Responsible NLP Checklist

Paper title: *Probing Political Ideology in Large Language Models: How Latent Political Representations Generalize Across Tasks*
Authors: *Tianyi Zhang*

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?
See Section 6 (Discussion) and Section 7 (Conclusion), which discuss risks such as unintended bias amplification during ideological interventions, and the potential misuse of steering mechanisms in politically sensitive applications.

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

☑ B1. Did you cite the creators of artifacts you used?
Citations are provided in Section 2 (Related Work) and Section 3 (Methodology), including Kim et al. for the probing framework and Hugging Face for the LLaMA model.

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
The paper references publicly available models such as LLaMA-2 and GPT-4o-mini in Section 3 (Methodology) and Section 4 (Results). These models are used under academic research licenses and commercial APIs (e.g., OpenAI terms of service).

☑ 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 artifacts (e.g., model checkpoints, synthetic data) are used for academic research and aligned with intended use, as discussed in Section 3.1 and Section 3.2.

☑ 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?
See Section 3.2 (Downstream Tasks). All political statements are synthetically generated and do not contain real PII.

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.?
The dataset creation process and domains (policy topics, personas) are described in Section 3.2 (Downstream Tasks). Statements are simulated from U.S. Congress members across six political domains.

☑ 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?
Dataset statistics (e.g., number of samples = 240, domains = 6) are reported in Section 3.2 (Bias Detection Task).

☑ **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?
The parameter size are included in the model names. The computational budget is negligible as the computation would only take 30-45 minutes on a RTX3090 workstation.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Described in Section 3 (Methodology).

☑ 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?
Results are shown in Section 4 (Results) and visualized in Table 1, Figure 1 through 3.

N/A 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?
(left blank)

☒ **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?
The use of GPT-4o-mini for generating synthetic statements is acknowledged in Section 3 (Methodology). AI assistants were also used to polish the language of writing.