PlotTwist: A Creative Plot Generation Framework with Small Language Models

Abhinav Thorat[*]
Sony Research India
abhinav.thorat@sony.com

Ravi Kolla[*]
Sony Research India
ravi.kolla@sony.com

Jyotin Goel[†]
Sony Research India
jyotin.goel@sony.com

Madhav Kataria[†]
Sony Research India
kataria.madhav@sony.com

Niranjan Pedanekar
Sony Research India
niranjan.pedanekar@sony.com

Abstract

Creative plot generation presents a fundamental challenge for language models: transforming a concise premise into a coherent narrative while satisfying multiple narrative constraints, including global coherence, character development, pacing, tone consistency, and emotional progression. Although recent Large Language Models (LLMs) demonstrate strong fluency on general-purpose tasks, they require preference alignment to perform well on domain-specific tasks such as creative plot generation. However, conducting such alignment at the scale of frontier LLMs is computationally prohibitive, significantly limiting accessibility and practical deployment. To address this, we present PLOTTWIST, a structured framework that enables Small Language Models (SLMs) with $\leq$ 3B active parameters to generate high-quality, premise-conditioned plots competitive with frontier systems of vastly greater parameter scale. Our approach decomposes generation into three specialized components: (1) an Aspect Rating Reward Model, trained via a novel Positive–Negative prompting strategy; (2) a Mixture-of-Experts (MoE) plot generator aligned via Direct Preference Optimization (DPO); and (3) an Agentic Evaluation module containing a cross-family jury that emulates human critical judgment for unbiased, independent post-hoc assessment. Extensive experiments demonstrate that PLOTTWIST consistently outperforms all baselines, including frontier models, across multiple Narrative Quality Dimensions (NQDs), achieving higher win rates against all but the strongest baseline, with which it remains competitive. Further validation confirms strong sensitivity to narrative quality, as the framework reliably distinguishes plots derived from critically acclaimed versus widely panned screenplays. Together, these results establish structured, preference-based alignment as a resource-efficient approach to high-quality creative plot generation. Project Page.

[*] Equal contribution.
[†] Work done during an internship at Sony Research India.

1 Introduction

Writers across film studios, streaming platforms, and publishing houses constantly face the same challenge: transforming a concise creative premise into a compelling narrative outline under tight deadlines. A showrunner needs to develop episode arcs for a new series. A screenwriter must pitch three distinct treatments by week's end. An educator requires diverse story examples for a creative writing course. In each case, the task is not simply to generate text, but to craft plots that exhibit coherent structure, believable character arcs, consistent tone, and emotionally resonant turning points, qualities that distinguish professional storytelling from arbitrary event sequences (Teleki et al., 2025; Yao et al., 2019). While experienced writers navigate these demands through years of training and intuition, computational systems offer no comparable mechanism for structured creative assistance, a gap this work seeks to address.

The challenge of creative plot generation with generative models extends beyond surface-level text production. Unlike summarization or question-answering, where local context often suffices, plot generation demands long-horizon reasoning over concise conditioning signals. A promising premise, such as "`a romantic comedy set in the modern tech startup era`", provides minimal concrete guidance to writers, yet must expand into a causally connected sequence of events spanning setup, development, climax, and resolution. The narrative must maintain global coherence while ensuring that early character motivations align with later decisions, that tonal shifts feel earned rather than arbitrary, and that pacing sustains engagement across the entire arc. These requirements pose significant difficulties for standard autoregressive language models, which optimize token-level likelihoods and lack explicit mechanisms for enforcing discourse-level constraints. Prior work has shown that hierar-

Proceedings of the 3rd Workshop on Language Understanding in the Human-Machine Era (LUHME 2026), pages 144–172
October 28, 2026 ©2026 Association for Computational Linguistics

chical planning and explicit structural decomposition can improve narrative consistency (Fan et al., 2018; Gurung and Lapata, 2025; Teleki et al., 2025; Yao et al., 2019) but such approaches typically assume large model capacities, task-specific supervision, or relaxed efficiency constraints, leaving open the question of whether effective plot generation is achievable under computational constraints.

Recently, LLMs have demonstrated impressive fluency across creative writing tasks, yet their success comes at a steep cost. Frontier models such as GPT-4.1, Claude Sonnet 4, and Gemini 2.0 Flash operate at vastly greater parameter scales, demanding substantial computational infrastructure for both training and inference, costs that are ultimately passed on to end users. Beyond raw cost, the literature consistently shows that task-specific alignment yields meaningfully better performance than relying on general-purpose models alone (Ouyang et al., 2022; Sun et al., 2023). In creative writing, this alignment imperative is compounded by a deeper challenge: scale alone does not reliably resolve long-horizon coherence. Even the largest models exhibit narrative drift, inconsistent characterization, and structural incoherence when generating extended plots without additional inductive biases (Fan et al., 2018; Yao et al., 2019). Achieving professional-grade plot generation thus requires targeted alignment to the creative domain, yet for models of this scale, such alignment is computationally prohibitive, particularly when the resulting system is intended for a narrow, specialized use case.

This observation motivates a natural question: can Small Language Models (SLMs), defined here as models with $\leq$ 3B active parameters per token, generate creative plots of comparable quality to frontier LLMs when aligned using an appropriate structural scaffolding? We hypothesize that the key lies not in model scale, but in externalizing narrative structure into explicit evaluative and training signals. Rather than relying on a monolithic model to implicitly learn all aspects of narrative quality through token prediction, we propose decomposing the generation process into specialized, modular components. This architectural separation enables SLMs to leverage explicit guidance where large models rely on emergent capabilities, effectively trading model capacity for structured workflow design. To operationalize this approach, we introduce PLOTTWIST, a three-component framework for concise premise-conditioned plot generation with SLMs. The first component is an *Aspect Rating Reward Model* that evaluates plots across five Narrative Quality Dimensions (NQDs): character development, tone consistency, pacing, narrative coherence, and emotional turning points. The second component is a *Mixture-of-Experts (MoE) Plot Generator* based on Qwen-3-30B-A3B (3B active parameters), trained via Direct Preference Optimization (DPO) (Rafailov et al., 2023) on preference pairs derived from the aspect rating reward model. The third component is an *Agentic Evaluation Module with a Cross-Family Jury* that operates independently of the training pipeline, providing post-hoc assessment through structured, weakness-focused criteria. The key contributions of this work are as follows.

- **Structured Workflow using SLMs for Plot Generation.** We propose a modular framework comprising an Aspect Rating Reward Model, a DPO-trained MoE Plot Generator, and an independent Agentic Evaluation Module.
- **Positive–Negative Prompting for Aspect Rating Reward Modeling.** We introduce a novel prompting strategy that mitigates positivity bias in LLM-based evaluation, enabling reliable aspect-level supervision across five NQDs.
- **Agentic Evaluation using a Cross-Family Jury.** We develop a cross-family jury of five open-weight LLMs, together with a pre-registered evaluation protocol incorporating structured deliberation, a held-out judge, and reliability analyses.
- **External Validation of Evaluation Components.** We demonstrate that both the reward model and the agentic evaluator reliably distinguish acclaimed from critically panned plots across all NQDs.
- **Efficient, Quality-Adaptive Plot Generation.** We show that PLOTTWIST outperforms all baselines across multiple NQDs and achieves higher per-premise win rates against all but the strongest baseline, with which it remains competitive, despite using only 3B active parameters. Moreover, PLOTTWIST exhibits principled intervention scaling across quality strata, lightly refining strong narratives while substantially restructuring weak ones rather than uniformly inflating scores.

2 Related Work

Story and Plot Generation. Early neural methods introduced hierarchical generation, decomposing stories into premise and continuation stages (Fan

et al., 2018), while Plan-and-Write frameworks explicitly separated outline planning from surface realization (Yao et al., 2019). More recently, Agents' Room (Huot et al., 2024) simulates professional writing rooms. Such systems remain tethered to frontier-scale computation, and a recent survey notes that structured, quality-aware generation remains an open problem (Teleki et al., 2025).

Preference Alignment and Efficient Models. Reinforcement Learning from Human Feedback (RLHF) is effective for aligning LLMs with human preferences, but the training pipeline can be computationally expensive. Direct Preference Optimization (DPO) (Rafailov et al., 2023) offers a simpler, more stable alternative, making it attractive for resource-constrained settings, as is sparse computation in MoE architectures (Shazeer et al., 2017; Fedus et al., 2022).

Evaluation of Creative Text and Reliability of LLM Judges. Evaluating creative generation remains challenging. Despite their strong surface-level fluency, LLM-generated stories often lack authentic creativity (Chakrabarty et al., 2024). Consequently, recent work has increasingly adopted LLM-as-a-Judge frameworks (Zheng et al., 2023) and dimension-specific benchmarks for creative writing (Zheng et al., 2025; Kim and Oh, 2025). However, LLM judges exhibit systematic biases, including same-family preference (Panickssery et al., 2024; Ye et al., 2024), verbosity bias (Dubois et al., 2024), and limited agreement with expert evaluations (Fein et al., 2025). Existing mitigation strategies include cross-family judge panels (Verga et al., 2024), structured deliberation (Chan et al., 2024), and evidence-based justifications (Jiang et al., 2025); nevertheless, judge errors remain correlated across models (Kohli, 2026), and conventional agreement measures can be distorted by skewed score distributions.

3 Problem Formulation

We consider premise-conditioned plot generation with SLMs[1]: given a concise, high-level premise specifying the narrative setting, genre, and thematic constraints (e.g., `a romantic comedy set in the modern tech startup era`), the model must produce a plot approaching professionally authored narrative quality. Drawing on computational narrative modeling and affective narratology (Chakrabarty et al., 2024; Hogan, 2011), we assess quality along five Narrative Quality Dimensions (NQDs): *narrative coherence* (global logical consistency and causal connectivity), *character development* (meaningful character evolution), *pacing* (distribution of narrative progression), *tone consistency* (stylistic alignment), and *emotional turning points* (effectiveness of major affective transitions). Together, these dimensions span the structural, temporal, stylistic, character-centric, and affective aspects of narrative quality while keeping the evaluation space compact. Our objective is a structured SLM-based workflow that generates premise-conditioned plots exhibiting strong performance across all NQDs.

