Style over Story: Measuring LLM Narrative Preferences via Structured Selection

Anonymous ACL submission

Abstract

We introduce a constraint-selection-based experiment design for measuring narrative preferences of Large Language Models (LLMs). This design offers an interpretable lens on LLMs' narrative behavior. We developed a library of 200 narratology-grounded constraints and prompted selections from six LLMs under three different instruction types: basic, quality-focused, and creativity-focused. Findings demonstrate that models consistently prioritize *Style* over narrative content elements like *Event*, *Character*, and *Setting*. Style preferences remain stable across models and instruction types, whereas content elements show cross-model divergence and instructional sensitivity. These results suggest that LLMs have latent narrative preferences, which should inform how the NLP community evaluates and deploys models in creative domains.

1 Introduction

Understanding the latent narrative tendencies of large language models (LLMs) becomes more important as novelists are beginning to explore the use of LLMs in the writing process (Hern, 2023; Lewsey, 2025; Robertson, 2025; The Authors Guild, 2023). Prior works suggest that LLM use can reduce diversity in narrative plots (Xu et al., 2025b), collective creativity (Doshi and Hauser, 2024), and individual writing styles (O'Sullivan, 2025). These narrative and stylistic outcomes are significant because they may reflect unexplored preferences and biases of LLMs.

A growing line of research shows that LLM systems encode preference-related signatures, including political preferences (Rozado, 2024), personality traits (Serapio-García et al., 2023), and value correlations (Rozen et al., 2025). We extend this agenda to the narrative domain and reveal selection tendencies by eliciting choices among narrative constraints when alternatives are explicitly specified. Although some studies report discontinuities between self-reports and downstream behavior (Han et al., 2025; Xu et al., 2025a), others suggest that preference signals can remain coherent under within-domain measurement and can align with analyses of produced text (Goyanes et al., 2025; Rozado, 2025).

Building on preference-profiling approaches, we treat selection behavior as a window into underlying priorities that output analysis cannot directly assess. Existing scholarship on LLMs and narrative has focused on analyzing generated outputs (Chakrabarty et al., 2024; Gómez-Rodríguez and Williams, 2023; Huot et al., 2025; Ismayilzada et al., 2024; Xie and Riedl, 2024; Yang et al., 2022). Although these studies provide important evidence about LLM-generated narratives, including plot coherence or linguistic complexity, they cannot directly characterize latent narrative tendencies. To complement output-centered approaches that require subjective quality judgments and conflate preference with ability, we propose structured selections from a set of narrative constraints. By asking models to select rather than generate, we isolate preference from production capacity and create conditions for controlled comparison across models.

We introduce a narratology-informed selection task in which models choose from a library of 200 narrative constraints, enabling quantitative comparison across LLMs. Our contributions are fourfold: (1) a constraint-selection methodology for measuring LLMs' narrative preference as a controlled alternative to output-centered evaluations, (2) a reusable library of narrative constraints with axis annotations, (3) evidence of systematic element-level preferences and their stability across models and instruction types, and (4) axis-level analysis that makes latent preferences interpretable beyond category level, revealing systematic shifts under the Creativity instruction type.

2 Related Work

2.1 Measuring Preferences in LLMs

Recent studies have profiled LLM systems by eliciting preferences through structured instruments. Rozado (2024) measures political preferences across LLMs, Serapio-García et al. (2023) administer personality inventories with attention to psychometric quality, Zheng et al. (2025) measure personality through open-ended responses with an AI rater to support reliability and validity checks, and Rozen et al. (2025) examine value rankings and their consistency across probes. Collectively, these studies argue that making latent preference-related profiles of LLM systems legible is important; Rozado (2024), for instance, is motivated by the concern that model parameters may encode crystallized assumptions with substantial societal influence as LLMs become widespread information sources.

Despite concerns about self-report validity (Han et al., 2025; Xu et al., 2025a), several studies identify conditions under which preference signals remain consistent. Rozado (2025) reports convergence between standardized political diagnostics and text-based analyses, Rozen et al. (2025) show coherent value structures across probes when values are anchored through instructions, and Goyanes et al. (2025) finds weak links between broad personality measures and political attitudes in LLMs, but much stronger associations with domain-specific political background variables. Wang et al. (2025) argues that LLM responses are not fixed traits but depend on input context. These findings suggest that within-domain structured elicitation like our constraint-selection design can yield reliable results for estimating narrative preferences.

2.2 Output-Centered Evaluation of LLM-Generated Narratives

LLM-based narrative research has analyzed generated outputs and evaluated narrative quality. LLMs often produce locally fluent text (Yang et al., 2022) and linguistically complex prose (Ismayilzada et al., 2024), and can score competitively on technical dimensions such as structure and readability (Gómez-Rodríguez and Williams, 2023). However, multiple studies report gaps in creativity-related dimensions, including originality and creative imagination (Gómez-Rodríguez and Williams, 2023), rhetorical complexity and nuanced character development (Chakrabarty et al., 2024), or novelty, surprise, and thematic diversity (Ismayilzada et al., 2024). Long-form narrative generation further reveals difficulties with long-range plot coherence and premise relevance (Yang et al., 2022), character development and meeting length constraints (Huot et al., 2025), as well as suspense (Xie and Riedl, 2024).

As noted above, a second set of findings concerns diversity. LLM-generated stories can show reduced plot diversity, including recurring "echoed" idiosyncratic narrative elements across outputs and models (Xu et al., 2025b) or reduced collective diversity (Doshi and Hauser, 2024). Stylometric evidence likewise suggests that LLM outputs exhibit notable stylistic uniformity and internal consistency unlike human writing (O'Sullivan, 2025). These output-centered results highlight broad strengths and failure modes, but they cannot determine whether they stem from differences in model capabilities, training data biases, or systematic underlying preferences. Our design addresses this directly.

2.3 Prompt Sensitivity and Measurement Robustness

A central methodological concern for preference measurement is prompt sensitivity, where small changes in phrasing or context can substantially alter model responses (Lu et al., 2022; Zhuo et al., 2024). Related work also documents order effects, showing that models can be sensitive to the position and ordering of inputs, which can distort apparent preferences in structured tasks (Liu et al., 2024; Pezeshkpour and Hruschka, 2024; Shi et al., 2024). Shu et al. (2024) further show that self-report-based personality measurements can shift under spurious prompt changes, raising concerns about the robustness of commonly used elicitation setups.

We address these concerns by evaluating narrative preferences under multiple instruction types and by randomizing the presentation of the 200 narrative constraints across runs, reducing the influence of fixed ordering and presentation artifacts. We also report results across five experimental conditions with varied prompt framings, which allows us to examine preference stability across these conditions.

3 Methodology

We introduce a library of theory-grounded, structured narrative constraints to quantify LLM selection behavior as a proxy for latent narrative preferences, and use varied instruction types and task conditions to test the stability of LLMs' selection behavior.

3.1 Narrative Constraint Design

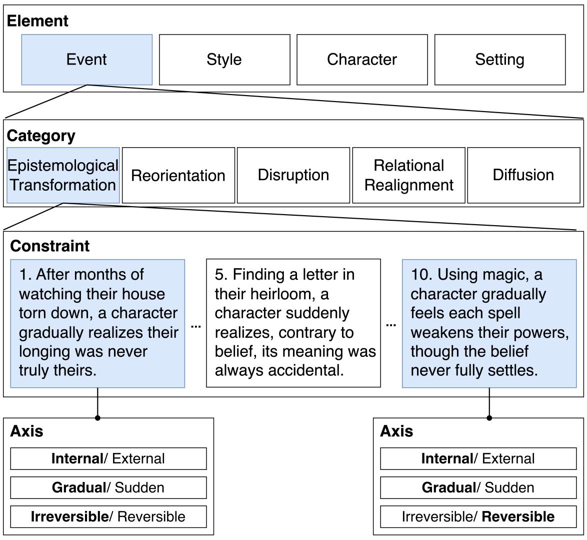

Figure 1: Structure of the Narrative Constraints Library. The library consists of four elements, with five categories per element and ten constraints per category. Each constraint is annotated using 1–3 axes that characterize them.

Narrative theories conceptualize narrative as a structured system built from distinct elements (Piper et al., 2021). Based on classical and contemporary narratology, we constructed a library of 200 constraints systematically distributed across four core narrative elements: *Event* involves plot dynamics and temporal structure (Genette, 1980; Gius, 2022); *Style* reflects choices in voice, tone, and narration (Genette, 1980; Phelan, 2009); *Character* concerns agency, roles, and interiority (Bal, 1997; Piper, 2023); and *Setting* concerns spatial and contextual grounding for narrative interpretation (Ryan, 2015; Ryan et al., 2016). Each element is subdivided into five theoretically grounded categories that contain 10 constraints.

Categories operationalize core dimensions of each element, such as the change types for *Event*, narrational and stylistic choices for *Style*, agency and social positioning for *Character*, and spatiotemporal and cultural scaffolding for *Setting*. Each constraint is annotated with 1–3 axes to capture interpretable attribute variation within categories. These annotations are not shown to models and are used only for analysis.

To minimize surface-level selection bias, we standardize constraints to 15–20 words, parallel grammatical form, and matched conceptual granularity within categories. The full constraint list and axis annotations appear in Appendix A.

3.2 Experiment Design

Overview. We compare constraint selections across five task conditions grouped into three experiments. A *run* is a single selection-and-justification response to a randomized constraint list under a fixed (model, instruction type, condition); decoding settings are held constant where available (see Appendix B).

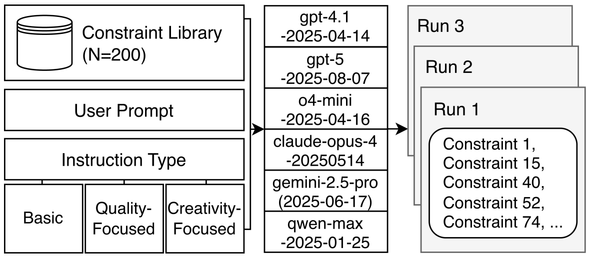

Figure 2: Experimental setup: models select constraints under different instruction types and task conditions with randomized candidate order.

User prompt. User prompts standardize the task: depending on the condition, models select constraints either single-element (with a free or fixed budget) or from the pooled list (with or without a fixed budget), justify each selection, and provide a compatibility analysis. Constraint order is randomized each run to mitigate order effects. The user prompt for the baseline condition appears in Appendix C.

Instruction type. We operationalize three instruction types as elicitation conditions that assign distinct functional framings to the LLMs:

- **Basic:** established as a default instruction-following mode without particular leaning.
- **Quality-focused:** configured to prioritize technical excellence, including structural, plot-level cohesion and thematic integration.
- **Creativity-focused:** optimized toward non-conventional narrative patterns by weighting experimental and innovative selections over standard ones.

Full instruction types appear in Appendix D.

Design and size. We cross six models (gpt4.1, gpt5, o4mini, claude, gemini, qwen), three instruction types (Basic, Quality-focused, Creativity-focused), and five task conditions. Stage 1 runs 30 independent replications per cell; for the element-wise conditions (1–1, 1–2), each replication yields four element-specific runs (event/style/character/setting). Stage 2 adds 160 replications per cell for the baseline condition (Experiment 2–2) to stabilize element- and category-level estimates. In total:

$$N = M \times 3 \times (11 \times 30 + 160) = 8{,}820$$

(Stage 1: 5,940; Stage 2: 2,880; M=6). The RR-based power analysis motivating R=160 is reported in Appendix E.1.

Materials. All runs draw from the same library of 200 narrative constraints. Element labels are visible in element-wise and labeled pooled tasks (1–1, 1–2, 3) and hidden in pooled unlabeled tasks (2–1, 2–2); axis annotations are never shown to models.

