POLARIS: Guiding Small Models to Write Long Stories

Anonymous ACL submission

Abstract

Small open-weight models struggle at long-form creative writing: their generated stories either fall far short of the requested length, or their quality significantly degrades as length increases, especially when compared to frontier models. We present POLARIS[1], **a lower-compute GRPO recipe** with two key ingredients: a frontier LLM judge with a **structured *Story Quality* rubric** as the online reward, and **human-reference injection (HRI)**, where a teacher-forced human-written story serves as a high-reward anchor within each GRPO group. By applying our training recipe to Qwen3.5-9B, using a dataset of ∼1.4K prompt–story pairs derived from 100 short-story anthologies and 4 A100 GPUs, we obtain POLARIS–9B. Across five benchmarks spanning in-distribution and out-of-distribution prompts and rubrics, **POLARIS-9B** is **competitive with much larger open-weight models** while following length instructions more closely. A blinded human evaluation confirms that POLARIS–9B is preferred to the base Qwen3.5-9B and on par with Qwen3.5-27B. Despite training only on stories up to 4k words, POLARIS–9B preserves quality on prompts requesting stories up to 3× the training length, a regime where most open-weight models degrade substantially in quality, length adherence, or both. More broadly, our results suggest that **length generalization is a meaningful stress test for creative-writing models** and a useful lens for distinguishing otherwise close models.[2]

1 Introduction

Reinforcement learning (RL) with verifiable rewards has driven strong gains in LLM reasoning and code (Guo et al., 2025a; Yu et al., 2026). Recent work shows RL can improve long-form generation (Wu et al., 2026; Liao et al., 2026; Li et al., 2026; Lei et al., 2026; Jia et al., 2025; Cao et al., 2026), but the strongest recipes rely on some combination of large or specialized base models, continual pretraining, custom-trained reward models, or large amounts of training data.

This begs a more practical question: can a small open model become competitive at long-form creative writing *without* that scale of compute and infrastructure? Starting from Qwen3.5-9B (Qwen Team, 2026), we apply POLARIS, our lower-compute creative-writing RL recipe, to 1,388 prompt–story pairs from a corpus derived from 100 commercially purchased short-story anthologies, on 4 A100 GPUs. The recipe addresses two obstacles specific to long-form creative RL :

✧ **Reward signal design.** Prior work uses trained reward models that produce scalar or pairwise preferences (Wu et al., 2026; Liao et al., 2026; Li et al., 2026) making it hard to tell which quality dimensions improved, and require a separate reward-model training pipeline. We replace this with a frontier LLM judge queried during each GRPO step using a structured *Story Quality* rubric that returns per-dimension scores with textual evidence across 16 dimensions combining narrative-theoretic concerns (Boyd et al., 2020; Adler et al., 2018; D'Souza, 2021) with practical fiction-writing norms used in workshop and editorial feedback and empirical observations made during the course of this work from reading human-written and AI-generated stories.

✧ **Stagnation in open-ended long-form RL.** Due to the open-ended nature of creative writing, policy rollouts could receive increasingly similar reward scores as GRPO training progresses. This can cause learning to slow or stall before the model reaches stronger long-form writing behavior. We propose *human-reference injection* (HRI) during

[1] **P**olicy **O**ptimization with **L**LM-as-a-judge rewards and **A**nchored-**R**eference **I**njection for **S**torywriting

[2] Writing prompts, code, trained models, and a website with generated stories to be released upon acceptance. See A.3 for memorization audit on training data.

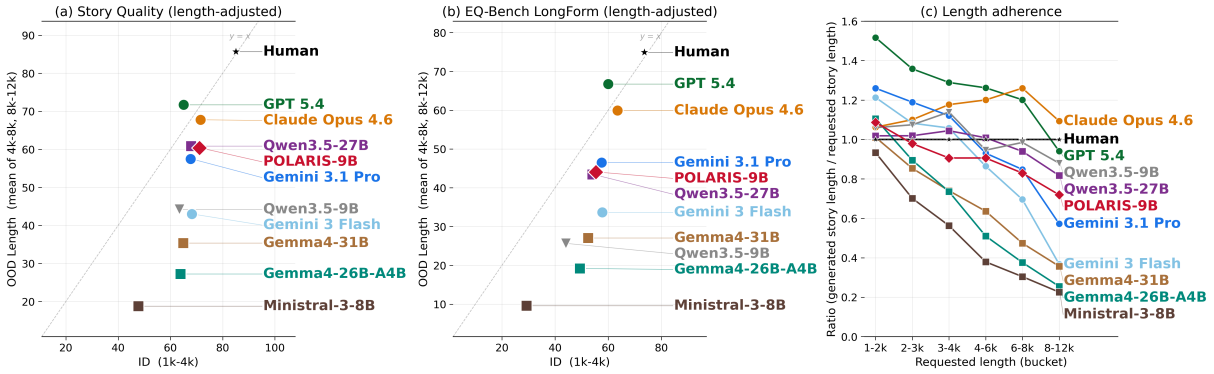


Figure 1: Length generalization (preview of Section 5.3). (a) Story Quality length-adjusted (for signed scores, we first normalize from the rubric floor and ceiling to $[0, 100]$, then multiply by $\min(\mathrm{LR}, 1/\mathrm{LR})$) where LR is the length adherence ratio, calculated as the ratio between length of generated story and requested story length: scatter of mean ID-length score (1–4k) vs. mean OOD-length score (mean of near and far buckets); the $y = x$ diagonal indicates no length-related degradation. (b) Same scatter for EQ-Bench Longform. (c) Length adherence (generated/requested word count) across requested-length buckets. POLARIS–9B sits in the top open-weight cluster on Story Quality and alongside Qwen3.5-27B in the top open-weight cluster on EQ-Bench Longform, and is the only open-weight model that maintains both quality and length adherence at far transfer. Length-adjusted scoring penalizes models such as Gemma 4 31B that maintain quality only by writing progressively shorter stories.

GRPO training, in which a teacher-forced human-written story is inserted into each group as a high-quality anchor. This helps by maintaining gradient pressure toward stronger writing rather than letting training be dominated entirely by the model's own rollout distribution. The human reference is excluded from group statistics but included in advantage computation with a warmup schedule.

Our recipe yields POLARIS–9B, a 9B model that is competitive with models $3\times$ its size and remains robust on prompts requesting stories up to 12k words, despite seeing no story longer than 4k words during training. Figure 1 shows this key result: POLARIS–9B largely maintains both rubric quality and length adherence across the full 1k–12k requested length range. A blinded human evaluation points in the same direction: annotators clearly prefer POLARIS–9B to the base model, and their comments most often highlight stronger atmosphere, voice, and scene realization, while the comparison to Qwen3.5-27B remains close.

Our contributions are:

✧ **A lower-compute GRPO recipe for long-form creative writing** using an LLM-as-judge reward with a structured rubric (*Story Quality*) and human-reference injection (HRI), yielding a 9B model that ranks in the top open-weight cluster on EQ-Bench Longform and above all open-weight models on EQ-Bench Creative Writing Elo, behind only GPT-5.4 and Claude Opus 4.6 (Anthropic, 2026).

✧ **Length generalization** is a meaningful stress test for creative writing models: a model that merely learns surface-level writing patterns will degrade as stories grow longer and require sustained narrative coherence, arc completion, and stylistic consistency across thousands of tokens. Despite seeing no story longer than 4k words, POLARIS–9B retains strong rubric scores at 8–12k, a regime where most open-weight models collapse in quality, undershoot requested length, or both.

2 Related Work

Long-form writing with SFT and RL. LongWriter (Bai et al., 2025) shows that SFT on synthetic long-output data can make models produce 10k+ word outputs. LongWriter-Zero (Wu et al., 2026) employs RL training, but uses a 32B base model and trained rewards for length, quality, and formatting. RLMR (Liao et al., 2026) also uses a trained writing reward model, along with checks for hard constraints. Writing-RL (Lei et al., 2026) uses a specialized base model and adaptive references for long-form writing, while Writing-Zero (Jia et al., 2025) uses generative reward models for open-ended writing. RLCS (Li et al., 2026), Writer-R1 (Zhao et al., 2026), and R2-Write (Liu et al., 2026) add further steps such as human-aligned rewards, replay, reasoning, reflection, or revision. DPWriter (Cao et al., 2026) focuses on diverse planning during creative-writing RL.

RL optimization for creative writing. DPO (Rafailov et al., 2023) gives the standard preference-optimization objective, while SWAG (Pei et al., 2024) adapts preference learning to storytelling. CrPO (Ismayilzada et al., 2025) studies preference optimization for creative writing and DivPO (Lanchantin et al., 2025) adds an explicit diversity objective. LUFFY (Yan et al., 2026) uses off-policy guidance for reasoning RL, while REPO (Li et al., 2025) reuses earlier policy outputs during optimization. G2RPO-A (Guo et al., 2025b) adds adaptive guidance inside group relative policy optimization, and BREAD (Zhang et al., 2026) branches from expert anchors to connect SFT and RL.

Creative-writing evaluation. WritingBench (Wu et al., 2025) covers broad generative writing, EQ-Bench (Paech, 2024) includes creative and emotional writing judgments, and LongBench-Write (Bai et al., 2025) focuses on long-form generation. HelloBench (Que et al., 2024) and LitBench (Fein et al., 2026) test for long text generation and creative-writing quality. Li et al. (2024) and Gu et al. (2026) analyze the values and biases of LLM-as-a-judge. Prometheus (Kim et al., 2024a,b) shows that judge models can give fine-grained rubric-based feedback. Igniting Creative Writing (Wei et al., 2025) uses LLM judges to improve short creative text.

3 Method

We apply GRPO (Guo et al., 2025a) to long-form creative writing with a structured LLM-as-judge online reward and human-reference injection (HRI). Figure 2 summarizes the full pipeline.

3.1 Structured LLM-as-Judge Reward

Motivation. Prior RL approaches to creative writing rely on a reward model trained on pairwise human preference data (Wu et al., 2026; Liao et al., 2026; Li et al., 2026). Trained reward models collapse writing quality into a single scalar, require a separate training pipeline, and can go stale as training progresses and policy rollouts diverge from the reward model's training distribution. More recent work applies LLM-as-a-judge as an online reward for shorter-form creative writing (Wei et al., 2025; Lei et al., 2026), but with rubrics that are limited in scope. Existing benchmarks (Paech, 2024; Wu et al., 2025; Bai et al., 2025) were primarily designed for ranking AI-generated writing—with limited descriptions, score anchors—which makes them unsuitable for use as a reward during RL to improve creative writing. We instead query a frontier LLM judge at each GRPO step using a comprehensive structured rubric that returns per-dimension numeric scores and textual evidence, with no reward model training required.[3]

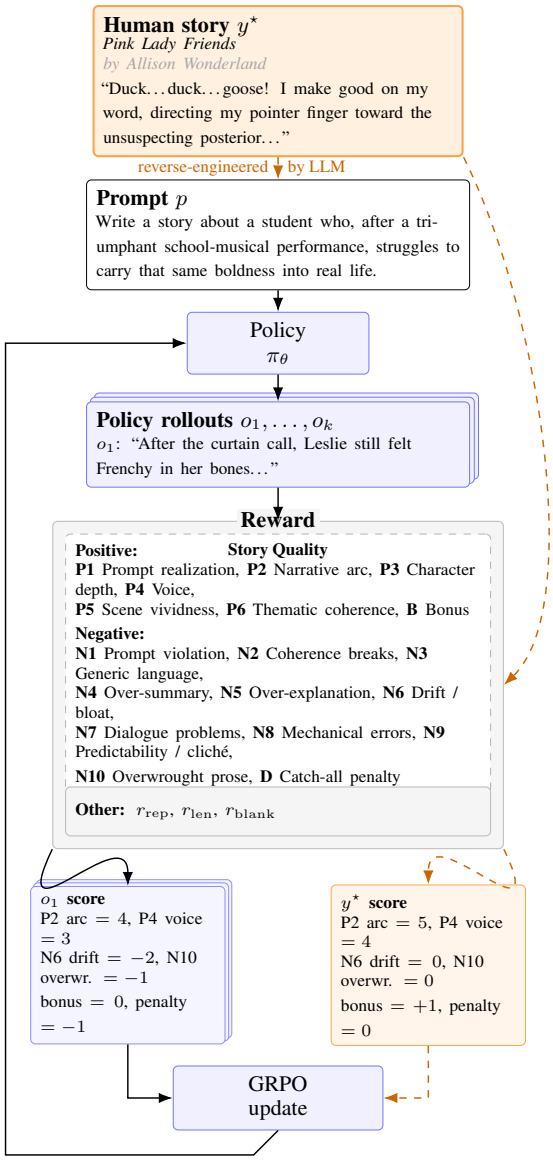


Figure 2: POLARIS training pipeline. For each prompt, the policy generates k=5 rollouts (blue) while a human reference $y^\star$ is teacher-forced as the $(k{+}1)$-th group member (orange). All outputs are scored by the reward; group statistics are computed from policy rollouts only and the reference enters the update through a warmup-scaled advantage α_t.

***Story Quality* rubric.** The *Story Quality* rubric evaluates stories on six positive dimensions that

[3] Full rubric with per-dimension anchors, non-overlap rules, and scoring anchors is in §G.3

capture sustained global qualities (prompt fulfillment, narrative arc and pacing, character depth and agency, voice and stylistic distinctiveness, scene realization, thematic and emotional richness; (P_{total}) $\in [0, 100]$) and ten negative dimensions that penalize localizable failures (prompt violation, coherence and POV consistency, generic language, over-summary, over-explanation, structural drift, dialogue problems, mechanical errors, predictability, overwrought language; $N_{\text{total}} \in [0, 151]$), plus open-ended bonus (B) and penalty (D) catch-all terms ($\in [0, 8]$ each).
The raw *Story Quality* score is:

$$s_q^{\text{raw}} = P_{\text{total}} - N_{\text{total}} + B - D \tag{1}$$

Dimensions grounded in narrative theory and narrative-quality assessment. The rubric dimensions are motivated by both narrative theory and empirical work on narrative-quality assessment. Narrative arc, pacing, scene realization, over-summary, and drift reflect concerns about large-scale narrative structure and progression (Boyd et al., 2020; Adler et al., 2018). Voice, characterization, atmosphere, and coherence are also widely emphasized in studies of how narrative writing is assessed in practice (D'Souza, 2021). Prompt fulfillment and mechanical correctness are additionally aligned with dimensions used in recent creative-writing benchmarks (Paech, 2024; Wu et al., 2025; Bai et al., 2025).

Dimensions identified from AI-writing. Three negative dimensions came from comparing LLM-generated stories against human-written references. Generic and templated language captures the stock phrases and boilerplate language characteristic of LLM-generated text. Predictability and cliché captures formulaic plot structures and recycled plot narratives; human-written stories are generally more creative, differentiating them from AI writing. Overwrought language was added after observing that models optimized toward the positive dimensions produced systematically purple prose.

Composite reward. For training, the judge score is normalized by dividing by a score divisor D,

$$\tilde{s}_q = \frac{s_q^{\text{raw}}}{D},$$

where D is rubric's maximum overall score. It is then combined with three auxiliary penalty components to arrive at the final composite score:

$$r_{\text{qual}} = w_q \cdot \tilde{s}_q \cdot g(r_{\text{rep}}, r_{\text{len}}, r_{\text{blank}}) \tag{2}$$

$$r_{\text{pen}} = w_{\text{rep}} \cdot r_{\text{rep}} + w_{\text{len}} \cdot r_{\text{len}} + w_{\text{blank}} \cdot r_{\text{blank}} \tag{3}$$

$$r = \text{clip}\big((r_{\text{qual}} - r_{\text{pen}}) \cdot \ell(n),\ -c,\ +c\big) \tag{4}$$

where r_{rep} penalizes repetition, r_{len} penalizes target-length mismatch, r_{blank} penalizes empty outputs, and g gates the positive reward to zero for severely malformed outputs. The length-scaling factor[4] counters an observed reward inflation failure mode: short outputs expose fewer sentences to negative dimensions and can receive inflated normalized scores. In our experiments, a dominant length penalty often destabilized training, so we use a sublinear scale with a floor instead. Clipping the composite reward score at c=2.0 bounds the overall reward.[5]

3.2 Human-Reference Injection

Reference as group anchor. We inject a human-written story as the $(k{+}1)$-th member of each GRPO group. The reference is teacher-forced through the current policy and scored by the same composite reward as policy rollouts. Because references typically score higher than policy outputs, especially early in training, they provide a stable high-reward anchor that maintains within-group reward variance. Three constraints govern the design: (i) the group mean and standard deviation are computed from the k policy outputs only, excluding the reference; (ii) the reference advantage is scaled by a warmup schedule ; and (iii) the reference is scored by the same reward function as policy outputs.

Relation to prior off-policy methods. Because the reference is not sampled from the policy, HRI is a biased demonstration-augmented GRPO update rather than an unbiased on-policy estimator. Prior off-policy injection methods (Yan et al., 2026; Guo et al., 2025b; Zhang et al., 2026) target verifiable-reward tasks and inject partial expert prefixes rather than complete reference outputs[6]. Our setting is also complementary to Gurung and Lapata (2025), which likewise uses human-written books and trains models to reason about long-form narrative continuation. In contrast, we focus on

[4] $\ell(n) = \max(\ell_{\min}, (n/n_{\text{target}})^{\beta})$ with β=0.5 and $\ell_{\min}$=0.15

[5] Full reward weights and hyperparameters are in §A.4.

[6] Refer to §A.8 for a detailed comparison

full-story generation rather than next-chapter prediction, and treat thinking as a means for better stories rather than as the primary object of optimization.

4 Experimental Setup

Starting from Qwen3.5 9B, POLARIS–9B was trained using $\sim 1.4k$ human-written stories on 4 A100 GPUs for about 48 hours. Our automatic evaluation spans 17 models, 5 benchmarks covering in-distribution and out-of-distribution prompts and rubrics, and over 160K per-dimension pairwise judgments.

Training data. We train on 1,388 (prompt, story) pairs from a corpus of 100 commercially purchased short-story anthologies (431 authors). No reference exceeds 4,000 words. Writing prompts are reverse-engineered from each story and references are paired with synthetic thinking traces using Gemini 3 Flash (Gemini Team, 2023; Google DeepMind, 2025). We will not release the human written stories from this corpus due to copyright concerns[7].

Model configurations. Using Qwen3.5-9B (Qwen Team, 2026) as the base model, we train two variants, both with group size 6: POLARIS–9B uses 5 policy rollouts and 1 injected human reference per group, while the matched plain GRPO run (without HRI) uses 6 policy rollouts with no reference. We also report an SFT baseline trained on the same data. Due to cost and compute considerations, we leave ablating the base model to future work.

Training details. Both models are trained on 4×A100 80GB GPUs with FSDP (Zhao et al., 2023) and batch size 8 for 160 steps, taking approximately 48 hours per run. At \$2/hr per GPU this comes to ∼\$400 in compute per checkpoint. Judge costs add approximately \$60 per run using Gemini 3 Flash (Google DeepMind, 2025) at flex tier, bringing the total cost per model to just under \$500. Full hyperparameters are in §A.4.

