When Reasoning Supervision Hurts: TTCW-Based Long-Form Literary Review Generation

Jinlong Liu **Mohammed Bahja** **Mark Lee**
School of Computer Science, University of Birmingham, United Kingdom
jxl2069@student.bham.ac.uk;{m.bahja,m.g.lee}@bham.ac.uk

Abstract

Automatic evaluation of long-form literary writing remains challenging, as generic LLM-as-Judge approaches may not fully capture creativity-related dimensions such as originality and flexibility. Although the Torrance Test of Creative Writing (TTCW) provides a structured creativity framework, and prior work has demonstrated reference-based TTCW evaluation at the pairwise level, no large-scale dataset exists for long-form TTCW-based literary review generation. We address this gap by constructing a dataset of 263,911 long-form stories, each annotated with scalar scores and meta-synthesised review comments across 14 TTCW-based dimensions. Using this dataset, we fine-tune Qwen3 models at two scales, 4B and 8B, under two conditions: with and without reasoning content. Results show that non-reasoning fine-tuning achieves stronger and more stable performance, with the best setting reaching an evaluation score of 0.6820. Further analysis shows that reasoning-supervised models are more prone to parse failures, often continuing with irrelevant or repetitive reasoning-style text rather than completing the required 14-metric review report. These results suggest that, for fixed-format rubric-based review generation, reasoning supervision is not straightforwardly beneficial, and precise metric-aligned scoring remains challenging even after task-specific fine-tuning.[1]

1 Introduction

In recent years, LLM-as-Judge has become increasingly common and has shown promising reliability across multiple evaluation settings (Bonomo et al., 2025; Chiang and Lee, 2023; Liu et al., 2023). At the same time, more benchmarks and resources have been proposed for long-form literary or narrative evaluation, including ABSEVAL (Liang et al., 2024), STORYWARS (Du and Chilton, 2023), and COLLABSTORY (Venkatraman et al., 2025). In parallel, Chakrabarty et al. (2024) introduce TTCW as a structured creativity-oriented evaluation framework, and Li et al. (2025a) further propose a reference-based TTCW evaluator.

However, current work still lacks a public dataset for long-form TTCW-based review generation in a reference-free setting. Existing long-form literary evaluation resources do not provide TTCW-grounded review reports as supervision, and existing TTCW-based evaluation work has not released a large-scale dataset focused on literary review generation. This leaves a gap for training judge-style models that must produce both metric-aligned scores and review comments under a structured rubric.

To address this gap, we construct a large TTCW-based literary review dataset by converting the original TTCW binary questions into scalar rating questions from 1 to 10. We ask three reviewer models to score all 14 TTCW metrics independently for each story, evaluate reviewer quality through score distribution, discrimination, and metric-isolation analyses, remove the weakest reviewer, and then use a separate model to synthesise the remaining metric-wise comments into final review reports. The resulting dataset contains 263,911 rows of long-form stories in the 4K–8K word range, each paired with a complete TTCW-based review report.

Using this dataset, we further study whether reasoning supervision improves performance on this structured rubric-based review task. We compare models fine-tuned with and without reasoning content, and find that the non-reasoning setting performs better overall. The results suggest that, for fixed-format review generation with explicit score prediction, reasoning content does not improve performance and may instead reduce output stability. Our main contributions are as follows:

- We construct a large TTCW-based dataset for

[1]Code available at https://github.com/Vince-Liuss/TTCW-based-Review

long-form literary review generation by converting the original binary TTCW questions into scalar rating-based review supervision.
- We design a dataset construction pipeline that performs metric-wise reviewer scoring, reviewer-quality filtering, and comment synthesis to produce complete TTCW review reports for long-form stories.
- We provide an empirical comparison of reasoning and non-reasoning fine-tuning on this structured review task, and show that non-reasoning supervision performs better in our setting.

2 Related Work

We review two strands: (i) *LLM-as-Judge* for open-ended text evaluation, and (ii) *long-form literature* resources and metrics from the past two years. We then identify a supervision gap around the TTCW.

2.1 LLM-as-Judge

Bonomo et al. (2025) introduce LITERARYQA, a cleaned subset of NarrativeQA focused on literary works, and conduct a meta-evaluation showing that n-gram metrics correlate weakly with human judgments, whereas LLM judges—including small open-weight models—recover human-like rankings under a reference-based protocol. Chiang and Lee (2023) evaluate "LLM-as-evaluator" by giving models the same instructions and items used in human studies; model ratings track expert judgments and remain stable across prompt formatting and sampling choices. Liu et al. (2023) propose G-EVAL, where GPT-4 as the judge achieves Spearman $\rho = 0.514$ with human on summarization, illustrating that rubric-prompted judging can reach competitive human alignment.

Discourse-level analyses highlight where generic judges may miss narrative structure. Tian et al. (2024) analyze story arcs, turning points, and affect; baseline arc identification is near random for mid-tier models, improves for frontier models, but remains below human; explicitly modeling arcs/affect boosts narrative diversity, suspense, and arousal.

TTCW operationalizes creativity as a product via 14 binary tests across Fluency, Flexibility, Originality, and Elaboration (Chakrabarty et al., 2024). Reported per-test interrater agreement is moderate, while aggregate agreement is strong, supporting TTCW as a reproducible *set*-based evaluation protocol. Recent surveys catalog limitations of LLM-as-Judge (e.g., sentiment, token, and context/culture biases) and outline reliability practices (e.g., pairwise comparisons, bias controls).

2.2 Long-Form Literature Resources and Metrics

Scripted and collaborative narratives. Liang et al. (2024) propose ABSEVAL with MCSCRIPT (1,500 tasks) and report closer alignment with human judgments than single-LLM setups; top systems include strong chat models, and the agentic framework improves agreement with human evaluators. Du and Chilton (2023) release STORYWARS (40k human-authored collaborative stories; 12 task types, 101 tasks). Venkatraman et al. (2025) build COLLABSTORY (32k LLM-coauthored stories) and show that standard baselines struggle on authorship-related tasks; fine-tuned Transformers perform strongly on boundary authorship verification.

Character cognition and inner thought. Xu et al. (2025) present ROLETHINK (6,058 instances from 76 books) for character-thought generation; MIRROR (memory retrieval + chain-of-thought) outperforms baselines. Gold (original monologues) is harder than silver (expert analyses), indicating sensitivity to reference fidelity and memory access.

