Responsible NLP Checklist

Paper title: *Who Holds the Pen? Caricature and Perspective in LLM Retellings of History*
Authors: *Lubna Zahan Lamia, Mabsur Fatin Bin Hossain, Md Mosaddek Khan*

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 in Section 9 of the paper. While our study does not deploy models in real-world systems, the methods and dataset we release could be misused to construct persuasive yet distorted simulations of historical events. LLM-generated narratives with exaggerated or biased persona framings could spread disinformation, reinforce one-sided interpretations, or simulate credibility in sensitive contexts. Additionally, uncritical application of our metrics may risk over-generalizing identity traits or confirming existing biases. We recommend applying our approach only in controlled research settings and emphasize the need for context-aware safeguards if adapted elsewhere.

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

☑ B1. Did you cite the creators of artifacts you used?
Yes, the citations are in Section 3 and Section 4 of the paper. We cite the original creators of all external artifacts used in our analysis, including Sentence-BERT (all-mpnet-base-v2), RoBERTa-MNLI, Detoxify, spaCy, the API Ninjas Historical Events API, and the LLMs GPT-4o, Claude 3.7 Sonnet, and Gemini 2.0 Flash. Sentence-BERT, RoBERTa-MNLI, Detoxify, and spaCy are cited in the Methodology section (Section 3), while the API and LLMs are cited in the Experimental Settings section (Section 4). Where applicable, version names and URLs are provided to ensure transparency and traceability. Full references are included in the bibliography.

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
Yes. All license and terms of use information is provided in detail in the supplementary materials README.md file under Section 1: Licensing and Terms of Use. All external tools used in our analysis (e.g., Sentence-BERT, RoBERTa-MNLI, Detoxify, spaCy) are publicly available under permissive licenses such as Apache 2.0 and MIT. We accessed GPT-4o, Claude 3.7 Sonnet, and Gemini 2.0 Flash through their respective platforms in accordance with their terms of use. Historical event descriptions were obtained via the API Ninjas Historical Events API under its publicly documented terms of service and supplemented with content from publicly available news archives and institutional repositories, used in line with applicable terms. Our final dataset contains only publicly available historical events and includes no private or restricted material. The dataset and analysis code are

provided as supplementary materials under a CC BY 4.0 International License. A full LICENSE file is also included in the supplementary materials package.

☑ 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. A detailed usage statement, data privacy, licensing information, and terms of use are included in the supplementary materials README.md file under Section 1: Licensing and Terms of Use. Our use of external artifactssuch as Sentence-BERT, RoBERTa-MNLI, Detoxify, and spaCyis consistent with their intended use and permissive licenses (Apache 2.0, MIT). GPT-4o, Claude 3.7 Sonnet, and Gemini 2.0 Flash were used for response generation in accordance with their publicly documented terms of service. Historical event data was obtained via the API Ninjas Historical Events API under its stated usage conditions. All tools and datasets were used solely for research purposes and were neither modified nor deployed commercially. The dataset we created consists of GPT-4o-generated responses grounded in publicly known historical events and includes no personal or identifiable information. While the materials are made available without restriction, we urge caution in applying them to real-world identity simulation or sensitive contexts, as generative outputs may encode subtle narrative biases that could misrepresent or inadvertently stereotype marginalized or vulnerable populations.

☑ 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. Data privacy considerations are detailed in the supplementary materials README.md file under Section 1: Licensing and Terms of Use. Our dataset includes some references to named public figures (e.g., political leaders, governors) drawn from publicly available historical records. These references are contextually appropriate and do not involve private individuals or sensitive personal data. All narrative outputs were generated by GPT-4o and are not sourced from user-generated content. To safeguard against offensive language, we used the Detoxify model to screen all outputs and manually reviewed samples for tone and quality. Given the public nature of the source material and the absence of private or identifiable data, no further anonymization was necessary.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Yes. The dataset is described in Section 4: Experimental Settings of the paper, and the linguistic phenomena analyzed are detailed in Section 5: Results and Discussions. We document our dataset in the paper: it contains 197 historical events from 20002025, all based on English-language data. Each event includes responses generated by GPT-4o for two role-based personasone directly affected and one indirectly affected. The dataset spans domains such as politics, war, public health, law, environmental disasters, etc. While no demographic metadata is attached, the personas reflect diverse sociopolitical positions and institutional roles. Linguistic features include exaggeration, contradiction, grammatical agency (passivization), and toxicity, supporting downstream analysis of bias and caricature in LLM-generated narratives.

