## Responsible NLP Checklist

Paper title: *Addition in Four Movements: Mapping Layer-wise Information Trajectories in LLMs*
Authors: *YaoYan*

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 Section 8(Limitations). We state that the work does not introduce new harmful capabilities. Main risks are: (i) methodological over-interpretation of correlational evidence from linear probes and logit-lens; (ii) limited but possible dual-use in revealing internal features that could inform adversarial red-teaming; (iii) overgeneralization beyond multi-digit addition or beyond the single model studied. We mitigate these by clearly framing results as correlational, documenting scope and assumptions, releasing reproducibility materials only, and avoiding any instructions or artifacts intended for misuse.*

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

☑ B1. Did you cite the creators of artifacts you used?
*Yes see Section 2 (Related Work) and Appendix A (Reproducibility). We cite all third-party artifacts used, including the base model (e.g., Llama-3-8B-Instruct), related tools (linear probes, logit-lens), and any datasets or libraries. Full bibliographic entries are in References.*

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
*Yes see Appendix A and Abstract. We release our code and probe checkpoints under the Apache-2.0. We release scripts to generate the synthetic arithmetic datasets; we do not redistribute any third-party model weights or datasets. Users must obtain the base model under its original license, and all third-party libraries are used under their respective OSS licenses.*

☑ 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)?
*Yes see Section 4 (Experimental Setup) and Appendix A. Our use of all third-party artifacts is within their stated research/intended use. Prompts are purely synthetic arithmetic expressions; no attempts are made to extract training data or circumvent provider restrictions. We do not deploy or redistribute restricted assets.*

*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.*

☑ 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?
*Yes see Section 3 (Data). Our datasets consist solely of synthetic numeric strings and arithmetic operators generated by scripts. They contain no human-authored text, names, or real-world content; therefore there is no PII or offensive material, and no anonymization is required.*

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
*Yes see Appendix A (Artifact Documentation) and the repository README. We document the purpose and scope of the artifacts, supported models, dataset generation scripts, input/output formats, configurable parameters (e.g., digit length, number of addends, carry patterns), evaluation procedure, versioning and dependency/hardware info, license, and known limitations. A lightweight dataset card is included in the README.*

☑ 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?
*Yes see Section 3 (Methodology) and Appendix. We report dataset sizes and train/dev/test splits, length distributions of numbers, carry/no-carry proportions, prompt templates, and evaluation metrics. We also provide scripts to regenerate the datasets with user-specified sizes and random seeds to match our reported statistics.*

☑ **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?
*Yes see [Section 3: Methodology] and [Appendix A]. We report the base model and parameter count (e.g., Llama-3-8B-Instruct, ~8B params), context length and precision, and that we do inference only (no fine-tuning). For linear-probe training we describe feature dimensionality, classifier size, batch size, number of steps/epochs, and hardware. We also give an estimate of total compute 100 and the type/number of 1A100*

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*Yes see [Section 3: Methodology] and [Appendix A]. We detail prompt templates and decoding settings for the LLM, random seeds, and the full configuration for linear probes/logit-lens: optimizer, learning rate, weight decay, epochs/steps, early-stopping criteria, and feature layer selection. When search is used, we list ranges and the validation metric; best settings are chosen on validation only.*

☑ 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?
*Yes see [Section 4:Experiments and Results] . For probe training we report mean standard deviation over [N] random seeds and state N explicitly. For deterministic analyses (e.g., logit-lens curves) we clarify that results are from a single run. We report sample sizes (number of examples) and describe whether we aggregate by micro accuracy over all sums; where relevant we include confidence intervals or bootstrapped estimates.*

☑ 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?
*Yes We put it in the open source github repository mentioned in the abstract*

☒ **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.?
*N/A no human subjects or annotators were involved; all data consist of synthetic numeric strings and operators, so no instructions were administered.*

☒ 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)?
*N/A we did not recruit or pay any participants; no human subjects were used.*

☒ 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)?
*N/A No human subjects or curated human data were used. All datasets consist of synthetic numeric strings and arithmetic operators generated by scripts; therefore no consent is applicable.*

☒ N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*N/A This work involves no human-subjects research or collection/processing of personal data. We analyze a pretrained model offline using fully synthetic arithmetic data, so IRB/ethics review is not required.*

☒ N/A D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
*N/A No annotators or participants were involved at any stage; all data are synthetically generated.*

☑ **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 only for manuscript polishing (grammar/clarity) and code hygiene (formatting/linting/docstrings); no experiments, analyses, or results were produced by AI, and all content was authored and verified by the authors.*