SCREENWRITINGBENCH: EVALUATING LLMS ACROSS THE SCREENWRITING WORKFLOW WITH DIMENSION-SPECIFIC EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Professional screenwriting develops through planning, drafting, and revision across several artifact types. Existing writing benchmarks sample only parts of this workflow, making it difficult to assess whether large language models (LLMs) can support the successive tasks of screenwriting. Dramatic quality also lacks an objective ground truth and calls for a different form of judgment from task completion or artifact compliance. To address these gaps, we introduce ScreenWritingBench, a collection of 279 tasks constructed and validated by professional screenwriters. The tasks span artifacts from proposals to scripts and operations from initial generation to revision; 219 include dialogue history. We evaluate responses along five dimensions: Task Completion, Contextual Consistency, Artifact Compliance, Basic Lexical Defects, and Dramatic Quality. The first four use rubrics, whereas Dramatic Quality places each response relative to eight expert-ordered anchors. An evaluation of 18 models reveals substantial variation across dimensions. We report both an overall leaderboard and dimension-level scores to characterize these differences.

1 INTRODUCTION

Large language models (LLMs) have become increasingly capable and are now used for script cowriting, novel-to-script adaptation, and staged script generation (Mirowski et al., 2023; Lin et al., 2025; Lei et al., 2025). Real-world screenwriting, however, involves a complex production workflow spanning the full creative process, including story development, script creation, and revision. Creative writing is inherently subjective and lacks definitive ground-truth references (Fein et al., 2025). Evaluation therefore requires both clearly defined dimensions and judging procedures suited to their different requirements. An overall score alone cannot explain performance across these dimensions.

Existing benchmarks have two major gaps. First, evaluating professional screenwriting requires coverage of tasks across the full creative workflow, but existing benchmarks cover only parts of this process. General writing benchmarks aim to cover diverse domains, including educational and academic writing (Wu et al., 2025; Feng et al., 2026). Creative-writing benchmarks focus on story or fiction generation (Fein et al., 2025). Screenwriting-specific benchmarks, however, define a narrower range of tasks. For example, DramaBench evaluates the continuation of a partially provided script, whereas CML-Bench evaluates the generation of script segments from summaries (Ma et al., 2025; Zheng et al., 2025).

Second, evaluation criteria need to be organized and adapted to diagnose the distinct requirements encountered across the screenwriting workflow. WritingBench generates criteria for individual writing queries (Wu et al., 2025), while DramaBench evaluates script continuation along six dimensions (Ma et al., 2025). These settings do not by themselves establish how to distinguish and judge the requirements of tasks across the workflow. For example, a response may satisfy the user's current request yet contradict information established in the dialogue history. Task completion alone would not capture this inconsistency. Scripts must also follow professional conventions and avoid basic language defects, while a well-formatted, readable script can remain dramatically weak (Ma et al., 2025). These requirements call for separate judgments: explicit constraints can be checked against rubrics, whereas dramatic quality requires assessment of open-ended creative outcomes.

To address the workflow gap, we introduce **ScreenWritingBench**, a collection of 279 tasks constructed and validated by professional screenwriters. The tasks cover creation and revision at stages

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

throughout the workflow, from proposals and story synopses through character profiles, character biographies, and scene breakdowns to scripts. Each task evaluates a response to a specific request; 219 tasks include dialogue history to retain the context needed for writing or revision.

Accordingly, we propose a screenwriting evaluation framework that distinguishes among task fulfillment, contextual consistency, adherence to professional conventions, basic linguistic correctness, and dramatic effectiveness. We assess these aspects through five complementary dimensions, each with an appropriate evaluation procedure. **Task Completion** asks whether the current request was fulfilled; **Contextual Consistency** checks agreement with prior context and internal coherence; **Artifact Compliance** checks whether the output follows the professional conventions of the requested artifact type; and **Basic Lexical Defects** identifies observable surface errors. Evaluation across these four dimensions is guided by explicit rubrics. **Dramatic Quality** assesses dramatic effectiveness through comparison with eight expert-ordered anchors, because dramatic effectiveness cannot be fully captured by a checklist. To accommodate higher-quality future responses, the anchor set can be extended with candidate anchors generated through refinement and screened independently by two judges. In the extension experiment, adding a higher-quality ninth anchor left the relative ranking of all eight models unchanged, providing evidence for the extensibility of the anchor-based evaluation approach. By combining five complementary dimensions with corresponding evaluation procedures, our proposed framework provides fine-grained scoring for each screenwriting task.

Evaluating 18 models reveals why these dimensions need to be reported alongside the overall leaderboard. Models with high Contextual Consistency scores can still fall short on Task Completion, and strong performance on Basic Lexical Defects can coexist with low Dramatic Quality scores. These patterns expose differences in model capabilities that an overall score alone would obscure.

Our contributions are threefold:

- We introduce ScreenWritingBench, built and validated by professional screenwriters, covering creation and revision across workflow stages in single-turn and multi-turn settings.
- We organize screenwriting evaluation into five complementary dimensions and pair each with an appropriate judging procedure, enabling fine-grained scoring for each task. For Dramatic Quality, multiple anchors are ranked by experts according to quality, and higher-quality anchors can be added to extend the reference range.
- By evaluating 18 models, we reveal differences in task fulfillment, consistency, language validity, and dramatic quality, demonstrating the value of dimension scores.

2 Related Work

2.1 Benchmarks for Writing and Screenplay Generation

General writing benchmarks cover diverse products and scenarios. WritingBench spans six domains and constructs criteria for query-specific requirements; EssayBench targets four Chinese essay genres; and HoWToBench extends Chinese writing evaluation to twelve genres (Wu et al., 2025; Gao et al., 2025; Feng et al., 2026). Creative writing has received particular attention because of its open-ended and subjective nature. Wordcraft supports interactive story writing by letting writers request and revise model suggestions (Yuan et al., 2022). TTCW complements preference-based evaluation with expert assessments of fluency, flexibility, originality, and elaboration (Chakrabarty et al., 2024). CreataSet compares responses across open-domain creative tasks, whereas LitBench models human preferences over pairs of short stories (Cao et al., 2025; Fein et al., 2025).

Screenplay generation, as a professional branch of creative writing, has also motivated specialized systems and benchmarks. Dramatron hierarchically develops a logline into characters, story beats, locations, and dialogue, while R^2 adapts novels into screenplays through causal plot graphs; both primarily evaluate particular generation systems rather than provide standardized benchmarks of general screenwriting capability (Mirowski et al., 2023; Lin et al., 2025). More recently, DramaBench evaluates drama-script continuation along six dimensions, whereas STAGE tests character-state tracking, cross-scene reasoning, and in-script role-playing over full screenplays (Ma et al., 2025; Tian et al., 2026). Nevertheless, these benchmarks remain centered on a single artifact type or task family. ScreenWritingBench instead evaluates multi-turn instruction–response pairs covering the whole screenplay generation pipeline including diverse artifacts generation, revision and evaluation.