[1]We distinguish models by active parameter count rather than total parameters, and refer to models with at most 3B active parameters per token as SLMs, even when implemented as MoE architectures with larger total parameter counts.

4 Proposed Methodology

As shown in Figure 1, PLOTTWIST comprises three modules: an Aspect Rating Reward Model that scores plots across the NQDs (Section 4.1), a preference-aligned Plot Generator (Section 4.2), and an independent Agentic Evaluation module with a cross-family jury (Section 4.3).

4.1 Aspect Rating Reward Model

Our objective is to develop a reward model that produces aspect-level ratings and can be used to guide the plot generator model. We begin by constructing a dataset comprising plots paired with their corresponding ratings across the considered NQDs, and then fine-tune an LLM on this dataset to assign continuous-valued scores to plots. For a given plot-p and aspect-a, we use $r_a(p)$ to denote the rating of plot-p along aspect-a in NQD.

4.1.1 Aspect rating dataset construction

As there are no existing datasets that provide fine-grained ratings of plots across the considered NQDs, we construct such a dataset synthetically using LLMs. To the best of our knowledge, the only widely available human-provided score is the IMDb rating for movies. However, this rating serves as a holistic assessment that encapsulates many creative elements simultaneously. Consequently, it cannot be used as a direct proxy for individual aspect-specific ratings, but it can serve as a coarse aggregate indicator of the overall quality of the plot. Note that IMDb ratings are not used in model training; they are employed solely for data curation and stratification purposes. We begin by

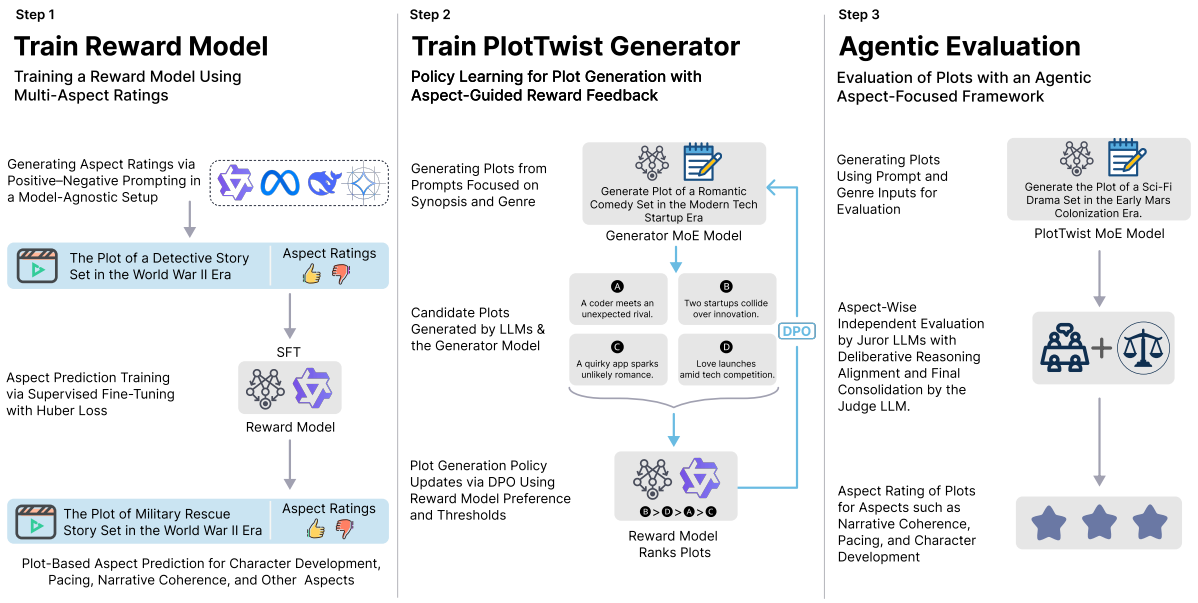

Figure 1: PlotTwist Framework

randomly sampling 5000 movies from the MovieLens (Harper and Konstan, 2015) dataset spanning a broad range of IMDb ratings to ensure diversity. For each movie, we then scrape its corresponding plot from Wikipedia [2]. Next, we employ LLMs to generate synthetic aspect-level ratings for each plot. We emphasize that LLMs are used solely in the reward model, while the final plot generator model, described in the following section, is implemented as an SLM.

For each plot p, we generate ratings for all aspects in NQDs using positive-negative prompting [3] in a model agnostic setup, described below. This prompting strategy mitigates the positive bias often observed in LLM-based evaluation (Zheng et al., 2023) and reduces reliance on potentially memorized IMDb ratings, yielding stronger correlations with external quality indicators and enabling more accurate and balanced plot evaluation. This style of prompting mitigates the inherent positive bias often observed in LLMs (Zheng et al., 2023) and yields stronger correlation with external indicators, enabling more accurate and balanced critique of plots.

We first take five LLMs, namely `Qwen-2.5-7B`, `Llama-3.3-70B`, `Llama-3.1-8B`, `DeepSeek-14B` and `Gemma-27B`, to avoid model bias in the aspect rating. We then prompt them with a plot to output a rating, on a scale of 1-10, for each aspect by only considering the positives present in the plot along that aspect, denoted as $r^{+}_{a,m}(p)$ where m identifies the LLM. Similarly, we prompt LLMs to output a rating, on a scale of 1-10, by only considering the negatives present in the plot along each aspect, denoted as $r^{-}_{a,m}(p)$. Note that, if an aspect is well captured in the plot then we expect $r^{+}_{a,m}(p)$ and $r^{-}_{a,m}(p)$ to be high and low respectively. Then, the final aspect rating of a plot is calculated as $r_a(p) = \sum_m \left(r^{+}_{a,m}(p) - r^{-}_{a,m}(p)\right)$.

4.1.2 Supervised Fine-Tuning (SFT)

We fine-tune Qwen-3-32B with 4-bit quantization on the aspect-rating dataset using a joint language-modeling and regression objective. The token-level Cross-Entropy (CE) loss supervises generation of the target response, whereas the Huber loss penalizes deviations between predicted and target aspect ratings. For an input sequence $x_{1:T}$, target sequence $y_{1:T}$, and model parameters θ, the CE loss is defined as

$$\mathcal{L}_{CE}(\theta) = -\frac{1}{T}\sum_{t=1}^{T} \log p_\theta\left(y_t \mid x_{1:t}\right).$$

Let r denote the residual between a predicted aspect rating and its target value. The Huber loss is defined as

$$\mathcal{L}_\delta(r) = \begin{cases} \frac{1}{2}r^2, & |r| \le \delta \\ \delta\left(|r| - \frac{1}{2}\delta\right), & |r| > \delta \end{cases},$$

where we set $\delta = 1$ in all experiments.

[2] To meet the token output and computation requirements we select only movies with plots of at most 4000 words.

[3] For reference, both positive and negative prompts for all aspects are given in the Appendix A.

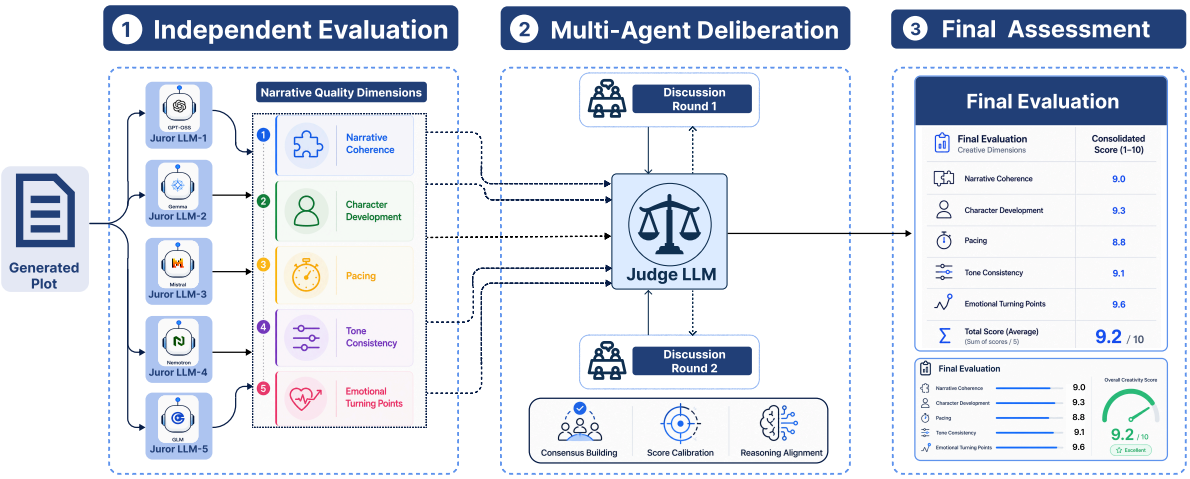

Figure 2: The Cross-Family Jury - Agentic Evaluation Framework

4.2 Plot Generator Model

We adopt Qwen-3-30B-A3B, an MoE architecture that enables increased model capacity and expert specialization while maintaining efficient inference (Fedus et al., 2022; Shazeer et al., 2017). Although the MoE model has a total parameter count of 30B, only 3B parameters are active per token, which classifies it as an SLM under our definition. To further align the model toward producing higher-quality plots, we employ Direct Preference Optimization (DPO), a Reinforcement Learning from Human/AI Feedback (RLHF/RLAIF) approach that directly optimizes preference objectives without requiring an explicit reward model or on-policy reinforcement learning. Given that the Qwen-3-30B-A3B model already exhibits strong instruction-following capabilities, we omit SFT for instruction alignment and focus exclusively on preference-based optimization.

