StoryScope: Investigating idiosyncrasies in AI fiction

Jenna Russell, **Rishanth Rajendhran**, **Mohit Iyyer**, **John Wieting**
University of Maryland, College Park, Google DeepMind
{jennarus, rishanth, miyyer}@umd.edu, jwieting@google.com

Abstract

As AI-generated fiction becomes increasingly prevalent, questions of authorship and originality are becoming central to how written work is evaluated. While most existing work in this space focuses on identifying surface-level signatures of AI writing (e.g., word choice, syntactic structure), we ask instead whether AI-generated stories can be distinguished from human ones *without relying on stylistic signals*, focusing on discourse-level narrative choices such as character agency and chronological discontinuity. We propose StoryScope, a pipeline that automatically induces a fine-grained, interpretable feature space of discourse-level narrative features across 10 dimensions (e.g., plot, agents, temporal structure). We apply StoryScope to a parallel corpus of 10,272 writing prompts, each written by a human author and five LLMs (Claude, DeepSeek, Gemini, GPT, and Kimi), yielding 61,608 stories, each ~5,000 words, and 304 extracted features per story. Narrative features alone achieve 93.2% macro-F1 for human vs. AI detection and 68.4% macro-F1 for six-way authorship attribution, retaining over 97% of the performance of models that include stylistic cues. A compact set of 30 *core* narrative features captures much of this signal: AI stories over-explain themes and favor tidy, single-track plots while human stories frame protagonists' choices as more morally ambiguous and have increased temporal complexity (e.g., flashbacks, nonlinear structure). Per-model *fingerprint* features enable six-way attribution: for example, Claude produces notably flat event escalation, GPT likes using gossip as a plot mechanism, and Gemini defaults to external character description. We find that AI-generated stories cluster in a shared region of narrative space, while human-authored stories exhibit greater diversity. More broadly, these results suggest that differences in underlying narrative construction, not just writing style, can be used to separate human-written original works from AI-generated fiction. We release the StoryScope code, 10,272 writing prompts, 51,336 AI-generated narratives, and narrative features for each story to support future work on narrative analysis and AI authorship.[1]

1 Introduction

AI fiction is already under our noses. In March 2026, Hachette, a major publishing house, pulled the horror novel *Shy Girl* after it was flagged as ∼78% AI-generated, the first commercially published novel canceled over AI allegations. Nearly 20% of a sample of 14,000 self-published Amazon novels were flagged by Pangram (Emi & Spero, 2024) as largely AI-generated, a figure that jumped 41% year-over-year.[2] Overall, readers are increasingly being misled into purchasing AI-generated books attributed to human authors. If authors are unwilling to self-disclose AI usage, how can we address this issue?

At first glance, this appears to be a detection problem: can we determine whether a given story was written by human or machine? Existing AI detectors (Hans et al., 2024; Adam et al., 2026; Thai et al., 2026) primarily rely on stylistic signals such as word choice and sentence

[1] Code and data: https://github.com/jenna-russell/storyscope
[2] https://www.nytimes.com/2026/03/19/books/ai-fiction-shy-girl.html

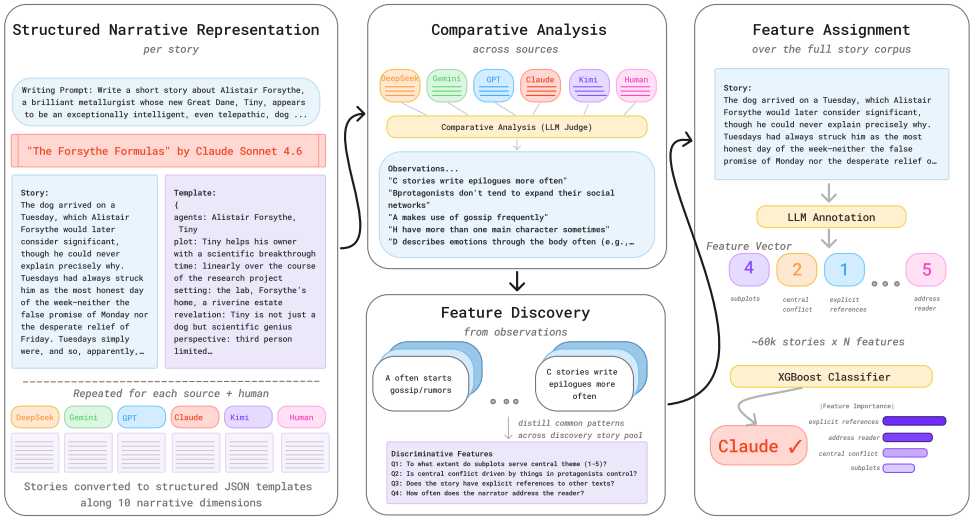

Figure 1: Overview of the StoryScope pipeline. Stories are converted into structured templates, then compared across sources writing to the same prompt to induce discriminative narrative features, and finally featurized across the full corpus for downstream detection and authorship experiments. Story inspired by "Tiny and the Monster" (Sturgeon, 1983).

structure, and for good reason: these cues are highly discriminatory. AI-generated text systematically overuses em-dashes, words like "delve" and "tapestry," and other surface-level patterns that even simple classifiers detect reliably (Sun et al., 2025; Shaib et al., 2026). That said, AI style is increasingly fleeting: GPT-5.4 significantly reduced em-dash usage,[3] and fine-tuning to mimic human style drops AI detection rates on creative writing from 97% to 3% (Chakrabarty et al., 2026). Discourse-level narrative features (e.g., plot structure, character agency, information revelation), which we refer to simply as *narrative features* throughout, are far harder to "humanize," as changing them requires significant structural rewrites rather than simple post-hoc edits (Namuduri et al., 2025).

As AI seeps into the writing industry, the question of *what constitutes original work* shifts from how a story is written to how it is conceived. Settled U.S. legal precedent requires that protected works show a minimal degree of originality (U.S. Supreme Court, 1991); recent guidance from the U.S. Copyright Office clarifies that eligibility depends on sufficient *human* creative control (U.S. Copyright Office, 2023). To measure this, we use statistical rarity in a feature space of narrative decisions as a proxy for originality, where less common combinations reflect the broader notion of originality invoked by Torrance (1966) and copyright law. We hypothesize that humans and AI models make systematically distinct narrative choices, and that these differences persist even when stylistic cues are removed.

To test this, we introduce StoryScope, a pipeline that automatically induces interpretable narrative features grounded in the NarraBench taxonomy (Hamilton et al., 2025), spanning 10 dimensions of narrative structure (character, plot, setting, temporal structure, etc.). Applied to a parallel corpus of 10,272 prompts each written by a human author and five LLMs, StoryScope yields 304 features and 61,608 featurized stories on which we train classifiers. Our stories average roughly 5,000 words depending on source, enabling extraction of fine-grained narrative features that shorter texts cannot support.

The narrative features alone, withholding all stylistic features (e.g., sentence-level rhythm, figurative language density), achieve a macro-F1 of 93.2% for the binary human vs. AI detection task, retaining over 97% of the performance of a model with both narrative and stylistic features. In the six-way authorship attribution task (identifying which source wrote

[3] https://x.com/sama/status/1989193813043069219?s=20

a story), narrative features get a macro-F1 of 68.4% without style, compared to 77.3% with style. When we represent each story as a vector of narrative features, the five AI models occupy a tight cluster that is well-separated from human stories, showing that AI models have converged on a shared narrative space that is systematically separated from human storytelling, and that these changes remain after editing stories for style (macro-F1 93.9%). Human stories are, on average, rarer in narrative feature space (mean rarity percentile 0.71 vs. 0.49 for AI). Each model also exhibits a unique narrative "fingerprint": a set of features on which it diverges from the other AI sources and enables fine-grained attribution.

Our results suggest that the narrative choices underlying AI-generated fiction are distinguishable from those of human authors, even when surface style is removed. Because these features reflect structural decisions rather than lexical ones, they may prove more durable as models continue to evolve. We release the StoryScope code, 10,272 writing prompts, 51,336 AI-generated narratives, and narrative features for each story, generated at the cost of $4.4k USD, to support future work on AI authorship and narrative analysis.[4]

2 The StoryScope pipeline

In this section, we first describe the creation of a dataset that enables comparative analysis of narrative choices, before detailing our pipeline's operation and post-hoc interpretation.

Data. Our dataset consists of human-written short stories paired with multiple AI-generated mirrors with similar plots and characters. We extract 10,272 human-written stories from Books3 (Presser, 2020).[5] To generate mirrored AI stories, we reverse-engineer writing prompts from each human story by prompting Gemini 2.5 Flash (Gemini Team, 2025a) to infer the underlying premise (Li et al., 2024).[6] We generate stories from five LLMs: Gemini 3 Flash (Gemini Team, 2025b), Kimi K2.5 (Kimi Team, 2026), DeepSeek V3.2 (DeepSeek-AI, 2025), Claude Sonnet 4.6 (Anthropic, 2026), and GPT-5.4 (OpenAI, 2026). Together with the human-authored story, this yields six *sources*, a term we use throughout to refer to each origin uniformly.[7]

2.1 StoryScope

Manually annotating the narrative decisions made within these 60K+ stories (averaging 4,753 words) is infeasible at scale.[8] Instead, we leverage LLMs to perform structured analysis over our corpus of stories to create interpretable narrative features. The StoryScope pipeline consists of three stages (Figure 1): (1) structured intermediate representations, (2) cross-source comparison, and (3) discourse-level feature discovery, guided by the narrative taxonomy introduced in NarraBench (Hamilton et al., 2025). Templates convert prose into structured narrative fields, forcing downstream stages to reason over narrative content rather than stylistic surface. The comparative analysis stage distills 600 stories' worth of structured representations into cross-source observations (compressing ∼2.7M tokens of raw text into ∼686K tokens) so that feature discovery can generalize from already-identified patterns rather than simultaneously reading, comparing, and formalizing.

Theoretical grounding. We transform each story into a template-based representation that preserves narrative elements while abstracting away surface wording, grounding this representation in NarraBench's taxonomy of narrative dimensions rooted in literary

[4]We do not release human-written stories due to copyright concerns.

[5]We use Books3 *strictly* for academic purposes to understand the narrative differences in recent human-written and AI-generated text and inform discussions on AI-detection, authorship, and copyright policy. Dataset statistics found in Table 4.

[6]Full prompt-extraction template in Figure 6. Gemini 2.5 Flash was used during writing prompt construction in June 2025. We identify and extract stories from short story anthologies identified via running Gemini 2.5 Flash over titles and blurbs in Books3.

[7]All source identities are anonymized in LLM-facing prompts (see §B).

[8]Across the 61,608-story corpus, mean length is 4,753 words. Models refused to generate 24 stories.

#	Feature	Question	Dim.	Type	Response Options
AI					
1	Thematic Explicitness and Moralizing	How explicitly does the story articulate its themes or morals?	SIT	scale	1–5
2	Agency in Resolution → protagonist choice	Is resolution driven by protagonist's choices or external events?	PLT	cat	protagonist_choice, mixed, external_fate
3	Narratorial Thematic Commentary → yes	Does the narrator explicitly comment on themes beyond characters' perspectives?	SIT	binary	no, yes
Human					
1	Intertextual Strategy Types → explicit named reference	What kinds of intertextual engagement does the story employ?	SIT	multi	explicit named, retelling, pastiche, myth/religion, self referential
2	Depth of Recontextualization After Surprise	How extensively does a revelation force reinterpretation of earlier scenes?	REV	scale	1 (none)–5 (complete rereading)
3	Degree of Chronological Discontinuity	How often does the narrative jump across time?	TMP	scale	1–5

Table 1: Examples of core features, see AI core features in Table 14, human in Table 15, and all 30 with human/AI means in Table 16. NarraBench dimensions include situatedness (SIT), plot (PLT), revelation (REV) and temporal structure (TMP).

theory. We adopt ten of its twelve aspects: *Agent, Social Network, Event, Plot, Structure, Setting, Time, Revelation, Perspective*, and *Style*.[9] Each dimension consolidates an established line of narratological work (see Table 5 of Hamilton et al., 2025). StoryScope searches *within* those dimensions for the specific decisions that separate sources. For example, its temporal dimension reflects long-standing narratological distinctions around narrative order, duration, and frequency from Genette (1980), while plot arc is heavily inspired by Tian et al. (2024).