Long-context generation and long-text modeling. Liu et al. (2024) introduce LONGGENBENCH for long-context *generation* (logical flow); higher-baseline models degrade less, and within-series scaling (e.g., LLaMA-3, Qwen2) reduces the performance drop. Guan et al. (2022) propose LOT (Chinese long text) and show that LONGLM pretrained on 120G novels substantially outperforms similar-sized baselines on understanding and generation, with high agreement for human-labeled understanding tasks. Yang and Jin (2025) introduce LONGSTORYEVAL (600 books; avg. 121k tokens), derive aspect criteria from reader critiques, and report that NOVELCRITIQUE aligns best with human ratings overall and on most aspects.

Stress tests and judge models. He et al. (2023) design synthetic stress tests that expose blind spots in model-based metrics, recommending metric combinations and robustness probes. Judge models fine-tuned for evaluation include PANDALM, which recovers a large fraction of GPT-3.5/4's evaluation ability on its testbed and improves base models under its tuning regimen (Wang et al., 2024),

and THEMIS, a reference-free evaluator trained with consistency verification and rating-oriented preference alignment, reporting the best average performance across six Natural Language Generation(NLG) tasks in its setup (Hu et al., 2024). Wu et al. (2025) present WRITINGBENCH (six domains, 100 subdomains) with a fine-tuned critic; some English prompts request emulation of non-English figures (e.g., "write a story as Li Bai"), which can produce translationese-like prose rather than native English literary writing and complicate cross-domain comparability.

Creativity-targeted evaluation. Li et al. (2025a) propose a reference-based TTCW evaluator and report improved alignment (pairwise accuracy up to 0.75). Complementary work on creative reward shaping (RLAIF) reports strong agreement with human judgments in constrained creative settings (e.g., Chinese greetings) and underscores the role of principled judge prompts or reward models (Wei et al., 2025). Bias analyses for complex evaluation contexts find auxiliary-information-induced vulnerabilities in LLM judges, motivating explicit robustness checks (Li et al., 2025b).

2.3 Gap: Supervision for TTCW-Grounded Evaluation

Despite recent progress, there is still no public long-form dataset with TTCW-labelled supervision for automated judges. Existing judge models are typically trained on generic rubrics, while long-form literary benchmarks do not provide TTCW-grounded review supervision. As a result, current evaluation settings may capture surface quality more easily than creativity-related dimensions such as originality and flexibility. We address this gap by constructing a TTCW-based dataset for long-form literary review generation and using it to study structured rubric-based evaluation.

3 Dataset Preparation

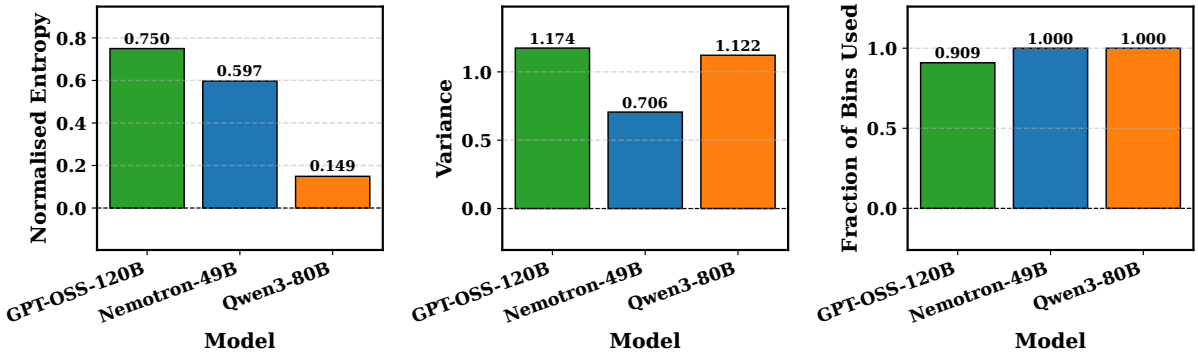
We first reformulate the original TTCW metric questions from binary judgments to scalar ratings on a 1–10 scale, embedding explicit score anchors in the system instruction so that all reviewer models operate under the same rubric, shown as Table 1. To minimise cross-metric interference and reduce the risk of a reviewer collapsing multiple criteria into a single latent judgment, we evaluate the 14 metrics independently rather than jointly; the full metric list is provided in the Appendix.

We select three recent and capable reviewer models: `Llama-3_3-Nemotron-Super-49B-v1_5` (Bercovich et al., 2025), `Qwen3-Next-80B-A3B-Instruct` (Team, 2025) (non-reasoning mode), and `gpt-oss-120b` (OpenAI, 2025). For source fiction, we use the `WritingPrompts` corpus (Fan et al., 2018). Because many stories fall below the length threshold suitable for long-form evaluation, we remove samples exceeding 8K words and use `Gemma-3-27b-it` (Team et al., 2025b) to regenerate stories from the original prompts, treating human-written stories as references, to obtain samples in the 4K–8K word range. Each reviewer model then evaluates every story one metric at a time, and `GLM-4.5-Air` (Team et al., 2025a) serves as a meta-synthesis model that consolidates the per-metric reviews into a single coherent review per story. All models are run with `temperature = 0`.

Before finalising the dataset, we assess reviewer suitability using three diagnostics: *score distribution*, which detects score concentration or ceiling effects; *discrimination score*, which measures whether a reviewer uses the score scale sufficiently to distinguish among stories; and *metric isolation*, which examines whether the 14 TTCW metrics are treated as distinct criteria rather than collapsed into a single latent quality judgement. The results are shown in Fig. 1, Fig. 3, Fig. 2, and Fig. 4.

Compared with `gpt-oss-120b` and `Llama-3_3-Nemotron-Super-49B-v1_5`, `Qwen3-Next-80B-A3B-Instruct` shows weaker reviewer suitability. Fig. 1 and Fig. 3 show strong score concentration, low score entropy, and limited score variation, indicating weak practical discrimination. Although Fig. 2 shows lower inter-metric correlations for Qwen3-80B, this is not sufficient evidence of better metric isolation, because the model also lacks meaningful variation in score usage. We therefore interpret its low correlation pattern together with its score-distribution behaviour as evidence of unreliable reviewer performance. Full 14-metric heatmaps are shown in Fig. 4.

