

Responsible NLP Checklist

Paper title: *InsLogicBench: An Argumentation Logic Grounded Benchmark for Complex Insurance Claims Adjudication*

Authors: *Jin Liu, Yunpeng Liu, Keyi Wang, Jie Shi, Xiao Xu, Wenkang Huang, Xingzhong Xu, Xin Liang, Yanghua Xiao*

How to read the checklist symbols:

- the authors responded 'yes'
- the authors responded 'no'
- 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?

See the "Ethics Statement" section (Page 9), which explicitly states the dataset does not constitute professional advice, and the "Limitations" section (Page 9)

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?

See the "Ethics Statement" section (Page 9). The paper states: "All claim queries and scenarios in the dataset are synthetically generated and do not involve any real-world insurance claims or private information of individuals."

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?

See Section 4.1 "Dataset Statistics" and Table 2 (Page 5).

C. Did you run computational experiments?

C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

See Section 5.1 "Experimental Settings" (Page 6) and Appendix B.5 for detailed prompts and setup.

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?

See Section 5 "Experiments", Table 4 (Page 6), and Table 5 (Page 7) which report Precision, Recall, F1, Coverage, and Accuracy.

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?

- 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.?

See Appendix D "Human Evaluation Rubrics" and Tables 9, 10, 11.

- 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)?

See Appendix A "Expert Recruitment and Compensation", which states participants were compensated at approximately \$30 per hour.

- 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)?

See Appendix A "Expert Recruitment and Compensation", which states: "All experts provided informed consent regarding the usage of their evaluation data for research purposes."

- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The study involved professional experts for data quality evaluation (paid consultation) rather than human subjects research requiring strict IRB approval. The recruitment details are provided in Appendix A.

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?

The paper explicitly describes using LLMs (e.g., in Section 3.2 and Section 3.3) as part of the data synthesis pipeline to extract clauses, assign truth values, and generate text narratives.