Structured narrative representations. For each story, we prompt GPT-5.1 (OpenAI, 2025) to extract a structured template organized along these ten dimensions. The extraction uses a zero-shot prompt with a detailed JSON schema specifying the expected fields for each dimension (e.g., character names, roles, and motivations for Agent; causal chains and key events for Event).[10] The templates serve as a controlled intermediate representation (Pham et al., 2024; Wang et al., 2023): by converting prose into structured fields, the subsequent comparison stage reasons over narrative content rather than stylistic surface (see §B for details).

Comparative story analysis across sources. To identify systematic narrative differences, we perform pairwise comparisons across sources writing from the same prompt. We construct a discovery pool of 600 stories (and their corresponding structured representations) over 100 parallel randomly selected writing prompts. The discovery pool is held out from the main corpus and kept small because this stage uses GPT-5.1 with high reasoning effort over long inputs; 600 stories balance narrative diversity against cost. For each prompt, we present all six templates to GPT-5.1 and ask it to produce a structured comparative analysis: a structured JSON containing per-source dimension notes (e.g., how each source handles character motivation), cross-source comparisons highlighting where sources diverge, and an executive summary of recurring patterns.[11] The comparative analyses serve as the raw material from which we derive our feature taxonomy.

Feature discovery. We extract interpretable narrative features from the comparative analyses by prompting GPT-5.1 to propose discriminative features within each NarraBench dimension.[12] Each expert prompt is grounded in a specific NarraBench aspect and instructs the model to propose features as closed-form questions with discrete answer choices. We

[9]We exclude *Paratext* and *Motivation*, which depends on external context unavailable in our setting.
[10]Templating prompt depicted in Figure 8.
[11]We batch multiple prompts per analysis (mean 3.1), details in §B.
[12]We use 10 specialized expert prompts, one per NarraBench dimension (see §C).

constrain features to five response types (categorical, ordinal, scale, binary, and multi-select) to support interpretable downstream modeling. To improve coverage, we run the discovery process three times and take the union of all proposed features, yielding 408 candidates. We then deduplicate via embedding-based clustering, retaining the feature nearest each cluster centroid, resulting in $d = 304$ features after merging 65 clusters.[13] Each feature is defined as a specific axis of narrative variation (e.g., the degree to which character motivation is stated explicitly versus left implicit, or the extent to which a story's timeline departs from chronological order). The final taxonomy spans five response types: categorical (124), ordinal (59), scale (45), binary (44), and multi-select (32).

2.2 STORYSCOPE interpretability

Feature assignment. We apply the features to the full dataset of 61,608 stories across all 10,272 prompts. For each story, we present the full text along with the relevant feature definitions to Gemini 3 Flash (with minimal thinking) and obtain a value for each of the 304 features,[14] yielding a raw narrative-assignment vector $\mathbf{z} = (z_1, \ldots, z_d)$ with $d = 304$ semantic features for every story in the corpus. The classifier does not consume $\mathbf{z}$ directly: after encoding each feature according to its type (one-hot for categorical, multi-hot for multi-select, numeric for ordinal and scale, binary as-is), we obtain an encoded input vector $\mathbf{x} \in \mathbb{R}^D$, where D depends on the feature subset and the expansion induced by categorical and multi-select features.

Defining core and fingerprint features. Not all features contribute equally to distinguishing sources. To identify which features are robust human-AI markers versus source-specific cues, we train XGBoost classifiers (Chen & Guestrin, 2016). Gradient-boosted trees pair naturally with SHAP (Lundberg & Lee, 2017) to yield exact, per-feature importance decompositions, letting us trace every prediction back to specific narrative decisions. We assess importance and stability through bootstrap SHAP analysis (B=50 iterations with prompt-level resampling) and assign each feature one of three roles:[15] (1) **Core features:** stable and important in the *binary* human-vs-AI task, with a strong, consistent separation that holds across all five AI models (examples in Table 1). Each is signed as human-leaning or AI-leaning based on the direction of the gap; (2) **Fingerprint features:** identified from the six-way *multiclass* task. A feature qualifies when its SHAP importance is concentrated in a single source class, and that source's observed values visibly differ from others; (3) Features satisfying neither set of criteria are excluded from role-based analyses.

3 Experiments

Classification. We train XGBoost classifiers on encoded feature vectors $\mathbf{x}$ over 52,434 stories (8,739 prompts) for two tasks: (1) binary detection (human vs. AI) and (2) six-way authorship attribution (human, Claude, GPT, DeepSeek, Gemini, Kimi), evaluated on a held-out test set of 1,377 prompts (8,262 stories) with macro-F1 and AUPRC as the primary binary metrics.[16] Nominal features are one-hot encoded, multi-select features are multi-hot encoded, and ordinal / scale features retain numeric encoding. For the full Narrative+Style this yields $D = 1108$ encoded columns; the Narrative variant uses $D = 958$, and Style Only uses $D = 129$. All evaluations use prompt-level grouping to prevent train/test leakage.

Feature variants. A central question of this work is whether narrative *choices* carry discriminative signal independent of surface style. To test this, we evaluate the following:

[13] Each feature is encoded with F2LLM-4B (Zhang et al., 2025b) and clustered at cosine similarity threshold 0.85, details in §B.

[14] Gemini 3 Flash has high repeatability over 5 independent runs (Krippendorff's $\alpha = 0.90$); human validation on 240 story-feature items has a mean human–model Cohen's $\kappa = 0.84$ (details in §C).

[15] This procedure yields 30 core features and 75 fingerprint features. Full selection criteria, equations, and numerical thresholds are detailed in §D.

[16] Hyperparameters are selected via grid search on the validation split, *Binary*: n_{est}=420, depth=8, λ=2.0, and 5:1 positive class weight; *six-way*: n_{est}=500, depth=7, λ=1.0, evaluation details in §D.

Narrative (257 features) includes all narrative features across nine NarraBench dimensions, excluding all 39 features in the style dimension plus 8 features from other dimensions flagged as style-related, 47 in total.[17] This is our primary model; it tests whether narrative structure alone suffices for the detection and attribution tasks. **Style Only** (39 features) uses only the style features. **Narrative + Style** (304 features) uses all features across all 10 NarraBench dimensions. **Core Only** (30 features) uses the subset of narrative features identified as universally important and stable (§D). **Core + Fingerprint** (101 features) adds source-specific fingerprint features to the core set.

Baselines. We compare against four text-based baselines that operate on raw story text rather than extracted features: (1) ModernBERT (Warner et al., 2025), a transformer fine-tuned on our training set (max length 512, 3 epochs); (2) Stylometric+XGB, XGBoost on 144 hand-crafted stylometric features (character n-grams, POS distributions, readability scores); (3) TF-IDF+XGB, XGBoost on 5,000 TF-IDF features; and (4) Binoculars (Hans et al., 2024), a zero-shot AI-text detector. These baselines establish an upper bound on raw-text detection; we are interested in understanding how closely our Narrative model performs.

4 Detecting AI from narrative features

We evaluate whether narrative features can separate human-written from AI-generated fiction. Narrative features alone achieve a macro-F1 of 93.2%, just 2.8 points below the model with both style and narrative features in macro-F1 (95% prompt-bootstrap CI 2.09–3.54). A set of 30 *core* features retains about 91% of the narrative model's macro-F1 (84.8%). We ask if StoryScope can still detect AI narratives after stylistic alterations, finding it robust to edits.

Style signals are stronger but narrative features close the gap. Models built using only the style features achieve comparable performance to the core-only model (85.8% and 84.8% macro-F1 respectively; Table 2), confirming that style remains a strong detection cue, especially in long-context settings.[18] Most automatic AI detection models have higher performance on longer text (Bao et al., 2024; Xu et al., 2025b). The narrative model is unchanged with length: on a human-length-matched test subset, it scores 93.2% macro-F1 both before and after matching (see §G). While the strongest supervised text-based baselines achieve near-perfect separation (≥99.7% macro-F1, with ModernBERT at 99.9%), zero-shot Binoculars is much weaker (55.9% macro-F1). Our narrative+style model nearly matches the strongest baselines (96.0% macro-F1), with the added benefit of full interpretability.[19] Narrative features alone recover 97% of the combined model's macro-F1.

Method	Size	F1	AP
Narrative features (ours)			
Narrative	257	93.2	.959
Core Only	30	84.8	.828
Core+FP	101	91.1	.934
Style ablations			
Narr. + Style	304	96.0	.982
Style Only	39	85.8	.867
Text-based baselines			
ModernBERT	—	99.9	1.00
Stylometric+XGB	144	99.8	.999
TF-IDF+XGB	5,000	99.7	.999
Binoculars	—	55.9	.404
Edited Stories			
LAMP	—	93.9	.988

Table 2: Binary human vs. AI classification results. F1 = macro-F1 (%); AP = AUPRC.

Core features retain most of the performance. Based on just 30 core features, our core model retains substantial human vs. AI separation, reaching 84.8% macro-F1 and 0.828 AUPRC. Representative core-feature definitions were introduced earlier in Table 1; see §I for the complete lists. Adding fingerprint features raises performance to 91.1% macro-F1 and 0.934 AUPRC, recovering much of the gap to the full narrative model while using less than half of the features. Our results suggest that the human-AI boundary lives in a

[17] We flagged features using an LLM audit of a feature's reliance on stylistic cues, details in §B.

[18] See all feature ablations in §E.

[19] We also ask if model style is affected by memorization of famous stories, finding that after filtering out likely memorized stories performance remains the same (see §F).

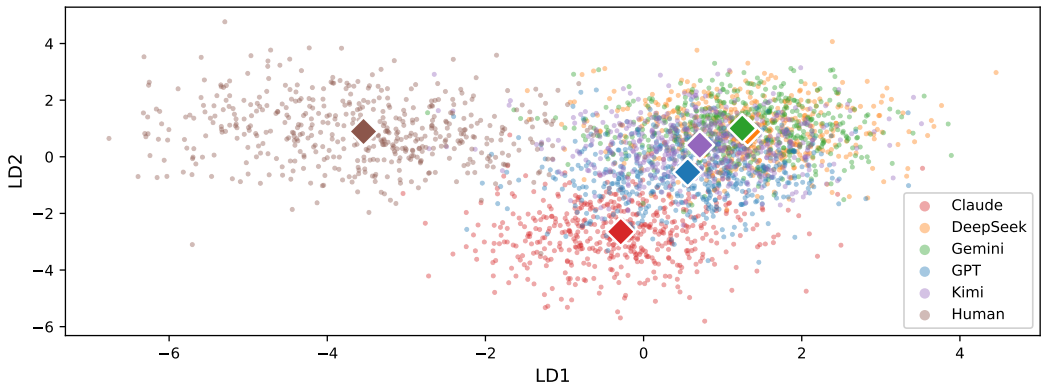

Figure 2: Projection of narrative feature vectors onto the first two linear discriminant components. Human writing occupies a distinct region; the five AI models cluster together. Claude is the most distinct of the five AI models, Gemini and DeepSeek the nearest neighbors.

compact subset of narrative decisions, while fingerprint features contribute model-specific separability.

4.1 What core narrative choices separate human from AI writing?

AI over-explains its themes. AI stories are more explicit and moralizing (roughly 20% higher on a 1-5 scale), with tighter thematic unity and more central moral questions. Narrators explicitly explain the story's theme 77% of the time, versus 52% for humans: a grieving character's arc will typically end with the narrator stating the lesson learned. AI dialogue serves philosophical debate more often (59% vs. 34%), and references to other works tend to be vague allusions (72% vs. 50%) rather than specific, named references. The pattern is one of over-determination: AI spells out meaning rather than trusting the reader to infer it.