☑ 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, we mention these in Section 3 and Section 4 of our paper. Our dataset includes 197 historical events (Section 4, Experimental Settings, of the paper), totaling 985 generated responses across target-simulation, default-persona, and default-topic conditions. For individuation, we trained a classifier using contextual embeddings with an 80/20 train/test split (stratified by label) and reported accuracy on the held-out set (Section 3, Methodology, of the paper). For exaggeration, contradiction,

grammatical agency, and toxicity metrics, we report descriptive statistics (e.g., means, standard deviations, quartiles) across the full dataset.

☑ **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. Model sizes and computational resources are reported in the supplementary materials README.md file under Section 2: Model Sizes and Computational Resources. We did not train or fine-tune any large language models. All primary generations were produced using GPT-4o, which served as the main model for our analysis. The exact number of parameters for GPT-4o has not been publicly disclosed by OpenAI. Additional generations were produced using Claude 3.7 Sonnet and Gemini 2.0 Flash solely for comparative evaluation; the parameter counts for these models have also not been disclosed by Anthropic and Google, respectively. For downstream analysis, we used pretrained models: Sentence-BERT all-mpnet-base-v2 (~110M parameters), RoBERTa-large-MNLI (~355M), Detoxify (based on RoBERTa-base, ~125M), and spaCy en_core_web_sm (lightweight, no word vectors). All computations were conducted on a single NVIDIA T4 GPU on Google Colab. The total GPU usage for the entire project was approximately 3 hours. Model training was required only for the individuation classifier; all other analyses used inference only.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Yes, they are discussed in Section 2.5: Data Preprocessing in the supplementary materials README.md file and Section 4: Experimental Settings of the paper. Our experiments did not involve model training or hyperparameter tuning. All pretrained models (e.g., S-BERT, RoBERTa-MNLI, Detoxify) were used with their default configurations as released by the original authors. Our focus was on applying these models for evaluation rather than optimization. Minimal preprocessing was applied where necessarycertain metrics used tokenization and lowercasing, while others processed raw text directly. The applied preprocessing steps are documented in the supplementary materials README.md file (Section 2.5: Data Preprocessing). The overall experimental pipeline is described in Section 4: Experimental Settings of the paper.

☑ 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, they are discussed in Section 5: Results and Discussions, the Appendix of the paper, and Section 2.4: Total GPU Usage of the anonymized supplementary materials README.md file. For the individuation metric, we trained a classifier using contextual embeddings with an 80/20 train/test split (stratified by label) and reported accuracy on the held-out set along with the mean value and 95% confidence interval, as described in Section 5. For the exaggeration, contradiction, grammatical agency, and toxicity metrics, we report descriptive statistics (means, standard deviations, quartiles), t-values from significance testing, and Cohens d effect sizes across the full dataset. Key values are reported in Section 5, with full statistical tables and tests provided in the Appendix. As noted in Section 2.4 of the supplementary materials README.md file, no repeated runs or random seed averaging were performedeach reported result reflects a single run.

☑ 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. These are reported in Section 3: Methodology, Bibliography of the paper, and Section 2: Model Sizes and Computational Resources of the supplementary materials README.md file. We used established libraries including spaCy (for dependency parsing), Sentence-BERT (all-mpnet-base-v2), RoBERTa-large-MNLI, and Detoxify. All tools were used with default settings and are cited in the

Methodology section of the paper. Parameters of the models are reported in the supplementary materials README.md file under Section 2: Model Sizes and Computational Resources. We did not modify any of these libraries. Version details (where available) and references are provided in the Bibliography of the paper.

☒ **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?
This is declared in the supplementary materials README.md file (Section 3: AI Writing Assistance Declaration). AI was used for grammar and fluency enhancement, predictive text suggestions, and literature search assistance. According to ACL guidelines (https://2023.aclweb.org/blog/ACL-2023-policy/), these specific uses do not require formal disclosure; this information is provided voluntarily for completeness and transparency.