Responsible NLP Checklist

Paper title: *Better and Worse with Scale: How Contextual Entrainment Diverges with Model Size*
Authors: *Dikshant Kukreja, Kshitij Sah, Gautam Gupta, Avinash Anand, Rajiv Ratn Shah, Zhengkui Wang, Aik Beng Ng, Erik Cambria*

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 study is a behavioral analysis of existing, publicly released pretrained models (Cerebras-GPT, Pythia) evaluated on existing public benchmarks (LRE, Brown corpus). We introduce no new models, datasets, or deployment artifacts. The work characterizes a known vulnerability (contextual entrainment) with the aim of informing more robust retrieval and context-curation practices; the findings do not provide uplift for misuse beyond what is already documented in Niu et al. (2025) and related RAG literature.

☑ **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?
We use the LRE dataset (Hernandez et al., 2024), consisting of factual relational queries over public entities (e.g., country capitals, company headquarters), and the Brown corpus (Francis and Kuera, 1979), a standard academic reference corpus. Both are established public benchmarks containing only publicly available information about public entities; no private-individual PII or offensive content is involved, and no new data was collected from humans.

☑ 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?
Appendix A reports the complete dataset construction methodology, total sample counts (4,265,204 samples), per-condition breakdowns (1,012,847 Related; 1,038,192 Counterfactual; 1,087,453 Irrelevant; 1,126,712 Random), the 100,000-samples-per-relation cap, and the source corpus for Random tokens.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 2 specifies the model families and sizes evaluated (Cerebras-GPT 111M13B; Pythia 410M12B),

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.

the entrainment metrics, and the scaling-law estimation procedure (log-log linear regression with $R > 0.8$ and $p < 0.01$ as evidence thresholds). Appendices A and B detail dataset construction and context-condition templates. We evaluate pretrained checkpoints without further training, so no hyperparameter search was performed over our own training run.

☑ 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?
Elaboration: All scaling-law fits are reported with the exponent b, 95% confidence intervals, R, and p-values (Tables 1, 4, 6). Figures 1, 2, 46, and 10 show 95% confidence and prediction intervals for each fit. Raw entrainment measurements per model size are provided in Tables 3 and 5.

☒ **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)

☒ **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 assistant use was limited to routine editorial assistance (grammar, phrasing, LaTeX formatting) comparable to standard writing tools like Grammarly, and did not extend to generating scientific claims, experimental design, analysis, or interpretation of results. We considered this below the threshold warranting explicit disclosure in the paper text.