We therefore exclude `Qwen3-Next-80B-A3B-Instruct` from the synthesis stage and retain the remaining two reviewer models. The final dataset contains 263,911 rows and is designed for long-context literary review generation: each input story is in the 4K–8K word range, and each output contains a complete TTCW-based review report. This makes the task substantially longer than standard short-form evaluation settings. In fine-tuning, the non-reasoning version uses a maximum context



(a) Normalised Score Entropy(Higher is better) (b) Per-Metric Score Variance(Higher is better) (c) Score Bin Coverage(Higher is better)

Figure 1: Discrimination score comparison across reviewer models. `Gpt-oss-120b` exhibits the strongest criterion-sensitive score usage, with the highest normalised entropy and per-metric variance. `Llama-3_3-Nemotron-Super-49B-v1_5` is intermediate. `Qwen3-Next-80B-A3B-Instruct`, despite full bin coverage, has extremely low entropy, indicating strong score concentration and weaker practical discrimination.

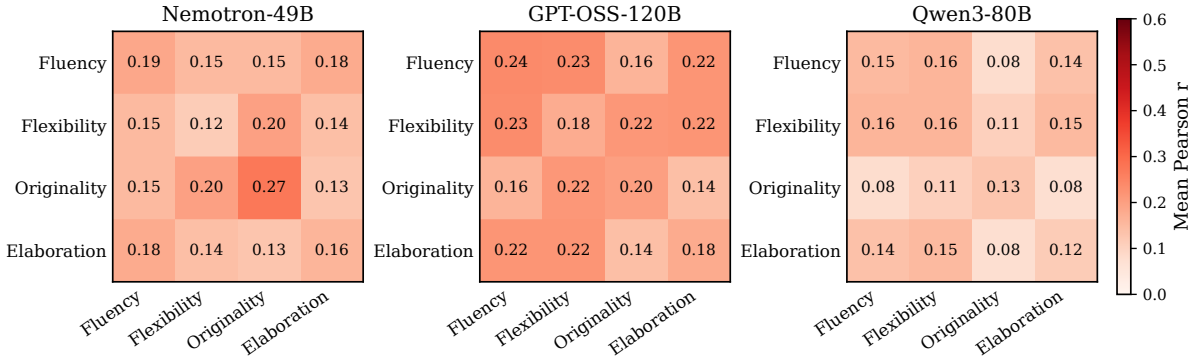


Figure 2: Compact group-level inter-metric correlation comparison across reviewer models. The original 14 TTCW metrics are aggregated into four TTCW dimensions: Fluency, Flexibility, Originality, and Elaboration. Diagonal cells report mean within-dimension off-diagonal Pearson correlation, while off-diagonal cells report mean cross-dimension Pearson correlation. Qwen3-80B shows comparatively low group-level correlations, but this does not indicate stronger reviewer quality in our setting; combined with its weak score-distribution behaviour and strong score concentration, it suggests limited practical discrimination across samples. We therefore exclude Qwen3-80B from the final synthesis stage and retain GPT-OSS-120B and Nemotron-49B. Full 14-metric correlation heatmaps are provided in Figure 4.

length of 16,384 tokens, while the reasoning version requires 32,768 tokens because it additionally includes reviewer-style reasoning traces. [2]

Sample Validation. To further assess the quality of the meta-synthesised reviews, we conduct an automatic sample-level validation using `NVIDIA-Nemotron-3-Super-120B-A12B` (NVIDIA, 2025), a recent reasoning-oriented judge model. We randomly sample 50 stories and pair each story with its 14 metric-specific review comments, resulting in 700 story–metric review pairs. For each pair, we ask the validation model three binary questions commonly used to assess NLG quality:

1. **Faithfulness:** Does the review only make claims that are consistent with the story's actual content, without introducing details, events, or characterisations not present in the story?
2. **Coherence:** Is the review logically organised and internally consistent, with no contradictory statements?
3. **Relevance:** Does the review focus on specific aspects of this story rather than making observations that could apply to almost any story?

The results are reported in Table 2. The validation results show that the meta-synthesised reviews are highly relevant to the corresponding stories and

[2]Dataset available at `https://huggingface.co/datasets/VibrantVista/TTCW-Based-Review`

Component	Full Prompt Content
System Prompt	You are an experienced fiction editor preparing manuscripts for publication. You evaluate writing like you would for a serious author revising toward publication. General rules: - You judge ONLY the specific craft metric you are given. - You must ground every judgment in concrete evidence from the story (scenes, passages, beats). Do not make abstract claims with no textual anchor. - You must give both strengths (what should be preserved) and revision priorities (what should change first). Editors do both. - You must assign a score using the rubric below. You are allowed to use any integer 1–10. Do not collapse to the middle just to be "safe." If the work is excellent for this metric, score high. If it is weak, score low. This instruction overrides any instinct to hedge. Scoring rubric (used for ANY metric): 10 = Publication-ready control of this metric. Consistent, intentional, supports the story's emotional/structural goals. 8 = Strong. The metric is working in most places; only light refinements needed. 6 = Mixed. The core skill is present but unreliable. Some scenes undercut the intended effect. Needs targeted revision. 4 = Weak. The metric frequently misfires or distracts. Multiple sections need significant rewrite. 2 = Fundamentally not working. The intended effect is mostly lost for this metric. 1 = Essentially absent or actively damaging the story. You may still choose any integer 1–10. The descriptions are anchors, not the only valid scores. Your required output format: Reasons: [The detailed reasoning you used to arrive at your score, including specific examples from the story] Score: [single integer 1–10] You must follow that exact formatting.
Fluency1 Prompt	{story} ###Expanded Expert Measure 'Compression/stretching of time' in fiction writing, also known as pacing, refers to the manipulation of time in storytelling for dramatic effect, pacing, or other narrative purposes. Essentially, it's about controlling the perceived speed and rhythm at which a story unfolds. Compression of time refers to when events that take a long time (hours, days, weeks, or even years) are summarized or condensed into a brief narrative span. For example, a writer might compress several years of a character's life into a few paragraphs to quickly convey important changes or developments. On the other hand, stretching of time is when a brief moment or event is drawn out over pages or chapters. It's often used to create suspense, emphasize details, or delve deeper into a character's thoughts and feelings. For example, the few seconds it takes for a dropped glass to hit the floor might be stretched out with detailed descriptions of the action, reactions, and thoughts of characters involved. Storytime refers to the time within the world of the story, while real-world time refers to the time it takes for the reader to read the story. A skilled writer can manipulate the relationship between these two to affect the pacing of the narrative, either speeding it up (compression) or slowing it down (stretching). This technique plays a crucial role in shaping the reader's experience and engagement with the story. Given the story above, list out the scenes in the story in which time compression or time stretching is used, and argue for each whether it is successfully implemented. Then overall, give your reasoning about the question below and give an answer from 1 to 10, where 10 is the best score and 1 is the worst score. Q) How appropriate and balanced does the manipulation of time in terms of compression or stretching feel?