Human authors subvert linearity. AI stories exhibit tighter causal chains, more protagonist-driven resolutions (69% vs. 46%), and far fewer subplots (79% "no subplots" vs. 57%). AI resolutions favor internal understanding or acceptance (47% vs. 27%), whereas humans are more comfortable with ambiguous endings. Humans use more time jumps, flashbacks and flash-forwards, and nonlinear structure to delay key revelations. AI favors single-track narratives with fewer loose ends; human stories are messier, with time jumps and disjointed causal chains—a human mystery might open at the funeral and spiral backward through decades, while AI tells the same story from first clue to the grand reveal.

AI over-writes the body and senses. AI overwhelmingly conveys emotion through physical sensations and bodily metaphors (81% vs. 38% human), deploys more smell-based imagery (82% vs. 57%), and uses setting as a reflection of characters' inner states more heavily. AI pays closer attention to physical environment and characters' inner mental states, showing emotion through bodies and environments rather than naming feelings directly. Where a human author might write that a character "felt afraid," AI renders fear as a tightening chest, cold sweat, and dimming lamplight. Humans use explicit emotion labels 29% of the time versus just 8% for AI.

Human authors engage the outside world. Humans reference specific texts and authors at nearly double the AI rate (47% vs. 24%) and balance explicit with implicit references more evenly (37% "balanced mix" vs. 16%), whereas AI generally sticks to vague allusions and avoids naming real brands, places, or works. Humans break the fourth wall far more often (67% vs. 39%) and address the reader directly more frequently (28% vs. 7%). Human writing acknowledges its audience as a co-participant (e.g., an aside to "you, dear reader"); AI writes as though no one is watching.

AI writing has less diverse narrative features. Human stories draw from a broader narrative repertoire: they span more locations, carry more dialogue relative to narration, integrate more subplots into overarching themes (42% vs. 21%), and present morally ambivalent protagonists more often (59% vs. 38%), resisting the pull to a narrow set of defaults.

4.2 Do narrative features still work after stylistic changes?

Style is the most obvious surface-level cue distinguishing AI generations from human writing, and style-based detectors are known to be brittle to paraphrasing and light editing (Saakyan et al., 2026). We ask whether narrative features are robust to such edits: if an AI-generated story is stylistically edited to remove surface artifacts, does our narrative classifier still detect it?

We test this using Chakrabarty et al. (2025)'s span-level rewriting framework (LAMP), which identifies and rewrites seven categories of AI writing artifacts (e.g., cliché, redundant exposition, purple prose) using 25 few-shot examples from professional writers. We apply this to 278 Gemini-generated stories, using Gemini itself as the rewriter, and evaluate our narrative classifier on these edited stories.

Surface editing barely affects narrative detection. After span-level artifact removal, the narrative model detects edited Gemini stories at 93.9% macro-F1 (AUPRC 0.988), compared to 95.5% macro-F1 (AUPRC 0.996) on the original unedited stories, a drop of only 1.6 points. This near-zero effect indicates that narrative features are largely orthogonal to the surface prose artifacts that LAMP targets: editing out clichéd phrasing or purple prose does not alter the structural narrative choices (causal linearity, thematic explicitness, sensory over-description) that drive our classifier.

5 Pinpointing each source's writing style

The binary task asks whether a story is human or AI; we now ask if we can use narrative features to detect exactly *which* source wrote a story. Using the same features and classifier, we train a six-way model over all six sources. Our narrative model reaches 68.4% macro-F1, indicating that narrative structure alone still carries weighty authorship signal, though the AI-authorship boundaries overlap much more than the binary human-AI boundary.

Narrative features retain substantial attribution signal. Table 3 shows macro-F1 and accuracy results for six-way attribution. Our narrative model achieves 68.4% macro-F1, well above the 16.7% chance baseline but below the narrative+style model at 77.3%. Narrative exceeds Style Only by 8.0 macro-F1 points (95% prompt-bootstrap CI 6.7–9.2), while Core+Fingerprint exceeds Core Only by 16.8 points (15.7–17.9), consistent with style and fingerprint features adding complementary attribution signal. Text-based baselines remain far stronger on raw text, all reporting macro-F1s of at least 99.5%.

Method	Size	F1	Acc.
Narrative features (ours)			
Narrative	257	68.4	68.4%
Core Only	30	46.5	46.8%
Core+FP	101	63.4	63.6%
Style ablations			
Narr. + Style	304	77.3	77.3%
Style Only	39	60.4	60.5%
Text-based baselines			
ModernBERT	—	99.8	99.8%
Stylometric+XGB	144	99.6	99.6%
TF-IDF+XGB	5,000	99.5	99.5%

Table 3: Six-way authorship attribution on the test set. F1 = macro-F1 (%). As AI models converge, the attribution task proves much more difficult than the binary detection task.

Claude and GPT are the most distinctive AI models. Table 12 breaks down per-class F1 for the six-way task. Human is the most distinctive source (93.0% F1 with style, 88.5% without), followed by Claude (89.3% / 77.1%) and GPT (82.1% / 73.0%). DeepSeek, Gemini, and Kimi form a more confused cluster (65.8–66.8% F1 with style). The gap widens in the narrative-only model: human and Claude retain strong separability while the bottom three models fall to 55.2–59.6%, suggesting that

narrative structure alone is less diagnostic for distinguishing among models that make similar storytelling choices.

AI convergence separates from human narratives. The five AI sources occupy overlapping regions of narrative feature space, separate from human authors (Figure 2). The six most-confused source pairs in our narrative-only six-way classifier are exclusively AI↔AI (Figure 3); the largest pair (Gemini↔DeepSeek, 222 and 207 stories) dwarfs the most common human misclassification (Human→Kimi, 46). Working in the z-scored encoded feature space with Euclidean distance, mean human-AI centroid distance is 1.6× the mean AI-AI centroid distance (6.6 vs. 4.3). Even the closest human-AI centroid pair is farther apart than the most distant AI-AI pair (6.2 vs. 6.0), indicating that human stories occupy a distinct region rather than merely a broader version of the AI cluster. Human stories are also more dispersed: their mean distance to the human centroid is 22% greater than the average AI radius (33.2 vs. 27.4) and their median 10-nearest-neighbor radius is 1.13× larger (33.1 vs. 29.2).

Human narratives are rarer. We measure per-story rarity as the mean Euclidean distance to a story's 25 nearest neighbors. Human stories have a higher mean rarity percentile than AI stories (0.71 vs. 0.49; Cohen's $d = 0.83$ indicating a substantial difference), and humans are overrepresented in the rarest tail: 24.7% of all human stories fall in the top 10% rarest stories corpus-wide, compared to just 7.1% of AI stories (top 1%: 3.0% human vs. 0.6% AI). At the prompt level, the human story is ranked the rarest of all six story versions 57.8% of the time (vs. 16.7% by chance; per-source distributions in §H).

Claude keeps it cool. Claude has the most distinctive narrative profile of the five LLMs. Its stories are defined by restraint: event intensity escalates less than in any other source, and narrative voice is the most uniform. Claude takes a reverent/continuist approach to literary tradition, honoring and extending storytelling conventions rather than subverting or challenging them (62% of Claude stories vs. 39–56% across other sources). It favors epilogues and avoids dream sequences, producing careful, consistent stories that favor quiet endings over "avalanche" endings.

GPT likes to gossip. GPT centers on socially-oriented storytelling: gossip and rumor as a plot mechanism (64% vs. 44–55% for other sources), a tendency to frame stories as reflections on events from years or decades ago, and ensemble-heavy social networks matching human levels. GPT subverts expectations more than other AI (41% vs. 27–36%) and leaves reconciliations ambiguous.

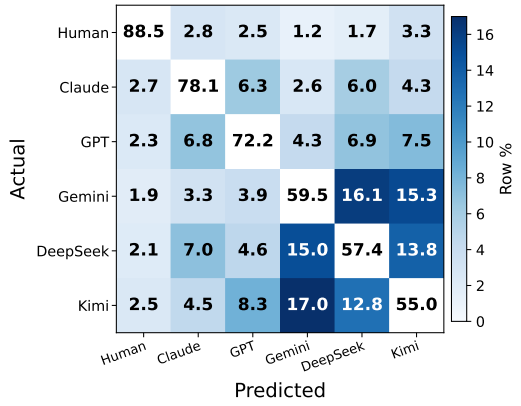

Figure 3: Confusion matrix for authorship attribution (narrative model) as a percentage (%). Misclassifications concentrate among AI models, particularly DeepSeek–Gemini–Kimi.

Gemini, DeepSeek, and Kimi are 'triplets'. Despite forming a more confused cluster, each still has individual quirks. DeepSeek front-loads crucial context that other sources leave until later. Gemini produces the tidiest endings, extended denouements, and the bleakest settings (88% tagged bleak and oppressive). Kimi has the fewest fingerprints and lowest F1, sitting at the generic center of the AI distribution with no distinctive narrative choices.

6 Related work

AI slop and detection. LLM text exhibits systematic lexical and syntactic patterns (Shaib et al., 2026; Kobak et al., 2025; Shaib et al., 2025), and parallel-corpus studies show LLMs converge on a narrow grammatical and rhetorical style distinct from human variation (Reinhart et al., 2025). Detection

methods range from stylometric and linguistic features (Li et al., 2025; O'Sullivan, 2025; Tripto et al., 2025) to ML classifiers on creative fiction (McGlinchey & Barclay, 2024; Najjar et al., 2025; McGovern et al., 2025; Huang et al., 2024), with expert humans also achieving near-perfect accuracy (Russell et al., 2025). Detecting partial AI edits to human text remains harder (Thai et al., 2026; Chen et al., 2025; He et al., 2025).

Originality and creativity. LLM outputs are less novel across n-gram (Padmakumar et al., 2026), psychometric (Chakrabarty et al., 2024; Koivisto & Grassini, 2023; Gómez-Rodríguez & Williams, 2024), and creativity-theoretic measures (Boden, 2004; Colton & Wiggins, 2012; Franceschelli & Musolesi, 2023). Narrative theory provides structural grounding for generation and understanding (Liu et al., 2026a): LLMs can extract discourse-level features such as POV, temporality, and focalization (Piper & Bagga, 2024) to better understand novelty (Namuduri et al., 2026), and prompt-based comparisons reveal that human stories contain more cultural nuance, emotional ambiguity, and unexpected twists than GPT-generated counterparts (Beguš, 2024). Network analyses reveal simpler social structures in AI fiction (Nonaka & Perry, 2025). Fine-tuning on literary corpora produces preferred text (Chakrabarty et al., 2026) but activates verbatim memorization (Liu et al., 2026b); constrained generation from human text fragments further blurs the authorship boundary (Pham et al., 2025); and injecting human-written stories as reward anchors during reinforcement learning improves small models' long-form story writing (Rajendhran et al., 2026). Across benchmarks (Hou et al., 2026; Paech, 2024; Zhang et al., 2025a; Saakyan et al., 2026; Wadhwa et al., 2026; Nguyen & Singla, 2025), LLMs consistently reduce collective diversity (Doshi & Hauser, 2024; Jiang et al., 2025), with repeated plot elements across generations (Xu et al., 2025a).

7 Conclusion

We introduce StoryScope, a pipeline for extracting interpretable narrative features at scale, and show that these features alone achieve 93.2% macro-F1 for human vs. AI detection and 68.4% for six-way attribution across 61,608 stories, retaining over 97% of the performance of models that also include stylistic signals. A compact set of 30 core features captures much of this separation: AI stories are systematically more thematically explicit, causally tidy, and temporally linear, while human stories show greater structural diversity and occupy a rarer, more dispersed region of narrative space. As surface-level signatures become increasingly transient, often removed by newer model versions or simple post-hoc edits, narrative features offer a more durable basis for authorship analysis, since altering them requires significant structural rewrites. Our features provide a measurable proxy for narrative uniqueness that complements existing detection tools.