For DPO training, we first construct a dataset of pairwise plot preferences (Figure 3 in Appendix B). Each sample in the dataset consists of a premise, a pair of plots generated under that premise, and a preference ordering between them. We generate premise descriptions for each of the 5,000 plots, as mentioned in Section 4.1, using the Gemma-27B model. For each premise, we prompt the base Qwen-3-30B-A3B MoE model, along with several frontier models (Claude Sonnet 4, Gemini 2.0 Flash, and GPT-4.1), to generate plots conditioned on the same premise. All generated plots are evaluated using the reward model described in Section 4.1, with aspect-level ratings averaged to obtain a final reward for each plot. To align our MoE model toward higher-quality plot generation, we retain only those samples in which a frontier model achieves the highest reward score, exceeds a score threshold of 8, and outperforms the next-best model by a margin of at least 0.5. Applying this procedure across all 5,000 premises yields 160 high-confidence preference samples. Although modest in size, this dataset is intentionally curated to ensure reliable preference signals, and prior work has shown that DPO can effectively leverage a small number of high-quality preference pairs (Djuhera et al., 2026; Wu et al., 2024). We subsequently perform DPO on the Qwen-3-30B-A3B model using this dataset, yielding the final plot generator model. Computational infrastructure details are provided in Appendix H.

4.3 Agentic Evaluation of Plots using Grounded Cross-Family Jury

Following plot generation, reliable validation of creative quality is essential. Although the aspect rating reward model provides structured supervision across the predefined NQDs, it remains a predictive model optimized for aspect-level signals derived from training data. Relying solely on this model risks evaluation bias, as it may reward patterns correlated with learned signals rather than reflect broader narrative soundness. Although direct evaluation by domain experts remains the gold standard for assessing plot quality, it is often impractical due to limited availability, time, and cost. Therefore, following prior work (Teleki et al., 2025; Kim and Oh, 2025), we adopt an independent agentic evaluation framework, described below and illustrated in Figure 2. Our agentic evaluation framework performs structured, multi-criteria assessment

across the NQDs independently of the aspect rating reward model. It employs a panel of five independent Juror LLMs to assess each generated plot across five NQDs. Each juror performs its evaluation independently to encourage diverse and unbiased assessments. Their individual judgments are then provided to a Judge LLM, which consolidates the evaluations through two rounds of structured deliberation with the jurors, resolving disagreements and calibrating scores where necessary. The Judge LLM produces the final consensus score for each NQD, yielding a robust, consistent, and transparent assessment while reducing the variance inherent in single-model evaluations.

To ensure evaluation that is reliable, consistent, and independent of the identity of the LLM applying it, each NQD is specified through a rubric of ten explicit, instruction-level criteria that translate abstract narrative concepts into concrete and observable failure modes; each criterion is scored on a $[0, 1]$ scale, and the criterion scores are summed to yield an aspect score out of ten. Narrative coherence, for example, is assessed by identifying breakdowns in logical progression, causal relationships, and structural consistency, including plot holes, contradictions, and incoherent world-building. The remaining four dimensions are specified in the same style; the full rubrics and exact prompts appear in Appendix C.

5 Experiments

We first validate both the proposed Aspect Rating Reward Model and the Cross-Family Jury-based Agentic Evaluation framework. Next, we evaluate the PlotTwist under varying conditions and introduce the baselines. Finally, we compare PlotTwist against these baselines and present ablation studies.

5.1 Aspect Rating Reward Model and Agentic Evaluation: Validation

We verify that both components assign higher scores to high-quality plots than to low-quality ones. Since IMDb ratings are an imperfect proxy for plot quality, we instead use critically acclaimed plots from the 101 Greatest Screenplays of All Time (GSAT) and critically panned plots from the Golden Raspberry Screenplay Awards (Razzies). Owing to the class imbalance (37 Razzie vs. 94 GSAT)[4], we employ repeated balanced subsampling, fixing the Razzie set and sampling an equal number of GSAT films over 1,000 runs.

Aspect Rating Reward Model. The reward model consistently separates GSAT and Razzie plots across all NQDs. GSAT plots achieve an overall mean score of 8.28 compared with 7.21 for Razzie plots, yielding a mean difference of $+1.07$ (95% CI $[0.96, 1.19]$), with GSAT outperforming Razzie plots in 100% of subsampling runs. The largest improvements are observed in pacing ($+1.41$), emotional turning points ($+1.13$), and narrative coherence ($+1.12$), while character development ($+0.75$) and tone consistency ($+0.96$) also exhibit clear separation. All NQDs show large effect sizes (Cohen's $d = 2.77$–4.68), and Welch's t-tests confirm statistically significant differences across all dimensions, including the aggregate score ($t = 9.69$, $p = 3.41 \times 10^{-18}$).

Agentic Evaluation. The Cross-Family Jury-based agentic evaluator exhibits similar behavior, assigning consistently higher scores to GSAT plots across all NQDs. The overall mean score increases from 6.58 (Razzie) to 8.33 (GSAT), corresponding to a mean difference of $+1.75$ (95% CI $[1.66, 1.85]$), with GSAT outperforming Razzie plots in 100% of subsampling runs. The largest separations occur in narrative coherence ($+2.64$), tone consistency ($+1.99$), and pacing ($+1.91$), followed by character development ($+1.22$) and emotional turning points ($+0.98$). All NQDs exhibit very large effect sizes (Cohen's $d = 4.36$–9.34), and Welch's t-tests confirm statistically significant differences across all dimensions, including the aggregate score ($t = 16.54$, $p = 2.96 \times 10^{-38}$).

Together, these results demonstrate that both the proposed reward model and the Cross-Family Jury-based agentic evaluator reliably distinguish critically acclaimed from critically panned plots across all NQDs.

5.2 Quality-Stratified Analysis of PlotTwist

We consider 160 films partitioned into four IMDb-defined quality strata: Excellent (IMDb > 8), Good ($7 <$ IMDb ≤ 8), Mid ($6 <$ IMDb ≤ 7), and Low (IMDb ≤ 6). For each source plot, we derive a premise using Gemma-27B and use it to condition PlotTwist. The final dataset contains 160 complete original–generated pairs (40 Excellent, 40 Good, 40 Mid, and 40 Low). Within each stratum, we compare paired original and generated plots across the five NQDs under the same Cross-Family Jury

[4]The imbalance arises because the Razzie Awards were established only in 1981, the Worst Screenplay award is not presented every year, and our 4,000-word plot-length filter further reduces the available Razzie pool.

protocol as Section 5.4; uncertainty is estimated by resampling films rather than individual scores.

The overall improvement increases monotonically as source quality decreases: $+0.31$ points for Excellent films (95% CI $[0.09, 0.57]$; paired Cohen's $d_z = 0.38$), $+0.65$ for Good films ($[0.51, 0.79]$; $d_z = 1.40$), $+0.98$ for Mid films ($[0.81, 1.17]$; $d_z = 1.68$), and $+1.28$ for Low films ($[1.09, 1.48]$; $d_z = 2.01$). Within the Excellent stratum, the clearest gains occur in character development ($+0.50$, $[0.20, 0.88]$) and narrative coherence ($+0.40$, $[0.13, 0.72]$), whereas pacing remains unchanged within uncertainty (-0.02, $[-0.21, 0.19]$). In the Low stratum, the largest improvements are observed in narrative coherence ($+1.82$) and tone consistency ($+1.40$), with generated plots outperforming their paired originals on at least 97.5% of films across every NQD. This monotonic trend indicates quality-adaptive generation: PlotTwist primarily refines already strong narratives while progressively restructuring weaker ones (full per-stratum results are provided in Appendix F).

5.3 Baselines

We evaluate PlotTwist against baselines selected to provide a holistic comparison across three orthogonal axes: model scale, architectural design, and plot generation paradigm (see Appendix E for a complete overview of all models and their roles in the framework). Closed-source frontier models (GPT-4.1(OpenAI, 2025), Claude Sonnet 4(Anthropic, 2025), and Gemini 2.0 Flash(DeepMind, 2024)) and large open-weight models (Llama-3.3-70B and Qwen-3-32B (Team, 2025b)) test whether the structure of the framework can substitute for raw scale; instruction-tuned, reasoning-distilled, and reasoning-optimized models of comparable scale (Qwen-2.5-14B (Team, 2024), DeepSeek-R1-14B (DeepSeek-AI, 2025), Mistral Small 2501 (Team, 2025a), and Phi-4 Mini (Microsoft, 2025)) separate the contribution of preference optimization from sparse activation and capacity; and two narrative-specific paradigms, Agents' Room (Huot et al., 2024) and WizardLM (TheBloke et al., 2023), situate PlotTwist among purpose-built story generation systems.

5.4 PlotTwist Performance Evaluation

Jury composition in Agentic Evaluation. Since LLMs have been shown to exhibit bias toward outputs generated by models from the same family (Panickssery et al., 2024; Ye et al., 2024), we construct a jury comprising five open-weight LLMs from five model families different from our aspect rating reward model (Qwen): GPT-OSS-120B(OpenAI et al., 2025), Gemma-4-31B(Team et al., 2026), Mistral-Medium-3.5-128B, Nemotron-3-Super-120B(NVIDIA et al., 2025), and GLM-4.6V (Team et al., 2025), with Llama-3.3-70B serving as the Judge LLM. Each juror applies the same ten-criterion rubric described in Section 4.3, ensuring that only the underlying judge model varies across the panel (implementation details are provided in Appendix G).

Evidence grounding. Each criterion score must cite a span quoted verbatim from the plot being scored, turning written justifications (Jiang et al., 2025) into mechanically checkable claims: between 94% and 99% of the roughly 478,000 quoted spans verify against the source text. Grounding is measured rather than enforced, avoiding the selection bias of dropping cells (Appendix G). We compare PlotTwist against the eleven baselines described in Section 5.3 using the Cross-Family Jury Agentic Evaluation framework (Figure 2). Evaluation is performed on a held-out set of 160 premises sampled from the 5,000-premise dataset.

Mean scores. PlotTwist and Claude Sonnet 4 achieve the highest overall jury mean scores, 8.36 and 8.35, respectively (Table 1), with PlotTwist leading in narrative coherence, tone consistency, and pacing. Given the marginal difference in overall scores, we rely on the per-premise win rate (Table 2) as the primary comparison metric.