Table 1: Full shared system prompt and the full `Fluency1` prompt used in dataset construction. The metric description in `Fluency1` follows the original TTCW criterion wording from Chakrabarty et al. (2024), while the scoring instruction and output format are adapted for our review-generation setting.

moderately faithful to the story content. This indicates that most reviews focus on story-specific evidence rather than producing generic literary comments. However, the coherence pass rate is much lower. This does not necessarily mean that the reviews are irrelevant or ungrounded; instead, it reflects the difficulty of the meta-synthesis step. The input to synthesis contains reviewer outputs from multiple models, often with step-by-step reasoning and overlapping or partially inconsistent observations. The synthesis model must remove unnecessary reasoning traces, select the most useful evidence, and organise the remaining points into a concise metric-level comment. This is a complex transformation, and the low coherence score suggests that current models still struggle to consistently produce well-organised synthesised reviews in this setting.

Validation Dimension	Pass Rate	Passed / Total
Faithfulness	71.86%	503 / 700
Coherence	20.43%	143 / 700
Relevance	97.29%	681 / 700

Table 2: Sample-level validation of meta-synthesised review comments. We randomly sample 50 stories and evaluate all 14 metric-specific reviews for each story, yielding 700 story–metric review pairs. Each pair is judged for faithfulness, coherence, and relevance.

4 Experiment

This experiment investigates whether reasoning content improves model performance on the literary review generation task. To incorporate reasoning supervision, we use the raw outputs of the two retained reviewer models as reasoning traces, allowing the target model to learn from multiple reviewer-style reasoning processes. Due to com-

Overall Comparison								
Variant	**Qwen3-8B**				**Qwen3-4B**			
	Parse	**Score Acc.**	**BERTScore F1**	**Eval**	**Parse**	**Score Acc.**	**BERTScore F1**	**Eval**
Without reasoning content + thinking off	**1.0000**	**0.6744**	**0.6895**	**0.6820**	0.9998	0.6713	0.6885	0.6798
Without reasoning content + thinking on	**1.0000**	0.6738	**0.6895**	0.6816	0.9996	0.6715	0.6884	0.6797
With reasoning content + thinking on	0.8708	0.6319	0.6826	0.5723	0.8592	0.6295	0.6818	0.5633
With reasoning content + thinking off	0.9880	0.6519	0.6842	0.6600	0.4270	0.6224	0.6820	0.2785

Full Comparison Across Model Scale, Training Setting, and Test-Time Thinking								
Metric	**8B Without Reasoning**		**8B With Reasoning**		**4B Without Reasoning**		**4B With Reasoning**	
	Score Acc.	**BERTScore F1**	**Score Acc.**	**BERTScore F1**	**Score Acc.**	**BERTScore F1**	**Score Acc.**	**BERTScore F1**
Narrative Pacing (Compression/Stretching)	**0.5948**	**0.6722**	0.5537	0.6666	0.5980	0.6709	0.5494	0.6661
Scene vs Exposition Balance	**0.6592**	**0.6952**	0.6504	0.6865	0.6578	0.6959	0.6472	0.6853
Language Proficiency & Literary Devices	**0.4962**	**0.6804**	0.4446	0.6724	0.4942	0.6788	0.4483	0.6725
Narrative Ending Quality	**0.7105**	**0.6920**	0.6786	0.6847	0.7076	0.6913	0.6716	0.6842
Understandability & Coherence	**0.8304**	**0.6854**	0.7785	0.6777	0.8278	0.6844	0.7883	0.6769
Perspective & Voice Flexibility	**0.7888**	**0.6946**	0.7527	0.6871	0.7904	0.6939	0.7547	0.6860
Emotional Flexibility (Interiority/Exteriority)	**0.7843**	**0.7047**	0.7099	0.6950	0.7781	0.7037	0.7081	0.6939
Structural Flexibility (Surprising Turns)	**0.7760**	**0.6741**	0.6830	0.6684	0.7743	0.6724	0.6931	0.6672
Originality in Theme and Takeaway	**0.5930**	**0.6821**	0.5458	0.6751	0.5861	0.6807	0.5433	0.6745
Originality in Thought (Cliché Avoidance)	**0.6197**	**0.6971**	0.5728	0.6897	0.6205	0.6970	0.5498	0.6890
Originality in Form/Structure	**0.5848**	**0.6907**	0.5597	0.6845	0.5782	0.6897	0.5493	0.6835
World-Building and Sensory Believability	**0.7585**	**0.7096**	0.7297	0.7038	0.7524	0.7078	0.7160	0.7032
Character Development Depth	**0.5645**	**0.6936**	0.5401	0.6889	0.5557	0.6920	0.5475	0.6886
Rhetorical Complexity (Surface vs Subtext)	**0.6814**	**0.6816**	0.6471	0.6756	0.6774	0.6800	0.6466	0.6748

Table 3: Full comparison of reasoning and non-reasoning fine-tuning across Qwen3-8B and Qwen3-4B. The upper block reports all four decoding settings: models trained without reasoning content are evaluated with thinking disabled and enabled, and models trained with reasoning content are also evaluated with thinking enabled and disabled. These cross-mode results test whether the decoding mode alone can recover performance when it differs from the training setting. The lower block reports per-metric score accuracy and BERTScore F1 for the main matched comparison: non-reasoning training with thinking disabled versus reasoning training with thinking enabled. Overall, non-reasoning supervision remains stronger and more stable, while reasoning-supervised models show lower score accuracy and reduced parse reliability, especially under cross-mode decoding.

putational constraints, we restrict fine-tuning to an 8B model with LoRA, and choose `Qwen3-8B` (Team, 2025) as the base, given its broad community adoption and native support for both reasoning and non-reasoning modes. Using this base, we train two variants: one fine-tuned without reasoning content and one fine-tuned with reasoning content, and evaluate both under identical decoding conditions (`temperature = 0`). The training configuration is: learning rate 2×10^{-4}, `lora_r` $= 64$, and `lora_alpha` $= 128$. All experiments are conducted on a node equipped with four NVIDIA L40S GPUs (48GB VRAM each), two AMD EPYC 9334 32-Core processors, and 1TB RAM.