2.2 Evaluation of Creative Writing

LLM-as-judge for creative writing presents unique challenges due to its inherent subjectivity and lack of ground truth (Fein et al., 2025; Cao et al., 2025). Existing methods can be broadly grouped into three categories (Zheng et al., 2023). **Pointwise scoring** evaluates each response independently using fixed, category-specific, or query-dependent criteria. Evaluators such as Auto-J and Prometheus make this approach scalable, while WritingBench generates criteria for individual writing requests (Li et al., 2024; Kim et al., 2024a; Wu et al., 2025). G-Eval structures assessment through generated evaluation steps, and Prometheus 2 supports both direct scoring and pairwise ranking with user-defined criteria (Liu et al., 2023; Kim et al., 2024b). FLASK decomposes response quality into instruction-specific skill sets, making individual strengths and weaknesses more interpretable (Ye et al., 2024). IFEval focuses on objectively verifiable instructions, while FollowBench incrementally adds content, situation, style, format, and example constraints (Zhou et al., 2023; Jiang et al., 2024). These approaches help distinguish adherence to explicit requirements from broader response quality. However, a single scoring procedure may conflate instruction fulfillment, formal correctness, and creative quality. **Pairwise evaluation** compares two responses and aggregates preferences into rankings, as adopted by LitBench and CreataSet (Fein et al., 2025; Cao et al., 2025). Comparative NLG assessment also studies how relative judgments can be converted into rankings and how presentation order affects those judgments (Liusie et al., 2024). Although effective for subjective quality, its results depend on the compared model pool. **Reference-guided and listwise evaluation** positions responses relative to fixed examples; LINKAGE, for instance, ranks an answer within an ordered list of varied-quality references (Yang et al., 2024). ScreenWritingBench combines these paradigms by evaluation level: rubric-based judgments diagnose task fulfillment and professional validity, while reference-guided listwise ranking assesses holistic dramatic quality against examples spanning quality levels for the same task.

3 ScreenWritingBench

As noted above, existing writing benchmarks capture only parts of the screenwriting workflow, lacking systematic coverage of the full spectrum of screenwriting artifacts—from proposals and story synopses through character profiles, character biographies, and scene breakdowns to scripts. This gap leaves existing benchmarks unable to comprehensively assess model capabilities in professional screenwriting scenarios. Motivated by this observation, the data construction of ScreenWritingBench adopts workflow-wide coverage as its core principle, requiring that tasks encompass every stage of the screenwriting process. Building upon this full-coverage foundation, we design the following four-stage data construction pipeline.

3.1 Data Construction

Figure 1 shows four stages of task construction: Task Simulation, Filtering and Expert Validation, Construct and Validate Task Dimensions, and Rubric and Anchor Generation.

Task Simulation. The initial dialogues are generated by simulating screenwriting scenarios. Senior screenwriters freely construct tasks within the full spectrum of screenwriting artifacts and ten predefined task axes. This process keeps each task within the designated categories while yielding concrete creative goals, drafts, and revisions, ensuring diversity.

Filtering and Expert Validation. Simulated dialogues cannot be used as benchmark tasks without screening: some have incomplete context, an unclear user intent, missing turns, repeated content, incoherent logic, or truncated responses. An LLM-based pass checks each candidate for five properties: completeness, explicit intent, repeated content within the candidate, coherence, and truncation. Professional screenwriters then validate the filtered tasks. This process yields the final set of 279 tasks.

Construct and Validate Task Dimensions. For the validated tasks, Step 3 of Figure 1 constructs second-order task dimensions through bottom-up clustering under the ten predefined task axes. Professional screenwriters then validate the dimension labels. The resulting system defines the benchmark's scope and categories and supplies labels for generating category-aware rubrics in the next stage.

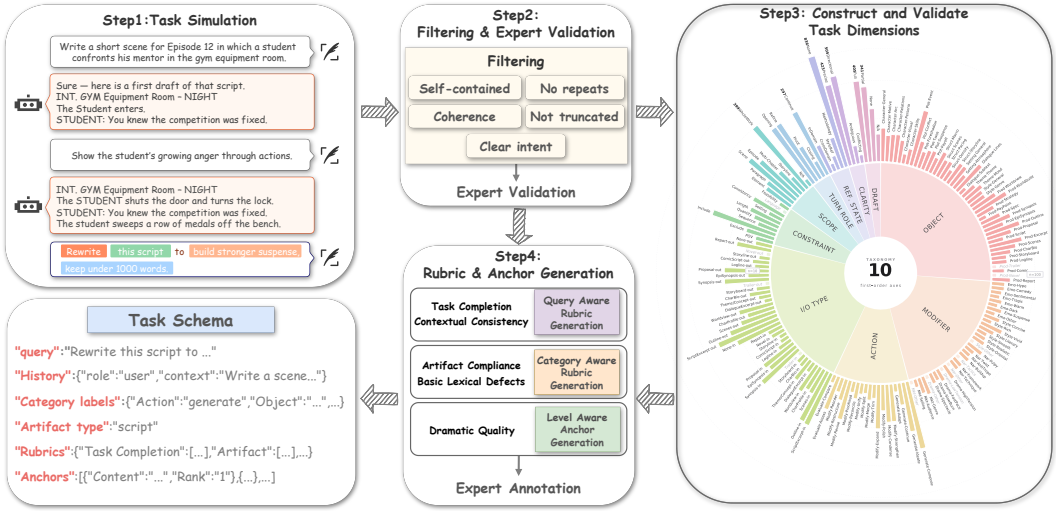


Figure 1: Construction of a benchmark task: screenwriters simulate a writing process (Step 1); candidate tasks are filtered and validated by experts (Step 2); second-order task dimensions are clustered from the validated tasks and their labels are verified by experts (Step 3); rubrics and anchors are generated (Step 4). Each accepted task is stored as a complete record.

Rubric and Anchor Generation. For each validated task, the evaluation materials for all five dimensions are generated at this stage (Step 4 of Figure 1). For Task Completion and Contextual Consistency, we generate query-aware rubrics tailored to the task's dialogue history and current query, which are then sampled and checked by professional screenwriters. For Artifact Compliance and Basic Lexical Defects, professional screenwriters collaborate with LLMs to generate category-aware rubrics that generalize within the same artifact or defect category. For Dramatic Quality, we sample eight candidate anchors (A–H) per task and have experts rank them by quality. Section 4 details how these materials are used in evaluation.

