I CAN'T BELIEVE LLMS STILL CAN'T WRITE DRAMA: MULTI-DIMENSIONAL FAILURES IN SCRIPT CONTINUATION

Shijian Ma & Yan Lin
University of Macau
Macau SAR, China
{mc36473, yanlin}@um.edu.mo

Yunqi Huang
University College London
London, United Kingdom
yunqi.huang.23@ucl.ac.uk

ABSTRACT

Despite remarkable advances in creative text generation, we find that state-of-the-art large language models exhibit systematic and surprising failures in drama script continuation. Through DramaBench, a six-dimensional evaluation framework applied to 8,824 model-script evaluations across 8 frontier models, we uncover three "I Can't Believe It's Not Better" findings: (1) **No universal winner**: No model excels across all quality dimensions—GPT-5.2 leads in narrative efficiency but ranks 4th in emotional depth, while Qwen3-Max excels at emotion but ranks 6th in narrative; (2) **LLM-as-Judge fails for creative writing**: Human-LLM evaluator agreement is non-significant for 2 of 5 dimensions (Narrative Efficiency: $r = 0.07$, Character Consistency: $r = -0.04$), challenging the reliability of automated creative writing evaluation; (3) **Persistent logic failures**: Even the best models show 2–5% logic contradiction rates, with 17.6% of scripts containing at least one factual violation. These findings suggest that drama script continuation remains an unsolved challenge requiring targeted architectural or training innovations beyond scale.

1 INTRODUCTION

Large language models have achieved impressive results across diverse creative writing tasks, from poetry generation to story completion. Given these successes, one might expect that drama script continuation—continuing an existing screenplay while maintaining character voice, plot coherence, and dramatic structure—would be a solved problem for frontier models.

It is not.

We present findings from DramaBench, a comprehensive evaluation of 8 state-of-the-art LLMs on 1,103 drama scripts across six quality dimensions. Our results reveal systematic failures that persist across model families and scales:

- **Specialization without generalization**: Models develop distinct "skill profiles"—GPT-5.2 excels at maintaining logical consistency but produces emotionally flat continuations, while Qwen3-Max generates rich emotional arcs but introduces more plot contradictions.
- **Unreliable automated evaluation**: The commonly-used LLM-as-Judge paradigm shows no significant correlation with human judgments for narrative efficiency and character consistency, calling into question the validity of automated creative writing benchmarks.
- **Stubborn logic errors**: Despite near-perfect format compliance (0% error rate across all models), logic consistency shows the highest variance (2–5% error rates) with 17.6% of scripts containing violations—suggesting that surface-level competence masks deeper reasoning failures.

Table 1: Model rankings across six dimensions. No model ranks in top 3 for all dimensions. Bold indicates best; underline indicates worst.

Model	Narr.	Char.	Emot.	Logic	Confl.	Avg
GPT-5.2	**1st**	**1st**	4th	**1st**	2nd	**1.8**
Qwen3-Max	6th	2nd	**1st**	5th	7th	4.2
Gemini-3-Pro	2nd	7th	8th	4th	**1st**	4.4
Claude Opus 4.5	4th	5th	7th	3rd	5th	4.8
DeepSeek v3.2	7th	3rd	5th	2nd	6th	4.6

These negative results matter because they identify specific capability gaps that scaling alone has not addressed, and they challenge assumptions underlying current evaluation practices for creative AI systems.

2 Background: DramaBench Framework

DramaBench evaluates drama script continuation across six independent dimensions using an "LLM Labeling + Statistical Analysis" methodology. Rather than asking LLMs to directly score quality (which introduces known biases (Zheng et al., 2023)), we use LLMs to extract categorical labels that are converted to objective metrics.

Dimensions: (1) *Format Standards*: Rule-based Fountain format compliance; (2) *Narrative Efficiency*: Ratio of plot-advancing beats to total beats; (3) *Character Consistency*: Out-of-character (OOC) dialogue rate; (4) *Emotional Depth*: Presence of emotional arcs and complex emotions; (5) *Logic Consistency*: Factual contradiction rate with context; (6) *Conflict Handling*: Whether dramatic conflicts escalate, twist, or are dropped.

Scale: 1,103 scripts $\times$ 8 models = 8,824 evaluations, with 252 statistical significance tests and human validation on 188 scripts.

3 Related Work

Story and Drama Generation. Early benchmarks like ROCStories (Mostafazadeh et al., 2016) evaluate story understanding via cloze tests but lack screenplay-specific requirements. Drama generation systems include Dramatron (Mirowski et al., 2023), which uses hierarchical LLM prompting, and IBSEN (Zhou et al., 2024), a director-actor agent framework. However, these are generation *systems* rather than *evaluation benchmarks*—they lack standardized test sets or quantitative model comparisons.

LLM-as-Judge Evaluation. The LLM-as-Judge paradigm (Zheng et al., 2023) has become popular for scalable evaluation, but recent surveys (Gu et al., 2024) identify systematic biases including position bias, verbosity preference, and self-enhancement. Our finding that human-LLM agreement fails for 2/5 creative dimensions adds domain-specific evidence to these concerns.

