## Responsible NLP Checklist

Paper title: *The Dog the Cat Chased Stumped the Model: Measuring When Language Models Abandon Structure for Shortcuts*
Authors: *Sangmitra Madhusudan, Kaige Chen, Ali Emami*

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.*

N/A A2. Did you discuss any potential risks of your work?
*(left blank)*

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

☑ B1. Did you cite the creators of artifacts you used?
*Section 2.1 and Appendix cite the psycholinguistic norms we used for noun selection. Section 4 properly cites all models we evaluated.*

N/A B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
*We created a dataset of grammatically correct sentences using common English words. Since these are purely linguistic constructions without copyrighted content, standard licensing doesn't really apply here. We release the dataset publicly for research use!*

☑ 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 2 explains our dataset is for evaluating how models understand syntactic structures. We use existing models (Section 4) consistent with their intended purpose of language understanding evaluation.*

N/A 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?
*Our dataset contains only grammatically constructed sentences about generic entities (like animals, occupations, vehicles) with no personally identifying information or offensive content.*

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?

*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.*

*Section 2 and Appendix detail the linguistic phenomena (center-embedding), domains covered, and complete noun/verb inventories used.*

☑ 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?
*Table 2 reports dataset composition: 360 sentences generating 9,720 questions, broken down by complexity level and question type.*

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

N/A C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
*We evaluate publicly available models through their APIs. Providers don't disclose model parameters, and our inference-only evaluation costs are minimal.*

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*Section 4 reports our experimental settings: temperature=0, max tokens=16,000, standardized prompts. No hyperparameter search needed for deterministic evaluation.*

☑ 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 4 specifies we average 10 runs for non-thinking models. Section 5 reports statistical significance tests and standard measures.*

☑ 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 3 details using spaCy en_core_web_sm for lemmatization and all-MiniLM-L6-v2 for semantic similarity (threshold 0.9).*

☑ **D. Did you use human annotators (e.g., crowdworkers) or research with human subjects?**

☑ 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.?
*Section 4 provides the study URL with full instructions. Participants completed sentence comprehension questions following the form's homepage instructions.*

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)?
*We had 24 volunteer participants. Since they are unpaid volunteers in a brief linguistic study, payment info wasn't applicable.*

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)?
*The study involved only sentence comprehension tasks without any personal data collection.*

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*Our study only involved reading sentences and answering comprehension questions; this typically qualifies as exempt from IRB review since there's no risk and no personal information collected.*

N/A D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
*We didn't collect demographic information as it wasn't relevant for our sentence comprehension task.*

☒ **E. Did you use AI assistants (e.g., ChatGPT, Copilot) in your research, coding, or writing?**

N/A E1. If you used AI assistants, did you include information about their use?
*(left blank)*