The Complete Task Definition. A complete task record contains the final query, preceding dialogue history if present, labels for its task dimensions and artifact type, the applicable rubrics, and eight ranked anchors. The artifact-type label determines which compliance criteria apply (Section 4).

3.2 Task Diversity

The tasks span the creative workflow from proposals and synopses through character profiles, character biographies, and scene breakdowns to scripts and their revision. Their diversity is not limited to artifact names: ten task dimensions and 172 axis-specific tags describe writing actions, the objects and scope of those actions, input and output forms, constraints, and interaction settings. These properties can combine within a task, such as revising a scene breakdown in light of an earlier draft. The suite also includes 219 tasks with dialogue history and 60 without it. History-conditioned tasks capture iterative work in which a current request builds on earlier exchanges. Appendix B summarizes the dimensions and representative tags in Table 9.

The mean input length is 6,670 tokens. Distributions of input length in characters and dialogue turns appear in Appendix B (Figures 6 and 7).

4 Evaluation Methodology

A screenwriting response may perform the requested operation yet contradict an established character motivation. It may also meet formal requirements while remaining dramatically weak. This observation reveals that the quality of a screenwriting response is inherently multidimensional. Starting from the practical needs of screenwriters in multi-turn dialogue, we decompose screenwriting quality into five complementary aspects: whether the requested operation is fulfilled, whether the response remains consistent with prior context, whether it conforms to artifact conventions, whether it is free of surface-

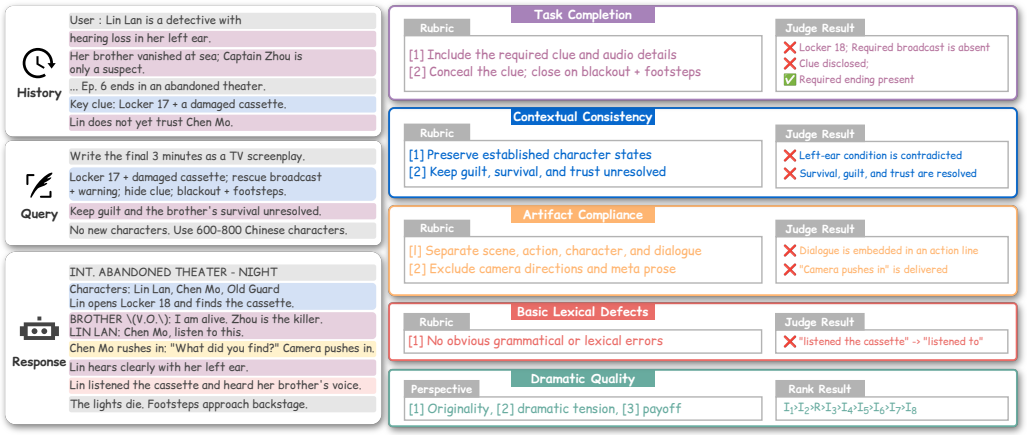


Figure 2: Worked example of the five-dimensional evaluation.

level linguistic defects, and whether it possesses dramatic merit. These five aspects are qualitatively heterogeneous—they measure distinct facets of writing that cannot substitute for one another. Furthermore, they differ in scope: the first two are meaningful only within an individual task, the next two constitute norms that generalize across task categories, and the last transcends specific tasks and formats to assess dramatic creation in the abstract. This progression from task-specific to normative to abstractly qualitative naturally yields a three-layer structure.

Building upon this three-layer structure, we adopt differentiated evaluation strategies tailored to the nature of each layer. For the Task Layer (TC, CC), the evaluation content is strongly tied to the context and the user query; therefore, we employ query-aware rubrics, extracting an independent rubric list for each task. For the Norm Layer (AC, BLD), where conventions exhibit a degree of universality, we design category-aware rubrics that generalize across different artifact types and common surface-level issues. Finally, for the Quality Layer (DQ), we argue that a checklist-based approach cannot adequately capture dramatic quality, while direct LLM-as-judge scoring suffers from instability. We therefore generate anchors at different quality levels for each task and adopt a comparative approach across multiple dimensions to derive the dramatic quality of the response under evaluation. For a task, let H denote its dialogue history, Q the current user query, and R the model response.

Figure 2 follows a deliberately flawed screenplay response through the five dimensions. The example shows how each judgment traces back to different evidence.

4.1 Evaluation Framework

4.1.1 Task Layer

Task Completion. TC evaluates whether R delivers what Q requests. Given H and Q, we generate query-aware rubrics. The judge checks whether R performs mandatory writing operations, includes indispensable plot points, and follows explicit constraints. Starting from 100, TC deducts ten points for each unmet requirement, five for each redundant addition, and ten for each deviation from the writing goal, with a minimum score of zero.

Contextual Consistency. CC evaluates whether R is consistent with H and Q and internally coherent. Its rubric is derived from the dialogue history and current query. Each violated rubric item incurs a ten-point deduction. We also check R for internal contradictions and add the corresponding deductions. The deductions are summed to produce the CC score, with a minimum of zero.

4.1.2 NORM LAYER

Artifact Compliance. AC evaluates whether R follows the professional conventions of the artifact type associated with the task. That type selects the applicable rubric: screenplays, outlines, scene breakdowns, and episode outlines have distinct content and format criteria. Each rubric item is assigned a violation level from 0 to 3 based on the frequency and proportion of violations. A level of 2 or higher crosses the red line; higher levels incur larger deductions, which accumulate across violated items. Let D_{str} and D_{cnt} denote the total format and content deductions.

The AC score subtracts their sum from 100, with a minimum of zero: $S_{\text{AC}} = \max(0, 100 - D_{\text{str}} - D_{\text{cnt}})$. Figure 3 illustrates the evaluation process for a screenplay.

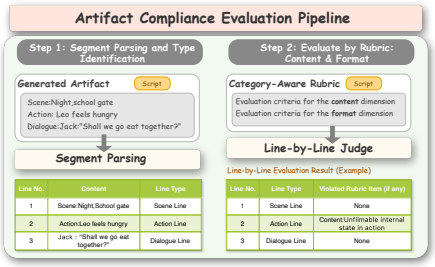


Figure 3: Artifact Compliance evaluation (script example).

Basic Lexical Defects. BLD checks R for observable surface defects: garbled or anomalous characters, verbatim repetition, inappropriate mixing of writing systems, typographical or grammatical errors, and truncated generation. The current request and history help distinguish defects from intentional quotations or inherited wording; formatting conventions remain under AC. The score is binary: 100 if no defect is detected, and 0 if any defect is found or the response is empty. BLD measures the response's surface-level validity.

