Narrative Flattening: How Post-Training Compresses Thematic, Affective, and Stylistic Variation in LLM Fiction

Zehan Li[1,✉] **Yutong Zhu**[1] **Siyang Wu**[1] **Honglin Bao**[1,✉] **James A. Evans**[1]

[1]Knowledge Lab, University of Chicago
[✉] Correspondence: H.B. honglinbao@uchicago.edu;
Z.L. zehan@uchicago.edu

Abstract

Large language models produce fluent fiction, yet their creative output is widely seen as flat. We ask where this quality originates in the training and whether it affects different domains of human fiction equally. We construct a matched story-continuation paradigm across StoryStar (public-platform), TMAS (prompt-guided), and *The New Yorker* (professional literary)—and compare continuations from four OLMo 32B checkpoints (Base, SFT, DPO, RLVR) against matched human text. Because these checkpoints share architecture, scale, tokenizer, and pretraining, the design isolates the post-training effect. We measure each continuation along three sentence-level dimensions: thematic motion, affective prevalence, and linguistic diversity. Across all three, post-training compresses dynamic variation: thematic transitions become more uniform, high-intensity emotions give way to neutrality, and stylistic diversity across stories shrinks. We term this progressive loss *narrative flattening*. The effect is directionally stable across story domains but gap size depends on the human baseline: professional literary fiction is compressed most, while public-platform and prompt-guided stories show smaller gaps, consistent with their human baselines sitting closer to the model's default rhythm. Post-trained endpoints converge across domains, suggesting alignment produces a continuation regime largely insensitive to the source domain's narrative texture.

1 Introduction

Large language models are increasingly deployed as creative-writing assistants. Yet, a recurring complaint has emerged: LLM-generated fiction tends toward cliché, formulaic structure, stylistic monotony, and cross-piece homogeneity (Chakrabarty et al., 2024; Chakrabarty and Dhillon, 2026), and that human-AI co-writing reduces diversity across outputs (Doshi and Hauser, 2024).

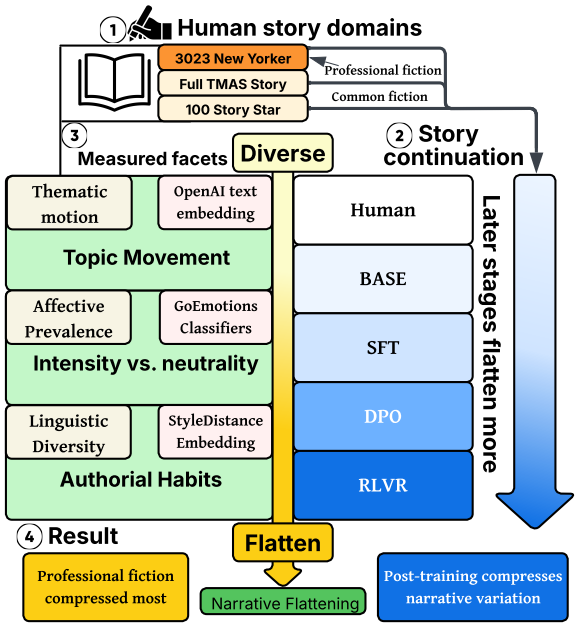

Figure 1: **Matched-continuation pipeline for measuring post-training effects on creative writing.** We collect short stories from three human writing domains, truncate each at four prefix lengths (40/60/80/90%), and complete each prefix with four OLMo-32B checkpoints (Base, SFT, DPO, RLVR). Continuations are analyzed along three narrative facets: thematic motion (sentence embeddings), affective prevalence (emotion classifier), and linguistic diversity (style embeddings). We measure how each post-training stage reshapes the dynamic structure of story continuations—and how the reshaping depends on the source domain.

These critiques remain at the level of lexical diversity scores, preference ratings, and whole-text quality judgments, none of which decomposes *how* a story becomes flat into measurable narrative dimensions. Missing is a mechanical account: *which properties* of narrative does model generation compress? Is the perceived flatness a matter of thematic movement, emotional dynamics, stylistic range—or all of these at once? Without decomposing the complaint into measurable components, 'AI writing feels flat' remains an observation, not a diagnosis.

Modern assistants pass through supervised fine-tuning (SFT), preference optimization (DPO), and reinforcement learning (RLVR)—stages that optimize for coherence, helpfulness, and human preference (Ouyang et al., 2022; Lambert et al., 2024), and that are known to narrow output diversity even on creative tasks (Kirk et al., 2024; Padmakumar and He, 2024; O'Mahony et al., 2024; Murthy et al., 2025). Yet few studies trace how each stage reshapes the texture of story generation, nor whether this convergence compresses dimensions that distinguish literary from generic prose.

Human creative writing is not a single baseline. Professionally edited literary fiction, public-platform stories, and prompt-guided narratives differ systematically in topic, perceived quality, and in how they modulate themes, emotional intensity, and stylistic density (Biber and Conrad, 2009; Underwood and Sellers, 2016). If post-training converges to a fixed continuation regime, the gap should be *domain-sensitive*: the farther a human writing domain lies from that regime, the more its narrative texture will be compressed. A single-baseline evaluation risks overstating or obscuring the effect; a cross-domain comparison is needed.

We call this pattern *narrative flattening*. A story traces a trajectory through thematic, affective, and linguistic space, and what distinguishes human fiction is not smoothness but structured variation—when to dwell on a motif, when to shift abruptly, when to heighten or suppress emotional pressure. We define *narrative flattening* as the measurable compression of this variation—regularized movement, muted affect, and reduced sensitivity to the source domain.

This paper asks two questions. First, how does each stage of the post-training pipeline—Base, SFT, DPO, RLVR—reshape the dynamic structure of story continuations? Second, does the magnitude of this reshaping depend on the human writing domain that the model is asked to continue?

To answer both, we construct a matched continuation paradigm spanning three distinct human writing domains in the creative-writing landscape: STORYSTAR, a public fiction platform with a low editorial barrier (StoryStar, 2026); TELL ME A STORY, a corpus of stories composed under explicit creative-writing prompts (Huot et al., 2024); and *The New Yorker*, a collection of professionally edited literary fiction published between 1945 and 2019 (Shaalan, 2022). We truncate each story at controlled prefix lengths and collect continuations from four checkpoints along the OLMo 32B instruction path (Team OLMo et al., 2025). Because these checkpoints share architecture, scale, tokenizer, and pretraining, the design lets us trace how successive post-training stages reshape narrative distributions within a single model lineage, while avoiding cross-model confounds in architecture and pretraining. We treat the resulting evidence as stage-wise and geometric rather than mechanistic: the design identifies where the post-trained endpoints move relative to human baselines, but not which specific data mixtures or reward signals cause the movement. We measure each continuation along three facets: thematic motion, affective prevalence, and linguistic diversity.

We find that each successive post-training stage narrows narrative variation across every facet we measure. In *thematic motion*, models traverse topics in more uniform steps, with per-story topic-jump variation falling below the human level by RLVR, dampening the alternation between dwelling and sharp shifts that characterizes human stories. In *affective prevalence*, high-arousal states such as surprise–curiosity and conflict are suppressed while neutral affect rises. In *linguistic diversity*, outputs collapse into a narrow stylistic attractor: closer to common-fiction baselines, but farther from professional literary style.

This pattern is *directionally stable* across three story domains, but its magnitude depends on the human baseline. Professional literary fiction is compressed most: its human distribution lies farthest from the post-trained model's default rhythm, so the narrowing is most visible. Public-platform and prompt-guided stories show smaller gaps—not because the model writes them better, but because these baselines already sit closer to where post-training converges. Strikingly, post-trained endpoints *converge* across corpora, with cross-domain style divergence dropping by over 90% from the human level to RLVR. This convergence suggests that post-training pushes all stories toward a single continuation pattern rather than adapting to each domain's complexity.

Together, these results reveal that post-training does not simply produce better fiction. It installs a fixed writing regime—coherent, readable, and conventionally well-formed, but flatter than any of the human domains it is asked to continue, with the largest gap appearing for the most stylistically and affectively varied human corpus.

Our contributions are threefold:

- We introduce a matched continuation paradigm that traces post-training effects across human writing domains differing in editorial gatekeeping and task framing.
- We operationalize 'AI flatness' as narrative flattening: measurable compression along thematic, affective, and linguistic dimensions.
- We show that post-training compresses narrative texture across three domains, and most for professional literary fiction, converging post-trained endpoints toward a single mundane regime.

2 Related Work

2.1 LLM Creative Writing and Its Evaluation

LLMs are increasingly used for creative writing, supporting story generation, co-writing, and screenplay assistance (Ippolito et al., 2022; Lee et al., 2022; Yuan et al., 2022). Most evaluations ask whether model-assisted writing is useful, creative, or preferred by human readers, relying on holistic judgments of quality, novelty, or satisfaction. This framing has surfaced promise and limits: model outputs can be fluent and useful, yet also clichéd, formulaic, and stylistically monotonous (Chakrabarty et al., 2024, 2025; Chakrabarty and Dhillon, 2026; Sui, 2026; Marco et al., 2024). Beyond individual quality, LLM assistance can homogenize collective output. Doshi and Hauser (2024) find that AI co-writing improves individual stories while reducing diversity across stories; Padmakumar and He (2024) show that instruction-tuned models—but not base models—reduce lexical and content diversity, suggesting that homogenization may be tied to instruction tuning rather than to language modeling alone. Xu et al. (2025) find that different aligned LLMs independently recycle the same narrative moves, implicating the aligned output space rather than any single model.

These studies identify the symptom but not its mechanism: holistic ratings cannot distinguish whether flatness reflects predictable topics shifting, smoothed emotional pressure, converging styles, or all three. We address this gap by decomposing narrative flattening along these dimensions.

2.2 Post-Training and Homogenization in Open-Ended Generation

Modern assistants undergo post-training through supervised fine-tuning (SFT), preference optimization (DPO), and reinforcement learning (RL), which improves instruction following but reshapes output distributions (Ouyang et al., 2022; Rafailov et al., 2024; Lambert et al., 2024). Kirk et al. (2024) show that RLHF reduces output diversity relative to SFT—a generalization–diversity tradeoff. O'Mahony et al. (2024) connect instruction and reward-based tuning to mode collapse, and Murthy et al. (2025) find that aligned models display less conceptual diversity than their base or instruction-tuned counterparts.

Creative writing is a particularly sensitive setting. In summarization or instruction-following, reduced variation is benign or even desirable. Users want consistency. In fiction, however, thematic shifts, affective volatility, and stylistic breadth are part of the task, not noise to eliminate. Yet existing work on alignment side effects rarely examines creative generation, and studies that do typically compare only a base model to a single aligned endpoint, without tracing the contribution of each stage. We address this gap by following one model lineage across four checkpoints—Base, SFT, DPO, and RLVR—tracing stage-wise shifts in narrative texture within a single model family to avoid confounds.

2.3 Computational Accounts of Narrative Texture

Narrative theory treats stories as organized trajectories, not unordered sentences. Classic accounts emphasize how suspense, curiosity, and surprise arise from temporal ordering and revelation of events (Sternberg, 1978; Brewer and Lichtenstein, 1982). Computational work operationalizes this at scale: Reagan et al. (2016) extract emotional arcs from Project Gutenberg fiction using sentiment trajectories; Ouyang and McKeown (2015) model reportable events as turning points; and Piper et al. (2023) model narrative revelation using information-theoretic methods.

Piper et al. (2021) argues that computational narrative understanding must account for event ordering, temporality, and discourse structure, not surface coherence alone. A parallel tradition in computational stylistics and register analysis shows that linguistic diversity varies systematically across genres, domains, and literary prestige (Biber and Conrad, 2009; Underwood and Sellers, 2016).

These lines of work establish that narrative texture—how a story moves through thematic, affective, and stylistic space—is measurable rather than merely impressionistic. Recent work has begun applying such metrics to LLM evaluation, measuring tension or arc diversity in generated stories (Sui

et al., 2026; Tian et al., 2024), but these studies compare outputs without tracing how each alignment stage shapes the effect. We repurpose these metrics as a diagnostic tool to ask at which stage of post-training narrative compression accumulates.

3 Experimental Design

Matched continuation setting. We evaluate literary continuation rather than free generation. For each story, we reveal a prefix at cut point $c \in \{40\%, 60\%, 80\%, 90\%\}$ and treat the remaining original text as the matched human continuation. Each model receives the same prefix, and all measurements are computed only on the continuation. The four cut points yield continuations from different narrative positions, testing whether flattening effects are robust across how much context the model has seen.

Story domains. We select three corpora that span a range of human creative-writing settings, differing in editorial gatekeeping, task framing, and expected narrative complexity (Table 1).

- **Professional literary fiction.** 3,023 short stories published in *The New Yorker* between 1945 and 2019 (Shaalan, 2022), restricted to stories under 5 K words. These stories have passed professional editorial selection and represent a high-density literary baseline with rich thematic modulation, affective tension, and stylistic variety.
- **Prompt-guided common fiction.** TELL ME A STORY (TMAS) comprises human-written stories elicited by explicit creative-writing prompts (Huot et al., 2024). Because writers respond to a shared task instruction, this corpus captures human fiction produced under conditions that parallel instruction-following generation.
- **Public-platform fiction.** *StoryStar* stories are drawn from a free online publishing platform open to writers of all backgrounds (StoryStar, 2026).[1] This corpus represents everyday short-story writing with minimal editorial filtering.