Multi-Dimensional NLG Evaluation. UniEval (Zhong et al., 2022) frames evaluation as boolean QA across coherence, consistency, and fluency, but does not address screenplay-specific dimensions like character voice or conflict escalation. DramaBench is the first to combine drama-specific multi-dimensional metrics with rigorous human validation.

4 Negative Result #1: No Model Excels Universally

Table 1 reveals a striking pattern: **excellence in one dimension often comes at the cost of another**. GPT-5.2 ranks 1st in Narrative, Character, and Logic, but only 4th in Emotional Depth—its continuations are coherent but emotionally "safe." Gemini-3-Pro is best at Conflict Handling but *worst* at Emotional Depth (8th)—it escalates dramatic tension mechanically without nuanced emotional payoff. Qwen3-Max leads Emotional Depth but ranks 6th and 7th in Narrative and Conflict—rich emotions come with meandering plots.

Table 2: Human-LLM evaluator agreement (188 scripts). Two dimensions show no significant correlation.

Dimension	Metric	Agreement
Logic Consistency	Pearson r	0.48***
Emotional Depth	Cohen's κ	0.53 (substantial)
Conflict Handling	Cohen's κ	0.42 (moderate)
Narrative Efficiency	Pearson r	0.07 (n.s.)
Character Consistency	Pearson r	-0.04 (n.s.)

This suggests that current training approaches create trade-offs between quality dimensions rather than improving them jointly. A user seeking "the best model for screenplay writing" will be disappointed—the answer depends entirely on which quality dimension matters most.

5 Negative Result #2: LLM-as-Judge Fails for Creative Writing

We validated our LLM evaluator (Qwen3-Max) against human expert annotations. Table 2 reveals a concerning finding: **2 of 5 content dimensions show no significant human-LLM agreement**.

Why does this matter? The LLM-as-Judge paradigm is increasingly used to evaluate creative writing at scale (Zheng et al., 2023). Our results suggest this approach may produce rankings that do not reflect human quality judgments for Narrative Efficiency ($r = 0.07$) and Character Consistency ($r = -0.04$). LLMs and humans disagree on what constitutes a "plot-advancing" beat, and the slight *negative* correlation for character consistency suggests LLMs may use different criteria entirely—what appears as "out-of-character" to an LLM might be "character development" to a human (see Appendix for case study).

This finding challenges the validity of any creative writing benchmark relying solely on LLM evaluation for these dimensions.

6 Negative Result #3: Persistent Logic Failures

Perhaps the most surprising finding is the persistent gap between *surface competence* and *deep reasoning*: All 8 models achieve 0% format error rate on Fountain screenplay format, yet logic consistency error rates range from 2.0% (GPT-5.2) to 5.3% (GLM-4.6), with 17.6% of scripts containing at least one factual violation.

Example failure (MiniMax M2): Context establishes protagonist is in surgery. Continuation: "*Ranran's eyes SNAP open in her bedroom—normal, undamaged.*" The model teleports the character and erases established injuries.

This pattern—perfect format, imperfect logic—suggests that models learn *syntactic* screenplay conventions without fully grounding continuations in the *semantic* constraints established by context. The 2–5% error rate has remained stable across model generations, indicating this is not simply a scale problem. We provide a detailed error taxonomy in the Appendix.

7 Discussion

Our negative results point to three hypotheses: **(1) Training data bias**: Screenplay format is explicitly marked, making it easy to learn, while implicit logical constraints require inference across distributed context. **(2) Evaluation gaming**: If models are trained with LLM-as-Judge feedback, they may optimize for dimensions where LLM evaluation is reliable while dimensions with poor human-LLM agreement remain undertrained. **(3) Attention limits**: Maintaining character consistency requires tracking persona information across long contexts, which may reflect attention failures.