We evaluate model outputs along two axes: *stability* and *performance*. For stability, we use the parse rate $p \in [0, 1]$, which measures whether the model can generate a complete report in the required format. If any of the 14 metric reports is missing or malformed, the whole output is treated as invalid.

For performance, we evaluate both score prediction and review text generation. Score quality is measured by the mean absolute error (MAE) between predicted and reference scores:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |\hat{y}_i - y_i| \,,$$

where $\hat{y}_i$ is the predicted score and y_i is the reference score. We then transform MAE into a bounded score:

$$s_{\text{MAE}} = e^{-\text{MAE}},$$

so that lower MAE gives a value closer to 1.

Review text quality is measured using BERTScore F1, denoted as:

$$s_{\text{BERTScore-F1}} \in [0, 1].$$

This measures the semantic similarity between generated and reference review comments.

Finally, we combine parse stability, score quality, and review similarity into the final evaluation score:

$$S_{\text{eval}} = p \cdot (0.5\, s_{\text{MAE}} + 0.5\, s_{\text{BERTScore-F1}}) \,.$$

This formulation rewards models only when they are both structurally parseable and semantically close to the reference outputs.

The results are reported in Table 3. The clearest difference between settings is parse stability. Models fine-tuned without reasoning content are highly reliable across both model scales and decoding modes: Qwen3-8B achieves a parse rate of 1.0000 with both thinking disabled and enabled, while Qwen3-4B remains close to perfect with parse rates of 0.9998 and 0.9996. This shows that the fixed 14-metric report format is learnable when the model is trained directly on the final review output.

By contrast, reasoning-supervised models are substantially less stable when thinking is enabled. Their parse rates drop to 0.8708 for Qwen3-8B and 0.8592 for Qwen3-4B, indicating that reasoning traces make the model more likely to violate the required output structure. The cross-mode results further clarify this effect. For Qwen3-8B, disabling thinking after reasoning fine-tuning improves the parse rate from 0.8708 to 0.9880 and raises the final evaluation score from 0.5723 to 0.6600. This suggests that part of the instability comes from the generated thinking content itself. However, the same intervention is harmful for Qwen3-4B, where the parse rate drops from 0.8592 to 0.4270, suggesting that the smaller model becomes more dependent on the reasoning-style generation pattern learned during fine-tuning.

Manual inspection supports this interpretation. Failed outputs are usually not caused by a single malformed metric field. Instead, when the reasoning process fails, the model often fails at the sequence level: it may leak reasoning-style content into the final answer, introduce unrelated intermediate text, repeat early report sections, or stop before producing the full 14-metric review. Since our parser requires every metric report to be present and correctly formatted, these sequence-level failures directly explain the lower parse rates of reasoning-supervised models.

The score accuracy gap is smaller than the parse-rate gap but remains consistent. The best score accuracy is obtained by Qwen3-8B without reasoning content, while reasoning-supervised variants remain lower under both decoding modes. We attribute this weaker score prediction mainly to the increased difficulty of the reasoning setting: it requires the model to process longer sequences, learn reviewer-style reasoning traces, and still output calibrated rubric-aligned scores. LoRA may further limit this adaptation because only a small fraction of parameters is updated, but the main source of parse instability appears to be the mismatch between reasoning-style generation and strict fixed-format report generation.

5 Conclusion

We construct a large-scale TTCW-based literary review dataset with scalar metric scores and metric-wise review comments for long-form stories. Using this dataset, we study whether reasoning supervision improves structured review report generation. Our results show that non-reasoning fine-tuning consistently achieves stronger and more stable performance across both Qwen3-8B and Qwen3-4B. In particular, reasoning-supervised models are more prone to parse failures caused by format leakage, repetitive generation, and incomplete metric reports. These findings suggest that reasoning traces are not automatically beneficial for fixed-format rubric-based evaluation, especially when the model must produce precise scores and complete structured outputs under long-context constraints. Future work should test whether higher-quality reasoning traces, larger models, or stronger adaptation methods beyond LoRA can make reasoning supervision more effective for this setting.

Limitations

This work has several limitations. First, dataset construction does not involve human annotators, so the supervision signal is entirely model-generated and may contain bias, scoring noise, or synthesis errors. Second, all experiments are conducted on the Qwen3 model family, which limits the generalisability of our findings to models with different architectures or reasoning behaviours. Third, we only study 4B and 8B models, so it remains unclear whether larger models with stronger long-context and instruction-following capabilities would show the same pattern. Finally, we use LoRA-based parameter-efficient fine-tuning rather than full fine-tuning, which may constrain fine-grained rubric-based score learning.

References

Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, Ido Shahaf, Oren Tropp, Ehud Karpas, Ran Zilberstein, Jiaqi Zeng, Soumye Singhal, Alexander Bukharin, Yian Zhang, Tugrul Konuk, and 114 others. 2025. Llama-nemotron: Efficient reasoning models. *Preprint*, arXiv:2505.00949.

Tommaso Bonomo, Luca Gioffré, and Roberto Navigli. 2025. LiteraryQA: Towards effective evaluation of long-document narrative QA. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 34074–34095, Suzhou, China. Association for Computational Linguistics.

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*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Yulun Du and Lydia Chilton. 2023. StoryWars: A dataset and instruction tuning baselines for collaborative story understanding and generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3044–3062, Toronto, Canada. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Jian Guan, Zhuoer Feng, Yamei Chen, Ruilin He, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2022. LOT: A story-centric benchmark for evaluating Chinese long text understanding and generation. *Transactions of the Association for Computational Linguistics*, 10:434–451.

Tianxing He, Jingyu Zhang, Tianle Wang, Sachin Kumar, Kyunghyun Cho, James Glass, and Yulia Tsvetkov. 2023. On the blind spots of model-based evaluation metrics for text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12067–12097, Toronto, Canada. Association for Computational Linguistics.

Xinyu Hu, Li Lin, Mingqi Gao, Xunjian Yin, and Xiaojun Wan. 2024. Themis: A reference-free NLG evaluation language model with flexibility and interpretability. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15924–15951, Miami, Florida, USA. Association for Computational Linguistics.

Ruizhe Li, Chiwei Zhu, Benfeng Xu, Xiaorui Wang, and Zhendong Mao. 2025a. Automated creativity evaluation for large language models: A reference-based approach. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 21475–21488, Suzhou, China. Association for Computational Linguistics.