Win rates. Following a pre-specified analysis plan, the primary endpoint is the per-premise paired win rate of PlotTwist against each baseline (ties counted as 0.5), evaluated using premise-level bootstrap 95% confidence intervals and Holm-corrected exact sign tests (Holm, 1979) (Table 2). PlotTwist achieves win rates of at least 0.95 against every open-weight baseline, 0.94 against Gemini 2.0 Flash, 0.91 against GPT-4.1 (95% CI $[0.86, 0.95]$), and 0.73 against Agents' Room, all remaining statistically significant after Holm correction for multiple comparisons ($p < 10^{-8}$). The GPT-4.1 result is consistent across jurors: each prefers PlotTwist on a majority of premises, and leave-one-juror-out analysis maintains a win rate of at least 0.85 (Appendix G). In contrast, the comparison with Claude Sonnet 4 is inconclusive, with a win rate of 0.56 (95% CI $[0.49, 0.64]$); neither the sign test ($p = 0.13$) nor the equivalence test (Lakens, 2017) establishes superiority or equivalence. We there-

Model	Character Development	Tone Consistency	Pacing	Narrative Coherence	Emotional Turning Points	Overall
Qwen-2.5-14B	7.41 ± 0.59	7.53 ± 0.43	7.52 ± 0.31	7.69 ± 0.48	7.95 ± 0.39	7.62 ± 0.33
Claude Sonnet 4	$\mathbf{8.30 \pm 0.41}$	8.23 ± 0.35	8.09 ± 0.24	8.49 ± 0.38	$\mathbf{8.63 \pm 0.24}$	8.35 ± 0.24
Gemini 2.0 Flash	8.01 ± 0.41	7.98 ± 0.32	7.91 ± 0.19	8.27 ± 0.32	8.40 ± 0.29	8.11 ± 0.24
GPT-4.1	7.88 ± 0.52	8.14 ± 0.29	7.99 ± 0.19	8.37 ± 0.34	8.41 ± 0.28	8.16 ± 0.24
Llama-3.3-70B	7.11 ± 0.69	7.39 ± 0.55	6.94 ± 0.56	6.89 ± 0.75	7.56 ± 0.53	7.18 ± 0.48
DeepSeek-R1-14B	7.29 ± 0.66	7.67 ± 0.42	7.53 ± 0.31	7.61 ± 0.48	7.89 ± 0.41	7.60 ± 0.35
Phi-4 Mini	6.58 ± 0.94	6.95 ± 1.27	6.27 ± 1.55	6.21 ± 1.54	6.88 ± 1.07	6.58 ± 1.18
Qwen3-32B	7.66 ± 0.57	8.10 ± 0.34	7.62 ± 0.39	7.83 ± 0.57	8.28 ± 0.31	7.90 ± 0.33
Mistral Small 24B	7.38 ± 0.54	7.70 ± 0.42	7.49 ± 0.31	7.57 ± 0.53	7.85 ± 0.42	7.60 ± 0.34
Agents' Room	8.15 ± 0.38	8.15 ± 0.37	7.97 ± 0.37	8.40 ± 0.34	8.57 ± 0.38	8.25 ± 0.26
WizardLM-30B	6.52 ± 0.80	7.07 ± 0.55	6.81 ± 0.55	6.53 ± 0.70	6.97 ± 0.62	6.78 ± 0.51
PLOTTWIST	8.11 ± 0.60	$\mathbf{8.42 \pm 0.41}$	$\mathbf{8.14 \pm 0.33}$	$\mathbf{8.51 \pm 0.44}$	8.61 ± 0.32	$\mathbf{8.36 \pm 0.35}$

Table 1: PlotTwist Performance Evaluation. Results are reported as mean $\pm$ standard deviation over 160 test premises using the Cross-Family Jury Agentic Evaluation framework. The best and second-best results in each column are shown in bold and underlined, respectively.

Baseline	Mean	WR	95% CI	p_{Holm}
Claude Sonnet 4	8.35	**0.563**	[0.488, 0.638]	0.13
Agents' Room	8.25	**0.734**	[0.663, 0.800]	4.4×10^{-9}
GPT-4.1	8.16	**0.906**	[0.863, 0.950]	2.0×10^{-27}
Gemini 2.0 Flash	8.11	**0.938**	[0.900, 0.975]	1.3×10^{-32}
Qwen-3-32B	7.90	**0.987**	[0.969, 1.000]	2.1×10^{-43}
Qwen-2.5-14B	7.62	**0.950**	[0.913, 0.981]	6.4×10^{-35}
Mistral Small 24B	7.60	**0.994**	[0.981, 1.000]	2.4×10^{-45}
DeepSeek-R1-14B	7.60	**0.994**	[0.981, 1.000]	2.4×10^{-45}
Llama-3.3-70B	7.18	**0.988**	[0.969, 1.000]	1.2×10^{-43}
WizardLM-30B	6.78	**0.994**	[0.981, 1.000]	2.4×10^{-45}
Phi-4 Mini	6.58	**0.994**	[0.981, 1.000]	2.4×10^{-45}

Table 2: Per-premise paired win rates of PlotTwist against each baseline under the Cross-Family Jury framework (ties counted as 0.5), with premise-level bootstrap 95% CIs and Holm-corrected exact sign tests. Mean: the baseline's overall mean from Table 1.

fore conclude only that PlotTwist is competitive with Claude Sonnet 4.

Ablations. The jury results show that PlotTwist's gains are not attributable to model scale, architecture, or generation paradigm. It outperforms similarly sized instruction-tuned models (Qwen2.5-14B: 7.62; DeepSeek-R1 14B: 7.60), the dense Qwen3-32B baseline (7.90) despite using roughly one-tenth as many active parameters per token (win rate 0.99), and the multi-agent Agents' Room framework (8.25), achieving a higher overall mean (8.36) with a win rate of 0.73 using a single model and inference pass. To isolate the effect of preference optimization, we evaluate the base Qwen-3-30B-A3B model before and after DPO using the development evaluator, observing an overall score increase from 8.03 to 8.81 (+0.78) after alignment on 160 high-confidence preference pairs.

Reliability. Raw inter-juror agreement is moderate (Krippendorff's $\alpha = 0.41$ (Krippendorff, 2004)), largely reflecting calibration differences rather than ranking disagreement. Consistent with this, Gwet's AC2 is 0.89 (Gwet, 2008), recentering juror scores increases α to 0.64, and system rankings are highly consistent across jurors (mean pairwise Spearman 0.97, Kendall's $W = 0.98$). Accounting for correlated judgments, the panel provides an effective sample size of approximately 1.3 independent judges rather than five (Kohli, 2026), motivating the per-judge and leave-one-judge-out analyses.

Deliberation and final judge. To probe shared error, score disagreements of at least two points between any pair of jurors are resolved through two round of structured deliberation, followed by adjudication from a held-out sixth-family judge, Llama-3.3-70B. Neither step alters the conclusions; accordingly, we report the independent first-round panel throughout. The system ranking is likewise robust to removing same-family judges, enforcing grounded evaluation, and all 113 analysis specifications (Appendix G).

6 Conclusion

We presented PLOTTWIST, a modular framework for premise-conditioned plot generation with SLMs. By combining aspect-rating reward modeling, preference optimization, and an independent cross-family jury agentic evaluation, PLOTTWIST enables efficient, high-quality plot generation while maintaining reliable and interpretable assessment of narrative quality. Extensive experiments demonstrate that PlotTwist consistently outperforms strong baselines despite using only 3B active parameters. Together, these findings un-

derscore the value of structured preference-based alignment as a scalable and effective alternative to brute-force model scaling for creative text generation with limited-capacity language models.

Limitations

As is standard in recent work on open-ended generation (Zheng et al., 2023), our assessment of plot quality relies on LLM judges rather than expert annotation. The jury protocol is designed to address the known risks of this choice: judges are drawn from model families disjoint from the generator, every score must cite verifiable evidence, and reliability is reported alongside the results, with the reward model additionally validated against human-derived labels. LLM judges nevertheless agree only moderately with expert readers on creative writing (Chakrabarty et al., 2024; Fein et al., 2025), and two caveats are quantified in Section 5.4: the comparison with Claude Sonnet 4 remains statistically unresolved at 160 premises, and correlated juror errors reduce the number of effectively independent opinions the panel provides. Our experiments are confined to English, synopsis-style plots scored on five craft-oriented dimensions, and the reward and preference data are constructed offline with the aid of larger models; extending the framework to other narrative forms, languages, and notions of quality such as novelty is a natural direction for future work.

References

Anthropic. 2025. Claude 4 model card and system safety. https://www.anthropic.com/news/claude-4.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. 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*, pages 1–34.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2024. ChatEval: Towards better LLM-based evaluators through multi-agent debate. In *International Conference on Learning Representations*.

Google DeepMind. 2024. Introducing gemini 2.0: Our new ai model for the agentic era. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Aladin Djuhera, Farhan Ahmed, Swanand Kadhe, Syed Zawad, Heiko Ludwig, and Holger Boche. 2026. When data is the algorithm: A systematic study and curation of preference optimization datasets. In *International Conference on Learning Representations*, volume 2026, pages 150085–150130.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2024. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. *Preprint*, arXiv:2404.04475.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 889–898.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

Daniel Fein, Sebastian Russo, Violet Xiang, Kabir Jolly, Rafael Rafailov, and Nick Haber. 2025. LitBench: A benchmark and dataset for reliable evaluation of creative writing. *Preprint*, arXiv:2507.00769.

Alexander Gurung and Mirella Lapata. 2025. Learning to reason for long-form story generation. *arXiv preprint arXiv:2503.22828*.

Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19.

Patrick Colm Hogan. 2011. *Affective narratology: The emotional structure of stories*. U of Nebraska Press.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

F. Huot, R. K. Amplayo, J. Palomaki, A. S. Jakobovits, E. Clark, and M. Lapata. 2024. Agents' room: Narrative generation through multi-step collaboration. *arXiv preprint arXiv:2410.02603*.

Shufan Jiang, Sizhou Chen, Chios Chen, Chi Zhang, Xiao-Lei Zhang, and Xuelong Li. 2025. HAMLET: A hierarchical and adaptive multi-agent framework for live embodied theatrics. *Preprint*, arXiv:2507.15518.

Sungeun Kim and Dongsuk Oh. 2025. Evaluating creativity: Can llms be good evaluators in creative writing tasks? *Applied Sciences*, 15(6):2971.

Leslie Kish. 1965. *Survey Sampling*. John Wiley & Sons, New York.

