KI - Künstliche Intelligenz
https://doi.org/10.1007/s13218-026-00907-1

EDITORIAL

# Can You Trust an LLM Even If You Prompted It Yourself?

**Bernd Ludwig**[1]



Recently, I used a LLM to analyse coherence between turns in the MultiWOZ 2.3 corpus. For example, I looked at this question-answer pair:

**Expert**: I found 1 cheap hotel for you that includes parking. Do you like me to book it?

**Customer**: Yes, please. 6 people 3 nights starting on Tuesday.

For a human reader, the customer provides a direct answer to the expert's question and adds more information for relevant parameters of the requested booking action. Although looking simple, explaining why these turns are coherent, is quite complex as each of them consists of two speech acts and only with a holistic view one can completely explain why the customer's answer is exactly matching the information need verbalized in the expert's question.

Analyses such as this are typical application cases for LLMs. Therefore, I prompted a recent LLM to identify communicative intents in both turns and got this response:

**Expert**: states the existence of items, options, or choices. Offers to perform a booking action.

**Customer**: answers the expert's question positively. Requests the expert to act.

On the other hand, MultiWOZ dialogue come with dialog action annotations. For the customer's answer, the annotation is Hotel-Inform. It can be defined as: The customer communicates preferences for a hotel. Obviously, this annotation ignores details in the original utterance. Is it at least compatible with the richer annotation above? For an answer, I asked the LLM "Is the corpus annotation entailed in the LLM annotation?" The response was "No." – quite unsatisfying from a discourse pragmatic point of view as in the customer's request in fact preferences are communicated.

So maybe it's all about wording in the LLM prompt? When I added the statement "If speakers request others to take action, they also communicate preferences.", the LLM this time responded "Yes."

Such observations raise doubts about the reliability of LLMs as judges or annotators. In fact, recent publications on the one hand point to a real gain from using LLMs to label large datasets. In dialogue preference labeling, relevance judgment, and some forms of expert text coding, model labels can reach useful agreement with human annotations at far lower cost. But the same literature also shows that the method is not yet a drop in replacement for gold labels. The central problem is not only bias in the usual sense, but measurement instability. Small prompt changes, response ordering, stylistic cues, and task framing can move labels enough to change conclusions, especially on subjective or frontier cases.

Researchers have identified prompt sensitivity, inconsistency, and biases for position, length, style, and social and authority related biases as causes for wrong LLM responses.

Therefore, a new research agenda is urgent. A more credible science of LLM judging should start by treating prompt robustness as a first-class benchmark target. Judge quality should be reported over sets of independently written prompts, not a single template chosen by the experimenter. The goal is not only high agreement, but stable agreement under reasonable prompt variation.

The second priority is selective automation rather than full replacement. The most promising near-term pattern is to let models label easy or high confidence cases and escalate uncertain cases to humans. That approach fits the evidence better than the stronger claim that human annotation can simply be removed.

Third, replacement claims should be backed by formal statistical tests on calibration subsets, together with uncertainty intervals that account for judge error. This would shift the field away from anecdotal claims of strong agreement and toward explicit evidence that a model is good enough for a stated use.

Fourth, work on subjective tasks should stop forcing all disagreement into a single point label. For dialogue quality and interpretive text annotation, the better goal is often to

✉ Bernd Ludwig
bernd.ludwig@ur.de

[1] Faculty of Informatics and Data Science, University of Regensburg, Regensburg, Germany

Published online: 04 May 2026

model distributions of plausible labels and patterns of disagreement rather than collapse them away.

Finally, evaluation should move toward hard frontier cases, where the judged systems are strong, differences are small, and the costs of measurement error are highest. Those are the settings in which the limitations of current LLM judges are most exposed, and where progress would matter most.

**Funding** Open Access funding enabled and organized by Projekt DEAL.



**Bernd Ludwig**