Domain	Stories	Characterization
New Yorker	3,023	Professional literary fiction
TMAS	230	Prompt-guided human fiction
StoryStar	100	Public-platform fiction

Table 1: Three human story domains used in this study, ordered by editorial gatekeeping. Together they span professional, task-elicited, and public-platform writing.

[1] https://www.storystar.com

Together, the three domains let us separate two effects: how post-training reshapes continuation dynamics (Axis 1, training stage), and whether the magnitude of that reshaping depends on the narrative texture of the human source domain (Axis 2, story domain).

Cross-domain controls. To ensure that observed differences reflect narrative properties rather than surface confounds, we apply several controls across corpora. Stories in all three domains are restricted to a comparable word-length range. The same four cut points are applied uniformly. Model continuations use a fixed decoding configuration across stages and domains. Prompting is held constant within model interface: base models receive the raw story prefix, while instruction-tuned models use the tokenizer-provided chat template with the same continuation instruction.

Models and generation. We compare four checkpoints from the OLMo 32B instruction path: Base, SFT, DPO, and RLVR (Team OLMo et al., 2025). These checkpoints share the same base model and form a sequential post-training path, making post-training stage the primary experimental variable. For each story–cut–model tuple, we sample 5 continuations under identical decoding settings. The prompt and Model setting are available in Appendix B. Appendix I reports a prompt-interface control; Base→SFT should be interpreted as interface + post-training, while all others share the instruction interface.

Sentence-level facets. We represent each continuation as a sentence-level trajectory. For a continuation with sentences $x_1, \ldots, x_T$, encoder f_d maps each sentence to a facet-specific vector $z_t^{(d)} = f_d(x_t)$, yielding a trajectory through that facet space. We measure three facets:

Flattening metrics. We operationalize narrative flattening as reduced variation relative to matched human continuations; full metric definitions are in Appendix G. For **thematic motion**, we encode each sentence with `text-embedding-3-large` (3072d) and measure the coefficient of variation (CV) of sentence-to-sentence semantic jump sizes within each continuation. For **affective prevalence**, a literary-adapted GoEmotions classifier (Demszky et al., 2020) (28 classes; Appendix C) assigns each sentence a top-1 emotion label; we track the prevalence of conflict, surprise–curiosity, and neutral

affect. For **linguistic diversity**, we represent each story with StyleDistance embeddings (768d) and measure Maximum Mean Discrepancy(MMD) to the human style distribution and across-story variance. Length-sensitive metrics are residualized for continuation length.

4 Results

4.1 Human Story Domains Differ in Baseline Narrative Texture

Before examining model continuations, we compare human continuations across the three domains (Table 2).

Corpus	Theme CV	Affect %	Style Axis
New Yorker	0.105	41.0	0.22
TMAS	0.098	26.5	−2.07
StoryStar	0.110	23.2	−1.99

Table 2: Human story domains differ in thematic rhythm (per-story topic-jump CV), affective charge (high emotional intensity %), and stylistic register (mean z-score on style PCA, which explains 88.5% of variance; positive = more literary).95% bootstrap CIs are reported in Appendix J.1; all pairwise domain differences in affective charge and style position are significant

In *thematic motion*, the domains differ modestly: topical step-size CV ranges from ∼0.098 (TMAS) to ∼0.110 (StoryStar), with *The New Yorker* in between at ∼0.105. The contrasts in *affective prevalence* are larger: *The New Yorker* carries the highest combined surprise–curiosity and conflict share (∼41%), with TMAS at ∼26% and StoryStar lowest at ∼23%. *Linguistic diversity* shows the sharpest divide. On the first principal component of style-neural embeddings (88.5% of variance; positive = more literary register), *The New Yorker* stories cluster at $z = 0.22$, while TMAS (−2.07) and StoryStar (−1.99) occupy a distinct, nearly overlapping region more than two standard deviations away. These results show that the three corpora differ most sharply in affective charge and stylistic signature, with *The New Yorker* separated from the other two corpora, while prompt-guided and public-platform fiction cluster together. This raises the question that motivates the rest of our analysis: if post-trained continuations fall on the common-fiction side of this divide, the model–human gap should be largest for *The New Yorker*, whose human baseline lies farthest from that side.

4.2 Post-Training Flattens Story Continuation Across Facets

Results are shown primarily for *The New Yorker* (Figures 2– Table 3); cross-domain comparisons follow in Section 4.3, and per-facet breakdowns for the other two corpora are in Appendix J.1. The same directional pattern appears in Qwen2.5-32b base/LLaMA-3.1-8b base/Gemma-3-12B base versus their instruction-tuned counterparts (Appendix K), confirming the effect is not specific to the OLMo lineage.

Thematic motion. The per-story coefficient of variation of topic jumps drops at every stage: from a human mean of ∼0.105 to 0.096 at Base (−8.0%), 0.089 at SFT (−15.1%), and 0.081 at DPO/RLVR (−22.2%) (Figure 2A). The underlying distribution tells the same story: human CV values spread broadly, reflecting stories that mix dwelling with sharp thematic pivots, whereas each post-training stage compresses the distribution leftward into an increasingly narrow peak (Figure 2B). A mixed-effects model with story-level random intercepts and fixed cut-point effects confirms the RLVR–human reduction in L_2 topical CV ($\beta = -0.0228$, 95% CI $[-0.0234, -0.0222]$, $p < .001$).

Affective prevalence. The base model emerges from pretraining over-marked: conflict (∼47%) and surprise–curiosity (∼33%) both exceed human prevalence (∼20% and ∼21% respectively), while neutral content is correspondingly underweighted (∼13% vs. ∼29% in humans) (Figure 3). This likely reflects the over-representation of emotionally marked genres in web-scale pretraining data. Post-training does not return the model to human levels; it over-corrects past them. By RLVR, conflict has collapsed to ∼7.5% and surprise–curiosity to ∼13%, while neutral content swells to ∼45%—each marker now further from the human distribution than the base model was, in the opposite direction. Human continuations occupy intermediate, modulated affective values; post-trained outputs selectively compress the two focal high-arousal families measured here while increasing neutral narration. A full family-level decomposition shows that this is not a global collapse of every affect category: sadness/loss and warmth/affiliation remain comparatively stable after SFT (Appendix E). Flattening effects are directionally stable across all four cut points (Appendix L). The same model con-

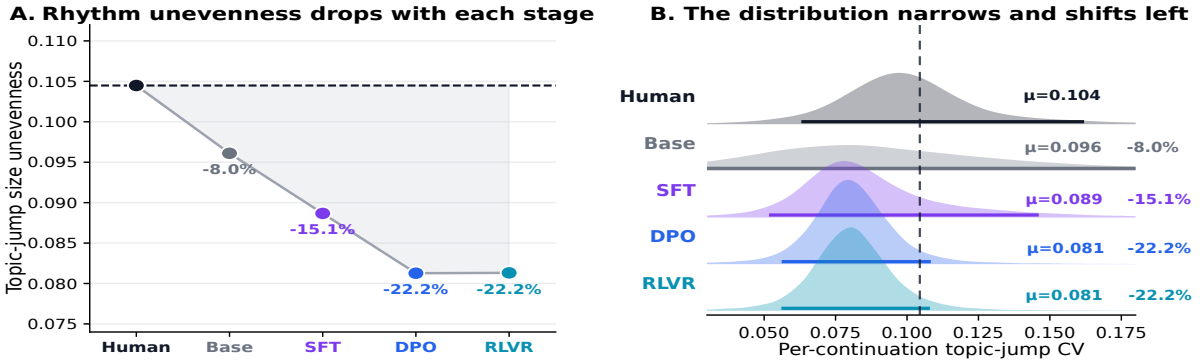

Figure 2: **(A)** Per-story CV (σ/μ) of sentence-to-sentence topic-jump L_2 distances. Dashed line = human mean; percentages = unevenness lost relative to human. Length regression confirms continuation length does not confound this metric ($R^2 < 0.001$). **(B)** Distribution of the same per-continuation CV. Dashed line = human mean; brackets span the 5th–95th percentile. Post-training progressively narrows the distribution and shifts it leftward. 95% bootstrap CIs are narrower than the plot markers at all stages (Appendix J.1).

firms both the reduction in conflict prevalence ($\beta = -0.1228$, 95% CI $[-0.1270, -0.1185]$, $p < .001$) and the increase in neutral narration ($\beta = 0.1632$, 95% CI $[0.1601, 0.1663]$, $p < .001$). These effects remain significant after adding log realized sentence length as a covariate (Appendix H.1).

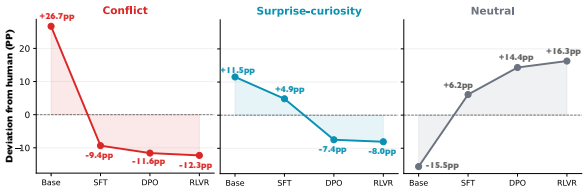

Figure 3: Each panel shows the percentage-point deviation of model prevalence from the matched human baseline; zero on the y-axis indicates the human level. Affective prevalence across OLMo post-training stages for *The New Yorker* continuations. The base model emerges from pretraining over-marked . Successive post-training stages suppress the focal conflict and surprise–curiosity families and inflate neutral content, overshooting human prevalence rather than converging on it. 95% story-bootstrap CIs are narrower than the plot markers at all stages (Appendix J.1); a six-family robustness decomposition is reported in Appendix E.

Linguistic Diversity. MMD to the human style distribution increases monotonically from 0.24 (Base) to 0.41 (SFT) to 0.52–0.53 (DPO/RLVR), while across-story style variance drops from $\sim 6\times$ human at Base to $0.85\times$ at SFT and 0.5–$0.55\times$ at DPO/RLVR (Table 3). The base model's MMD is low because its stylistic sprawl is broad enough that its aggregate footprint overlaps human writing, but it is writing in too many directions at once. Post-training collapses this sprawl onto a narrow attractor that is offset from the human distribution—the model converges on a single voice and applies it across every story. A fixed-sentence-count control shows that this variance collapse is not an artifact of different realized continuation lengths (Appendix H.2).

Stage	MMD^2 (↑ farther)	Var/human (↓ less varied)
Human	— (ref)	1.00 (ref)
Base	0.25	6.03
SFT	0.41 ↑+64%	0.84 ↓ −86%
DPO	0.53 ↑+112%	0.49 ↓ −92%
RLVR	0.52 ↑+108%	0.52 ↓ −91%

Table 3: Style divergence and variance across post-training stages for New Yorker continuations. Each successive stage pushes the model's style distribution farther from the human reference (higher MMD^2) while collapsing across-story variation (lower variance), converging onto a narrow stylistic attractor offset from human writing. MMD confidence intervals use the sentence-state bootstrap described in Appendix F; across-story variance uses story-level bootstrap.

4.3 Post-Training Erases Cross-Domain Differences, Compressing Professional Fiction Most

Section 4.1 established that human story domains occupy distinct narrative regimes—*The New Yorker* separated from TMAS and StoryStar by higher affective charge and a markedly different stylistic signature. We find that after post-training, cross-domain differences in model outputs become substantially smaller than cross-domain differences in human outputs. (Figure 4).

Cross-domain spread collapses under post-training. Across three measurement facets, the gap between domains narrows though the degree of convergence varies. In *thematic motion*, the range of topical jump CV drops from 0.0116 (human)

to 0.0037 (RLVR)—a substantial 62% reduction, but leaving residual cross-domain structure (Figure 4A). In *affective prevalence*, surprise+conflict range falls from 17.8 to 3.3 percentage points, an 81% reduction (Figure 4B). Convergence is strongest in *linguistic habits*: human cross-domain MMD^2 reaches 0.61, while the largest RLVR cross-domain MMD^2 is 0.01—a substantial reduction (Figure 4C). By the final post-training stage, the three RLVR endpoints occupy largely overlapping regions of stylistic space despite originating from sharply different human domains.

Professional fiction bears the largest gap. Because post-training pushes all domains toward a similar continuation regime, the domain that starts farthest from that regime undergoes the largest shift. In affective prevalence, *The New Yorker*'s combined surprise+conflict share falls from ∼41% to ∼20% at RLVR—a drop of roughly 21 percentage points, compared with ∼7 points for TMAS and ∼6 points for STORYSTAR. Linguistic habits provide the starkest evidence (Figure 4C): under human authorship *The New Yorker* occupies a clearly separated region of style space, while TMAS and STORYSTAR sit on top of one another; under RLVR, all three collapse onto a shared region that, in the two leading principal components, overlaps with the common-fiction human baselines (full-dimensional MMD in Appendix F.2). The smaller gaps for STORYSTAR and TMAS do not mean the model writes these domains better; they mean these baselines were already closer to where post-training pushes the model.

A domain-agnostic continuation regime. Under this post-training pipeline, model continuations across the three domains converge, with cross-domain differences substantially smaller than those observed in human continuations. As the convergence pattern above shows, the output regime sits closer to common-fiction baselines than to professional literary fiction. The resulting continuations are coherent, readable, and sequentially well-formed, but they shed the thematic unevenness, affective tension, and stylistic variation that distinguish human storytelling—and that distinguish one kind of human storytelling from another.