AI disclosure. Large language models and coding agents (Claude Code and Codex) are used to aid with and polish writing and generate some tables and plots.

Acknowledgments

We thank Chau Minh Pham, who contributed heavily to brainstorming and early experiments of this work. We thank the University of Maryland Computational Linguistics and Information Processing (CLIP) Lab for their feedback and support. This project was partially supported by awards IIS-2046248 and IIS-2312949 from the National Science Foundation (NSF). We thank Google for a Cloud Credit award and Pangram Labs for an OpenRouter credit award, both of which enabled this research.

Ethics Statement

We acknowledge the copyright issues related to the Books3 dataset (Presser, 2020) and do not endorse its use for model training or commercial text generation. The use of this dataset in our paper is restricted to academic purposes only and is meant to understand the

narrative differences in human-written and AI-generated text to help inform discussions on AI-detection, authorship, and copyright policy.

References

George Alexandru Adam, Alexander Cui, Edwin Thomas, Emily Napier, Nazar Shmatko, Jacob Schnell, Jacob Junqi Tian, Alekhya Dronavalli, Edward Tian, and Dongwon Lee. GPTZero: Robust detection of LLM-generated texts, 2026. URL https://arxiv.org/abs/2602.13042.

Alibaba Cloud. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5, 2026. Model release.

Anthropic. Claude system card. Technical report, Anthropic, 2026. URL https://www-cdn.anthropic.com/bbd8ef16d70b7a1665f14f306ee88b53f686aa75.pdf.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. Fast-detectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Bpcgcr8E8Z.

Nina Beguš. Experimental narratives: A comparison of human crowdsourced storytelling and AI storytelling. *Humanities and Social Sciences Communications*, 11:1392, 2024. doi: 10.1057/s41599-024-03868-8. URL https://www.nature.com/articles/s41599-024-03868-8.

Margaret A. Boden. *The Creative Mind: Myths and Mechanisms*. Routledge, 2 edition, 2004.

Book Industry Study Group. BISAC subject headings list. https://www.bisg.org/complete-bisac-subject-headings-list, 2024. Industry standard for categorizing books by subject; accessed 2026-03-31.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://papers.nips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 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*, pp. 30:1–30:34. ACM, 2024. doi: 10.1145/3613904.3642731. URL https://dl.acm.org/doi/10.1145/3613904.3642731.

Tuhin Chakrabarty, Philippe Laban, and Chien-Sheng Wu. Can AI writing be salvaged? mitigating idiosyncrasies and improving human-AI alignment in the writing process through edits. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1210:1–1210:33. ACM, 2025. doi: 10.1145/3706598.3713559. URL https://dl.acm.org/doi/10.1145/3706598.3713559.

Tuhin Chakrabarty, Jane C. Ginsburg, and Paramveer Dhillon. Readers prefer outputs of AI trained on copyrighted books over expert human writers, 2026. URL https://arxiv.org/abs/2510.13939.

Jiaqi Chen, Xiaoye Zhu, Tianyang Liu, Ying Chen, Xinhui Chen, Yiwen Yuan, Chak Tou Leong, Zuchao Li, Tang Long, Lei Zhang, Chenyu Yan, Guanghao Mei, Jie Zhang, and Lefei Zhang. Imitate before detect: Aligning machine stylistic preference for machine-revised text detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23559–23567, 2025. doi: 10.1609/aaai.v39i22.34525. URL https://doi.org/10.1609/aaai.v39i22.34525.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939785. URL https://doi.org/10.1145/2939672.2939785.

Simon Colton and Geraint A. Wiggins. Computational creativity: The final frontier? In *Proceedings of the 20th European Conference on Artificial Intelligence (ECAI 2012)*, volume 242 of *Frontiers in Artificial Intelligence and Applications*, pp. 21–26. IOS Press, 2012. doi: 10.3233/978-1-61499-098-7-21.

DeepSeek-AI. DeepSeek-V3.2: Pushing the frontier of open large language models, 2025. URL https://arxiv.org/abs/2512.02556.

Aniket R. Doshi and Oliver P. Hauser. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(5):eadk1234, 2024. doi: 10.1126/sciadv.adk1234.

Bradley Emi and Max Spero. Technical report on the Pangram AI-generated text classifier, 2024. URL https://arxiv.org/abs/2402.14873.

Giorgio Franceschelli and Mirco Musolesi. On the creativity of large language models. *arXiv preprint arXiv:2304.00008*, 2023. URL https://arxiv.org/abs/2304.00008.

Gemini Team. Gemini: A family of highly capable multimodal models, 2023. URL https://arxiv.org/abs/2312.11805.

Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025a. URL https://arxiv.org/abs/2507.06261.

Gemini Team. Gemini 3 Flash model card. Technical report, Google DeepMind, 2025b. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf.

Gérard Genette. *Narrative Discourse: An Essay in Method*. Cornell University Press, Ithaca, NY, 1980.

Carlos Gómez-Rodríguez and Paul Andrew Williams. The unlikely duel: Evaluating creative writing in LLMs through a unique scenario. In *Proceedings of the XX Conference of the Spanish Association for Artificial Intelligence (CAEPIA 2024)*, pp. 225–226, 2024. URL https://www.grupolys.org/biblioteca/GomWil2024a.pdf.

Sil Hamilton, Matthew Wilkens, and Andrew Piper. NarraBench: A comprehensive framework for narrative benchmarking, 2025. URL https://arxiv.org/abs/2510.09869.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Spotting LLMs with Binoculars: Zero-shot detection of machine-generated text, 2024. URL https://openreview.net/forum?id=iARAKITHTH.

Yongxin He, Shan Zhang, Yixuan Cao, Lei Ma, and Ping Luo. DETree: DEtecting human-AI collaborative texts via tree-structured hierarchical representation learning. In *Advances in Neural Information Processing Systems*, 2025.

Zhaoyi Joey Hou, Bowei Alvin Zhang, Yining Lu, Bhiman Kumar Baghel, Anneliese Brei, Ximing Lu, Meng Jiang, Faeze Brahman, Snigdha Chaturvedi, Haw-Shiuan Chang, Daniel Khashabi, and Xiang Lorraine Li. CreativityPrism: A holistic evaluation framework for large language model creativity, 2026. URL https://arxiv.org/abs/2510.20091.

Baixiang Huang, Canyu Chen, and Kai Shu. Authorship attribution in the era of LLMs: Problems, methodologies, and challenges. *arXiv preprint arXiv:2408.08946*, 2024. URL https://arxiv.org/abs/2408.08946.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, and Yejin Choi. Artificial hivemind: The open-ended homogeneity of language models (and beyond). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=saDOrrnNTz.

Kimi Team. Kimi K2.5: Visual agentic intelligence, 2026. URL https://arxiv.org/abs/2602.02276.

Dmitry Kobak, Rita González-Márquez, Emőke-Ágnes Horvát, and Jan Lause. Delving into LLM-assisted writing in biomedical publications through excess vocabulary. *Science Advances*, 11(27):eadt3813, 2025. doi: 10.1126/sciadv.adt3813. URL https://www.science.org/doi/10.1126/sciadv.adt3813.

Mika Koivisto and Simone Grassini. Best humans still outperform artificial intelligence in a creative divergent thinking task. *Scientific Reports*, 13:13601, 2023. doi: 10.1038/s41598-023-40858-3.

Siyuan Li, Aodu Wulianghai, Xi Lin, Guangyan Li, Xiang Chen, Jun Wu, and Jianhua Li. StyleDecipher: Robust and explainable detection of LLM-generated texts with stylistic analysis, 2025. URL https://arxiv.org/abs/2510.12608.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Omer Levy, Luke Zettlemoyer, Jason E Weston, and Mike Lewis. Self-alignment with instruction backtranslation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=1oijHJBRsT.

David Y. Liu, Aditya Joshi, and Paul Dawson. Narrative theory-driven LLM methods for automatic story generation and understanding: A survey, 2026a. URL https://arxiv.org/abs/2602.15851.

Xinyue Liu, Niloofar Mireshghallah, Jane C. Ginsburg, and Tuhin Chakrabarty. Alignment whack-a-mole: Finetuning activates verbatim recall of copyrighted books in large language models, 2026b. URL https://arxiv.org/abs/2603.20957.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jeremy Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2:56–67, 2020. doi: 10.1038/s42256-019-0138-9.

Andrea Cristina McGlinchey and Peter J. Barclay. Using machine learning to distinguish human-written from machine-generated creative fiction, 2024. URL https://arxiv.org/abs/2412.15253.

Hope McGovern, Rickard Stureborg, Yoshi Suhara, and Dimitris Alikaniotis. Your large language models are leaving fingerprints. In Firoj Alam, Preslav Nakov, Nizar Habash, Iryna Gurevych, Shammur Chowdhury, Artem Shelmanov, Yuxia Wang, Ekaterina Artemova, Mucahid Kutlu, and George Mikros (eds.), *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pp. 85–95, Abu Dhabi, UAE, January 2025. International Conference on Computational Linguistics. URL https://aclanthology.org/2025.genaidetect-1.6/.

Ayat A. Najjar, Huthaifa I. Ashqar, Omar A. Darwish, and Eman Hammad. Leveraging explainable AI for LLM text attribution: Differentiating human-written and multiple LLM-generated text. *Information*, 16(9):767, 2025. doi: 10.3390/info16090767. URL https://www.mdpi.com/2078-2489/16/9/767.

Ramya Namuduri, Yating Wu, Anshun Asher Zheng, Manya Wadhwa, Greg Durrett, and Junyi Jessy Li. QUDsim: Quantifying discourse similarities in LLM-generated text. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=zFz1BJu211.

Ramya Namuduri, Manya Wadhwa, Anshun Asher Zheng, Greg Durrett, and Junyi Jessy Li. Genie: A fine-grained measure for novelty, 2026. URL https://arxiv.org/abs/2606.12790.

Manh Hung Nguyen and Adish Singla. Divergent-convergent thinking in large language models for creative problem generation, 2025. URL https://arxiv.org/abs/2512.23601.

Hiroshi Nonaka and K. E. Perry. Evaluating LLM story generation through large-scale network analysis of social structures. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*, 2025. URL https://openreview.net/forum?id=Ae4UeGI021.

OpenAI. GPT-5.1 instant and GPT-5.1 thinking system card. https://cdn.openai.com/pdf/4173ec8d-1229-47db-96de-06d87147e07e/5_1_system_card.pdf, November 2025.

OpenAI. GPT-5.4 thinking system card. Technical report, OpenAI, 2026. URL https://deploymentsafety.openai.com/gpt-5-4-thinking/gpt-5-4-thinking.pdf.

James O'Sullivan. Stylometric comparisons of human versus AI-generated creative writing. *Humanities and Social Sciences Communications*, 12:1708, 2025. doi: 10.1057/s41599-025-05986-3.

Vishakh Padmakumar, Chen Yueh-Han, Jane Pan, Valerie Chen, and He He. Measuring LLM novelty as the frontier of original and high-quality output. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=i7QNKZioN6.

Samuel J. Paech. EQ-Bench: An emotional intelligence benchmark for large language models, 2024. URL https://arxiv.org/abs/2312.06281.

Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. TopicGPT: A prompt-based topic modeling framework. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2956–2984, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.164. URL https://aclanthology.org/2024.naacl-long.164/.

Chau Minh Pham, Jenna Russell, Dzung Pham, and Mohit Iyyer. Frankentext: Stitching random text fragments into long-form narratives, 2025. URL https://arxiv.org/abs/2505.18128.

Andrew Piper and Sunyam Bagga. Using large language models for understanding narrative discourse. In Yash Kumar Lal, Elizabeth Clark, Mohit Iyyer, Snigdha Chaturvedi, Anneliese Brei, Faeze Brahman, and Khyathi Raghavi Chandu (eds.), *Proceedings of the 6th Workshop on Narrative Understanding*, pp. 37–46, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wnu-1.4. URL https://aclanthology.org/2024.wnu-1.4/.

