broader effort to develop a general world knowledge evaluation framework. We ask three questions: (1) do LLMs and humans presented with the same prompts perform similarly? (2) how does prompt-based performance compare to traditional logprob-based evaluations? (3) do humans and LLMs yield consistent judgments in different prompt-based settings?

Prompt-based evaluations offer an easy way to query the knowledge of an LLM. Yet answering a prompt correctly requires not only knowing the answer, but also correctly interpreting the prompt and, in the case of multiple choice prompts, mapping the answer onto a corresponding choice option. As a result, prompt-based evaluations might underestimate the knowledge available to an LLM (Hu and Levy, 2023; Hu et al., 2024).

An alternative approach to evaluating LLMs is calculating the log probability of a response (either by itself or given the appropriate context). Relative log probabilities have been successfully used to distinguish grammatical and ungrammatical sentences (e.g., Warstadt et al., 2020), plausible and implausible events (Kauf et al., 2023), and relevant vs. irrelevant object properties (Misra et al., 2023). Yet raw log probabilities reflect a number of factors that might not be relevant for the task, including sentence length, word frequency (Kauf et al., 2023), and the number of possible paraphrases (Holtzman et al., 2021). Thus, prompt-based evaluation might provide a more task-sensitive approach to evaluating model knowledge.

We evaluate LLM performance on a straightforward yet carefully designed test of commonsense social relations knowledge (Figure 1). We compare humans and models on two tasks: relative sentence plausibility in a forced binary choice task

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."

Figure 1: Three different evaluation strategies. Both models and humans were tested on CHOICE and LIKERT prompts, with identical instructions. Models were additionally tested via log probability (LOGPROBS) scoring.

	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**

Table 1: Accuracy scores for different model and metric combinations, with LOGPROBS consistently outperforming the rest.

	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

Table 2: Response consistency across two task settings in humans is close to 1 whereas in all models it is considerably lower.

(CHOICE) and absolute plausibility using a 5-point Likert scale (LIKERT), which is then used to derive relative plausibility estimates. For LLMs, we also calculate the probability scores assigned to target sentences given either a plausible or an implausible context sentence.

2 Methods

Benchmark. The social relations commonsense knowledge set is part of a bigger, cognitively inspired benchmark called Elements of World Knowledge (EWoK; Ivanova et al., in prep). Item design in this benchmark is inspired by the minimal pairs framework but takes that approach a step further. An item consists of 2 minimal pair context sentences (e.g., C_1: "Aalok and Ben like each other" / C_2: "Aalok and Ben hate each other") and two target sentences (e.g., T_1: "Aalok and Ben are friends" / T_2: "Aalok and Ben are enemies"). The two target concepts juxtaposed here are *friend* and *enemy*. In any item, $p(T_1|C_1) > p(T_1|C_2)$ and $p(T_2|C_1) < p(T_2|C_2)$. Thus, base target probabilities $p(T_1)$ and $p(T_2)$ cannot serve as plausibility cues: a model has to rely on context to establish plausibility. We used 175 items, resulting in 350 CHOICE judgments and 700 LIKERT and LOGPROBS scores (one per context/target combination).

Evaluation. In CHOICE, participants (humans and models) are presented with C_1 and C_2, followed by a single target (T_1or T_2). They are then asked to select the context that better matches the target (i.e., maximizes $p(T|C)$). In LIKERT, participants (humans and models) read C and T presented together and are asked to rate the plausibility of that sequence on a $1-5$ scale. In LOGPROBS, we use tokenwise log probabilities from the LLM to calculate $p(T|C)$ as a sum of conditional log probabilities of each word piece: $\sum_{i=1}^{n} \log P(t_i \mid C, t_{<i})$, where t_i are word pieces composing the target T. See Appendix A for details on human data collection and Appendix B for details on prompt-based scoring.

3 Results

Alignment with humans. We compared human judgments with GPT2_XL (as a basic control) and 4 MPT models: MPT_7B, MPT_7B-chat, MPT_30B, and MPT_30B-chat. We found that LOGPROBS resulted in above-chance performance for all models, including the older GPT2_XL, base models MPT_7B and MPT_30B, and fine-tuned models MPT_7B-chat and MPT_30B-chat, with model size and fine-tuning both contributing to small improvements in performance. Prompt-based performance was almost always at chance.

