Responsible NLP Checklist

Paper title: *Hopscotch: Discovering and Skipping Redundancies in Language Models*
Authors: *Mustafa Eyceoz, Nikhil Shivakumar Nayak, Hao Wang, Ligong Han, Akash Srivastava*

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?
The method may slightly alter model behavior due to block removal, but introduces no new risks beyond those of the base models. Section 6 discusses limitations.

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

☑ B1. Did you cite the creators of artifacts you used?
See "References" starting on page 6, or section 2 "Related Work" starting on page 1

☒ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
Section 7 (Ethical Considerations) discusses the intention to broaden accessibility of large language models by reducing inference costs. While specific license terms are not detailed in the paper body, artifacts developed (e.g., scaling method, code, weights) are assumed to follow standard open-source practices and will be released under a permissive license (e.g., Apache 2.0) upon publication.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 7 (Ethical Considerations) describes how the method is intended to democratize access to large language models by reducing inference costs. The method was applied solely to publicly available LLMs (e.g., Llama-3.1-8B-Instruct, Qwen2.5-7B-Instruct) and used in a manner consistent with their licenses and intended research uses. No violation of model usage conditions occurred.

☒ 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?
Section 4.1 (Benchmarks and Setup) confirms that all training and evaluation datasets (e.g., GSM8K, ARC, SocialIQA) are standard publicly available benchmarks and do not contain any personally identifying or offensive content. No user-contributed or private data was used.

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.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3 (The Hopscotch Method) and Appendix A.2 (Hyperparameters and Training Details) provide documentation of the algorithm, model scaling parameters, training strategies, and benchmark coverage. Specifics about models (e.g., LLaMA, Qwen) and data used are given in Section 4 (Numerical Experiments).

☑ 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 A.2 (Appendix: Hyperparameters and Training Details) specifies the training data sample count and evaluation library

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

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4.6 (Effects on Model Efficiency) covers model parameters and changes to base memory requirement

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section A.2 (Appendix: Hyperparameters and Training Details) specifies the learning rate, optimizer (Adam), batch size, and training steps used.

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?
(left blank)

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation, such as NLTK, SpaCy, ROUGE, etc.), did you report the implementation, model, and parameter settings used?
Section A.2 (Appendix: Hyperparameters and Training Details) specifies the learning rate, optimizer (Adam), batch size, and training steps used. The implementation used HuggingFace Transformers and PyTorch, which are standard libraries for this research. All package versions and code to be provided in open source implementation for 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.?
(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)

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)?
(left blank)

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

☒ N/A D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
(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?
We used ChatGPT to detect and correct grammar issues.