4.1.3 QUALITY LAYER

Dramatic Quality. DQ evaluates the dramatic quality of model responses. Because this subjective dimension cannot be verified against an explicit rubric checklist, we instead adopt a comparative approach: each task is equipped with anchor responses, and an LLM judge ranks the model response R among them, considering character design, dramatic tension, relationships, and originality. The DQ score depends on R's position in this ordering: the more anchors it surpasses, the higher its score (Appendix C). To test whether multiple anchors distinguish responses more finely than one anchor or no-anchor scoring, we compare these conditions in Appendix C. On the tested sample, eight anchors spread responses across nine score levels, while single-anchor scoring produces many ties and both no-anchor conditions concentrate scores near the top of their scales. This shows greater score discrimination in this comparison, although the no-anchor conditions also use different criteria or scales. Stronger anchors can be added when the observed scores approach the top of the scale to preserve room for distinguishing stronger responses (Appendix D).The DQ score is $S_{\text{DQ}} = 100 \cdot N_{\text{below}}/8$, where N_{below} counts the anchors ranked strictly below R.

4.2 SCORE AGGREGATION

The five dimensions each receive a weight of 0.2. The overall score is their weighted sum:

$$S_{\text{overall}} = 0.2\, S_{\text{TC}} + 0.2\, S_{\text{CC}} + 0.2\, S_{\text{AC}} + 0.2\, S_{\text{BLD}} + 0.2\, S_{\text{DQ}}, \tag{1}$$

5 EXPERIMENTS

5.1 EVALUATION SETUP

Models. We evaluate a diverse suite of 18 representative LLMs, encompassing leading proprietary systems as well as a select number of open-source alternatives. Our selection features models from the GPT series (GPT-5.1-2025-11-13, GPT-5.2-2025-12-11, GPT-5.4-2026-03-05, GPT-5.5-2026-04-24, GPT-5.6-sol), Claude series (Claude-Opus-4.6, Claude-Opus-5), Qwen series (Qwen3.5-122B-A10B, Qwen3.6-plus, Qwen3.8-27B, Qwen3.8-max-2026-09-02), GLM series (GLM-5.1, GLM-5.2, GLM-5.3), and Kimi series (Kimi-K2.6, Kimi-K3), alongside other leading models like Gemini-3.1-pro-preview and DeepSeek-V4-pro-2026-04-24. Table 1 uses shortened run identifiers for the dated models.

Evaluation protocol. For Completion and Contextual Consistency, we employ DeepSeek-V4-flash as the evaluation model. For Artifact Compliance and Basic Lexical Defects, we use Qwen3.7-plus.

For Dramatic Quality, we use DeepSeek-V4-pro-2026-04-24. For each model response, we calculate the five dimension scores and their equal-weight overall score using Equation 1. A model's reported score is the mean of its per-task scores.

5.2 Overall Results

Table 1: Main leaderboard on the benchmark task set. Each dimension column is the mean per-task score; Artifact Compliance (AC) combines its format and content components. Overall is the equal-weight mean of the five dimensions. Shaded bold cells mark the top three positions in each score column, including ties.

Rank	Model	TC	CC	AC	BLD	DQ	Overall
1	Claude-Opus-5	**79.66**	90.00	78.18	87.50	**90.91**	**85.25**
2	GPT-5.6-sol	**77.95**	93.58	78.38	**92.05**	**83.66**	**85.13**
3	Qwen3.8-max	**79.00**	**94.75**	76.32	85.00	**73.32**	**81.68**
4	Kimi-K3	72.61	82.56	79.38	**92.05**	69.18	79.15
5	GLM-5.3	53.10	**95.54**	**80.36**	86.90	67.41	76.66
6	GPT-5.2	62.90	93.07	77.10	86.26	61.21	76.11
7	GPT-5.5	61.93	88.92	74.74	89.77	65.06	76.09
8	GPT-5.4	57.18	93.82	75.35	87.40	62.79	75.31
9	Claude-Opus-4.6	60.34	84.66	77.10	85.23	66.48	74.76
10	Qwen3.6-plus	68.40	88.26	79.65	86.26	45.18	73.55
11	GLM-5.1	68.41	87.73	79.38	89.77	39.35	72.93
12	GPT-5.1	59.55	91.65	75.40	73.86	59.38	71.97
13	DeepSeek-V4-pro	56.83	81.79	75.67	82.82	62.64	71.95
14	GLM-5.2	72.58	**93.96**	**79.73**	81.25	31.74	71.85
15	Kimi-K2.6	56.25	80.51	78.13	71.59	63.35	69.97
16	Gemini-3.1-pro-preview	62.16	83.01	79.49	**96.59**	25.57	69.36
17	Qwen3.8-27B	48.31	87.54	79.14	89.62	8.70	62.66
18	Qwen3.5-122B-A10B	42.06	82.37	**80.45**	91.98	14.69	62.31

Table 1 reports the overall leaderboard and the mean score for each of the five evaluation dimensions.

5.3 Capability Analysis

Table 1 reveals that no single model dominates across all five evaluation dimensions; instead, different models exhibit distinct capability profiles, confirming that screenwriting competence is not a monolithic skill but a composite of heterogeneous abilities.

In Task Completion (TC), Claude-Opus-5 achieves the highest score (79.66), demonstrating superior instruction-following and content-delivery capabilities. At the other extreme, Qwen3.5-122B-A10B scores only 42.06—the lowest in the benchmark. This divergence indicates that more capable models possess stronger abilities in understanding both long-term and short-term user needs across multi-turn dialogues and in adhering to those needs in their responses.

For Contextual Consistency (CC), GLM-5.3 leads (95.54), followed closely by Qwen3.8-max (94.75), with GLM-5.2 in third at 93.96. The overall score distribution is relatively narrow, yet both GLM-5.3 and GLM-5.2 maintain high scores among the top-tier models, indicating that the GLM family exhibits strong capability in preserving multi-turn dialogue settings while avoiding significant self-contradictions in its outputs.

Artifact Compliance (AC) reveals a notable phenomenon: scores range narrowly between 74.74 and 80.45, with no substantial gap among models. This is because no model strictly adheres to artifact conventions. Notably, Qwen3.5-122B-A10B, which ranks last overall (62.31), produces the most format-compliant artifacts (80.45). This suggests that adherence to screenplay formatting conventions—scene headings, action descriptions—can be achieved through surface-level pattern matching and does not require deep narrative competence.

In Basic Lexical Defects (BLD), Gemini-3.1-pro-preview achieves the highest score (96.59), demonstrating robust avoidance of garbled text, verbatim repetition, and inappropriate language mixing. This dimension reflects the stability of a model's output during conversation.



