Responsible NLP Checklist

Paper title: *Demystifying Uncertainty in LLMs: Active Calibration between Concepts and Human Evaluations*
Authors: *Pengqi Li, Lizhong Ding, Zhehao Zhou, Chunhui Zhang, Jiarun Fu, Hao Li, Ye Yuan, Guoren Wang*

How to read the checklist symbols:

☑ the authors responded 'yes'

☒ the authors responded 'no'

N/A the authors indicated that the question does not apply to their work

☐ the authors did not respond to the checkbox question

For background on the checklist and guidance provided to the authors, see the Responsible NLP Checklist page at ACL Rolling Review.

☑ **A. Questions mandatory for all submissions.**

☑ A1. Did you describe the limitations of your work?
This paper has a Limitations section.

☑ A2. Did you discuss any potential risks of your work?
Potential risks are discussed in Section 6 (Limitations).

☑ **B. Did you use or create scientific artifacts? (e.g. code, datasets, models)**

N/A B4. Did you discuss the steps taken to check whether the data that was collected/used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect/anonymize it?
(left blank)

☑ B6. Did you report relevant statistics like the number of examples, details of train/test/dev splits, etc. for the data that you used/created?
Relevant dataset statistics and experimental configurations are described in Section 3.2 (n-gram Study), Section 4 (Active Calibration with Human), and Section 5.2 (Interactive Learning Strategy).

☑ **C. Did you run computational experiments?**

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
The experimental setup and relevant hyperparameters are detailed in Section 4 (Active Calibration with Human) and Section 5.2 (Interactive Learning Strategy).

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Descriptive statistics and aggregation methods are reported in Section 3.2 (n-gram Study), Section 4 (Active Calibration with Human), and Section 5.2 (Interactive Learning Strategy).

The Responsible NLP Checklist used at ACL Rolling Review is adopted from NAACL 2022, with the addition of ACL 2023 question on AI writing assistance and further refinements based on ARR practice. ACL 2026 used a subset of ARR checklist form.

☒ **D. Did you use human annotators (e.g., crowdworkers) or research with human subjects?**

N/A D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
(left blank)

N/A D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
(left blank)

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating (e.g., did your instructions explain how the data would be used)?
Consent and data-use considerations are addressed implicitly through the choice of publicly available, licensed datasets: The IMDb non-commercial dataset used in Section 3.2 (n-gram Study) is explicitly released for research purposes under a non-commercial license (see References, IMDb 2024). The AR-Bench benchmark employed in Section 4 (Active Calibration with Human) is an open academic dataset (Zhou et al., 2025a) designed for evaluating reasoning and interaction, and does not involve private or personally identifiable data. We additionally evaluate on AmbigQA and HotpotQA (Sections 5.2 / Appendix F), both of which are publicly released research benchmarks for question answering; we use them solely for research evaluation, without collecting any new user data, and do not access or process any private or personally identifiable information beyond what is contained in the original datasets.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
(left blank)

☑ **E. Did you use AI assistants (e.g., ChatGPT, Copilot) in your research, coding, or writing?**

☑ E1. If you used AI assistants, did you include information about their use?
AI assistants (large language models) were used as part of the experimental and writing workflow in two distinct ways: Multiple LLMs (Qwen2.5, LLaMA3, DeepSeek-R1, GPT-4o) were employed within the experiments as policy and response agents to simulate interactive humanLLM dialogues (see Section 4, Active Calibration with Human, and Section 5.2, Interactive Learning Strategy). Their use is clearly documented with model names, parameter scales, and inference settings. Besides, AI writing assistance (e.g., for language polishing and grammar correction) was used under human supervision to improve readability and consistency.