TitleTrap: Probing Presentation Bias in LLM-Based Scientific Reviewing

Shurui Du
University of Minnesota, Twin Cities, USA
du000288@umn.edu

Abstract

Large language models (LLMs) are now used in scientific peer review, but their judgments can still be influenced by how information is presented. We study how the style of a paper's title affects the way LLMs score scientific work. To control for content variation, we build the TitleTrap benchmark using abstracts generated by a language model for common research topics in computer vision and NLP. Each abstract is paired with three titles: a branded colon style, a plain descriptive style, and an interrogative style, while the abstract text remains fixed. We ask GPT-4o and Claude to review these title–abstract pairs under the same instructions. Our results show that title style alone can change the scores: branded titles often receive higher ratings, while interrogative titles sometimes lead to lower assessments of rigor. These findings reveal a presentation bias in LLM-based peer review and suggest the need for better methods to reduce such bias and support fairer automated evaluation.

1 Introduction

Large language models (LLMs) are increasingly used as *automatic reviewers* in scientific evaluation, helping conferences and journals screen submissions and offer initial feedback (Gu et al., 2024). Recent studies further show that LLM review scores can be shifted by seemingly superficial factors such as prompt order or verbosity (Ye et al., 2024; Shi et al., 2025).

One prominent cue is the *paper title*. Human studies show that title phrasing can shape first impressions and perceived novelty, sometimes even influencing acceptance decisions (Jamali and Nikzad, 2011). Titles often carry stylistic signals, such as branded colon-style patterns ("X: A Framework for Y") or interrogative forms ("Can We Do Z?"), which may guide attention for both humans and machines.

If LLM reviewers respond to such cues, their scores may reflect *presentation bias* rather than content quality, potentially misleading automated pipelines and downstream human decisions.

We introduce TitleTrap, a controlled benchmark to study this effect. Using a language model, we generate scientific abstracts on common NLP and vision topics and create three title variants for each: (1) *branded colon-style*; (2) *plain descriptive*; (3) *interrogative*. We also compare reviews under two input settings: *title only* vs. *title + abstract*, and disentangle the effects of title *format* from *content*.

We prompt leading LLMs (GPT-4o and Claude) to review each variant under identical instructions. With abstracts fixed, any score differences arise from title framing or input condition.

Our results show that title style can significantly shift LLM review scores: branded titles often score higher, while interrogative ones tend to reduce perceived rigor. These findings reveal a persistent presentation bias in LLM-based reviewing and highlight the need for mitigation strategies to ensure fairer automated evaluation.

2 Related Work

2.1 LLMs for Scientific Evaluation and Peer Review

LLMs are increasingly explored as tools for assisting or even simulating peer review. Zhou et al. (Zhou et al., 2024) benchmarked GPT-3.5/4 for score prediction and review generation, finding persistent weaknesses on long papers and fine-grained critique. Tyser et al. (Tyser et al., 2024) developed *OpenReviewer* with watermarking and long-context prompting but observed over-confident and inflated scoring. Yu et al. (Yu et al., 2024) proposed the *SEA* framework with standardized data and self-correction, improving review quality across conference datasets. Chen et al. (Chen et al., 2025)

Proceedings of the 5th Workshop on Evaluation and Comparison of NLP Systems, pages 119–125
December 23, 2025 ©2025 Association for Computational Linguistics

studied LLM-assisted review with 24 HCI reviewers, reporting reduced workload but little quality gain without human oversight. Jin et al. (Jin et al., 2024) modeled review as a multi-agent process, revealing authority and conformity biases.

These works show that LLMs can accelerate review but remain influenced by contextual and presentation cues. We focus on a subtler yet practical factor: how a paper's *title framing* can bias LLM judgments even with identical abstract content.

2.2 Title Framing and Presentation Effects in Human Review

Human peer review is shaped by cognitive and social biases (Lee et al., 2013), including the classic *framing effect* (Tversky and Kahneman, 1981). Similar effects appear in clinical and decision-making contexts (Malenka et al., 1993; Gong et al., 2013).

