Responsible NLP Checklist

Paper title: *Calibrated Progressive Distillation: Co-Designing Curriculum and Target Mixing for Knowledge Distillation of Large Language Models*
Authors: *Mengxiang Zhang, Lingyuan Liu*

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?
Our work presents foundational research on knowledge distillation methodology for compressing large language models. It does not include the types of risks outlined in the ARR guideline.

☑ **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 datasets used in this work are publicly available, widely-used benchmarks in the NLP community. These datasets have been vetted by their respective creators and are commonly used for LLM evaluation. They do not contain personally identifying information or offensive content.

☑ 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 5.1 and Section 5.2 report the number of examples and train/validation/test splits for all datasets used. Further dataset descriptions and statistics are provided in Appendix D.2.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5 reports key hyperparameters. Appendix D.3 provides full details including grid search over learning rates and batch sizes, number of training epochs per task, hardware, and all baseline-specific hyperparameter settings.

☑ 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?
Section 5.1 states that results are averaged over five random seeds. Statistical significance is assessed via Wilcoxon signed-rank tests ($p < 0.05$), as reported in Section 5.1. Tables 12 and all appendix

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.

tables clearly present mean scores across random seeds. The evaluation details (including the use of five independent generation runs with distinct random seeds) are described in Appendix D.4.

☒ **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.?
No human annotators or participants were involved in this research.

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 annotators or participants were involved in this research.

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 annotators or participants were involved in this research. All datasets used are publicly available benchmarks.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No human subjects research was conducted. All data used are publicly available benchmarks.

☑ **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 purely for language polishing of the paper (paraphrasing and refining the authors' original content), consistent with Category (a) of the ACL AI Assistance in Authorship guidelines - Assistance purely with the language of the paper does not need to be disclosed.