Guneet Kohli. 2026. Nine judges, two effective votes: Correlated errors undermine llm evaluation panels. *Preprint*, arXiv:2605.29800.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*, 2 edition. Sage Publications, Thousand Oaks, CA.

Daniël Lakens. 2017. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362.

Microsoft. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

NVIDIA, :, Aaron Blakeman, Aaron Grattafiori, Aarti Basant, Abhibha Gupta, Abhinav Khattar, Adi Renduchintala, Aditya Vavre, Akanksha Shukla, Akhiad Bercovich, Aleksander Ficek, Aleksandr Shaposhnikov, Alex Kondratenko, Alexander Bukharin, Alexandre Milesi, Ali Taghibakhshi, Alisa Liu, Amelia Barton, and 340 others. 2025. Nvidia nemotron 3: Efficient and open intelligence. *Preprint*, arXiv:2512.20856.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

OpenAI. 2025. Gpt-4.1 system card. https://openai.com/index/gpt-4-1/.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Uri Simonsohn, Joseph P. Simmons, and Leif D. Nelson. 2020. Specification curve analysis. *Nature Human Behaviour*, 4(11):1208–1214.

Jiuding Sun, Chantal Shaib, and Byron C Wallace. 2023. Evaluating the zero-shot robustness of instruction-tuned language models. *arXiv preprint arXiv:2306.11270*.

5 Team, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, and 152 others. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *Preprint*, arXiv:2508.06471.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, and 282 others. 2026. Gemma 4 technical report. *Preprint*, arXiv:2607.02770.

Mistral AI Team. 2025a. Mistral small 3 (2501) release. https://mistral.ai/news/mistral-small-3/.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Qwen Team. 2025b. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Maria Teleki, Vedangi Bengali, Xiangjue Dong, Sai Tejas Janjur, Haoran Liu, Tian Liu, Cong Wang, Ting Liu, Yin Zhang, Frank Shipman, et al. 2025. A survey on llms for story generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 13954–13966.

TheBloke, Eric Hartford, and Kaiokendev. 2023. Wizardlm-uncensored-supercot-storytelling-30b-gptq. https://huggingface.co/TheBloke/WizardLM-Uncensored-SuperCOT-StoryTelling-30B-GPTQ.

Pat Verga, Sebastian Hofstätter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. *Preprint*, arXiv:2404.18796.

Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. 2024. β-dpo: Direct preference optimization with dynamic β. *Advances in Neural Information Processing Systems*, 37:129944–129966.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7378–7385.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V. Chawla, and Xiangliang Zhang. 2024. Justice or prejudice? quantifying biases in LLM-as-a-judge. *Preprint*, arXiv:2410.02736.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Mingzhe Zheng, Dingjie Song, Guanyu Zhou, Jun You, Jiahao Zhan, Xuran Ma, Xinyuan Song, Ser-Nam Lim, Qifeng Chen, and Harry Yang. 2025. Cml-bench: A framework for evaluating and enhancing llm-powered movie scripts generation. *Preprint*, arXiv:2510.06231.

A Reward Model Prompts

Aspect 1: Narrative Coherence

Positive Prompt (r_a^+)

You are a professional movie critic whose **only** output must be a **single JSON object** with exactly one integer field (0–10):

`Narrative Coherence:` **Field Definition (Positive Focus):**

Narrative clarity, logical plot progression, coherent world-building, strong cause–effect relationships, and well-integrated subplots.

Strict output rules:

1. Output only a valid JSON object.
2. Include only `Narrative_Coherence`.
3. Integer value from 0 to 10.
4. Score generously.

`### MoviePlot: { }`

Negative Prompt (r_a^-)

You are a professional movie critic whose **only** output must be a **single JSON object** with exactly one integer field (0–10):

`Narrative Coherence:` **Field Definition (NegativeFocus):**

Confusing storytelling, plot holes, inconsistent world-building, disconnected subplots, or illogical character decisions.

Strict output rules:

1. Output only a valid JSON object.
2. Include only `Narrative_Coherence`.
3. Integer value from 0 to 10.
4. 0 = no issues, 10 = severe issues.

`### MoviePlot: { }`

Aspect 2: Emotional Turning Points

Positive Prompt (r_e^+)

You are a professional movie critic whose **only** output must be a **single JSON object** with exactly one integer field (0–10):

`Emotional Turning Points:` **Field Definition (Positive Focus):**

Powerful emotional moments, effective turning points, meaningful revelations, and emotionally satisfying narrative shifts.

Strict output rules:

1. Output only JSON.
2. Include only `Emotions_Turning_Points`.
3. Integer 0–10.
4. Score generously.

MoviePlot: { } | ### Review:

Negative Prompt (r_e^-)

You are a professional movie critic whose **only** output must be a **single JSON object** with exactly one integer field (0–10):

`Emotional Turning Points:` **Field Definition (Negative Focus):**

Flat emotional arcs, forced turning points, unearned twists, or moments that fail to engage the audience.

Strict output rules:

1. Output only JSON.
2. Include only `Emotions_Turning_Points`.
3. Integer 0–10.
4. 0 = no issues, 10 = severe issues.

MoviePlot: { } | ### Review:

Aspect 3: Tone Consistency

Positive Prompt (r_t^+)

You are a professional movie critic whose **only** output must be a **single JSON object** with exactly one integer field (0–10):

`Tone Consistency:` **Field Definition (Positive Focus):**

Successful maintenance of mood, atmosphere, and stylistic coherence throughout the story. Effective emotional consistency, well-maintained genre conventions, and smooth transitions between story beats. Intentional tonal shifts are rewarded when they serve the narrative purpose.

Strict output rules:

1. Output only a valid JSON object.
2. Include only `Tone_Consistency`.
3. Integer value from 0 to 10.
4. Score generously.

MoviePlot: { } | ### Review:

Negative Prompt (r_t^-)

You are a professional movie critic whose **only** output must be a **single JSON object** with exactly one integer field (0–10):

`Tone Consistency:` **Field Definition (Negative Focus):**

Jarring mood shifts, inconsistent atmosphere, conflicting stylistic elements, genre incoherence, or awkward tonal transitions that disrupt immersion or emotional continuity.

Strict output rules:

1. Output only a valid JSON object.
2. Include only `Tone_Consistency`.
3. Integer value from 0 to 10.
4. 0 = no issues, 10 = severe issues.

MoviePlot: { } | ### Review:

Aspect 4: Character Development

Positive Prompt (r_c^+)

You are a professional movie critic whose **only** output must be a **single JSON object** with exactly one integer field (0–10):

`Character Development:` **Field Definition (Positive Focus):**

Compelling character arcs, meaningful growth, clear motivations, well-developed relationships, authentic character voices, and satisfying character journeys. Emphasis is placed on characters who evolve, learn, or change meaningfully over the course of the story.

Strict output rules:

1. Output only a valid JSON object.
2. Include only `Character_Development`.
3. Integer value from 0 to 10.
4. Score generously.

`### MoviePlot: { } | ### Review:`

Negative Prompt (r_c^-)

You are a professional movie critic whose **only** output must be a **single JSON object** with exactly one integer field (0–10):

`Character Development:` **Field Definition (Negative Focus):**

Weak or static character arcs, lack of growth, unclear motivations, poorly developed relationships, inconsistent character voices, or unsatisfying character journeys. Emphasis is placed on characters who remain static, act illogically, or fail to develop meaningfully.

Strict output rules:

1. Output only a valid JSON object.
2. Include only `Character_Development`.
3. Integer value from 0 to 10.
4. 0 = no issues, 10 = severe issues.

`### MoviePlot: { } | ### Review:`

Aspect 5: Pacing

Positive Prompt (r_p^+)

You are a professional movie critic whose **only** output must be a **single JSON object** with exactly one integer field (0–10):

`Pacing:` **Field Definition (Positive Focus):**

Effective narrative rhythm, well-balanced scene progression, appropriate timing of plot events, and smooth transitions that maintain momentum and audience engagement. Emphasis is placed on pacing that supports tension, emotional beats, and story clarity.

Strict output rules:

1. Output only a valid JSON object.
2. Include only `Pacing`.
3. Integer value from 0 to 10.
4. Score generously.

MoviePlot: { } | ### Review:

Negative Prompt (r_p^-)

You are a professional movie critic whose **only** output must be a **single JSON object** with exactly one integer field (0–10):

`Pacing:` **Field Definition (Negative Focus):**

Uneven or inconsistent pacing, excessive slowdowns or rushed segments, poorly timed plot events, unnecessary filler scenes, or abrupt transitions that disrupt narrative flow or emotional impact.

Strict output rules:

1. Output only a valid JSON object.
2. Include only `Pacing`.
3. Integer value from 0 to 10.
4. 0 = no issues, 10 = severe issues.

MoviePlot: { } | ### Review:

B DPO Data Curation

Figure 3 illustrates the preference-pair curation procedure described in Section 4.2.

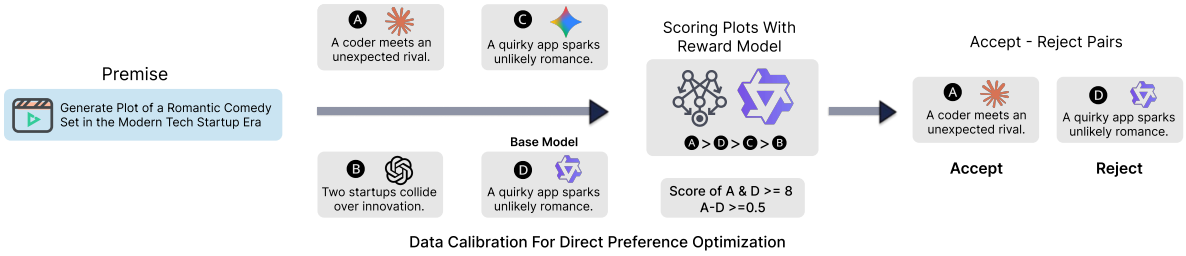

Figure 3: Dataset curation for performing DPO

C Agentic Evaluation System Prompts

System Prompt 1: Narrative Coherence Evaluation

Task Overview

Evaluate a movie plot's narrative structure and logical consistency using a 10-criteria framework. Assign precise numerical scores reflecting coherence quality.

