Responsible NLP Checklist

Paper title: *Beyond Human Labels: A Multi-Linguistic Auto-Generated Benchmark for Evaluating Large Language Models on Resume Parsing*
Authors: *Zijian Ling, Han Zhang, Jiahao Cui, Zhequn Wu, Xu Sun, Guohao Li, Xiangjian He*

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. We discuss potential risks such as biases in synthetic data generation, misuse of resume parsing in recruitment, and the possibility of generating fake resumes. To mitigate these, we applied debiasing techniques, anonymized sensitive fields, and restricted all artifacts to non-commercial research use under CC BY-NC 4.0.

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

☑ B1. Did you cite the creators of artifacts you used?
References are provided in Section 3.33.4 and the Appendix.

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
We specify licensing and terms for all artifacts. ResumeBench and ResumeBench-Mix are released strictly for non-commercial research and educational purposes under the CC BY-NC 4.0 license; uses of third-party datasets comply with their original terms. See Section Ethical considerations for details.

☑ 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. We discuss artifact use and intended scope in Section 3.4 (Integration of Real-World Resume Data) and in the Ethical Considerations section, where we clarify that existing datasets were used only within their research terms and that ResumeBench and ResumeBench-Mix are released strictly for non-commercial research use.

☑ 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. As described in the Ethical Considerations section, all names, emails, and contact details in

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.

ResumeBench were synthetically generated to avoid personally identifiable information. Sensitive fields were anonymized and manually reviewed to ensure no offensive or inappropriate content remained. For ResumeBench-Mix, only publicly available resumes were used, and this subset is restricted to research evaluation with no external release.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Yes. Documentation of the artifacts is provided in Section 3.33.5 (Dataset Statistics, Integration of Real-World Resume Data, and Dataset Analysis) and in Appendices MQ, where we describe coverage of domains, languages, templates, schema design, and data distributions with illustrative examples.

☑ 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. Dataset statistics are documented in Section 3.33.5, where we report the number of resumes, templates, domains, and languages in ResumeBench. Additional details on ResumeBench-Mix, including real vs. synthetic sample counts and language distribution, are provided in Appendix Q (Table 21).

☑ **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. We report model sizes in Section 4 (Experiments). We also specify the computing setup and report inference latency in Appendix L.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Yes. Section 4 (Experiments) details the evaluation setup. We also discuss hyperparameters such as temperature settings in Appendix C.

☑ 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. We report descriptive statistics in Section 4 (Experiments) and Section 5 (Results), where we present mean scores for SR, KM Ratio, TED, ROUGE-L, and BERTScore across models. Additional breakdowns by language and template are provided in Appendices FJ, and data distribution statistics are given in Appendix Q.

☑ 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 document the packages and parameter settings used (Section 4, Appendix O).

☒ **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.?
No human subjects or external annotators were involved. The dataset consists of synthetic resumes generated via our pipeline and publicly available resumes; we did not recruit participants or collect new human data.

☒ 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 human subjects or annotators included in this study.

☒ 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)?
Our dataset does not involve human subjects or personal data. Synthetic resumes were generated via a privacy-compliant pipeline, and ResumeBench-Mix incorporates only publicly available resumes, which are restricted to research evaluation without external release. Therefore, no additional consent procedures were required.

☒ N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Our work does not involve human subjects research requiring IRB approval.

☒ N/A D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No human annotators are involved in this study.

☑ **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 assistance only for grammar checking. No AI assistants were used for research design, coding, data generation, or analysis.