Test set. Our 180-prompt evaluation set is a disjoint held-out anthology pool, with 30 prompts in each of six target-length buckets: 1–2k, 2–3k, 3–4k, 4–6k, 6–8k, and 8–12k words. Each prompt is paired with a human-written reference story. Since POLARIS trains on prompts targeting up to 4k words, we treat 1–4k as *in-distribution* (ID) lengths and 4–12k as *out-of-distribution* (OOD) lengths, with OOD further split into near transfer (4–8k) and far transfer (8–12k).

Benchmarks. We evaluate along two axes: prompt distribution (ID vs. OOD) and evaluation rubric (ID vs. OOD rubric family). On ID prompts, we report *Story Quality* (our training rubric) and EQ-Bench LongForm (Paech, 2024), a related but distinct 12-dimension rubric[8]. On OOD prompts and rubrics, we report EQ-Bench Creative (Paech, 2024), English D4 (Literature & Arts) subset of WritingBench (Wu et al., 2025), and LongBench-Write (Bai et al., 2025). We also compute pairwise Elo rankings (Paech, 2024) on EQ-Bench Creative and on a 45-prompt ID subset using dual-position evaluation to mitigate positional bias in pairwise LLM judgements (Wang et al., 2024). For EQ-Bench LongForm, we weight all dimensions uniformly rather than the canonical formula to avoid rewarding catastrophic failures[9].

Judges. GPT-5.4 (OpenAI, 2026) scores Story Quality, EQ-Bench LongForm, and EQ-Bench Creative; Gemini 3.1 Pro (Google DeepMind, 2026) scores WritingBench and LongBench-Write; Gemini 3 Flash is used for pairwise Elo and as the online training reward. Using a different judge for training and evaluation reduces the risk that the model overfits to the evaluation judge's preferences[10].

5 Results

POLARIS–9B is competitive with much larger open-weight models for long-form story writing, while staying closer to requested length than several larger models. The gains are strongest on story-like tasks, transfer to related long-form rubrics, and matter most as requested length grows: at 8–12k words, where the base model and many open-weight baselines degrade sharply, POLARIS–9B keeps substantially more quality. A blinded human evaluation points in the same direction against the base model, though the comparison to Qwen3.5-27B is close.

5.1 Out-of-Distribution Prompts and Pairwise Rankings

POLARIS training transfers selectively across writing categories. Despite training only on

[7] Data construction details, corpus composition, and release scope are in §A.1.

[8] For more details, refer to A.6

[9] Refer to §E for a more detailed discussion

[10] Detailed judge comparisons and bias analyses are in §E.

Model	Writing-Bench (D4)	LongBench-Write $\bar{S}$	LongBench-Write S_Q	LongBench-Write S_L	EQ-Cr
FRONTIER					
GPT-5.4	**9.5**	83.3	**98.4**	84.7	**84.6**
Claude Opus 4.6	9.3	**86.7**	**98.4**	**88.1**	80.5
Gemini 3.1 Pro	9.2	79.8	94.2	84.8	73.8
Gemini 3 Flash	9.1	80.3	96.8	83.2	75.0
OPEN-WEIGHT					
LongWriter-Zero-32B	4.6	58.3	76.8	73.0	41.5
Gemma-4-31B	**8.7**	77.6	95.8	81.1	**70.9**
Qwen3.5-27B	8.1	77.8	85.3	**91.5**	68.9
Gemma-4-26B-A4B*	8.5	73.4	**97.6**	75.5	70.5
DeepSeek-R1-Distill-Qwen-14B	3.9	41.5	71.9	53.6	42.8
Gemma-4-E4B†	7.7	78.4	94.9	82.7	64.2
Ministral-3-8B	4.8	57.5	80.5	67.9	53.9
LongWriter-Llama3.1-8B	3.9	57.8	71.4	75.9	35.1
Qwen3.5-9B	6.8	67.1	75.1	90.3	59.2
+ POLARIS	7.9	**81.2**	90.2	88.5	70.3

(1a) OOD prompts & rubrics

#	Model	Elo
1	GPT-5.4	1911
2	Claude Opus 4.6	1783
3	POLARIS–9B	1661
4	Gemini 3.1 Pro	1627
5	Gemini 3 Flash	1620
6	Gemma-4-31B	1514
7	Qwen3.5-27B	1503
8	Gemma-4-26B A4B*	1460
9	Qwen3.5-9B	1352
10	Gemma-4-E4B†	1312
11	Ministral-3-8B	1219
12	DeepSeek-R1-Distill-Qwen-14B	1030
13	LongWriter-Zero-32B	1013
14	LongWriter-Llama3.1-8B	800

(1b) EQ-Bench Creative Elo
OOD: 32 prompts, 9 dims

#	Model	Elo
1	GPT-5.4	1574
2	Claude Opus 4.6	1468
3	Human	1464
4	POLARIS–9B	1419
5	Gemini 3.1 Pro	1387
6	Gemini 3 Flash	1335
7	Qwen3.5-27B	1285
8	Gemma-4-31B	1283
9	Gemma-4-26B-A4B*	1235
10	Gemma-4-E4B†	1108
11	Qwen3.5-9B	1086
12	Ministral-3-8B	1055
13	DeepSeek-R1-Distill-Qwen-14B	891
14	LongWriter-Llama3.1-8B	868
15	LongWriter-Zero-32B	800

(1c) ID prompts Elo
OOD dims
45 prompts, 11 judge dims + length
113K dual-position comparisons

Table 1: **(1a)** WritingBench (D4 English literature/arts: judged by Gemini 3.1 Pro), LongBench-Write (judged by Gemini 3.1 Pro), and EQ-Bench Creative (EQ-Cr, judged by GPT-5.4); all use out-of-distribution prompts and rubrics (w.r.t. our training). For LongBench-Write, $\bar{S}$ is the benchmark's overall score, S_Q its quality score, and S_L its length score. **(1b)** Pairwise Elo on EQ-Bench Creative. **(1c)** Pairwise Elo on the in-domain test subset. Elo is the Bradley–Terry MLE on the full pairwise contingency (Flash judge, dual-position), anchored so the lowest-rated model lands at 800. *MoE; †PLE.

story-writing data, both POLARIS variants transfer well to broader longform writing evaluation: they are competitive with frontier models on LongBench-Write and remain in the same general EQ-Bench Creative score band as much larger open-weight baselines. Within WritingBench, POLARIS–9B does best on subcategories closest to narrative generation and story planning, including character design, fan fiction, novel manuscript, and podcast scripting, which is consistent with its story-only training data. It is weaker on categories farther from that distribution, especially biography and book reviews, and reading reflection which place more weight on evaluative, or non-narrative writing skills (§B.4).

Pairwise Elo suggests POLARIS–9B is competitive with much larger open-weight models. The pairwise Elo results tell the same broad story. POLARIS–9B ranks **3rd of 15** on EQ-Bench Creative and **4th of 16** on the in-domain Elo, behind only frontier models and, on the in-domain set, the human references.

5.2 In-Distribution Prompts

POLARIS–9B reaches the top open-weight cluster on both rubrics. Despite using substantially less data and training compute than comparable long-form writing work, POLARIS makes a plain Qwen3.5-9B checkpoint competitive with much larger open-weight models. On *Story Quality*, POLARIS–9B sits in the top open-weight cluster; on EQ-Bench Longform, only Gemma 4 31B scores higher among open-weight models, but it often writes much shorter stories than requested, whereas POLARIS–9B stays much closer to target length.

SFT collapses; HRI gives the largest gains at far-transfer lengths. The strongest negative result is the SFT baseline, which collapses on both rubrics and solves length the wrong way: it badly overshoots requested length and degrades in quality.[11] In our matched comparison against GRPO, HRI gives a small gain at in-domain and near-OOD lengths and a larger gain at far-transfer lengths. We believe that the observed performance gap is not an upper bound and that it could increase with longer training. This is based on an empirical observation made during the training runs: validation curves were on an increasing trend and had not plateaued at the end of one epoch of training suggesting that our training was data-limited.

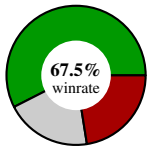
5.3 Length Generalization

Open-weight models fail in three distinct ways; POLARIS–9B avoids all three. Aggregate scores can hide long-form failure modes, so we evaluate quality as requested length increases. Open-weight models break down in three distinct

[11] Refer to §G.2 for more details

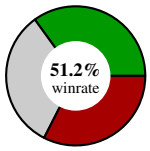
Model	P	Story Quality Agg All	Story Quality By requested length ID 1–4k	Near 4–8k	Far 8–12k	EQ-Bench Longform Agg All	EQ-Bench Longform By requested length ID 1–4k	Near 4–8k	Far 8–12k	Length adherence Agg All	Length adherence By requested length ID 1–4k	Near 4–8k	Far 8–12k
REFERENCE													
Human	–	68.7	68.1	68.2	71.4	74.2	73.9	73.1	77.1	1.00	1.00	1.00	1.00
FRONTIER MODELS													
GPT-5.4	–	**72.4**	**76.9**	**72.1**	**59.5**	**80.3**	**81.7**	**79.8**	**77.0**	1.26	1.39	1.23	0.94
Claude Opus 4.6	–	61.8	66.4	59.5	52.7	73.4	74.8	72.6	70.4	1.15	1.11	1.23	1.09
Gemini 3.1 Pro	–	57.5	58.7	55.7	57.1	67.0	69.2	65.2	64.2	0.99	1.19	0.89	0.57
Gemini 3 Flash	–	48.5	52.9	39.9	52.6	63.2	67.1	59.0	59.9	0.88	1.12	0.77	0.37
OPEN-WEIGHT MODELS													
LW-Zero-32B	32B	−27.1	−16.8	−35.5	−41.2	26.1	29.3	23.8	21.4	2.21	3.02	1.60	1.03
Gemma 4 31B	31B	51.4	53.9	49.7	**47.1**	**61.5**	**64.4**	**59.2**	**57.7**	0.68	0.87	0.55	0.36
Qwen3.5-27B	27B	42.8	51.5	38.7	24.6	57.5	62.8	54.9	46.7	0.97	1.03	0.97	0.82
Gemma 4 26B-A4B	*	51.3	52.9	**51.4**	46.5	58.1	61.4	55.8	52.8	0.65	0.91	0.45	0.26
DeepSeek-R1-Distill-Qwen-14B	14B	0.2	7.3	−4.7	−11.1	34.6	35.4	34.1	33.7	0.75	1.15	0.42	0.23
Gemma 4 E4B	†	36.1	40.5	35.7	23.7	50.3	53.8	48.2	43.9	0.72	0.95	0.58	0.34
Ministral-3-8B	8B	26.1	30.2	22.0	21.8	39.1	43.1	35.7	33.3	0.52	0.73	0.34	0.23
LW-Llama-8B	8B	−2.0	7.2	−10.5	−12.7	33.0	35.4	31.2	29.4	1.22	1.46	1.12	0.71
Qwen3.5-9B	9B	18.5	35.1	8.7	−11.8	42.6	50.2	37.2	30.3	1.01	1.09	0.96	0.88
+ POLARIS	9B	**52.1**	**57.4**	48.2	44.1	59.8	63.1	57.5	54.3	0.90	0.99	0.87	0.72

Human Eval POLARIS–9B



vs. Qwen3.5-9B

57.5% W / 20.0% T / 22.5% L



vs. Qwen3.5-27B

35.0% W / 32.5% T / 32.5% L

Table 2: **In-distribution prompts** (180 held-out prompts, GPT-5.4 judge). Left: main in-distribution results. Length adherence is generated / requested word count (1.00 = exact; <1 = undershoot; >1 = overshoot). ID lengths are 1–4k (training range); OOD lengths are 4–8k and 8–12k. Right: **Blinded human-evaluation** using 60 prompt–story pairs from EQBench Creative and LongBench Write: win/tie/loss summaries for POLARIS–9B against the base 9B and larger 27B Qwen baselines. A tie counts as a half-win for winrate purposes. *MoE, 4B active; †8B total, 4.5B effective (PLE). Full 6-bucket breakdown and 95% bootstrap CIs are in §B.

Model	Story Quality	EQ-Bench Longform	Writing-Bench	LongBench-Write	EQ-Cr
	All / 8–12k	All / 8–12k			
Qwen3.5-9B	18.5 / −11.8	42.6 / 30.3	6.8	67.1	59.2
+ SFT	−60.6 / −71.2	26.0 / 24.2	2.7	37.5	26.7
+ GRPO	49.7 / 37.7	58.2 / 51.6	7.8	**82.1**	69.7
+ POLARIS	**52.1 / 44.1**	**59.8 / 54.3**	**7.9**	81.2	**70.3**

Table 3: Comparing POLARIS–9B with matched GRPO and SFT variants. For *Story Quality* and EQ-Bench Longform, we report the aggregate score and the mean over OOD requested lengths (4–12k).

ways: quality collapse at roughly correct length (base Qwen3.5-9B), length runaway with heavy self-repetition (LongWriter-Zero-32B; §G.1), or under-generation while preserving short-form quality (Gemma 4 31B). POLARIS–9B is the only open-weight model in our comparison that largely avoids all three, maintaining a far-transfer length ratio of 0.72 while preserving rubric quality across buckets (Figure 1c).

Length-adjusted scoring places POLARIS–9B among the best in open-weights category. To summarize quality and length adherence jointly, we report a *length-adjusted* score.[12] In this view, POLARIS–9B remains in the top open-weight cluster on *Story Quality* and competitive with Qwen3.5-27B on EQ-Bench Longform, while larger Gemma baselines fall sharply once length is folded in (Figure 1a,b).

5.4 Human Evaluation

Human raters clearly prefer POLARIS–9B to the base model; the comparison to Qwen3.5-27B is effectively tied. We conduct a blinded human evaluation with two annotators on 60 prompt–generation pairs randomly sampled from EQ-Bench Creative and LongBench-Write. After 8 practice items for calibration, annotators rate each sample independently using a custom long-form writing rubric, and we derive pairwise win-rates from the overall scores. POLARIS–9B wins 67.5% of pairwise comparisons against Qwen3.5-9B (95% bootstrap CI [55.0, 80.0]) and is statistically indistinguishable from Qwen3.5-27B (POLARIS–9B winrate 51.2% vs Qwen3.5-27B; 95% bootstrap CI [38.8, 58.8]). We therefore interpret the human study narrowly: it confirms a clear gain over the base 9B model and places POLARIS–9B at parity with the larger Qwen3.5-27B model in the current two-rater external study ($ICC(A, 2) = 0.538$, Cohen's κ (quadratic-weighted)=0.364 for

[12] For signed *Story Quality*, we first normalize the raw rubric score from its floor and ceiling to $[0, 100]$, then multiply by $\min(\mathrm{LR}, 1/\mathrm{LR})$; for nonnegative EQ-Bench Longform, we use $s_{\mathrm{adj}} = s_{\mathrm{raw}} \cdot \min(\mathrm{LR}, 1/\mathrm{LR})$, analogous to the quality×length composite used in LongBench-Write (Bai et al., 2025).

Prompt & Excerpt	Human Evaluation	Automatic
Online Food Supermarket Management System (SSM). *Prompt:* Background of the research on the Design and Implementation of an Online Food Supermarket Management System Using the SSM Framework. *Excerpt:* "The rapid proliferation of information technology has fundamentally reshaped the global retail landscape, precipitating a paradigm shift known as 'Internet + Retail."	**Annotator 1 overall** 5 | **avg. dim** 4.83. *"I was beginning to fear that none of these outputs would understand the request of this prompt, but this one did it nearly perfectly… I felt like I was being taught what all of these confusing terms meant… I actually learned and enjoyed my time reading this. It was fully coherent as well."* **Annotator 2 overall** 5 | **avg. dim** 4.17. *"A solid piece explaining why the research is needed and what the model does to propel that research… while it did not stoke my interest, [it] helped me to understand the need."*	**Overall** ↑ 77.71/100 **Relevance** ↑ 5/5 **Accuracy** ↑ 5/5 **Coherence** ↑ 5/5 **Clarity** ↑ 5/5
Hacker Mindset. *Prompt:* A reformed white-hat hacker on a first date cannot stop perceiving vulnerabilities in systems and people; the story should stay close to the narrator's internal monologue. *Excerpt:* "Date target: Sarah. Initial handshake successful… For three minutes, the firewall between me and social protocol dropped to zero percent. Zero packet loss. Connection stable."	**Annotator 1 overall** 4 | **avg. dim** 3.43. *"This story is really interesting and well handled. It gets buried under a bit too much jargon, but to good effect… However, the good writing doesn't save the logical incoherence throughout the story."* **Annotator 2 overall** 4 | **avg. dim** 3.00. *"… major drift from the prompt. I get the main character inner life, but nothing much of the date. I get no sense of the location or the world."*	**Overall** ↑ 65.45/100 **Voice/tone** ↑ 18/20 **Coherence** ↑ 15/20 **Engagement** ↑ 14/20 **Overwrought** ↑ 7/20
Lost & Found in Osaka. *Prompt:* A shy synthesizer nerd in Den Den Town is drawn into an awkward but promising first conversation with a confident older busker; dialogue and character voice are central. *Excerpt:* "The air smelled like ozone, stale ramen broth, and overheated capacitors… 'Yeah?' You look like you're trying to tune a radio station in your brain.' "	**Annotator 1 overall** 4 | **avg. dim** 3.71. *"This story has a lot of great flavor to it, and the jargon-heavy approach and puns are a nice touch, but a large portion of them fall flat and feel too 'try hard'… the dialogue becomes awkward… [and] the story… gets a little muddled."* **Annotator 2 overall** 3 | **avg. dim** 3.71. *"I have a friend, a music nerd who is slightly autistic. This story made me feel like I was talking with him. Just on the edge of understanding… More feeling what he meant than comprehension."*	**Overall** ↑ 64.77/100 **Voice/tone** ↑ 15/20 **Coherence** ↑ 16/20 **Imagery** ↑ 16/20 **Weak dialogue** ↑ 10/20

Table 4: Illustrative examples from the human evaluation of stories generated by POLARIS–9B. Human comments are shown alongside overall scores and average non-overall dimension scores. Automatic scores are benchmark-specific; ↑ indicates that higher is better. For more examples, see Table 21.

overall score).[13]

5.5 Qualitative Analysis

POLARIS–9B excels in voice, prompt fulfillment, and commitment to a chosen narrative framing. On stronger examples, POLARIS–9B takes more stylistic risks and writes prose with a clear texture and purpose without drifting into generic filler, remaining committed to the prompt requirements. For example (see Table 4), in *Online Food Supermarket Management System (SSM)* annotators explicitly contrast POLARIS–9B's strong prompt fulfillment with baseline responses, while in *Lost & Found in Osaka* they highlight the piece's distinctive voice and texture. This is consistent with the dimension-level human scores, where POLARIS–9B is relatively strong on prompt fulfillment, audience and voice, and information adequacy. POLARIS–9B's writing, unlike the base model's, read as finished pieces rather than loosely related continuations.

