Responsible NLP Checklist

Paper title: *Can Large Language Models Outperform Non-Experts in Poetry Evaluation? A Comparative Study Using the Consensual Assessment Technique*
Authors: *Piotr Sawicki, Marek Grzes, Dan Brown, Fabricio Goes*

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?
Section 11, titled "Limitations," directly addresses the potential risks of the work. Specifically, it discusses the "broader ethical and social implications," including "perpetuating algorithmic biases present in LLM training data and fostering creative homogenization." A similar discussion on the risk of reinforcing "homogenized notions of literary quality" due to shared training data biases is also present in the final paragraph of Section 9, "Discussion."

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

☑ B1. Did you cite the creators of artifacts you used?
Yes, all creators of the scientific artifacts used in this study were cited. Dataset: The 90-poem dataset, created by Lamb et al. (2015), is introduced and cited in multiple sections, with a full description in Section 4. Models: The Large Language Models used for the evaluationClaude-3-Opus (Anthropic, 2024) and GPT-4o (OpenAI, 2024)are introduced and cited in Section 1 and further specified in Section 3. Full citations for all artifacts are provided in the References section.

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
Yes, the terms for the artifacts were discussed in two key sections: Section 4 (Dataset): This section explicitly addresses the distribution terms for the 90-poem dataset. It states that "copyright issues prevent us from reproducing the full dataset" but clarifies that "the full dataset will be made available to researchers upon request." It also notes that the "Good" category of poems is publicly archived, with links provided in Appendix A.4. Section 11 (Limitations): This section discusses the nature of the models used, identifying them as "two proprietary models, Claude-3-Opus and GPT-4o." This implicitly addresses their terms of use, as they are not open-source and are governed by the licenses of their respective creators.

☑ 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)?

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.

Yes, the paper's use of existing artifacts is consistent with their intended use, and the intended use of the new artifact (our methodology) is clearly specified. Existing Artifacts: Dataset: The use of the Lamb et al. (2015) dataset is entirely consistent with its original purpose, which was to serve as a benchmark for creativity evaluation research. Our study, as described in Section 1 and Section 4, directly builds upon the original work by re-evaluating this same dataset with a new technique (LLMs instead of humans). Models: The LLMs (Claude-3-Opus and GPT-4o) are general-purpose tools designed for a wide array of language understanding and generation tasks. Using them for the nuanced evaluation of text, as described in Section 3, falls squarely within their intended and widely accepted use cases. Created Artifact: Methodology: The primary artifact created in this work is the CAT-inspired evaluation methodology itself. Its intended useas a framework for the automated assessment of poetry and other creative worksis the central topic of the entire paper. This is specified in the Abstract, the Introduction (Section 1), and is thoroughly explored in the Discussion (Section 9) and Conclusion (Section 10).

☒ 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?

A dedicated discussion on this topic was not included because the nature of the dataset did not present significant risks in these areas. The dataset consists entirely of publicly published poems sourced from established literary journals and a public website for amateur poetry. Personally Identifying Information: The only identifying information in the dataset is the names of the poets, which are part of the public attribution of their creative work. This is not considered sensitive PII requiring anonymization in the context of a literary dataset. Furthermore, as mentioned in Section 3, a key protocol of our CAT-inspired methodology is "creator anonymity," ensuring that the evaluators (the LLMs) were blinded to the author's identity during the assessment to prevent bias. Offensive Content: The poems were selected based on their publication venue's prestige, not their content. As they were sourced from standard, moderated literary outlets, a specific pre-screening for offensive content was not deemed necessary for the scope of this research.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?

Yes, the paper provides documentation for all key artifacts used and created. Dataset Documentation: The dataset's domain (poetry), language (English-language), and sources are thoroughly documented in Section 4. This section details the three quality tiers ("Good," "Medium," "Bad") and the specific publication venues they were sourced from. Further, Section 2.1 discusses the linguistic phenomena that differentiate the poem categories, referencing the work of Kao and Jurafsky (2015) on stylistic differences (e.g., Imagist influence vs. 19th-century style). Appendix A.4 provides direct links to the "Good" poems, offering concrete, documented examples from the corpus. Methodology Documentation: The novel CAT-inspired evaluation methodology is documented in detail in Section 3. The exact prompts used for all five evaluation criteria are provided verbatim in Appendix A.1, ensuring full transparency and reproducibility. The scoring methods are also explicitly defined in Section 7. Model Documentation: The specific versions of the LLMs used (Claude-3-Opus and GPT-4o) are precisely identified in Section 3. The scope of their application is documented in Section 11 (Limitations), which notes that the study is constrained to English-language poems, thereby documenting the language context for the models' use in this work.