Paper titles also guide attention and expectations. Linguistic studies show disciplinary differences in title style (Haggan, 2004), and Hartley (Hartley, 2007) emphasized their rhetorical as well as descriptive functions. Bibliometric analyses reveal that question-style titles increase downloads but reduce citations, while colon-style titles tend to be longer with only modest impact (Jamali and Nikzad, 2011).

These findings suggest titles frame novelty and importance beyond the content itself. We build on this literature to test whether LLM reviewers exhibit similar presentation-driven biases.

2.3 Bias and Robustness in LLM-based Evaluation

The reliability and fairness of LLM-as-a-Judge systems has become a key concern. Gu et al. (Gu et al., 2024) survey common biases and call for standardized protocols. Ye et al. (Ye et al., 2024) quantify position, verbosity, and persona effects, showing persistent sensitivity to superficial cues. Dietz et al. (Dietz et al., 2025) warn that over-reliance on LLM judgments risks reinforcing biases. Shi et al. (Shi et al., 2025) show that minor order changes can flip model decisions due to position bias.

Together, these studies highlight that LLM-based evaluation is still vulnerable to non-substantive presentation factors. We extend this perspective by isolating the influence of the paper's *title* and showing it systematically shifts LLM review scores.

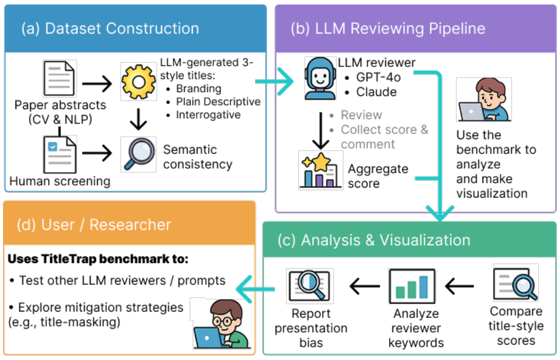
(a) Dataset Construction
Paper abstracts (CV & NLP)
LLM-generated 3-style titles:
• Branding
• Plain Descriptive
• Interrogative
Human screening
Semantic consistency
(b) LLM Reviewing Pipeline
LLM reviewer
• GPT-4o
• Claude
• Review
• Collect score & comment
Aggregate score
Use the benchmark to analyze and make visualization
(d) User / Researcher
Uses TitleTrap benchmark to:
• Test other LLM reviewers / prompts
• Explore mitigation strategies (e.g., title-masking)
(c) Analysis & Visualization
Report presentation bias
Analyze reviewer keywords
Compare title-style scores

Figure 1: Overview of the TITLETRAP workflow. (a) Benchmark construction with controlled title styles and human screening. (b) LLM reviewing with GPT-4o and Claude. (c) Analysis of score differences and reviewer comments.

3 Dataset and Methods

Figure 1 illustrates the TITLETRAP workflow, including benchmark construction, LLM-based reviewing, and analysis.

3.1 Benchmark Construction

We built TITLETRAP from scratch to study presentation bias. Instead of sampling real papers, we used a language model to generate short, research-style abstracts in computer vision (CV) and natural language processing (NLP), similar in spirit to synthetic benchmarks for controlled evaluation such as SciBench (Wang et al., 2024). Prompts encouraged typical problem–method–result structure, and human annotators screened outputs for coherence and plausibility.

For each abstract we produced three title styles:

1. **Branded / Colon-style**: with a coined term (e.g., "TitleTrap: A Benchmark for. . . ").

2. **Plain Descriptive**: standard academic style.

3. **Interrogative**: phrased as a research question.

To disentangle stylistic *format* from coined *content*, we created sub-variants: either fixing the term but changing the format, or keeping the format but swapping the term.

Items were reviewed in two modes: (i) *Title-only* to test pure framing; (ii) *Title+Abstract* to test framing with technical content.

The final benchmark includes 50 CV and 50 NLP abstracts, each with three title variants and title-only versions, enabling systematic analysis of presentation effects as advocated in prior work on

Table 1: Key experimental conditions in TITLETRAP.