Figure 4: Mean overall scores by task action category. Comparing a model across categories shows where its performance rises or falls during different screenwriting actions.

Dramatic Quality (DQ) exhibits the largest score spread among all five dimensions, with Claude-Opus-5 substantially outperforming all others (90.91), followed by 83.66 and 73.32 in second and third place. This dimension—encompassing character design, dramatic tension, conflict and relationships, and creativity—is the most discriminating dimension. Claude-Opus-5 leads in both TC and DQ, which explains its first-place overall ranking (85.25).

Several cross-dimension trade-offs emerge from the leaderboard. The TC–AC divergence in Qwen3.5-122B-A10B (42.06 vs. 80.45) reveals a "form over substance" pattern, where the model produces reasonably formatted but substantively inadequate outputs. The CC–TC gap in GLM-5.3 (95.54 vs. 53.10) indicates that a model may generate fluent, contextually coherent text while failing to genuinely fulfill the user's explicitly requested tasks. The BLD–DQ gap in Gemini-3.1-pro-preview (96.59 vs. 25.57) further confirms that freedom from surface-level defects does not equate to dramatic quality. These trade-offs underscore the necessity of multi-dimensional evaluation: collapsing heterogeneous judgments into a single score would obscure practically important differences across model capabilities.

5.4 Task-Category Analysis

The task taxonomy in Section 3.2 provides action and artifact-type categories. For each category, we average a model's per-task overall scores over tasks with that label. Figures 4 and 5 show how performance varies across the two category axes.

The task taxonomy introduced in Section 3.2 organizes tasks along two complementary axes: action type (what the model is asked to do) and artifact type (what kind of written artifact is expected). Figures 4 and 5 decompose model performance along these two axes respectively, revealing that aggregate leaderboard scores can mask substantial within-model variation across categories.

Action-Type Analysis (Figure 4). Figure 4 reports mean overall scores for three action categories—Generate, Modify, and Other—across all evaluated models. Generate (producing original screenwriting content from scratch) and Modify (revising or rewriting existing material) exhibit correlation with each other, indicating that these two task types share certain commonalities in their demands on the model. Meanwhile, the Modify floor (57.2) is lower than the Generate floor (61.7), indicating that revision-oriented tasks are more challenging than generation for lower-ranked models. This is consistent with the intuition that effective modification requires not only writing ability but also faithful comprehension of source material and targeted application of changes. For instance, GPT-5.4 scores 74.26 on Generate tasks but drops to 68.41 on Modify tasks, a 5.85-point decline that highlights the additional cognitive load imposed by revision. Top-tier models (e.g., Claude-Opus-5, GPT-5.6-sol) maintain relatively flat profiles across Generate and Modify, suggesting that their advantage lies not only in raw generation quality but also in the ability to preserve and adapt existing content.

Artifact-Type Analysis (Figure 5). Figure 5 reports performance along five artifact categories (Planning artifacts, Story outlines, Character profile / biography, Scene planning, and Scripts) plus an Other category. Overall, the score trends across artifact types exhibit a certain degree of correlation with overall scores. Character-related tasks show the widest score range (85.5 to 51.8), identifying character design and development as the most discriminating artifact type. Writing compelling characters demands an understanding of motivation and interpersonal dynamics that goes beyond structural or formatting competence. The 33.7-point spread indicates that this category effectively separates models capable of nuanced character portrayal from those producing only generic, flat characterizations.

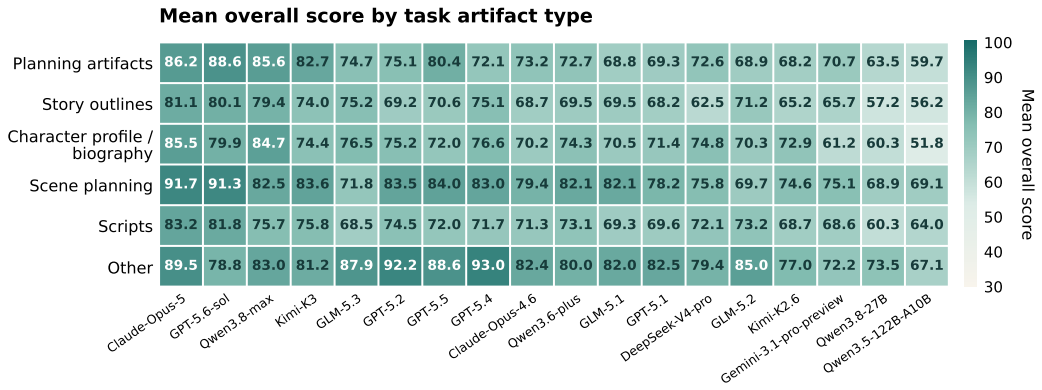


Figure 5: Mean overall scores by task artifact type. Each cell summarizes a model's performance on tasks associated with that artifact category.

Script tasks, which require full screenplay-format writing, range from 83.2 to 60.3, exhibiting the narrowest score range among all artifact types, indicating that most models can not produce reasonably adequate screenplays.

5.5 Agreement with Expert Annotations

Table 2: Agreement between evaluation results and expert screenwriters' annotations for each dimension and overall.

Dimension	Agreement	Dimension	Agreement
Task Completion	81.48%	Basic Lexical Defects	87.33%
Contextual Consistency	58.33%	Dramatic Quality	58.89%
Artifact Compliance	96.5%	Overall	64%

Table 2 shows higher agreement with expert screenwriters on Task Completion, Artifact Compliance, and Basic Lexical Defects, where judgments rely more on explicit requirements or observable violations. Agreement is lower for Contextual Consistency and Dramatic Quality, which involve more contextual interpretation or subjective judgment. This pattern suggests that the evaluator is more consistent with experts on determinate checks than on judgments requiring interpretation.

6 Conclusion

This paper makes three contributions. First, we introduce ScreenWritingBench, a benchmark of 279 tasks constructed and validated by professional screenwriters. Its tasks cover the full screenwriting workflow, from proposals and story synopses through character profiles, character biographies, and scene breakdowns to scripts and their revision. The suite includes single-turn and multi-turn tasks; 219 include dialogue history. Second, we develop a five-dimensional evaluation of Task Completion, Contextual Consistency, Artifact Compliance, Basic Lexical Defects, and Dramatic Quality. The task, norm, and quality layers organize the description of these dimensions. Each dimension uses a suitable evaluation procedure: the first four use explicit rubrics, while Dramatic Quality compares responses with eight expert-ordered anchors for finer score discrimination and supports anchor extension. Third, we evaluate 18 models and report both an overall leaderboard and scores on all five dimensions.