☑ 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, the paper reports all relevant statistics for the dataset and its use in the experiments. Core Dataset Statistics: Section 4 (Dataset) specifies the total number of examples (90 poems) and the exact split into three categories ("Good," "Medium," "Bad"), with 30 poems per category. Since this study uses pre-trained LLMs for evaluation rather than training a new model, traditional

train/test/dev splits are not applicable. Experimental Subset Statistics: The paper provides detailed statistics on how the data was sampled for the in-context evaluations: In Experiment 2 (Section 7.2), it is stated that 10 randomized versions of the full 90-poem dataset were used, meaning each poem was evaluated 10 times. In Experiment 3 (Section 7.3), it is detailed that 100 unique subsets of 15 poems were created, each composed of 5 randomly selected poems from each of the three categories. This section even includes a statistical analysis of the sampling process, reporting the expected frequency (16.7) and the actual range of appearances (9 to 25) for each poem. In Experiment 4 (Section 8), the setup is described as using 1 subset of 15 poems, which was evaluated 10 times to measure reliability.

☑ **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?

A discussion of these details was not included because the study utilizes proprietary, closed-source Large Language Models (Claude-3-Opus and GPT-4o) accessed via their respective APIs. Number of Parameters: The exact number of parameters for these state-of-the-art models has not been publicly disclosed by their creators (Anthropic and OpenAI). Computational Budget & Infrastructure: The experiments were conducted by querying external, third-party services. As a result, we did not have access to or control over the underlying computing infrastructure. This makes it infeasible to report specific hardware details or metrics like GPU hours, as the computations were not run on local or dedicated hardware managed by us.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

Yes, the paper details the experimental setup and the key hyperparameter values used. Experimental Setup: The overall methodology, including the CAT-inspired framework, the use of randomized batches, the models, and the five evaluation criteria, is described in Section 3. The exact prompts provided to the LLMs, which are a critical part of the setup, are documented in Appendix A.1 and Appendix A.2. Hyperparameter Values: The single most important hyperparameter for these experiments, temperature, is explicitly discussed. In Section 3, it is stated: "...we simulate a panel of independent raters by setting the temperature hyperparameter to 1 for all queries." The rationale for this choice is also explained. Best-Found Values: While a formal hyperparameter search was not the focus, the paper does mention that the choice of the 1-5 rating scale was based on initial testing. In Section 7, it notes that the prompts "ask the LLMs to evaluate every poem in a set on a scale 1 5, as these values worked best in our exploratory experiments." This indicates that the selected value was found to be effective for the task.

☑ 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, the paper transparently reports summary and descriptive statistics derived from multiple experimental runs, not single runs. Averaged Results: The paper explicitly states that its primary results are based on averaged scores. In Section 7.2, for the 90-poem experiment, it notes: "...the numerical scores for each poem were averaged, and the poems were ranked based on their average scores." In Section 7.3, for the 15-poem experiments, it is stated: "...scores for each poem were averaged across all subsets where they appeared..." This makes it clear that the Spearman's Rank Correlation (SRC) values reported in Table 3 are summary statistics from a large set of experiments (10 runs for the 90-poem setup and 100 subsets for the 15-poem setup). Statistical Analysis (ANOVA): Section 7.3.1 and its corresponding tables (Table 4 and Table 6 in the Appendix) report detailed descriptive statistics. They present the mean evaluation scores for each of the three ground truth categories ("Good," "Medium," "Bad"), along with F-values and p-values to describe the variance between groups. Reliability Statistics: Section 8 and its appendices are dedicated to analyzing the results

from a set of repeated experiments. Table 7 explicitly shows the raw scores from 10 individual runs for a sample subset and then reports the Average Score. Table 8 reports the Intraclass Correlation Coefficient (ICC), a descriptive statistic that summarizes the consistency across those 10 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?
(left blank)

☒ **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.?
(left blank)

☒ 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)

☒ 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)

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
(left blank)

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
The paper did not report detailed demographic or geographic characteristics of the human judges from the Lamb et al. (2015) study. The primary characteristic of the human annotators relevant to our study was their level of expertise, which was consistently reported. Throughout the paper, particularly in the Abstract, Introduction (Section 1), and Section 5, they are identified as "non-expert" human judges. This classification is the critical baseline for our central claim that LLMs can surpass the performance of this specific group. While further demographic details were not provided, the key distinguishing feature necessary for the paper's comparative analysisthe non-expert status of the human judgeswas clearly and repeatedly stated.

☑ **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?
Yes, the entire paper is a detailed report on the use of AI assistantsspecifically Claude-3-Opus and GPT-4oas the primary tools for conducting the research. The methodology for their use as "surrogate judges" in poetry evaluation is the central contribution of the work. Full details of their implementation are provided in several sections: Section 3 (Methodology): This section describes the complete experimental framework, including the specific models used (Claude-3-Opus 2024-02-29 version and GPT-4o 2024-05-13 version), the rationale for their use, and the hyperparameter settings (temperature=1). Appendix A.1 and A.2: These appendices provide the full, verbatim prompts given to the AI assistants for both the in-context and without-context experiments, ensuring complete transparency and reproducibility of their use. Therefore, the use of AI assistants in this research is not merely an incidental part of the process but is the core subject of the investigation, and it is documented exhaustively throughout the manuscript.