Responsible NLP Checklist

Paper title: *Hallucination Detection in LLMs Using Spectral Features of Attention Maps*
Authors: *Jakub Binkowski, Denis Janiak, Albert Sawczyn, Bogdan Gabrys, Tomasz Jan Kajdanowicz*

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?
The potential risk of the work are discussed in the Limitations section at the end of the paper, as they are strongly connected to the them.

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

☑ B1. Did you cite the creators of artifacts you used?
Creators of the datasets are cited in Section 4.1, creators of software are cited in Appendix C. In addition, proper credits for code are given in the repository whenever applicable

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
License for artifacts is discussed in Appendix D

☑ 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)?
Intended use for artifacts is discussed in Appendix D.

☒ 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 leveraged widely used benchmarks which were published at peer-review conferences, and besides basic analysis we did not further check whether data contains personally identifying information.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We provide README in the linked repository. Appendix E documents statistics of generated hallucinations datasets.

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.

☑ 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?
Section 4.1, Appendix D contains QA dataset statistics (number of examples, token distribution), Appendix E contains class distribution for hallucination detection datasets generated in this work.

☑ **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?
Appendix C mentions computational resources used in the work, Appendix L provides analysis of cost and time of the method, including API cost estimation

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix C describes the experimental setup, Appendix G.3 presents the best found values of hyperparameters, repository shared in the paper contains full hyperparameter configurations.

☑ 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?
Motivational study relied on statistical tests with p-values presented in Figure 1, and in Table 3. The main results for method comparison were obtained from a single run, and it is mentioned in the paper (caption for Table 1). For ablation in Section 6.3, we provide error bars representing standard deviation across several runs.

☑ 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 4.2 contains parameters for logistic regression used, otherwise we use default hyperparameters. Other important configuration values can be found in the repository.

☒ **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)

N/A D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
(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?
Acknowledgement section and Appendix C