Responsible NLP Checklist

Paper title: *Evaluating Structure-Aware Retrieval and Safety in Statute-Centric Legal QA*
Authors: *Kyubyung Chae, Jewon Yeom, Jeongjae Park, Seunghyun Bae, Ijun Jang, Hyunbin Jin, Jinkwan Jang, Taesup Kim*

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.

N/A A2. Did you discuss any potential risks of your work?
See the Ethical Considerations section, where we discuss potential risks such as hallucinated or overconfident statutory interpretations under incomplete evidence in safety-critical legal QA.

☑ **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?
All data used in this work are derived from publicly available statutory and regulatory texts. These documents do not include personally identifying information or user-generated content, and therefore no additional anonymization or filtering procedures were necessary

☑ 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 3

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4, 5 and Appendix A, B

N/A 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?
This benchmark evaluates fixed pretrained language models under a deterministic inference setting. As no multi-run experimental setup is involved, descriptive statistics such as error bars are not applicable.

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.?
The annotators were members of the same research group, and the instructions were provided informally through in-person discussions and internal communication channels (e.g., Slack). As there was no standalone written instruction document, we did not include full instruction text in the paper.

☒ 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)?
All annotators were graduate students from the same research group who voluntarily participated in the annotation process. No monetary compensation was involved, and participation did not affect academic evaluation.

☒ 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)?
The dataset does not include personal data or user-generated content from individuals; annotators contributed only to the construction and verification of statutory QA pairs. Therefore, consent procedures for personal data collection are not applicable.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The study involved minimal-risk annotation by members of the same research group and did not collect or release personal data. Therefore, institutional ethics review was not required.

☑ **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 were used in two controlled stages of this work. First, in Section 3.1, GPT-4o was employed for optical character recognition (OCR) and PDF parsing to transcribe non-machine-readable statutory content (e.g., scanned tables, mathematical formulas, and annexed technical diagrams) into structured text. All outputs were subsequently manually verified and corrected by the authors in a human-in-the-loop pipeline to ensure zero information loss in safety-critical data. Second, in Section 5, large language models were used as automated judges (LLM-as-a-Judge) to assess answer correctness and safety-related behaviors under predefined evaluation prompts. The evaluation protocol and criteria were fixed in advance, and all results were reviewed and interpreted by the authors.