The overall leaderboard summarizes model performance, while the dimension scores show where each model is stronger or weaker. The highest-ranked model also leads Dramatic Quality, yet its performance varies across the other dimensions. Variation across the five dimensions shows that screenwriting capability is heterogeneous and benefits from a multidimensional reading of the results.

REPRODUCIBILITY STATEMENT

The appendix provides scoring rules, rubric specifications, and the definitions used in the ranking analyses.

AI USE STATEMENT

AI assistance was used for manuscript editing and translation, data-consistency checks, figure-label editing, and LaTeX preparation.

REFERENCES

Qian Cao, Xiting Wang, Yuzhuo Yuan, Yahui Liu, Fang Luo, and Ruihua Song. Evaluating text creativity across diverse domains: A dataset and large language model evaluator. *arXiv preprint arXiv:2505.19236*, 2025. URL https://arxiv.org/abs/2505.19236.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024. doi: 10.1145/3613904.3642731. URL https://arxiv.org/abs/2309.14556.

Daniel Fein, Sebastian Russo, Violet Xiang, Kabir Jolly, Rafael Rafailov, and Nick Haber. Litbench: A benchmark and dataset for reliable evaluation of creative writing. *arXiv preprint arXiv:2507.00769*, 2025. URL https://arxiv.org/abs/2507.00769.

Andrew Zhuoer Feng, Cunxiang Wang, Yu Luo, Lin Fan, Yilin Zhou, Zikang Wang, Xiaotao Gu, Jie Tang, Hongning Wang, and Minlie Huang. HoWToBench: Holistic evaluation for LLM's capability in human-level writing using tree of writing. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, 2026. URL https://arxiv.org/abs/2604.19071.

Fan Gao, Dongyuan Li, Ding Xia, Fei Mi, Yasheng Wang, Lifeng Shang, and Baojun Wang. Essaybench: Evaluating large language models in multi-genre chinese essay writing. *arXiv preprint arXiv:2506.02596*, 2025. URL https://arxiv.org/abs/2506.02596.

Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. Followbench: A multi-level fine-grained constraints following benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024. URL https://arxiv.org/abs/2310.20410.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024a. URL https://arxiv.org/abs/2310.08491.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024b. URL https://arxiv.org/abs/2405.01535.

Hang Lei, Shengyi Zong, Zhaoyan Li, Ziren Zhou, Hao Liu, and Liang Yu. Beyond direct generation: A decomposed approach to well-crafted screenwriting with LLMs. *arXiv preprint arXiv:2510.23163*, 2025. URL https://arxiv.org/abs/2510.23163.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. Generative judge for evaluating alignment. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.05470.

Zefeng Lin, Yi Xiao, Zhiqiang Mo, Qifan Zhang, Jie Wang, Jiayang Chen, Jiajing Zhang, Hui Zhang, Zhengyi Liu, Xianyong Fang, and Xiaohua Xu. R^2: A LLM based novel-to-screenplay generation framework with causal plot graphs. *arXiv preprint arXiv:2503.15655*, 2025. URL https://arxiv.org/abs/2503.15655.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://aclanthology.org/2023.emnlp-main.153/.

Adian Liusie, Potsawee Manakul, and Mark J. F. Gales. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, 2024. URL https://arxiv.org/abs/2307.07889.

Shijian Ma, Yunqi Huang, and Yan Lin. Dramabench: A six-dimensional evaluation framework for drama script continuation. *arXiv preprint arXiv:2512.19012*, 2025. URL https://arxiv.org/abs/2512.19012.

Piotr Mirowski, Kory W. Mathewson, Jaylen Pittman, and Richard Evans. Co-writing screenplays and theatre scripts with language models: An evaluation by industry professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023. doi: 10.1145/3544548.3581225. URL https://arxiv.org/abs/2209.14958.

Qiuyu Tian, Yiding Li, Yingce Xia, Fengyi Chen, Youyong Kong, Fan Guo, Yuyao Li, Jinjing Shen, Zhijing Xie, Yiyun Luo, Xin Zhang, and Zequn Liu. STAGE: A full-screenplay benchmark for reasoning over evolving stories. *arXiv preprint arXiv:2601.08510*, 2026. URL https://arxiv.org/abs/2601.08510.

Yuning Wu, Jiahao Mei, Ming Yan, Chenliang Li, Shaopeng Lai, Yuran Ren, Zijia Wang, Ji Zhang, Mengyue Wu, Qin Jin, and Fei Huang. Writingbench: A comprehensive benchmark for generative writing. *arXiv preprint arXiv:2503.05244*, 2025. URL https://arxiv.org/abs/2503.05244.

Sihui Yang, Keping Bi, Wanqing Cui, Jiafeng Guo, and Xueqi Cheng. LINKAGE: Listwise ranking among varied-quality references for non-factoid QA evaluation via LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 6985–7000, 2024. URL https://aclanthology.org/2024.findings-emnlp.410/.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. FLASK: Fine-grained language model evaluation based on alignment skill sets. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2307.10928. Spotlight.

Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. Wordcraft: Story writing with large language models. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, 2022. URL https://dl.acm.org/doi/10.1145/3490099.3511105.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Datasets and Benchmarks Track*, 2023. URL https://arxiv.org/abs/2306.05685.

Mingzhe Zheng, Dingjie Song, Guanyu Zhou, Jun You, Jiahao Zhan, Xuran Ma, Xinyuan Song, Ser-Nam Lim, Qifeng Chen, and Harry Yang. CML-Bench: A framework for evaluating and enhancing LLM-powered movie scripts generation. *arXiv preprint arXiv:2510.06231*, 2025. URL https://arxiv.org/abs/2510.06231.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023. URL https://arxiv.org/abs/2311.07911.

A Rubric and Anchor Categories

This appendix records the rubric categories used for Task Completion, Contextual Consistency, Artifact Compliance, and Basic Lexical Defects, with one section for each rubric-based dimension. Task Completion and Contextual Consistency use task-specific rubrics derived from the dialogue history and current query. Artifact Compliance and Basic Lexical Defects use category-aware rubrics; Dramatic Quality uses eight expert-ordered anchors per task.

A.1 Task-Completion Rubric Categories

Task-Completion rubrics cover formal constraints, requested writing actions, and essential plot points. Table 3 lists the violation categories used by Task Completion.

Table 3: Task-Completion rubric categories.

Dimension	Code	Definition
TC	F1	Required content is absent.
	F2	The requested deliverable is incomplete.
	F3	Required content is altered contrary to the task.
	F4	Content prohibited by the task is present.
	F5	The response promises compliance without delivering it.