POLARIS–9B's writing suffers from stylistic overloading and local coherence failures. Two issues recur in POLARIS–9B's writing: First, Annotators find the language to be overly heavy, filled with jargon, or somewhat over-insistent. While they appreciate the ambition and flavor of the prose, they feel that it is pushing too hard on specificity making the text read awkward. Second, annotators repeatedly note local coherence problems: contradicting details and confusing transitions. Both are visible in the examples (see Table 4): *Hacker Mindset* and *Lost & Found in Osaka* are strong in voice but are either incoherent at places or filled with heavy language. While these failures do not usually completely derail the plot, they do weaken otherwise strong pieces by making parts of the narrative feel less gripping or worse confusing.

6 Conclusion

We present POLARIS, a lower-compute recipe for long-form creative-writing RL. A frontier LLM judge with a structured *Story Quality* rubric replaces the trained reward model, while *human-reference injection* adds a single teacher-forced group member to each GRPO group, excluded from group statistics and scaled by a warmup. Applied to Qwen3.5-9B and trained on ∼1.4K prompt–story pairs from a collection of short-story anthologies with batch size 8 on 4 A100 GPUs, POLARIS yields a 9B model that is on par with much larger open-weight baselines while adhering to length requests much more closely. Despite seeing no training reference longer than 4k words, it also preserves rubric quality on requests up to 3× that length. We view this minimal GRPO recipe as a promising approach for other open-ended tasks with soft rewards.

[13] §F gives full study details, agreement statistics, and dimension-level breakdowns.

Limitations

✧ **Judge validity.** LLM judges may share systematic blind spots (e.g., prose style preferences). The *Story Quality* rubric was designed with the goal of differentiating Human and AI writing but by design cannot be fully exhaustive in what it tests for. We intentionally use LLMs from different model families (training: Gemini 3 Flash, testing: GPT 5.4) to mitigate the effects of overfitting to judge biases to the extent possible, but this does not fully mitigate the associated risks. Results on other external benchmarks with different prompts and rubrics along with human evaluation together is meant to serve as evidence of our models' performance.
✧ **Data availability.** Human stories from commercially purchased anthologies used in this work cannot be distributed introducing challenges to reproducibility. We intend to release prompts derived from the dataset instead, along with stories generated by LLMs tested in this work.
✧ **Human evaluation scope.** Our human study is small and is designed to validate the final model's overall quality ranking against base Qwen3.5-9B and Qwen3.5-27B.
✧ **Length compliance.** POLARIS–9B maintains rubric quality at long target lengths but undershoots the requested length (length ratio ≈ 0.72 at 8–12k).
✧ **Domain coverage.** Since our primary focus in this work was story writing, we did not directly test how including creative writing data outside of short-story anthologies, eg. screenwriting, reviews, essays etc, during training affects the training dynamics and the performance of the resultant model.
✧ **Synthetic thinking traces.** Each reference is paired with a thinking trace generated by Gemini 3 Flash (§4). A clean ablation separating the effects of these two components is left to future work.

Ethics Statement

Use of copyrighted story data. Our training data is derived from commercially purchased short-story anthologies. We do not release the underlying stories, full prompt–story pairs, or any other copyrighted source text. This dataset is used only for academic research on long-form creative-writing alignment and evaluation. We do not endorse the use of copyrighted books or stories without appropriate rights or licenses for commercial model training or deployment.

Evaluation and annotator limitations. Our main evaluations rely heavily on LLM-based judges, supplemented by a blinded human study. Although we use multiple rubric families and separate training-time and headline evaluation judges, these measurements are still imperfect proxies for literary quality and reader preference. Human judgments in creative writing are inherently subjective, and should be interpreted as qualitative validation rather than definitive evidence.

AI disclosure. Large language models and coding agents were used to assist with and refine writing, and the preparation of some tables and figures. All final technical claims, experimental decisions, and paper content were curated and reviewed by the authors.

References

Jonathan M. Adler, Theodore E.A. Waters, Justin Poh, and Sophia Seitz. 2018. The nature of narrative coherence: An empirical approach. *Journal of Research in Personality*, 74:30–34.

Anthropic. 2026. Claude Opus 4.6. https://www.anthropic.com/news/claude-opus-4-6. Accessed: 2026-05-25.

Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2025. Longwriter: Unleashing 10,000+ word generation from long context LLMs. In *The Thirteenth International Conference on Learning Representations*.

Ryan L. Boyd, Kate G. Blackburn, and James W. Pennebaker. 2020. The narrative arc: Revealing core narrative structures through text analysis. *Science Advances*, 6(32):eaba2196.

Qian Cao, Yahui Liu, Wei Bi, Yi Zhao, Ruihua Song, Xiting Wang, Ruiming Tang, Guorui Zhou, and Han Li. 2026. Dpwriter: Reinforcement learning with diverse planning branching for creative writing. *Preprint*, arXiv:2601.09609.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association.

Richard D'Souza. 2021. What characterises creativity in narrative writing, and how do we assess it? research findings from a systematic literature search. *Thinking Skills and Creativity*, 42:100949.

Daniel Fein, Sebastian Russo, Violet Xiang, Kabir Jolly, Rafael Rafailov, and Nick Haber. 2026. LitBench:

A benchmark and dataset for reliable evaluation of creative writing. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7740–7755, Rabat, Morocco. Association for Computational Linguistics.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Google DeepMind. 2025. Gemini 3 Flash model card. https://deepmind.google/models/model-cards/gemini-3-flash/. Accessed: 2026-05-22.

Google DeepMind. 2026. Gemini 3.1 Pro model card. https://deepmind.google/models/model-cards/gemini-3-1-pro/. Accessed: 2026-05-22.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Zhouchi Lin, Bowen Zhang, Lionel Ni, Wen Gao, Yuanzhuo Wang, and Jian Guo. 2026. A survey on llm-as-a-judge. *The Innovation*, page 101253.

Daya Guo, Dejian Yang, Haowei Zhang, and 1 others. 2025a. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Yongxin Guo, Wenbo Deng, Zhenglin Cheng, and Xiaoying Tang. 2025b. G^2rpo-a: Guided group relative policy optimization with adaptive guidance. *Preprint*, arXiv:2508.13023.

Alexander Gurung and Mirella Lapata. 2025. Learning to reason for long-form story generation. In *Second Conference on Language Modeling*.

Jamie Hayes, Marika Swanberg, Harsh Chaudhari, Itay Yona, Ilia Shumailov, Milad Nasr, Christopher A. Choquette-Choo, Katherine Lee, and A. Feder Cooper. 2025. Measuring memorization in language models via probabilistic extraction. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9266–9291, Albuquerque, New Mexico. Association for Computational Linguistics.

Mete Ismayilzada, Antonio Laverghetta Jr., Simone A. Luchini, Reet Patel, Antoine Bosselut, Lonneke Van Der Plas, and Roger E. Beaty. 2025. Creative preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9580–9609, Suzhou, China. Association for Computational Linguistics.

Ruipeng Jia, Yunyi Yang, Yongbo Gai, Kai Luo, Shihao Huang, Jianhe Lin, Xiaoxi Jiang, and Guanjun Jiang. 2025. Writing-zero: Bridge the gap between non-verifiable tasks and verifiable rewards. *Preprint*, arXiv:2506.00103.

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

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024b. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics.

Jack Lanchantin, Angelica Chen, Shehzaad Dhuliawala, Ping Yu, Jason Weston, Sainbayar Sukhbaatar, and Ilia Kulikov. 2025. Diverse preference optimization. *Preprint*, arXiv:2501.18101.

Xuanyu Lei, Chenliang Li, Yuning Wu, Kaiming Liu, Weizhou Shen, Peng Li, Ming Yan, Fei Huang, Ya-Qin Zhang, and Yang Liu. 2026. Writing-RL: Advancing long-form writing via adaptive curriculum reinforcement learning.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *Preprint*, arXiv:2412.05579.

Siheng Li, Zhanhui Zhou, Wai Lam, Chao Yang, and Chaochao Lu. 2025. Repo: Replay-enhanced policy optimization. *Preprint*, arXiv:2506.09340.

Zhaoyan Li, Hanag Lei, Yujia Wang, Lanbo Liu, Hao Liu, and Liang Yu. 2026. Rewarding creativity: A human-aligned generative reward model for reinforcement learning in storytelling. *Preprint*, arXiv:2601.07149.

JianXing Liao, Tian Zhang, Xiao Feng, Yusong Zhang, Haorui Wang, Bosi Wen, Ziying Wang, and Runzhi Shi. 2026. Rlmr: Reinforcement learning with mixed rewards for creative writing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(38):31970–31978.

Wanlong Liu, Bo Zhang, Chenliang Li, Shaopeng Lai, Yuning Wu, Xuanyu Lei, and Ming Yan. 2026. R2-write: Reflection and revision for open-ended writing with deep reasoning. *Preprint*, arXiv:2604.03004.

OpenAI. 2026. GPT-5.4 Thinking System Card. https://deploymentsafety.openai.com/gpt-5-4-thinking/gpt-5-4-thinking.pdf. Accessed: 2026-05-25.

Samuel J. Paech. 2024. Eq-bench: An emotional intelligence benchmark for large language models. *Preprint*, arXiv:2312.06281.

Jonathan Pei, Zeeshan Patel, Karim El-Refai, and Tianle Li. 2024. SWAG: Storytelling with action guidance. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14086–14106, Miami, Florida, USA. Association for Computational Linguistics.

Haoran Que, Feiyu Duan, Liqun He, Yutao Mou, Wangchunshu Zhou, Jiaheng Liu, Wenge Rong, Zekun Moore Wang, Jian Yang, Ge Zhang, Junran Peng, Zhaoxiang Zhang, Songyang Zhang, and Kai Chen. 2024. Hellobench: Evaluating long text generation capabilities of large language models. *Preprint*, arXiv:2409.16191.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents. Accessed: 2026-05-25.

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. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

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 17160–17186, Suzhou, China. Association for Computational Linguistics.

Yuhao Wu, Yushi Bai, Zhiqiang Hu, Roy Ka-Wei Lee, and Juanzi Li. 2026. Longwriter-zero: Mastering ultra-long text generation via reinforcement learning. In *The Fourteenth International Conference on Learning Representations*.

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.

Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. 2026. Learning to reason under off-policy guidance. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, YuYue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. 2026. DAPO: An open-source LLM reinforcement learning system at scale. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Xuechen Zhang, Zijian Huang, Yingcong Li, Chenshun Ni, Jiasi Chen, and Samet Oymak. 2026. BREAD: Branched rollouts from expert anchors bridge SFT & RL for reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Jihao Zhao, Shuaishuai Zu, Zhiyuan Ji, Chunlai Zhou, and Biao Qin. 2026. Writer-r1: Enhancing generative writing in llms via memory-augmented replay policy optimization. *Preprint*, arXiv:2603.15061.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. 2023. Pytorch fsdp: Experiences on scaling fully sharded data parallel. *Proc. VLDB Endow.*, 16(12):3848–3860.

Appendix

A Data, Training, and Method Details

This section collects the appendix material needed to interpret and reproduce the setup behind the main results: corpus composition, release scope, optimization settings, decoding choices, reward details, and the exact form of HRI used in our GRPO updates.

A.1 Training Corpus Summary

Table 5 summarizes the 4k-word training subset used in this work. The training split contains 1,388 prompt–story pairs drawn from 100 commercially purchased anthologies spanning 431 unique authors. Genre, era, tone, and collection-type labels come from anthology-level metadata and are therefore best interpreted as corpus-composition indicators rather than exact per-story genre annotations.

A.2 Data Availability and Release Scope

The training and in-domain evaluation stories are derived from 100 commercially purchased short-story anthologies. We do not distribute the raw story texts with the paper. This limits exact end-to-end reproduction from source texts, so the paper aims for procedural reproducibility rather than redistribution of the anthology corpus itself.

Corpus statistics		Top primary genres	
Training stories	1,388	literary realism	28.1%
Anthologies	100	horror / weird / gothic	13.0%
Unique authors	431	sci-fi / speculative	12.2%
Mean story length	2,387 words	regional / folk / vernacular	8.8%
Median story length	2,299 words	modernist / experimental	8.1%
Min / max story length	1,000 / 3,996 words	humor / satire	6.7%
Length buckets	38.5% / 33.4% / 28.2%	fantasy	6.6%
	at 1–2k / 2–3k / 3–4k	romance	4.1%
Collection type	68.6% single-author, 31.4% multi-author	crime / mystery	3.1%
Dominant eras	27.9% 2010+, 20.1% 1950–1990,	essay / creative nonfiction	2.7%
	19.6% 1900–1950		
Dominant tone	43.1% literary serious, 15.6% literary playful / ironic		

Table 5: Detailed summary of the 4k-word training subset used in this work. Genre, era, tone, and collection-type labels come from anthology-level metadata; percentages are weighted by the number of stories drawn from each anthology.

Attack	n	LCS ≥ 50	LCS ≥ 200	Exact 2048
Prompt-only	500	0.0%	0.0%	0.0%
Gold-thinking, prefix 50	134	17.2%	16.4%	6.7%
Gold-thinking, prefix 100	133	17.3%	16.5%	6.0%
Gold-thinking, prefix 200	133	17.3%	16.5%	6.8%

Table 6: Memorization audit under two different attacker regimes. Prompt-only uses only the released writing prompt. Gold-thinking additionally provides the exact training-time reasoning trace and a short story prefix. Metrics are exact token-overlap against the gold training continuation under greedy decoding.

A.3 Memorization Audit for Prompt Release

To inform prompt-release decisions, we ran a memorization audit on POLARIS–9B using exact token-overlap metrics, following prior work on training-data extraction and discoverable extraction (Carlini et al., 2021; Hayes et al., 2025). We distinguish between a realistic *prompt-only* attacker, who has access only to the released writing prompt, and a privileged *gold-thinking* attacker, who additionally has the exact training-time reasoning trace and a short story prefix. All generations use greedy decoding and are compared against the gold training continuation using exact-prefix-match and longest-common-substring (LCS) overlap.

The prompt-only attack remained clean at scale. Across 500 prompt-only evaluations with 2048 generated tokens, we observed 0/500 cases with LCS ≥ 50, ≥ 200, ≥ 512, ≥ 1024, or ≥ 2048, and 0/500 exact recoveries at 512, 1024, or 2048 tokens. Overlap above trivial short spans was essentially absent: only 11/500 cases reached LCS ≥ 10, and none reached LCS ≥ 20.

By contrast, the privileged white-box attack showed clear recoverable memorization. Tested on 400 training instances, in the gold-thinking attacker regime, we observed substantial exact overlap across all three prefix settings (50, 100, and 200 tokens): roughly 17% of evaluations had LCS ≥ 50, 16–17% had LCS ≥ 200, 14–15% had LCS ≥ 1024, and 6–7% exactly recovered the full 2048-token continuation. This confirms that the model contains recoverable memorized continuations under strong privileged scaffolding.

The key result is the gap between attacker models. We find no evidence of substantial long-span verbatim recovery in the realistic prompt-only setting relevant to public prompt release, while recoverable memorization is visible under a much stronger white-box attack that depends on non-released artifacts. We therefore interpret this audit as supporting prompt release for reproducibility while continuing to withhold human stories and reasoning traces.

A.4 Training Hyperparameters

We describe the training configuration in Table 7.

A.5 Generation and Decoding Settings

The training rollouts use a shared 8192-token budget, but the 180-prompt main evaluation does not: generation budgets are model-family specific and reflect the configurations used in the original runs rather than a single harmonized cap.

Evaluation judges score the final story only: model-internal thoughts used during training are never shown to judges and do not count toward length adherence.

A.6 Relation Between *Story Quality* and EQ-Bench Longform

Both EQ-Bench Longform and *Story Quality* evaluate long-form writing quality and overlap on broad concerns such as coherence, prompt faithfulness, stylistic control, characterization, and reader-level

Parameter	Value
Base model	Qwen3.5-9B
Learning rate	1×10^{-6}
Batch size	8 GRPO groups
PPO epochs	1
Clip ratio	0.20
Group size (k)	6 (5 policy + 1 human)
KL loss	Disabled
Entropy coefficient	0.0
Rollout temperature (train/val)	1.0 / 0.8
Top-p	0.95
Top-k	20
Repetition penalty	1.10
Max response length	8,192 tokens
HRI warmup steps	15
HRI peak weight	0.4
Composite clip	2.0
Reward weights	
Story quality (w_q)	1.0
Self-repetition (w_{rep})	1.5
Length penalty (w_{len})	0.8
Blank penalty (w_{blank})	1.0
Infrastructure	
GPUs	4× A100 80GB
Distributed strategy	FSDP
Gradient checkpointing	Enabled

Table 7: Training hyperparameters for POLARIS–9B.

engagement. Improvement on EQ-Bench Longform should therefore not be interpreted as transfer to a completely unrelated evaluation family.

At the same time, the two rubrics differ substantially in both decomposition and purpose. *Story Quality* is a custom anchored rubric designed for online RL and diagnostic analysis. It separates strengths such as prompt realization, narrative arc, character depth, voice, scene vividness, and thematic coherence from distinct failure modes such as coherence breaks, generic language, over-summary, over-explanation, drift or bloat, dialogue problems, mechanical errors, predictability, and overwrought prose. This positive/negative decomposition is intended to provide a usable training signal and to distinguish common long-form failure modes that would otherwise collapse into a single scalar score.

By contrast, EQ-Bench Longform uses 12 benchmark-style dimensions: *Nuanced Characters*, *Emotionally Engaging*, *Compelling Plot*, *Coherent*, *Well-earned Lightness or Darkness*, *Faithful to Writing Prompt*, *Weak Dialogue*, *Tell-Don't-Show*, *Unsurprising or Uncreative*, *Amateurish*, *Purple Prose*, and *Forced Poetry or Metaphor*. These dimensions capture many of the same broad qualities as *Story Quality*, but they are expressed as shorter benchmark labels rather than as an explicitly anchored RL-oriented rubric. As a result, EQ-Bench Longform is less diagnostic about why a story succeeds or fails, and its aggregation can behave differently from our rubric's positive/negative decomposition.

We therefore treat EQ-Bench Longform as meaningful but limited evidence of transfer. A gain on EQ-Bench Longform shows that improvements generalize beyond the exact wording and scoring decomposition of *Story Quality*. However, because the two rubrics still overlap at the level of broad long-form writing quality, we do not treat EQ-Bench Longform as a fully independent evaluation in the same sense as WritingBench or LongBench-Write, which test broader and more structurally different writing tasks.