Shawn Presser. Books3, 2020. URL https://twitter.com/theshawwn/status/1320282149329784833.

Rishanth Rajendhran, Jenna Russell, Mohit Iyyer, and John Wieting. POLARIS: Guiding small models to write long stories, 2026. URL https://arxiv.org/abs/2606.04095.

Alex Reinhart, Ben Markey, Michael Laudenbach, Kachatad Pantusen, Ronald Yurko, Gordon Weinberg, and David West Brown. Do LLMs write like humans? variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences of the United States of America*, 122(8):e2422455122, 2025. doi: 10.1073/pnas.2422455122. URL https://www.pnas.org/doi/10.1073/pnas.2422455122.

Jenna Russell, Marzena Karpinska, and Mohit Iyyer. People who frequently use ChatGPT for writing tasks are accurate and robust detectors of AI-generated text. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5342–5373, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.267. URL https://aclanthology.org/2025.acl-long.267/.

Arkadiy Saakyan, Najoung Kim, Smaranda Muresan, and Tuhin Chakrabarty. Death of the novel(ty): Beyond n-gram novelty as a metric for textual creativity. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=z2idLjqzBe.

Chantal Shaib, Vinith Menon Suriyakumar, Byron C Wallace, and Marzyeh Ghassemi. Learning the wrong lessons: Syntactic-domain spurious correlations in language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=oBikm5Rshc.

Chantal Shaib, Tuhin Chakrabarty, Diego Garcia-Olano, and Byron C. Wallace. Measuring AI "slop" in text, 2026. URL https://arxiv.org/abs/2509.19163.

Theodore Sturgeon. Tiny and the monster. In Isaac Asimov and Martin H. Greenberg (eds.), *Isaac Asimov Presents The Great SF Stories 9 (1947)*. DAW Books, 1983. Originally published in Astounding Science Fiction, May 1947.

Mingjie Sun, Yida Yin, Zhiqiu Xu, J Zico Kolter, and Zhuang Liu. Idiosyncrasies in large language models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=FCZ3jVzmTZ.

Katherine Thai, Bradley Emi, Elyas Masrour, and Mohit Iyyer. EditLens: Quantifying the extent of AI editing in text. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=gOkitaPCfZ.

Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. Are large language models capable of generating human-level narratives? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17659–17681, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.978. URL https://aclanthology.org/2024.emnlp-main.978/.

Ellis Paul Torrance. *Torrance Tests of Creative Thinking: Norms-Technical Manual: Verbal Tests, Forms A and B; Figural Tests, Forms A and B*. Personnel Press, 1966.

Nafis Irtiza Tripto, Saranya Venkatraman, Mahjabin Nahar, and Dongwon Lee. Beyond checkmate: Exploring the creative choke points for AI generated texts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 11953–11970, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.600. URL https://aclanthology.org/2025.emnlp-main.600/.

U.S. Copyright Office. Copyright registration guidance: Works containing material generated by artificial intelligence, 2023. https://www.copyright.gov/ai/ai_policy_guidance.pdf.

U.S. Supreme Court. Feist Publications, Inc. v. Rural Telephone Service Co. 499 U.S. 340, 1991. U.S. Supreme Court.

Manya Wadhwa, Tiasa Singha Roy, Harvey Lederman, Junyi Jessy Li, and Greg Durrett. CREATE: Testing LLMs for associative creativity, 2026. URL https://arxiv.org/abs/2603.09970.

Zihan Wang, Jingbo Shang, and Ruiqi Zhong. Goal-driven explainable clustering via language descriptions. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10626–10649, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.657. URL https://aclanthology.org/2023.emnlp-main.657/.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2526–2547, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.127. URL https://aclanthology.org/2025.acl-long.127/.

Weijia Xu, Nebojsa Jojic, Sudha Rao, Chris Brockett, and Bill Dolan. Echoes in AI: Quantifying lack of plot diversity in LLM outputs. *Proceedings of the National Academy of Sciences*, 122(35):e2504966122, 2025a. doi: 10.1073/pnas.2504966122. URL https://www.pnas.org/doi/abs/10.1073/pnas.2504966122.

Yihuai Xu, Yongwei Wang, Yifei Bi, Huangsen Cao, Zhouhan Lin, Yu Zhao, and Fei Wu. Training-free LLM-generated text detection by mining token probability sequences. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=vo4AHjowKi.

Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. NoveltyBench: Evaluating creativity and diversity in language models. In *Second Conference on Language Modeling*, 2025a. URL https://openreview.net/forum?id=XZm1ekzERf.

Ziyin Zhang, Zihan Liao, Hang Yu, Peng Di, and Rui Wang. F2LLM technical report: Matching SOTA embedding performance with 6 million open-source data, 2025b. URL https://arxiv.org/abs/2510.02294.

Statistic	Value
Prompts / human stories	10,272
Target words (mean / median)	6,242 / 5,000
Target words (IQR)	3,000–7,000
Human story words (mean / median)	6,403 / 5,035
Human story words (IQR)	2,930–7,650

Table 4: Human-corpus summary. All stories extracted from Books3.

Model	n	Coverage	Mean words	Median words	Mean Δ words	Mean abs. % err.	Within 10%
Gemini 3 Flash	10,261	99.9%	3,155	2,967	-3,089	36.9%	19.4%
GPT-5.4	10,272	100.0%	6,651	6,515	409	40.2%	9.6%
Claude Sonnet 4.6	10,259	99.9%	6,817	6,187	575	16.6%	38.4%
DeepSeek V3.2	10,272	100.0%	2,946	2,693	-3,296	39.9%	12.8%
Kimi K2.5	10,272	100.0%	3,274	3,005	-2,968	35.2%	15.6%

Table 5: Length and coverage statistics for AI-generated stories in the finalized six-model dataset. Word-target adherence is computed against the explicit "approximately N words" instruction embedded in each prompt; Δ is signed model-word-count minus target-word-count.

A Data

Information about our human-written corpus is depicted in Table 4. Information on average length of generated stories, as well as how well models adhered to length-based instructions, is in Table 5 and shown in Figure 4. Story generation cost roughly $2800 USD in total, varying widely by model. Extracting features from Gemini 3 Flash (Gemini Team, 2025b) over the entire story-corpus cost roughly $1600 USD. All stories were generated with a maximum length of 128,000 tokens, except for Gemini 3 Flash, whose maximum token limit is 65,536.

B STORYSCOPE details

Blinding protocol. Throughout the pipeline, source identities are anonymized in all LLM-facing prompts. Within pairwise template comparisons, presentation order is randomized to reduce positional bias while keeping the comparisons reproducible.

Are templates truly needed? We ran early-stage testing on the full pipeline with and without template extraction, the only difference being whether the comparative analysis operates on templates or raw story text. The two variants discover qualitatively different features: comparing the top-20 discriminative features, only 6 overlap. The direct pipeline's features are style-heavy (humor usage, vocabulary register, allusion types, dominant imagery), while the template pipeline's are structure-heavy (emotional arcs, relationship trajectories, event density, flashback usage).

Meta-analysis batching details. The cross-author comparative analysis operates over 100 development prompts, each with six source templates. We batch prompt sets greedily by estimated token count, treating the combined size of the six templates for a prompt as the batching unit; this yields 32 batches, averaging 3.1 prompts per batch. A parallel Gemini 3 Flash run permits fewer, larger batches but produces substantially fewer features, so we retain the GPT-5.1 outputs.

Feature deduplication. Feature discovery produces 408 candidate features, many of which capture overlapping concepts (e.g., "metanarrative self-awareness" and "self-reflexive narration"). We deduplicate them with embedding-based clustering: each feature is represented by its name, question, and detection method, encoded with F2LLM-4B (Zhang et al., 2025b), then clustered with single linkage at cosine threshold 0.85. We keep the feature nearest each cluster centroid as the representative, reducing the taxonomy from 408 to 304 features (25.5% reduction) and merging 65 multi-feature clusters. We use 0.85 because 0.90 left many clear

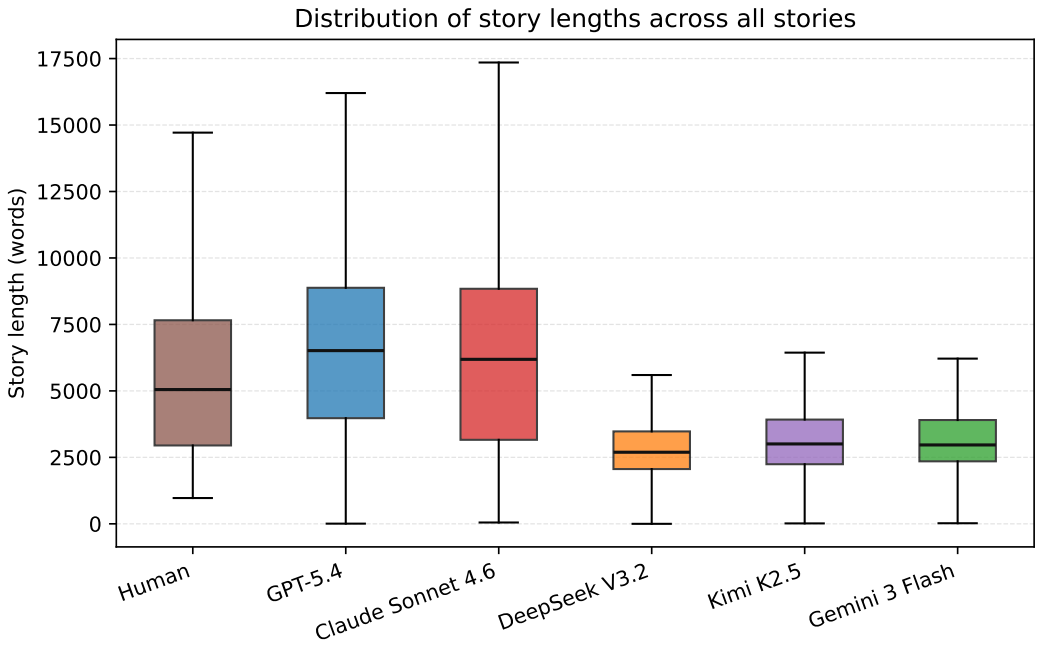

Figure 4: Boxplots of story lengths across all stories in the finalized six-source dataset, shown separately for the human corpus and each model. Boxes show the interquartile range, center lines show medians, and whiskers show the non-outlier range.

duplicates unmerged (372 features total), whereas thresholds above 0.85 quickly stopped merging semantically overlapping features that differed only slightly in wording.

Feature taxonomy details. The final taxonomy contains $d = 304$ interpretable narrative features spanning five response types. Table 6 summarizes the final distribution of feature types and gives one full illustrative feature entry for each type.

Style boundary. Style features concern sentence- and phrase-level texture (diction, syntax, rhythm, figurative density, tonal register, and prose-level allusion), whereas non-style features concern narrative content and structure (events, causality, closure, character roles, relationship topology, setting, and temporal arrangement). For gray-zone cases, our rule is simple: if a feature can only be answered from prose texture, it is style; if it is primarily about narrative content and prose choices are incidental, it is non-style. *Sensory Density*, *Depth of Interior Access*, and *Chronological Discontinuity* fall on the non-style side of this rule. We conducted a style-dependence audit over all 304 taxonomy features, prompting GPT-5.4 (OpenAI, 2026) to rate each as `high`, `medium`, or `low` style dependence. For the strict narrative variant used in the final numbers, the exclusion set contains 47 features. This leaves 257 features in the strict non-style taxonomy used by the final narrative runs.

C Application of features

Why not single shot application? We compared single-call application against aspect-based application over 12 story files. Aspect-based application produced far more complete feature vectors: average coverage rose from 68.4% of features to 95.4% of features. The remaining aspect-based misses were concentrated in a small number of harder-to-apply features, mostly in setting, whereas single-call application showed broad systematic dropout, especially in revelation and temporal-structure features. Later dimensions are omitted more often in the single-call setup, while the aspect-based variant preserves near-complete coverage by restricting each call to one narrative dimension.