A.2 Contextual-Consistency Rubric Categories

Contextual Consistency uses rubrics derived from the dialogue history and current query for conflicts with established story facts (C1–C4 in Table 4). The same dimension also checks contradictions within the response using the nine categories in Table 5. These categories apply across tasks.

Table 4: Contextual-Consistency rubric categories for established story facts.

Dimension	Code	Definition
CC	C1	An outcome conflicts with the established story.
	C2	A character or world state conflicts with the established story.
	C3	Details from incompatible story versions are mixed.
	C4	An established story invariant is violated.

Table 5: Nine categories of contradictions within a response (I1–I9), used in Contextual Consistency evaluation.

Code	Violation type	Definition
I1	Factual contradiction	Facts asserted at different points of R cannot both hold.
I2	Cross-module contradiction	Mutually inconsistent statements across the separate modules of the same deliverable.
I3	State discontinuity	An entity's state breaks discontinuously between passages of R.
I4	Chronology confusion	The internal time line of R is inconsistent or disordered.
I5	Unmotivated repetition	A story element recurs in R without motivation.
I6	Knowledge contradiction	A character's knowledge state contradicts what R establishes about who knows what.
I7	Naming confusion	Characters or objects are named inconsistently within R.
I8	Numerical contradiction	Quantities (counts, amounts, dates) asserted at different points of R conflict.
I9	Point-of-view mechanism error	The narrative point-of-view mechanism is applied inconsistently or incorrectly.

A.3 Artifact-Compliance Rubric Categories

The full Artifact-Compliance rubric index contains nine base rule sets (Table 6). Table 7 gives one representative format violation and one content violation per set. The character set covers both profiles and

biographies. The scene-planning sets distinguish coarse scene breakdowns from detailed scene outlines.

Task artifact labels in Section 3.2 and base rubric sets are not in one-to-one correspondence. Proposals combine four sets: Logline (12 rules), World-building (10), Character profile / biography (16), and Story synopsis (15), totaling 53 component rules already included in the 138-rule base inventory. The index repeats the Script set under Dialogue excerpt and contains an empty Other-artifact entry; neither adds a base rule set. Three recommendation-only rules (two for Script and one for Episode outline) are excluded from the table counts.

Table 6: Format/structure and content criteria in the nine base Artifact-Compliance rule sets. The 138 criteria exclude recommendations and repeated component sets.

Variant	Rules	Composition
Script	19	8 format + 11 content
Story outline	13	9 format + 4 content
Episode outline	14	10 format + 4 content
Coarse scene breakdown	19	11 format + 8 content
Detailed scene outline	20	13 format + 7 content
Story synopsis	15	9 format + 6 content
World-building	10	4 format + 6 content
Logline	12	3 format + 9 content
Character profile / biography	16	9 format + 7 content

Table 7: Representative format and content violations for the nine base Artifact-Compliance sets. These examples do not replace the full rule sets counted in Table 6.

Artifact	Format violation	Content violation
Script	Inconsistent character names: a character's core name must remain identical throughout the script; e.g., the same character is called "Li Ming" and "Wang Ming."	An action line states an unfilmable thought, e.g., "Li realizes he was betrayed."
Story outline	Events are presented as numbered list items instead of narrative paragraphs.	Character information is listed as attributes, e.g., "Age: 30; trait: timid."
Episode outline	An episode contains an internal subheading, e.g., "Plotline A."	The narrative comments on its own function, e.g., "This scene builds tension."
Coarse scene breakdown	Scene numbers skip a number, e.g., Scene 1 followed by Scene 3.	A character listed in the scene heading never appears in its synopsis.
Detailed scene outline	A scene heading omits a definite time or location.	An entire scene merely states the result, e.g., "Zhao persuades the landlord," without developing actions and counteractions.
Story synopsis	The synopsis uses scene numbers, e.g., "Scene 1," instead of continuous prose.	The synopsis gives the setup but omits the central conflict and its resolution.
World-building	Markdown markup appears in the plain-text description, e.g., "**Politics**."	The setting's era cannot be inferred from the text.
Logline	The logline is presented as unrelated bullet items.	The protagonist's goal is absent, e.g., "A detective lives in a coastal town."
Character profile / biography	The biography gives basic information but omits the character's life history.	The profile contains script dialogue, e.g., "Li: Leave now!"

A.4 BASIC-LEXICAL-DEFECT CATEGORIES (B1–B6)

Basic Lexical Defects uses six category-aware criteria for observable surface defects in Chinese responses. The detector receives the response and may consult the current request and dialogue history to identify intentional quotations or inherited wording; it does not receive expert annotations. Every positive finding requires a verbatim evidence span from the response. Formatting conventions are evaluated under Artifact Compliance rather than BLD. Table 8 defines the six classes and their boundaries.

Table 8: Six Basic Lexical Defect categories (B1–B6) and their decision boundaries.

Class	**Definition and boundary**
B1 Garbled or anomalous characters	Corrupted text whose meaning cannot be restored by replacing a foreign word with Chinese: meaningless symbol runs, repeated-symbol blocks, encoding failures, or stray characters at fixed positions. Unrecoverable foreign-script text also belongs here. Normal punctuation, standard script symbols, writing placeholders, and formatting markup do not count. Recoverable foreign-script text belongs to B3.
B2 Mechanical repetition	A sentence or passage reproduced verbatim within the same response, including the original and its duplicate. Repeated sentence patterns, verbose restatements, and text intentionally carried over from the user or dialogue history do not count.
B3 Writing-system mixing	A foreign word or script inserted into otherwise Chinese text when its meaning is recoverable and a Chinese substitution restores fluency. This includes English, recoverable Japanese/Korean/Cyrillic text, and traditional characters in simplified-Chinese prose. Common loanwords and abbreviations, screenplay terms, proper names, titles, and wording requested or inherited from the dialogue history are exempt. The B1/B3 boundary is whether meaning can be restored.
B4 Typos	A particular wrong, missing, extra, or transposed character can be located and corrected. This includes mismatched stated counts, misspelled fixed expressions, and incorrect official names. Deliberate wordplay, character speech errors used as a narrative device, and wording inherited from the request or earlier turns are excluded. Traditional characters belong to B3 rather than this class.
B5 Ungrammatical sentences	A sentence is difficult to parse because of broken syntax, incompatible collocations, missing constituents, unclear subjects, or disordered word order; spoken dialogue that cannot be delivered naturally may also qualify. Ordinary colloquial ellipsis and stylistic choices are exempt. Logical or plot-level incoherence is evaluated elsewhere. A specific wrong character belongs to B4 instead.
B6 Generation truncation	The response ends abruptly in the middle of a sentence or passage without closure. An intentional open ending, a completed requested excerpt, or a complete final sentence followed by missing requested sections does not count; the last case concerns Task Completion.