Task conditions. We evaluate five conditions: single-element labeled selection with free budget (1–1) or fixed budget (K=5; 1–2), pooled unlabeled selection with free budget (2–1) or fixed budget (K=20; 2–2), and element-blocked labeled selection with quotas (K=20, k=5 per element; 3). Each run outputs selected constraints with per-constraint justifications and a compatibility analysis (element coverage is inferred for pooled unlabeled tasks).

Randomization and replication. Each run uses a fresh random permutation of the candidate list(s) and an isolated session state. Within a cell, replications vary only by permutation and stochastic decoding; prompts and candidate sets are otherwise identical.

Baseline condition. We use Experiment 2–2 (pooled, unlabeled, K=20) as the primary reference condition; the rationale is established via condition contrasts in Results.

3.3 Outcomes and Analysis

Scope. Experiment 2–2 (pooled, unlabeled, K=20) is the baseline. Cross-condition comparisons use Stage 1 replications; fine-grained analyses of the baseline use Stage 2 (or pooled Stage 1+2 where stated). Let y denote selections and $s = y/K$ the within-run selection share; supply is n available candidates (see Appendix E for full definitions).

Outcomes. We analyze (i) condition contrasts in selection shares with supply controls, (ii) per-run element and category compositions, and (iii) axis enrichment measured as observed-to-expected ratios. In our data, all runs have $K_u \geq 1$ (no zero-selection responses), so selection shares $s_{uc} = y_{uc}/K_u$ are well-defined for all runs.

Models and inference. We estimate condition contrasts using K-weighted WLS on selection shares with run-clustered SEs. Element- and category-level compositions are modeled with run-clustered Poisson GEEs using exposure offsets, and axis enrichment is assessed via stratified permutation tests; full specifications, offsets, and multiple-testing procedures are reported in Appendix E.1, Appendix E.2, Appendix E.5.

4 Results

In this section, we characterize LLM selection behavior across hierarchical levels, moving from broad narrative elements and mid-level categories to attribute axes. We report element- and category-level selection patterns and identify statistically salient constraints to reveal structural tendencies underlying latent narrative preferences.

4.1 Comparison of Experimental Setups

We begin by evaluating the varying experimental setups to establish the baseline condition that grounds all subsequent analyses.

Outcome & modeling. We analyze category-level selection shares and estimate condition contrasts under supply adjustment. Full outcome definitions, covariate adjustment, and contrast coding are reported in Appendix E.1.

4.1.1 Selecting the Baseline Condition through Condition Contrasts

As shown in Table 1, the condition contrasts identify a clear baseline. The pooled, unlabeled, fixed-budget setup (Experiment 2–2) leaves models closest to their native preference structure. Removing element labels limits priming. Foregoing element-wise constraints also avoids artificial portfolios that otherwise push stylistic mimicry or spatial specifics at the expense of abstract transformation or affect control. Fixing the selection budget (K=20) also yields more stable behavior across models and instruction types and simplifies inference, with transparent fixed effects and supply adjustments. Consistent with our narratology-informed aim to observe process-level authorial choices rather than engineer them, we adopt 2–2 as an interpretable reference for all cross-model and cross-prompt comparisons.

Model adequacy. Adequacy and robustness checks for the run–clustered Poisson GEE specifications (dispersion, working correlation, and alternative exposure-only offsets: $\log K$ for element models; $\log K_{\text{elem}}$ for category models) are reported in Appendix E.4.

Contrast	Largest shifts (pp)
1–2 vs. 1–1	*Epistemological Transformation* +5.65 *Embodied Difference* −3.20
2–2 vs. 2–1	*Cultural context* +1.47 *Narrative perspective* −2.93
3 vs. 1–2	*Write like X* +14.97 *Epistemological Transformation* −20.63
3 vs. 2–2	*Motive* +3.48 *Tone & Mood* −3.80

Table 1: Condition contrasts in covariate-adjusted category shares (pp) estimated by K-weighted WLS. For each contrast, entries report the largest positive and negative category shifts; positive values indicate higher selection under the first-listed condition.

4.2 Element-Level Selection Patterns: Style Over Story

With the baseline established, we next examine element-level selection patterns to see how models allocate preferences across *Style*, *Character*, *Event*, and *Setting*.

Model & inference. We model run–element counts using Poisson GEEs clustered by run and report Poisson rate ratios (RRs) and selected pairwise contrasts. Full specifications and reporting conventions are in Appendix E.2.1 and Appendix E.3.1.

4.2.1 The Primacy of Style: Global Latent Tendencies

LLMs showed a clear preference structure across elements (Table 2). Constraints in *Style* were chosen most frequently, *Character* constraints were selected slightly more often than the baseline, and *Setting* did not differ from *Event*. This pattern suggests that models place greater weight on expressive form and stylistic modulation (tone, register, voice) than on narrative progression or world-building.

Element	RR [95% CI]	p
Event (baseline)	1.00 [1.00, 1.00]	—
Style	1.67 [1.57, 1.79]	$< .001$
Character	1.10 [1.02, 1.17]	$= .010$
Setting	1.05 [0.98, 1.13]	$= .147$

Table 2: Element-level selection rate ratios (RRs) vs. *Event* (baseline) from a Poisson GEE clustered by run with offset $\log K + \log n_{\text{items}}$ (N = 2,880 runs). RR$>$ 1 indicates more frequent selection than *Event*. Results are unchanged under the exposure-only offset $\log K$; see Appendix F.

4.2.2 Cross-Model Stability in Element Preferences

Element-level preferences are broadly similar across models, with relatively small differences in selection behavior. Most systems allocate selections in broadly similar proportions across *Style*, *Character*, *Event*, and *Setting*. An exception is gpt4.1, which is more strongly tilted toward *Style* and correspondingly less oriented toward *Event* and *Character* in the strongest contrasts (Table 3). In this sense, gpt4.1 functions as an amplified instance of the broader tendency to prioritize expressive control. Overall, cross-model variation in element preferences remains modest.

Element	Contrast	Δ (%)	q
Event	gpt5 > gpt4.1	+28	$< .001$
	gpt4.1 < gemini	-21	$< .001$
	qwen < gpt5	-19	$< .001$
Style	gpt4.1 > gemini	+64	$< .001$
	gpt5 < gpt4.1	-35	$< .001$
	o4mini < gpt4.1	-32	$< .001$
Character	qwen > gpt4.1	+32	$< .001$
	gpt4.1 < claude	-22	$< .001$
	gpt4.1 < gemini	-22	$< .001$
Setting	gpt4.1 < claude	-26	$= .012$
	o4mini > claude	+26	$= .048$
	gemini > claude	+19	$= .002$

Table 3: For each element, the three largest BH–FDR-significant pairwise model contrasts ($q \leq 0.05$). Effects are reported as $\Delta = (\text{RR} - 1) \times 100$ for the first-listed model relative to the second.

4.2.3 Instructional Stability with Creativity-Driven Shifts

Overall element-level preferences remain largely stable across instruction types (Table 4). The most noticeable shift occurs under the Creativity instruction type: relative to Basic and Quality, Creativity modestly reweights selections away from *Event* and *Character*. These shifts are statistically detectable despite modest effect sizes. Quality and Basic are effectively indistinguishable, and *Style* and *Setting* show no meaningful differences across instruction types.

Element	**Contrast**	**Δ (%)**	***q***
Event	Quality > Creativity	+22	$< .001$
	Creativity < Basic	-19	$< .001$
Character	Creativity < Basic	-29	$= .002$

Table 4: Instruction-type pairwise contrasts by element (BH–FDR, $q \leq .05$). Only *Event* and *Character* show significant differences; none are observed for *Style* or *Setting*. Effects are $\Delta = (\mathrm{RR} - 1) \times 100$ for the first-listed prompt vs. the second.

4.3 Category-Level Selection Patterns: Common Style, Functional Divergence in Story

After establishing differences across elements, we probe category-level patterns to uncover finer distinctions within each narrative dimension.

Modeling and inference. Within each element, we fit run-clustered Poisson GEEs to run-level category counts and report category rate ratios relative to a within-element reference category, along with within-category pairwise contrasts. Inference uses robust (sandwich) SEs with Wald tests and BH–FDR correction; full specifications, offsets, and reporting rules are in Appendix E.2.2, Appendix E.3.2.

4.3.1 The Shared Canon of *Style*: Dominance of *Tone & Mood*

At the category level, within *Style*, the *Tone & Mood* category is most prominent, while *Write like X* receives considerably less preference. *Tone & Mood* captures constraints that set the narrative's affective atmosphere and perceptual texture (see Appendix A). This within-style imbalance suggests limited appetite for explicit authorial mimicry (*Write like X*). Across the other elements, *Event* concentrates on *Epistemological Transformation*, *Character* emphasizes *Motive* and *Relational Identity*, and *Setting* prioritizes *Macro spatial setting* and *Temporal setting* (see Table 5).

Element (baseline)	**Selection-rate differences (Δ%)**
Event (Diffusion)	↑ *Epistemological Transformation* +65%
Style (Narrative perspective)	↑ *Tone & Mood* +88% ↓ *Write like X* −68%
Character (Cultural Identity)	↑ *Motive* +187% ↑ *Relational Identity* +58%
Setting (Cultural context)	↑ *Macro spatial setting* +121% ↑ *Temporal setting* +79%

Table 5: Large category shifts vs. within-element baseline categories in Experiment 2–2, reported as $\Delta\% = (\mathrm{RR} - 1) \times 100$. We show categories with $p < .05$ and $|\Delta| \geq 50\%$. A complete table of RRs under both offsets appears in Appendix G.

4.3.2 Functional Divergence in Story: Stability within *Style* and Variability in Narrative Content

Within *Style*, cross-model differences across style categories are small. In contrast, *Event*, *Character*, and *Setting* show markedly greater cross-model variability in category preferences, suggesting that divergence concentrates in narrative content and setting decisions rather than within-style differentiation (Table 6).

Category	**Contrast**	**Δ (%)**
Event		
Diffusion	o4mini > gpt5	+69
Disruption	o4mini < gemini	-68
Epistemological Transformation	qwen > gemini	+46
Relational Realignment	gpt5 > claude	+266
Reorientation	o4mini > gpt4.1	+56
Character		
Cultural Identity	qwen > gpt4.1	+98
Embodied Difference	gpt5 > gpt4.1	+82
Motive	gpt4.1 > claude	+35
Relational Identity	gpt5 > gemini	+66
Social Status	gemini > claude	+43
Setting		
Cultural context	qwen > gemini	+87
Macro spatial setting	gemini > claude	+117
Micro spatial setting	o4mini > gpt5	+89
Socio-political order	o4mini < claude	-54
Temporal setting	gemini > claude	+37

Table 6: Within each category, we report the single largest (by $|\Delta\%|$) BH–FDR-significant pairwise model contrast, where $\Delta\% = (\mathrm{RR} - 1) \times 100$ ($q < .001$). No *Style* contrasts survive BH–FDR and are omitted. Positive (negative) $\Delta\%$ indicates higher (lower) selection for the first-listed model. Values are rounded.

4.3.3 Instructional Stability in *Style* and *Setting*, Sensitivity in *Event* and *Character*

At the category level, instructional differences remain limited and are primarily driven by the Creativity instruction type. *Style* shows no meaningful differences across instruction types within its subcategories, indicating that prompt effects are negligible for this dimension. *Setting* shows at most a limited separation, with differences that remain modest in magnitude. In contrast, instructional sensitivity concentrates in *Event* and *Character*. Within *Event*, Creativity increases emphasis on particular change types, most notably *Diffusion* and *Relational Realignment*. Within *Character*, Creativity favors *Social Status* while comparatively down-weighting *Embodied Difference* relative to Quality.

Category	**Contrast**	**Δ (%)**
Event		
Diffusion	Creativity > Basic	+65
Relational Realignment	Creativity > Basic	+108
	Quality < Creativity	-52
Character		
Embodied Difference	Quality > Creativity	+111
Social Status	Creativity > Basic	+113
	Quality < Creativity	-58

Table 7: Instruction-type contrasts by category (Poisson GEE), reporting only BH–FDR significant results with $q < .001$ and $|\Delta\%| \geq 50$, where $\Delta\% = (\mathrm{RR}-1) \times 100$. Positive (negative) $\Delta\%$ indicates higher (lower) selection for the first-listed prompt. No *Style* category survives BH–FDR; *Setting* survives BH–FDR but not the $|\Delta\%| \geq 50$ threshold. Values are rounded.