Factor	Settings
Input mode	Title-only / Title+Abstract
Title style	Branded / Plain / Interrogative
Format vs. Content	Format fixed / Term fixed
Domains	CV / NLP
Models	GPT-4o / Claude
Scoring	Clarity, Originality, Significance

peer-review robustness (Zhou et al., 2024; Tyser et al., 2024; Yu et al., 2024).

3.2 LLM Reviewer Setup

We prompted GPT-4o and Claude with a standardized rubric for *clarity*, *originality*, and *significance*, following practices similar to other LLM-based reviewing frameworks (Jin et al., 2024; Chitale et al., 2025). For each input, models scored all three titles (1–5), selected the best one, and gave brief justifications. Prompts concealed the study purpose to avoid priming. We collected one review per case due to computational limits, leaving multi-run averaging for future work.

3.3 Evaluation and Analysis

We focused on the factors summarized in Table 1 and tested their influence on review outcomes. Paired statistical tests were used to assess significance, and we also analyzed reviewer comments to understand how titles affected reasoning, consistent with the analytic approaches advocated for evaluating LLM-as-a-Judge reliability (Shi et al., 2025; Ye et al., 2024).

4 Experiments and Results

4.1 Overall Score and Preference Patterns

Figure 2 reports the average scores for clarity, originality, and significance across the three title styles (A: branded / colon-style; B: plain descriptive; C: interrogative), along with the proportion of times each was chosen as the preferred option. Branded titles (A) consistently scored highest on all three metrics and were selected as the preferred choice in over 80% of cases. Plain descriptive titles (B) received the lowest scores and were rarely preferred, while interrogative titles (C) occupied a middle position, sometimes attracting modest preference.

These results indicate that even when abstracts remain unchanged, the surface framing of a title exerts a measurable and systematic effect on LLM judgments.

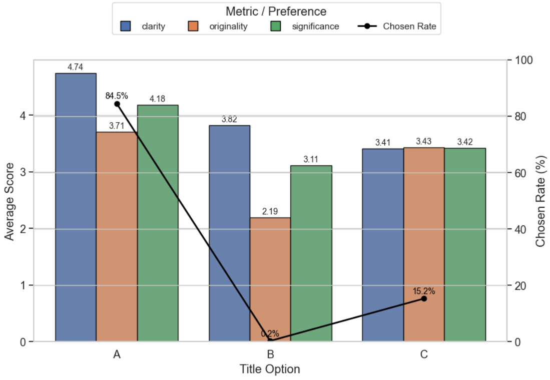

Figure 2: Overall average scores for clarity, originality, and significance under each title option (A/B/C). The black line shows the proportion of times each option was selected as the preferred title.

Table 2: Chosen-title rate (%) across model–mode settings.

Model & Mode	A (%)	B (%)	C (%)
Claude | Title+Abstract	100.0	0.0	0.0
Claude | Title-only	73.0	1.0	26.0
GPT-4o | Title+Abstract	99.0	0.0	1.0
GPT-4o | Title-only	66.0	0.0	34.0

4.2 Model- and Mode-Specific Differences

We next analyzed how results varied across model type and input mode. Figure 3 shows the clarity scores broken down by Claude and GPT-4o, under title-only and title+abstract conditions. Both models favored branded titles, but the effect was stronger for Claude in the title+abstract setting, suggesting that stylistic cues interact with richer content.

Table 2 summarizes the chosen-title rates. Branded titles dominated in all conditions, particularly when abstracts were included. Interrogative titles gained some traction only in the title-only mode, implying that question-style framing may draw attention when no further technical context is available.

4.3 Qualitative Analysis of Reviewer Comments

To better understand these quantitative patterns, we examined the textual review comments. Figure 4 shows the polarity-weighted frequency of selected terms.