B TASK-AXIS TAXONOMY

B.1 TASK-TAXONOMY OVERVIEW

The task dimensions characterize operations, artifacts, interaction settings, and instruction conditions. Across the ten axes, the taxonomy defines 172 axis-specific tags; the shared 19-term artifact vocabulary is counted once for input and once for output. Table 9 summarizes all ten axes with representative tags. A task may have multiple tags on an axis.

Table 9: Ten task dimensions and representative tags.

Axis	**Typical tags**	**Definition**
Action	Generate, Modify, Evaluate	The operation the task asks for.
Object	Plot, Character, Creative System	The entity the task operates on.
Scope	Global, Scene, Paragraph	The granularity of the operation.
I/O Type	Script, Synopsis, Outline	Input and delivered artifact forms.
Turn Role	First Turn, Continuation, Pivot	The turn's role in the dialogue.
Ref. State	None, Creative Norms	Whether references are provided.
Inst. Clarity	Precise, Directional, Open	How explicit the instruction is.
Draft State	Complete Draft, Partial Draft	The draft the task starts from.
Constraint	Required Elements, Consistency	Explicit requirements on output.
Modifier	Literary Style, Emotional Tone	Stylistic or tonal requirements.

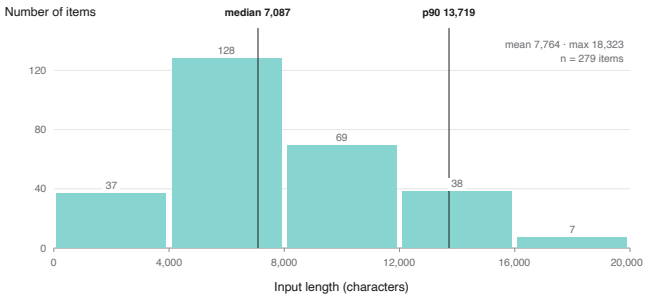


Figure 6: Input length distribution of the 279 tasks, grouped in 4,000-character bins. The median is 7,087 characters and the 90th percentile is 13,719.

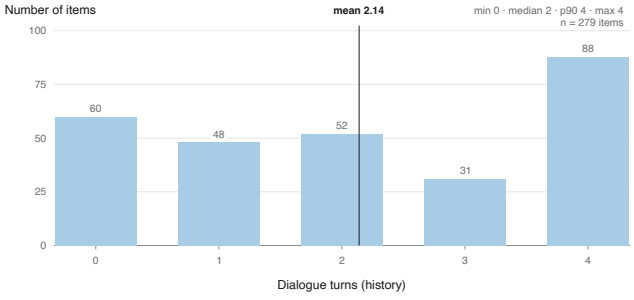


Figure 7: Distribution of the number of dialogue turns per item.

C Comparing Anchor Counts

We compare four Dramatic Quality scoring conditions on a subset of 97 tasks from the 279-task benchmark and eight models selected from the main leaderboard, yielding 776 task–model responses (775 valid rubric scores). The comparison tests whether multiple expert-ordered anchors provide finer score resolution than one anchor or no anchors. The eight-anchor condition inserts each response among the task's eight ordered anchors; the single-anchor condition compares it directly with one anchor. The no-anchor rubric condition uses the pass rate of task-specific checklist criteria for Dramatic Quality. The no-anchor direct condition retains the main ranking criteria in a pointwise prompt and adds five absolute-quality rating bands. Thus, the two no-anchor conditions also change the evaluation procedure or scale; differences among all four conditions cannot be attributed to anchor count alone.

Figure 8 shows the resulting score distributions. Eight anchors place responses across all nine available score levels, whereas one anchor assigns 0 for a loss, 50 for a tie, and 100 for a win, producing many tied scores. The two no-anchor controls concentrate scores near the top of their scales. On this sample, multiple anchors provide finer score resolution than one anchor; the distribution alone does not establish greater agreement with expert judgment. Appendix D examines the addition of a ninth anchor.

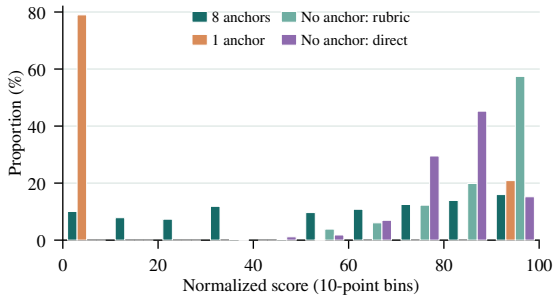


Figure 8: Score distributions for 97 subset tasks and eight models. Scores are normalized to 0–100 and plotted in ten-point bins; eight anchors yield nine levels at 12.5-point intervals.

D Constructing the A9 Anchor

To test whether the Dramatic Quality anchor set can be extended, we construct a ninth anchor (A9) from the original eight expert-ordered anchors. The experiment uses subset tasks for which an A9 candidate is available and eight models selected from the main leaderboard. For each task, the initial candidate is drawn from existing model responses and improved through self-refinement. Two judge models, DeepSeek-V4-pro and Claude-Opus-4.6, then assess it independently against the nine options. The candidate qualifies as A9 only if both judges rank it first.

Table 10 and Figure 9 report the matched Dramatic Quality scores before and after adding A9. Scores decrease by 4.2 points on average, while the ordering of the eight evaluated models remains unchanged.

Table 10: Matched A9 scores (0–100 scale).

Measure	Eight anchors	Nine anchors
Top model score	81.5	76.4
Bottom model score	18.9	17.0
Top–bottom spread	62.6	59.4
Models changing rank	0 of 8	
Mean score change	−4.2 points	

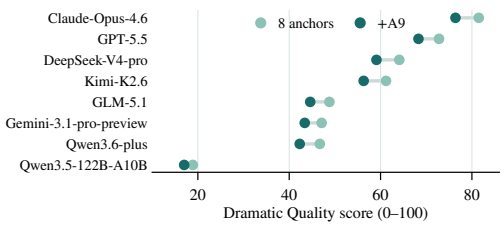


Figure 9: Dramatic Quality with A9; model ordering is unchanged.

E LIMITATIONS

The benchmark currently consists primarily of Chinese-language tasks For the subjective Dramatic Quality dimension, evaluation results show 58.89% agreement with expert annotations, leaving room for improvement. Moreover, the evaluation depends on the quality of the judging models, rubrics, and scoring anchors. The quality of these components is, in turn, constrained by the expertise of the human reviewers. Substantial room remains for improving alignment between the evaluation method and expert reviewers' judgments. These limitations motivate broader language coverage and stronger expert validation in future work.