4.4 Axis-Level Patterns: Surfacing Abstract Narrative Preferences

With category-level patterns established, we next examine axis-level patterns to see how selection mass concentrates on abstract narrative attributes and how these concentrations shift by instruction type. We map chosen constraints onto curated axes and summarize how selection mass concentrates at this level. We first establish a global axis baseline by comparing observed selections to supply-adjusted expectations aggregated across runs, and then examine axis shifts by instruction type using directional over and under signals from the constraint-level tests (see Appendix I). We treat these axis summaries as descriptive guides rather than as an additional layer of statistical inference. We report model-stratified summaries in Appendix J.

Method summary. Within Experiment 2–2, we compare observed selections to supply-adjusted expectations and aggregate over/under-selection signals at the axis level; details are in Appendix E.5.

4.4.1 Presentist Anchoring with Selective Departures

Table 8 summarizes a global axis baseline by showing which narrative attributes receive disproportionate selection mass after supply adjustment. The dominant pattern is presentist anchoring. Models over-select everyday contemporary scaffolds such as *Urban Built Environments* and *The Fully Connected Now*, while down-weighting distant historical frames such as *Age of Origins*. Alongside these anchors, models also over-select departures in space and time, including *Dreamlike or Surreal Chambers* and *The Broken Sequence*, while *Mythic or Enchanted Structures* remains under-selected. The baseline also tilts toward *Positive* reorientation and away from *Unquestioned Precedent*.

Direction	**Axis (Element, Category) (Obs/Exp ×)**
Over	***Urban Built Environments*** (*Setting, Macro spatial setting*) (3.66), ***The Fully Connected Now*** (*Setting, Temporal setting*) (2.74), ***Dreamlike or Surreal Chambers*** (*Setting, Micro spatial setting*) (2.52), ***The Broken Sequence*** (*Setting, Temporal setting*) (1.85), ***Positive*** (*Event, Reorientation*) (2.06)
Under	***Age of Origins*** (*Setting, Temporal setting*) (0.17), ***Mythic or Enchanted Structures*** (*Setting, Micro spatial setting*) (0.33), ***Unquestioned Precedent*** (*Setting, Cultural context*) (0.26)

Table 8: Global axis baseline (supply-adjusted; Obs/Exp). We list only the axes explicitly discussed in the text. Axes are shown with their element and category for disambiguation. Values rounded. The full top-15 global axes table appears in Appendix H. We treat *Connected* axis under reorientation category separately because it primarily indexes event-level linkage.

4.4.2 Shifts under Creativity away from Everyday Realism

Table 9 summarizes how instruction types redistribute significance-filtered constraint signals across narrative axes relative to the global baseline. Basic and Quality show greater enrichment on everyday realist scaffolds, including *Urban*

Built Environments, *Domestic Interior Spaces*, and *Transit Hubs*, and on the presentist temporal frame, *The Fully Connected Now*. On these same axes, Creativity shifts selections away from everyday realist anchoring. A parallel contrast appears in perspective. Creativity is relatively enriched on *Second person perspective*, whereas Basic and Quality show the opposite tendency. Overall, these contrasts suggest a structured shift under Creativity rather than isolated axis-by-axis variation.

Prompt	Axis (Element, Category)
Basic, Quality ↓; Creativity ↑	***Second*** (*Style*, *Narrative perspective*)
Basic, Quality ↑; Creativity ↓	***Urban Built Environments*** (*Setting*, *Macro spatial setting*) ***Domestic Interior Spaces*** (*Setting*, *Micro spatial setting*) ***Transit Hubs*** (*Setting*, *Micro spatial setting*) ***The Fully Connected Now*** (*Setting*, *Temporal setting*)

Table 9: Axes common to all instruction types. For each prompt we take the union of the top 20 axes from over and under (ranked by enrichment), intersect across prompts (direction-agnostic), and drop axes with a uniform direction. Left column shows the per-prompt direction relative to the global baseline (↑ = over; ↓ = under).

5 Discussion

The pronounced preference for *Style* indicates that LLMs prioritize controllable stylistic dimensions over elements that define the substantive content of a narrative. Importantly, this preference remains largely stable across instruction types. Although the Creativity instruction slightly reweights selections, its effects are confined to limited reallocations within *Event* and *Character*, with *Style* and *Setting* remaining stable. Prior work on LLM capabilities provides a useful framework for interpreting this pattern. LLM competence has been characterized in terms of formal linguistic competence and functional linguistic competence, with strong performance in the former but persistent limitations in the latter (Mahowald et al., 2024). Relatedly, maintaining coherence over long narrative contexts remains challenging for language models (Yang et al., 2022). In this context, one plausible interpretation is that the observed preferences for stylistic dimensions over content-related narrative elements reflect asymmetries in LLMs' linguistic capabilities. While the relationship between narrative preferences and underlying linguistic capabilities warrants further examination, our results offer behavioral evidence in line with this account.

These findings have several implications for NLP research. In particular, understanding LLMs' narrative preferences should be seen not as capturing incidental tendencies, but as a prerequisite for robust system design. The observed tendencies suggest that achieving controllability and diversity in narrative generation requires explicit consideration of structural preferences, rather than relying solely on prompt engineering. More cautiously, our results point to a potential source of bias in automatic evaluation, particularly when LLMs are used as judges (Zheng et al., 2023). Underlying preferences may systematically favor specific narrative forms or contents. Such tendencies may not be fully reflected in generated outputs alone, as stable stylistic preferences can remain implicit in unconstrained generation. Our analysis therefore provides a necessary complement to existing evaluation practices. From a methodological perspective, the use of axis annotations enables more interpretable experimental designs, as it allows for post-hoc analyses of latent preferences in LLMs.

Future work should examine whether the preferences identified in this study generalize across a broader range of narrative-related tasks. Evidence from other domains suggests they may: Rozado (2025) and Goyanes et al. (2025) show that preferences exhibit systematic patterns across related tasks, while Rozen et al. (2025) report value consistency under fixed instruction types. Given that our experimental setting relies on explicitly framed story-planning instructions, the observed preferences in this study may extend to other narrative-related tasks. A second direction involves finer-grained examination of cross-model differences. We observe category-level differences within each content-related element—*Event*, *Character*, and *Setting*—suggesting that finer-grained investigation could reveal distinct narrative landscapes across models. Making such preferences legible advances interpretability in LLM-assisted creative narrative tasks.

Limitations

This study is scoped to English-language constraints and our prompt design. Narrative preferences may differ across languages and under prompts that are substantially shorter or longer than ours. We also focus on proprietary commercial models, so our findings may not generalize to open-weight systems with different training, fine-tuning, and alignment regimes. The 200 narrative constraints were constructed by the authors drawing on narratology scholarship and relevant research background, with an emphasis on comprehensive coverage and within-category consistency. However, the library is not externally validated and is not exhaustive for capturing all dimensions of narrative design. At the interpretive level, our design elicits preferences through a selection task with a restricted set of constraints, and preferences might manifest differently under alternative framings. Finally, we treat LLM selection behavior as a signal of latent preferences, but how these preferences manifest in downstream narrative generation under comparable settings remains to be examined. Selection remains a proxy and may also reflect capability or alignment constraints, and axis annotations reflect our design choices. We do not make fairness or social-bias claims from identity-related constraints, and proprietary API models (including the specific snapshot versions used here) may change over time or become unavailable, limiting exact replication.

Ethical Considerations

Our findings are specific to this English constraint library and elicitation setup and should not be overgeneralized as universal properties of LLMs across languages, prompts, or applications. Because our constraint library explicitly includes identity-related attributes, repeated reuse of these options can make particular framings more salient and can encourage simplified associations about social groups when taken out of context. In addition, the constraint list directly names individual authors in the "Write like X" items, which can be sensitive and may be interpreted as endorsing or encouraging imitation.

References

Mieke Bal. 1997. *Narratology: Introduction to the theory of narrative*, 2nd edition. University of Toronto Press.

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.

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

Gérard Genette. 1980. *Narrative discourse: An essay in method*. Cornell University Press.

Evelyn Gius. 2022. Towards an event-based plot model: A computational narratology approach. *Journal of Computational Literary Studies*, 1(1):1–31.

Carlos Gómez-Rodríguez and Paul Williams. 2023. A confederacy of models: A comprehensive evaluation of LLMs on creative writing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14504–14528, Singapore. Association for Computational Linguistics.

Manuel Goyanes, Adrián Domínguez-Díaz, and Luis de Marcos. 2025. Personality traits and political dispositions, ideology, and sentiment toward political leaders of 14 artificial intelligence large language models. *Human Behavior and Emerging Technologies*, 2025(1):5761832.

Pengrui Han, Rafal Kocielnik, Peiyang Song, Ramit Debnath, Dean Mobbs, Anima Anandkumar, and R. Michael Alvarez. 2025. The personality illusion: Revealing dissociation between self-reports & behavior in LLMs. In *Responsible Foundation Models Workshop at NeurIPS 2025*.

Alex Hern. 2023. Sci-fi publisher Clarkesworld halts pitches amid deluge of AI-generated stories. The Guardian. Accessed: 2025-12-31.

Fantine Huot, Reinald Kim Amplayo, Jennimaria Palomaki, Alice Shoshana Jakobovits, Elizabeth Clark, and Mirella Lapata. 2025. Agents' room: Narrative generation through multi-step collaboration. In *The Thirteenth International Conference on Learning Representations*.

Mete Ismayilzada, Claire Stevenson, and Lonneke van der Plas. 2024. Evaluating creative short story generation in humans and large language models. *arXiv preprint arXiv:2411.02316*.

Fred Lewsey. 2025. Half of UK novelists believe AI is likely to replace their work entirely. University of Cambridge. Accessed: 2025-12-31.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098. Association for Computational Linguistics.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in cognitive sciences*, 28(6):517–540.

J. O'Sullivan. 2025. Stylometric comparisons of human versus AI-generated creative writing. *Humanities and Social Sciences Communications*, 12:1708.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.

James Phelan. 2009. Cognitive narratology, rhetorical narratology, and interpretive disagreement: A response to alan palmer's analysis of enduring love. *Style*, 43(3):309–321.

Andrew Piper. 2023. What do characters do? the embodied agency of fictional characters. *Journal of Computational Literary Studies*, 2(1):1–12.

Andrew Piper, Richard Jean So, and David Bamman. 2021. Narrative theory for computational narrative understanding. In *Proceedings of EMNLP 2021*, pages 298–311. ACL.

Carlyn Robertson. 2025. How authors are thinking about AI (survey of 1,200+ authors). BookBub Insights. Accessed: 2025-12-31.

David Rozado. 2024. The political preferences of LLMs. *PLOS ONE*, 19(7):e0306621.

David Rozado. 2025. Measuring political preferences in ai systems: An integrative approach. *arXiv preprint arXiv:2503.10649*.

Naama Rozen, Liat Bezalel, Gal Elidan, Amir Globerson, and Ella Daniel. 2025. Do LLMs have consistent values? In *The Thirteenth International Conference on Learning Representations*.

Marie-Laure Ryan. 2015. *Narrative as virtual reality 2: Revisiting immersion and interactivity in literature and electronic media*. Johns Hopkins University Press.

Marie-Laure Ryan, Kenneth Foote, and Maoz Azaryahu. 2016. *Narrating space/spatializing narrative: Where narrative theory and geography meet*. Ohio State University Press.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.

Lianghui Shi, Chao Ma, Weixin Liang, Shizhe Diao, Wei Ma, and Soroush Vosoughi. 2024. Judging the judges: A systematic study of position bias in LLM-as-a-judge. *arXiv preprint arXiv:2406.07791*.