Type	Count	Feature	Details
Categorical	124	Dominant narrative person (*Perspective*)	**Question:** What is the primary grammatical person used by the main narrator across the story? **Description:** Identify the default pronoun/verb alignment in narrative (not dialogue): 'I/we', 'you', 'he/she/they'. If no dominant usage, mark 'mixed_no_dominant'. **Answer choices:** 'first_person'; 'second_person'; 'third_person'; 'mixed_no_dominant'.
Ordinal	59	Secondary character density (*Agents*)	**Question:** How many distinct secondary/supporting characters appear in the story? **Description:** Include minor or functional characters; exclude major cast from `AGENT_ID_001`. Map counts to bins. **Answer choices:** '0–1'; '2–3'; '4–7'; '8+'.
Scale	45	Dyadic vs Group Scene Emphasis (*Social Networks*)	**Question:** To what extent do key interactions occur in dyads versus groups? **Description:** Evaluate major scenes. Rate 1 = mostly dyads, 3 = balanced, 5 = mostly groups. **Answer choices:** '1'; '2'; '3'; '4'; '5'.
Binary	44	Fourth-Wall Breaking (*Situatedness*)	**Question:** Does the narration directly address the reader? **Description:** Look for explicit audience address (e.g., 'dear reader'). Ignore in-world dialogue. **Answer choices:** 'no'; 'yes'.
Multi-select	32	Global Narrative Schema (*Events*)	**Question:** Which high-level narrative schemas are central? **Description:** Identify structural patterns such as journey, mystery, transformation, ordeal, slice-of-life, trial, heist, or framed recollection. **Answer choices:** 'quest/journey'; 'investigation/mystery'; 'transformation/redemption'; 'siege/ordeal'; 'slice_of_life'; 'trial/test/game'; 'heist/caper'; 'frame_confession/memoir'.

Table 6: Distribution of feature types in the feature taxonomy. An example feature for each type, with the question, description, and answer given, to illustrate the type of features searched for.

Comparison	Encoded exact agreement (%)	Cohen's κ
Annotator 1 vs model	91.67	0.91
Annotator 2 vs model	79.86	0.77
Mean human vs model	85.76	0.84
Human vs human	76.85	0.74

Table 7: Human validation on 240 story-feature items across 12 stories.

Extractor reliability. We evaluate the reliability of Gemini 3 Flash (minimal thinking) as the production feature extractor via a repeated-measures design. This yields 300 total extraction outputs and, for fully present items, 600 within-item run pairs. We use the minimal thinking variant due to the much lower cost. Since we give full stories as part of the input context, having thinking on makes Gemini think for a very long time, even for the simple feature assignment tasks. On a smaller test of full vs. minimal thinking, the minimal thinking caused a 0.45% reduction in performance on the binary task and an 11.61% increase in performance on the six-way attribution task. Aggregate inter-run agreement across the five runs is Krippendorff's α (nominal) = 0.90, mean pairwise Cohen's κ = 0.89.

Human validation against annotators. We ran a small human validation study over 12 stories, 240 story-feature items per annotator. Two annotators completed the full set. We report agreement in that encoded representation. Against the model assignments, Cohen's $\kappa = 0.91$ for *annotator 1* and Cohen's $\kappa = 0.77$ for *annotator 2*; the mean human–model Cohen's κ is therefore 0.84 (see Table 7). Human–human encoded agreement on the same subset is Cohen's $\kappa = 0.74$.

D Experimental setup

Model selection. We compared linear classifiers, random forests, and XGBoost on the validation split (100 prompts, 600 stories). XGBoost consistently outperformed the alternatives on both binary and multiclass tasks, and integrates cleanly with SHAP-based interpretability analyses (Lundberg & Lee, 2017; Lundberg et al., 2020), so we adopt it as the default classifier throughout.

Baseline setup. We evaluate four external baselines that operate on raw story text, without access to the LLM-extracted narrative features.

1. *Stylometric baseline.* We extract 144 surface features from each story, including sentence-, word-, and paragraph-length statistics, document-level counts, vocabulary richness metrics, 100 function-word frequencies, punctuation rates, dialogue features, and readability indices.
2. *TF-IDF baseline.* We fit a unigram/bigram TF-IDF vectorizer. The resulting matrix is passed through the same XGBoost sweep and train/val/test protocol as the stylometric baseline.
3. *ModernBERT baseline.* We fine-tune ModernBERT-base (Warner et al., 2025) directly on raw story text using the same train/val/test split.
4. *Binoculars.* We run Binoculars in accuracy mode with all other settings left at their defaults.

Feature encoding. For the binary task we set XGBoost's `scale_pos_weight`$= 5$ to match the 5:1 AI-to-human class ratio. Multiclass uses uniform class weights. Features are encoded with one-hot columns for nominal and binary types and explicit integer encoding for ordinal and scale types.

Evaluation protocol. The final split layout is 7,383 train / 1,405 val / 1,384 test prompts. We tune hyperparameters on the validation split, then retrain each final model on train+val (8,788 prompts; 52,707 stories) and report all final results on the held-out test set (1,384 prompts, 8,301 stories). For binary classification, we report macro-F1 and AUPRC.[20] For six-way attribution, we report macro-F1 and accuracy.

Feature role assignment. A feature is marked *important* if $\bar{s}_j$ exceeds the top-quartile threshold on mean absolute SHAP across features. Among important features, j is labeled *stable-important* iff (i) stab_j is at least the median stability score among important features and (ii) its mean SHAP exceeds the 95th-percentile null-label baseline with permutation $p \leq 0.10$; otherwise it is *unstable-important*. Non-important features with above-median stability are labeled *stable-weak*, and the remainder are *noise*.

Core features come from the *binary* SHAP analysis and must satisfy all of the following: *stable-important* quadrant membership, null significance, $\text{stab}_j \geq 0.55$, $\text{top25}_j \geq 0.60$, absolute human–AI mean-value gap at least 0.20, and cross-model AI spread at most 0.35.

E Dimension-level ablations

Table 8 reports two ablations. *Dimension-only* retrains the classifier on a single NarraBench dimension; *leave-one-out* retrains it on all narrative features except that dimension's, both under the protocol of §D. No dimension is individually sufficient (best: *agents* at 80.2% binary macro-F1) and none is individually necessary (worst removal: *agents* at -1.2), so the signal is redundant across correlated dimensions. This is expected, since AI-generated fiction has a recognizable narrative profile: e.g., more linear plots, tighter causal structure, more explicit thematic framing, and less diverse character and social structure often co-occur.

[20] With 5:1 class imbalance a trivial all-AI classifier achieves 83.3% accuracy, so we omit accuracy from the binary table and rely on macro-F1 and AUPRC.

Dimension only			*Leave-one-dimension-out*		
Dimension	**Binary**	**Six-way**	**Dimension**	**Binary**	**Six-way**
Agents	80.2	46.0	Agents	92.0 (−1.2)	66.3 (−2.1)
Situatedness	77.3	41.3	Situatedness	92.3 (−0.8)	66.1 (−2.3)
Plot	74.5	40.5	Perspective	92.6 (−0.6)	67.3 (−1.2)
Perspective	71.3	32.3	Setting	92.8 (−0.4)	66.8 (−1.7)
Setting	70.6	38.4	Events	92.8 (−0.4)	67.4 (−1.0)
Social networks	67.9	38.3	Plot	92.8 (−0.4)	67.5 (−0.9)
Events	67.8	37.4	Temporal structure	93.0 (−0.2)	65.4 (−3.1)
Revelation	67.2	33.2	Social networks	93.0 (−0.2)	67.8 (−0.6)
Temporal structure	62.0	37.7	Revelation	93.5 (+0.3)	67.7 (−0.7)
All narrative	*93.2*	*68.4*	*All narrative*	*93.2*	*68.4*

Table 8: Per-dimension ablations of the narrative model, macro-F1 (%). Left: trained on one NarraBench dimension only. Right: trained on all narrative features except that dimension, with the change relative to the full narrative model in parentheses.

F Memorization risk in story generation

Because the human corpus includes stories that are likely to be widely available on the public web and in Books3-adjacent collections, we run a conservative contamination-risk audit over the reverse-engineered outputs. Memorization could lead to the inclusion of style atypical to AI writing. We flag *exact overlap* when the pair shares at least one exact 13-gram beyond split-matched shuffled-human controls, as done in prior work (Brown et al., 2020), and *near-verbatim overlap* when paired 8-gram coverage is at least 5%, the pair shares at least four distinct 8-grams, and the longest exact common span is at least 30 tokens, again beyond shuffled-human controls. We measure output overlap rather than formal training-set membership, so we treat it as a sensitivity analysis rather than proof that a model saw a specific story during pretraining. Risk rates in Table 9.

Model	**Exact 13-gram**	**Near-verbatim**
GPT-5.4	64 (0.63%)	7 (0.07%)
DeepSeek V3.2	58 (0.57%)	13 (0.13%)
Kimi K2.5	44 (0.43%)	7 (0.07%)
Gemini 3 Flash	119 (1.17%)	32 (0.32%)
Claude Sonnet 4.6	67 (0.66%)	14 (0.14%)

Table 9: Memorization-risk rates by model.

The memorization screen flags 352 of 50,672 evaluated generated-story comparisons (0.70%) as high risk. Under the conservative prompt-level ablation rule, this removes 169 of 10,139 audited prompts (1.67%). The strongest cases are concentrated in canonical or heavily reprinted texts such as *The Yellow Wallpaper*, *The Legend of Sleepy Hollow*, *Heart of Darkness*, *The Call of Cthulhu*, and other Poe, Doyle, Wilde, and Dickens stories.

Dropping the high-risk prompts leaves the main detectability conclusions essentially unchanged. In the filtered rerun, the Narrative + Style model improves slightly from 96.00 to 96.06 macro-F1 in binary detection and from 77.29 to 77.56 macro-F1 in six-way attribution; the Narrative-only model shifts from 93.18 to 93.23 in binary and from 68.42 to 68.68 in multiclass (see Table 10).

G Additional analysis

Does length matter for narrative detection? Human stories are substantially longer than AI stories on the held-out final test split (human mean / median = 6418/4973 words; AI mean / median = 4523/3355). To test whether binary detectability is primarily a length artifact, we run two additional analyses. We freeze the final binary classifiers from §4.

Variant	Original test macro-F1	Low-risk test macro-F1	Δ
Narrative + Style (binary)	96.00	96.06	+0.06
Narrative + Style (multiclass)	77.29	77.56	+0.27
Narrative (binary)	93.18	93.23	+0.05
Narrative (multiclass)	68.42	68.68	+0.26

Table 10: Sensitivity of the final detectability results after dropping all high-risk prompts.

Model	Setting	Macro-F1 (%)	Human-F1	Bal. Acc.	AUPRC
Length only	Full test	55.9	0.330	0.597	0.261
Style Only	Full test	85.8	0.766	0.871	0.867
Style Only	Length-matched test	86.8	0.859	0.869	0.956
Narrative	Full test	93.2	0.886	0.932	0.959
Narrative	Length-matched test	93.2	0.929	0.932	0.990
Narrative + Style	Full test	96.0	0.933	0.956	0.982
Narrative + Style	Length-matched test	95.7	0.956	0.957	0.996

Table 11: Held-out final-test length-confound audit. "Length only" is a logistic regression trained on train+val story word counts alone. "Length-matched test" reuses the frozen final binary models but evaluates them on a 2,754-story subset (1,377 human, 1,377 AI) formed by stratifying AI stories into human-derived length deciles and sampling without replacement.

and re-evaluate them on a decile-stratified length-matched test subset with 2,754 stories (1,377 human, 1,377 AI) and nearly identical medians (human 4,973 vs. AI 4,951 words). The Narrative model is unchanged (93.2% macro-F1 before and after matching), indicating that the main separation is not driven by the trivial human-longer / AI-shorter length contrast. The Style Only model is similarly stable, moving from 85.8% to 86.8% macro-F1 after matching. Because that matched-subset comparison alone does not establish perfect length invariance, we also evaluate the same frozen models separately on short, medium, and long bands defined by human-story tertiles within the matched subset. Narrative macro-F1 remains 91.6 / 94.3 / 93.7 across these bins, Narrative + Style remains 95.0 / 96.2 / 95.9, and Style Only remains 86.4 / 86.5 / 87.4, showing little variation with length overall (Table 11).