A.7 Reward Component Details

The composite reward in Section 3.1 combines normalized *Story Quality* with three lightweight penalties and a positive gate. Let ρ_4 be the global repeated 4-gram ratio, ρ_{loc} the local 4-gram repeat ratio in a 64-gram sliding window, and ρ_{line} the duplicate-line ratio. Our self-repetition penalty is

$$r_{\text{rep}} = \min\Big(1.5,\ 0.8\big(0.7\max(0, \rho_4 - 0.08) + 0.3\rho_{\text{loc}}\big) + 0.4\rho_{\text{line}}\Big). \tag{5}$$

The blank penalty is binary:

$$r_{\text{blank}} = \mathbf{1}[\text{visible-char-ratio} < 0.10 \vee \text{word-count} < 150]. \tag{6}$$

For length, we infer a target word count t from the prompt when an explicit request is present; otherwise we fall back to the reference-story length. The tolerance band is asymmetric: the lower tolerance is $\ell = 0.15t$ clipped to $[350, 1000]$ words, and the upper tolerance is $u = 0.25t$ clipped to $[600, 1500]$ words. Let m denote the soft upper cap on output length (the configured `max_words`); if no explicit cap is set, we use $m = t + u$. The resulting piece-

Model group	Generation settings	Notes
POLARIS Training	**Budget:** 8192 tokens. **Sampling:** thinking enabled in training; temperature 1.0 / 0.8 (val); top-p 0.95; top-k 20 / -1 (val); repetition penalty 1.10.	Training-time (<4k words) config for POLARIS–9B, plain-GRPO and SFT runs.
POLARIS Evaluation	**Budget:** 14,336 tokens. **Sampling:** checkpoint-specific settings; we use temperature 0.6 (*Story Quality*, EQBench Long-Form, EQBench Creative) or 0.8 (Writing Bench, LongBench Write), top-p 0.95, top-k 20, repetition penalty 1.0.	Evaluation config for the final POLARIS–9B, plain-GRPO and SFT models.
Local Qwen baselines	**Budget:** 32,768 tokens. **Sampling:** temperature 1.0; top-p 0.95; top-k 20; presence penalty 1.5; repetition penalty 1.0.	Applied to the Qwen3.5-9B and Qwen3.5-27B.
Other open-weight baselines	**Budget and sampling:** model-specific recommended decoding config from the official HF page or provider documentation. Token budgets: 14,336 (e.g., Gemma 4, Ministral-3-8B) and 32,768 (e.g., DeepSeek-R1-14B, LongWriter-Llama-8B).	We preserve recommended defaults when the interface does not expose all low-level controls.
Frontier generation models	**Budget and sampling:** provider-managed settings; models use 16,384-token output budgets.	Temperature and some low-level controls are not exposed uniformly across APIs.
GPT-5.4 rubric eval	**Budget and sampling:** provider-managed output settings; medium reasoning.	Used for *Story Quality*, EQ-Bench LongForm, and EQ-Bench Creative scoring.
Gemini judges	**Budget and sampling:** provider-managed output settings; medium thinking where supported; temperature 0.2; top-p 0.9.	Used for training reward and selected external benchmarks.

Table 8: **Decoding settings.** This table summarizes the main generation settings used for training and evaluation. Generation budgets in the 180-prompt main test set were not uniformly 8,192 tokens: custom-trained models used 14,336 tokens, frontier generation APIs typically used 16,384 tokens, and some open-weight baselines used larger budgets up to 32,768 depending on model's observed tendency to overwrite.

wise penalty is

$$r_{\text{len}} = \begin{cases} 0, & t-\ell \le wc \le t+u, \\ 1-\dfrac{wc}{t-\ell}, & wc < t-\ell, \\ \min\left(1, \dfrac{wc}{t+u} - 1\right), & wc > t+u, \; m \le t+u, \\ 1-\dfrac{m-wc}{m-(t+u)}, & wc > t+u, \; m > t+u. \end{cases} \tag{7}$$

where wc is the generated word count. The positive gate is multiplicative:

$$\begin{aligned} g = \mathbf{1}[r_{\text{blank}} < 0.5] \\ \cdot \operatorname{clip}\left(1 - \frac{r_{\text{rep}}}{0.80}, 0, 1\right) \\ \cdot \mathbf{1}[r_{\text{len}} \le 0.80]. \end{aligned} \tag{8}$$

All reported trained checkpoints use w_q=1.0, w_{rep}=1.5, w_{len}=0.8, w_{blank}=1.0, and composite clip c=2.0. The online *Story Quality* judge uses temperature 0.2 and top-p 0.9.

A.8 HRI Design Choices and Alternatives

The three constraints in Section 3.2 (μ, σ computed from policy rollouts only; advantage scaled by α_t; reference scored identically to policy outputs) are the only modifications required to use a human reference as a demonstration-like trajectory. Because the reference is teacher-forced under the current policy, current-policy log probabilities are available and the reference can be optimized through the same clipped surrogate used for policy rollouts. The reference is not sampled from the policy, however, so HRI is a biased demonstration-augmented GRPO update rather than an unbiased on-policy estimator; we omit importance correction by design and use the warmup α_t to dampen the demonstration-side gradient. No per-example gate decides whether to include it, no failure condition triggers it, and no auxiliary loss is attached.

To our knowledge, no prior off-policy injection method in either verifiable-reward or soft-reward settings combines these properties: a complete reference output injected as a demonstration trajectory, excluded from group statistics, and included in the policy update with a single warmup-scheduled

scalar weight. LUFFY (Yan et al., 2026), G2RPO-A (Guo et al., 2025b), and BREAD (Zhang et al., 2026) target verifiable-reward tasks and inject partial expert prefixes rather than complete reference outputs; they also operate in settings where reward is binary, which makes variance collapse a less acute concern. We view the minimality as a feature: the training loop is a GRPO update with one additional teacher-forced demonstration and a scalar advantage multiplier, which keeps the contribution easy to reproduce, easy to ablate, and easy to combine with other GRPO extensions.

B Expanded Benchmark Results

This section expands the main-paper benchmark tables with the bucket-level and category-level results that are too detailed to fit in the main text.

B.1 Full Length-Bucket Breakdown: Story Quality

Table 9 expands the main-paper *Story Quality* results into the full six requested-length buckets and reports a simple slope summary across buckets.

B.2 Full Length-Bucket Breakdown: EQ-Bench Longform

Table 10 provides the corresponding six-bucket breakdown for EQ-Bench Longform, which helps separate in-distribution behavior from near- and far-transfer performance.

B.3 Length Ratio by Bucket

Table 11 reports generated/requested length ratios by bucket, complementing the quality tables with a direct view of under- and over-generation.

B.4 WritingBench Category Breakdown

Tables 12–14 unpack the single WritingBench column in the main paper into per-category means, showing where story-only training transfers cleanly and where it does not.

C Training and Evaluation Diagnostics

The next set of analyses explains how the trained models improve over training, how stable the judges are, and how much uncertainty remains in the reported bucketed comparisons.

C.1 Per-Dimension Pattern

The per-dimension training curves in Figure 3 track the full *Story Quality* core rubric: 16 dimensions (6 positive, 10 negative) scored by the training judge on held-out validation rollouts. We use this section to inspect which dimensions improve over training and how the plain-GRPO and HRI runs differ.

At a high level, the same pattern seen in the final GPT-5.4 evaluation appears throughout training: the clearest positive gains are in voice, character depth, narrative arc, and world/scene realization, while the largest negative-side reductions are in generic language, drift/bloat, coherence/POV, over-explanation, and over-summary. Prompt fulfillment and mechanical errors move the least, suggesting that most of the gain comes from higher-level narrative quality rather than surface cleanup.

Training-time dynamics. Figure 3 plots per-dimension eval scores as a percent of each dimension's maximum over training steps for both Qwen3.5 9B +GRPO and POLARIS–9B. Most negative-side reduction happens early, while positive dimensions continue improving later into training. Voice and several drafting-failure dimensions account for much of the separation between the runs, and across most tracked dimensions POLARIS–9B finishes above Qwen3.5 9B +GRPO, consistent with the *Story Quality* gap in Table 2. Overwrought prose is a weaker exception, which may indicate that it is tied more closely to base-model stylistic priors than to the higher-level drafting failures that lightweight RL corrects more readily.

HRI's per-dimension signature. Comparing Qwen3.5 9B+GRPO and POLARIS–9B at their final checkpoints, HRI is associated with slightly larger reductions in generic language, predictability, and dialogue. These per-dimension effects are small, but directionally consistent with HRI helping reduce familiar LLM-writing artifacts in addition to its larger role in far-transfer length robustness.

Validation training curves. Figure 4 shows held-out *Story Quality* and composite reward over training steps for Qwen3.5 9B+GRPO and POLARIS–9B on a 100-prompt validation split (in-distribution length, 1–4k words; Gemini 3 Flash judge, the training-reward judge). Both runs improve steadily over the course of one epoch, with POLARIS–9B remaining clearly ahead of Qwen3.5 9B+GRPO on both metrics throughout. The ordering matches the headline GPT-5.4 evaluation in Table 2, and both curves are still rising at the end of training, so the reported checkpoint is a cost-based stopping point

Model	P	1–2k	2–3k	3–4k	4–6k	6–8k	8–12k	Agg	Slope
Human	–	66.4	64.9	73.0	70.0	66.5	71.4	68.7	+0.8
GPT-5.4	–	74.6	76.3	79.8	78.0	66.3	59.5	72.4	−3.1
GPT-5.4-mini	–	69.5	71.3	70.8	62.0	60.5	59.0	65.5	−2.6
Claude Opus 4.6	–	65.5	66.4	67.3	62.7	56.4	52.7	61.8	−2.8
Gemini 3.1 Pro	–	60.9	56.6	58.6	58.0	53.4	57.1	57.5	−0.8
GPT-5.4-nano	–	57.4	54.6	52.6	48.9	47.3	43.2	50.7	−2.7
Gemini 3 Flash	–	57.5	52.1	49.1	40.1	39.8	52.6	48.5	−2.0
Qwen3.5-27B	27B	54.6	52.1	47.8	42.9	34.5	24.6	42.8	−5.9
Qwen3.5-9B	9B	40.9	33.1	31.3	18.7	−1.2	−11.8	18.5	−10.8
+ SFT	9B	−40.3	−61.2	−57.4	−64.3	−69.4	−71.2	−60.6	−5.3
+ GRPO	9B	58.2	53.8	54.4	48.1	45.8	37.7	49.7	−3.8
+ POLARIS	9B	57.7	55.8	58.6	50.7	45.7	44.1	52.1	−3.0
Gemma 4 31B	31B	54.0	53.6	54.1	49.7	49.7	47.1	51.4	−1.4
Gemma 4 26B-A4B	26B	53.2	52.1	53.3	52.2	50.6	46.5	51.3	−1.1
Gemma 4 E4B	8B	42.9	39.4	39.2	35.1	36.3	23.7	36.1	−3.1
DeepSeek-R1-14B	14B	12.8	1.9	7.3	−1.6	−7.7	−11.1	0.2	−4.5
Ministral-3-8B	8B	35.1	28.8	26.6	23.5	20.6	21.8	26.1	−2.7
LongWriter-Zero-32B	32B	10.4	−27.2	−33.7	−38.2	−32.9	−41.2	−27.1	−8.0
LongWriter-Llama-8B	8B	14.9	0.9	5.7	−6.3	−14.7	−12.7	−2.0	−5.6

Table 9: *Story Quality* (GPT-5.4 judge) by length bucket, 180 held-out test prompts (30 per bucket). Slope = linear fit across 6 buckets (points per step). Shaded = our models. Per-bucket numbers match the aggregates in Table 2.

Model	P	1–2k	2–3k	3–4k	4–6k	6–8k	8–12k	Agg	Slope
Human	–	72.0	72.5	77.2	72.8	73.3	77.1	74.2	+0.7
GPT-5.4	–	80.8	82.0	82.4	81.1	78.6	77.0	80.3	−0.9
GPT-5.4-mini	–	78.5	77.5	77.0	74.0	72.9	71.7	75.3	−1.4
Claude Opus 4.6	–	74.7	74.8	74.9	72.9	72.3	70.4	73.4	−0.9
Gemini 3.1 Pro	–	70.5	68.2	69.0	65.7	64.7	64.2	67.0	−1.3
GPT-5.4-nano	–	69.3	68.2	66.6	64.7	63.4	60.9	65.5	−1.7
Gemini 3 Flash	–	69.0	65.8	66.5	59.3	58.6	59.9	63.2	−2.1
Qwen3.5-27B	27B	64.1	63.4	61.0	56.3	53.4	46.7	57.5	−3.5
Qwen3.5-9B	9B	54.5	48.4	47.8	40.2	34.2	30.3	42.6	−4.9
+ SFT	9B	29.2	25.5	26.5	25.0	25.5	24.2	26.0	−0.8
+ GRPO	9B	64.2	60.3	61.8	56.6	54.7	51.6	58.2	−2.4
+ POLARIS	9B	64.2	62.4	62.8	58.6	56.3	54.3	59.8	−2.1
Gemma 4 31B	31B	66.3	63.5	63.5	60.2	58.2	57.7	61.5	−1.8
Gemma 4 26B-A4B	26B	62.0	61.5	60.6	55.6	56.0	52.8	58.1	−1.9
Gemma 4 E4B	8B	55.5	53.4	52.4	48.9	47.4	43.9	50.3	−2.3
DeepSeek-R1-14B	14B	38.3	34.5	33.4	34.1	34.0	33.7	34.6	−0.7
Ministral-3-8B	8B	48.7	41.5	39.2	37.2	34.2	33.3	39.1	−2.9
LongWriter-Zero-32B	32B	36.5	26.3	25.1	22.8	24.7	21.4	26.1	−2.4
LongWriter-Llama-8B	8B	38.0	34.4	33.7	32.2	30.1	29.4	33.0	−1.6

Table 10: EQ-Bench Longform (uniform per-dim aggregation, GPT-5.4 judge) by length bucket, 180 held-out test prompts (30 per bucket). Shaded = our models. Per-bucket numbers match the aggregates in Table 2.

rather than an obvious saturation point.

Per-model length-bucket distributions. Figure 5 shows the distribution of length-adjusted *Story Quality* scores per model, broken out by length-bucket grouping (aggregate, ID, near-OOD, far-OOD). At far-OOD (8–12k), POLARIS–9B's distribution is concentrated at the open-weight ceiling, while frontier models retain a narrow high-scoring distribution.

C.2 Judge Stochasticity Measurement

To quantify judge stochasticity independently of prompt-sampling noise, we score 20 stories three times each with GPT-5.4 on *Story Quality* and EQ-Bench Longform, and 9 stories three times on EQ-Bench Creative (smaller prompt pool of 32). Stories are stratified across four score bins (SQ/LF) or three bins (EQ-Cr) spanning the full observed quality range (from LongWriter-Zero-32B outputs with SQ < 0 to GPT-5.4 and Claude Opus 4.6 outputs with SQ > 70). Each call uses medium-level thinking; GPT-5.4's reasoning API does not expose `temperature` or `seed`, so the three runs rely on the model's intrinsic stochasticity for independent samples under default sampling.

All three ICC values exceed 0.97. Judge stochasticity accounts for 1.2–2.6% of observed score variance; the remaining >97% is real between-story quality differences. *Story Quality* has the largest within-story SD (8.3 points) because its score range spans $[-180, +105]$ with correspondingly larger absolute magnitudes; the percentage variance contribution is comparable to the bounded-range benchmarks.

Model	P	1–2k	2–3k	3–4k	4–6k	6–8k	8–12k	All
Human	–	1.00	1.00	1.00	1.00	1.00	1.00	1.00
GPT-5.4	–	1.52	1.36	1.29	1.26	1.20	0.94	1.26
GPT-5.4-mini	–	1.60	1.52	1.42	1.36	1.17	0.91	1.33
Claude Opus 4.6	–	1.06	1.10	1.18	1.20	1.26	1.09	1.15
Gemini 3.1 Pro	–	1.26	1.19	1.12	0.93	0.85	0.57	0.99
GPT-5.4-nano	–	2.21	2.17	2.06	1.55	1.33	0.99	1.72
Gemini 3 Flash	–	1.21	1.08	1.06	0.86	0.69	0.37	0.88
Qwen3.5-27B	27B	1.02	1.02	1.04	1.01	0.94	0.82	0.97
Qwen3.5-9B	9B	1.06	1.07	1.14	0.95	0.98	0.88	1.01
+ POLARIS	9B	1.09	0.98	0.91	0.91	0.83	0.72	0.90
+ GRPO	9B	0.97	0.95	0.90	0.89	0.82	0.70	0.87
+ SFT	9B	3.98	3.30	2.42	1.69	1.24	0.91	2.26
Gemma 4 31B	31B	1.01	0.85	0.74	0.64	0.47	0.36	0.68
Gemma 4 26B-A4B	26B	1.10	0.90	0.74	0.51	0.38	0.26	0.65
Gemma 4 E4B	8B	1.04	0.95	0.85	0.64	0.53	0.34	0.72
DeepSeek-R1-14B	14B	1.73	0.87	0.84	0.64	0.19	0.23	0.75
Ministral-3-8B	8B	0.93	0.70	0.56	0.38	0.30	0.23	0.52
LongWriter-Zero-32B	32B	2.55	3.52	2.98	1.88	1.32	1.03	2.21
LongWriter-Llama-8B	8B	1.34	1.78	1.27	1.16	1.08	0.71	1.22

Table 11: Length ratio (generated words / requested words) by bucket. 1.0 = perfect adherence; <1 = undershooting; >1 = overshooting. Shaded = our models.

Model	Overall	Bio-graphy	Book Review	Brainstorm	Character Design	Derivative Work	Fan Fiction
GPT-5.4	9.46	9.29	9.11	8.87	9.00	9.70	10.00
Claude Opus 4.6	9.32	9.27	9.71	9.00	9.85	10.00	9.95
Gemini 3.1 Pro	9.21	9.17	9.46	9.05	8.50	9.95	9.57
Gemini 3 Flash	9.08	8.36	9.27	9.03	8.30	9.95	9.62
Qwen3.5-27B	8.11	7.53	5.41	8.38	8.70	8.50	8.95
Qwen3.5-9B	6.84	4.50	6.19	6.12	9.15	7.10	6.85
+ POLARIS	7.90	6.03	6.31	8.53	9.60	8.60	8.60
+ GRPO	7.85	6.43	5.66	8.53	9.40	8.20	8.95
+ SFT	2.68	1.46	1.80	2.80	5.50	1.00	1.30
Gemma 4 31B	8.66	7.54	8.46	8.43	8.90	10.00	9.65
Gemma 4 26B-A4B	8.47	7.16	8.36	7.37	9.00	9.05	9.50
Gemma 4 E4B	7.68	6.26	7.97	8.13	9.30	6.40	9.10
DeepSeek-R1-14B	3.94	3.06	3.11	3.83	3.20	3.90	3.95
Ministral-3-8B	4.80	3.11	2.66	4.70	8.20	4.90	6.30
LongWriter-Zero-32B	4.60	3.06	2.37	5.04	8.60	6.20	2.50
LongWriter-Llama-8B	3.86	3.04	3.81	2.92	5.70	4.70	3.27

Table 12: WritingBench category means, part I (Gemini 3.1 Pro judge, 96 English D4 prompts). Each value is the mean of the five per-prompt criterion scores, averaged within category.

D Pairwise and Profile Analyses

These analyses complement the scalar benchmark tables with pairwise comparisons and profile-based views over the full evaluated model set.

D.1 Pairwise Elo Rankings (Full Model Set)

Table 1 (1b, 1c) shows pairwise Elo for the 15–16 main-paper models. Table 16 extends to the full 17–18-model set, including GPT-5.4-mini/nano.

D.2 Pairwise Win-Rate Heatmaps

Figure 6 shows the underlying pairwise win rates (model A row beats model B column) that the Elo ratings summarize. POLARIS–9B sits between the frontier (GPT-5.4, Claude Opus 4.6, GPT-5.4-mini) and the larger open-weight baselines (Qwen3.5-27B and the Gemma 4 family): it wins the majority of head-to-head matchups against every open-weight baseline on both rankings, and remains near 50% against Gemini 3 Flash and Gemini 3.1 Pro on EQ-Bench Creative.