Evaluation Methodology

Each criterion is scored from 0–1 (increments of 0.1 allowed). Scores are summed for a total out of 10.

Scoring Criteria

Plot Structure and Logic (4 points)

1. Plot Progression
Logical beginning–middle–end flow.

2. Causal Connectivity
Events arise naturally from prior actions.

3. Plot Integrity
No plot holes or contradictions.

4. Conflict Focus
A sustained central conflict drives the story.

Character Integration (3 points)

5. Protagonist Consistency 6. Supporting Character Function 7. Resolution Authenticity

Narrative Flow and Unity (3 points)

8. Pacing Appropriateness 9. Thematic Integration 10. Tonal Consistency

Output Format

```
1. Plot Progression: X.X
2. Causal Connectivity: X.X
3. Plot Integrity: X.X
4. Conflict Focus: X.X
5. Protagonist Consistency: X.X
6. Supporting Character Function: X.X
7. Resolution Authenticity: X.X
8. Pacing Appropriateness: X.X
9. Thematic Integration: X.X
10. Tonal Consistency: X.X
TOTAL: X.X/10
```

System Prompt 2: Emotional Turning Point Evaluation

Task Overview

Identify and evaluate the primary emotional turning point of the narrative using a 10-criteria framework focused on emotional impact and character change.

Scoring Criteria

Conflict & Character Foundation (4 points)

1. Conflict Resolution
Addresses or reframes central conflict.

2. Character Believability
Emotion aligns with established arc.

3. Character Transformation
Meaningful internal change.

4. Emotional Satisfaction
Emotionally resonant payoff.

Narrative Construction (3 points)

5. Narrative Causality 6. Thematic Crystallization 7. Relationship Impact

Technical & Structural Elements (3 points)

8. Cinematic Execution 9. Structural Necessity 10. Audience Alignment

Output Format

```
1. Conflict Resolution: X.X
2. Character Believability: X.X
3. Character Transformation: X.X
4. Emotional Satisfaction: X.X
5. Narrative Causality: X.X
6. Thematic Crystallization: X.X
7. Relationship Impact: X.X
8. Cinematic Execution: X.X
9. Structural Necessity: X.X
10. Audience Alignment: X.X
TOTAL: X.X/10
```

System Prompt 3: Character Development Evaluation

Task Overview

Evaluate protagonist character development using a 10-criteria framework assessing motivation, arc progression, and narrative function.

Scoring Criteria

Core Character Elements (4 points)

1. Motivation Clarity
Clear goals and desires.

2. Behavioral Consistency
Actions align with personality.

3. Character Arc
Believable transformation.

4. Psychological Depth
Emotional and psychological complexity.

Character Foundation (3 points)

5. Backstory Integration 6. Audience Connection 7. Character Distinctiveness

Narrative Function (3 points)

8. Relationship Dynamics 9. Plot Agency 10. Thematic Alignment

Output Format

```
1. Motivation Clarity: X.X              2. Behavioral Consistency: X.X
3. Character Arc: X.X                   4. Psychological Depth: X.X
5. Backstory Integration: X.X           6. Audience Connection: X.X
7. Character Distinctiveness: X.X       8. Relationship Dynamics: X.X
9. Plot Agency: X.X                     10. Thematic Alignment: X.X
TOTAL: X.X/10
```

System Prompt 4: Pacing Analysis Evaluation

Task Overview

Assess narrative pacing using a 10-criteria framework measuring rhythm, momentum, and emotional timing.

Scoring Criteria

1. Premise Establishment Speed	**2. Structural Foundation**
3. Pacing Consistency	**4. Event Frequency**
5. Scene Purposefulness	**6. Tension Management**
7. Transition Quality	**8. Emotional Beat Timing**
9. Climax Timing	**10. Genre–Tone Alignment**

Output Format

```
1. Premise Establishment Speed: X.X     2. Structural Foundation: X.X
3. Pacing Consistency: X.X              4. Event Frequency: X.X
5. Scene Purposefulness: X.X            6. Tension Management: X.X
7. Transition Quality: X.X              8. Emotional Beat Timing: X.X
9. Climax Timing: X.X                   10. Genre-Tone Alignment: X.X
TOTAL: X.X/10
```

System Prompt 5: Tone Consistency Evaluation

Task Overview

Evaluate tonal coherence using a 10-criteria framework assessing atmosphere, stylistic unity, and emotional continuity.

Scoring Criteria

1. Initial Atmosphere Establishment	**2. Scene-to-Scene Consistency**
3. Tonal Relief Integration	**4. Earned Tone Shifts**
5. Dialogue Style Consistency	**6. Visual Reinforcement**
7. Stakes Alignment	**8. Comedy/Drama Balance**
9. Ending Consistency	**10. Motif and Symbol Unity**

Output Format

```
1. Initial Atmosphere Establishment: X.X   2. Scene-to-Scene Consistency: X.X
3. Tonal Relief Integration: X.X           4. Earned Tone Shifts: X.X
5. Dialogue Style Consistency: X.X         6. Visual Reinforcement: X.X
7. Stakes Alignment: X.X                   8. Comedy/Drama Balance: X.X
9. Ending Consistency: X.X                 10. Motif and Symbol Unity: X.X
TOTAL: X.X/10
```

System Prompt 6: Grounded Output Contract (Independent Scoring)

Applied in

Round independent scoring. Appended verbatim to each of the five rubrics above, so every juror returns one structured, evidence-grounded record per NQD.

Instruction (verbatim)

Return ONLY a single JSON object and nothing else, with the exact shape below. List every rubric criterion in order. Every `evidence_quote` must be copied verbatim from the plot text, not paraphrased. Do not add any text outside the JSON object.

Output schema

```
{"nqd": "<the nqd key>",
 "criteria": [{"name": "<criterion name from the rubric above>",
               "score_0_1": <number between 0 and 1>,
               "evidence_quote": "<a span copied verbatim from the plot>",
               "rationale": "<one sentence>"}],
 "total_0_10": <sum of the criterion scores, between 0 and 10>,
 "overall_rationale": "<two sentences>"}
```

System Prompt 7: Deliberation Instruction (Juror Revision)

Applied in

Both deliberation rounds, on cells the independent panel contests (a spread of at least two points between juror scores). Appended to the same rubric; each juror sees its own prior assessment and the anonymized, order-shuffled arguments and cited spans of the other jurors, with peer scores hidden to prevent vote-matching.

Instruction (verbatim)

You already assessed this plot on the dimension defined by the rubric above. Below the plot you are shown (a) your own prior assessment and (b) the anonymized arguments of the other independent jurors, with the verbatim spans they cited. Reconsider your assessment in light of their EVIDENCE. Change your score only if a peer cites stronger grounded evidence than you did; do NOT change it to match a majority or because a peer sounds confident. If your original judgement still holds, keep your score and say why. Stay blind to anything outside the plot text (there is no premise, genre label, or reference answer).

Output schema

```
{"nqd": "<the nqd key>",
 "revised_total_0_10": <0 to 10>,
 "changed": <true|false>,
 "criteria": [{"name": "<criterion>",
               "score_0_1": <0 to 1>,
               "evidence_quote": "<verbatim span from the plot>",
               "rationale": "<one sentence>"}],
 "rebuttal": "<2 to 3 sentences: what you kept or moved, and which grounded peer argument did or did
     not change your mind>"}
```

System Prompt 8: Consolidation Instruction (Judge LLM)

Applied in

Final consolidation by the held-out Judge LLM (Llama-3.3-70B) after the two deliberation rounds, once per contested NQD. The judge reads the plot, the anonymized independent assessments, and both rounds of revisions before issuing one consolidated score.

Instruction (verbatim)

You are the presiding judge. Below the plot you are shown the anonymized, grounded assessments of several independent jurors on the dimension defined by the rubric above, each with the verbatim spans it cited. Produce ONE final score for this dimension by WEIGHING the arguments on how well each is grounded in evidence actually present in the plot, not by averaging the jurors and not by counting how many agree. Discount any claim whose cited span does not support it or is absent from the plot. Name the main point of disagreement and how the evidence resolves it. Stay blind to anything outside the plot text.

Output schema

```
{"nqd": "<the nqd key>",
 "final_total_0_10": <0 to 10>,
 "key_disagreement": "<the main split and how you resolved it>",
 "synthesis_rationale": "<3 to 4 sentences citing the decisive grounded evidence>",
 "criteria": [{"name": "<criterion>",
               "score_0_1": <0 to 1>,
               "evidence_quote": "<verbatim span>",
               "rationale": "<one sentence>"}]}
```

D Plot Generation with PlotTwist

Plot Generation with PlotTwist

Task Description

The following example demonstrates long-form creative movie plot generation under the *PlotTwist* setting. The task emphasizes narrative coherence, concealed truths, emotional stakes, and morally consequential decision-making, serving as a qualitative illustration of model behavior.

Premise

Generate a movie plot that follows a dedicated detective as he investigates a murder that implicates his beloved, forcing him to navigate a web of concealed truths, blackmail, and the desperate lengths one will go to protect a dark secret.

PlotTwist Generated Plot

Detective Aris Thorne, a man whose sharp eyes and even sharper instincts have solved countless cases but whose heart remains scarred by the unsolved murder of his wife Elara five years prior, finds his world shattered when the body of City Councilman Silas Croft—a powerful figure known for his ruthless political maneuvering and rumored ties to the city's underbelly—is discovered in his opulent penthouse with a single, precise stab wound to the heart.

The evidence points damningly toward Lena Voss, Aris's former lover and now a devoted social worker fighting for the city's forgotten children. Witnesses report a heated argument between Lena and Croft hours before his death, and a fragmented security recording captures her leaving the building, her face pale with fear. Aris, torn between duty and lingering love, cannot believe she is capable of murder—yet the physical evidence is overwhelming. A rare antique switchblade, identical to one Aris gifted Lena years earlier, is found clutched in Croft's hand. Even more troubling is a cryptic encrypted file on Croft's laptop, labeled `Project Chimera`, implicating Lena in a scheme involving missing children.