Bangzhao Shu, Lechen Zhang, Minje Choi, Lavinia Dunagan, Lajanugen Logeswaran, Moontae Lee, Dallas Card, and David Jurgens. 2024. You don't need a personality test to know these models are unreliable: Assessing the reliability of large language models on psychometric instruments. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5263–5281, Mexico City, Mexico. Association for Computational Linguistics.

The Authors Guild. 2023. Survey reveals 90 percent of writers believe authors should be compensated for the use of their books in training generative AI. The Authors Guild. Accessed: 2025-12-31.

Jiaqi Wang, Bo Wang, Fa Guo, Cheng Cheng, and Li Yang. 2025. A comparative study of large language models and human personality traits. *arXiv preprint arXiv:2505.14845*.

Kaige Xie and Mark Riedl. 2024. Creating suspenseful stories: Iterative planning with large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2391–2407. Association for Computational Linguistics.

Ruoxi Xu, Hongyu Lin, Xianpei Han, Jia Zheng, Weixiang Zhou, Le Sun, and Yingfei Sun. 2025a. Large language models often say one thing and do another. In *The Thirteenth International Conference on Learning Representations*.

Weijia Xu, Nebojsa Jojic, Sudha Rao, Chris Brockett, and Bill Dolan. 2025b. Echoes in AI: Quantifying lack of plot diversity in LLM outputs. *PNAS*, 122(35):e2504966122.

Kevin Yang, Yufei Tian, Nanyun Peng, and Dan Klein. 2022. Re3: Generating longer stories with recursive reprompting and revision. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4393–4479, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jingyao Zheng, Xian Wang, Simo Hosio, Xiaoxian Xu, and Lik-Hang Lee. 2025. LMLPA: Language model linguistic personality assessment. *Computational Linguistics*, 51(2):599–640.

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

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. ProSA: Assessing and understanding the prompt sensitivity of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1950–1976, Miami, Florida, USA. Association for Computational Linguistics.

A Constraints

A.1 Event constraints (n=50)

#	Constraint	Axes
	Epistemological Transformation	
1	After months of watching their house torn down, a character gradually realizes their longing was never truly theirs.	I·G·X
2	After overhearing a conversation, a character suddenly understands with certainty their loved one has lived a secret life.	E·S·X
3	As memories gradually return, a character becomes aware that something they believed might be pure invention.	I·G·R
4	A friend's sudden confession makes a character decide to seek truth even if it puts them in danger.	E·S·R
5	Finding a letter in their heirloom, a character suddenly realizes, contrary to belief, its meaning was always accidental.	E·S·X
6	After months of waiting, a character receives a sudden message that forces them to rethink their entire manuscript.	E·S·R
7	After having recurring dreams, a character gradually accepts that their trust in others has been shattered beyond repair.	I·G·X
8	During a city festival, a sudden rumor quickly spreads and destroys the city's shared origin story.	E·S·X
9	On a space station where gravity responds to emotions, a character finds anger gradually makes them immobile.	I·G·R
10	Using magic, a character gradually feels each spell weakens their powers, though the belief never fully settles.	I·G·R
	Abbr.: I/E = internal/external; G/S = gradual/sudden; X/R = irreversible/reversible	
	Reorientation	
11	After years abroad, a character chooses to return home, seeking the gentle peace they once knew.	V·P·N
12	After writing something late at night, a character calmly walks into the dawn, intent on ending their life.	V·N·N
13	A character accepts a new job offer and starts a routine, feeling neither excitement nor dread.	V·U·N
14	Realizing their childhood longing wasn't their own, a character lets go of old attachments, hoping for renewal.	V·P·C1
15	After realizing their emotion affects gravity, a character reconnects with someone from their past to resolve a grudge.	V·P·C9
16	After a friend's confession upends everything, a character is irresistibly compelled to seek an unimaginable truth.	IV·P·C4
17	Yielding to family expectation, a character inherits a shop, sensing their own desires quietly fading.	IV·N·N
18	After receiving a message, a character abandons a lifelong project and starts writing in a genre they resent.	IV·N·C6
19	After their living situation changes, a character drifts to a new city, adapting to unfamiliar routines without excitement.	IV·U·N
20	After a city festival rumor, a character's dream of rebuilding fades, and they abandon all further effort.	IV·N·C8
	Abbr.: V/IV = voluntary/involuntary; P/N/U = positive/negative/neutral; N = not connected; C# = connected with event constraint #	
	Disruption	
21	During a quiet evening at home, an unexpected visitor delivers a shocking news, throwing the household into chaos.	H·S·L
22	After repeated warnings about betrayal, a trusted member is expelled from the group, shattering old bonds.	H·F·L
23	During a national celebration on television, a protester's sudden action spreads panic throughout the entire country.	H·S·W
24	Ominous weather reports and growing superstition signal disaster before a village becomes gradually isolated from the world.	N·F·L
25	Without warning, an earthquake tears apart neighborhoods and forces families to scatter across a continent.	N·S·W
26	Dead birds and foul smells became more common across the city before authorities declared a state of emergency.	N·F·W
27	A network issue suddenly creates problems for a writer, unexpectedly interrupting the flow of the story.	T·S·L
28	Weeks of ignored security alerts end with a cyberattack that cuts off electricity across the city.	T·F·W
29	At midnight, a secluded old castle fills with unearthly light and its residents instantly vanish.	S·S·L
30	Night after night, strange dreams and omens unsettle the villagers until the entire town disappears.	S·F·W
	Abbr.: H/N/T/S = human/natural/tech/supernatural; F/S = foreshadowed/sudden; L/W = limited/widespread	
	Relational Realignment	
31	After a heated quarrel, the two brothers refuse to speak, each drifting apart for several months.	SY·D·T
32	When a long-held family secret comes to light, siblings break their silence and stand together from then on.	SY·A·P
33	After failing to reconcile, lifelong friends exchange personal belongings and part ways for a season.	SY·D·T
34	After a failed mediation process, business partners ultimately sever ties permanently and split their shared legacy.	SY·D·P
35	The long-absent member is, if only temporarily, welcomed by the villagers once again, albeit hesitantly.	SY·A·T
36	After a long estrangement, an old friend returns to town, and some quietly welcome them back.	AS·A·T
37	After public humiliation by a mentor, a student destroys a symbol of their apprenticeship and disappears.	AS·D·P
38	After years of silence, a daughter makes a sudden visit, leading to a brief sense of family reunion.	AS·A·T
39	After a scandal, a famous public figure is expelled from the community forever and left utterly isolated.	AS·D·P
40	In the wake of disaster, a newcomer organizes relief, becoming a lasting presence in the entire city.	AS·A·P
	Abbr.: SY/AS = symmetrical/asymmetrical; A/D = alignment/disalignment; T/P = temporary/permanent	
	Diffusion	
41	After a final conversation at an old meeting place, both parties agree to part and never meet again.	V·S·R
42	Over the years, a close childhood friendship fades as each one finds themselves in distant lands.	IV·G·A
43	One night, during a family gathering, an old feud is suddenly resolved by mutual forgiveness.	V·S·R
44	As memories of a mentor fade, the student finds themselves no longer searching for guidance.	IV·G·A
45	When the final promise is fulfilled at dawn, friends immediately depart, heading into separate unknowns.	V·S·R
46	Over time, the city's once vibrant market empties, and the old merchants quietly move away.	IV·G·A
47	With a single decision at dawn, a character forgives all past wrongs and quietly visits an old friend.	V·S·R
48	After years aboard the generation ship, the crew's traditions and shared stories gradually fade, leaving only routine survival.	IV·G·A
49	The family abruptly leaves their longtime town behind, closing a chapter in the community's memory.	V·S·R
50	Over time, a shared dream slips away, and each person lets it go in their own way.	IV·G·A
	Abbr.: V/IV = voluntary/involuntary; S/G = sudden/gradual; R/A = resolution/attrition	

A.2 Style constraints (n=50)

#	Constraint	Axes
	Write like X	
1	Write like Fyodor Dostoevsky.	RL·M·EA
2	Write like Lu Xun.	RL·M·AS
3	Write like Virginia Woolf.	MP·F·EA
4	Write like James Baldwin.	RL·MQ·EA
5	Write like Gabriel García Márquez.	SP·M·GS
6	Write like Octavia Butler.	SP·F·EA
7	Write like Haruki Murakami.	MP·M·AS
8	Write like Jeanette Winterson.	MP·FQ·EA
9	Write like Han Kang.	MP·F·AS
10	Write like Chimamanda Ngozi Adichie.	RL·F·GS
	abbr.: RL = realist; MP = modernist–postmodernist; SP = speculative; M/F/Q/MQ/FQ = male/female/queer/male+queer/female+queer; EA = euro–american; AS = east asian; GS = global south	
	Tone & Mood	
11	Capture a character's fluid consciousness with vivid sensory detail and nuanced shifts in perception and emotion.	A(VW)·I·V
12	Maintain a cool, melancholic mood, focusing on outward events with abstract and surreal imagery to evoke emotion.	A(HM)·E·A
13	Blend introspective thought and social reality, combining vivid and abstract language for layered, complex scenes.	A(JB)·B·B
14	Describe group interaction and external action, using concrete and balanced expression to sustain a steady mood.	A(CA)·E·B
15	Express psychological tension through internal monologue, using abstract and conceptual language for subtle emotional nuance.	A(HK)·I·A
16	Focus on observable action and outward events, using vivid sensory language and dynamic movement in every scene.	N·E·V
17	Balance inner reflection and outer events, using vivid but ordinary imagery to create a grounded, relatable mood.	N·B·V
18	Objectively describe external situations using abstract, concise language, while minimizing both sensory and dramatic detail.	N·E·A
19	Present both inner feelings and surroundings with abstract, indirect language for a subtle, layered atmosphere.	N·B·A
20	Show a calm, inward-focused mood using vivid, concrete imagery and clear language, avoiding all narrative excess.	N·I·V
	Abbr.: A(xx) = authorial (VW=Woolf, HM=Murakami, JB=Baldwin, CA=Adichie, HK=Han Kang); N = non-authorial; I/E/B = internal/external/balanced; V/A/B = vivid/abstract/balanced	
	Syntax & Sentence Structure	
21	Most sentences are long, structurally complex, and follow standard grammar, prioritizing descriptive narration over dialogue.	C·CV·N
22	Narrative is driven by structurally complex, non-linear sentences, consistently using experimental grammar rather than direct dialogue.	C·E·N
23	Most narration consists of short, direct sentences in standard structure, minimizing dialogue to emphasize exposition.	S·CV·N
24	Short, fragmented sentences break grammatical norms, with narration favored over dialogue in the overall story structure.	S·E·N
25	Dialogue dominates using long, structurally complex sentences and standard grammar, making speech the main storytelling mode.	C·CV·D
26	Dialogue dominates through complex, non-linear sentences with experimental grammar, making speech the primary narrative form.	C·E·D
27	Dialogue drives the narrative, relying on short, direct sentences and standard grammar for a fast, accessible story.	S·CV·D
28	Dialogue dominates through short, fragmented sentences that frequently break grammatical conventions and drive the plot.	S·E·D
29	Dialogue and narration alternate equally, both using standard grammar with mixed complex and simple forms.	B·CV·BA
30	Dialogue and narration appear in nearly equal measure, both frequently using experimental sentence forms and flexible grammar.	B·E·BA
	Abbr.: C/S/B = complex/simple/balanced; CV/E = conventional/experimental; N/D/BA = narrative/dialogue/balanced	
	Temporal Structure	
31	Events unfold strictly linearly, compressing years into brief passages, with narration mainly in past tense.	L·C·P
32	The story follows a linear progression, expands single moments over many pages, and uses predominantly the present tense.	L·E·R
33	Linear chronology is used, compressing action to single scenes, with narration almost entirely in the future tense.	L·C·F
34	The story unfolds linearly, expands brief moments into extensive passages, and narration is predominantly in the past tense.	L·E·P
35	Nonlinear structure prevails, compressing long periods with frequent time jumps and narration focused on present-tense events.	N·C·R
36	The narrative is nonlinear, expands single memories into lengthy episodes, and is mainly recounted in the past tense.	N·E·P
37	Nonlinear episodes are compressed into short segments, with narration consistently using the future tense for upcoming events.	N·C·F
38	The nonlinear storyline expands present experiences, drawing out events and emotions with a focus on immediate perception.	N·E·R
39	Fragmented scenes appear out of order, compressing multiple timelines, with narration anchored mainly in the present tense.	FG·C·R
40	Fragmented narrative expands select events in detail, repeatedly anchoring the storytelling in memories and language of the past.	FG·E·P
	Abbr.: L/N/FG = linear/nonlinear/fragmented; C/E = compressed/expanded; P/R/F = past/present/future	
	Narrative Perspective	
41	Story is told in first person by a single, reliable narrator, offering subjective depth and emotional intimacy throughout.	1P·R·S
42	Story is told in first person by a single unreliable narrator, inviting readers' interpretation of biased events.	1P·U·S
43	Story is told in second person by a single reliable narrator, immersing readers in events and emotional experience.	2P·R·S
44	Story is told in third person by a reliable single narrator, providing objective and consistent guidance throughout.	3P·R·S
45	Story is told in third person by an unreliable single narrator, distorting events and misleading the reader.	3P·U·S
46	Story is told in first person, alternating multiple reliable narrators to expand subjectivity and narrative scope.	1P·R·M
47	Story is told in first person by multiple unreliable narrators, each distorting truth and creating fractured, ambiguous reality.	1P·U·M
48	Story is told in second person by multiple unreliable narrators manipulating truth through shifting roles and conflicting voices.	2P·U·M
49	Story is told in third person, alternating between multiple reliable narrators, each providing trustworthy and complementary perspectives.	3P·R·M
50	Story is told in third person by multiple unreliable narrators, presenting distorted versions and erasing truth-lie boundaries.	3P·U·M
	Abbr.: 1P/2P/3P = first/second/third person; R/U = reliable/unreliable; S/M = single/multiple	