Does topic matter for detection? We also examined whether performance varies by story topic. We prompted Qwen3.5-9B (Alibaba Cloud, 2026) to give a Book Industry Standards and Communications (BISAC) code (Book Industry Study Group, 2024) given a story text, and apply this over the test set. Stories are categorized into 12 topics but only 6 have at least 20 test prompts. For the Narrative model, binary macro-F1 ranges from 90.0% (mystery/detective) to 96.2% (historical), while six-way macro-F1 ranges from 65.8% (mystery/detective) to 70.8% (science fiction).

	Hum.	Cla.	GPT	Gem.	DS	Kimi
Narr. + Style	0.93	0.89	0.82	0.67	0.66	0.67
Narrative	0.89	0.77	0.73	0.60	0.57	0.55

Table 12: Per-class F1 for six-way attribution. Human writing is easiest to classify, while Gemini, DeepSeek, and Kimi have the lowest performance.

To test whether mean detection performance differs by topic, we computed prompt-level test accuracy within each prompt's majority topic and ran a Kruskal–Wallis omnibus test over the six topics with at least 20 test prompts (literary, science fiction, horror, action/adventure, fantasy, mystery/detective). We find no significant topic-wise differences: H=4.69, p=0.46. So topic shifts the point estimates somewhat, especially for multiclass attribution, but we do not find strong evidence that detectability changes systematically by topic.

H Per-story rarity analysis

The group-level centroid and dispersion statistics above characterize source regions but do not directly measure whether individual human stories are more structurally unusual. To operationalize the originality definition from §1, we compute per-story rarity as the mean

Tail	Human	Claude	GPT	DeepSeek	Kimi	Gemini
Top 1%	42	5	5	11	6	14
Top 5%	180	33	34	58	64	45
Top 10%	340	66	71	127	135	88

Table 13: Composition of the rarest test-set stories by source. Counts are out of 8,262 test stories (1,377 per source). Human stories are overrepresented in all tails relative to a $\frac{1}{6}$ baseline, but AI stories are present throughout.

Euclidean distance to a story's $k = 25$ nearest neighbors in the pooled train+val narrative space. Rarity percentiles are computed against the train+val rarity distribution: a story at the 90th percentile sits in a sparser region of narrative space than 90% of all train+val stories. All effect sizes here are computed over individual stories, not per-prompt summaries.

Figure 5 shows the per-source rarity distributions on the held-out test set. Human stories are shifted toward higher rarity (mean percentile 0.71 vs. 0.49 for AI, AUC = 0.73). Using the standard sample-size-weighted pooled SD over the 1,377 human and 6,885 AI test stories gives Cohen's $d = 0.83$, with sample SDs of 0.227 for human stories and 0.274 for pooled AI stories and a pooled SD of 0.267.

At the prompt level, the human story is rarer than all five AI alternatives 57.8% of the time (mean margin = 0.96 raw distance units). The rarest test-set tail is mixed: the top 1% contains 42 human and 41 AI stories; the top 5% contains 180 human vs. 234 AI; the top 10% contains 340 human vs. 487 AI stories across five models (Table 13).

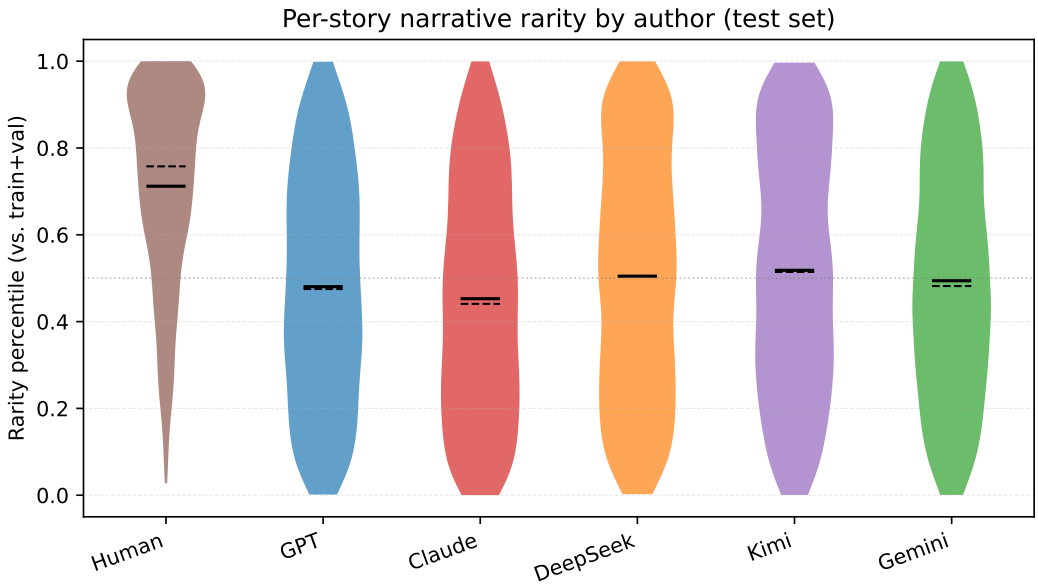

Figure 5: Per-story narrative rarity percentiles by author on the held-out test set. Solid lines show means; dashed lines show medians. Human stories are shifted toward higher rarity but all distributions overlap substantially.

I Core and fingerprint features

Table 14 lists the 20 features that reliably characterize AI writing in the binary task, ranked by a core score computed as mean SHAP multiplied by stability score and by one plus the absolute human–AI gap for the qualifying encoded column.

#	Feature	Question	Dim	Type	Response Options
1	Thematic Explicitness and Moralizing	How explicitly does the story articulate its themes or morals?	SIT	scale	1–5
2	Dominant Emotional Expression → embodied	How are characters' emotions most commonly conveyed?	AGENT	cat	explicit labels, embodied metaphors, behavioral cues, ambiguous
3	Thematic Unity	To what extent do subplots and flourishes serve a central thematic concern?	PLT	scale	1–5
4	Dominant Sensory Modalities → olfactory	Which sensory modalities does the story most frequently engage?	SET	multi	visual, auditory, olfactory, tactile, gustatory, kinesthetic
5	Character Introduction → external description	What narrative device primarily introduces the central character?	AGENT	cat	external_desc, in-action, in-dialogue, inner_thought, others_reports
6	Setting as Psychological Mirror	To what degree does physical environment mirror characters' inner states?	SET	scale	1–5
7	Continuity of Main Causal Chain	How continuous is the single causal chain from inciting incident to ending?	EVT	scale	1–5
8	Sensory Density	How dense is sensory description across the narrative?	SET	scale	minimal–lush
9	Agency in Resolution → protagonist choice	Is resolution driven by protagonist's choices or external events?	PLT	cat	protagonist_choice, mixed, external_fate
10	Narratorial Thematic Commentary → yes	Does the narrator explicitly comment on themes beyond characters' perspectives?	SIT	binary	no, yes
11	Opening Spatial Grounding	How clearly does the opening ground the reader in a specific physical setting?	SET	ord	none/vague, minimal, clear local, clear local+global
12	Dialogue Function → philosophical debate	What main functions does dialogue serve?	PER	multi	advance plot, reveal character, worldbuilding, philosophical, comic
13	Spatial Granularity Level	How fine-grained is the story's depiction of physical space?	SET	ord	very_low–high
14	Subplot Integration → no subplots	How directly do subplots echo the central theme?	PLT	cat	no_subplots, thematically_parallel, contrasting, independent
15	Moral / Philosophical Weighting	How heavily does the story foreground moral or philosophical questions?	SIT	scale	1–5
16	Reference Explicitness → implicit echoes	Are intertextual gestures primarily explicit or diffuse?	SIT	cat	none, explicit named, implicit echoes, balanced mix
17	Environmental and Ecological Emphasis	How prominent is the natural environment or ecology in the narrative?	SET	scale	1–5
18	Mode of Resolution → internal understanding	Is the main event chain resolved through internal acceptance or external action?	EVT	cat	resolved externally, resolved internally, unresolved
19	Pre-Threat Character Investment	How much does the story build investment before major jeopardy?	REV	scale	1–5
20	Depth of Interior Access	How deep into characters' inner life does narration go?	PER	scale	1–5

Table 14: Core AI-characterizing features (20). Dim = NarraBench dimension prefix; Type: scale = 1–5 Likert, cat = categorical, ord = ordinal, multi = multi-select, binary = yes/no. Arrow (→) indicates the specific option value that is elevated for AI.

#	Feature	Question	Dim	Type	Response Options
1	Intertextual Strategy Types → explicit named reference	What kinds of intertextual engagement does the story employ?	SIT	multi	explicit named, retelling, pastiche, myth/religion, self referential
2	Frequency of Direct Reader Address	How often does the text directly address the reader?	PER	ord	never, occasional asides, frequent/structural
3	Reference Explicitness → balanced mix	Are intertextual gestures explicit or diffuse?	SIT	cat	none, explicit named, implicit echoes, balanced mix
4	Depth of Recontextualization After Surprise	How extensively does a revelation force reinterpretation of earlier scenes?	REV	scale	1 (none)–5 (complete re-reading)
5	Dialogue-to-Narration Proportion	What proportion of text is direct dialogue vs. narration?	PER	scale	1 (no dialogue)–5 (dialogue dominates)
6	Fourth-Wall Permeability	To what extent does the story break the boundary between story-world and reader?	SIT	ord	1 (no breaking)–4 (radical violations)
7	Subplot Integration → thematically parallel	How directly do subplots echo the central theme?	PLT	cat	no subplots, thematically parallel, contrasting, independent
8	Degree of Chronological Discontinuity	How often does the narrative jump across time?	TMP	scale	1–5
9	Location Variety Scope	How many distinct physical locales does the story inhabit?	SET	ord	single–multiworld
10	Anachrony Intensity	How heavily does the narrative rely on flashbacks or flashforwards?	TMP	scale	1 (absent)–5 (dominant anachronic)
11	Moral Polarity Toward Protagonist → ambivalent	Does the narrative frame the protagonist's choices as morally clear or ambiguous?	PLT	cat	clearly positive, ambivalent/mixed, clearly negative
12	Dominant Emotional Expression → explicit labels	How are characters' emotions most commonly conveyed?	AGENT	cat	explicit labels, embodied metaphors, behavioral cues, ambiguous
13	Nonlinear Framing for Delayed Disclosure	To what extent does the story use time jumps to stage revelations?	REV	scale	1 (linear)–5 (heavily fragmented)

Table 15: Core human-characterizing features (13), Arrow (→) indicates the specific option value elevated for human authors. Abbreviations as in Table 14.

Table 15 lists the 13 features that reliably characterize human writing, ranked by the same core score.

Table 16 presents all 30 core features organized by interpretive theme, with human and AI mean values and gaps. This table is referenced from §4.

The number of fingerprints varies widely across sources (Human: 32, Claude: 26, GPT: 11, Gemini: 11, DeepSeek: 7, Kimi: 3). Table 17 provides a compact summary.

J Prompts

Given a source story, this prompt asks the model to produce a single concise writing prompt that preserves the story's key characters, setting, and thematic direction while leaving room for variation. Prompt is in Figure 6.