As Aris investigates further, the case reveals a labyrinth of concealed truths. Croft was not merely corrupt; he was the linchpin of an extensive blackmail network exploiting vulnerable families. The encrypted files reveal that Croft had been threatening Lena with exposure—not only of her past, but of evidence tied directly to Elara's death. Aris had long believed Elara's death was a tragic accident and had quietly suppressed details to protect her reputation.

The blackmail was never solely about Lena—it was about Aris himself. Lena ultimately murdered Croft in desperation to stop him from exposing the truth, protect the children he was trafficking, and prevent Aris's buried past from destroying them both. Aris finds himself ensnared between the city's elite, an ambitious District Attorney eager to frame Lena, Croft's remaining associates intent on silencing her, and his own conscience, which recognizes that Lena's act was born of sacrifice rather than malice.

Aris uncovers the final truth: Elara had been investigating Croft's operations as a journalist and was killed in a staged car accident when she came too close. In a climactic confrontation at the rain-lashed docks, Aris dismantles the remnants of Croft's network, but the true reckoning is internal. Choosing love over career, he confesses to the murder to protect Lena, sacrificing his badge and freedom.

As rain washes over the city, Aris walks away from everything he once was, the weight of concealed truths finally lifted. What remains is the memory of Lena—not as a suspect, but as the woman he loved—and the understanding that truth and love, when intertwined, can demand a devastating cost.

E Models and Their Roles in the PlotTwist Framework

Overview of all language models employed across training, evaluation, and baseline comparison stages of the PlotTwist system.

PlotTwist Core Components

Qwen-3-30B-A3B (MoE) — PlotTwist Plot Generator

MoE backbone with **3B active parameters**; preference-aligned via **Direct Preference Optimization (DPO)** to generate high-quality premise-conditioned plots. Only 3B parameters are active per token despite a total count of 30B, classifying it as an SLM under the paper's definition.

Qwen-3-32B (4-bit) — PlotTwist Aspect Rating Reward Model

Fine-tuned via **regression-aware SFT** using a weighted combination of **cross-entropy loss** and **Huber loss** to predict continuous aspect-level narrative quality scores across all five NQDs.

Cross-Family Jury — PlotTwist Agentic Evaluation Module

Independent post-hoc evaluation module operating separately from the training pipeline. A panel of five open-weight **Juror LLMs** drawn from five model families disjoint from the Qwen family of the generator and reward model: **GPT-OSS-120B**, **Gemma-4-31B**, **Mistral-Medium-3.5-128B**, **Nemotron-3-Super-120B-A12B**, and **GLM-4.6V**. Each juror independently scores all five **NQDs** using ten-criterion, weakness-focused rubrics with verbatim evidence grounding. Score disagreements of at least two points are resolved through **two rounds of structured deliberation**, consolidated by a held-out **Judge LLM**, **Llama-3.3-70B**, drawn from a sixth model family disjoint from both the panel and the generator.

Ensemble Models for Positive-Negative Aspect Rating

Qwen-2.5-7B **Llama-3.3-70B** **Llama-3.1-8B** **DeepSeek-14B** **Gemma-27B**

Five-model ensemble used to generate **synthetic aspect-level ratings** via positive–negative prompting. Each model outputs both a positive score $r^{+}_{a,m}(p)$ and a negative score $r^{-}_{a,m}(p)$ per aspect, aggregated as:

$$r_a(p) = \sum_m \left(r^{+}_{a,m}(p) - r^{-}_{a,m}(p)\right)$$

Model diversity across the ensemble mitigates individual model bias in the rating construction process. **Gemma-27B** additionally generates premise descriptions from movie plots for DPO dataset curation.

DPO Candidate Plot Generators

GPT-4.1

Frontier model used to generate candidate plots per premise for reward-model scoring and **DPO preference pair construction**. Retained as accepted plot only when it achieves the highest reward score (≥ 8) and outperforms the next-best model by a margin of at least 0.5.

Claude Sonnet 4

Frontier model used to generate candidate plots per premise for reward-model scoring and **DPO preference pair construction**. Retained as accepted plot only when it achieves the highest reward score (≥ 8) and outperforms the next-best model by a margin of at least 0.5.

Gemini 2.0 Flash

Frontier model used to generate candidate plots per premise for reward-model scoring and **DPO preference pair construction**. Retained as accepted plot only when it achieves the highest reward score (≥ 8) and outperforms the next-best model by a margin of at least 0.5.

Baselines: Model Scale

Llama-3.3-70B

Large open-weight baseline. Tests performance at high parameter count **without task-specific alignment**, providing an upper-bound reference for scale alone.

Qwen3-32B (Dense)

Dense baseline from the same model family as the PlotTwist generator. Used to **isolate performance gains attributable to DPO alignment** from those due to the MoE architecture.

Qwen2.5-7B-Instruct **Qwen2.5-14B-Instruct**

Small instruction-tuned baselines used for **scale ablation**, confirming that PlotTwist's gains arise from methodology rather than model size.

Baselines: Architectural Design

DeepSeek-R1 14B

MoE and reasoning-oriented baseline. Contrasts **sparse activation without preference alignment**, isolating the contribution of DPO from that of expert routing.

Phi-4 Mini Instruct

Compact reasoning-optimized model baseline. Evaluates whether **small reasoning-focused architectures** can match structured preference-aligned generation.

Mistral Small 2501 24B

Reasoning-optimized baseline assessing **structured temporal progression** and narrative coherence in mid-scale models.

Baselines: Generation Paradigm

Agents' Room

Multi-agent collaborative narrative generation baseline. Decomposes the writing process into specialized planning and writing agents communicating via a shared scratchpad to maintain long-term coherence across the plot.

WizardLM-StoryTelling-30B

Monolithic instruction-tuning baseline. Relies on the Evol-Instruct methodology to embed narrative constraints directly into model weights rather than resolving them through external orchestration or preference alignment.

F Quality-Stratified Analysis: Per-Stratum Results

This appendix provides the detailed statistics for the quality-stratified analysis summarized in Section 5.2. For each quality stratum, Table 12 reports mean jury scores for the original and PlotTwist-generated plots, bootstrap 95% confidence intervals over paired differences, paired effect sizes (Cohen's d_z), the probability that a generated plot outperforms its paired original, and Welch's t-tests for secondary statistical validation.

Excellent Category (IMDb > 8). Generated plots exhibit only modest improvements over already strong originals. Character development ($+0.50$), narrative coherence ($+0.40$), and tone consistency ($+0.33$) show the clearest gains, whereas emotional turning points improve only marginally ($+0.32$) and

Stratum	NQD	Orig.	Gen.	Δ	95% CI	d_z	$P(\text{gen} > \text{orig})$	Welch p
Excellent	Narrative coherence	8.24	8.64	$+0.40$	$[0.13, 0.72]$	0.41	0.70	0.012
($n = 40$)	Emotional turning points	8.37	8.69	$+0.32$	$[0.02, 0.72]$	0.28	0.55	0.088
	Character development	7.72	8.22	$+0.50$	$[0.20, 0.88]$	0.44	0.75	0.021
	Pacing	8.18	8.16	-0.02	$[-0.21, 0.19]$	-0.04	0.38	0.828
	Tone consistency	8.18	8.51	$+0.33$	$[0.17, 0.50]$	0.62	0.75	4.47×10^{-4}
	Overall	**8.14**	**8.45**	$\mathbf{+0.31}$	$[\mathbf{0.09}, \mathbf{0.57}]$	**0.38**	**0.65**	**0.023**
Good	Narrative coherence	7.70	8.55	$+0.85$	$[0.64, 1.08]$	1.18	0.85	3.28×10^{-8}
($n = 40$)	Emotional turning points	8.19	8.65	$+0.46$	$[0.34, 0.59]$	1.11	0.85	2.88×10^{-7}
	Character development	7.36	8.22	$+0.86$	$[0.66, 1.08]$	1.24	0.88	1.05×10^{-8}
	Pacing	7.68	8.12	$+0.43$	$[0.28, 0.59]$	0.87	0.80	1.50×10^{-6}
	Tone consistency	7.83	8.46	$+0.64$	$[0.47, 0.81]$	1.16	0.95	4.15×10^{-10}
	Overall	**7.75**	**8.40**	$\mathbf{+0.65}$	$[\mathbf{0.51}, \mathbf{0.79}]$	**1.40**	**0.93**	$\mathbf{2.91 \times 10^{-11}}$
Mid	Narrative coherence	7.20	8.54	$+1.34$	$[1.08, 1.63]$	1.52	0.98	5.39×10^{-10}
($n = 40$)	Emotional turning points	8.08	8.65	$+0.57$	$[0.43, 0.73]$	1.17	0.95	1.92×10^{-7}
	Character development	7.03	8.18	$+1.15$	$[0.91, 1.39]$	1.44	0.93	1.80×10^{-8}
	Pacing	7.44	8.10	$+0.67$	$[0.48, 0.89]$	1.01	0.85	6.48×10^{-8}
	Tone consistency	7.25	8.40	$+1.15$	$[0.96, 1.35]$	1.82	1.00	1.59×10^{-13}
	Overall	**7.40**	**8.37**	$\mathbf{+0.98}$	$[\mathbf{0.81}, \mathbf{1.17}]$	**1.68**	**0.98**	$\mathbf{2.11 \times 10^{-11}}$
Low	Narrative coherence	6.67	8.50	$+1.82$	$[1.52, 2.16]$	1.76	0.98	1.04×10^{-13}
($n = 40$)	Emotional turning points	7.68	8.58	$+0.90$	$[0.69, 1.13]$	1.26	0.98	3.39×10^{-9}
	Character development	6.59	7.84	$+1.25$	$[1.00, 1.53]$	1.46	0.98	2.04×10^{-9}
	Pacing	7.15	8.15	$+1.00$	$[0.83, 1.20]$	1.63	1.00	1.62×10^{-12}
	Tone consistency	6.94	8.35	$+1.40$	$[1.19, 1.61]$	2.05	0.98	7.28×10^{-16}
	Overall	**7.01**	**8.28**	$\mathbf{+1.28}$	$[\mathbf{1.09}, \mathbf{1.48}]$	**2.01**	**1.00**	$\mathbf{7.95 \times 10^{-15}}$

Table 12: Per-stratum jury statistics for original and PlotTwist-generated plots. Δ denotes the mean paired difference between generated and original plots, with bootstrap 95% confidence intervals. The statistic d_z is the paired Cohen's effect size, and $P(\text{gen} > \text{orig})$ is the proportion of films for which the generated plot receives a higher score than its paired original.

pacing remains unchanged (-0.02). These results indicate that PlotTwist primarily performs conservative refinement when the original narrative quality is already high.

