Qeios

Peer Review

Review of: "LLM Confidence Evaluation Measures in Zero-Shot CSS Classification"

Nicolas Baumard[1]

1. Cognitive Sciences, PSL Research University, Paris, France

This paper explores methods for assessing uncertainty in large language models (LLMs) during zero-shot classification tasks. Developing robust uncertainty quantification techniques is crucial for ensuring the accuracy *and* trustworthiness of LLMs, particularly in computational social science and computational humanities. Overall, this is an important methodological contribution to enhance the reliability of LLMs. We have two minor pieces of feedback, one suggestion, and two open questions.

Minor feedback:

1. The reliance on multiple LLMs in the Confidence Ensemble method may limit its feasibility (despite its capability). Evaluating its performance with fewer (2 randomly picked) models would provide insights into its robustness and, more importantly, its utility in scenarios where access to multiple LLMs is restricted.

2. The study focuses on discrete classification tasks, which are important for many applications. However, tasks involving continuous annotations (e.g., sentiment intensity) are very common in computational social science and computational humanities, and could also benefit from uncertainty quantification methods. Testing the proposed techniques in these contexts would significantly extend their utility. If it is not feasible, the authors could mention this point as a current limitation and avenue for future research.

Suggestion:

We recommend the inclusion of a comparative table summarizing the five UQ methods discussed in the paper, to serve as a quick-reference tool. This table could provide: a succinct definition or

 doi.org/10.32388/R6BM7W

description of the method; the data required for it; the source of the model (whether the confidence is reported by the model or derived from its internal calculations, which seems one of the main differences between the different UQ methods); maybe the strengths and limitations of each method (there are, it seems, computational constraints specific to each method); maybe a “best use case”. We believe this kind of table could help make this paper more actionable. Maybe adding intuitive explanations or examples for each UQ metric would make the results more accessible (it would be important at least for statistical measures like AUC).

Open questions:

1. The results suggest that self-report methods underperform compared to the Confidence Ensemble and Confidence Score. But, given the ease of implementation of self-report methods (both quantitative and qualitative), to what extent do the researchers think these methods could be viable in resource-limited frameworks?

2. As a broader question, do the authors believe that improvements in AI "metacognition"—the ability of models to assess and report their own confidence more accurately—could make self-report methods a more competitive option in the future? For instance, if LLMs were to incorporate better uncertainty estimation mechanisms at the architecture level, could these methods eventually rival more computationally intensive approaches like the Confidence Ensemble?

This review was written in collaboration with Edgar Dubourg.

Declarations

Potential competing interests: No potential competing interests to declare.