Responsible NLP Checklist

Paper title: *E-Verify: A Paradigm Shift to Scalable Embedding-based Factuality Verification*
Authors: *Zeyang Liu, Jingfeng Xue, Xiuqi Yang, Wenbiao Du, Jiarun Fu, Junbao Chen, Wenjie Guo, Yong Wang*

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 does not pose any ethical risks. It introduces a scalable and efficient factuality verification framework without involving human data, sensitive information, or decision-making systems, and thus does not raise concerns related to fairness, privacy, or misuse.

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

☑ B1. Did you cite the creators of artifacts you used?
We cite all creators of models, datasets, and tools used in our work. Please refer to Sections 4 and 5, as well as Appendix A and B, for detailed references.

☒ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
We only use publicly available open-source models and datasets (e.g., Qwen2, BGE, wiki-bio-hallu, CNN), all of which are covered by permissive licenses suitable for academic research. Since no proprietary or restricted artifacts were used or redistributed, an explicit license discussion was deemed unnecessary.

☑ 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 reused artifacts (models, datasets, tools) were used strictly for academic research, in accordance with their intended purposes as stated by the original creators. We do not redistribute or repurpose any artifacts beyond research. Please refer to Appendix A and B for usage details.

☒ 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 our work are publicly available, curated resources (e.g., Wikipedia, news articles) commonly used in academic research. They do not contain personally identifying information or offensive content, and thus no additional anonymization or filtering was necessary.

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 the domains, languages, and use cases of all datasets used in our experiments. Section 5.1 and Appendix B.1 describe the coverage and construction of wiki-en-sentences, wiki-bio-hallu, CNN, and REVEAL datasets, including distinctions between synthetic and real-world factual errors, as well as domain and linguistic scope.

☑ 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 including examples and sources. Section 5.1 and Appendix B.1 describe the scale and construction of each dataset, such as wiki-en-sentences, wiki-bio-hallu (simple and hard subsets), CNN, and REVEAL.

☑ **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?
We specify the computing infrastructure used for training and inference (Appendix B), and we also reported parameter counts. All models used are standard open-source checkpoints (e.g., BERT-base, Qwen2.5-0.5B) and the training was conducted with limited GPU resources.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We detail the experimental setup and hyperparameter choices in Appendix A.1 and A.2, including optimizer settings, learning rates, batch sizes, number of training epochs, and fine-tuning strategies for both the Decomposer and Checker modules.

☒ 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 report deterministic results from single runs using fixed seeds for all models. Due to computational constraints and the stable performance of our framework across datasets, we did not conduct multiple runs to compute averages or standard deviations. We acknowledge this as a potential area for future robustness 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?
While we describe the main tools used for sentence segmentation (Stanza), model training (Transformers, PyTorch), and inference (vLLM), we did not explicitly report the implementation details of evaluation metrics or preprocessing utilities, as they follow standard practice and default settings commonly adopted in prior work.

☑ **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.?
We did not include the full annotation instructions, as all annotations were conducted by domain-expert colleagues within the research team. The task involved factuality judgment of atomic claims based on Wikipedia-style content and posed no risk to annotators. Given the annotators' familiarity with the task and absence of external participants, formal instructions or disclaimers were not deemed necessary.

☒ 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 external recruitment or payment was involved. All annotations were conducted by members of the research team as part of the projects internal effort. As such, compensation and demographic considerations do not apply.

☒ 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)?
All data used in this study is sourced from publicly available datasets (e.g., Wikipedia, news articles) that do not contain personal or user-generated content requiring consent. No private or sensitive data was collected or used, and all annotations were performed on public-domain texts by in-house researchers.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Our work did not require ethics board approval, as all data is publicly available and annotations were conducted in-house by qualified researchers on non-sensitive, Wikipedia-style content. No personal, user-generated, or private data was collected, and no external participants were involved.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
The annotation was conducted by research team members with domain expertise. As the task involved objective factuality labeling over public Wikipedia-style content, and did not involve subjective or culturally sensitive judgments, reporting annotator demographics was not deemed necessary.

☑ **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 (e.g., ChatGPT and GPT-4o) in two capacities: (1) for improving the clarity and fluency of manuscript writing during the revision process, and (2) for generating synthetic data used in training the decomposer module. These uses are documented in Section 5.1 and Appendix A.1.