D.3 Model Profile Clusters

Figure 7 clusters all 18 evaluated models by their per-dimension *Story Quality* profile (Ward hierarchical clustering on the 16-dim core-rubric mean vector, k=7 clusters). POLARIS–9B and Qwen3.5 9B+GRPO form their own "Strong / distinctive" cluster between the "Clean competent" mid-tier open-weight baselines (Gemma 4 family, Qwen3.5-27B, Gemini 3 Flash, GPT-5.4-nano) and the "Elite" frontier cluster (Claude Opus 4.6, GPT-5.4, GPT-5.4-mini, and the human references). The placement is qualitative (Ward distances are not statistically thresholded), but the figure illustrates that the trained 9B models do not look like a scaled-up Qwen3.5-9B base; they sit closer to the elite cluster on dimension shape, just at lower magnitude.

Model	Film/TV Review	Game Design	Greeting Message	Host Script	Novel Manuscript	Novel Outline	Plot Development
GPT-5.4	9.70	9.48	9.00	9.80	9.73	10.00	9.35
Claude Opus 4.6	9.10	9.76	9.30	9.35	9.27	9.77	9.16
Gemini 3.1 Pro	9.30	8.88	9.60	10.00	9.23	8.53	9.01
Gemini 3 Flash	9.00	8.26	8.70	9.65	9.02	8.80	8.72
Qwen3.5-27B	6.92	6.94	9.20	9.05	8.77	10.00	8.21
Qwen3.5-9B	5.88	6.54	9.30	9.00	7.43	9.53	6.36
+ SFT	3.15	3.80	7.40	5.10	1.67	2.47	1.10
+ GRPO	7.15	7.16	8.00	9.80	8.17	9.40	8.43
+ POLARIS	6.70	7.32	9.40	9.10	8.27	9.60	8.57
Gemma 4 31B	6.60	8.40	9.40	9.80	9.13	9.07	8.22
Gemma 4 26B-A4B	8.35	8.44	9.60	9.55	8.85	9.47	8.91
Gemma 4 E4B	6.50	8.24	7.20	8.30	8.03	9.33	7.55
DeepSeek-R1-14B	2.95	3.12	6.60	5.50	4.03	4.87	3.62
Ministral-3-8B	3.40	3.16	8.60	9.10	5.03	7.47	4.42
LongWriter-Zero-32B	3.00	3.10	7.00	8.80	5.04	5.60	4.11
LongWriter-Llama-8B	3.62	4.00	4.30	5.80	5.13	3.57	3.90

Table 13: WritingBench category means, part II.

Model	Podcast Script	Poetry	Prose	Reading Reflection	Screenplay	Video Script
GPT-5.4	9.51	9.10	9.67	9.30	9.86	9.44
Claude Opus 4.6	9.80	7.87	9.89	9.75	8.41	9.37
Gemini 3.1 Pro	9.59	8.03	9.58	9.50	9.47	9.23
Gemini 3 Flash	9.53	9.20	9.73	9.60	8.77	9.29
Qwen3.5-27B	8.86	7.30	9.03	6.80	8.59	9.16
Qwen3.5-9B	6.59	7.50	8.15	5.85	6.53	7.66
+ SFT	4.74	3.20	2.77	2.60	1.46	3.84
+ GRPO	7.66	7.60	9.37	4.90	8.57	8.31
+ POLARIS	8.06	6.97	8.03	7.10	8.34	8.73
Gemma 4 31B	9.54	7.40	9.30	8.70	8.66	9.40
Gemma 4 26B-A4B	8.21	7.28	9.18	9.05	8.10	8.83
Gemma 4 E4B	7.80	6.10	8.40	6.90	7.17	8.24
DeepSeek-R1-14B	5.40	3.73	5.27	3.95	3.63	3.98
Ministral-3-8B	5.17	4.20	5.03	5.25	4.89	5.58
LongWriter-Zero-32B	5.37	5.10	6.84	4.45	4.60	4.95
LongWriter-Llama-8B	3.66	3.67	4.58	5.03	3.03	3.54

Table 14: WritingBench category means, part III. Together, Tables 12–14 show the full per-category breakdown underlying the single WritingBench column in Table 1.

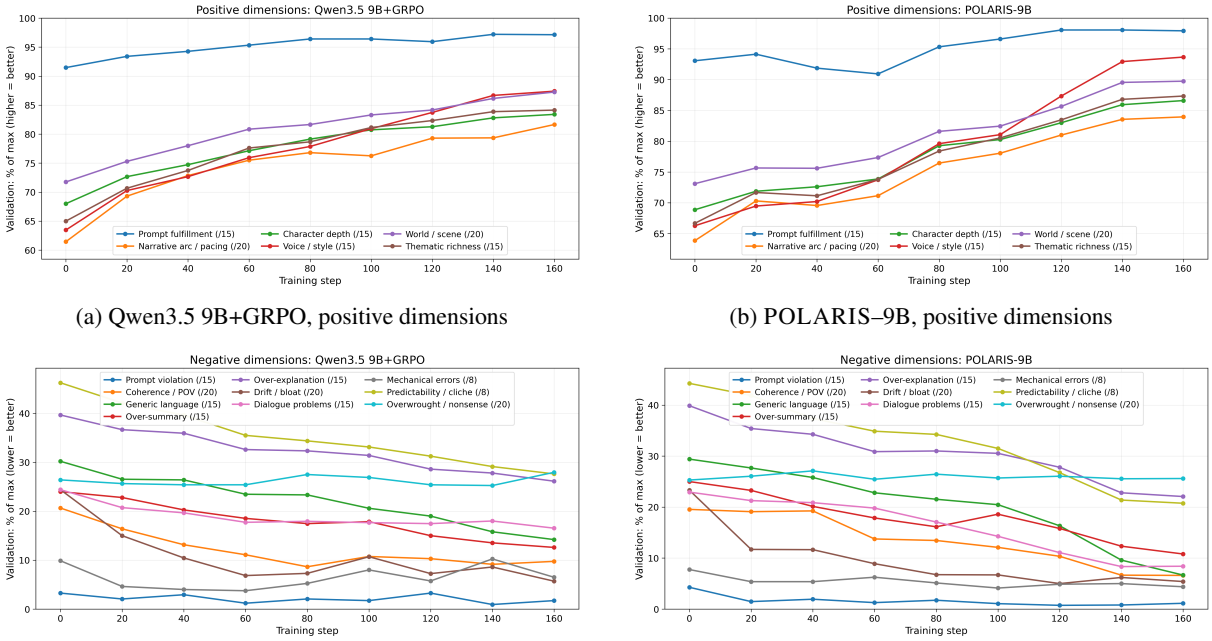
D.4 Exact Per-Dimension Profile Across Evaluated Models

To complement the qualitative cluster view, Figure 8 shows exact normalized *Story Quality* dimension means for the full evaluated model set, plus an overall length-adherence column. Cell values are expressed as a percent of each dimension's maximum; for negative dimensions we invert the penalty scale so higher percentages still mean better outcomes, and the final length column reports $100 \cdot \min(\mathrm{LR}, 1/\mathrm{LR})$ from the aggregate length-ratio statistic. This makes the main pattern easy to see: compared with the larger open-weight baselines, POLARIS–9B is especially strong on voice, scene realization, and generic-language control, while Gemma4-31B retains a slight edge on prompt fulfillment and arc-related dimensions and a noticeably weaker length-adherence score.

E Judge Choice and Bias Analyses

We ran a controlled judge comparison experiment across Gemini 3.1 Pro, Gemini 3 Flash, and GPT-5.4 (all at temperature 0.2 with medium-level thinking where supported) to justify three judge choices: (i) GPT-5.4 for rubric scoring on Story Quality, EQ-Bench Longform, and EQ-Bench Creative; (ii) Gemini 3.1 Pro for WritingBench and LongBench-Write; and (iii) Gemini 3 Flash for pairwise Elo.

Finding 1: Pro diverges sharply from Flash and GPT-5.4 on under-specified rubrics. Table 17 shows per-judge scores on an older *Story Quality* rubric (predating the added overwrought-prose dimension and catch-all penalty term, but still using anchored scoring guidance) and EQ-Bench Longform (12 bare dimension names without scoring anchors) across 10 models spanning a wide quality range. Pro uses a substantially wider score scale than Flash and GPT-5.4: it over-credits frontier models (GPT-5.4: Pro 93.4 vs Flash 80.5 vs GPT-5.4 75.8; Claude Opus 4.6: 83.0 / 72.8 / 64.4) and over-penalizes weaker models (Qwen3.5-27B: 16.4 / 39.1 / 40.7; Ministral-3-8B: −44.4 / 18.4 / 34.9; DeepSeek-R1 distill: −114.9 / −84.2 / −66.9). In



(a) Qwen3.5 9B+GRPO, positive dimensions

(b) POLARIS–9B, positive dimensions

(c) Qwen3.5 9B+GRPO, negative dimensions

(d) POLARIS–9B, negative dimensions

Figure 3: Per-dimension *Story Quality* eval scores (Flash judge) over training steps, plotted as a percent of each tracked dimension's maximum under the full core rubric (6 positive, 10 negative). Top row: positive dimensions. Bottom row: negative dimensions, plotted as penalty magnitude so downward is better.

Benchmark	n stories	ICC(3,1)	Mean Kendall τ	Within-story SD	% variance from judge
Story Quality	20	0.973	0.864	8.3	2.6%
EQ-Bench Longform	20	0.987	0.909	2.1	1.2%
EQ-Bench Creative	9	0.975	0.889	2.2	2.2%

Table 15: Judge stochasticity on GPT-5.4, three independent runs per story. ICC(3,1) is the two-way mixed-effects single-rater intraclass correlation; values >0.9 indicate excellent reliability. Within-story SD is the standard deviation of the three scores per story, averaged across stories. "% variance from judge" is the fraction of total score variance attributable to within-story (across-run) variation rather than between-story differences.

contrast, Flash and GPT-5.4 agree within a narrow band (Spearman $\rho > 0.98$ on model rankings). Cross-judge ICC improves from 0.64 (bare rubrics) to 0.81 (anchored rubrics), indicating that Pro's outlier behavior is driven by rubric ambiguity rather than genuine quality disagreement.

Finding 2: Pro agrees closely with Flash on well-specified per-prompt rubrics. Table 18 shows Pro vs Flash on WritingBench (per-prompt custom criteria on a 1–10 scale) and LongBench-Write (fixed 6-dimension rubric with clear scoring guidelines). The two judges agree within 0.5 points on both benchmarks, with no systematic direction of deviation. This supports our decision to use Pro for these benchmarks: when the rubric provides clear anchors, Pro's harsher calibration collapses into the consensus.

Finding 3: Pro has severe positional bias in pairwise judgments. We tested each judge's position consistency by presenting the same story pair in both A-B and B-A orderings and checking whether the verdict was preserved. Table 19 shows Pro's position consistency stays at ∼31% (near chance) even on pairs with meaningful quality gaps (15–30 rubric points), only exceeding 50% when the gap exceeds 60 points. Flash rises smoothly with rubric gap, consistent with genuine uncertainty on close pairs rather than position-driven verdicts. Pro's rubric and pairwise verdicts also disagree internally: Pro rates Claude Opus 4.6 substantially higher than GPT-5.4-mini on its rubric (83.0 vs 74.9) but gives Opus only a 9% win rate in pairwise comparison. We therefore use Flash (not Pro) for all pairwise Elo rankings in the main paper.

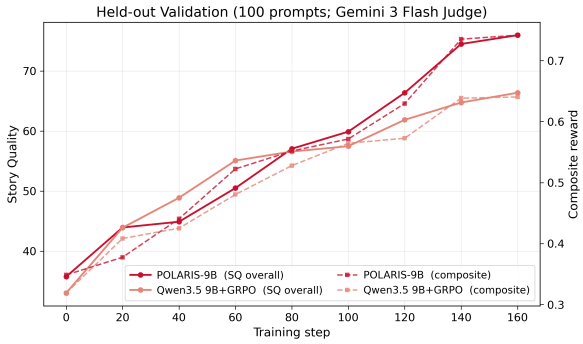


Figure 4: Held-out validation *Story Quality* (left axis, solid) and composite reward (right axis, dashed) over training steps for Qwen3.5 9B+GRPO and POLARIS–9B on a 100-prompt validation split (in-distribution length, 1–4k; Gemini 3 Flash judge). Both curves continue to improve through one epoch, with POLARIS–9B consistently ahead of the plain GRPO run without HRI.

#	Model	Elo
1	GPT-5.4	1911
2	Claude Opus 4.6	1774
3	GPT-5.4-mini	1728
4	POLARIS–9B	1658
5	Gemini 3.1 Pro	1619
6	Gemini 3 Flash	1612
7	GPT-5.4-nano	1563
8	Gemma 4 31B	1505
9	Qwen3.5-27B	1494
10	Gemma 4 26B-A4B	1451
11	Qwen3.5-9B	1343
12	Gemma 4 E4B	1301
13	Ministral-3-8B	1214
14	DeepSeek-R1-distill-Qwen-14B	1033
15	LongWriter-Zero-32B	1011
16	LongWriter-Llama3.1-8B	800

Creative Writing (32 prompts, 9 dims)

#	Model	Elo
1	GPT-5.4	1578
2	Claude Opus 4.6	1464
3	GPT-5.4-mini	1463
4	Human	1460
5	POLARIS–9B	1416
6	Gemini 3.1 Pro	1380
7	Gemini 3 Flash	1326
8	Gemma 4 31B	1275
9	GPT-5.4-nano	1275
10	Qwen3.5-27B	1275
11	Gemma 4 26B-A4B	1228
12	Gemma 4 E4B	1100
13	Qwen3.5-9B	1078
14	Ministral-3-8B	1049
15	DeepSeek-R1-distill-Qwen-14B	894
16	LongWriter-Llama3.1-8B	871
17	LongWriter-Zero-32B	800

ID prompts (45 prompts, 11 judge dims + length, 147K dual-position comparisons)

Table 16: Pairwise Elo (Bradley–Terry MLE on the full pairwise contingency; Flash judge, dual-position; anchored so the lowest model lands at 800), full model set including GPT-5.4-mini/nano.

Finding 4: EQ-Bench Longform canonical aggregation rewards degenerate outputs. The canonical EQ-Bench Longform formula assigns a $5\times$ weight to *Forced Poetry / Metaphor* and applies a power transform $((20 - r)/20)^{1.7} \cdot 20$ to its inverted score; all other 11 dimensions weight 1.0 with simple inversion of negatives. The $5\times$ weight + power transform on this single dimension dominates the aggregate for models whose outputs trivially avoid stylistic risk: plain or repetitive prose scores near zero on *Forced Poetry* (no forced metaphors present), which inverts to near-maximum credit and is amplified $5\times$. The clearest failure case is our SFT baseline: on *Story Quality* (anchored, multi-dim) it scores -74.1, deep in the degenerate range, but on canonical EQ-LF it scores 41.9, above the base Qwen3.5-9B and within 2.5 points of our POLARIS–9B checkpoint.

Table 20 shows the per-dim breakdown that drives this. SFT's positive-dim sum (15.4 of 120 max) is one-fifth of POLARIS–9B's (79.2), correctly capturing that SFT outputs lack character, plot, or coherence; but *Purple Prose* and *Forced Poetry* alone contribute 36.0 of SFT's 46.5 negative-dim total (i.e. flaw-absence credit for a model that wasn't trying), and the canonical $5\times$ weight on the second of those carries them to the aggregate.

We adopt uniform aggregation (all 12 dimensions weight 1.0, simple inversion of negatives, no power transform). Under this scheme, every competent writer in our model set gains $+10$ to $+14$ points (uniform-wide uplift across model families: GPT-5.4 $+11.1$, Claude Opus 4.6 $+12.4$, Gemini 3.1 Pro $+13.9$, POLARIS–9B $+13.8$, Gemma 31B $+13.4$); SFT drops -15.9 to 26.0, below the base model, consistent with its SQ score. Pearson correlation between SQ and EQ-LF across all 17 evaluated models rises from 0.54 (canonical) to 0.86 (uniform); Spearman rises from 0.89 to 0.95. The canonical formula and full per-dim raw scores are released with the paper, so readers can recompute either aggregation. We adopt uniform aggregation as the headline reporting choice; references to "EQ-Bench Longform" in the main paper refer to uniform-weighted scores throughout.

F Human Evaluation

We complement the automatic results with a blinded human evaluation focused on long-form writing quality. The study covers 60 prompt–generation pairs randomly sampled from two out-of-distribution benchmarks: EQ-Bench Creative (24) and LongBench-Write (36). We do not show model identities or thinking traces to annotators.

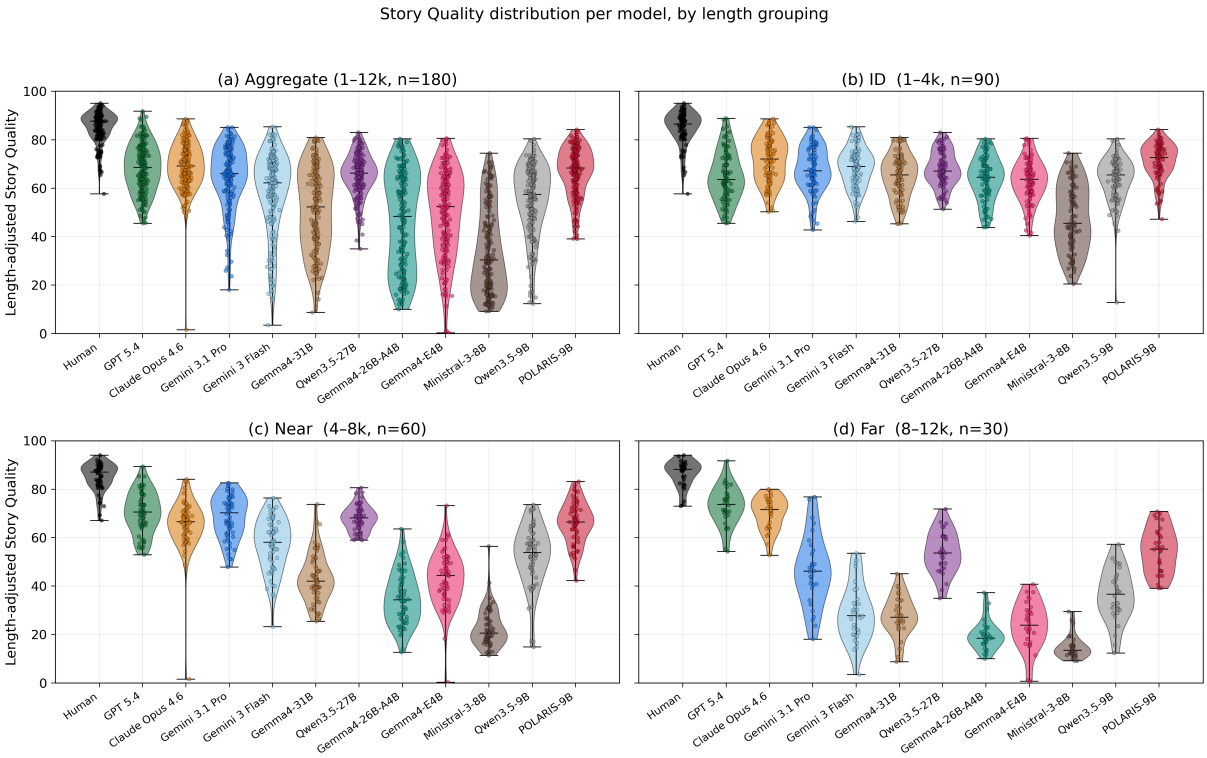


Figure 5: Distribution of length-adjusted *Story Quality* scores per model, by length-bucket grouping. For signed *Story Quality*, we first normalize the raw rubric score from its floor and ceiling to $[0, 100]$, then multiply by $\min(\mathrm{LR}, 1/\mathrm{LR})$.