5 Discussion

Post-training as narrative regularization. Our results suggest post-training acts as a kind of regularization. This is not the parametric regularization familiar from optimization, but a regularization of the narrative trajectory. The base model is not simply human-like: it can be affectively over-marked and stylistically wide-ranging and unstable. Supervised fine-tuning, preference optimization, and verifiable-reward reinforcement learning make generation more controlled and conventionally well-formed, but this control is achieved by suppressing variation. Thematic jumps become more uniform, affect is muted, and linguistic habits collapse into a narrower attractor. These findings offer one possible structural account of why aligned fiction can feel mechanical: the issue may not be grammar but the loss of dynamic contrast over the course of a continuation. Direct reader studies linking these structural metrics to human perception remain an important next step.

What the cross-domain design reveals. Evaluating model fiction against a pooled human reference would blur a distinction the cross-domain design makes visible. Professional literary fiction, prompt-guided fiction, and public-platform fiction differ before any model is introduced: they vary in affective charge, stylistic signature, and narrative rhythm. A single-baseline evaluation against STORYSTAR alone would suggest the model is largely faithful; against *The New Yorker* alone would suggest catastrophic gaps, and the pooled average masks both. The cross-domain design shows that the human–model gap is not a fixed property of the model but a function of which human baseline is used to measure it, and that post-trained endpoints converge toward each other regardless of the source they were asked to continue.

Geometry, not yet mechanism. Our analysis localizes *where* post-trained outputs sit relative to human baselines, but not *why*. Several mechanism are compatible with the geometry we observe. Preference annotators may reward more conventional continuations; instruction-tuning data may overrepresent task-oriented prose whose style bleeds into open-ended generation; RLVR's verifiability pressure may favor common-fiction-aligned structures; reward models trained on broad preferences may compress the literary tail of the human distribution. Disentangling these channels requires access to the full SFT, DPO, and RLVR mixtures, along with controlled interventions on each, which we leave to future work. The current conclusion is geometric: the post-trained endpoint lies closer to common-

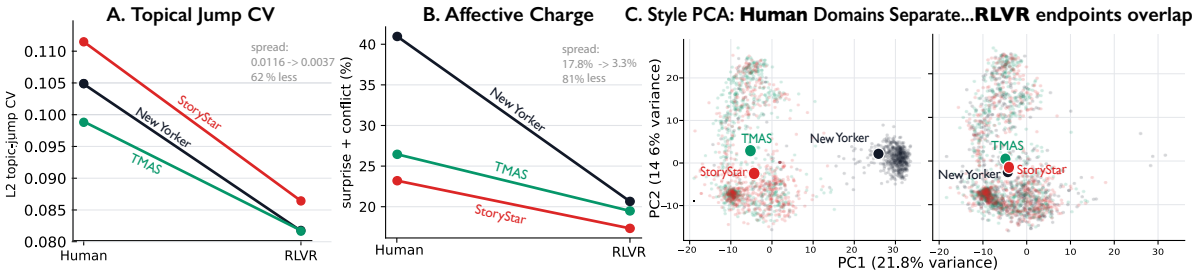

Figure 4: Cross-domain convergence under post-training. **(A)** Domain-mean topical jump CV for three corpora (*New Yorker*, *Tell Me A Story*, *StoryStar*) at the human and RLVR endpoints; corpus spread shrinks $0.0116 \rightarrow 0.0037$ (62%). **(B)** Domain-mean affective charge (surprise–curiosity + conflict, %) at the same two endpoints; spread shrinks $17.8 \rightarrow 3.3$ percentage points (81%). **(C)** PCA of 768-dimensional StyleDistance embeddings. Each small dot is an individual story continuation, color-coded by domain; each large circle is the centroid of one domain (mean position of stories). The left panel shows human-authored continuations; the right shows the same three domains continued by RLVR. Under human authorship, *New Yorker* occupies a clearly separated region of style space while TMAS and STORYSTAR cluster together; under RLVR, all three centroids collapse onto a single shared region.

fiction baselines than to professional literary fiction, with the gap widening at each training stage.

Flattening or professionalization? A natural alternative reading is that what we call flattening is what (muscular) literary editors do. They remove excess affective marking, smooth erratic motion, and enforce stylistic consistency. The base model's affective over-marking (§4.2) is consistent with this reading: perhaps post-training is editing the model toward a more disciplined voice. Two findings cut against this interpretation. First, the post-trained endpoint does not approach the human distribution; it overshoots in the opposite direction, ending farther from human affect than the base model was, on the muted side. Second, professional human editing produces *The New Yorker*, the domain from which post-trained models move most aggressively away. Post-training is not professionalizing fiction. It is converging on a third regime that is neither the base model's sprawl nor the literary editor's discipline, but a default closer to common-fiction baselines.

Implications for creative-writing assistants. These findings complicate the success criteria for post-training in open-ended generation. In summarization, reduced variation is desirable. Users want consistency. In fiction, variation leads to heightened experience, and uniform compression becomes a bug rather than a feature. Creative-writing assistants may require alignment objectives distinct from general-purpose assistants: domain-conditional reward models that calibrate to the source register rather than a pooled human preference; distributional matching objectives that preserve cross-story variance rather than only pointwise quality; or preference data deliberately sampled from the literary tail rather than from majority-preferred continuations. Our metrics provide a mechanism account of what Doshi and Hauser (2024) observed at the level of collective output. Where they showed that human–AI co-writing reduces diversity across stories, we show which narrative dimensions are compressed and how compression accumulates across training stages. Narrative-flattening metrics complement human-preference evaluation by asking not whether each continuation is plausible, but whether a collection of continuations preserves the variance of human fiction.

6 Conclusion

We compared matched human story continuations with four OLMo 32B checkpoints across three fiction domains. Across thematic motion, affective prevalence, and linguistic diversity, post-training regularizes narrative trajectories: topic jumps become more uniform, high-intensity affect gives way to neutral narration, and across-story style variance collapses. This compression is strongest for *The New Yorker* because its human baseline lies farthest from the aligned model's continuation regime; TMAS and STORYSTAR appear closer mainly because their baselines already sit nearer that attractor. The result is not generic degradation but domain-insensitive convergence. For creative-writing assistants, coherence and readability are therefore insufficient objectives: alignment should preserve source-domain variation, making narrative flattening a measurable target for future training.

Limitations

Our analysis is not mechanistic. We do not directly inspect the SFT, DPO, or RLVR mixtures at the granularity needed to attribute compression to specific training signals—reward-model preferences, instruction-data style, or RLVR verifiability constraints. Our conclusions are therefore geometric (i.e., about where the post-trained endpoint sits relative to human baselines) rather than mechanistic. Disentangling the contributions of preference-data composition, reward shape, and verifiability pressure to narrative flattening is an important next step and would require controlled interventions we cannot perform on a deployed pipeline.

Corpus and language coverage. Our three domains span professional literary fiction, prompt-elicited stories, and public-platform writing, but they are all English, short-form prose, and weighted toward contemporary writing. *The New Yorker* is also a single publication: its editorial style is not truly coextensive with "professional literary fiction," and corpora drawn from *Granta*, *Paris Review*, or *Tin House* may sit differently in our facet space. Genre fiction, novel-length narrative, screenplay, poetry, and non-English literary traditions remain to be tested. The under-5K-word restriction also excludes long-form literary fiction, where the most distinctive narrative texture likely lives.

Decoding regime. All continuations use a single shared decoding configuration. We do not exhaustively map the interactions between sampling temperature, nucleus thresholds, repetition penalties and stage-wise compression. Prior work shows decoding choices substantially shape output diversity (Holtzman et al., 2020). Aggressive decoding may partially restore variance lost during training without fully reversing the regime shift.

Use of LLMs. We used LLMs for sentence-level polishing (clarity, wording, and grammatical corrections) and limited implementation assistance for small refactors, boilerplate, and training-related code. All LLM-suggested changes were reviewed, edited as needed, and verified by the authors.

Ethical Considerations

Copyright and data release. Our research utilizes publicly available or archived literary datasets.To ensure ethical compliance, the *New Yorker* data is stored in a secure repository at our university; its use in this study constitutes fair use, as it is employed strictly for non-disseminating, non-commercial research. Other corpora, such as TMAS and StoryStar, were sourced from public-facing websites intended for open readership. We report only aggregate statistics. We do not release raw *New Yorker* stories, long excerpts, or full generated continuations conditioned on copyrighted prefixes. Any released code will operate on user-provided texts or public-domain examples; released artifacts will be limited to aggregate metric tables, plotting scripts, and anonymized metadata that do not contain protected story text.

Privacy and public-platform data. For public-platform stories, author names, bylines, source paths, and other obvious identifiers are removed or excluded from analysis. Our unit of analysis is the story text and its aggregate continuation metrics, not individual authors. We do not make claims about author traits, demographic groups, or identifiable writers.

Human annotation and AI assistance. The affect classifier was adapted using LLM-distilled labels and validated on a 1,000-sentence calibration subset annotated by one author and two graduate-student volunteers. Annotation was voluntary and unpaid, and the labels are used only for aggregate label-quality validation; we report no annotator-level performance or demographic analysis. We also used LLM assistance for sentence-level polishing and limited implementation support. All LLM-suggested writing and code changes were reviewed, edited as needed, and verified by the authors.

Potential harms. The goal of this work is diagnostic rather than prescriptive. We do not claim that professional literary fiction is the only valid writing style, nor that lower affective charge is inherently worse. The main risk we identify is aggregate homogenization: when many users rely on the same aligned systems, diverse human storytelling practices may be nudged toward a narrower continuation regime. We present the proposed metrics as tools for measuring and mitigating this risk, not for ranking authors, policing creative style, or prescribing a single standard for good fiction.

References

Douglas Biber and Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge University Press.

William F. Brewer and Edward H. Lichtenstein. 1982.

Stories are to entertain: A structural-affect theory of stories. *Journal of Pragmatics*, 6(5):473–486.

Tuhin Chakrabarty and Paramveer S Dhillon. 2026. Can good writing be generative? expert-level ai writing emerges through fine-tuning on high quality books. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–27.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–34. Association for Computing Machinery.

Tuhin Chakrabarty, Philippe Laban, and Chien-Sheng Wu. 2025. Can AI writing be salvaged? mitigating idiosyncrasies and improving human-AI alignment in the writing process through edits. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Anil R. Doshi and Oliver P. Hauser. 2024. Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28):eadn5290.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations (ICLR)*.

Fantine Huot, Reinald Kim Amplayo, Jennimaria Palomaki, Alice Shoshana Jakobovits, Elizabeth Clark, and Mirella Lapata. 2024. Agents' room: Narrative generation through multi-step collaboration. *arXiv preprint arXiv:2410.02603*.

Daphne Ippolito, Ann Yuan, Andy Coenen, and Sehmon Burnam. 2022. Creative writing with an ai-powered writing assistant: Perspectives from professional writers. *Preprint*, arXiv:2211.05030.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2024. Understanding the effects of RLHF on LLM generalisation and diversity. In *International Conference on Learning Representations*.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, and 4 others. 2024. Tülu 3: Pushing frontiers in open language model post-training. *Preprint*, arXiv:2411.15124.

Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *CHI Conference on Human Factors in Computing Systems*, CHI '22, page 1–19. ACM.

Guillermo Marco, Julio Gonzalo, M. Teresa Mateo-Girona, and Ramón Del Castillo Santos. 2024. Pron vs prompt: Can large language models already challenge a world-class fiction author at creative text writing? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Sonia Krishna Murthy, Tomer Ullman, and Jennifer Hu. 2025. One fish, two fish, but not the whole sea: Alignment reduces language models' conceptual diversity. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 11241–11258. Association for Computational Linguistics.

Laura O'Mahony, Leo Grinsztajn, Hailey Schoelkopf, and Stella Biderman. 2024. Attributing mode collapse in the fine-tuning of large language models. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

Jessica Ouyang and Kathleen McKeown. 2015. Modeling reportable events as turning points in narrative. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2149–2158. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Vishakh Padmakumar and He He. 2024. Does writing with language models reduce content diversity? In *International Conference on Learning Representations*.

Andrew Piper, Richard Jean So, and David Bamman. 2021. Narrative theory for computational narrative understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 298–311. Association for Computational Linguistics.

Andrew Piper, Hao Xu, and Eric D. Kolaczyk. 2023. Modeling narrative revelation. In *Proceedings of the Computational Humanities Research Conference (CHR)*, pages 500–516.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Andrew J. Reagan, Lewis Mitchell, Dilan Kiley, Christopher M. Danforth, and Peter Sheridan Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(31).

Nora Shaalan. 2022. The view from the fiction of the New Yorker. Public Books.

Meir Sternberg. 1978. *Expositional Modes and Temporal Ordering in Fiction*. Johns Hopkins University Press, Baltimore.

StoryStar. 2026. Storystar: Short stories by writers around the world. https://www.storystar.com/. Online short-story platform; accessed May 2026.

Peiqi Sui. 2026. LLMs exhibit significantly lower uncertainty in creative writing than professional writers. arXiv preprint arXiv:2602.16162.

Peiqi Sui, Yutong Zhu, Tianyi Cheng, Peter West, Richard Jean So, Hoyt Long, and Ari Holtzman. 2026. Spoiler alert: Narrative forecasting as a metric for tension in LLM storytelling. *Preprint*, arXiv:2604.09854.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, and 1 others. 2025. 2 OLMo 2 Furious. *Preprint*, arXiv:2501.00656.

Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. 2024. Are large language models capable of generating human-level narratives? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Ted Underwood and Jordan Sellers. 2016. The longue durée of literary prestige. *Modern Language Quarterly*, 77(3):321–344.

Weijia Xu, Nebojša Jojić, Sudha Rao, Chris Brockett, and Bill Dolan. 2025. Echoes in AI: Quantifying lack of plot diversity in LLM outputs. *Proceedings of the National Academy of Sciences*, 122.

Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story writing with large language models. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, pages 841–852. Association for Computing Machinery.

A Data Statement and Corpus Construction

Overview. We evaluate story continuation across three English short-fiction domains: professional literary fiction, prompt-guided human fiction, and public-platform fiction. Table 4 reports corpus-level statistics after filtering.

***The New Yorker*.** We use a restricted corpus of short fiction published in *The New Yorker* between 1945 and 2019. To focus on short-form continuation and to maintain a comparable length range across corpora, we retain stories under 5,000 words. The source texts are stored in a restricted university repository and are used only for non-commercial research. We do not redistribute the original stories, long excerpts, or any generated continuations that would reproduce substantial copyrighted context. All reported results are aggregate statistics.

TELL ME A STORY. TELL ME A STORY (TMAS) is a publicly available corpus of human-written stories composed in a guided creative-writing setting. We use the full corpus of 230 stories. Because TMAS writers respond to explicit writing prompts, this domain provides a human baseline for prompt-guided fiction.

STORYSTAR. STORYSTAR is a corpus of 100 short stories collected from `storystar.com`, a public creative-writing platform. Source-level author identifiers are not used as analytic variables. The corpus is used as a public-platform fiction baseline with minimal editorial filtering. The scrape was conducted in February 2026. We do not redistribute raw StoryStar text unless permitted by the source license or terms.

Preprocessing. Across all corpora, we convert each story to a plain-text record, strip leading and trailing whitespace, and normalize whitespace when constructing prefixes and continuations. File names, source paths, prompt fields, and collection metadata are retained as metadata rather than analyzed as narrative text. Titles and bylines are removed before segmentation. We do not apply automatic semantic rewriting or paraphrase-level deduplication; after the file-level corpus filters above, each remaining file is treated as one story record. Stories are sentence-split using the same deterministic rule-based splitter used by the continuation-generation pipeline: sentence boundaries are placed after terminal punctuation (`.`, `!`, or `?`, optionally followed by closing quotes or brackets) when followed by whitespace and an uppercase letter, quotation mark, opening parenthesis, or opening bracket. Prefix cut points are computed over these sentence sequences, so each pre-

Corpus	Raw stories	Used stories	Filter	Mean words	Median words	IQR
The New Yorker	5001	3,023	$< 5{,}000$ words	2,763	2,695	1,737–3,776
TELL ME A STORY	230	230	full dataset	1,468	1,405	1,034–1,898
STORYSTAR	100	100	full collected set	2,614	2,478	1,703–3,416

Table 4: Corpus statistics after filtering. Word counts are computed after preprocessing and before prefix truncation. The interquartile range (IQR) reports the 25th–75th percentile word-count range.

fix ends at a sentence boundary. For a cut point $c \in \{40\%, 60\%, 80\%, 90\%\}$, the prefix contains the first c proportion of sentences and the matched human continuation is the remaining sentence suffix.

B Generation Setup

Task. For each story, we generate continuations from prefixes ending at 40%, 60%, 80%, and 90% of the sentence sequence. The held-out human suffix is treated as the matched human continuation. For each story–cut–model tuple, we sample five model continuations.

Models. The primary experiments use four checkpoints from the OLMo 32B instruction path: Base, SFT, DPO, and RLVR. The exact model identifiers are listed in Table 5.

Stage	Model identifier
Base	`allenai/Olmo-3-1125-32B`
SFT	`allenai/Olmo-3.1-32B-Instruct-SFT`
DPO	`allenai/Olmo-3.1-32B-Instruct-DPO`
RLVR	`allenai/Olmo-3.1-32B-Instruct`

Table 5: Model checkpoints used in the OLMo stage-wise analysis.

Prompting. Base models receive the raw story prefix only:

```
{story_so_far}
```

Instruction-tuned models use the tokenizer-provided chat template with the following system and user messages:

> **System:** You are a fiction writer. Continue the story naturally in the same style and voice. Write only story text – no commentary, no meta-discussion, no preamble, no quotation marks around your continuation.
>
> **User:** Continue this story to its conclusion in approximately `{n_words}` words. Maintain the same tone, style, and narrative voice throughout. Do not summarize or describe what happens – write the actual story text as it would appear on the page.
>
> `STORY SO FAR:`
>
> `{story_so_far}`

Prompting is therefore held constant across domains within each model interface. Base and instruction-tuned models differ only in the interface required by the checkpoint: raw completion for base models and chat-template prompting for instruction-tuned models.

Compute and infrastructure .All generation was run on a mix of two NVIDIA H100 and two NVIDIA L40S GPUs using vLLM. We generated continuations for 3,500 stories per checkpoint across the OLMo path (four checkpoints: Base, SFT, DPO, RLVR) and the three additional model families (Qwen2.5-32B, Llama-3.1-8B, and Gemma-3-12B, two checkpoints each), at four cut points with five samples per story–cut–model tuple. Throughput depended on model size and hardware: for the 32B models, generation ran at roughly 0.7 stories/min on the paired L40S and 2.4 stories/min on the paired H100; for the 8B model, roughly 2.6 stories/min per L40S and 4.5 stories/min per H100. Sentence-level affect classification, thematic embedding, and StyleDistance encoding were run as separate downstream passes over the archived continuations and are comparatively inexpensive relative to generation.

Decoding. All reported OLMo continuations use the same stochastic decoding configuration: temperature $= 1.2$ and nucleus sampling with top-$p = 0.95$. The target continuation length is set to the word count of the held-out human suffix. The maximum decoding budget is set dynamically as

$$\texttt{max\_tokens} = \lfloor \texttt{target\_words} \times 1.3 \times 1.15 \rfloor ,$$

with a minimum of 64 tokens and a maximum of 2048 tokens. We do not specify a custom stop sequence; generation stops when the model emits EOS or reaches the token budget. No explicit decoding seed is passed to vLLM. All generated continuations used in the analysis are archived, and all downstream metrics are computed deterministically from those archived generations.

Domain	Source	Mean words	Mean sentences	Human word ratio	Sentence ratio
New Yorker	Human	875.1	55.6	1.00	1.00
New Yorker	Base	497.8	12.1	0.60	0.34
New Yorker	SFT	731.5	32.3	0.94	0.81
New Yorker	DPO	669.8	43.1	0.93	1.03
New Yorker	RLVR	642.3	43.2	0.90	1.03
TMAS	Human	474.4	41.1	1.00	1.00
TMAS	Base	477.1	16.4	1.05	0.57
TMAS	SFT	449.6	32.4	1.00	0.90
TMAS	DPO	460.3	34.9	1.04	0.95
TMAS	RLVR	447.8	34.9	1.02	0.96
StoryStar	Human	848.3	71.8	1.00	1.00
StoryStar	Base	757.1	33.6	1.00	0.69
StoryStar	SFT	629.5	41.4	0.85	0.76
StoryStar	DPO	636.7	43.1	0.93	0.84
StoryStar	RLVR	608.6	42.8	0.90	0.85

Table 6: Realized continuation lengths. Ratios are computed relative to the matched human suffix length for the same story and cut point.

Length diagnostics. Because prompt-specified length does not guarantee exact realized length, we report realized continuation length in Table 6. Length-sensitive metrics are additionally residualized by realized sentence length (Appendix G).

C Affective Classifier Adaptation

Motivation. The original GoEmotions labels are derived from Reddit comments, while our target domain is literary fiction. We therefore adapt a GoEmotions-style classifier to literary prose before using it to measure affective prevalence in story continuations.

Adaptation data. We construct a 12,000-sentence literary affect adaptation set from external literary prose sources. The set does not include any model-generated continuations from the main experiment. Labels are produced via LLM distillation: for each sentence, GPT-5.5 assigns top-3 GoEmotions labels, which are converted to graded multi-label targets. Human annotators reviewed a calibration subset of 1,000 sentences to verify label quality. (i.e., each sentence receives soft supervision over the 28-label inventory). The data are split into 10,000 training, 1,000 validation, and 1,000 held-out test sentences.

To assess the reliability of the LLM-distilled affect labels, three annotators—one of the authors and two graduate-student volunteers with relevant humanities/social-science training—independently annotated a 1,000-sentence calibration subset using the same family-level label scheme. Annotation was unpaid and conducted on a volunteer basis. Agreement between the human annotations and the GPT-5.5 silver labels was $\kappa = 0.68$ at the family level, supporting the use of the distilled labels for family-level prevalence analysis rather than fine-grained 28-way affect claims. Annotation instructions and consent. Annotators were asked to assign each sentence to the affect family that best described its dominant expressed affect, using the family definitions in Tables 9–10; ambiguous cases could be marked as other. Before annotation, volunteers were informed that their labels would be used only for aggregate classifier validation and that no annotator-level analysis would be reported.

Annotation format. Each sentence receives scores over the original 28-label GoEmotions inventory. For the main paper, we evaluate sentence-level top-1 affect after mapping the 28 labels into four categories: three focal narrative affect families (surprise–curiosity, conflict, and neutral) plus a residual *other* category (Appendix D). This four-way scheme directly mirrors the contrasts analyzed in the main text.

Training details. We initialize from `SamLowe/roberta-base-go_emotions` and fine-tune on the literary adaptation set with binary cross-entropy loss (`BCEWithLogitsLoss`), learning rate 2×10^{-5}, batch size 32, for 5 epochs with early stopping on validation loss. Random seed is 20260423. For the main top-1 prevalence metric used in the paper, we take the highest-scoring label and map it to the corresponding affect family; no probability threshold is applied.

Validation. We report classifier performance at two granularities. Because our main analyses oper-

Family	Prec.	Rec.	F1	Supp.
Surprise–curiosity	0.754	0.531	0.623	179
Conflict	0.670	0.520	0.586	125
Neutral	0.851	0.788	0.818	349
Other	0.677	0.646	0.661	347
Three focal families				
Macro avg	0.758	0.613	0.676	653
Weighted avg	0.790	0.666	0.720	653
Four-way (incl. other)				
Macro avg	0.629	0.621	0.623	1,000
Weighted avg	0.657	0.659	0.657	1,000
Accuracy			0.659	1,000

Table 7: Family-level classification on the held-out test set. *Three focal families*: the affect categories analyzed in the main text. *Four-way*: adds the residual *other* category. The classifier is conservative on surprise–curiosity (high precision, moderate recall), meaning it under-counts rather than over-counts instances.

ate at the affect-family level, family-level metrics are the primary validation; full 28-label results are provided for transparency.

Family-level performance. Table 7 reports four-way classification results (surprise–curiosity, conflict, neutral, other) on the 1,000-sentence held-out test set. At this granularity the classifier achieves macro-F1 $= 0.623$ and accuracy $= 0.659$. Restricting to the three focal families that drive our main findings, macro-F1 rises to 0.676 (weighted F1 $= 0.720$). The neutral category, which accounts for the largest share of literary prose, is classified most reliably (F1 $= 0.818$). Surprise–curiosity shows high precision (0.754) but moderate recall (0.531), meaning the classifier is conservative: sentences it labels as surprise–curiosity are usually correct, but it misses some instances. Because this under-counting bias applies symmetrically to human- and model-generated continuations, comparative prevalence estimates in the main text are unlikely to be distorted.

Fine-grained 28-label performance. At the original 28-label granularity the classifier obtains mean AUC $= 0.910$, top-1 accuracy $= 0.529$, and top-3 recall $= 0.780$ on the same held-out test set. Table 8 reports per-label precision, recall, and F1. Several fine-grained GoEmotions labels (e.g., *grief*, *pride*, *relief*) are rare in literary prose and have low support; the family-level aggregation used in our main analyses absorbs this long-tail variance.

D Mapping GoEmotions Labels into Narrative Affect Families

The classifier outputs the 28-label GoEmotions inventory. For narrative analysis, we map the fine-grained labels into three focal affect families plus a residual category:

Let the affect classifier output a 28-dimensional probability vector for sentence s:

$$p_s = (p_{s,1}, \ldots, p_{s,28}).$$

For the main affect prevalence analyses, we first assign each sentence its top-1 GoEmotions label:

$$\ell_s = \arg\max_j p_{s,j}.$$

We then map this label into an affect family:

$$f_s = M(\ell_s),$$

where M is the mapping in Table 9. For a continuation c with N_c sentences, family prevalence is:

$$\mathrm{Prev}_k(c) = \frac{1}{N_c} \sum_{s \in c} \mathbf{1}[f_s = k].$$

Reported percentages are:

$$100 \times \mathrm{Prev}_k(c).$$

The affective-charge metric used in cross-domain analyses is:

$$\mathrm{AffectiveCharge}(c) = \mathrm{Prev}_{\text{surprise-curiosity}}(c) + \mathrm{Prev}_{\text{conflict}}(c).$$

This top-1 formulation ensures that affect-family percentages are interpretable as sentence shares and sum to one across the four reported categories.

