LLMs as assistants or tests subjects? A case study on gender association word norms.

Matthijs Westera
Leiden University
m.westera@hum.leidenuniv.nl

Introduction It has become fairly commonplace in linguistics to use instruct-tuned LLMs as a virtual assistant/annotator/rater, with prompts containing explicit task instructions. This taps into their accumulated statistics from the (increasingly complex and opaque) training regime about how to follow (similar) instructions. Language models used in this manner are capable of displaying behaviors that highly correlate with that of human annotators/raters [1, 2]. Alternatively, one can try to use LLMs, especially base (as opposed to instruction-tuned) models, more like naive participants in a psycholinguistic experiment, by passing in stimuli and observing a model's response without (much) explicit instruction. This taps into their accumulated statistics about patterns in language in general, not specific to an 'instruction following' context.

The aforementioned distinction somewhat resembles that between, on the human side, explicit knowledge and tacit knowledge. Some works have shown that the 'tacit knowledge' of LLMs, specifically their word probabilities, can provide a better fit to human data than LLM responses obtained from explicit instructions, for instance with regard to reading times [3] and plausibility judgments [4]. Unsurprisingly, instruct-tuning an LLM, while making it more suited for instruction following, can make it behave less like an ordinary human language user (ibid.).

Aim and method To contribute to the aforementioned insights we compare explicit to more 'tacit' ways of using LLMs, for predicting existing word ratings on gender association [5]. We compare 'explicit' prompts like (some details omitted) "On a scale from 1-7, how masculine is the word 'plumber'?", with more 'tacit' prompts like "The main character of this story ... plumber ... . Initially, [he/she]" (where the (relative) predicted probabilities of the gendered pronouns are taken as the measured variable). We compare several Llama models (base vs. instruct, 8B and 70B variants, 4bit-quantized), as well as the proprietary GPT-4o. Moreover, for the explicit rating task (using our Python library ChoiceLLM, https://pypi.org/project/choicellm), we compared a scale-rating approach to a comparative approach (e.g., "Of these four words, which is the most stereotypically masculine?"), as well as different frames (e.g., "how masculine?" vs. "how feminine?").

Results and discussion Explicitly prompted, the models display rather different rating distributions (Figure 1); Llama 70B base and instruction-tuned models perform similarly in terms of Pearson correlation to human ratings (Figure 2, left), and slightly outperform GPT-4o, perhaps because the latter has been tuned (more) to avoid gender bias. Llama 8B base is worse than 8B instruct, with the latter only slightly below the 70B models, suggesting that instruct-tuning can make up for limited model capacity. Combining both frames ('how masculine' and 'how feminine') offers a slight advantage for the smaller models and for GPT-4o (with the caveat that the human ratings against which models were evaluated were masculine-framed). In contrast to these results of explicit prompting, the 'tacit' way of prompting base models results in substantially smaller correlations of .57 and .52 for the 8B and 70B base models, respectively (not shown in the plot). Lastly, Figure 2 (right) shows that with explicit prompting, asking for comparative judgments as opposed to scalar judgments can aid weaker models (but requires many prompts to aggregate over).

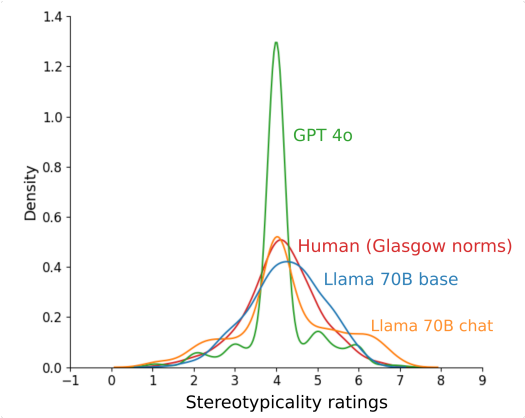

Figure 1: Stereotypicality word rating distributions (KDE) of models vs. human.

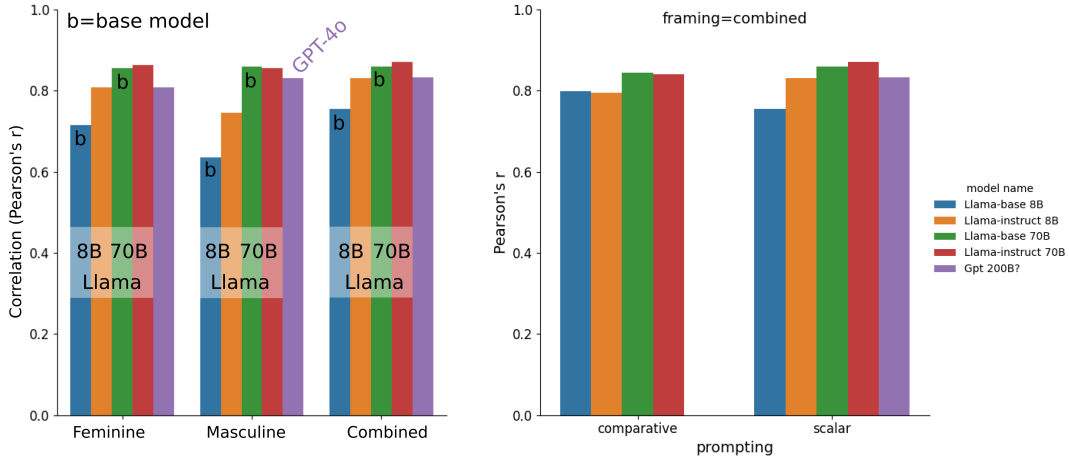

Figure 2: Stereotypicality correlation to human ratings, comparing different models and different framings (scalar prompting, left); and scalar vs. comparative prompting (right).

References

[1] Martínez, G., Molero, J. D., González, S., Conde, J., Brysbaert, M., & Reviriego, P. (2025). Using large language models to estimate features of multi-word expressions: Concreteness, valence, arousal. *Behavior Research Methods*, *57*(1), 1–11.

[2] Almeida, G. F., Nunes, J. L., Engelmann, N., Wiegmann, A., & de Araújo, M. (2024). Exploring the psychology of LLMs' moral and legal reasoning. *Artificial Intelligence*, *333*, 104145.

[3] Kuribayashi, T., Oseki, Y., & Baldwin, T. (2023). Psychometric predictive power of large language models. *arXiv preprint arXiv:2311.07484*.

[4] 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*. https://doi.org/CoRRabs/2403.14859

[5] Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., & Sereno, S. C. (2019). The Glasgow Norms: Ratings of 5,500 words on nine scales. *Behavior research methods*, *51*, 1258–1270.