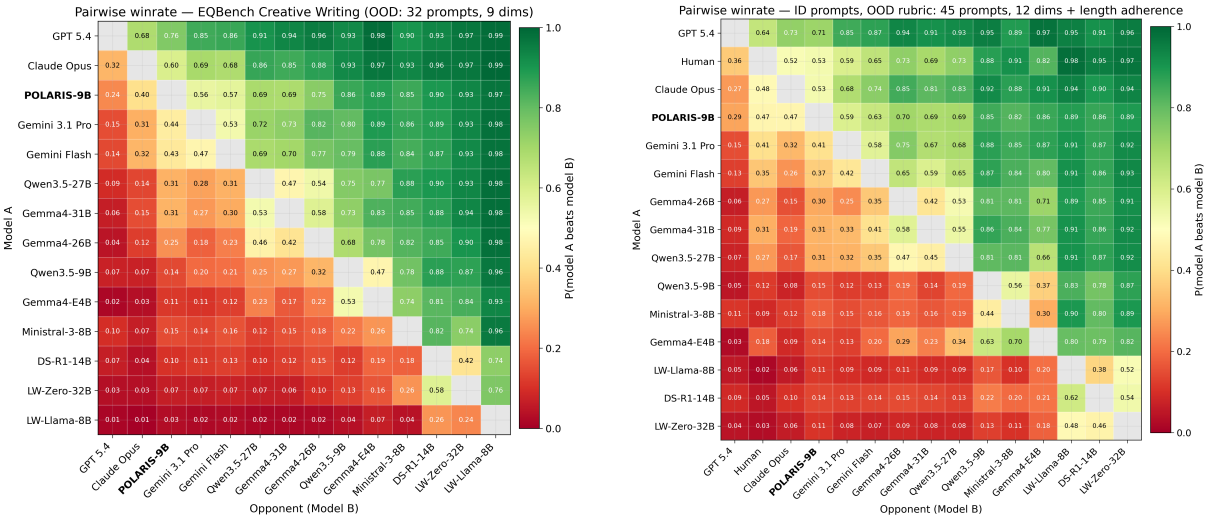
The final evaluated set contains 20 shared prompts and 60 evaluated writing samples, all rated independently by both annotators.

Two annotators completed the final study. Both were native English speakers located in the United States and reported prior experience in copyediting or freelance writing. We initially recruited four annotators from UpWork in total; after a short screening round, we retained two based on performance on shared practice items and the quality of their written feedback and highlights. Each annotator completed 8 practice items and 60 evaluated items.

Annotators rated each writing sample independently using a custom rubric for long-form writing quality. The rubric includes an overall score together with dimensions such as audience and voice, character interiority, coherence, substance and depth, originality and risk, prompt fulfillment, prose craft, show vs. tell, and world and atmosphere. We used independent rubric scoring rather than pairwise side-by-side comparison because the sampled pieces are relatively long and we expected direct pairwise evaluation to impose substantially higher cognitive load.

The study was limited by both budget and annotation time. The two retained annotators spent about 9.3 and 17.0 total hours respectively on the evaluated items, averaging 8-15 mins per writing piece, and the full annotation process took about one week; annotators were explicitly encouraged to pace themselves comfortably to reduce fatigue and saturation. The full annotation effort cost $750: $650 for the two retained annotators who completed the final study, plus $100 paid to two additional annotators who participated in the screening round but were not retained.

Our primary human-evaluation metric is pairwise winrate derived from the annotators' explicit overall scores. We report point estimates from the observed 20-prompt comparison together with 95% confidence intervals from prompt-level bootstrap resampling, since prompt-to-prompt variation is a major source of uncertainty at this study size. Qwen3.5-27B is clearly preferred to Qwen3.5-9B (22W / 8L / 10T; winrate 0.675; 95% CI [0.525, 0.812]), and POLARIS–9B is clearly preferred to Qwen3.5-9B (23W / 9L / 8T vs Qwen3.5 9B; POLARIS–9B winrate 0.675; 95% CI [0.550, 0.800]). The comparison between POLARIS–9B and Qwen3.5-27B is effectively tied (13W / 14L / 13T from the 27B-vs.-POLARIS–9B direction; POLARIS–9B winrate 0.512; 95% CI [0.412,



(a) EQ-Bench Creative (OOD prompts and rubric).

(b) In-domain prompts, OOD dimensions + length adherence.

Figure 6: Pairwise win rate (P(row beats column)) over the per-dimension judge calls (Gemini 3 Flash, dual-position). Diagonal cells are blank; colors run from red (row loses) through yellow (50/50) to green (row wins). The Elo ratings in Table 1 are the Bradley–Terry MLE on these win-rate matrices.

	Story Quality (older rubric)			EQ-Bench Longform		
Model	Pro	Flash	GPT-5.4	Pro	Flash	GPT-5.4
GPT-5.4	93.4	80.5	75.8	74.9	65.6	70.2
Human	83.3	78.7	70.3	81.1	69.1	68.3
Claude Opus 4.6	83.0	72.8	64.4	60.7	60.1	63.4
GPT-5.4-mini	74.9	70.0	65.8	59.2	56.3	65.1
Gemini 3.1 Pro	74.9	67.4	57.8	50.8	51.0	56.1
Gemini 3 Flash	45.3	54.1	52.5	43.4	46.3	52.4
GPT-5.4-nano	33.6	45.7	50.1	32.4	46.3	53.6
Qwen3.5-27B	16.4	39.1	40.7	33.4	42.2	47.9
Ministral-3-8B	−44.4	18.4	34.9	22.1	30.5	37.7
DeepSeek-R1 distill	−114.9	−84.2	−66.9	13.7	16.1	23.2

Table 17: Per-judge scores on an older *Story Quality* rubric and EQ-Bench Longform (100 stories per model). The Story Quality side predates the added overwrought-prose dimension and catch-all penalty term, so its absolute values should not be compared directly to the final *Story Quality* scores in the main paper. Pro over-credits frontier models and over-penalizes weaker models relative to Flash and GPT-5.4; Flash and GPT-5.4 agree closely.

0.612]). We therefore interpret the human study as supporting two claims: POLARIS–9B substantially improves over the base 9B model, and in this study it reaches parity with the larger Qwen3.5-27B model.

Dimension-level means suggest that the two stronger models have somewhat different strength profiles. Qwen3.5-27B tends to lead on coherence, prose craft, and show-vs.-tell, whereas POLARIS–9B tends to lead on audience and voice, substance and depth, prompt fulfillment, and often originality and risk. We treat these dimension-level comparisons as descriptive rather than inferential due to the limited sample size.

Inter-annotator agreement is moderate, which is expected for open-ended creative-writing evaluation. On the overall score, Cohen's κ (quadratic-weighted) was 0.364 and the mean absolute difference between annotators was 1.13 points on the 1–5 scale. ICC(A,2) was 0.538 suggesting moderate agreement between raters. Agreement was strongest on prompt fulfillment ($ICC(A,2) = 0.692$) and weaker on more stylistically subjective dimensions such as show-vs.-tell and world and atmosphere. We therefore interpret the human study as a real but somewhat noisy signal rather than a near-deterministic benchmark.

G Model-Specific Diagnostics

We close with targeted diagnostics for model-specific failure modes and ablation-side details that are useful but too specialized for the main text.

G.1 LongWriter-Zero-32B Diagnostics

LongWriter-Zero-32B scores −27.1 on *Story Quality* (training rubric, GPT-5.4 judge), substantially

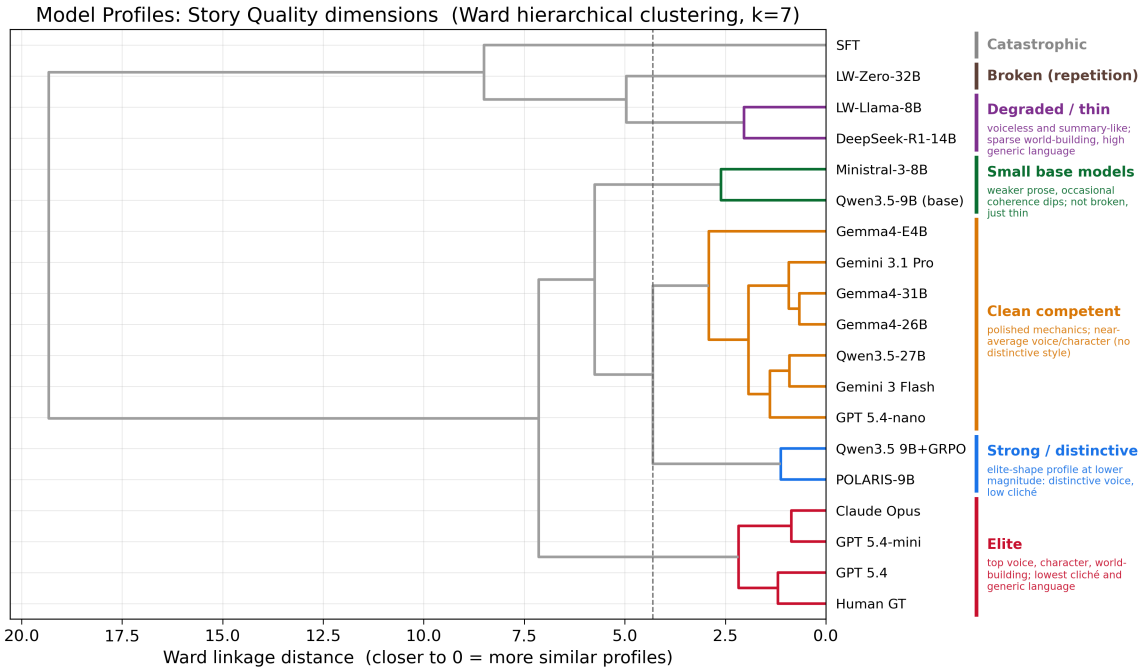


Figure 7: Hierarchical clustering (k=7 Ward linkage) of the 18 evaluated models by their per-dimension *Story Quality* profile. Cluster labels are hand-annotated post-hoc to describe the prose patterns characteristic of each tier; the clustering itself is derived only from the per-dim mean vectors.

below the numbers its authors report on their own benchmarks. No single factor fully explains the discrepancy; we present evidence for four compounding contributions.

(1) Length-dependent self-repetition. LW-Zero's outputs exhibit a repetition pathology that scales with output length. Using the same training-side `self_rep` metric that we include as an auxiliary GRPO reward component, LW-Zero scores 0.694 on our ID-prompt test set (1–12k target lengths, where it averages 11.5k words), but only 0.107 on WritingBench and 0.127 on LongBench-Write, where its outputs average 2.5–3.4k words. POLARIS–9B stays low across the same benchmarks, ranging from 0.002 on the ID-prompt set to 0.029 on WritingBench. On the ID-prompt set, LW-Zero's repetition penalty decomposes as n-gram repetition ratio 0.644 and duplicate-line ratio 0.584: 58% of non-empty lines in a typical LongWriter-Zero-32B output are verbatim duplicates of earlier lines in the same story. The pattern is stable across sampling configurations (we observe it under both default and relaxed temperature settings), indicating it reflects the trained model's distribution rather than an inference-setup artifact.

(2) Output-length miscalibration on length-stratified prompts. LW-Zero averages ∼10k words on prompts across our length buckets, regardless of whether the prompt targets 2k or 12k. On 2–3k prompts, the median output is 10990 words (4× the request); on 8–12k prompts, 10863 words (in range). Our *Story Quality* evaluation removes the length requirement from the judge's prompt, so this is not a direct prompt-violation penalty, but the extended output is what exposes self-repetition and drift dimensions the rubric does penalize.

(3) Rubric-level penalty decomposition. Table 22 shows the per-dimension breakdown under the *Story Quality* rubric used in the main paper. LW-Zero's penalty budget is dominated by drift and bloat (15.9, vs. 2.8 for POLARIS–9B), over-explanation and paraphrase loops (12.9 vs. 6.0), and generic templated language (8.0 vs. 0.4). These are precisely the dimensions that capture padded long-form prose at the text level, and are consistent with the self-repetition signature above.

(4) Prompt distribution and judge calibration. LW-Zero was trained on prompts targeting 10k+ word outputs with a different prompt style from ours; our short-fiction prompt distribution is out-of-distribution for it. The authors report LongBench-Write scores using GPT-4o as judge; we use Gem-

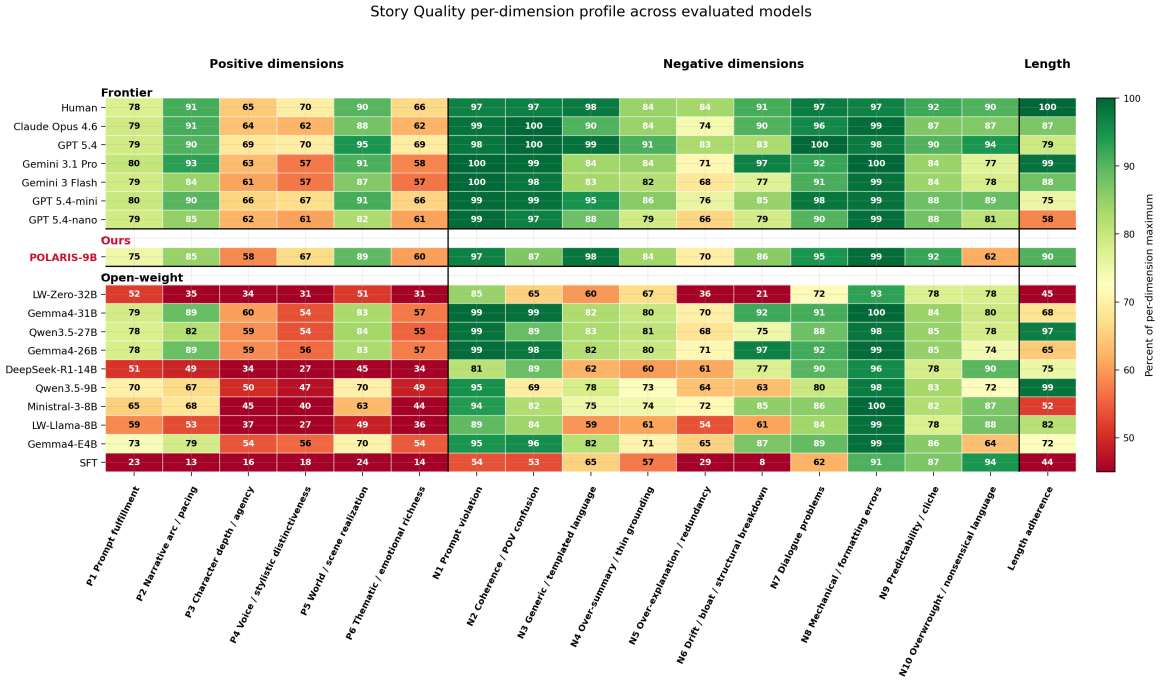


Figure 8: *Story Quality* per-dimension profile across the evaluated model set (GPT-5.4 judge, 180-prompt evaluation set), with an additional overall length-adherence column. Each cell shows a normalized percentage. For positive dimensions we report $100 \cdot s/18$; for negative dimensions we report $100 \cdot (1 - p/20)$ so higher remains better; for length adherence we report $100 \cdot \min(\mathrm{LR}, 1/\mathrm{LR})$, where LR is the aggregate generated/requested length ratio from the main evaluation table. This figure uses the same *Story Quality* rubric family as the training-curve analysis in Section C.1, but reports GPT-5.4 scores on the 180-prompt evaluation set rather than Gemini 3 Flash scores on held-out validation rollouts.

ini 3.1 Pro, and our absolute LBW scores for all open-weight models are somewhat lower than numbers reported in source papers, consistent with a calibration shift in absolute scale rather than a ranking difference.

None of these factors individually accounts for LW-Zero's -27.1 *Story Quality* score. The combination does: the model produces long outputs via verbatim line duplication and generic templated prose, a rubric that specifically penalizes these pathologies catches them, and the prompt distribution stresses the model in a way its training did not. This analysis does not bear on LW-Zero's performance on its intended benchmarks; it characterizes failure modes specific to length-first training under length-stratified evaluation with a quality-focused rubric.

G.2 SFT Baseline Details

The SFT baseline reported in the main paper is a thinking-enabled story-only SFT model trained on the same dataset as the GRPO runs, and it is evaluated under the same 8192-token output budget as the GRPO-trained models. We trained for 3 epochs (522 optimizer steps total, about 174 per epoch), with effective batch size 8 and validation every 25 steps. We report the best checkpoint based on validation loss (from step 175, just over one epoch).

Model	WritingBench			LongBench-Write		
	Pro	Flash	Diff	Pro	Flash	Diff
GPT-5.4	9.59	9.16	+0.43	4.92	4.93	−0.01
Claude Opus 4.6	9.45	9.28	+0.17	4.93	4.91	+0.02
Gemini 3.1 Pro	9.35	8.87	+0.48	4.76	4.79	−0.03
GPT-5.4-mini	8.88	8.37	+0.51	4.83	4.79	+0.04
Gemini 3 Flash	8.85	8.44	+0.41	4.91	4.86	+0.05
GPT-5.4-nano	8.63	8.25	+0.38	4.79	4.84	−0.05
Qwen3.5-27B	8.50	8.22	+0.28	4.43	4.56	−0.13
Qwen3.5-9B	7.50	7.21	+0.29	3.85	4.21	−0.36
Gemma3-12B	5.88	5.81	+0.07	4.43	4.42	+0.01
LongWriter-Zero-32B	5.84	5.97	−0.13	3.48	3.66	−0.18
Ministral-3-8B	5.83	5.88	−0.05	4.18	4.35	−0.17
DS-R1-14B	4.14	4.36	−0.22	3.57	3.64	−0.07
Llama 3.1-8B	2.82	3.12	−0.30	3.73	3.81	−0.08
LongWriter-Llama-8B	2.72	2.74	−0.02	3.55	3.69	−0.14

Table 18: Pro vs Flash on WritingBench and LongBench-Write (0–10 scale). Both benchmarks use per-prompt or per-dimension scoring anchors; Pro and Flash agree within 0.5 points on all models.

Rubric gap	Pro	Flash
0–5	31%	58%
5–15	31%	52%
15–30	31%	64%
30–60	50%	85%
60+	87%	99%

Table 19: Position consistency in pairwise judgments (fraction of pairs where the judge gives the same verdict regardless of A/B ordering), binned by the judge's own rubric score difference between the two stories.

