Responsible NLP Checklist

Paper title: *Guaranteed Guess: A Language Modeling Approach for CISC-to-RISC Transpilation with Testing Guarantees*
Authors: *Ahmed Heakl, Sarim Hashmi, Chaimaa Abi, Celine Lee, Abdulrahman Mahmoud*

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?
See Section 7 (Limitations). We discuss risks including failure under compiler optimizations (e.g., -O2), incomplete semantic guarantees due to reliance on unit tests, and possible misuse of transpiled binaries in security-critical contexts without thorough validation.

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

☑ B1. Did you cite the creators of artifacts you used?
See Section 3 (Data Collection) and Conclusion. We used AnghaBench and Stackv2 (public, permissively licensed). We cite AnghaBench (Da Silva et al., 2021) and The Stack v2 (Kocetkov et al., 2022) explicitly.

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
In 3.1, we state that Stackv2 consists of permissively licensed code. AnghaBench is open-source (cited with proper attribution). Our released models and datasets will also be open-sourced under permissive terms (GitHub page linked).

☑ 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)?
All datasets (Stackv2, AnghaBench) are research benchmarks released for code modeling. We use them consistently for training and evaluation of LLMs in 3.1.

☒ 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 3.1 Data Collection. The datasets (AnghaBench, Stackv2) contain only open-source code and benchmarks. No personal, identifying, or offensive content is present.

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.?
We document dataset sources, compilation methods, and preprocessing steps in 3.1 Data Collection. Benchmarks (HumanEval, BringUpBench) are also described in 4.2 Evaluation with task coverage and properties.

☑ 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?
We report dataset size (1.32M samples, ~21B tokens) and filtering criteria in 3.1 Data Collection. Benchmark statistics, including program lengths and token counts, are detailed in 4.2 Evaluation and Figure 2.

☑ **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?
See 3.2 Training. We trained models (0.5B, 1.3B, 1.5B parameters) on A100 GPUs (40GB), for 2 epochs over 1.3M samples. Training took ~3 days with mixed precision and gradient accumulation.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Hyperparameters (batch size 24, lr=2e5, AdamW, cosine schedule, context window 16k, RoPE extrapolation to 32.7k, beam size 8) are reported in 3.2 Training and 4.1 Setup. Ablations in 5.4 Ablation Study further detail their effects.

☒ 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?
We did not report error bars or variance across multiple training runs, since experiments were computationally expensive. Instead, we reported single-run results with deterministic evaluation (see 5 Results and Analysis).

☑ 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?
See 4.1 Setup. We describe using DeepSpeed ZeRO-3, FlashAttention2, liger kernels, and vLLM with their default/recommended parameters. We also specify training optimizer (AdamW with weight decay=0.001, lr=210, cosine schedule).

☒ **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 AI assistants to improve readability, polish wording, and prepare figures in LaTeX/TikZ. They were not used for designing experiments, producing results, or generating research ideas. All research contributions and code implementations are original to the authors.