Branded titles (A) consistently elicited positive

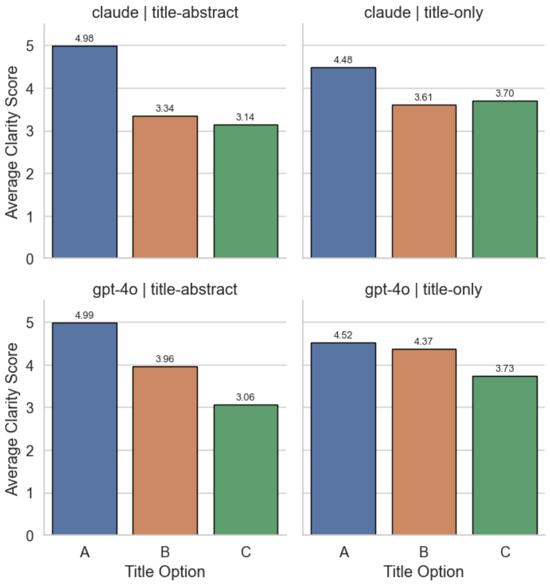

Figure 3: Average clarity scores by model (Claude vs. GPT-4o) and input mode. Branded titles (A) consistently lead to higher clarity scores, with stronger effects for Claude when abstracts are included.

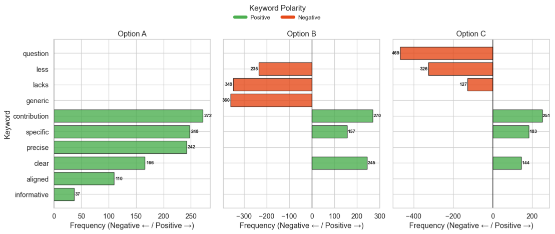

Figure 4: Keyword polarity analysis of reviewer comments for each title style. Branded titles receive more positive descriptors, while interrogative titles elicit more negative ones.

descriptors such as *contribution*, *specific*, *precise*, and *clear*, suggesting that reviewers inferred focus, credibility, and novelty even without additional content. **Plain descriptive titles (B)** were often associated with negative terms such as *generic*, *less*, or *lacks*, but still attracted some positive descriptors like *contribution* and *clear*, indicating that they were seen as accurate yet uninspiring. **Interrogative titles (C)** triggered the highest frequency of negative terms, especially *question*, along with *less* and *lacks*, reflecting skepticism toward rigor and completeness, particularly in the title-only setting.

These observations highlight that title framing not only shapes first impressions but also colors how the abstract is interpreted. A branded format can signal the existence of a concrete framework, a plain descriptive title may be perceived as safe but unremarkable, and a question-style title often amplifies uncertainty even when the underlying content is identical.

5 Discussion and Limitations

5.1 Implications of Title Effects

Our findings show that LLM reviewers are sensitive to surface presentation. Branded or colon-style titles received higher scores than descriptive or interrogative ones despite identical abstracts, indicating reliance on superficial cues. Such sensitivity risks amplifying presentation bias and incentivizing strategic title wording, underscoring the need for review protocols that mitigate framing effects.

5.2 Understanding the Mechanism

Keyword patterns suggest that branded titles convey focus and credibility, while interrogative titles evoke uncertainty. This may reflect biases from training data—where high-impact papers often use branded titles—or simple heuristic shortcuts. Further controlled experiments with synthetic or counterfactual titles could help separate these factors.

5.3 Limitations and Future Work

Our study covered only two domains (CV and NLP), two LLM reviewers, and one prompt style; results may vary across other domains, models, and instructions.

Another limitation is the use of *synthetic abstracts* generated by a language model. This ensured control over content but may not fully capture the complexity of real submissions. Future benchmarks could mix synthetic and human-written abstracts for greater ecological validity.

Finally, we did not examine interactions with human reviewers. Future work should explore human–AI joint review to assess whether human oversight mitigates or amplifies such biases, and test mitigation strategies such as title masking or structured content-only review.

6 Conclusion

We presented TITLETRAP, a benchmark for probing how paper titles influence LLM-based reviewing. With fixed abstracts, we found that branded titles tended to raise, while interrogative titles often lowered, review scores. This highlights a persistent presentation bias in automated reviewing and underscores the need for mitigation to support fairer scientific evaluation.

Acknowledgments

We thank the anonymous reviewers for their helpful feedback and suggestions. This work represents the author's first independently conducted research project.