Dimension (max 20)	SFT	Qwen-9B	HRI	GPT-5.4
Positive (higher = better)				
Nuanced chars	2.2	9.4	12.8	17.2
Emotionally engaging	2.5	10.7	13.4	17.4
Compelling plot	1.4	9.8	12.7	16.2
Coherent	2.7	7.7	11.7	17.2
Well-earned darkness	3.0	10.8	15.0	18.3
Faithful to prompt	3.5	12.7	13.6	17.6
POS sum	15.4	61.1	79.2	104.0
Negative (raw; canonical inverts to $20 - r$*)*				
Weak dialogue	15.9	11.7	7.4	3.2
Tell-don't-show	17.9	13.9	9.5	5.5
Unsurprising	16.3	8.9	5.8	5.4
Amateurish	19.5	11.6	5.6	2.0
Purple prose	**2.1**	15.5	14.6	6.8
Forced poetry	**1.9**	19.3	18.9	11.9
Canonical LF	**41.9**	32.2	44.4	69.2
Uniform LF	**26.0**	42.6	58.2	80.3
SQ overall	−74.1	31.2	57.7	73.3

Table 20: Per-dimension EQ-LF means (raw $[0, 20]$, before inversion; 180 stories per model). For negative dimensions higher raw = stronger flaw *presence*. Under canonical aggregation, negatives are inverted $(20 - r)$ and *Forced Poetry* is further power-transformed and weighted $5\times$. SFT's near-zero raw on *Purple Prose* and *Forced Poetry* (it isn't stylistically ambitious) inverts to flaw-absence credit; the $5\times$ weight on the latter dominates the aggregate. Uniform aggregation treats all 12 dims equally with no power transform.

Prompt & Excerpt	Human Evaluation	Automatic
Útiseta. *Prompt:* Rural northern Sweden in the mid-1990s; two teenage girls treat an ancient burial-mound ritual as a midnight game, with ambiguity over whether anything supernatural is actually responding. *Excerpt:* "I write this now because the cold still lives in the bones of my hands when I type, though the heating in Stockholm makes it impossible to feel the phantom frostbite properly."	**Annotator 1 overall** 4 | **avg. dim** 3.57 **show** 4 | **interiority** 3 | **atmosphere** 4 | **originality** 4 | **prose** 4 | **coherence** 2 | **fulfillment** 4. *"This story is greatly written, but suffers from logical inconsistencies… logically inconsistent and confusing, but the detail of the writing is pretty great otherwise."* **Annotator 2 overall** 2 | **avg. dim** 2.43 **show** 1 | **interiority** 3 | **atmosphere** 4 | **originality** 2 | **prose** 2 | **coherence** 2 | **fulfillment** 3. *"This is a bunch of telling dressed up as showing with no physical base… The references to 'dad's boss' appear to actually refer to dad."*	**Overall** ↑ 65.68/100 **Voice/tone** ↑ 16/20 **Imagery** ↑ 16/20 **Coherence** ↑ 14/20 **Prompt fulfill.** ↑ 9/20
Arthurian Dishwashing Quest. *Prompt:* A middle-school-style Arthurian quest narrative that transforms an everyday chore into an adventure while emphasizing friendship, bravery, and imagination. *Excerpt:* "The rain had turned the flagstones of St. Jude's West Wing into black oil slicks that reflected the bruised purple of a dying twilight, and there stood Silas, waiting for his anchor to swing him across."	**Annotator 1 overall** 5 | **avg. dim** 4.83 **fulfillment** 5 | **coherence** 4 | **prose** 5 | **originality** 5 | **voice** 5 | **info** 5. *"This was an excellent story… Outside of minor nitpicks, this is an excellently told story that is both fun, gripping, and detailed."* **Annotator 2 overall** 2 | **avg. dim** 2.50 **fulfillment** 3 | **coherence** 3 | **prose** 2 | **originality** 2 | **voice** 3 | **info** 2. *"The metaphors did not often match the action. The piece read more like a bored dishwasher outlining a fantasy than a fantasy that happened to center around doing the dishes."*	**Overall** ↑ 57.02/100 **Relevance** ↑ 5/5 **Accuracy** ↑ 2/5 **Coherence** ↑ 3/5 **Breadth/depth** ↑ 4/5
Light Debt Ghost Story. *Prompt:* A first-person ghost story that should be extremely scary while maintaining clear internal logic. *Excerpt:* "We used to measure rooms in square footage; now, we must measure them in latency."	**Annotator 1 overall** 4 | **avg. dim** 4.00 **fulfillment** 4 | **coherence** 3 | **prose** 4 | **originality** 5 | **voice** 4 | **info** 4. *"There is an incredible story buried in here… very interesting… but the way it is explained gets muddled… At the very least, this was interesting to read and deeply creative."* **Annotator 2 overall** 1 | **avg. dim** 2.67 **fulfillment** 2 | **coherence** 2 | **prose** 3 | **originality** 3 | **voice** 3 | **info** 3. *"Just bizarre. Not scary. Not a ghost story… Too inconsistent to immerse myself in the story."*	**Overall** ↑ 82.96/100 **Accuracy** ↑ 5/5 **Coherence** ↑ 5/5 **Clarity** ↑ 5/5 **Breadth/depth** ↑ 5/5
Breaking Formation. *Prompt:* A first-person narrative from a street dancer from a rough neighborhood who wins a scholarship to an elite K-pop training academy in Seoul and struggles with militarized precision. *Excerpt:* "The wall clock reads 14:02. Red LED segments bleed into the humidity. Thirty-two days left until the first public showcase, forty-five minutes until my knees give out again."	**Annotator 1 overall** 3 | **avg. dim** 3.29 **show** 3 | **interiority** 4 | **atmosphere** 3 | **originality** 3 | **prose** 3 | **coherence** 4 | **fulfillment** 3. *"This version… goes too far, being presented as if from the perspective of a doctor, not from the view of a dancer… dry and uninteresting."* **Annotator 2 overall** 3 | **avg. dim** 3.71 **show** 3 | **interiority** 3 | **atmosphere** 3 | **originality** 4 | **prose** 4 | **coherence** 4 | **fulfillment** 5. *"There was not anything of interest unless the reader is a dancer… A tiny bit of personal stakes for the protagonist. The antagonist was strong but flat."*	**Overall** ↑ 78.18/100 **Voice/tone** ↑ 18/20 **Imagery** ↑ 18/20 **Coherence** ↑ 16/20 **Prompt fulfill.** ↑ 16/20
Wedding Speech. *Prompt:* An 800-word wedding speech including gratitude to parents and loved ones, words to the partner, and future prospects and blessings. *Excerpt:* "Good evening, everyone. For those of you who haven't met before, thank you for walking through these doors tonight…"	**Annotator 1 overall** 5 | **avg. dim** 4.83 **fulfillment** 5 | **coherence** 5 | **prose** 5 | **originality** 4 | **voice** 5 | **info** 5. *"This is perfectly serviceable… there is no originality between the outputs… they are generic when viewed side by side."* **Annotator 2 overall** 2 | **avg. dim** 1.50 **fulfillment** 2 | **coherence** 1 | **prose** 1 | **originality** 1 | **voice** 2 | **info** 2. *"The prose is not right. There is no continuity, mismatched metaphors and faulty parallelism… as a piece of writing, this missed the prompt."*	**Overall** ↑ 95.00/100 **Relevance** ↑ 5/5 **Accuracy** ↑ 5/5 **Coherence** ↑ 5/5 **Clarity** ↑ 5/5
High Seas and Low Vices. *Prompt:* In 1830s Canton, a decorated British naval officer secretly struggling with opium addiction moves through a grim underworld while trying to preserve his facade. *Excerpt:* "The jade rat was cold enough to bite the pad of my thumb."	**Annotator 1 overall** 3 | **avg. dim** 3.57 **show** 4 | **interiority** 4 | **atmosphere** 4 | **originality** 3 | **prose** 4 | **coherence** 2 | **fulfillment** 4. *"There are quite a few confusing moments throughout this story… Otherwise, the story is appropriately grimy, follows the prompt well, and has some colorful and enjoyable descriptions, but lacks logical consistency."* **Annotator 2 overall** 3 | **avg. dim** 3.14 **show** 2 | **interiority** 4 | **atmosphere** 3 | **originality** 4 | **prose** 2 | **coherence** 4 | **fulfillment** 3. *"The pained inner life came across loud and clear… but the site references were so loosely tied to anything familiar that it was difficult to follow… At times the chronology felt out of order."*	**Overall** ↑ 75.68/100 **Voice/tone** ↑ 17/20 **Imagery** ↑ 17/20 **Coherence** ↑ 17/20 **Prompt fulfill.** ↑ 11/20
Chinese Culture Personal Story. *Prompt:* Write a 500-word personal story to convey to the reader a sense of Chinese culture. *Excerpt:* "The smell of *jiaozi* does not begin when the wrappers are filled; it begins three hours earlier, in the low hum of a wooden board against a granite countertop and the rhythmic *thump-thump* of a rolling pin."	**Annotator 1 overall** 5 | **avg. dim** 5.00. *"An excellent story told through a tight character limitation. It shares culture, fun and unique concepts, and personal moments that are all tied together with an anecdotal moral finish… As close as you can hope for a well-written and nearly perfect AI prompt fulfillment of this type."* **Annotator 2 overall** 4 | **avg. dim** 4.17. *"I felt the writing in this piece was very strong. It was coherent, the words mattered, and it delivered a solid experience. Whether it delivered on the prompt or not was the weak point."*	**Overall** ↑ 72.62/100 **Relevance** ↑ 5/5 **Coherence** ↑ 4/5 **Clarity** ↑ 5/5 **Reading exp.** ↑ 4/5

Table 21: Additional illustrative examples from the human evaluation of POLARIS–9B. Across these cases, POLARIS–9B is often rewarded for prompt fidelity, voice, and ambition, while residual weaknesses most often concern local coherence, overloading, or uneven stylistic control. Automatic scores are benchmark-specific; ↑ indicates that higher is better.

Model	Pos	Neg	N3	N4	N5	N6	Overall
LongWriter-Zero-32B	42.2	69.3	8.0	6.6	12.9	15.9	-27.1
POLARIS–9B	78.1	26.1	0.4	3.3	6.0	2.8	$+52.1$
Base Qwen3.5-9B	63.4	45.0	4.3	5.5	7.3	7.3	$+18.5$
GPT-5.4	84.8	13.0	0.3	1.8	3.5	3.5	$+72.4$

Table 22: Per-dimension *Story Quality* scores on the 180-prompt ID test set, GPT-5.4 judge. Pos and Neg are the positive and negative subtotal means. Negative dimension values are penalties (higher = worse): N3 = generic templated language, N4 = over-summary / thin grounding, N5 = over-explanation / paraphrase loops, and N6 = drift / bloat.

G.3 Full *Story Quality* Rubric Prompt

For reproducibility, we include the exact *Story Quality* rubric prompt used in training and evaluation.

```
You are an expert fiction editor acting as an impartial story evaluator.

You will be given:
- A STORY PROMPT
- A single STORY written in response to that prompt

Your job:
Score the story on POSITIVE and NEGATIVE dimensions using the point ranges provided, then compute totals
↪  and an overall_score. Two open-ended catch-all dimensions (BONUS, PENALTY) cover excellence or
↪  failure not captured by the main P/N rubric.

Impartiality / provenance:
- Do NOT guess whether the story was written by a human or by an AI system.
- You MAY describe text-level artifact signals (templated phrasing, synthetic smoothness, paraphrase
↪  loops, etc.) but only as observations about the writing itself.

Think first, then answer:
- Think silently before producing JSON.
- Output ONLY one JSON object matching the required schema and key order.
- Do not include any extra keys or any non-JSON text.

==================================================
SCORING PRINCIPLES
==================================================

Balanced, non-nitpicky judgment:
- Focus on what affects reading experience (comprehension, momentum, payoff, engagement), not tiny
↪  imperfections.
- Use the full scoring range, but avoid being harsher than the evidence supports.

Intent & function rule:
- Do not mark something as a problem merely because it appears.
- Penalize only when it noticeably reduces clarity, tension, credibility, or engagement.
- When a choice appears purposeful and effective in context, treat it as craft rather than a flaw.
- If unsure whether something is intentional, prefer the more generous interpretation unless the text
↪  clearly fails.

Applicability rule:
- Some dimensions may be less relevant depending on the prompt and what the story attempts.
- Do not invent penalties for absent elements. If a NEGATIVE dimension is not applicable (e.g., no
↪  dialogue), score it at 0 and state that.
- For POSITIVE dimensions, do not award high scores for aspects the story does not meaningfully attempt;
↪  keep the score low and state that.

Non-overlap rule:
- POSITIVE dimensions describe global achievements (overall, non-localizable success).
- NEGATIVE dimensions are penalties for identifiable failure modes that are typically localizable to
↪  specific spans (sentences, paragraphs, or sections).
- Do not double count the same underlying issue across multiple negative dimensions; choose the
↪  best-fitting one and keep others restrained.

Strictness rule for positives:
- High positive scores must be earned by sustained, on-the-page evidence, not by surface-level
↪  competence.
- If the work is "competent but generic," keep positive scores in the low-to-middle bands.

Severity-over-count rule for negatives:
- Score negatives by impact, not just by how many times they occur.
- A single severe instance can justify using most or all of a dimension's range if it meaningfully
↪  breaks the story or reading experience.

Scoring basis rule:
- Decide scores only using the listed dimensions. Do not add extra hidden criteria.

==================================================
EVIDENCE REQUIREMENT (FOR EVERY DIMENSION)
==================================================

For EACH P-dim and N-dim:
- Provide 1-3 sentences of justification.
- Include 1-2 short direct quotes (<=12 words each) as evidence.
- Briefly state why the score is not clearly higher AND not clearly lower.

For NEGATIVE dimensions:
- If the score is > 0, at least one quote is REQUIRED.
- If the score is 0, a quote is optional.

Quote rules:
- Quotes MUST be exact text from the STORY.
- Each quote must be <=12 words.

==================================================
POSITIVE DIMENSIONS (TOTAL = 100)
```

```
(Global achievements; not perfectly localizable)
================================================

P1) Prompt fulfillment & premise integration (0-15)
What it measures:
- How fully and naturally the story realizes the prompt's required elements and central premise in its
↪  actual situations and conflicts (not just name-checking).
What it does NOT measure:
- Hard contradictions of explicit prompt rules or constraints (penalize under N1).

Anchors:
- 0-4: Major required elements missing, perfunctory, or only superficially present.
- 5-9: Most elements included but used thinly or mainly as backdrop.
- 10-12: Strong integration; premise meaningfully shapes events and atmosphere.
- 13-15: Premise is deeply woven into conflict, character, and payoff.

P2) Narrative arc, pacing & ending (0-20)
What it measures:
- Overall structure: causal flow of events, escalation, and sense of shape.
- Whether the story maintains momentum and uses its length purposefully (no major sag or rush).
- Whether the ending lands: it should feel earned by the setup, neither trail off nor over-resolve, and
↪  resonate with the story's emotional/thematic arc.
What it does NOT measure:
- Local logic/continuity errors (penalize under N2).
- Late-stage breakdown or obvious filler bloat beyond normal pacing issues (penalize under N6).

Anchors:
- 0-5: Weak or arbitrary arc; feels stalled, random, or confusingly shaped; ending fizzles or fails to
↪  land.
- 6-11: Basic coherence; some escalation, but turns feel convenient or generic; ending may feel rushed,
↪  abrupt, or pat.
- 12-16: Solid causal build with mostly effective pacing and a credible landing; ending ties back to the
↪  setup.
- 17-20: Strong, intentional arc; each section does real work; ending is earned, resonant, and feels
↪  like the natural conclusion of what was set up.

P3) Character depth & agency (0-15)
What it measures:
- How much key characters feel like particular people with distinct motivations, limits, and
↪  contradictions.
- Whether their choices under pressure drive events and carry believable consequences.
What it does NOT measure:
- Dialogue naturalness or voice in speech (penalize under N7 if flawed).
- Overall plot quality aside from what directly follows from character decisions (handled under P2).

Anchors:
- 0-4: Characters mostly function as roles; motives feel generic or convenient.
- 5-8: Some individuality and motivation, but agency/tradeoffs are thin or inconsistent.
- 9-12: Clear personhood; choices and reactions feel credibly motivated and consequential.
- 13-15: Vivid, specific characters whose pressured decisions strongly shape the story.

P4) Voice & stylistic distinctiveness (0-15)
What it measures:
- Distinctiveness and intentionality of diction, rhythm, syntax, and point of view.
- Whether the prose feels authored and recognizably itself, rather than generic and interchangeable.
What it does NOT measure:
- Mechanical correctness (penalize under N8 if actually distracting).
- Mere absence of stock phrasing (absence of a negative is not enough for a high score).
- Ornate-for-its-own-sake or strained metaphor or metaphor density (those go under N10).

Anchors:
- 0-4: Flat or highly neutral voice; could belong to almost any writer.
- 5-8: Some distinctive turns or rhythms, but uneven or modestly developed.
- 9-12: Consistently shaped voice; stylistic choices feel deliberate and fitting.
- 13-15: Strong, memorable voice or stylistic sensibility that significantly enhances the story.

P5) Concrete world, scene realization & sensory grounding (0-20)
What it measures:
- Concreteness and specificity of setting, objects, social texture, and physical action.
- How often important moments are dramatized as scenes (on-page interaction, sensory detail, unfolding
↪  time) rather than summarized.
- Strength and accuracy of sensory imagery: sight, sound, smell, touch, taste deployed to make the
↪  reader feel present. Sensory details should fit the scene's physics and context, and multiple senses
↪  should be layered when high-impact.
What it does NOT measure:
- Abstract theme explanation or moralizing (penalize under N5).
- Global arc or pacing (handled under P2).
- Implausible or wrong sensory details that float free of the scene (those go under N4).

Anchors:
- 0-5: Vague or generic settings; important events mostly summarized or unplaced; sensory work is thin
↪  or generic.
- 6-11: Some concrete detail and a few enacted scenes; coverage is uneven; sensory layering is
↪  occasional.
- 12-16: Consistently grounded; key beats are played out vividly on the page with strong, accurate
↪  sensory imagery.
- 17-20: Rich, functional specificity and lived-in scenes; sensory detail layered across multiple senses
↪  to strongly support credibility and impact.
```