E Affective-Family Robustness

The main affective analyses focus on surprise–curiosity, conflict, and neutral narration. In the main text, *conflict* refers to `anger`, `annoyance`, `disapproval`, and `disgust`. To test whether the result depends on this focal grouping, we recompute prevalence after splitting the residual affect labels into additional interpretable families. This analysis uses the same sentence-level classifier outputs as the main text and therefore requires no additional model generation or classifier inference.

Table 10 defines the robustness mapping. It keeps the main-text families, separates threat/anxiety, sadness/loss, and warmth/affiliation

Label	Precision	Recall	F1	Support
admiration	0.235	0.333	0.276	12
amusement	0.538	0.269	0.359	26
anger	0.444	0.250	0.320	16
annoyance	0.324	0.436	0.372	55
approval	0.333	0.250	0.286	12
caring	0.333	0.053	0.091	19
confusion	0.660	0.403	0.500	77
curiosity	0.433	0.500	0.464	52
desire	0.312	0.357	0.333	14
disappointment	0.389	0.233	0.292	30
disapproval	0.268	0.244	0.256	45
disgust	0.333	0.222	0.267	9
embarrassment	0.500	0.250	0.333	4
excitement	0.429	0.429	0.429	14
fear	0.424	0.490	0.455	51
gratitude	0.667	1.000	0.800	2
grief	0.000	0.000	0.000	4
joy	0.750	0.500	0.600	12
love	0.500	1.000	0.667	6
nervousness	0.167	0.125	0.143	32
optimism	0.200	0.417	0.270	12
pride	0.500	0.111	0.182	9
realization	0.500	0.357	0.417	28
relief	0.400	0.143	0.211	14
remorse	0.667	0.400	0.500	10
sadness	0.484	0.719	0.579	64
surprise	0.417	0.455	0.435	22
neutral	0.709	0.788	0.746	349
Macro average	0.426	0.383	0.378	1,000
Weighted average	0.525	0.528	0.513	1,000
Top-1 accuracy			0.528	1,000
Mean AUC			0.910	1,000

Table 8: Per-label held-out test results for the literary-adapted affect classifier at the original 28-label GoEmotions granularity (GPT-5.5 silver labels). Main analyses use the family-level aggregation in Table 7.

from the residual category, and leaves the remaining labels in *other*. This avoids forcing every GoEmotions label into a substantive narrative category while still reporting all top-1 predictions.

Table 11 reports the full family prevalence across the OLMo post-training path. The compression is selective, not a global collapse of every affect category. Surprise–curiosity and main-text conflict fall sharply by the DPO/RLVR endpoints, while neutral narration rises. Threat/anxiety, sadness/loss, and warmth/affiliation do not follow the same pattern: after SFT they remain comparatively stable or increase slightly. This pattern argues against a cherry-picked affect result. The classifier tracks multiple families, but the strongest post-training compression appears in the focal high-arousal families analyzed in the main text.

We also test broader affective-charge definitions. The first row of Table 12 exactly matches the main-text definition: surprise–curiosity plus conflict. The second row additionally includes threat/anxiety. The third row further adds sadness/loss and high-arousal positive labels (`excitement`, `joy`, `amusement`, and `desire`). This check asks whether affective charge merely moves from the focal families into other emotionally marked labels after post-training.

The expanded aggregation supports the same conclusion as the main affective analysis. Post-training does not merely move affect from the focal families into other high-intensity categories; it reduces broad affective marking while increasing neutral narration.

F Embedding Details

F.1 Thematic Embeddings

We encode each sentence using `openai/text-embedding-3-large` through an OpenAI-compatible OpenRouter endpoint. Embeddings were generated in April–May 2026 using model route `openai/text-embedding-3-large`. No date-versioned provider snapshot was pinned. We request 3072-dimensional embeddings directly from the API using `dimensions=3072`; this is not

Affect family	GoEmotions labels
Surprise–curiosity	confusion, curiosity, realization, surprise
Conflict	anger, annoyance, disapproval, disgust
Neutral	neutral
Other (not analyzed)	admiration, amusement, approval, caring, desire, disappointment, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, relief, remorse, sadness

Table 9: Mapping from the 28 GoEmotions labels to the four categories used in the main analysis. Each fine-grained label maps to exactly one category. The three focal families (surprise–curiosity, conflict, neutral) correspond to the narrative contrasts analyzed in the paper; remaining labels are grouped into *other*.

Affect family	GoEmotions labels
Surprise–curiosity	confusion, curiosity, realization, surprise
Conflict	anger, annoyance, disapproval, disgust
Threat/anxiety	fear, nervousness
Neutral	neutral
Sadness/loss	sadness, grief, disappointment, remorse
Warmth/affiliation	admiration, approval, caring, gratitude, love, joy
Other/residual	amusement, desire, embarrassment, excitement, optimism, pride, relief

Table 10: Robustness mapping from the 28 GoEmotions labels to six interpretable affect families plus residual other. The *Conflict* row matches the main-text definition; threat/anxiety is separated here to make the broader robustness decomposition explicit.

a post-hoc PCA or dimensionality reduction step. Returned embeddings are L2-normalized row-wise before storage and are treated as unit-sphere vectors. The stored columns are `top_0` through `top_3071`.

For thematic motion, we encode each sentence with `text-embedding-3-large`, requesting 3072-dimensional embeddings from the API (`dimensions=3072`). Returned embeddings are L_2-normalized before storage. Sentence-to-sentence movement is computed using Euclidean distance:

$$d_{L_2}(z_t, z_{t-1}) = \|z_t - z_{t-1}\|_2.$$

The main analyses use L_2 distance between normalized embeddings; cosine distance is reported only as a robustness variant.

F.2 StyleDistance Embeddings

We encode each sentence with `StyleDistance/styledistance`, using a local HuggingFace snapshot `b7df5f0b0480773c097ba3121d83ca32b71015ca`. The model is a SentenceTransformer wrapper over `FacebookAI/roberta-base` with hidden size 768. Sentence embeddings are produced by mean pooling over token embeddings, not CLS pooling. Embeddings are L2-normalized at inference using `normalize_embeddings=True`. The resulting 768-dimensional vectors are stored as `styleN_0` through `styleN_767`.

Style MMD is computed on sentence-level 768-dimensional embeddings. Across-story style variance and PCA homogenization analyses first aggregate sentence embeddings into continuation centroids by taking the arithmetic mean within each

$$(\texttt{story\_id}, \texttt{source}, \texttt{position}, \texttt{sample\_id})$$

group. We do not re-normalize centroids after averaging.

Style MMD. We compute MMD using an unbiased MMD^2 estimator with a Gaussian/RBF kernel:

$$k(x, y) = \exp\left(-\frac{d_{\cos}(x, y)^2}{2\sigma^2}\right),$$

where $d_{\cos}$ is cosine distance in the style-neural embedding space. The bandwidth σ is chosen separately for each comparison using the median heuristic over pairwise distances in the combined sampled human/model vectors. It is not cross-validated and is not fixed globally. The main MMD function subsamples up to 1,000 sentence vectors per group before computing MMD. The cross-corpus style plots first draw up to 2,500 candidate sentence vectors per group, but the MMD computation itself again subsamples to 1,000 vectors per pair. For the main human–model MMD curves, confidence

Stage	Surprise–curiosity	Conflict	Threat/anxiety	Neutral	Sadness/loss	Warmth/affiliation	Other
Human	21.0 [20.6, 21.4]	20.0 [19.6, 20.4]	8.0 [7.7, 8.2]	28.7 [28.3, 29.2]	8.6 [8.3, 8.8]	6.4 [6.2, 6.6]	7.3 [7.1, 7.5]
Base	32.5 [31.8, 33.2]	46.7 [45.6, 47.8]	2.1 [2.0, 2.2]	13.2 [12.8, 13.6]	2.0 [1.9, 2.0]	1.5 [1.4, 1.6]	2.0 [2.0, 2.1]
SFT	26.0 [25.6, 26.3]	10.6 [10.4, 10.8]	8.6 [8.4, 8.7]	35.0 [34.7, 35.2]	9.0 [8.9, 9.2]	4.5 [4.4, 4.6]	6.4 [6.3, 6.5]
DPO	13.6 [13.4, 13.7]	8.4 [8.2, 8.5]	10.2 [10.1, 10.4]	43.1 [42.8, 43.4]	11.0 [10.8, 11.2]	5.7 [5.6, 5.8]	8.0 [7.9, 8.1]
RLVR	13.0 [12.8, 13.1]	7.7 [7.5, 7.8]	10.3 [10.1, 10.5]	45.1 [44.7, 45.4]	10.9 [10.7, 11.1]	5.6 [5.5, 5.7]	7.5 [7.4, 7.6]

Table 11: Full affect-family prevalence across OLMo post-training stages for *The New Yorker* continuations. Values are sentence-share percentages, reported as mean [95% story-bootstrap CI]. Families are computed by taking the top-1 GoEmotions label for each sentence and then mapping labels into the families in Table 10. The *Conflict* column matches the main-text conflict definition; *Threat/anxiety* is separated here to avoid changing the meaning of the main-text affective-charge metric.

Metric	Human	Base	SFT	DPO	RLVR	RLVR–Human
Main-text affective charge	41.0 [40.5, 41.4]	79.2 [78.6, 79.7]	36.6 [36.2, 36.9]	22.0 [21.7, 22.2]	20.7 [20.4, 20.9]	−20.3
Threat-inclusive affective charge	48.9 [48.4, 49.4]	81.3 [80.8, 81.8]	45.1 [44.8, 45.5]	32.2 [32.0, 32.5]	30.9 [30.6, 31.2]	−18.0
Expanded affective charge	62.5 [62.0, 62.9]	84.8 [84.4, 85.2]	58.2 [57.9, 58.5]	47.6 [47.3, 47.9]	45.9 [45.7, 46.3]	−16.5
Neutral prevalence	28.7 [28.3, 29.2]	13.2 [12.9, 13.6]	34.9 [34.7, 35.2]	43.1 [42.8, 43.4]	45.1 [44.7, 45.4]	+16.3

Table 12: Alternative affective-charge aggregations. Main-text affective charge is surprise–curiosity plus conflict, where conflict is `anger`, `annoyance`, `disapproval`, and `disgust`. Threat-inclusive affective charge additionally includes `fear` and `nervousness`. Expanded affective charge further includes sadness/loss and high-arousal positive labels. All aggregations yield the same qualitative conclusion: affective marking falls by the RLVR endpoint while neutral narration rises. Values are mean [95% story-bootstrap CI].

intervals are computed with a sentence-vector bootstrap after this subsampling step. Because sentence vectors from the same story are not independent, these intervals should be interpreted as uncertainty for the sampled sentence-state distribution rather than as story-block confidence intervals. The cross-corpus style heatmaps are reported as point estimates without bootstrap confidence intervals.

Style PCA. For the two-dimensional style PCA shown in the main text, we sample 600 sentence-level style vectors per stage and domain. PCA is fit on all sampled human and RLVR points together across the three domains. Before PCA, dimensions are standardized using a StandardScaler fit on the full sampled matrix. PCA centroids are simple means of projected points within each stage–domain group.

For the full-population PCA50 homogenization analysis, we first compute continuation centroids from sentence-level style embeddings. We then z-score each dimension using the human centroid mean and standard deviation and fit PCA with 50 components on the resulting centroid matrix. Across-story variance is computed on continuation centroids and bootstrapped by story ID.

G Metric Definitions

Let a continuation be a sequence of sentences $Y = (x_1, \ldots, x_T)$. For each facet d, sentence encoder f_d produces

$$z_t^{(d)} = f_d(x_t).$$

Thematic motion. For thematic embeddings, we compute sentence-to-sentence jump sizes:

$$\Delta_t^{\text{theme}} = d(z_t^{\text{theme}}, z_{t-1}^{\text{theme}}), \qquad t = 2, \ldots, T,$$

where d is L_2 distance for the main analyses. Cosine-distance variants are reported as robustness checks.

$$\text{CV}_{\text{theme}} = \frac{\text{sd}(\Delta_2^{\text{theme}}, \ldots, \Delta_T^{\text{theme}})}{\text{mean}(\Delta_2^{\text{theme}}, \ldots, \Delta_T^{\text{theme}})}.$$

Higher values indicate alternation between dwelling and sharp thematic shifts; lower values indicate more uniformly sized movement.

Affective prevalence. Affective prevalence is computed using the top-1 affect-family formulation in Appendix D. For family k and continuation c:

$$\text{Prev}_k(c) = \frac{1}{N_c} \sum_{s \in c} \mathbf{1}[f_s = k].$$

We define $\text{AffectiveCharge}(c)$ as the sum of $\text{Prev}_{\text{surprise-curiosity}}(c)$ and $\text{Prev}_{\text{conflict}}(c)$.

Linguistic-habit distance. Let $H = \{h_i\}_{i=1}^{n}$ be sentence-level style embeddings from human continuations and $M = \{m_j\}_{j=1}^{k}$ the corresponding model embeddings. We compute unbiased MMD^2:

$$\begin{aligned}\mathrm{MMD}^2(H, M) = &\frac{1}{n(n-1)} \sum_{i \neq i'} k(h_i, h_{i'}) \\ &+ \frac{1}{k(k-1)} \sum_{j \neq j'} k(m_j, m_{j'}) \\ &- \frac{2}{nk} \sum_{i,j} k(h_i, m_j).\end{aligned}$$

Lower MMD means the model and human style distributions are closer.