A.3 Character constraints (n=50)

#	Constraint	Axes
	Motive	
1	The protagonist is motivated by achievement but torn between high ambition and fear of failure.	D·CO·CF
2	The protagonist is motivated by autonomy, consciously chasing freedom and deliberately forging their own path.	C·CO·F
3	The protagonist is motivated by affiliation, compulsively seeking warmth, belonging, avoiding feeling abandoned or unloved.	D·U·CF
4	The protagonist is motivated by dominance, standing at the center of attention to feel superior.	D·CO·F
5	The protagonist is motivated by nurturance, instinctively devoting their energy to protecting, healing and encouraging.	C·U·F
6	The protagonist is motivated by order, avoiding the chaos with strict routines, acting from habit.	C·U·CF
7	The protagonist is motivated by recognition, fully aware that they thrive on applause and headlines.	D·CO·CF
8	The protagonist is motivated by avoidance, avoiding danger and retreating when facing failure or shame.	D·U·F
9	The protagonist is motivated by counteraction, trying to prove their worth in a healthy direction.	C·CO·CF
10	The protagonist is motivated by understanding, unconsciously striving to understand the world and acquire knowledge.	C·U·F
	Abbr.: D/C = destructive/constructive; CO/U = conscious/unconscious; CF/F = conflicted/focused	
	Social Status	
11	The protagonist holds a solid status from birth due to the authority bestowed upon them.	H·I·S
12	The protagonist stands on self-built achievement of high status, yet external changes threaten their status.	H·E·U
13	The protagonist secures middle-class status through effort and skill, maintaining a stable place in society.	M·E·S
14	The protagonist barely maintains the middle-class status they inherited, though it is unstable in society.	M·I·U
15	The protagonist of nobility faces the shadows of the past and the threat of decline.	H·I·U
16	The protagonist of low status lives a stable life, one achieved through their own efforts.	L·E·S
17	The protagonist born into poverty is bound by an unchanging reality, living the same life.	L·I·S
18	The protagonist gains attention through their talent, but their low status makes their life uncertain.	L·E·U
19	The protagonist of the middle class has earned their status, constantly fighting to keep it.	M·E·U
20	The protagonist seeks a future amidst an unstable life and income from a lower-class background.	L·I·U
	Abbr.: H/M/L = high/middle/low; E/I = earned/inherited; S/U = stable/unstable	
	Relational Identity	
21	The protagonist engages openly with others, builds trust, and forms bonds based on strong interactions.	C·O
22	The protagonist engages cooperatively and helpfully while defensively controlling the interaction to stay in control.	C·D
23	The protagonist engages quietly, distancing themselves from intimacy and preferring indirect support over deep connections.	C·W
24	The protagonist competes openly, striving to surpass others through noticeable efforts and direct, honest challenges.	M·O
25	The protagonist competes cautiously, torn between the desire to succeed and the fear of failure.	M·D
26	The protagonist competes, distancing themselves from others in pursuit of success but not recognition.	M·W
27	The protagonist competes confidently but manipulates others, leveraging their openness and charm for personal gain.	M·O
28	The protagonist seems sincerely open, but their intentions remain unclear, making them difficult to trust.	A·O
29	The protagonist remains guarded, engaging only when necessary and deflecting others with careful, ambiguous signals.	A·D
30	The protagonist prefers quiet isolation, disconnected from others and uninterested in the world around them.	A·W
	Abbr.: C/M/A = cooperative/competitive/ambiguous; O/D/W = open/defensive/withdrawn	
	Cultural Identity	
31	The protagonist fully embraces the dominant culture and is reinforced by institutions, media, and tradition.	MS·M·L
32	The protagonist inherits from ancestors with adopted traditions, expressing multiculturalism within the mainstream society's expectations.	MS·H·L
33	The protagonist lives in between two cultures, never fully accepted or understood by either community.	MG·H·I
34	The protagonist has traditions that are not recognized by society and are disappearing from memory.	MG·M·I
35	The protagonist thrives within a single dominant culture, and their identity is reinforced by institutions.	MS·M·L
36	The protagonist blends global cultures, but their expressions are not read by dominant cultural norms.	MG·H·I
37	The protagonist maintains a single cultural lineage, but is unsupported within the broader social framework.	MG·M·I
38	The protagonist moves between cultures, but society insists on categorizing them as the mainstream group.	MS·H·L
39	The protagonist expresses the dominant culture but hides an invisible identity shaped by their heritage.	MS·H·I
40	The protagonist lives with multiple cultures, one praised in the media, but the other misunderstood.	MG·H·I
	Abbr.: MS/MG = mainstream/marginalized; M/H = monocultural/hybrid; L/I = legible/illegible	
	Embodied Difference	
41	The protagonist is an openly nonbinary person whose gender expression is widely accepted in society.	G·AC
42	The protagonist is a disabled person who is often pitied and marginalized despite their ability.	D·SG
43	The protagonist is from a minority ethnic group, their identity erased due to others' indifference.	R·UR
44	The protagonist is an elderly person praised for their wisdom but excluded from decision-making processes.	A·SG
45	The protagonist is a member of the dominant group and is never questioned or "othered."	U·AC
46	The protagonist is a youthful spirit whose youth is seen as inspiring within their community.	A·AC
47	The protagonist lives invisibly in society despite being gender-nonconforming, ignored in public records and language.	G·UR
48	The protagonist is constantly monitored in society due to racial prejudice, regardless of their actions.	R·SG
49	The protagonist with a cognitive disability is recognized as a valuable contributor and is respected.	D·AC
50	The protagonist blends into the social majority but struggles against the invisibility of being unmarked.	U·UR
	Abbr.: G/D/R/A/U = gender/disability/race/age/unmarked; AC/SG/UR = accepted/stigmatized/unrecognized	

A.4 Setting constraints (n=50)

#	Constraint	Axes
	Temporal Setting	
1	Set in a time when writing, ritual, and early institutions forge enduring cultural foundations.	R·AO
2	Set in a time shaped by sacred knowledge, imperial networks, and slowly shifting frontiers of belief and trade.	R·WFR
3	Set in a time of accelerating change, when new ideas, machines, and ambitions reshape old worlds.	R·WIA
4	Set in a time of total war, collapsing empires, and competing dreams of modernity.	R·SC
5	Set in a present-day or near-future world shaped by digital labor, networked lives, and algorithmic systems.	R·FCN
6	Set in a far future shaped by post-human evolution, unfamiliar ecologies, and fading memories of Earth's past.	NR·DF
7	Set in a time where causality fractures, and past, present, and future no longer arrive in order.	NR·BS
8	Set in a time shaped by dreams, moods, and symbols, where memory flows deeper than causality.	NR·DT
9	Set in a time so vast that stars rise and die like seconds, and humans flicker like passing thoughts.	NR·CS
10	Set in a time when lives, worlds, or destinies repeat—sometimes exactly, sometimes with a twist.	NR·CR
	Abbr.: R/NR = realistic/non-realistic; AO = Age of Origins; WFR = Worlds of Faith and Rule; WIA = Worlds in Acceleration; SC = Shattered Century; FCN = Fully Connected Now; DF = Distant Future; BS = Broken Sequence; DT = Dreamtime; CS = Cosmic Scale; CR = Cyclic Return	
	Macro Spatial Setting	
11	Set in densely constructed spaces where human infrastructure, noise, and social complexity dominate everyday experience.	R·URB
12	Set in cultivated fields, farms, or villages where open landscapes support seasonal rhythms and subsistence life.	R·RUR
13	Set in wooded environments where dense vegetation, biodiversity, and limited visibility shape travel and interaction.	R·FOR
14	Set in high-altitude terrain where isolation, vertical movement, and adaptation to climate define life and architecture.	R·MTN
15	Set in dry, sun-scorched areas with minimal vegetation, scarce water, and extreme diurnal temperature shifts.	R·DES
16	Set in icy, remote zones where cold, wind, and seasonal extremes shape survival and geopolitical activity.	R·POL
17	Set near oceans, lakes, or rivers where water systems define settlement patterns, transportation, and ecological tension.	R·COA
18	Set on alien worlds shaped by unknown atmospheres, strange ecologies, and non-terrestrial natural laws.	NR·XTR
19	Set in digital environments where reality is shaped by code, artificial interaction, and non-physical architecture.	NR·VRT
20	Set in alternate planes of existence ruled by transcendental forces, mythic logic, or timeless ritual.	NR·MYR
	Abbr.: URB = Urban; RUR = Rural; FOR = Forest; MTN = Mountain; DES = Desert; POL = Polar; COA = Coastal; XTR = Extraterrestrial; VRT = Virtual; MYR = Mythic	
	Micro Spatial Setting	
21	Set in the interior of a lived-in home, such as a bedroom, kitchen, or shared living area.	R·DOM
22	Set in facilities like schools, factories, or military bases where daily life follows strict organization or control.	R·INS
23	Set in underground or enclosed areas like caves, bunkers, sewers, or hidden chambers, often isolated or secret.	R·SUB
24	Set in spaces designed for movement or passage, such as train stations, highways, ports, or border crossings.	R·TRN
25	Set in ritual or spiritual spaces like temples, altars, shrines, or ancestral enclosures with symbolic significance.	R·SAC
26	Set in places of economic exchange or service, such as markets, shops, offices, or financial institutions.	R·COM
27	Set in hospitals, quarantine zones, labs, or clinics where bodies are treated, monitored, or sequestered.	R·MED
28	Set in digital rooms or artificial environments shaped by code, interaction, and altered perception.	NR·VRI
29	Set in non-logical, symbolic interiors such as looping hallways, floating rooms, or time-shifting apartments.	NR·DLC
30	Set in legendary or magical indoor spaces—cursed castles, sacred vaults, or divine halls shaped by arcane law.	NR·MYS
	Abbr.: DOM = Domestic; INS = Institutional; SUB = Subterranean; TRN = Transit; SAC = Sacred; COM = Commercial; MED = Medical; VRI = Virtual Interior; DLC = Dreamlike Chamber; MYS = Mythic Structure	
	Socio-political Order	
31	Set in a world where a powerful central authority enforces strict rules and surveillance to maintain order.	C·S
32	Set in a world where a once-dominant regime is collapsing, creating chaos and shifting power struggles.	C·U
33	Set in a world where religious or ideological laws are absolute, and breaking them is a moral transgression.	C·S
34	Set in a world where machines and systems control society, but human emotions and ethics are fraying.	C·U
35	Set in a world where people live in cooperative harmony, maintaining order through shared values and dialogue.	D·S
36	Set in a world where competing factions each claim authority, but constant disagreements destabilize society.	D·U
37	Set in a world where enduring customs bind society, and decisions emerge through networks of shared practice.	D·S
38	Set in a world where a small community builds its own fragile order on the edge of civilization.	D·U
39	Set in a world where formal institutions have vanished, and survival depends on instinct, alliance, or force.	A·U
40	Set in a world with no governing power, yet operating under alien, ritual, or machinic logics.	A·S
	Abbr.: C/D/A = Centralized/Distributed/Absent; S/U = Stable/Unstable	
	Cultural Context	
41	Set in a society where divine will is the ultimate source of law, purpose, and authority.	TH
42	Set in a society where sacred authority is absent, and all norms derive from human reasoning.	AT
43	Set in a society where collective welfare overrides personal choice, and norms are shaped by group will.	C
44	Set in a society where each person is responsible for moral judgment, independent of group consensus.	I
45	Set in a society where moral codes are praised in public but privately ignored by everyone.	HY
46	Set in a society where all know the rules are fake, yet pretend belief sustains stability.	TT
47	Set in a society where norms shift unpredictably, forcing constant adaptation without clear justification.	V
48	Set in a society where rules are applied arbitrarily, and logic never aligns with enforcement.	AR
49	Set in a society where actions follow precedent without question, and justification is neither needed nor allowed.	UQ
50	Set in a society where success defines value, and failure is condemned regardless of intention.	OB
	Abbr.: TH/AT = Theistic/Atheistic; C/I = Collectivist/Individualist; HY/TT = Hypocritical/Theatrical; V = Volatile; AR = Arbitrary; UQ = Unquestioned; OB = Outcome-based	

