Responsible NLP Checklist

Paper title: *Beyond Memorization: Testing LLM Reasoning on Unseen Theory of Computation Tasks*
Authors: *Shlok Shelat, Jay Raval, Souvik Roy, Manas Gaur*

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?
Yes. Potential risks are discussed in Section 9. This work is primarily foundational and evaluative, focusing on formal languagetheoretic reasoning tasks. It does not involve human subjects, personal data, or deployment-facing systems. We identify no significant ethical risks associated with the released artefacts. A plausible risk lies in misinterpretation or overgeneralization of the results beyond the evaluated setting; accordingly, we explicitly limit our claims to regular-language formalisms and caution against extrapolation to broader linguistic or real-world domains.

☑ **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?
The datasets used in this study consist exclusively of symbolic, mathematical objects (e.g., regular expressions, automata, transition tables, and formal language specifications) and do not contain natural language text referring to individuals. As a result, they do not include personally identifiable information, offensive content, or sensitive attributes. Because no human-related data is present, anonymization and additional privacy-protection measures are not applicable. This is described explicitly in Section 9 (Ethical Considerations).

☑ 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 3.3 and Section 6. We report detailed dataset statistics, including the number of examples, dataset composition, difficulty categorisation, and construction procedures. These are described in Section 3 (Unseen DFA Construction Dataset), Section 3.3 (Dataset Statistics and Difficulty Level), Section 6 (Concluding Remarks) and Appendix B.2 (Dataset Composition and Sizes). No train, validation, or test splits are used, as the work focuses exclusively on evaluation rather than model training. All datasets are used in their entirety for benchmarking and analysis.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
The experimental setup is described in Section 4 (Evaluation Methodology) and Appendix D (Prompt Templates). All experiments are conducted via API-based access to pretrained language models, which are treated as black-box systems. As a result, no model training or hyperparameter optimisation (e.g., learning rates, regularisation, early stopping) is performed in this work. Decoding and execution settings (such as temperature and determinism controls) are fixed across all experiments to ensure comparability and are reported explicitly in the paper and appendices. Because the study focuses exclusively on evaluation rather than training or fine-tuning, hyperparameter search and best-found hyperparameter values are not applicable.

☑ 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?
Results and analysis are reported in Section 5 (Results and Analysis), including success rates across datasets, prompting strategies, models, and difficulty levels. These statistics are computed over fixed problem sets and clearly indicate aggregate performance (e.g., percentage of correctly constructed DFAs). All experiments are conducted under deterministic decoding settings (e.g., temperature set to zero), and results correspond to single, deterministic runs per model and prompt configuration rather than averages over multiple random seeds. This is stated explicitly to ensure transparency and reproducibility.

☒ **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.?
Not applicable. This study does not involve human participants or annotators.

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, and no compensation was provided.

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)?
No human participants were involved, and no personal data was collected. All datasets are either author-created or sourced from publicly available educational materials, as described in Section 9.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable.

☑ **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?
Section 9. Yes. AI assistants were used in a limited capacity during the development and debugging of auxiliary code. They were not used to generate datasets, ground truth solutions, experimental results, analyses, or conclusions, and did not contribute to the scientific claims of this work. All datasets, experimental designs, validations, and interpretations were carried out by the authors. The limited use of AI assistance does not affect the originality, correctness, or validity of the reported results.