Responsible NLP Checklist

Paper title: *Tree-of-Evidence: Efficient "System 2" Search for Faithful Multimodal Grounding*
Authors: *Micky C. Nnamdi, Benoit Louis Marteau, Yishan Zhong, J. Ben Tamo, May Dongmei 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 7 (Ethical Considerations), including clinical safety concerns (false negatives reducing care, false positives causing alarm fatigue), data privacy under the PhysioNet Credentialed Data Use Agreement, and bias/fairness risks arising from demographic and socioeconomic skew in MIMIC-IV. We emphasize that ToE is a research prototype and not intended for autonomous diagnosis or treatment planning.

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

☑ 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?
Discussed in Section 7 (Data Privacy and Compliance). All clinical data (MIMIC-IV, eICU) is de-identified prior to release by the data providers under HIPAA Safe Harbor. We accessed the data under the PhysioNet Credentialed Data Use Agreement and made no attempt to re-identify patients. LEMMA-RCA contains no personal data (microservice logs and metrics).

☑ 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?
Section 4.1 reports cohort sizes and splits. MIMIC-IV: N=74,829 ICU stays (train=52,597, validation=11,053, test=11,179) with class prevalences per task (E1: 11.5%, E2: 14.1%, E3: 7.4%, E4: 11.2%). eICU spans 208 hospitals. LEMMA-RCA has prevalence 22%.

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.1 (Reproducibility and Hyperparameters) lists all ToE hyperparameters (beam width B=8, max depth Smax=10, candidate pool Nts=24/Nnote=20, =1.0, =0.05, conf=0.9, suff=0.9, batch size 32, NVIDIA A100). Appendix E reports STE temperature sensitivity; Appendix I summarizes full hyperparameter sensitivity; Section 3.4 and Appendix D detail the search objective and its ablations.

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.

☑ 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?
All MIMIC-IV results report mean standard deviation across 5 random seeds (stated in Section 4.1 and repeated in Tables 2, 4, 5, 6, A1, A2). LLM baselines use bootstrap confidence intervals (Figure A3). eICU and LEMMA-RCA results are from a single seed, as noted in the Table 4 caption.

☒ **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.?
This work does not involve human annotators, crowdworkers, or research with human participants. All experiments use pre-existing de-identified clinical databases (MIMIC-IV, eICU) and a public microservice benchmark (LEMMA-RCA).

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)?
No human participants were recruited or compensated.

N/A 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)?
Data consent was handled by the original dataset providers (PhysioNet for MIMIC-IV and eICU). We accessed the data under the PhysioNet Credentialed Data Use Agreement.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
This research uses de-identified, publicly available datasets released under data use agreements and does not constitute human subjects research requiring additional IRB review at our institution. Use of MIMIC-IV and eICU is approved under the PhysioNet DUA.

☑ **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?
Yes. AI assistants were used for limited language polishing and drafting support. All scientific content, experiments, analysis, and final verification were performed by the authors.