References

Shiping Chen, Duncan Brumby, and Anna Cox. 2025. Envisioning the future of peer review: Investigating LLM-assisted reviewing using ChatGPT as a case study. In *Proceedings of CHIWORK '25: 4th Annual Symposium on Human-Computer Interaction for Work*, pages 1–18. ACM.

Maitreya Prafulla Chitale, Ketaki Mangesh Shetye, Harshit Gupta, Manav Chaudhary, and Vasudeva Varma. 2025. Autorev: Automatic peer review system for academic research papers. *arXiv preprint arXiv:2505.14376*.

Laura Dietz, Oleg Zendel, Peter Bailey, Charles L. A. Clarke, Ellese Cotterill, Jeff Dalton, Faegheh Hasibi, Mark Sanderson, and Nick Craswell. 2025. Principles and guidelines for the use of llm judges. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, pages 218–229. ACM.

Jingjing Gong, Yan Zhang, Zheng Yang, Yonghua Huang, Jun Feng, and Weiwei Zhang. 2013. The framing effect in medical decision-making: A review of the literature. *Psychology, Health & Medicine*, 18(6):645–653.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*. Version 5, last revised 9 Mar 2025.

Madeline Haggan. 2004. Research paper titles in literature, linguistics and science: Dimensions of attraction. *Journal of Pragmatics*, 36(2):293–317.

James Hartley. 2007. There's more to the title than meets the eye: Exploring the possibilities. *Journal of Technical Writing and Communication*, 37(1):95–101.

Hamid R. Jamali and Mahsa Nikzad. 2011. Article title type and its relation with the number of downloads and citations. *Scientometrics*, 88(3):653–661.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024. AGENTREVIEW: Exploring peer review dynamics with LLM agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*. Available at `https://agentreview.github.io/`.

Carole J. Lee, Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin. 2013. Bias in peer review. *Journal of the American Society for Information Science and Technology*, 64(1):2–17.

David J. Malenka, John A. Baron, Sarah Johansen, Jon W. Wahrenberger, and Jonathan M. Ross. 1993. The framing effect of relative and absolute risk. *Journal of General Internal Medicine*, 8(10):543–548.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2025. Judging the judges: A systematic study of position bias in llm-as-a-judge. *arXiv preprint arXiv:2406.07791*.

Amos Tversky and Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458.

Keith Tyser, Jason Lee, Avi Shporer, Madeleine Udell, Dov Te'eni, and Iddo Drori. 2024. Openreviewer: Mitigating challenges in LLM reviewing.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2024. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*. To appear at ICML 2024.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, and 1 others. 2024. Justice or prejudice? quantifying biases in llm-as-a-judge. *arXiv preprint arXiv:2410.02736*.

Jianxiang Yu, Zichen Ding, Jiaqi Tan, Kangyang Luo, Zhenmin Weng, Chenghua Gong, Long Zeng, Renjing Cui, Chengcheng Han, Qiushi Sun, Zhiyong Wu, Yunshi Lan, and Xiang Li. 2024. Automated peer reviewing in paper sea: Standardization, evaluation, and analysis. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. Accepted at EMNLP 2024.

Ruiyang Zhou, Lu Chen, and Kai Yu. 2024. Is LLM a reliable reviewer? a comprehensive evaluation of LLM on automatic paper reviewing tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9340–9351, Torino, Italia. ELRA and ICCL.

A Additional Details

A.1 Benchmark Overview

We built TITLETRAP to study title-framing bias under controlled conditions. We generated 100 synthetic research-style abstracts (50 CV, 50 NLP) with GPT-4o-mini using prompts that encouraged a standard *problem–method–result* structure. A trained researcher manually screened all model outputs for plausibility, mentions of standard datasets (e.g., ImageNet, COCO, Cityscapes), and consistency of structure, discarding or editing drafts that failed quality checks.

Each abstract was paired with:

- Three stylistic titles: (A) branded/colon style, (B) plain descriptive style, (C) interrogative style.
- Two input modes: title-only and title+abstract.
- Sub-variants fixing either formatting style or coined term to isolate stylistic versus lexical effects.