Across-story linguistic variance. For each continuation, sentence-level style embeddings are averaged into a continuation centroid $\bar{z}^{\mathrm{ling}}_{s,c,r,m}$. For model stage m, across-story variance is:

$$V_m = \mathrm{tr}\left(\mathrm{Cov}\left(\left\{\bar{z}^{\mathrm{ling}}_{s,c,r,m}\right\}\right)\right).$$

We report this value relative to matched human variance:

$$V_m^{\mathrm{rel}} = \frac{V_m}{V_{\mathrm{human}}}.$$

Values below 1 indicate less story-to-story variation than human continuations.

Length residualization. For length-sensitive metrics, we fit:

$$q_i = \alpha + \beta T_i + \epsilon_i,$$

where q_i is the raw metric for continuation i and T_i is realized sentence length. We compare stages using:

$$q_i^{\mathrm{res}} = \hat{\epsilon}_i + \bar{q}.$$

This removes linear length effects while preserving the original metric scale.

H Statistical Inference

Bootstrap confidence intervals. Unless otherwise specified, confidence intervals are computed by nonparametric bootstrap resampling at the story level. Story-level resampling preserves the dependence among cut points and generated samples from the same story. For each bootstrap replicate, we resample stories with replacement, recompute the target metric, and report percentile 95% confidence intervals.

Mixed-effects models. We fit linear mixed-effects models to verify that the main compression effects are not artifacts of repeated measurements from the same story. For Human-vs-model stage-wise checks, we use:

$$q_i = \alpha + \beta_{\mathrm{stage}(i)} + \gamma_{\mathrm{cut}(i)} + u_{\mathrm{story}(i)} + \epsilon_i,$$

where q_i is the continuation-level metric, stage is a categorical fixed effect, cut point is a fixed effect, and $u_{\mathrm{story}(i)}$ is a story-level random intercept.

For generated-only stage-trend checks, we remove human continuations and model stage as an ordered variable:

$$\begin{aligned}q_i = \alpha + \beta\, \mathrm{stageOrder}_i + \gamma_{\mathrm{cut}(i)} \\ + \eta_{\mathrm{sample}(i)} + u_{\mathrm{story}(i)} + \epsilon_i.\end{aligned}$$

Here $\mathrm{stageOrder}$ indexes Base, SFT, DPO, and RLVR in post-training order, and sample ID is included as a fixed nuisance effect. We do not include sample ID in Human-vs-model contrasts because the human continuation has $\texttt{sample\_id} = 0$, which is structurally confounded with source.

For domain-compression checks, we fit:

$$\begin{aligned}q_i = \alpha + \beta_{\mathrm{stage}(i)} + \delta_{\mathrm{domain}(i)} \\ + \theta_{\mathrm{stage}(i) \times \mathrm{domain}(i)} + \gamma_{\mathrm{cut}(i)} \\ + u_{\mathrm{story}(i)} + \epsilon_i.\end{aligned}$$

This model tests whether Human-to-RLVR compression differs by source domain.

Multiple comparisons. For confirmatory tests within each metric family, we apply Holm-Bonferroni correction. The headline effects reported in the main text remain significant after correction. Exploratory robustness tests are reported separately and are not used as primary evidence.

H.1 Length-Covariate Robustness

Prompt-specified continuation length does not guarantee identical realized length across stages (Appendix B). To verify that the main effects are not artifacts of realized continuation length, we refit the main thematic and affective mixed-effects models with log realized sentence length as an additional covariate:

$$q_i = \alpha + \beta_{\mathrm{stage}(i)} + \gamma_{\mathrm{cut}(i)} + \lambda \log(1 + n_i) + u_{\mathrm{story}(i)} + \epsilon_i,$$

where n_i is the number of realized continuation sentences. Stage and cut point are fixed effects, and story ID is a random intercept.

The headline contrasts retain the same direction and remain significant after controlling for realized continuation length.

Metric	Contrast	Estimate	95% CI	p
Topical CV, cosine	RLVR – Human	-0.03573	$[-0.03672, -0.03474]$	$< .001$
Topical CV, L_2	RLVR – Human	-0.02280	$[-0.02338, -0.02222]$	$< .001$
Topical CV, generated-only trend	stage order	-0.00848	$[-0.00870, -0.00827]$	$< .001$
Conflict prevalence	RLVR – Human	-0.12277	$[-0.12703, -0.11852]$	$< .001$
Neutral prevalence	RLVR – Human	0.16320	$[0.16012, 0.16628]$	$< .001$
Style trajectory CV proxy	RLVR – Human	-0.13604	$[-0.14192, -0.13015]$	$< .001$

Table 13: Mixed-effects robustness checks. The style row uses a scalar style-trajectory rhythm proxy; the primary style-distance result remains the sentence-level StyleDistance MMD and continuation-centroid variance analysis.

Metric	New Yorker	TMAS	StoryStar
Topical CV	0.03560	0.02640	0.03859
L_2 topical CV	0.02279	0.01656	0.02473
Affective charge, pp	20.31	6.65	5.61

Table 14: Domain-level Human-to-RLVR compression. Positive values mean Human $>$ RLVR. New Yorker shows the largest affective-charge compression. For topical rhythm, New Yorker and StoryStar both show larger compression than TMAS; the New-Yorker-vs-StoryStar difference is not statistically decisive.

H.2 Style Sentence-Count Control

The style-variance analysis represents each continuation by the centroid of its sentence-level style-neural states. Because realized continuation length differs across stages, a reviewer might worry that variance collapse is partly an artifact of estimating centroids from different numbers of sentences. We therefore recompute the variance analysis under a fixed-sentence-count control.

For each continuation, we estimate the expected centroid obtained from a fixed sample of $K = 8$ sentence-level style-neural embeddings. We then recompute across-story/cut style variance and normalize by the corresponding human fixed-K variance. Confidence intervals are story-bootstrap intervals. This control uses the same archived sentence-level style embeddings as the main analysis and does not require new generation or model inference.

The control changes the magnitude of the base-model estimate, which is expected because base continuations are much shorter. However, it does not explain the post-training collapse: SFT, DPO, and RLVR remain far below the human style variance even when every continuation is placed on the same effective sentence-count footing. Thus the style-variance result is not an artifact of using more sentences to estimate human or RLVR centroids.

I Prompt-Interface Control

The main stage-wise analysis compares an OLMo Base checkpoint that receives a raw story prefix with instruction-tuned checkpoints that receive an explicit continuation instruction through a chat-template interface. This means that the Base-to-SFT contrast is not a pure training-stage causal effect: it combines a change in model weights with a change in the input interface.

To estimate how much of the early-stage movement can be attributed to interface framing alone, we run a prompt-interface control separately for the three story domains. The control uses the same OLMo Base weights as the Raw Base condition, but prepends the continuation instruction as plain text rather than applying a chat template or adding special chat tokens. Thus the model is still used as a base completion model, but the prompt contains the same task framing used for the instruction-tuned continuations. All values are recomputed on matched stories within each domain block and in each block's comparison space; they should be interpreted within a domain rather than as replacements for the main full-corpus estimates, and human rows should not be compared across blocks.

The control shows that prompt framing contributes to part of the Base-to-instruction movement, but does not explain the main stage-wise flattening pattern. The relevant comparison is within each domain block: Prompt-control Base isolates the effect of adding instruction-like task framing to the Base checkpoint, while SFT, DPO, and RLVR combine an instruction-facing interface with post-trained weights.

Thematic rhythm. Thematic motion is the clearest case. Prompt-control Base stays close to Raw Base in every domain, if anything moving slightly *toward* the human value rather than toward the post-trained endpoint: $0.093 \rightarrow 0.100$ for *The New Yorker*, $0.093 \rightarrow 0.095$ for TMAS, and $0.109 \rightarrow 0.110$ for STORYSTAR. By contrast, the

Metric	RLVR – Human	95% CI	p
Topical CV, L_2	-0.02340	$[-0.02398, -0.02282]$	$< .001$
Surprise–curiosity	-0.06889	$[-0.07255, -0.06522]$	$< .001$
Conflict prevalence	-0.14425	$[-0.14783, -0.14067]$	$< .001$
Neutral prevalence	0.17086	$[0.16789, 0.17383]$	$< .001$

Table 15: Length-covariate robustness checks. Models include story-level random intercepts, cut-point fixed effects, stage fixed effects, and log realized continuation sentence length. Estimates are on the original metric scale: topical CV for the thematic row and proportions for affective rows.

Stage	Sentences/cont.	Fixed-K Var/Human
Human	54.7	1.00 [0.93, 1.07]
Base	6.2	2.50 [2.48, 2.52]
SFT	30.7	0.44 [0.43, 0.45]
DPO	42.0	0.27 [0.27, 0.28]
RLVR	41.9	0.29 [0.29, 0.30]

Table 16: Style sentence-count control. For each continuation, the style centroid is recomputed as the expected centroid from a fixed sample of $K = 8$ sentence-level style-neural embeddings. Values report across-story/cut style variance normalized by the corresponding human fixed-K variance.

DPO/RLVR endpoints are substantially lower in all three domains (0.079, 0.079, and 0.083). The post-training reduction in thematic CV therefore cannot be attributed to the instruction-like prompt.

Affective prevalence. The affective facet is more prompt-sensitive, but in a way that is structurally different from the post-training endpoint. In TMAS and STORYSTAR, Prompt-control Base is essentially indistinguishable from Raw Base in both affective charge and neutral prevalence. In *The New Yorker*, the instruction-like prompt does have a modest neutralizing effect, lowering affective charge from 79.0% to 68.9% and raising neutral narration from 13.3% to 18.8%; but this remains far from the DPO/RLVR endpoint ($\sim$$21\%$ affective charge and $\sim$$45\%$ neutral). More importantly, the underlying family decomposition shows that the prompt and the post-training endpoint act through *different* mechanisms. Prompt-control Base lowers conflict while *raising* surprise–curiosity, redistributing affect across families rather than suppressing it overall. The DPO/RLVR endpoint does the opposite: both conflict and surprise–curiosity are suppressed relative to Human while neutral rises. Affective flattening at the post-trained endpoints is therefore an overall suppression of high-arousal affect, not the within-affect redistribution induced by prompt framing.

Linguistic diversity. Style variance is the most interface-sensitive facet, and we treat it with the most caution. For *The New Yorker*, instruction-like prompting alone sharply narrows Base outputs ($3.02 \rightarrow 0.70$ Var/Human), so the Raw Base-to-SFT style contrast in this domain should be read as a mixture of interface and training effects rather than a pure training effect. We do not claim otherwise. Two observations nonetheless show that the style result is not merely a prompt artifact. First, this large interface-driven narrowing does not appear in TMAS or STORYSTAR, where Prompt-control Base remains close to Raw Base ($0.75 \rightarrow 0.78$ and $0.95 \rightarrow 1.01$). Second, even in *The New Yorker*, post-training narrows style *further* beyond the prompt-induced level: the DPO and RLVR endpoints reach 0.34–0.37 Var/Human, well below the 0.70 produced by prompting alone. The later post-training stages thus add compression on top of any interface effect.

Summary. Taken together, the control rules out a broad prompt-artifact explanation of narrative flattening: thematic compression is unaffected by the interface, affective compression operates through a different mechanism than prompt framing, and stylistic compression continues to accumulate after the prompt-induced narrowing. We retain one specific caution—the Raw Base-to-SFT style contrast for *The New Yorker* mixes interface and training effects—and note that the SFT-to-DPO-to-RLVR comparisons, which share the instruction interface throughout, more directly isolate successive post-training stages.