Feature	Human	AI	Gap
AI-elevated: Thematic over-determination			
Thematic Explicitness & Moralizing [s]	3.28	3.94	−0.65
Moral / Philosophical Weighting [s]	3.26	3.68	−0.42
Thematic Unity [s]	4.41	4.74	−0.33
Narratorial Thematic Commentary → yes	52%	77%	−25
Dialogue Function → philosophical debate	34%	59%	−25
Reference Explicitness → implicit echoes	50%	72%	−22
AI-elevated: Sensory & embodied performativity			
Emotional Expression → embodied	38%	81%	−42
Setting as Psychological Mirror [s]	3.58	4.07	−0.49
Environmental & Ecological Emphasis [s]	2.83	3.21	−0.38
Sensory Modalities → olfactory	57%	82%	−26
Sensory Density [s]	3.66	3.93	−0.26
Depth of Interior Access [s]	3.67	3.93	−0.26
AI-elevated: Structural streamlining			
Causal Chain Continuity [s]	3.92	4.20	−0.28
Spatial Granularity [o]	2.27	2.53	−0.26
Agency in Resolution → protagonist choice	46%	69%	−23
Character Introduction → external description	30%	52%	−22
Subplot Integration → no subplots	57%	79%	−22
Resolution Mode → internal understanding	27%	47%	−21
Opening Spatial Grounding [o]	2.12	2.33	−0.20
Pre-Threat Character Investment [s]	2.76	2.99	−0.23
Human-elevated: Intertextual richness			
Intertextual Strategy → explicit named reference	47%	24%	+23
Reference Explicitness → balanced mix	37%	16%	+21
Human-elevated: Reader engagement			
Fourth-Wall Permeability [o]	0.67	0.39	+0.28
Direct Reader Address [o]	0.28	0.07	+0.21
Human-elevated: Temporal complexity			
Depth of Recontextualization After Surprise [s]	3.28	2.95	+0.34
Chronological Discontinuity [s]	2.40	2.12	+0.28
Nonlinear Framing for Delayed Disclosure [s]	1.96	1.68	+0.28
Anachrony Intensity [s]	2.58	2.31	+0.27
Human-elevated: Narrative diversity			
Location Variety Scope [o]	1.34	1.08	+0.26
Dialogue-to-Narration Proportion [s]	2.95	2.70	+0.24
Subplot Integration → thematically parallel	42%	21%	+22
Moral Polarity → ambivalent/mixed	59%	38%	+21
Emotional Expression → explicit labels	29%	8%	+21

Table 16: Core features separating human from AI writing, grouped by theme. Features marked [s] are 1–5 Likert scales (values are means); [o] are ordinal scales (values are means over integer codes); → indicates a specific categorical/binary option (values are prevalence %). Gap = Human − AI; negative gaps indicate AI-elevated features. AI column averages across all five models.

Author	#	Feature	Dim	SHAP	Uniq.
Human	1	Character introduction → in-dialogue	AGENT	0.110	21.4
	2	Breadth of focalization → single focal	PER	0.083	15.7
	3	Narrator address mode → no direct address	SIT	0.069	12.6
	4	Overall revelation pacing → back-loaded	REV	0.094	7.7
	5	Literary ambition → crossover genre	SIT	0.116	6.8
	+ 27 more (visibility of withholding, atmospheric techniques, subplot density, naming, twist placement, . . .)				
Claude	1	Strength of event escalation	EVT	0.402	22.4
	2	Event-type diversity	EVT	0.491	10.7
	3	Ending temporal scope → epilogue/flashforward	TMP	0.096	8.9
	4	Dreams/visions as temporal distortion → no	TMP	0.116	7.7
	5	Setting mood → uncanny/haunted	SET	0.059	4.6
	+ 21 more (event density, conflict modality, relationship trajectory, heteroglossia, closure, . . .)				
GPT	1	Role of gossip and rumor → salient	SOC	0.200	22.1
	2	Narrator temporal distance → distant retrospective	TMP	0.119	6.8
	3	Reader expectation strategy → subverts	REV	0.098	3.9
	4	Iterative/habitual narration → no	TMP	0.144	3.2
	5	Reconciliation/forgiveness → partial/ambiguous	SOC	0.066	2.6
	+ 6 more (community salience, social emphasis, reader competence, individualization, psych. depth, . . .)				
Gemini	1	Protagonist social trajectory → expands	SOC	0.058	5.0
	2	Balance of speech → primarily direct	PER	0.119	3.6
	3	Global narrative schema → siege/ordeal	EVT	0.037	3.5
	4	Naming practice → named personal name	AGENT	0.033	3.2
	5	Global chronological structure → frequent flashbacks	TMP	0.060	3.1
	+ 6 more (secondary char. density, batch intro, authority stance, setting mood, community, closure)				
DeepSeek	1	Narrator presence/visibility	PER	0.153	4.1
	2	Emotional expression → behavioral cues	AGENT	0.069	3.6
	3	Plot vs. atmosphere orientation	SIT	0.117	2.9
	4	Backstory placement → evenly interleaved	TMP	0.096	2.7
	5	Embedded storytelling scenes	SIT	0.078	2.2
	+ 2 more (seasons/cyclical time, . . .)				
Kimi	1	Character introduction → in-action event	AGENT	0.163	3.7
	2	Narrative entry frame → in medias res	PLT	0.035	3.0
	3	Explicit trait labeling → no	AGENT	0.136	2.0

Table 17: Per-source fingerprint features, showing the top 5 (or all, if fewer) ranked by uniqueness ratio. SHAP = mean class SHAP importance; Uniq. = uniqueness ratio vs. next-best class.

Prompt used to generate benchmark writing prompts from source stories

```
You are a creative-writing instructor generating **writing prompts** for published short stories.

For the story below, craft one prompt that:
1. Begins with **"Write a short story"** (exact phrase).
2. Continue the sentence with **"where...", "about...", "following...", or "from the perspective
   of..."** and then introduce at least one key character or setting by name. This keeps the
   opening grammatically smooth.
   - If a first-person narrator is unnamed, refer to them as "the narrator" without inventing a
     new name. Otherwise, provide their name explicitly.
3. Conveys the story's distinctive **essence / theme / style**, giving the writer a clear sense of
   mood and direction, and includes **some concrete details** (character, location, object,
   striking event) as needed - don't overload with minutiae.
4. Offers enough narrative guidance to get the writer started (situation + conflict or question to
   explore) yet leaves room for their own twists.
5. Do **not** address the reader in second person; keep the prompt in third-person imperative (no
   "you/your").
6. Avoid vague hedge words (*maybe*, *perhaps*, *consider*) **and** absolutely do NOT use
   comparison phrases or qualifiers such as *like*, *much like*, *similar to*, *reminiscent of*,
   *in the style of*. Refer to the concrete names / details directly. Do not invent character
   names that do not appear.
7. Single paragraph <= 120 words.

Return ONLY the prompt text--**no extra commentary**. If anything else is included, keep it on the
same line separated by a single space.

STORY TO ANALYSE:
batch_text
```

Figure 6: Prompt used to generate benchmark writing prompts from source stories

We do not use a single writing template; each story is generated from a standalone narrative prompt in the Books3-derived prompt set, and the same prompt text is given to all six sources. Figure 7 shows one representative validation prompt.

Representative benchmark story-generation prompt

```
Write a short story following Dr. Temperance Brennan, a bioarchaeologist, as she reluctantly lends her
expertise to a Charlotte police investigation involving two cases of severely burned human remains. Delve into
 her meticulous process of analyzing charred bones, the shocking medical link she uncovers connected to Dr.
Keith Millikin and his patients, and how this unexpected plunge into forensic work challenges her academic
detachment, forever altering her perspective on the living and the dead, particularly after a personal loss
within the medical examiner's office. Your story must be approximately 18000 words long.
```

Figure 7: Representative benchmark story-generation prompt

After story generation, each story is converted into a JSON template with a single extraction prompt. Figure 8 shows the exact markdown template.

Prompt for NarraBench template extraction

```
You are a narrative analysis expert. Extract a comprehensive outline from the provided narrative text by
answering the questions below. Analyze the text systematically and provide specific evidence for each element
identified.

## Instructions
- Be objective and avoid interpretation beyond what the text explicitly or implicitly conveys
- Use null for information not present in the narrative or fields with no applicable items
- For trajectories and sequences, use arrows (->) to show progression: "state1 -> state2 -> state3". Remember,
   trajectories and sequences are not always linear, but always have to be complete.
- Keep descriptions concise but specific.
- **Scale guidance**:
  - **Global** fields require story-level analysis across the entire narrative
  - **Local** fields require scene-level or moment-specific analysis (indicate which scene/moment when
    relevant)

## Narrative
narrative_text

## Extraction object schema
Return a JSON object with this structure:

- `story`
  - `agents`
    - `major_characters`: list only major characters who drive the plot or have significant narrative
      importance
      - each item includes:
        - `name` `(string) [GLOBAL]`: use the character's full name as-is
        - `role` `(string) [GLOBAL]`: narrative and functional role, max 2 short clauses
        - `attributes` `(array of strings) [GLOBAL]`: descriptive traits or phrases
        - `emotion_trajectory` `(string) [GLOBAL]`: initial emotion -> progression -> final emotion
        - `motivation_trajectory` `(string) [GLOBAL]`: initial motivation -> progression -> final motivation
        - `trope` `(string) [GLOBAL]`: trope or archetype, if any
    - `supporting_characters`: meaningful minor characters who are not central to the plot
      - each item includes:
        - `name` `(string) [GLOBAL]`
        - `description` `(string) [GLOBAL]`: one-line role/significance summary
  - `social_network`
    - `relationships` `(list) [GLOBAL]`: enduring bonds, formatted as `A-B: relationship type and quality`
  - `events`
    - `sequence`: ordered list of concrete beat-level events `[LOCAL]`
      - each item includes `who`, `where`, `what`, and `when`
    - `causality` `(list) [GLOBAL]`: causal links formatted as `event1 -> event2: explanation`
    - `narrative_schema` `(string) [GLOBAL]`: higher-level pattern such as quest, revenge, or coming-of-age
  - `plot`
    - `themes` `(list) [GLOBAL]`
    - `summary` `(string) [GLOBAL]`: 2-3 sentence plot summary
    - `moral` `(string) [GLOBAL]`: one sentence if signaled; else `null`
    - `central_obstacle` `(string) [GLOBAL]`
    - `central_conflict` `(string) [GLOBAL]`
    - `narrative_archetype` `(string) [GLOBAL]`
    - `plot_arc` `(string) [GLOBAL]`: e.g., `rising action -> climax -> falling action`
  - `setting`
    - `locations` `(list) [LOCAL/GLOBAL]`: include scene-level and story-level locations, with scope noted
    - `time_period` `(string) [GLOBAL]`
    - `atmosphere` `(string) [GLOBAL]`

- `discourse`
  - `revelation`
    - `suspense` `(string) [GLOBAL]`: what key information is withheld?
    - `curiosity` `(string) [GLOBAL]`: what causal antecedents are withheld?
    - `surprises` `(list) [GLOBAL]`: what was revealed, and when?
  - `temporal_order`
    - `structure` `(string) [GLOBAL]`: choose `linear`, `nonlinear`, or `mixed`
    - `duration` `(string) [GLOBAL]`: overall time span
    - `flashbacks` `(list) [LOCAL]`: note which scenes are flashbacks
    - `time_jumps` `(list) [LOCAL]`: ellipses or leaps in time/place, with scene references
    - `scene_duration` `(list) [LOCAL]`: approximate duration of major scenes

- `narration`
  - `perspective`
    - `point_of_view` `(string) [GLOBAL]`: choose `1st person`, `2nd person`, `3rd person limited`, or
      `3rd person omniscient`
    - `focalization` `(list) [LOCAL]`: whose perspective we occupy by scene/section
    - `dialogue_speakers` `(list) [LOCAL]`: named speakers by scene/section
  - `style`
    - `allusions` `(list) [LOCAL]`: allusions plus scene/section context
    - `figurative_language` `(list) [LOCAL]`: examples plus scene context
    - `imagery` `(list) [LOCAL]`: vivid sensory descriptions plus scene context
    - `sentence_complexity` `(string) [GLOBAL]`
    - `evaluative_language` `(list) [LOCAL]`: judgmental/evaluative language plus scene context
```

Figure 8: Prompt for NarraBench template extraction