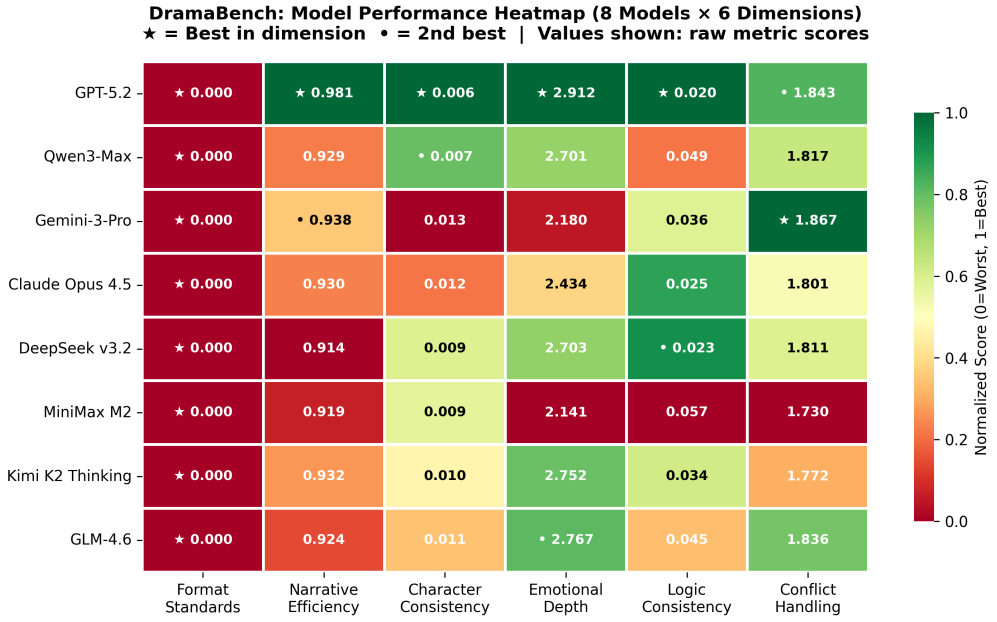

Figure 1: Model performance heatmap. Logic Consistency (rightmost column) shows the highest cross-model variance despite all models achieving 0% format error.

8 Limitations

Our study has several limitations. First, we use a single LLM evaluator (Qwen3-Max); multi-evaluator ensembles might improve reliability. Second, DramaBench contains only English scripts—cross-lingual generalization is untested. Third, we evaluate *continuation* rather than full-script generation. Finally, our human validation covered 17% of the dataset; larger-scale annotation would strengthen claims.

9 Conclusion

We present three "I Can't Believe It's Not Better" findings for LLM drama script continuation: (1) no model excels universally, (2) LLM-as-Judge evaluation fails for key creative dimensions, and (3) logic failures persist despite format mastery. These results suggest that drama script continuation—and perhaps creative writing more broadly—requires targeted innovations beyond scaling. We hope these negative results guide future research toward the specific capability gaps identified here.

LLM Usage Disclosure

This work evaluates eight large language models as the primary subjects of our benchmark. Additionally, we employ Qwen3-Max as an automated evaluator to extract categorical labels from model-generated continuations (e.g., identifying out-of-character dialogue, plot-advancing beats, and factual contradictions). All LLM API calls were made in December 2025 through OpenRouter. LLMs were used to assist with writing optimization and language polishing of this paper.

Ethics Statement

This research evaluates publicly available large language models on creative writing tasks using drama scripts. All evaluated models are accessible through official APIs or open-source releases. The drama scripts in DramaBench were collected from publicly available sources with appropriate usage rights. Our evaluation does not involve human subjects beyond the volunteer annotators who provided validation labels (188 scripts), and all annotators participated with informed consent. We acknowledge that our benchmark focuses exclusively on English-language drama, which may not

generalize to other languages or cultural contexts. The findings presented here aim to identify capability gaps in current LLMs to guide future research, not to make definitive claims about the creative potential of AI systems.

Reproducibility Statement

To facilitate reproducibility, we have open-sourced both our dataset and evaluation code. The DramaBench dataset, containing 1,103 drama scripts with context-continuation pairs, is available at `https://huggingface.co/datasets/FutureMa/DramaBench`. Our evaluation framework, including the LLM-based labeling pipeline and statistical analysis scripts, is available at `https://github.com/IIIIQIIII/DramaBench`. The repository includes detailed documentation of model API configurations, prompt templates, and hyperparameters used for all experiments.

References

Jiawei Gu, Xuhui Xu, Yuxuan Zhu, et al. A survey on LLM-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

Piotr Mirowski, Kory W Mathewson, Jaylen Pittman, and Richard Evans. Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals. *arXiv preprint arXiv:2209.14958*, 2023.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of NAACL-HLT*, pp. 839–849, 2016.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of EMNLP*, 2022.

Senyu Zhou et al. IBSEN: Director-actor agent collaboration for controllable and interactive drama script generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

A Appendix: Case Study - Character Consistency

To illustrate the human-LLM disagreement on character consistency, consider script_3350 (GPT-5.2). The context establishes a stoic military commander who never shows emotion. In the continuation:

> *COMMANDER CHEN wipes a tear from his eye. "I never told anyone... my son died in that battle."*

The LLM evaluator labeled this as **OOC** (out-of-character)—a stoic character crying violates the established persona. Human annotators disagreed, viewing this as **character development**—a breakthrough moment revealing hidden depth. This fundamental disagreement about what constitutes "consistency" vs "growth" explains the near-zero correlation.

B Appendix: Error Taxonomy

We classified 10,850 errors across all 8,824 evaluations. The most common failure modes are:

- **Dialogue-Action Imbalance** (1,354 errors, 15.3%): Continuations contain excessive dialogue without corresponding action beats, or vice versa.
- **Low Information Gain** (1,305 errors, 14.8%): Beats that neither advance plot nor develop character—pure padding.
- **Redundant Beats** (1,211 errors, 11.2%): Repetition of information already established in context.
- **Spatial/Temporal Contradictions** (892 errors, 8.2%): Characters appearing in wrong locations or events occurring in impossible sequences.

Model-specific patterns: GPT-5.2 shows minimal errors across categories. Qwen3-Max exhibits dialogue imbalance (243 occurrences) despite emotional strength. GLM-4.6 produces excessive prose (102 occurrences) and the highest logic violations.