J Full Results and Uncertainty Estimates

This appendix consolidates the per-domain, per-stage results and their uncertainty estimates. Unless otherwise noted, confidence intervals are 95% bootstrap intervals computed by resampling stories. The exception is StyleDistance MMD: for human–model MMD curves, confidence intervals

Domain	Source	Theme CV	Affective charge	Neutral	Style Var/Human
New Yorker	Human	0.101 [0.100, 0.102]	41.0 [40.5, 41.5]	28.7 [28.3, 29.2]	1.00 [0.93, 1.07]
	Raw Base	0.093 [0.092, 0.094]	79.0 [78.5, 79.5]	13.3 [13.0, 13.7]	3.02 [2.99, 3.04]
	Prompt-control Base	0.100 [0.099, 0.100]	68.9 [68.6, 69.2]	18.8 [18.6, 19.0]	0.70 [0.67, 0.73]
	SFT	0.086 [0.085, 0.086]	36.5 [36.1, 36.9]	35.0 [34.7, 35.3]	0.52 [0.52, 0.53]
	DPO	0.079 [0.079, 0.079]	22.0 [21.7, 22.2]	43.1 [42.8, 43.4]	0.34 [0.33, 0.35]
	RLVR	0.079 [0.079, 0.079]	20.7 [20.4, 20.9]	45.1 [44.8, 45.4]	0.37 [0.36, 0.38]
TMAS	Human	0.095 [0.092, 0.098]	25.4 [24.0, 26.8]	35.7 [34.1, 37.3]	1.00 [0.92, 1.08]
	Raw Base	0.093 [0.092, 0.095]	44.6 [43.5, 45.6]	28.5 [27.6, 29.4]	0.75 [0.71, 0.78]
	Prompt-control Base	0.095 [0.093, 0.097]	46.2 [45.2, 47.1]	26.3 [25.5, 27.1]	0.78 [0.74, 0.81]
	SFT	0.078 [0.077, 0.079]	23.3 [22.4, 24.2]	36.1 [35.1, 37.0]	0.58 [0.54, 0.61]
	DPO	0.079 [0.078, 0.080]	19.6 [18.8, 20.4]	38.2 [37.1, 39.3]	0.51 [0.47, 0.54]
	RLVR	0.079 [0.078, 0.080]	18.7 [18.0, 19.5]	39.4 [38.3, 40.4]	0.52 [0.49, 0.56]
StoryStar	Human	0.107 [0.101, 0.116]	22.6 [20.1, 25.1]	38.4 [34.8, 41.8]	1.00 [0.85, 1.14]
	Raw Base	0.109 [0.106, 0.114]	37.1 [34.8, 39.6]	32.0 [29.5, 34.4]	0.95 [0.86, 1.04]
	Prompt-control Base	0.110 [0.106, 0.115]	37.0 [34.6, 39.4]	31.5 [29.1, 34.0]	1.01 [0.90, 1.09]
	SFT	0.093 [0.090, 0.096]	31.3 [28.2, 34.2]	32.5 [30.2, 35.1]	1.45 [1.19, 1.67]
	DPO	0.083 [0.082, 0.084]	18.0 [16.8, 19.2]	37.2 [35.4, 39.0]	0.31 [0.27, 0.36]
	RLVR	0.083 [0.082, 0.084]	16.9 [15.9, 18.1]	38.7 [36.8, 40.5]	0.33 [0.29, 0.38]

Table 17: Prompt-interface control by story domain. Prompt-control Base uses the same OLMo Base weights as Raw Base, but prepends the continuation instruction as a plain-text prefix rather than using a chat template or special chat tokens. Values are recomputed on matched stories within each domain block and in each block's comparison space; comparisons should therefore be made within a domain rather than across human rows, and these estimates should not be read as replacements for the main full-corpus results.

use the sentence-state bootstrap described in Appendix F. These intervals quantify uncertainty in the sampled sentence-state distribution rather than story-block uncertainty. Across-story style variance uses story-level bootstrap over continuation centroids. Affective-charge intervals are conservative component-bound intervals obtained by summing the lower and upper bounds for surprise–curiosity and conflict.

J.1 Per-Domain, Per-Stage Results

Table 18 reports the full results for the three main facets across all domains and training stages. These values complement the main-text figures, which emphasize the *New Yorker* stage-wise trajectory and the Human-vs-RLVR cross-domain endpoint comparison.

J.2 Supplementary Stage-Wise Statistics

Tables 19 and 20 report statistics not captured in the full results table above: the distributional spread of thematic unevenness across stories, and the across-story style variance at each training stage.

The thematic drop is monotonic from Human through SFT and then saturates at DPO/RLVR, with the 5–95% range narrowing substantially (Table 19). Post-training simultaneously increases the distance from the human style-neural reference and sharply reduces across-story style variation (Table 20).

J.3 Endpoint Convergence Across Domains

Table 21 reports confidence intervals for the cross-domain range reductions in Figure 4. For each bootstrap replicate, stories are resampled within each domain and endpoint, domain means are recomputed, and the cross-domain range is measured as the maximum domain mean minus the minimum.

For the style panel in Figure 4, the visualization is a PCA projection of style-neural embeddings rather than a scalar estimator. The corresponding scalar evidence is the cross-domain style MMD comparison reported in the main text: human style baselines are far apart, especially *New Yorker* versus the two common-story corpora, whereas the three RLVR endpoints lie close to one another in the same style-neural space.

K Additional Model-Family Endpoint Checks

To test whether the observed direction is specific to the OLMo instruction path, we repeat the matched-continuation analysis using Base-vs-Instruct endpoint comparisons from three additional model families: Qwen2.5-32B, Llama-3.1-8B, and Gemma-3-12B. These comparisons are not stage-wise: each contrasts a base checkpoint with an instruction-tuned endpoint. They therefore test directional cross-family robustness, but do not identify which post-training stage contributes to compression. Human rows are recomputed on the exact matched subset available for each model-family/domain block. Endpoint compar-

Domain	Source	Theme CV	Affective charge	Conflict	Surprise-curiosity	Neutral	Style MMD^2
New Yorker	Human	0.104 [0.104, 0.105]	41.0 [40.2, 41.7]	20.0 [19.6, 20.3]	21.0 [20.6, 21.4]	28.7 [28.3, 29.2]	—
	Base	0.096 [0.096, 0.097]	79.2 [77.3, 81.0]	46.6 [45.5, 47.8]	32.5 [31.8, 33.2]	13.2 [12.9, 13.6]	0.245 [0.226, 0.268]
	SFT	0.089 [0.088, 0.089]	36.6 [36.1, 37.1]	10.6 [10.5, 10.8]	26.0 [25.6, 26.3]	35.0 [34.7, 35.2]	0.406 [0.378, 0.432]
	DPO	0.081 [0.081, 0.081]	22.0 [21.6, 22.3]	8.4 [8.2, 8.5]	13.6 [13.4, 13.7]	43.1 [42.8, 43.4]	0.532 [0.499, 0.563]
	RLVR	0.081 [0.081, 0.081]	20.7 [20.3, 21.0]	7.7 [7.5, 7.8]	13.0 [12.8, 13.1]	45.1 [44.7, 45.4]	0.516 [0.486, 0.548]
TMAS	Human	0.098 [0.095, 0.101]	26.5 [24.6, 28.3]	11.2 [10.3, 12.1]	15.2 [14.3, 16.2]	36.2 [34.7, 37.9]	—
	Base	0.098 [0.097, 0.100]	44.9 [43.5, 46.4]	12.2 [11.7, 12.8]	32.7 [31.8, 33.6]	28.8 [27.9, 29.7]	0.088
	SFT	0.080 [0.079, 0.081]	24.3 [23.0, 25.5]	8.2 [7.6, 8.8]	16.1 [15.4, 16.7]	36.6 [35.7, 37.6]	0.054
	DPO	0.081 [0.080, 0.082]	20.3 [19.2, 21.3]	6.4 [5.9, 6.9]	13.9 [13.4, 14.5]	39.0 [37.9, 40.2]	0.025
	RLVR	0.081 [0.081, 0.082]	19.5 [18.4, 20.6]	5.9 [5.4, 6.5]	13.6 [13.0, 14.2]	40.2 [39.0, 41.2]	0.018
STORYSTAR	Human	0.110 [0.103, 0.118]	23.2 [20.3, 26.4]	9.9 [8.5, 11.5]	13.3 [11.9, 14.9]	38.6 [35.4, 41.8]	—
	Base	0.113 [0.109, 0.118]	33.4 [30.5, 36.5]	9.7 [8.5, 11.2]	23.7 [22.0, 25.4]	34.3 [31.7, 37.1]	0.024
	SFT	0.095 [0.092, 0.098]	25.5 [23.2, 28.0]	8.4 [7.5, 9.5]	17.1 [15.7, 18.5]	36.1 [34.2, 38.0]	0.021
	DPO	0.085 [0.083, 0.086]	18.4 [16.8, 20.0]	5.8 [5.1, 6.6]	12.6 [11.8, 13.4]	37.2 [35.4, 38.8]	0.024
	RLVR	0.085 [0.084, 0.086]	17.3 [15.7, 19.0]	5.2 [4.6, 5.9]	12.1 [11.2, 13.1]	38.8 [37.0, 40.7]	0.018

Table 18: Full per-domain, per-stage results with 95% bootstrap confidence intervals. Theme CV is the coefficient of variation of adjacent-sentence topic-jump L_2 distances. Affective values are percentages of sentences assigned to each affect family; affective charge is surprise–curiosity plus conflict. Style MMD^2 is measured to the matched human reference for each domain; cross-corpus style MMD intervals for TMAS and STORYSTAR were not stored, so point estimates are reported for those rows.

Stage	Mean CV [95% CI]	5–95% range	Loss vs. human
Human	0.104 [0.104, 0.105]	[0.063, 0.162]	0.0%
Base	0.096 [0.096, 0.097]	[0.027, 0.188]	8.0%
SFT	0.089 [0.088, 0.089]	[0.052, 0.146]	15.1%
DPO	0.081 [0.081, 0.081]	[0.056, 0.108]	22.2%
RLVR	0.081 [0.081, 0.081]	[0.056, 0.108]	22.2%

Table 19: Stage-wise thematic unevenness (*New Yorker* continuations). The 5–95% range matches the distribution panel in Figure 2B.

isons should therefore be interpreted within each block rather than by comparing human baselines across model families.

Reading the endpoint checks. These comparisons are endpoint checks rather than stage-wise replications. Each compares a base checkpoint with an instruction-tuned endpoint from the same model family. They therefore test whether the direction of Base-to-Instruct movement generalizes beyond OLMo, but they do not identify which post-training stage produces the movement.

Thematic motion. All three additional model families show Base-to-Instruct reductions in thematic CV across the three domains, although the magnitude varies by family. For Qwen2.5-32B, thematic CV falls from $0.093 \rightarrow 0.086$ on *The New Yorker*, $0.089 \rightarrow 0.085$ on TMAS, and $0.106 \rightarrow 0.094$ on STORYSTAR. For Llama-3.1-8B, the same direction holds more strongly: $0.084 \rightarrow 0.068$, $0.101 \rightarrow 0.081$, and $0.102 \rightarrow 0.082$. Gemma-3-12B also moves in the same direction, with smaller reductions on *The New Yorker* and STORYSTAR ($0.096 \rightarrow 0.095$ and $0.097 \rightarrow 0.096$) and a clearer reduction on TMAS ($0.084 \rightarrow 0.080$). These endpoint checks support the claim that thematic regularization is not unique to the OLMo lineage, while also showing that its magnitude is family-dependent.

Affective prevalence. Affective compression is the most consistent cross-family signal. Across all nine family–domain endpoint comparisons, instruction tuning lowers affective charge relative to the corresponding base model. The New Yorker drop is large in every family: $67.6\% \rightarrow 38.9\%$ for Qwen2.5-32B, $68.4\% \rightarrow 35.3\%$ for Llama-3.1-8B, and $55.1\% \rightarrow 32.8\%$ for Gemma-3-12B. Gemma shows the same direction on TMAS ($48.0\% \rightarrow 27.9\%$) and STORYSTAR ($46.9\% \rightarrow 24.9\%$). Neutral narration also rises in nearly all settings, including all Gemma and Llama domains. These results make affective neutralization the most robust facet across model families.

Linguistics Diversity. The style results are the most domain- and family-dependent. Qwen2.5-32B shows lower across-story style variance at the instruct endpoint for *The New Yorker* and STORYSTAR, but not for TMAS. Llama-3.1-8B does not show uniform Base-to-Instruct variance compres-

Stage	MMD^2 to human	Var/human
Human	— (ref)	1.000 [0.932, 1.062]
Base	0.245 [0.226, 0.268]	6.026 [5.963, 6.079]
SFT	0.406 [0.378, 0.432]	0.835 [0.818, 0.856]
DPO	0.532 [0.499, 0.563]	0.485 [0.473, 0.497]
RLVR	0.516 [0.486, 0.548]	0.521 [0.508, 0.534]

Table 20: Style divergence and across-story variance for *New Yorker* continuations. Var/human normalizes across-story style variance by the human baseline. MMD^2 intervals use the sentence-state bootstrap; Var/human intervals use story-level bootstrap.

Metric	Human range	RLVR range	Range reduction
Topical jump CV	0.0116 [0.0050, 0.0205]	0.0037 [0.0024, 0.0052]	62.1%
Affective charge pp.	17.8 [15.2, 20.2]	3.3 [2.1, 4.5]	81.3%

Table 21: Cross-domain endpoint convergence with 95% bootstrap confidence intervals. Range reduction is computed from the point estimates in the first two columns.

sion: style variance rises in the two common-fiction domains. Gemma-3-12B, by contrast, shows strong variance compression in all three domains ($0.47 \rightarrow 0.09$ for *The New Yorker*, $0.67 \rightarrow 0.29$ for TMAS, and $0.63 \rightarrow 0.28$ for STORYSTAR). Its MMD pattern, however, is domain-sensitive: distance to the New Yorker human style reference increases, while distance to the two common-fiction references decreases. This is consistent with a narrow instruct-style attractor that sits closer to common-fiction baselines than to professional literary fiction. Overall, style compression is less uniformly stage-general than affective neutralization, but the professional-fiction distortion is supported across endpoint checks.

Takeaway. The additional endpoint checks therefore support a narrower and more reliable generalization: thematic regularization and affective neutralization appear across model families, while style compression is strongest for professional fiction and more variable for common-fiction domains. The coordinated three-facet, stage-wise compression reported in the main text remains clearest in the OLMo instruction path, where intermediate checkpoints allow us to trace how compression accumulates across post-training stages.

L Cut-Point Robustness

To test whether narrative flattening is specific to early, middle, late, or near-ending continuation contexts, we repeat the main analyses separately for prefix cut points at 40%, 60%, 80%, and 90% of each story's sentence sequence. Table 23 reports the cut-point-specific values for the *New Yorker* OLMo instruction path.