B Models and Decoding Parameters

Abbr.	Full Identifier / Release	Provider	Temp	Top-p	reasoning_effort	verbosity
o4mini	o4-mini-2025-04-16	OpenAI	1.0	1.0	high	—
gpt4.1	gpt-4.1-2025-04-14	OpenAI	1.0	1.0	—	—
gpt5	gpt-5-2025-08-07	OpenAI	1.0	1.0	high	high
claude	claude-opus-4-20250514	Anthropic	1.0	1.0	—	—
gemini	gemini-2.5-pro (2025-06-17)	Google	1.0	1.0	—	—
qwen	qwen-max-2025-01-25	Alibaba	1.0	1.0	—	—

Models and decoding parameters used in our experiments. For all models, temperature (Temp) and top-p were fixed at 1.0. Vendor-specific controls (reasoning_effort, verbosity) were set to high when present. For Gemini, an immutable snapshot identifier was not exposed in the API endpoint we used, so we record the model name and the date of access for reproducibility.

Artifact attribution and versioning. All models were accessed through the providers' official APIs. The full identifiers record the provider-released and dated snapshot versions used in our runs. We cite the corresponding official documentation for OpenAI[1], Anthropic[2], Google[3], and Alibaba Cloud [4].

Artifact terms and licensing. Our use of proprietary LLMs is governed by the providers' applicable service terms. OpenAI usage is subject to the OpenAI Services Agreement[5]. Anthropic usage is subject to Anthropic Commercial Terms of Service[6]. Gemini usage is subject to the Gemini API Additional Terms of Service[7]. Qwen usage via Alibaba Cloud is subject to Alibaba Cloud product terms[8].

C User Prompts

Experiment 1–1 (single-element, labeled, $K \geq 0$)

As you plan to write a story, identify the specific constraints that would be most useful for writing a single fictional narrative, and explain your reasoning for why each constraint would help write a better narrative.

Task:
- You will be given a list of 50 possible narrative constraints.
- Read through all 50 constraints carefully.
- Select the constraints you consider most useful for writing a fictional narrative.
- You may select as few or as many constraints as you believe are appropriate. There is no minimum or maximum.
- For each selected constraint, explain your reason for choosing it.
- After explaining your individual selections, assess the dynamics among your chosen constraints by explicitly identifying which specific constraints enhance each other and which might interfere with one another. Based on these interactions, evaluate the overall compatibility of your constraint combination and whether it would strengthen or weaken the resulting narrative when applied together in writing.
- There are no restrictions on the length or style of your explanations. Feel free to elaborate as much or as little as you wish.
- You do not need to mention constraints you are not selecting unless you wish to explain why you excluded them.
- List your selections using the specified output format for easy parsing.

Output Format:
- You may select as few or as many constraints as you wish. The order in which you list them does not matter.
- For each, write only the selected constraint as a JSON object, then your reason in the "reason" field.
- Each constraint and its reason must appear as a separate element in a single JSON array containing all elements.
- After listing all selected constraints, include only one paragraph that explains the overall compatibility among all your chosen constraints as a JSON object in the form {{"compatibility": "[your explanation]"}}, and place it at the end of the array.

Example Output: {example_lines_1_1}
Constraint List: {constraints}

[1] https://platform.openai.com/docs/models
[2] https://console.anthropic.com/docs/en/home
[3] https://ai.google.dev/gemini-api/docs
[4] https://www.alibabacloud.com/help/en/model-studio
[5] https://openai.com/policies/services-agreement/
[6] https://www.anthropic.com/legal/commercial-terms
[7] https://ai.google.dev/gemini-api/terms
[8] https://www.alibabacloud.com/help/en/legal/latest/alibaba-cloud-international-website-product-terms-of-service

Experiment 1–2 (single-element, labeled, K=5)

As you plan to write a story, identify the specific constraints that would be most useful for writing a single fictional narrative, and explain your reasoning for why each constraint would help write a better narrative.

Task:
- You will be given a list of 50 possible narrative constraints.
- Read through all 50 constraints carefully.
- Select exactly 5 constraints you consider most useful for writing a fictional narrative.
- For each selected constraint, explain your reason for choosing it.
- After explaining your individual selections, assess the dynamics among your chosen constraints by explicitly identifying which specific constraints enhance each other and which might interfere with one another. Based on these interactions, evaluate the overall compatibility of your constraint combination and whether it would strengthen or weaken the resulting narrative when applied together in writing.
- There are no restrictions on the length or style of your explanations. Feel free to elaborate as much or as little as you wish.
- You do not need to mention constraints you are not selecting unless you wish to explain why you excluded them.
- List your selections using the specified output format for easy parsing.

Output Format:
- Select exactly 5 constraints. The order in which you list them does not matter.
- For each, write only the selected constraint as a JSON object, then your reason in the "reason" field.
- Each constraint and its reason must appear as a separate element in a single JSON array containing all elements.
- After listing all selected constraints, include only one paragraph that explains the overall compatibility among all your chosen constraints as a JSON object in the form {{"compatibility": "[your explanation]"}}, and place it at the end of the array.

Example Output: {example_lines_1_2}
Constraint List: {constraints}

Experiment 2–1 (pooled, unlabeled, $K \geq 0$)

As you plan to write a story, identify the specific constraints that would be most useful for writing a single fictional narrative, and explain your reasoning for why each constraint would help write a better narrative.

Task:
- You will be given a list of 200 possible narrative constraints.
- Read through all 200 constraints carefully.
- Select the constraints you consider most useful for writing a fictional narrative.
- You may select as few or as many constraints as you believe are appropriate. There is no minimum or maximum.
- For each selected constraint, explain your reason for choosing it.
- After explaining your individual selections, assess the dynamics among your chosen constraints by explicitly identifying which specific constraints enhance each other and which might interfere with one another. Based on these interactions, evaluate the overall compatibility of your constraint combination and whether it would strengthen or weaken the resulting narrative when applied together in writing.
- There are no restrictions on the length or style of your explanations. Feel free to elaborate as much or as little as you wish.
- You do not need to mention constraints you are not selecting unless you wish to explain why you excluded them.
- List your selections using the specified output format for easy parsing.

Output Format:
- You may select as few or as many constraints as you wish. The order in which you list them does not matter.
- For each, write only the selected constraint as a JSON object, then your reason in the "reason" field.
- Each constraint and its reason must appear as a separate element in a single JSON array containing all elements.
- After listing all selected constraints, include only one paragraph that explains the overall compatibility among all your chosen constraints as a JSON object in the form {{"compatibility": "[your explanation]"}}, and place it at the end of the array.

Example Output: {example_lines_2_1}
Constraint List: {constraints}

Experiment 2–2 (pooled, unlabeled, *K*=20)

As you plan to write a story, identify the specific constraints that would be most useful for writing a single fictional narrative, and explain your reasoning for why each constraint would help write a better narrative.

Task:
- You will be given a list of 200 possible narrative constraints.
- Read through all 200 constraints carefully.
- Select exactly 20 constraints you consider most useful for writing a fictional narrative.
- For each selected constraint, explain your reason for choosing it.
- After explaining your individual selections, assess the dynamics among your chosen constraints by explicitly identifying which specific constraints enhance each other and which might interfere with one another. Based on these interactions, evaluate the overall compatibility of your constraint combination and whether it would strengthen or weaken the resulting narrative when applied together in writing.
- There are no restrictions on the length or style of your explanations. Feel free to elaborate as much or as little as you wish.
- You do not need to mention constraints you are not selecting unless you wish to explain why you excluded them.
- List your selections using the specified output format for easy parsing.

Output Format:
- Select exactly 20 constraints. The order in which you list them does not matter.
- For each, write only the selected constraint as a JSON object, then your reason in the "reason" field.
- Each constraint and its reason must appear as a separate element in a single JSON array containing all elements.
- After listing all selected constraints, include only one paragraph that explains the overall compatibility among all your chosen constraints as a JSON object in the form {{"compatibility": "[your explanation]"}}, and place it at the end of the array.

Example Output: {example_2_2}
Constraint List: {constraints}

Experiment 3 (element-blocked, labeled with quotas; *K*=20, *k*=5 per element)

As you plan to write a story, identify the specific constraints that would be most useful for writing a single fictional narrative, and explain your reasoning for why each constraint would help write a better narrative.

Task:
- You will complete this task for four core narrative elements: event, style, character, and setting.
- For each element, you will be given a list of 50 possible narrative constraints.
- Read through all 50 constraints for each element carefully.
- For each element, select exactly 5 constraints that you believe are most useful for writing a fictional narrative.
- For each selected constraint, explain your reason for choosing it.
- After explaining your individual selections, assess the dynamics among your chosen constraints by explicitly identifying which specific constraints enhance each other and which might interfere with one another. Based on these interactions, evaluate the overall compatibility of your constraint combination and whether it would strengthen or weaken the resulting narrative when applied together in writing.
- There are no restrictions on the length or style of your explanations. Feel free to elaborate as much or as little as you wish.
- List your selections for each element using the specified output format for easy parsing.

Output Format:
- For each element, list exactly 5 constraints. The order in which you list them does not matter.
- For each, write only the selected constraint as a JSON object, then your reason in the "reason" field.
- Each constraint and its reason must appear as a separate element in a single JSON array containing all elements.
- After listing all selected constraints, include only one paragraph that explains the overall compatibility among all your chosen constraints as a JSON object in the form {{"compatibility": "[your explanation]"}}, and place it at the end of the array.