Consistency. We also compared human and model results derived from different prompt-based tasks—CHOICE and LIKERT—applied to the same items (Table 2). Human results are highly consistent across the two tasks, with the exact match score of 0.96, whereas models are less consistent, in line with the prompt sensitivity issues highlighted in the literature (e.g., Kung et al., 2023).

Overall, we show that 1) LOGPROBS provide a better match to human plausibility judgments than prompts, even when prompts are exactly matched to human instructions, and 2) different prompt-based tasks produce consistent results in humans but not in models, indicating that prompt-based strategies may be an imperfect way for tapping into LLMs' world knowledge.

Limitations

We only present results from a handful of LLMs, although our exploratory experiments indicate that the results generalize to other LLM families. We also only present results from the social relations commonsense knowledge domain, which are consistent with results from other tasks (Hu and Levy, 2023) but the extent to which they apply to different task settings remains to be determined. Finally, our goal was to test LLMs on the exact same instructions as humans, and it is likely that careful prompt engineering will drastically improve their performance; that said, it is important to note that prompt engineering will need to be adjusted every time for a new task variant and a new model (Sclar et al., 2023), whereas the LogProbs approach remains the same for all.

Ethics Statement

Careful evaluations of LLM capabilities are essential for their responsible use in real-life applications. We do not foresee immediate ethically questionable applications of this work.

Acknowledgements

We thank Setayesh Radkani and Thomas Clark for help with creating social relations stimuli, as well as the rest of the EWoK team for insightful discussions.

Data and Code availability

To avoid benchmark contamination, we do not openly release our experimental materials at this time. We are happy to provide them upon request as a password-protected zip file. Analysis code and a general benchmark description are available at `https://github.com/neuranna/social-relations-prompts-logprobs`.

References

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7038–7051, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jennifer Hu and Roger Levy. 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*, pages 5040–5060, Singapore. Association for Computational Linguistics.

Jennifer Hu, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. 2024. Language models align with human judgments on key grammatical constructions.

Anna A. Ivanova, Aalok Sathe, Benjamin Lipkin, RT Pramod, Gabe Grand, Vivian Paulun, Setayesh Radkani, Thomas H. Clark, Carina Kauf, Jennifer Hu, Shari Liu, Ced Zhang, Roger Levy, Evelina Fedorenko, Josh Tenenbaum, Leshem Chosen, and Jacob Andreas. in prep. Elements of World Knowledge (EWoK): A cognitively inspired framework for evaluating basic world knowledge in large language models.

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.

Po-Nien Kung, Fan Yin, Di Wu, Kai-Wei Chang, and Nanyun Peng. 2023. Active instruction tuning: Improving cross-task generalization by training on prompt sensitive tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1813–1829, Singapore. Association for Computational Linguistics.

Kanishka Misra, Julia Rayz, and Allyson Ettinger. 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*, pages 2928–2949, Dubrovnik, Croatia. Association for Computational Linguistics.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*.

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.

Brandon T Willard and Rémi Louf. 2023. Efficient guided generation for large language models. *arXiv e-prints*, pages arXiv–2307.

A Appendix: Human Study

Participants were recruited using Prolific, an online experiment platform. Participants were screened based on self-reported fluency in English as well

as English being their primary language of use. We recruited a total of 30 participants across conditions. Of these, 18 reported identifying as 'female', 11 as 'male', and 1 preferred not to answer. Participants were assigned to either the LIKERT or the CHOICE condition. A total of 16 participants provided LIKERT-scale judgments, whereas 14 provided CHOICE responses to the items.

Each item in LIKERT was split into four sub-items: $(C_1, T_1), (C_1, T_2), (C_2, T_1), (C_2, T_2)$. Similarly, each CHOICE item was split into two sub-items: $(C_{\{1,2\}}, T_1), (C_{\{1,2\}}, T_2)$ as described earlier. Most LIKERT sub-items (C_x, T_y) received at least 4-5 judgments, with all items receiving at least 3 judgments. The average no. of ratings per sub-item were 4. Aggregated, a cumulative 16 people rated each item. All CHOICE sub-items received 7 judgments per $(C_{\{1,2\}}, T_y)$ pair. Participants never saw more than one sub-item of the same item (i.e., participants couldn't rate both (C_1, T_1) and (C_1, T_2) in the LIKERT study). The median RT for a sub-item in LIKERT was 9.2s, and that for CHOICE was 10.4s.