Good Category ($7 <$ IMDb ≤ 8). Generated plots improve consistently across all NQDs, with the largest gains in character development ($+0.86$), narrative coherence ($+0.85$), and tone consistency ($+0.64$). All dimensions exhibit large effect sizes and statistically significant improvements, indicating systematic enhancement of narratives with solid foundations but remaining structural limitations.

Mid Category ($6 <$ IMDb ≤ 7). Mid-quality narratives benefit substantially from PlotTwist. Narrative coherence ($+1.34$), tone consistency ($+1.15$), and character development ($+1.15$) show the largest improvements, accompanied by uniformly large effect sizes and high dominance probabilities. This quality range represents the regime where PlotTwist performs the most effective narrative restructuring.

Low Category (IMDb ≤ 6). Generated plots substantially outperform the originals across every NQD. The largest gains occur in narrative coherence ($+1.82$) and tone consistency ($+1.40$), followed by character development ($+1.25$). Generated plots outperform their paired originals on at least 97.5% of films across every dimension, indicating near-complete narrative regeneration for weak source plots.

Summary. Overall improvements increase monotonically across quality strata: $+0.31$ (Excellent), $+0.65$ (Good), $+0.98$ (Mid), and $+1.28$ (Low), demonstrating that PlotTwist adapts the magnitude of its intervention to the underlying quality of the source narrative rather than uniformly inflating evaluation scores.

G Jury Evaluation: Protocol, Win Rates, and Reliability

Composition and serving. Table 13 lists the jury. No juror shares a model family with another juror, with the plot generator, or with the single-judge Qwen3-32B evaluator used during development (Section 4.3). All five are open-weight models, served locally at pinned revisions and scored at temperature 0.1. Each run records every juror's model revision, quantization, serving engine, and decoding parameters in a manifest, so all scores can be regenerated.

Juror	Family	Precision
GPT-OSS-120B	OpenAI (open-weight)	MXFP4
Gemma-4-31B	Google	BF16
Mistral-Medium-3.5-128B	Mistral	Q6_K
Nemotron-3-Super-120B-A12B	NVIDIA	FP8
GLM-4.6V	Zhipu	FP8

Table 13: Jury composition. The five jurors span five model families, all disjoint from the Qwen family of the plot generator and of the single-judge development evaluator.

Grounded output contract. Each juror scores one NQD per call, using the corresponding rubric of Appendix C. The response is a structured JSON object giving, for each of the ten criteria, a score in $[0, 1]$, a one-sentence rationale, and the verbatim evidence span required by Section 5.4; the token-containment threshold of 0.92 tolerates minor quoting drift, and unverifiable citations are logged and reported rather than silently discarded. When a juror's stated total disagrees with the sum of its criterion scores by more than one point, we use the sum, so reported totals always reflect the cited evidence. Enforcing grounding at any verification threshold leaves the ranking of the twelve systems essentially unchanged (Spearman ≥ 0.99 against the full panel) and moves the two closest win rates by at most 0.04.

Per-judge win rates. Table 2 in Section 5.4 reports the panel-level win rates; here we break them down by judge. The GPT-4.1 comparison is unanimous at the level of individual jurors: each of the five prefers PlotTwist on a majority of premises, with per-judge win rates between 0.65 and 0.90, and removing any single juror leaves the panel estimate between 0.85 and 0.91, with every bootstrap interval excluding 0.5. The Claude Sonnet 4 comparison is contested: four jurors individually favor PlotTwist (win rates between 0.59 and 0.63), while Gemma-4-31B, the most conservative scorer on the panel, favors Claude (0.34). Removing any single juror keeps the estimate between 0.49 and 0.69, and no analysis choice we examined turns the comparison into a statistically significant loss.

Agreement by dimension. Table 14 breaks inter-juror agreement down by dimension. Krippendorff's α is lowest on the most subjective dimensions (emotional turning points, tone consistency), while the skew-robust Gwet's AC2 is high on every dimension, the pattern produced by judges who agree on order but differ in leniency. Recentering each judge's scores raises pooled α from 0.413 to 0.641, confirming that much of the nominal disagreement is a per-judge level offset. High agreement should not be read as five independent confirmations, however: the Kish effective judge count is 1.35 of 5 on raw scores, and 1.81 after removing system and premise effects (Kish, 1965; Kohli, 2026). The per-judge and leave-one-judge-out analyses in the previous paragraph are therefore the ones that carry evidential weight.

Same-family bias audit. To quantify the circularity that motivated the jury, a separate diagnostic run adds an in-family Qwen judge to the panel. Controlling for system identity and premise difficulty, judges score generators from their own family 0.82 points higher (premise-clustered bootstrap 95% CI $[0.80, 0.84]$), so same-family judging does inflate scores. PlotTwist does not benefit from this effect: its leniency-corrected same-family inflation is negative (-0.47), and removing all same-family cells from the panel leaves the ranking of the twelve systems unchanged (Spearman 1.0).

Specification-curve robustness. We recompute the two closest comparisons under every defensible combination of analysis choices (Simonsohn et al., 2020): juror subset (all five, or each juror removed), aggregation across jurors (mean, median, trimmed mean), score basis (overall mean or per-NQD majority vote), tie handling (ties as 0.5, dropped, or as losses), grounding enforcement, and length control (Dubois

NQD	Krippendorff's α	Gwet's AC2
Narrative coherence	0.504	0.889
Emotional turning points	0.280	0.827
Character development	0.405	0.863
Pacing	0.476	0.932
Tone consistency	0.366	0.911
Pooled	**0.413**	**0.887**

Table 14: Inter-juror agreement per NQD (interval-level Krippendorff's α; ordinal-weighted Gwet's AC2). The pooled AC2 95% CI is $[0.880, 0.893]$.

et al., 2024), for a total of 113 specifications per comparison. Against GPT-4.1, the win rate remains within $[0.80, 0.92]$ and every specification's 95% interval excludes 0.5. Against Claude Sonnet 4, the median win rate is 0.556, 94.7% of specifications lie at or above 0.5, and none yields a statistically significant loss.

Deliberation and final judge. Deliberation targets only the cells where the panel genuinely disagrees, defined as a spread of at least two points between scores. On each such cell, jurors are shown the anonymized, order-shuffled rationales and cited spans of their peers, with peer scores hidden to prevent vote-matching, and may revise their assessments across two rounds of structured discussion. A held-out final judge, Llama-3.3-70B, drawn from a sixth model family disjoint from both the panel and the generator, then reads the plot, the arguments, and the revised assessments, and issues its own verdict on each contested cell rather than counting votes (Chan et al., 2024); panel scores are retained on uncontested cells. Deliberation itself changes little: inter-juror agreement on overall scores moves only from $\alpha = 0.455$ to 0.458, two premise-level verdicts flip, and no baseline-level conclusion changes. The final judge reproduces the panel's conclusions, with win rates of 0.89 against GPT-4.1, 0.92 against Gemini 2.0 Flash, 0.74 against Agents' Room, and 0.57 against Claude Sonnet 4 ($p = 0.08$, inconclusive as in the main analysis). Although the judge shares a model family with the Llama-3.3-70B baseline, it prefers PlotTwist against that baseline on all but one premise, so its verdicts are not driven by family loyalty. Because deliberation neither materially raises agreement nor changes any conclusion, the consolidated post-deliberation scores and the independent round-one panel are effectively interchangeable.

H Computational Resources

All open-weight models used in this work were downloaded from the **Hugging Face Model Hub**[5] and executed on a dedicated compute cluster consisting of **4 × NVIDIA L40S GPUs**. The L40S is a high-performance data center GPU featuring 48 GB of GDDR6 memory per card, providing a total of **192 GB of aggregate GPU memory**, which was sufficient to accommodate the quantized and sparse model configurations employed throughout the framework.

Closed-source frontier models were **not** run on local infrastructure and were instead accessed exclusively via their respective commercial APIs.

Model	Execution	Usage in PlotTwist
PlotTwist Core Components — Local (4 × L40S)		
Qwen-3-30B-A3B (MoE)	Local	DPO fine-tuning and inference for PlotTwist Plot Generator
Qwen-3-32B (4-bit)	Local	SFT training of Aspect Rating Reward Model

[5] https://huggingface.co

Qwen-3-32B (16-bit)	Local	Inference for Agentic Evaluation module
Ensemble Rating Models — Local (4 × L40S)		
Qwen-2.5-7B	Local	Positive–negative aspect rating ensemble
Llama-3.1-8B	Local	Positive–negative aspect rating ensemble
DeepSeek-14B	Local	Positive–negative aspect rating ensemble
Gemma-27B	Local	Positive–negative aspect rating ensemble; premise generation
Llama-3.3-70B	Local	Positive–negative aspect rating ensemble
Open-Weight Baselines — Local (4 × L40S)		
Llama-3.3-70B	Local	Model scale baseline
Qwen3-32B (dense)	Local	Architectural design baseline
Qwen2.5-7B-Instruct	Local	Scale ablation baseline
Qwen2.5-14B-Instruct	Local	Scale ablation baseline
DeepSeek-R1 14B	Local	Architectural design baseline
Phi-4 Mini Instruct	Local	Architectural design baseline
Mistral Small 2501 24B	Local	Generation paradigm baseline
WizardLM-30B	Local	Generation paradigm baseline
Frontier Models — API Access Only		
GPT-4.1	OpenAI API	DPO candidate plot generation; model scale baseline
Claude Sonnet 4	Anthropic API	DPO candidate plot generation; model scale baseline
Gemini 2.0 Flash	Google API	DPO candidate plot generation; model scale baseline

Note: Agents' Room (Huot et al., 2024) was evaluated according to its original open-source implementation and run locally on the same infrastructure.