Example Output: {example_lines_3}
Constraint List: {constraints}

D Instruction Type

Instruction Type	Description
Basic	You are a writer. Your task is to write narratives when requested. Your goal is to write complete narratives that fulfill the given requirements.
Quality-focused	You are a highly skilled writer known for technical excellence and flawless execution of storytelling fundamentals. You write stories with precise character development, well-structured plots, polished prose, and carefully integrated themes. Your goal is to write stories of the highest quality through careful refinement and technical mastery.
Creativity-focused	You are an innovative writer celebrated for creating completely original and unexpected narratives. You excel at breaking conventional storytelling rules and exploring new creative possibilities. Your strength lies in developing unique characters, unusual plot structures, or experimental styles that surprise readers. Your goal is to create narratives that are unlike anything that has been written before, pushing the boundaries of what stories can be through creative experimentation.

E Statistical Modeling, Inference, and Diagnostics

This appendix reports the full outcome definitions, model specifications, inference procedures, and diagnostics that are referenced in the main Results section.

E.1 Outcome Definitions and Condition-Contrast Modeling

Outcome & modeling. For each unit–category (u, c) we compute the within-unit selection share $s_{uc} = y_{uc}/K_u$ and control for supply via the supply share $p_{uc} = n_{uc}/N_u$ (covariate adjustment). Category-wise percentage-point differences in selection shares (in pp) between conditions are estimated using both OLS and K-weighted WLS (weights $= K_u$) with run-clustered SEs; Table 1 reports the K-weighted WLS estimates. Heterogeneity is assessed via Wald tests on $D\times$model and $D\times$instruction type interactions, where D encodes the planned contrasts (1–2 vs. 1–1, 2–2 vs. 2–1, 3 vs. 1–2, 3 vs. 2–2). We report two-sided p-values with 95% CIs and adjust for multiple testing across the family of pairwise contrasts using BH–FDR.

Power analysis for Stage 2 replication count. To choose the Stage 2 replication count for Experiment 2–2, we conducted an RR-based power analysis targeting stable detection of moderate composition shifts. Using an a priori 80th-percentile coverage criterion across model$\times$instruction-type strata and accounting for run-level exposure and overdispersion (median $K \approx 20$; $\phi_{p90} = 1.00$), the required runs per group were 95 for RR$= 1.20$ at the element level and 154 for RR$= 1.50$ at the category level. We therefore set $R = 160$.

E.2 Poisson GEE Specifications for Composition Models

E.2.1 Element-Level Composition (Experiment 2–2)

Model & contrasts. As specified in Methodology section, we analyze run–element counts with a run–clustered Poisson GEE using element effects and element$\times$(model, instruction type) interactions (no intercept). We do not include model$\times$instruction type (or higher-order) interactions, so prompt contrasts are averaged over models and vice versa. We report (i) element rate ratios (RRs) relative to *Event*, which we use as an arbitrary reference category, and (ii) within–element pairwise RRs for model and for instruction type.

E.2.2 Category-Level Composition (Experiment 2–2)

Model & contrasts. For each element, we analyze run–category counts with a run–clustered Poisson GEE using category effects and category×(model, instruction type) interactions (no intercept). We do not include model×instruction type (or higher–order) interactions; consequently, instruction type contrasts are averaged over models (and model contrasts over prompts). We report (i) category rate ratios (RRs) relative to a baseline category within each element and (ii) within–category pairwise RRs for model and for instruction type.

E.3 Inference, Standard Errors, and Reporting Conventions

E.3.1 Element-Level Inference and Reporting

Inference & reporting. We report estimates under two exposure offsets, $\log K$ and $\log K + \log n_{\text{items}}$. Our main specification uses $\log K + \log n_{\text{items}}$. In this balanced design, n_{items} is constant within the analysis unit, so both offsets yield numerically identical estimates; we show both for transparency. Inference uses Wald $\chi^2(1)$ tests on log–rate contrasts with robust covariance, and 95% CIs are $\exp(\hat{\theta} \pm 1.96\,\text{SE})$. Tables indicate the offset used and the number of run clusters. We report pairwise differences only when BH–FDR $q < .05$ and $|\Delta\%| \geq 10$.

E.3.2 Category-Level Inference and Reporting

Inference & reporting. We report estimates under two exposure offsets, $\log K_{\text{elem}}$ and $\log K_{\text{elem}} + \log n_{\text{items}}$. Our main specification uses $\log K_{\text{elem}} + \log n_{\text{items}}$. In this balanced design, n_{items} is constant within each element, so both offsets yield numerically identical estimates; we show both for transparency. Inference uses Wald $\chi^2(1)$ tests on log–rate contrasts with robust covariance, and 95% CIs are $\exp(\hat{\theta} \pm 1.96\,\text{SE})$. Tables indicate the offset used and the number of run clusters. We report pairwise differences only when BH–FDR $q < .05$ and $|\Delta\%| \geq 10$. In the main text, Table 5 summarizes large category shifts vs. within-element baselines and reports only categories with $p < .05$ and $|\Delta| \geq 50\%$.

E.4 Model Adequacy and Robustness Checks

Model adequacy. Across all Poisson GEE fits, dispersion diagnostics were below unity (elements: Pearson $\chi^2/\text{df} = 0.567$, deviance/df $= 0.611$; categories: $\chi^2/\text{df} = 0.402$–0.664, deviance/df $= 0.454$–0.740), with many run clusters (elements: $n = 2{,}880$; categories: $n = 2{,}793$–$2{,}880$; runs with $K_{\text{elem}} = 0$ excluded from the corresponding element-specific *category* models) and numerically identical results when adding the supply offset $\log n$; we therefore report run-clustered robust (sandwich) SEs and use an exchangeable working correlation (independence for *Style*), with no Generalized Linear Model (GLM) fallback.

E.5 Axis-Level Permutation Test and Axis Aggregation

Model & test. Within Experiment 2–2, we assess constraint-level over- or under-selection via a Monte Carlo permutation test stratified by model × instruction type × element × category. The null assumes exchangeability across constraints conditional on each run's selection budget K_u and pool composition. For each constraint c, we compute the observed total $Y_c = \sum_u y_{uc}$ and the supply-adjusted expectation $\mathbb{E}[Y_c] = \sum_u K_u\,(n_{c,u}/N_u)$. We report $\text{share}_{\text{obs}} = Y_c / \sum_u K_u$, $\text{share}_{\text{exp}} = \mathbb{E}[Y_c] / \sum_u K_u$, $\text{RD}_{\text{share}} = \text{share}_{\text{obs}} - \text{share}_{\text{exp}}$, and a smoothed observed-to-expected ratio $\text{Obs/Exp} = (Y_c + 0.5)/(\mathbb{E}[Y_c] + 0.5)$; direction is defined by $\text{share}_{\text{obs}}$ vs. $\text{share}_{\text{exp}}$.

Inference & reporting. Two-sided p-values (and one-sided $p_{\text{over}}, p_{\text{under}}$) come from B=2000 permutations with a +1 correction; multiplicity is controlled within stratum by BH–FDR on p_{two} (significance at $q \leq .10$; fallback $p_{\text{two}} \leq .05$ in degenerate strata). Axis-level summaries map constraint signals to axis annotations and aggregate them on a shared observed/expected scale. For the *global axis baseline*, we pool $(Y_c, \mathbb{E}[Y_c])$ across all strata and then aggregate to axes, reporting observed and expected axis shares and their ratio. For *instruction-type contrasts*, we restrict to significant constraints within each stratum, propagate their over/under direction to axes, and compute within-direction axis shares along with enrichment ratios defined as $\text{Share}/\text{Global}$, where Global denotes the pooled baseline share within the same direction (pooled across instruction types and models). Summary tables apply minimum-support thresholds (MIN_SUPPORT) and Top-K/union rules for visualization. We set MIN_SUPPORT = 3, reporting an axis-direction only when it is supported by at least three distinct significant constraints, to reduce single-constraint artifacts and stabilize the descriptive summaries.

F Element-Level Selection Rate Ratios Relative to *Event*

	Offset=log K		Offset=log $K + \log n_{\text{items}}$	
Element	**RR [95% CI]**	p	**RR [95% CI]**	p
Style	1.67 [1.57, 1.79]	< .001	1.67 [1.57, 1.79]	< .001
Character	1.10 [1.02, 1.17]	.010	1.10 [1.02, 1.17]	.010
Setting	1.05 [0.98, 1.13]	.147	1.05 [0.98, 1.13]	.147

Element-level selection rate ratios relative to *Event* (pooled across models and prompts), shown side-by-side for two offsets ($\log K$ and $\log K + \log n_{\text{items}}$), where n_{items} is the number of available constraints per element per run (a specification robustness check). *N runs* $= 2,880$ for all elements.

G Category-Level Selection Rate Ratios vs. Within-Element Baseline Category

		Offset=log(K_{elem})		Offset=log(K_{elem}) + log(n_{items})		
Element	**Category**	**RR [95% CI]**	p	**RR [95% CI]**	p	**N runs**
Event	*Diffusion* (baseline)	1.00 [1.00, 1.00]	-	1.00 [1.00, 1.00]	-	2880
	Disruption	0.71 [0.61, 0.84]	< .001	0.71 [0.61, 0.84]	< .001	2880
	Epistemological Transformation	1.65 [1.45, 1.88]	< .001	1.65 [1.45, 1.88]	< .001	2880
	Relational Realignment	1.13 [0.97, 1.33]	.126	1.13 [0.97, 1.33]	.126	2880
	Reorientation	1.22 [1.06, 1.40]	.007	1.22 [1.06, 1.40]	.007	2880
Style	*Narrative perspective* (baseline)	1.00 [1.00, 1.00]	-	1.00 [1.00, 1.00]	-	2880
	Syntax & Sentence Structure	0.71 [0.65, 0.77]	< .001	0.71 [0.65, 0.77]	< .001	2880
	Temporal Structure	0.75 [0.70, 0.80]	< .001	0.75 [0.70, 0.80]	< .001	2880
	Tone & Mood	1.88 [1.77, 1.99]	< .001	1.88 [1.77, 1.99]	< .001	2880
	Write like X	0.32 [0.27, 0.38]	< .001	0.32 [0.27, 0.38]	< .001	2880
Character	*Cultural Identity* (baseline)	1.00 [1.00, 1.00]	-	1.00 [1.00, 1.00]	-	2880
	Embodied Difference	0.66 [0.56, 0.79]	< .001	0.66 [0.56, 0.79]	< .001	2880
	Motive	2.87 [2.56, 3.21]	< .001	2.87 [2.56, 3.21]	< .001	2880
	Relational Identity	1.58 [1.39, 1.81]	< .001	1.58 [1.39, 1.81]	< .001	2880
	Social Status	0.99 [0.85, 1.15]	.879	0.99 [0.85, 1.15]	.879	2880
Setting	*Cultural context* (baseline)	1.00 [1.00, 1.00]	-	1.00 [1.00, 1.00]	-	2880
	Macro spatial setting	2.21 [1.95, 2.50]	< .001	2.21 [1.95, 2.50]	< .001	2880
	Micro spatial setting	1.16 [1.01, 1.33]	.031	1.16 [1.01, 1.33]	.031	2880
	Socio-political order	0.96 [0.81, 1.14]	.643	0.96 [0.81, 1.14]	.643	2880
	Temporal setting	1.79 [1.58, 2.02]	< .001	1.79 [1.58, 2.02]	< .001	2880

Category-level selection rate ratios vs. the within-element baseline category (pooled across models/prompts), shown side-by-side for two offsets (Offset=log K_{elem} and Offset=log $K_{\text{elem}} + \log n_{\text{items}}$). Runs with $K_{\text{elem}} = 0$ are excluded from the corresponding element-specific category models; run clusters therefore vary by element (2,793–2,880).