Weiyuan Li, Xintao Wang, Siyu Yuan, Rui Xu, Jiangjie Chen, Qingqing Dong, Yanghua Xiao, and Deqing Yang. 2025b. Curse of knowledge: Your guidance and provided knowledge are biasing LLM judges in complex evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 14900–14924, Suzhou, China. Association for Computational Linguistics.

Sirui Liang, Baoli Zhang, Jun Zhao, and Kang Liu. 2024. ABSEval: An agent-based framework for script evaluation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12418–12434, Miami, Florida, USA. Association for Computational Linguistics.

Xiang Liu, Peijie Dong, Xuming Hu, and Xiaowen Chu. 2024. LongGenBench: Long-context generation benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 865–883, Miami, Florida, USA. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

NVIDIA. 2025. Nvidia nemotron 3: Efficient and open intelligence. White Paper.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

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. 2025a. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *Preprint*, arXiv:2508.06471.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2025b. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May,

and Nanyun Peng. 2024. Are large language models capable of generating human-level narratives? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17659–17681, Miami, Florida, USA. Association for Computational Linguistics.

Saranya Venkatraman, Nafis Irtiza Tripto, and Dongwon Lee. 2025. CollabStory: Multi-LLM collaborative story generation and authorship analysis. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3665–3679, Albuquerque, New Mexico. Association for Computational Linguistics.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2024. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. *Preprint*, arXiv:2306.05087.

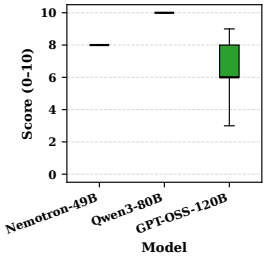
Xiaolong Wei, Bo Lu, Xingyu Zhang, Zhejun Zhao, Dongdong Shen, Long Xia, and Dawei Yin. 2025. Igniting creative writing in small language models: LLM-as-a-judge versus multi-agent refined rewards. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17171–17197, Suzhou, China. Association for Computational Linguistics.

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

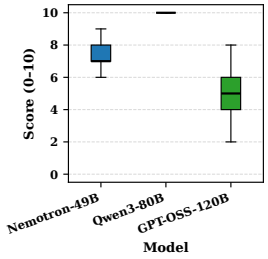
Rui Xu, Mingyu Wang, Xintao Wang, Dakuan Lu, Xiaoyu Tan, Wei Chu, and Xu Yinghui. 2025. Guess what I am thinking: A benchmark for inner thought reasoning of role-playing language agents. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15148–15168, Suzhou, China. Association for Computational Linguistics.

Dingyi Yang and Qin Jin. 2025. What matters in evaluating book-length stories? a systematic study of long story evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16375–16398, Vienna, Austria. Association for Computational Linguistics.

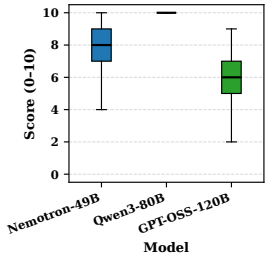
A Additional Plots



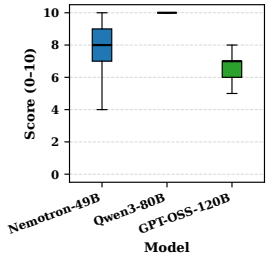
(a) Narrative Pacing



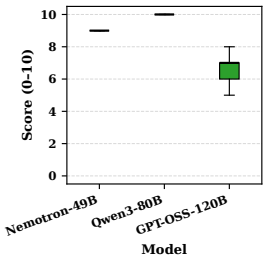
(b) Scene vs Exposition



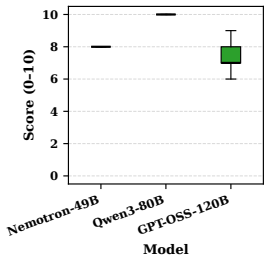
(c) Language Proficiency & Literary Devices