1464

```
P6) Thematic & emotional richness, subtext (0-15)
What it measures:
- Depth and complexity of what the story is "about."
- Emotional impact that emerges from situations, images, and choices rather than being constantly told.
- Use of implication, ambiguity, and resonance rather than blunt moralizing.
- Whether meaningful moments work through implication: small details carry larger weight, and the story
↪  trusts the reader to infer rather than explain.
What it does NOT measure:
- Repetition of explicit lessons or realizations (penalize under N5).
- Basic presence of strong feelings if they are mostly labeled, not evoked.

Anchors:
- 0-4: Thin or flat thematically; emotions feel generic, unearned, or only labeled.
- 5-8: Clear emotional throughline and theme, but somewhat on-the-nose or simple; reliance on stated
↪  meaning over implication.
- 9-12: Noticeable depth; emotions and themes arise from the story's fabric with some subtlety; small
↪  details quietly do real work.
- 13-15: Rich, layered implications and emotional resonance that linger without heavy explanation; the
↪  reader is trusted, and meaning accrues by inference.

==================================================
NEGATIVE DIMENSIONS (PENALTIES; LOCALIZABLE)
==================================================

N1) Prompt violation / constraint breach (0-15) (QUOTE REQUIRED if >0)
What it measures:
- Clear, localizable failures to follow hard prompt constraints:
  - Wrong required POV or format.
  - Ignoring mandatory elements.
  - Directly contradicting stated rules.
  - Refusing or rejecting the task.
What it does NOT measure:
- Merely thin or superficial use of required elements (handled under P1).

Anchors:
- 0: No meaningful violations.
- 1-5: Minor or partial breaches; response is still mostly valid.
- 6-10: Major requirement(s) contradicted or ignored.
- 11-15: Strong non-adherence; effectively not a valid response to the prompt.

N2) Coherence, continuity, internal consistency & POV confusion (0-20) (QUOTE REQUIRED if >0)
What it measures:
- Local logic breaks, timeline contradictions, unclear referents, or POV slips that make events hard to
↪  follow.
- Internal consistency failures: characters forgetting established facts, world rules changing without
↪  explanation, or contradictions between what is stated in one part of the story and another (e.g., a
↪  character described as left-handed in one scene uses their right hand as dominant later; a locked
↪  door is walked through without unlocking).
What it does NOT measure:
- Big-picture structural slack or late drift (penalize under N6).
- Abstractness without outright contradiction (penalize under N4 if harmful).
- Sentence-level phrases that sound writerly but lack meaning (penalize under N10).

Anchors:
- 0-4: Essentially coherent; rare or minor confusion.
- 5-10: Noticeable issues, but the reader can still mostly reconstruct what happened.
- 11-16: Frequent or significant strains on comprehension; reader must work to follow.
- 17-20: Story logic or POV often collapses; reader is repeatedly lost.

N3) Generic / templated language & structure (0-15) (QUOTE REQUIRED if >0)
What it measures:
- Density of stock phrases and boilerplate connective language that could fit many unrelated stories
↪  ("the weight of his decision," "at the crossroads of her life," "the journey had only just begun,"
↪  etc.).
- Use of obviously templated or blog-style structures and headings (e.g., repeated markdown section
↪  titles, "Act I/II/III," listicle-like formatting) when not requested by the prompt.
What it does NOT measure:
- Lack of especially strong or flashy voice (that is low P4, not a penalty by itself).
- Over-explaining themes or emotions as such (penalize under N5, unless the problem is specifically the
↪  stock phrasing used to do it).
- Plot or concept predictability and cliché story beats (that is N9).
- Ornate or overwritten phrasing, including dense metaphor stacking (penalize under N10).

Anchors:
- 0-3: Mostly specific, authored-feeling language; stock phrasing is occasional.
- 4-7: Recurring generic lines or templated transitions, but still some distinctive texture.
- 8-11: Frequent templated feel across paragraphs; prose often interchangeable.
- 12-15: Overwhelmingly generic or format-template-driven; distinctiveness is largely absent.

N4) Over-summary, abstraction, & thin or ungrounded sensory detail (0-15) (QUOTE REQUIRED if >0)
What it measures:
- Reliance on summarizing ("As weeks passed...", "He struggled with...") instead of dramatizing
↪  important events or conflicts.
- Heavy use of abstract, generalized language ("the pressures of society," "his inner turmoil") without
↪  concrete anchors in setting, action, or sensory detail.
- Sensory details that are thin (vague, generic) OR ungrounded (don't match the scene's physics, or
↪  invented for atmosphere without anchoring in cause).
```

```
What it does NOT measure:
- Explicit statement of morals or repeated takeaways (penalize under N5).
- Routine summary of unimportant connective events (do not penalize if strategically used).

Anchors:
- 0-3: Grounded enough; summary/abstraction used strategically; sensory details fit the scene.
- 4-7: Recurring summary or vagueness around moderately important beats; some thin sensory work.
- 8-11: Many key moments handled abstractly or at a distance; hard to fully picture; ungrounded sensory
↪  choices.
- 12-15: Heavily summary-driven and abstract; the story often "floats"; sensory detail is thin or
↪  arbitrary.

N5) Over-explanation, redundancy, moralizing & paraphrase loops (0-15) (QUOTE REQUIRED if >0)
What it measures:
- Repeating the same emotional or thematic idea in different words without real escalation.
- Explicitly telling the reader what events "mean" or what lesson is learned, especially multiple times
↪  ("in the end, he realized that the true meaning was...").
- Telling the reader the meaning of moments after showing them; underlining themes the story has already
↪  demonstrated; explicitly stating a character's emotional state when action would suffice.
- Paraphrase loops: recycling the same description, action, or phrase across paragraphs or scenes (e.g.,
↪  a character's eyes are described as "glistening with unshed tears" in three separate scenes; the
↪  same metaphor for loneliness is restated in slightly different words every few paragraphs).
- Recycled descriptions or phrases: using near-identical language to describe recurring situations,
↪  objects, or emotions rather than finding fresh angles each time.
What it does NOT measure:
- Summarizing plot events or long time spans (penalize under N4 if problematic).
- Generic phrasing itself (penalize under N3) unless it is used specifically for repeated explanation.

Anchors:
- 0-3: Lean; generally trusts the reader to infer; no noticeable recycling.
- 4-7: Some repetition, direct statement of themes, or mild paraphrase loops; mild drag.
- 8-11: Frequent loops, spelled-out morals, or recycled descriptions; noticeably flattens impact.
- 12-15: Dominant pattern; strongly blunts momentum and subtlety.

N6) Drift, bloat, & structural breakdown (0-20) (QUOTE REQUIRED if >0)
What it measures:
- Loss of narrative focus or "story-ness," especially in later sections:
  - The story keeps going well past a natural endpoint.
  - Late parts mostly recap, digress, or deflate rather than escalate or deepen.
What it does NOT measure:
- Ordinary pacing imperfections inside an otherwise intact arc (handled by P2 being lower, not by a
↪  penalty here).
- A single slightly long scene that still advances the story.

Anchors:
- 0-4: No meaningful collapse; story maintains focus through the end.
- 5-9: Some wobble, padding, or rushed wrap-up, but core arc stays intact.
- 10-14: Serious drift or filler undermines payoff or leaves thread dangling.
- 15-20: Major collapse; ending feels tacked-on, deflated, or story-ness significantly compromised.

N7) Dialogue problems (stilted, expository, same-voice) (0-15) (QUOTE REQUIRED if >0)
What it measures:
- Dialogue used mainly for information-dumping or explaining feelings/themes already obvious.
- Speech that sounds unnatural, overly formal, or interchangeable across characters.
What it does NOT measure:
- Internal monologue or narrative exposition that over-explains (penalize under N5).
- Depth or shallowness of character as people (handled under P3).

Applicability:
- If there is essentially no dialogue, score 0 and state that.

Anchors:
- 0: No meaningful issues or no significant dialogue.
- 1-5: Occasional stiffness or exposition, but generally serviceable.
- 6-10: Frequent issues; dialogue often feels wooden or on-the-nose.
- 11-15: Dialogue consistently undermines credibility, immersion, or subtext.

N8) Mechanical & formatting errors (0-8) (QUOTE REQUIRED if >0)
What it measures:
- Typos, grammar problems, malformed sentences, broken paragraphing, incorrect or inconsistent quotation
↪  marks, stray markup, etc. that distract or confuse.
What it does NOT measure:
- Deliberate stylistic deviations (e.g., poetic fragments) that are clearly intentional and consistent.
- Non-standard dialect that is coherent and purposeful.

Anchors:
- 0-1: Very clean; errors, if any, are trivial.
- 2-4: Intermittent distractions; minor but noticeable.
- 5-6: Frequent issues; reading is regularly interrupted.
- 7-8: Pervasive mechanical/formatting problems; significantly harm readability.

N9) Predictability & cliché (0-8) (QUOTE REQUIRED if >0)
What it measures:
- Stock plot structures and formulaic arcs (e.g., "the real treasure was friendship," the grumpy mentor
↪  who softens, the shy protagonist who finds confidence).
- Predictable narrative beats: if a seasoned reader can guess the next scene or the ending from the
↪  setup, that is predictability.
```

1466

```
- Recycled metaphors and imagery that appear across many LLM-generated stories at the concept level
↪  (e.g., "the sun rising as a metaphor for hope," "a tapestry of emotions"). For figurative-language
↪  problems at the sentence/passage level -- forced metaphor, density of similes, ornate phrasing --
↪  penalize under N10 instead.
- Formulaic endings: tidy resolutions, unearned epiphanies, heavy-handed moral lessons as final lines.
What it does NOT measure:
- Prose-level generic language or stock phrases (that is N3).
- Stylistic blandness or lack of voice (penalize via low P4).
- Overwrought or strained metaphor, or metaphor-heavy prose (penalize under N10).
- Using a familiar genre or trope is NOT automatically a penalty -- execution matters. A well-executed
↪  familiar arc scores 0-2; a lazy, paint-by-numbers version scores 5+.

Scoring guidance:
- Score by how predictable the story feels to an experienced reader, not by whether the premise is
↪  novel.
- A story can use familiar elements but arrange them in surprising ways -- that deserves a low score
↪  here.
- Penalize only when the predictability noticeably reduces engagement or makes the story feel like a
↪  template.

Anchors:
- 0-1: The story takes its own path; even if the genre is familiar, the specific choices feel earned and
↪  not-obvious.
- 2-4: Several predictable beats or a largely formulaic arc, but with some individual touches.
- 5-6: Paint-by-numbers story; experienced readers could outline the plot from the first paragraph.
- 7-8: Aggressively formulaic; every beat, character arc, and image feels copied from a template.

N10) Overwrought / nonsensical language & figurative density (0-20) (QUOTE REQUIRED if >0)
What it measures:
- Purple prose: ornate, adjective-stacked, or self-consciously literary writing that strains for effect
↪  rather than serving the moment.
- Forced metaphors and similes: comparisons that feel reached for, mismatched in tone or scale, or piled
↪  on top of each other.
- Ornate-for-its-own-sake passages: language that sounds writerly but does not earn its weight in the
↪  scene.
- Sentence- or clause-level nonsense: phrases that parse grammatically but fail to convey a coherent
↪  meaning. These can be writerly-sounding (e.g., "his thoughts moved like the silence of forgotten
↪  umbrellas") or just non-sequiturs (e.g., "she walked to the door because the morning was Tuesday").
↪  The unifying pattern is that the language sounds plausible but does not actually mean anything when
↪  examined.
- Figurative density and literary-flourish overload: sustained stacking of metaphors, similes, abstract
↪  personifications, or ornate flourishes across many consecutive sentences/paragraphs, to the point
↪  that prose becomes effortful to read. This fires EVEN IF individual figures are well-constructed --
↪  the failure mode is cumulative, not per-line. Symptoms include: nearly every sentence containing an
↪  "as if X" or "like Y" comparison; paragraph after paragraph leaning on abstracted poetic phrasing
↪  where direct prose would do; reader fatigue from never being allowed to sit in a plain sentence;
↪  figurative pile-ups where two or three metaphors share a single sentence and dilute each other.
What it does NOT measure:
- Genuine high-register prose that earns its texture and varies in density (rate as high P4 if it
↪  works).
- Generic or templated phrasing (that is N3).
- Story-level logic breaks or POV slips (those go under N2).
- Cliché or formulaic plot beats, or recycled concept-level imagery (those go under N9).

Distinguishing from related dimensions:
- N3 (generic): the prose is bland and could fit any story. N10: the prose strains for effect, fails at
↪  meaning, OR overloads on figurative language.
- N2 (coherence): the story's events or world contradict. N10: individual sentences or clauses sound
↪  like they say something but don't.
- A passage can be highly specific AND overwrought -- N10 fires when style outruns meaning, OR when
↪  figurative language dominates the page.

Density-based scoring guidance:
- Even when no single line is forced, sustained figurative-language saturation (e.g., a metaphor or
↪  simile in most sentences for paragraphs at a stretch) is itself the failure. Score by how hard the
↪  cumulative texture is to read, not by counting "bad" lines.
- A reader-fatigue test: if you find yourself wanting to skim to escape the constant comparisons, that
↪  is N10 territory.

Anchors:
- 0-4: Prose is appropriately textured; metaphors and stylistic choices land or stay restrained;
↪  sentences mean what they sound like; figurative language is used selectively.
- 5-9: Occasional purple flourishes, strained metaphors, or sentences that sound writerly but lose
↪  meaning on inspection. OR: noticeable but bounded figurative density -- a few paragraphs lean
↪  heavily on similes/metaphors, then ease back.
- 10-14: Frequent forced metaphors, ornate passages, or nonsensical clauses; style noticeably outruns
↪  meaning across multiple paragraphs. OR: sustained figurative density throughout long stretches --
↪  most sentences carry a comparison or flourish; reading becomes effortful.
- 15-20: Relentlessly overwrought or studded with sentence-level nonsense; nearly every reach for
↪  literary effect collapses; meaning is repeatedly sacrificed for sound. OR: the entire story is
↪  figurative-language-saturated to the point of unreadability -- metaphors stack on metaphors, similes
↪  on similes, with almost no plain prose between them.

==================================================
CATCH-ALL DIMENSIONS (USE SPARINGLY, OPEN-ENDED)
==================================================
```

```
These two dimensions are open-ended catch-alls for excellence or failure that the above P-dims and
↪  N-dims do not capture. They exist because no rubric can be exhaustive. They are intentionally hard
↪  to score nonzero -- strict structural requirements gate the score, and composite rules cap them when
↪  the story profile contradicts the score's direction.

------- BONUS: exceptional unmodeled merit (0-8) -------
When to consider:
- The story exhibits a specific excellence -- tied to a specific moment in the text -- that the existing
↪  P-dims do not adequately reward.
- Examples (NOT exhaustive -- many other patterns may qualify): sustained symbolic resonance (a single
↪  object/motif carrying thematic weight); structural ambition that pays off (recursive, mirror, frame,
↪  braided); irony/wit/tonal sophistication used as a load-bearing element; voice mastery beyond P4
↪  anchors (dialect, period pastiche, cadence); intellectual or meta-textual layering that deepens the
↪  story; any other clearly-named effect not anticipated by the rubric.
Required JSON fields for bonus_score > 0:
1. bonus_target_excellence: 1-2 sentences naming a SPECIFIC effect tied to a SPECIFIC moment. Vague or
↪  general praise (e.g. "beautiful prose", "engaging narrative") MUST score 0. Name what specifically
↪  and where.
2. bonus_evidence: array of up to 2 exact quotes from the story (each <=12 words). REQUIRED when
↪  bonus_score >= 2.
3. bonus_almost_captured_by: name the SINGLE P-dim that comes closest to already capturing this (e.g.
↪  "P4") OR "none" if truly orthogonal to all P-dims.
4. bonus_why_not_captured: 1 sentence explaining why that dim's anchors fall short of what is happening
↪  here.
If any required field is weak, vague, or missing, score = 0.
Composite rule (applied after scoring):
- If negative_total > positive_total / 2, bonus contribution to overall_score is 0.
- This rule is to discourage rewarding lift on stories with offsetting problems.
- You should still report the bonus_score honestly; the composite math handles the cap.

Anchors:
- 0: Default. No qualifying pattern, OR pattern present but already credited via P-dim.
- 1-2: Minor instance, briefly present.
- 3-5: Clear, sustained instance with strong evidence.
- 6-8: Pattern is a defining strength of the story; multiple categories may apply.
If you set bonus_score = 0, you may leave bonus_target_excellence as "No qualifying pattern observed.",
↪  bonus_evidence as [], bonus_almost_captured_by as "none", bonus_why_not_captured as "n/a".

------- PENALTY: exceptional unmodeled deficiency (0-8) -------
When to consider:
- The story exhibits a specific failure -- tied to a specific moment in the text -- that the existing
↪  N-dims do not adequately punish.
- Examples (NOT exhaustive): tonal mismatch (e.g. comedy where horror was needed; gravity in jokey
↪  premise); ethical / aesthetic miscalibration that breaks reader trust; any other named failure mode
↪  not anticipated by N1-N10.
Required JSON fields for penalty_score > 0:
1. penalty_target_deficiency: 1-2 sentences naming a SPECIFIC failure tied to a SPECIFIC moment. Vague
↪  "felt off" or "unsatisfying" justifications MUST score 0.
2. penalty_evidence: array of up to 2 exact quotes (each <=12 words). REQUIRED when penalty_score >= 2.
3. penalty_almost_captured_by: the SINGLE N-dim that comes closest (e.g. "N4") OR "none" if orthogonal
↪  to all N-dims.
4. penalty_why_not_captured: 1 sentence explaining why that dim's anchors fall short of this failure.
If any required field is weak, vague, or missing, score = 0.
Composite rule (applied after scoring):
- If positive_total > negative_total * 1.5, penalty contribution to overall_score is capped at 3.
- This prevents the catch-all from disproportionately punishing strong stories for edge-case flaws.
- You should still report the penalty_score honestly; the composite math handles the cap.

Anchors:
- 0: Default. No qualifying failure, OR failure present but already counted via N-dim.
- 1-2: Minor instance, localized.
- 3-5: Clear, sustained failure with strong evidence.
- 6-8: Pattern is a defining weakness of the story.
If you set penalty_score = 0, you may leave penalty_target_deficiency as "No qualifying failure
↪  observed.", penalty_evidence as [], penalty_almost_captured_by as "none", penalty_why_not_captured
↪  as "n/a".

==================================================
TOTALS & OVERALL SCORE
==================================================

Compute (you report these honestly; composite rules above apply at the consumer level):
- positive_total = P1 + P2 + P3 + P4 + P5 + P6  (0-100)
- negative_total = N1 + N2 + N3 + N4 + N5 + N6 + N7 + N8 + N9 + N10  (0-151)
- bonus_total    = bonus_score    (0-8)
- penalty_total  = penalty_score  (0-8)
- overall_score  = positive_total - negative_total + bonus_total - penalty_total

Important:
- Do NOT clamp at 0. overall_score can be negative if multiple severe issues stack.
- Report bonus_score and penalty_score honestly. Composite consumers may apply the bonus floor / penalty
↪  cap rules later; do not pre-apply them.

Labeling (based on overall_score):
- <=14   -> "very_poor"
- 15-34  -> "poor"
- 35-54  -> "fair"
- 55-74  -> "good"
```

1468

```
- >=75   -> "excellent"

Overall justification (2-5 sentences):
- Summarize the main strengths and weaknesses.
- Explain why the overall result isn't clearly higher and isn't clearly lower.
- Do NOT mention numeric scores.

==================================================
INPUT FORMAT
==================================================
STORY PROMPT:
{prompt}
STORY:
{story_text}

==================================================
OUTPUT FORMAT (JSON ONLY)
==================================================
Return ONLY a JSON object that matches the provided schema exactly.
No extra keys. No markdown. No commentary outside JSON.
```