A.2 Dataset Samples

Item 1 (CV)
`title_a`: *ImageFusion: Integrating Multi-Source Data for Enhanced Perception*
`title_b`: *ImageFusion for Enhanced Perception through Multi-Source Data Integration*
`title_c`: *Can ImageFusion Enhance Perception through Multi-Source Data Integration?*
Abstract: Introduces *ImageFusion*, a dual-stream framework fusing RGB, depth, and infrared for robust perception. On COCO, improves mean average precision by 4.5% over baselines and remains robust under adverse conditions.

Item 2 (CV)
`title_a`: *VisionNet: A Comprehensive Architecture for Visual Recognition*
`title_b`: *VisionNet as a Comprehensive Architecture for Visual Recognition*
`title_c`: *How Does VisionNet Function as a Comprehensive Architecture for Visual Recognition?*
Abstract: Presents *VisionNet*, integrating attention and residual connections. On ImageNet, achieves 3.2% top-1 accuracy gain over strong baselines, with robust transfer to other datasets.

A.3 LLM Reviewer Setup

We prompted GPT-4o and Claude with a standardized rubric (Clarity, Originality, Significance; 1–5 scale). Models rated all three titles for each abstract, selected the best one, and provided concise textual justifications. The prompts concealed the study's purpose to minimize priming effects. Single-run responses were collected due to computational constraints.

A.4 Prompt Templates

Generation Prompt (for synthetic benchmark):

> Generate 50 items of paper metadata in strict JSON array format. Each item must contain: - `id` (integer, starting at 1) - `field` ("CV") - `title_a`: Branding/colon-style title introducing a coined term or branded phrase (must use colon) - `title_b`: Plain descriptive academic title (must keep the same coined term but no colon) - `title_c`: Interrogative-style title phrased as a clear research question (must end with a question mark and keep the coined term) - `abstract`: A 180–220-word abstract in CVPR/ICCV/NeurIPS style, with background, method, experiments, contributions; mention at least one dataset; report at least one concrete performance result.
>
> Strict requirements: 1. All three titles describe the same paper. 2. Titles differ only in style, not in terminology. 3. The coined term must appear in all titles. 4. Abstract must be technically plausible and match the titles.

Evaluation Prompt (for LLM reviewer):

> You are serving as a peer reviewer for a major NLP conference. You will be given 3 titles (A, B, and C) for the same paper, along with its abstract. Evaluate them in the context of the abstract.
>
> Rate each title on: - Clarity (1–5) - Originality (1–5) - Significance (1–5)
>
> Choose the strongest overall title ("A", "B", or "C").
>
> Provide a JSON output: { "id": <int>, "round": "title+abstract", "scores": { "A": {"clarity": <int>, "originality": <int>, "significance": <int>}, "B": {...}, "C": {...} }, "choice": "A" | "B" | "C", "reasons": { "A": "2–3 sentences evaluating A", "B": "...", "C": "..." } }

A.5 Sample LLM Review Output

```
"id": 7,  "round": "title+abstract"
A: (5,4,5),   B: (4,2,3),   C: (3,3,3),   "choice": "A"
Reason A: Mentions FaceRecogNet, faithful and precise.
Reason B: Clear but generic, omits model name.
Reason C: Question framing feels less scholarly, misaligned with confident abstract.
```

A.6 Ethics and Data Release

All abstracts were synthetically generated and screened to remove personal or sensitive content. No real author names or affiliations were included. Following paper acceptance, we release:

- Benchmark data (100 abstracts × 3 titles × 2 modes)
- Prompt templates and code scripts
- Full JSON logs of LLM reviewer outputs

All data and code are released under a MIT license at https://github.com/ShuruiDu2002/titletrap-benchmark.

A.6 Reproducibility

We provide random seeds, YAML configuration files, description of the software environment, and analysis scripts for paired t-tests and visualization to facilitate reproducibility of our experiments.

A.7 Limitations and Broader Impact

Using synthetic abstracts allows controlled comparison but may not capture the full complexity of real submissions. Single-run LLM evaluations do not reflect stochastic variation. We encourage future work to combine human-written abstracts and study human–AI collaborative reviewing. The benchmark aims to reveal and help mitigate presentation bias in automated evaluation.