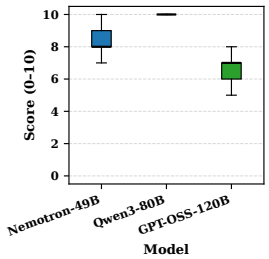
(d) Narrative Ending



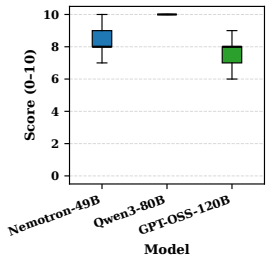
(e) Understandability & Coherence



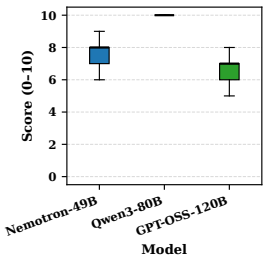
(f) Perspective & Voice Flexibility"s



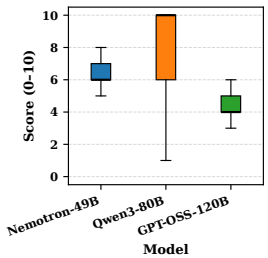
(g) Emotional Flexibility



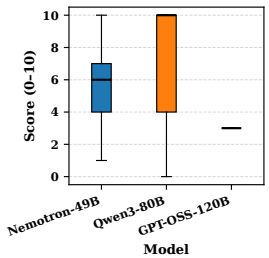
(h) Structural Flexibility



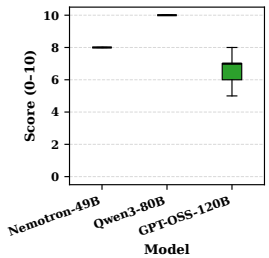
(i) Originality in Theme and Content



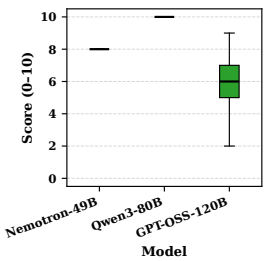
(j) Originality in Thought



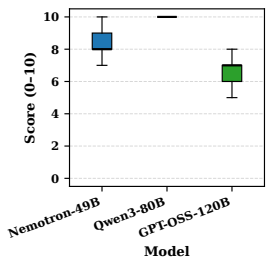
(k) Originality in Form



(l) World Building and setting

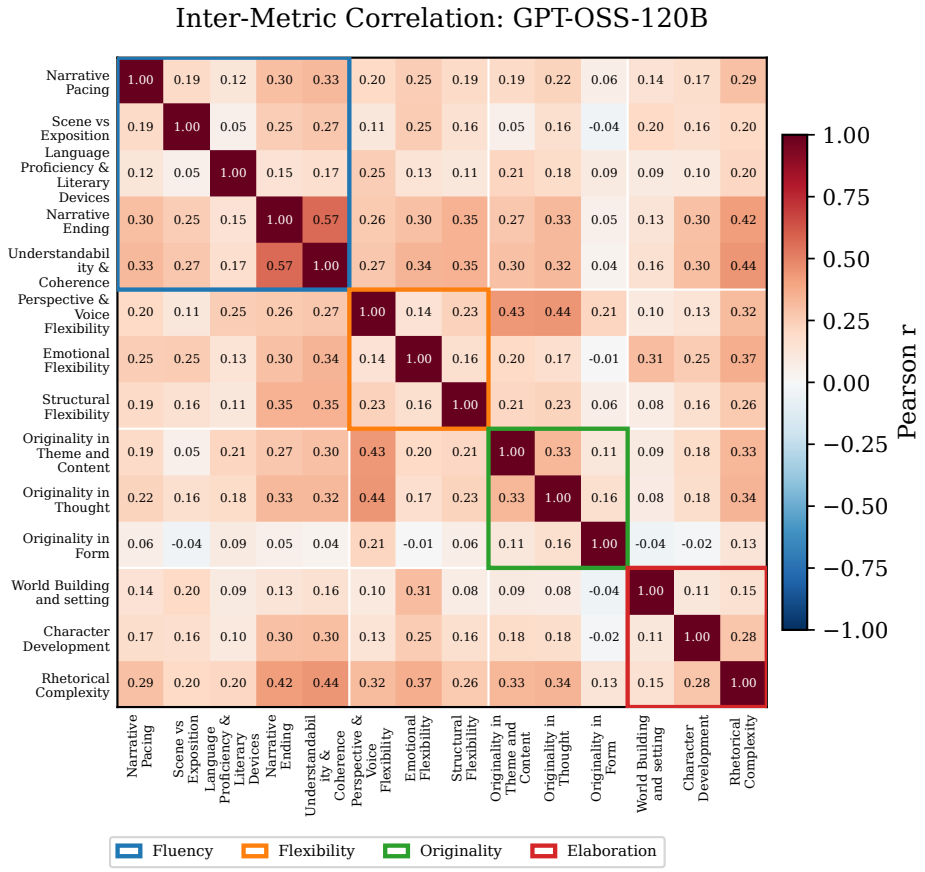


(m) Character Development

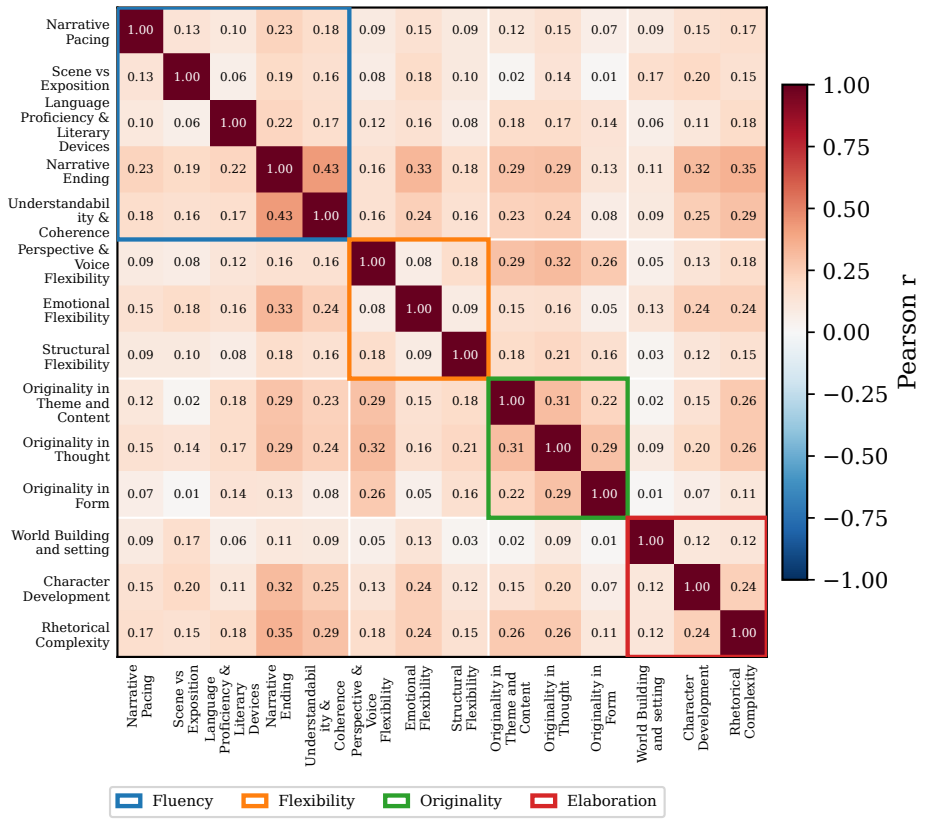


(n) Rhetorical Complexity

Figure 3: Score Distribution across all metrics

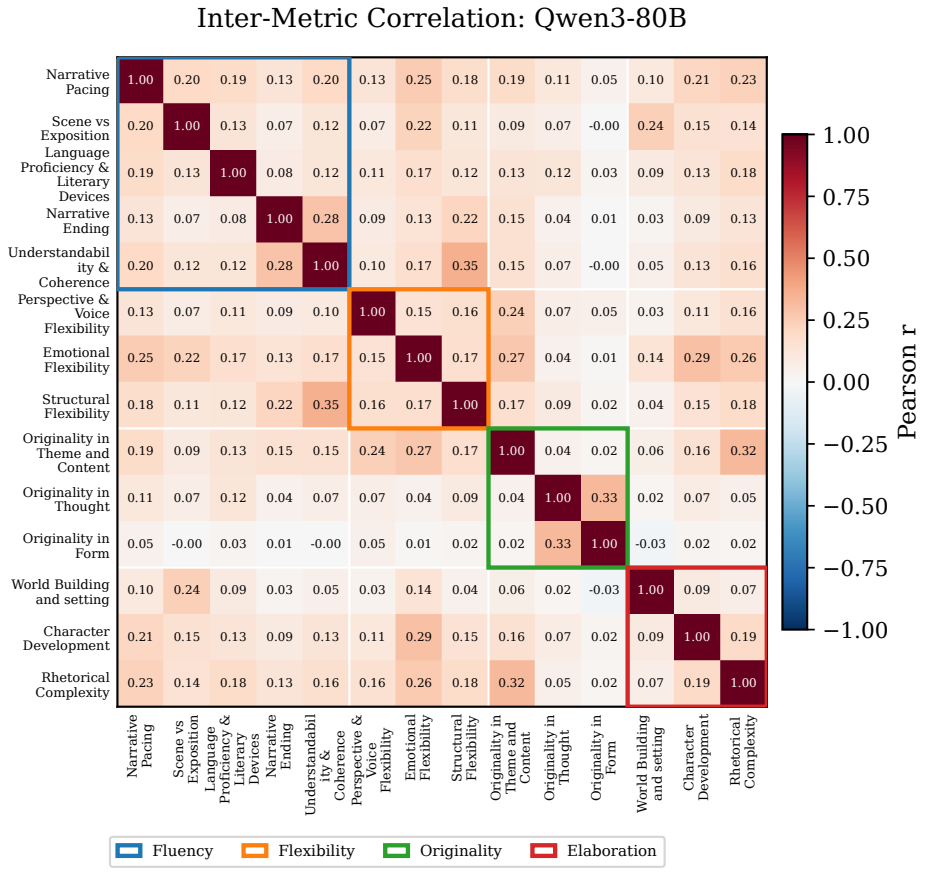


(a)



(b)

Figure 4: Inter-metric correlation heatmaps for the three reviewer models across the 14 independently scored fiction-review dimensions. These plots diagnose whether reviewer outputs preserve metric distinctions or exhibit cross-metric coupling; stronger widespread correlations suggest greater risk of rubric collapse into broader latent quality signals.



(c)

Figure 4: Inter-metric correlation heatmaps for the three reviewer models across the 14 independently scored fiction-review dimensions. These plots diagnose whether reviewer outputs preserve metric distinctions or exhibit cross-metric coupling; stronger widespread correlations suggest greater risk of rubric collapse into broader latent quality signals.(continued)

B Prompts

Our dataset-construction prompts are adapted from the expert evaluation criteria of the Torrance Test of Creative Writing (TTCW) proposed by Chakrabarty et al. (2024). Since the metric definitions largely follow the original TTCW descriptions, we report only the final instruction pattern and scoring question used for each metric.

Table 4: Final instruction patterns and scoring questions used in dataset construction. Full metric definitions follow the original TTCW criteria.

Metric	**Final prompt instruction**
Narrative Pacing	Given the story above, list out the scenes in the story in which time compression or time stretching is used, and argue for each whether it is successfully implemented. Then overall, give your reasoning about the question below and give an answer from 1 to 10, where 10 is the best score and 1 is the worst score. **Q)** How appropriate and balanced does the manipulation of time in terms of compression or stretching feel?
Scene vs. Exposition	Given the story above, answer the following question. Please first explain your reasoning step by step and then give an answer from 1 to 10, where 10 is the best score and 1 is the worst score. **Q)** How well does the story balance scene and summary/exposition, rather than relying heavily on one element?
Language Proficiency & Literary Devices	Given the story above, please list out all the metaphors, idioms and literary allusions, and for each decide whether it is successful or whether it feels forced or too easy. Then overall, give your reasoning about the question below and give an answer from 1 to 10, where 10 is the best score and 1 is the worst score. **Q)** How sophisticatedly does the story use idiom, metaphor, or literary allusion?
Narrative Ending Quality	Given the story above, answer the following question. Please first explain your reasoning step by step and then give an answer from 1 to 10, where 10 is the best score and 1 is the worst score. **Q)** How natural and earned does the end of the story feel, rather than arbitrary or abrupt?