H Top-15 Globally Over- or Under-selected Axes

Category	Axis	Obs	Obs/Exp (×)	Obs (%)	Exp (%)
Global — Over-selected					
Reorientation	*Connected (E4)*	324	3.89	0.56	0.14
Macro spatial setting	*Urban Built Environments*	508	3.66	0.88	0.24
Temporal setting	*The Fully Connected Now*	523	2.74	0.91	0.33
Micro spatial setting	*Dreamlike or Surreal Chambers*	357	2.52	0.62	0.25
Cultural context	*Theatrical Order*	611	2.36	1.06	0.45
Reorientation	*Connected (E1)*	189	2.27	0.33	0.14
Syntax & Sentence Structure	*Balanced*	488	2.07	0.85	0.41
Reorientation	*Positive*	684	2.06	1.19	0.58
Tone & Mood	*Authorial (Virginia Woolf)*	578	1.99	1.00	0.50
Tone & Mood	*Authorial (James Baldwin)*	834	1.92	1.45	0.75
Temporal setting	*The Broken Sequence*	352	1.85	0.61	0.33
Micro spatial setting	*Domestic Interior Spaces*	250	1.76	0.43	0.25
Embodied Difference	*Race-Marked*	206	1.69	0.36	0.21
Temporal setting	*Worlds in Acceleration*	322	1.69	0.56	0.33
Disruption	*Technological*	235	1.59	0.41	0.26
Global — Under-selected					
Temporal setting	*Age of Origins*	32	0.17	0.06	0.33
Cultural context	*Theistic*	45	0.18	0.08	0.45
Temporal Structure	*Future*	47	0.19	0.08	0.43
Reorientation	*Connected (E8)*	17	0.21	0.03	0.14
Temporal setting	*Worlds of Faith and Rule*	41	0.22	0.07	0.33
Reorientation	*Neutral*	40	0.24	0.07	0.29
Macro spatial setting	*Mountainous Regions*	34	0.25	0.06	0.24
Cultural Identity	*Legible*	69	0.25	0.12	0.48
Cultural context	*Unquestioned Precedent*	67	0.26	0.12	0.45
Macro spatial setting	*Polar and Glacial Frontiers*	36	0.26	0.06	0.24
Macro spatial setting	*Deserts and Arid Zones*	40	0.29	0.07	0.24
Temporal setting	*The Shattered Century*	60	0.31	0.10	0.33
Reorientation	*Negative*	105	0.32	0.18	0.58
Relational Identity	*Competitive*	146	0.33	0.25	0.77
Micro spatial setting	*Mythic or Enchanted Structures*	47	0.33	0.08	0.25

Top-15 globally *over*- or *under*-selected axes across all runs (descriptive; no significance filtering). Obs and Exp are aggregated observed and supply-adjusted expected counts; Obs/Exp $= (\mathrm{Obs}+0.5)/(\mathrm{Exp}+0.5)$. Obs (%) and Exp (%) report observed and expected axis shares. Values rounded.

I Top-5 Over- or Under-selected Axes by Instruction Type

Category	Axis	Support	Enrich (×)	Share (%)	Global (%)
Basic — Over-selected					
Embodied Difference	*Disability-Marked*	3	3.23	2.52	0.78
Micro spatial setting	*Domestic Interior Spaces*	4	3.20	2.50	0.78
Tone & Mood	*Authorial (James Baldwin)*	5	3.15	2.46	0.78
Tone & Mood	*Authorial (Virginia Woolf)*	5	3.15	2.46	0.78
Temporal setting	*Realistic*	12	3.12	4.88	1.56
Basic — Under-selected					
Write like X	*East Asian*	3	3.42	3.80	1.11
Write like X	*Male*	4	3.04	5.06	1.67
Tone & Mood	*Authorial (Chimamanda Adichie)*	4	2.66	1.48	0.56
Temporal setting	*The Dreamtime*	5	2.28	1.27	0.56
Tone & Mood	*Authorial (Haruki Murakami)*	5	2.28	1.27	0.56
Quality — Over-selected					
Reorientation	*Not Connected*	3	3.69	2.88	0.78
Tone & Mood	*Authorial (James Baldwin)*	5	3.37	2.63	0.78
Tone & Mood	*Authorial (Virginia Woolf)*	5	3.37	2.63	0.78
Macro spatial setting	*Aquatic and Coastal Environments*	4	3.32	2.60	0.78
Micro spatial setting	*Domestic Interior Spaces*	4	3.28	2.56	0.78
Quality — Under-selected					
Write like X	*East Asian*	4	4.50	5.00	1.11
Write like X	*Modernist-Postmodernist*	3	3.38	3.75	1.11
Write like X	*Male*	4	3.00	5.00	1.67
Tone & Mood	*Authorial (Chimamanda Adichie)*	4	2.51	1.39	0.56
Write like X	*Euro-American*	3	2.25	3.75	1.67
Creativity — Over-selected					
Macro spatial setting	*Extraterrestrial Terrain*	4	3.22	2.52	0.78
Macro spatial setting	*Otherworldly or Mythic Realms*	4	3.22	2.52	0.78
Cultural context	*Volatile Norms*	5	3.11	2.43	0.78
Reorientation	*Connected* (E9)	5	3.11	2.43	0.78
Temporal setting	*The Dreamtime*	5	3.11	2.43	0.78
Creativity — Under-selected					
Write like X	*Global South*	3	3.10	1.72	0.56
Write like X	*Realist*	11	2.84	6.32	2.22
Write like X	*Queer*	3	2.84	3.16	1.11
Write like X	*Male*	8	2.76	4.60	1.67
Cultural context	*Collectivist*	5	2.41	1.34	0.56

Top-5 Over- or Under-selected axes by instruction type. Support = pooled count of significant constraints mapped to the axis (summed across models). Enrich = Share / Global, where Share = axis share within direction and Global = pooled baseline share within the same direction (computed over significant constraints, pooled across instruction types and models). Values rounded.

J Model-Stratified Axis Summaries

J.1 Axes Shared Across Models

Category	Axis	claude	gemini	gpt4.1	gpt5	o4mini	qwen
Motive	Conscious	↑	↑	↑↓	↑	↑	↑↓
Motive	Constructive	↑	↑	↑	↑	↑↓	↑↓
Temporal setting	Realistic	↑	↑	↑↓	↑	↑	↑

Axes common to all models (Top-20 per model via over/under union; axes with a uniform direction across models excluded). Cells mark per-model direction: ↑=over-selected, ↓=under-selected. A paired mark (↑ ↓) indicates that the direction flips across instruction types (instruction-contingent).

J.2 Top-5 Over- or Under-selected Axes by Model (claude, gemini, gpt4.1)

Category	Axis	Support	Enrich (×)	Share (%)	Global (%)
claude — Over-selected					
Temporal setting	*Realistic*	4	3.61	5.63	1.56
Reorientation	*Connected (E4)*	3	3.37	2.63	0.78
Tone & Mood	*Authorial (James Baldwin)*	3	3.37	2.63	0.78
Tone & Mood	*Authorial (Virginia Woolf)*	3	3.37	2.63	0.78
Narrative perspective	*Unreliable*	4	2.98	9.30	3.12
claude — Under-selected					
Reorientation	*Neutral*	6	2.02	2.25	1.11
Macro spatial setting	*Deserts and Arid Zones*	3	2.02	1.12	0.56
Macro spatial setting	*Mountainous Regions*	3	2.02	1.12	0.56
Macro spatial setting	*Polar and Glacial Frontiers*	3	2.02	1.12	0.56
Micro spatial setting	*Sacred Grounds*	3	2.02	1.12	0.56
gemini — Over-selected					
Embodied Difference	*Stigmatized*	4	2.84	4.44	1.56
Temporal setting	*Realistic*	4	2.84	4.44	1.56
Cultural context	*Theatrical Order*	3	2.80	2.19	0.78
Micro spatial setting	*Dreamlike or Surreal Chambers*	3	2.80	2.19	0.78
Relational Identity	*Defensive*	3	2.80	2.19	0.78
gemini — Under-selected					
Cultural Identity	*Legible*	12	1.77	3.93	2.22
Relational Identity	*Competitive*	12	1.77	3.93	2.22
Reorientation	*Neutral*	6	1.77	1.97	1.11
Cultural context	*Atheistic*	3	1.77	0.98	0.56
Cultural context	*Collectivist*	3	1.77	0.98	0.56
gpt4.1 — Over-selected					
Temporal setting	*Realistic*	4	3.82	5.97	1.56
Reorientation	*Connected (E1)*	3	3.66	2.86	0.78
Reorientation	*Connected (E4)*	3	3.66	2.86	0.78
Tone & Mood	*Authorial (James Baldwin)*	3	3.66	2.86	0.78
Tone & Mood	*Authorial (Virginia Woolf)*	3	3.66	2.86	0.78
gpt4.1 — Under-selected					
Temporal Structure	*Future*	6	3.94	4.38	1.11
Reorientation	*Connected (E6)*	3	3.94	2.19	0.56
Temporal setting	*The Cosmic Scale*	3	3.94	2.19	0.56
Tone & Mood	*Authorial (Chimamanda Adichie)*	3	3.94	2.19	0.56
Tone & Mood	*External*	11	3.61	8.03	2.22

Top-5 Over- or Under-selected axes by model (claude, gemini, gpt4.1). Support = pooled count of significant constraints mapped to the axis (summed across instruction types). Enrich = Share / Global, where Share = axis share within direction and Global = pooled baseline share within the same direction (computed over significant constraints, pooled across instruction types and models). Values rounded.

J.3 Top-5 Over- or Under-selected Axes by Model (gpt5, o4mini, qwen)

Category	Axis	Support	Enrich (×)	Share (%)	Global (%)
gpt5 — Over-selected					
Macro spatial setting	*Urban Built Environments*	3	2.82	2.21	0.78
Micro spatial setting	*Transit Hubs*	3	2.82	2.21	0.78
Relational Identity	*Defensive*	3	2.82	2.21	0.78
Reorientation	*Connected (E1)*	3	2.82	2.21	0.78
Reorientation	*Connected (E4)*	3	2.82	2.21	0.78
gpt5 — Under-selected					
Embodied Difference	*Accepted*	4	1.70	3.77	2.22
Disruption	*Natural*	9	1.58	2.63	1.67
Motive	*Constructive*	9	1.58	2.63	1.67
Motive	*Focused*	9	1.58	2.63	1.67
Cultural Identity	*Legible*	12	1.58	3.51	2.22
o4mini — Over-selected					
Disruption	*Human*	4	2.88	4.49	1.56
Socio-political order	*Centralized*	9	2.82	6.62	2.34
Embodied Difference	*Race-Marked*	6	2.82	4.41	1.56
Temporal setting	*Realistic*	6	2.82	4.41	1.56
Macro spatial setting	*Urban Built Environments*	3	2.82	2.21	0.78
o4mini — Under-selected					
Motive	*Constructive*	9	1.83	3.05	1.67
Motive	*Focused*	9	1.83	3.05	1.67
Cultural Identity	*Legible*	12	1.83	4.07	2.22
Epistemological Transformation	*Gradual*	12	1.83	4.07	2.22
Epistemological Transformation	*Internal*	12	1.83	4.07	2.22
qwen — Over-selected					
Motive	*Constructive*	9	3.59	8.41	2.34
Write like X	*Male*	6	3.59	5.61	1.56
Write like X	*Speculative*	6	3.59	5.61	1.56
Embodied Difference	*Disability-Marked*	3	3.59	2.80	0.78
Reorientation	*Connected (E4)*	3	3.59	2.80	0.78
qwen — Under-selected					
Temporal Structure	*Future*	6	4.62	5.13	1.11
Reorientation	*Connected (E8)*	3	4.62	2.56	0.56
Temporal setting	*Age of Origins*	3	4.62	2.56	0.56
Tone & Mood	*Authorial (Chimamanda Adichie)*	3	4.62	2.56	0.56
Write like X	*East Asian*	5	3.85	4.27	1.11

Top-5 Over- or Under-selected axes by model (gpt5, o4mini, qwen). Support = pooled count of significant constraints mapped to the axis (summed across instruction types). Enrich = Share / Global, where Share = axis share within direction and Global = pooled baseline share within the same direction (computed over significant constraints, pooled across instruction types and models). Values rounded.