B Appendix: Prompt-based response scoring

We evaluate prompt-based generation in two ways: **free** and **constrained**. In the **free** paradigm, we elicit up to 20 tokens in the completion and look for the first occurrence of a valid response (a numeral between $1-2$ or $1-5$). In the **constrained** paradigm, we only allow completions from a pre-defined set of tokens, i.e., either "`[1-2]`" or "`[1-5]`", using a regex-guided constrained generation algorithm (Willard and Louf, 2023). Here, we report the results from the constrained evaluation approach, as it yielded higher LLM scores.

In the CHOICE task, participants directly indicate whether they prefer C_1 or C_2. In the LIKERT task, we compare the average Likert scores for C_1 and C_2 and select the context sentence with a higher score (both for humans and for LLMs).

C Appendix: Detailed results

In Figure 2, we show detailed response patterns corresponding to the accuracy scores in Table 1.

TASK INSTRUCTIONS (CHOICE)
In this study, you will see multiple examples. In each example, you will be given two contexts and a scenario. Your task is to read the two contexts and the subsequent scenario, and pick the context that makes more sense considering the scenario that follows. The contexts will be numbered "1" or "2". You must answer using "1" or "2" in your response. If you do not speak English or don't understand the instructions, please exit now and do not attempt this task—you will not be paid.
TASK INSTRUCTIONS (LIKERT)
"In this study, you will see multiple examples. In each example, you will be given a scenario. Your task will be to read the scenario and answer how much it makes sense. Your response must be on a scale from 1 to 5, with 1 meaning "makes no sense", and 5 meaning "makes perfect sense". "If you do not speak English or don't understand the instructions, please exit now and do not attempt this task—you will not be paid." Rate the scenario using a number from 1 to 5, with 1 meaning "makes no sense", and 5 meaning "makes perfect sense".

Table 3: Initial instructions presented in the online study. The exact same instructions were used in the prompts for LLMs.

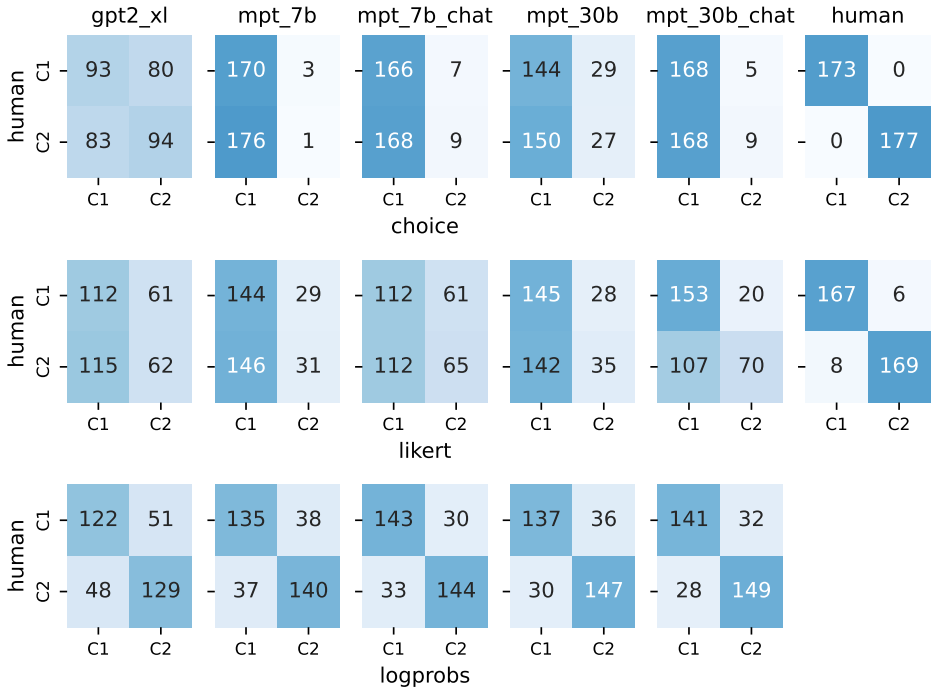

Figure 2: Confusion matrices (with human CHOICE data as the ground truth).