Understandability & Coherence	Given the story above, answer the following question. Please first explain your reasoning step by step and then give an answer from 1 to 10, where 10 is the best score and 1 is the worst score. **Q)** How well do the different elements of the story work together to form a unified, engaging, and satisfying whole?
Perspective & Voice Flexibility	Given the story above, answer the following question. Please first explain your reasoning step by step and then give an answer from 1 to 10, where 10 is the best score and 1 is the worst score. **Q)** How well does the story represent perspective and voice in a flexible and convincing way?
Emotional Flexibility	Given the story above, answer the following question. Please first explain your reasoning step by step and then give an answer from 1 to 10, where 10 is the best score and 1 is the worst score. **Q)** How well does the story achieve a balance between interiority and exteriority in a way that feels emotionally flexible?
Structural Flexibility	Given the story above, list each element in the story that is intended to be surprising. For each, decide whether the surprising element remains appropriate with respect to the entire story. Then overall, give your reasoning about the question below and give an answer from 1 to 10, where 10 is the best score and 1 is the worst score. **Q)** How well does the story contain turns that are both surprising and appropriate?
Originality in Theme and Takeaway	Given the story above, list out elements that are unique takeaways of this story for the reader. Then overall, give your reasoning about the question below and give an answer from 1 to 10, where 10 is the best score and 1 is the worst score. **Q)** How likely is it that an average reader of this story will obtain a unique and original idea from reading it?
Originality in Thought	Given the story above, are there any clichés in the story? If so, list out all the elements in this story that are cliché. Then overall, give your reasoning about whether the piece is negatively impacted by these clichés and give an answer from 1 to 10, where 10 is the best score and 1 is the worst score. **Q)** How original is the story as a piece of writing, without clichés?

Continued on next page

Metric	Final prompt instruction
Originality in Form	Given the story and the devices mentioned above, list each device used with a short explanation of whether it is successful or not. Then overall, give your reasoning about the question below and give an answer from 1 to 10, where 10 is the best score and 1 is the worst score. **Q)** How original is the story in its form?
World-Building and Sensory Believability	Given the story above, list out the elements in the story that call to each of the five senses. Then overall, give your reasoning about the question below and give an answer from 1 to 10, where 10 is the best score and 1 is the worst score. **Q)** How well does the writer make the fictional world believable at the sensory level?
Character Development Depth	Given the story above, list each character and the level of development. Then overall, give your reasoning about the question below and give an answer from 1 to 10, where 10 is the best score and 1 is the worst score. **Q)** How well does each character in the story feel developed at the appropriate complexity level, ensuring that no character is present merely to satisfy a plot requirement?
Rhetorical Complexity	Given the story above, answer the following question. Please first explain your reasoning step by step and then give an answer from 1 to 10, where 10 is the best score and 1 is the worst score. **Q)** How well do passages in the story involve subtext, and when subtext is present, how effectively does it enrich the story's setting rather than feel forced?