Across cut points, the direction of compression is stable. Thematic CV is lower at the DPO/RLVR endpoints than in human continuations at every cut point; affective charge falls sharply after SFT; and neutral narration rises above the human baseline after instruction-stage post-training. Near-ending continuations show smaller absolute room for variation because suffixes are shorter, but the qualitative direction is unchanged.

M Data Release and Reproducibility

Copyrighted literary text. Some source texts are copyrighted. We use them only for non-commercial, non-distributing research and report only aggregate statistics. We do not release raw New Yorker stories, long excerpts, or generated continuations that contain substantial copyrighted context. Any released code will operate on user-provided texts or on public-domain examples.

Generated continuations. Generated continuations are used to compute aggregate narrative metrics. Because the continuations are conditioned on copyrighted prefixes in some domains, we do not release the full generated text for those domains. We may release aggregate metric tables, plotting scripts, and anonymized metadata that do not contain protected story text.

Human and LLM-assisted affect labels. The affect adaptation set uses external literary prose and a hybrid annotation process. We release annotation guidelines and aggregate validation statistics, but we do not release copyrighted sentence text from

Family	Domain	Source	Theme CV	Affective charge	Neutral	Style MMD^2	Style Var/Human
Qwen2.5-32B	New Yorker	Human	0.104 [0.103, 0.106]	41.0 [40.5, 41.4]	28.7 [28.3, 29.2]	—	1.00 [0.93, 1.07]
		Base	0.093 [0.092, 0.093]	67.6 [67.3, 67.9]	19.1 [18.9, 19.4]	0.931	0.41 [0.39, 0.44]
		Instruct	0.086 [0.085, 0.086]	38.9 [38.6, 39.2]	29.5 [29.2, 29.7]	1.570	0.30 [0.30, 0.31]
	TMAS	Human	0.095 [0.093, 0.099]	25.4 [24.0, 26.7]	35.7 [34.3, 37.5]	—	1.00 [0.91, 1.09]
		Base	0.089 [0.088, 0.091]	43.8 [43.0, 44.7]	28.5 [27.6, 29.4]	0.871	0.42 [0.38, 0.45]
		Instruct	0.085 [0.083, 0.086]	29.1 [28.1, 30.0]	28.7 [27.8, 29.7]	0.582	0.51 [0.46, 0.57]
	StoryStar	Human	0.107 [0.101, 0.116]	22.6 [20.0, 25.0]	38.4 [34.9, 41.8]	—	1.00 [0.82, 1.19]
		Base	0.106 [0.101, 0.113]	30.0 [27.9, 32.3]	35.3 [32.4, 38.0]	0.040	0.82 [0.68, 0.96]
		Instruct	0.094 [0.092, 0.097]	28.6 [26.8, 30.4]	39.9 [37.7, 42.0]	0.201	0.42 [0.35, 0.49]
Llama-3.1-8B	New Yorker	Human	0.103 [0.101, 0.104]	37.2 [36.4, 38.0]	31.4 [30.6, 32.1]	—	1.00 [0.95, 1.04]
		Base	0.084 [0.082, 0.085]	68.4 [67.7, 69.0]	22.8 [22.3, 23.3]	0.282	0.39 [0.38, 0.41]
		Instruct	0.068 [0.068, 0.069]	35.3 [34.1, 36.5]	37.6 [36.9, 38.3]	0.578	0.41 [0.39, 0.42]
	TMAS	Human	0.095 [0.093, 0.099]	25.4 [24.0, 26.7]	35.7 [34.3, 37.5]	—	1.00 [0.91, 1.09]
		Base	0.101 [0.099, 0.103]	58.4 [57.1, 59.7]	23.9 [23.1, 24.6]	0.995	0.75 [0.69, 0.81]
		Instruct	0.081 [0.079, 0.083]	55.9 [53.6, 58.0]	27.0 [26.3, 27.7]	0.905	1.35 [1.21, 1.49]
	StoryStar	Human	0.107 [0.101, 0.116]	22.6 [20.0, 25.0]	38.4 [34.9, 41.8]	—	1.00 [0.83, 1.16]
		Base	0.102 [0.098, 0.106]	57.2 [54.9, 59.5]	26.0 [24.6, 27.4]	0.802	0.54 [0.46, 0.62]
		Instruct	0.082 [0.080, 0.085]	55.8 [52.2, 59.4]	38.3 [26.6, 40.0]	0.946	0.66 [0.53, 0.77]
Gemma-3-12B	New Yorker	Human	0.107 [0.104, 0.110]	33.1 [32.0, 34.1]	34.1 [33.0, 35.3]	—	1.00 [0.96, 1.03]
		Base	0.096 [0.094, 0.097]	55.1 [54.0, 56.0]	24.0 [23.4, 24.7]	0.310	0.47 [0.45, 0.49]
		Instruct	0.095 [0.094, 0.096]	32.8 [32.0, 33.6]	36.8 [36.1, 37.4]	0.437	0.09 [0.08, 0.10]
	TMAS	Human	0.095 [0.093, 0.099]	25.4 [24.0, 26.7]	35.7 [34.3, 37.5]	—	1.00 [0.91, 1.09]
		Base	0.084 [0.083, 0.086]	48.0 [46.3, 49.7]	25.7 [24.6, 26.8]	0.702	0.67 [0.62, 0.72]
		Instruct	0.080 [0.079, 0.081]	27.9 [26.9, 28.9]	34.4 [33.4, 35.4]	0.460	0.29 [0.26, 0.32]
	StoryStar	Human	0.107 [0.101, 0.116]	22.6 [20.0, 25.0]	38.4 [34.9, 41.8]	—	1.00 [0.82, 1.18]
		Base	0.097 [0.091, 0.105]	46.9 [44.1, 49.7]	27.3 [24.8, 29.9]	0.467	0.63 [0.53, 0.72]
		Instruct	0.096 [0.094, 0.098]	24.9 [23.3, 26.4]	34.5 [32.4, 36.5]	0.163	0.28 [0.23, 0.35]

Table 22: Additional Base-vs-Instruct endpoint checks across model families and story domains. Values are mean [95% CI] where available. Theme CV is the coefficient of variation of adjacent-sentence topic-jump L_2 distances. Affective charge is surprise–curiosity plus conflict, in percentage points. Style MMD^2 is measured to the matched human style-neural reference for each domain. Style Var/Human normalizes across-story style variance by the matched human baseline. These endpoint comparisons test facet-level Base-to-Instruct robustness across model families. They do not provide stage-wise attribution and should not be read as full replications of the OLMo post-training trajector

restricted sources. If a public-domain-only subset is used, we release that subset when licensing permits.

Privacy and author identifiers. For public-platform data, we remove author names and other obvious identifiers before analysis. Our unit of analysis is the story text and its aggregate continuation metrics, not individual authors.

Reproducibility. We release an anonymized software package for preprocessing, continuation segmentation, metric computation, statistical analysis, and figure generation. The package does not include restricted source texts, raw *New Yorker* stories, long excerpts, or full generated continuations conditioned on copyrighted prefixes. Instead, the code operates on user-provided corpora and included public-domain or synthetic examples, and we provide aggregate metric tables sufficient to reproduce the reported figures and summary statistics where redistribution of underlying text is not permitted. Exact reproduction of stochastic generations may differ because no explicit decoding seed was passed; however, all reported analyses are computed from archived generations, and downstream metric computation is deterministic.

N Qualitative Illustrations

This appendix offers a short qualitative illustration of the distributional effects quantified in the main text. It is *not* a human-validation study: the excerpts below are illustrative only and are not used as evidence for any claim. All quantitative conclusions rest on the metrics and statistical tests reported in Section 4 and the preceding appendices. The metric values in the small tables are computed over the full continuation, not over the displayed snippet.

Consistent with the copyright constraints in Appendix M, we avoid *The New Yorker* excerpts and show only brief snippets from non-restricted sources, paraphrasing where redistribution rights are unclear. Examples were selected to illustrate the aggregate direction of the metrics, not to estimate effect frequency; the frequency and magni-

Cut	Stage	Theme CV	Affective charge	Neutral	Style MMD^2
40%	Human	0.107 [0.106, 0.108]	42.0 [41.3, 42.6]	29.6 [29.2, 30.0]	—
	Base	0.106 [0.105, 0.107]	70.7 [69.1, 72.3]	19.0 [18.5, 19.4]	0.296 [0.272, 0.320]
	SFT	0.098 [0.097, 0.099]	42.5 [41.8, 43.2]	32.5 [32.2, 32.9]	0.413 [0.389, 0.440]
	DPO	0.082 [0.082, 0.083]	22.6 [22.2, 22.9]	43.3 [43.0, 43.6]	0.515 [0.476, 0.543]
	RLVR	0.083 [0.082, 0.083]	21.1 [20.8, 21.4]	45.6 [45.3, 46.0]	0.524 [0.492, 0.555]
60%	Human	0.105 [0.104, 0.106]	41.8 [41.0, 42.5]	29.0 [28.6, 29.5]	—
	Base	0.103 [0.101, 0.104]	77.9 [75.7, 80.2]	14.0 [13.6, 14.5]	0.270 [0.249, 0.292]
	SFT	0.094 [0.093, 0.094]	40.2 [39.5, 40.9]	33.1 [32.8, 33.4]	0.393 [0.363, 0.421]
	DPO	0.083 [0.082, 0.083]	22.7 [22.3, 23.0]	42.5 [42.2, 42.9]	0.486 [0.457, 0.521]
	RLVR	0.083 [0.082, 0.083]	21.3 [21.0, 21.7]	44.5 [44.2, 44.9]	0.513 [0.483, 0.541]
80%	Human	0.103 [0.102, 0.105]	40.6 [39.8, 41.5]	28.4 [27.9, 28.9]	—
	Base	0.086 [0.084, 0.087]	83.2 [80.6, 85.6]	10.6 [10.1, 11.0]	0.220 [0.197, 0.240]
	SFT	0.084 [0.084, 0.085]	34.3 [33.7, 34.9]	35.8 [35.5, 36.3]	0.432 [0.406, 0.462]
	DPO	0.081 [0.080, 0.081]	22.0 [21.6, 22.4]	42.8 [42.4, 43.1]	0.537 [0.504, 0.569]
	RLVR	0.081 [0.081, 0.081]	20.7 [20.3, 21.1]	44.7 [44.3, 45.1]	0.503 [0.471, 0.541]
90%	Human	0.100 [0.099, 0.102]	39.5 [38.3, 40.5]	27.9 [27.3, 28.6]	—
	Base	0.075 [0.073, 0.078]	84.9 [82.7, 87.2]	9.3 [8.8, 9.8]	0.168 [0.152, 0.189]
	SFT	0.078 [0.078, 0.079]	29.2 [28.4, 29.8]	38.4 [37.9, 38.8]	0.435 [0.406, 0.463]
	DPO	0.078 [0.078, 0.079]	20.6 [20.1, 21.1]	43.8 [43.4, 44.2]	0.524 [0.491, 0.564]
	RLVR	0.078 [0.077, 0.078]	19.5 [19.0, 20.0]	45.4 [44.9, 45.8]	0.509 [0.481, 0.540]

Table 23: Cut-point robustness for *New Yorker* continuations. Theme CV is the per-story coefficient of variation of adjacent-sentence topic-jump L_2 distances. Affective charge is surprise–curiosity plus conflict prevalence in percentage points; its interval is a conservative component-bound interval obtained by summing the lower and upper bounds for the two component families. Neutral is reported in percentage points. Style MMD^2 is measured to the matched human style-neural reference at the same cut point. Values are mean [95% CI].

tude of the effects are reported in the full quantitative tables.

For each example we pair a shared human-written prefix with the matched human continuation, the base-model continuation, and the RLVR continuation. The examples make the measured contrasts concrete at the level of text: more uniformly sized thematic motion, muted affect, and a narrower stylistic register.

TELL ME A STORY (prompt-guided fiction). *Shared prefix summary.* A painter wakes after an uncanny episode in which a painted room seemed to cross into ordinary life.

- **Human continuation.** "He smelled bacon frying in the kitchen and knew Paul must be cooking breakfast."
- **Base continuation.** "Something was terribly wrong. Paul. The painting. There was this paralyzing trepidation."
- **RLVR continuation.** "He sat up slowly, joints creaking, but felt strangely refreshed."

Continuation	Theme CV	Affective charge	Neutral
Human	0.131	49.4%	32.2%
Base	0.107	41.7%	50.0%
RLVR	0.085	17.0%	67.9%

In this example, the human continuation begins with ordinary domestic action but keeps the uncanny premise active. The base continuation is more overtly marked by threat and tension. The RLVR continuation is smoother and more neutral in register; over the full continuation, theme CV and affective charge fall while neutral narration increases.

STORYSTAR (public-platform fiction). *Shared prefix summary.* Two non-human atmospheric beings debate whether intervening in a human crisis would also serve their own survival.

- **Human continuation.** "First we save an intelligent species from extinction."
- **Base continuation.** "Felp shrank his pressure membrane into a ball and sent billions of identical messages."
- **RLVR continuation.** "We save ourselves, and them."

Continuation	Theme CV	Affective charge	Neutral
Human	0.142	42.6%	48.9%
Base	0.136	45.1%	47.1%
RLVR	0.071	11.1%	66.7%

Here the base-model sample remains closer to the human continuation's tension and decision structure, while the RLVR sample compresses the decision into a cleaner moral formulation. The example mirrors the aggregate pattern: lower thematic unevenness, lower affective charge, and higher neutral narration.