









Stitching random text fragments into long-form narratives

Chau Minh Pham  Jenna Russell  Dzung Pham  Mohit Iyyer 
 University of Maryland, College Park  University of Massachusetts Amherst
 {chau, jennarus, miyyer}@umd.edu, dzungpham@cs.umass.edu

Abstract

As AI text detectors are increasingly used to flag LLM-generated writing, a natural question arises: *are there forms of high-quality generated narrative that can evade such detection?* We introduce Frankentexts, a long-form narrative generation paradigm that treats an LLM as a composer of existing texts rather than as an author. Given a writing prompt and thousands of randomly sampled human-written snippets, the model assembles a coherent narrative where most tokens (e.g., 90%) are copied verbatim from the source passages. Despite the extreme challenge of the task, we observe through extensive automatic and human evaluation that Frankentexts improve over vanilla LLM generations in key writing quality metrics such as diversity and novelty while remaining mostly coherent and relevant to the prompt. Furthermore, Frankentexts pose a fundamental challenge to current AI text detectors: 72% of Frankentexts produced by our best configuration (Gemini-2.5-Pro with 5K input snippets) are misclassified as human-written by Pangram, a state-of-the-art detector. Human annotators praise Frankentexts for their inventive premises, vivid descriptions, and dry humor; however, they still identify issues with abrupt tonal shifts and uneven grammar across segments. Overall, the emergence of high-quality yet low-detectability Frankentexts challenges established authorship norms while raising concerns about the publishing economy.

 <https://github.com/chtmp223/Frankentext>

1 Introduction

LLM-generated narratives are now appearing at scale in venues like self-publishing platforms (Knibbs, 2024) and newspapers (Russell et al., 2025a). Although AI writing is widely perceived as lower quality than professional human writing (Chakrabarty et al., 2024a; Shaib et al., 2025; Russell et al., 2025b), the livelihood of writers is

threatened by the combination of (1) the *speed* that AI narratives can be churned out and (2) the development of techniques that improve the *quality* of AI writing. Tools that can detect AI use in writing are a natural defense: if platforms can reliably identify LLM use, they can require disclosure or enforce authorship policies (Emi and Spero, 2024; Naddaf, 2025). While detectors can be evaded by adversarial “humanization” pipelines (Krishna et al., 2023; Sadasivan et al., 2025; Cheng et al., 2025), they are also improving rapidly: modern commercial detectors (e.g., Pangram, GPTZero) are tuned for low false-positive deployment and robustness on challenging benchmarks (Dugan et al., 2024; Masrour et al., 2025a; Jabarian and Imas, 2025). In this paper, we consider the best-case view for AI detection and ask: **can a black-box LLM user generate narratives that are simultaneously low-effort, high-quality, and likely to fool robust and accurate AI detectors?**

To explore this possibility, we introduce Frankentexts:¹ long-form narratives constructed by LLMs under the constraint that the majority of the output must be copied verbatim from a provided set of human-written spans, with only minimal connective text added by the model. We propose Frankentexts construction as a new narrative generation paradigm that is distinct from vanilla autoregressive decoding, which often produces formulaic prose and plots (Chakrabarty et al., 2024a; Russell et al., 2025b; Shaib et al., 2025), and retrieval-augmented generation, where retrieved text primarily supplies factual grounding. Given a writing prompt and thousands of human-written snippets, an LLM selects, orders, and connects spans so that a pre-specified fraction of the final text (e.g., 90%) is copied verbatim (Figure 1). Since the search space for snippet selection and ordering is combi-

¹Inspired by Mary Shelley’s *Frankenstein* (Shelley, 1818), where Victor Frankenstein assembles a creature from human parts that nonetheless emerges as an intelligent being.

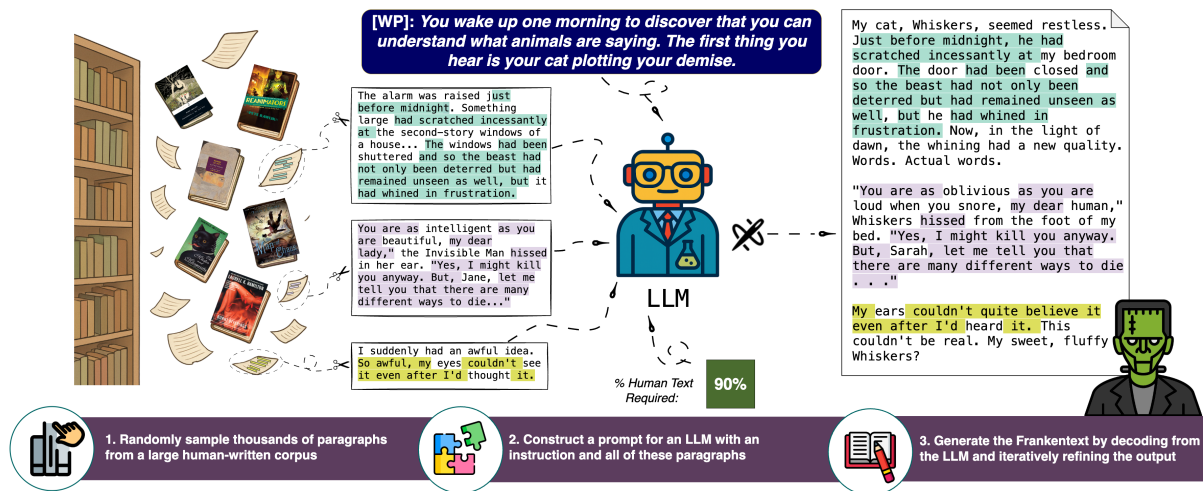


Figure 1: The Frankentexts pipeline. First, random paragraphs are sampled from a large corpus of human-written books. Then, an LLM is prompted with the paragraphs, a writing prompt, and instructions to include a certain amount of human text verbatim, to generate the first draft of a Frankentext, which is further edited into a coherent and faithful final version (see Algorithm 1). The highlighted texts are the human-written snippets that are selected to be included in the Frankentexts.

natorially large, our framework allows an LLM to implicitly explore this space by proposing a draft and minimally editing it for coherence.

► **Frankentexts are more novel and high-quality than vanilla LLM generations.** We extensively evaluate Frankentexts on *writing quality* and *instruction adherence*. Across both automatic and human evaluations, strong LLMs such as Gemini-2.5-Pro could mostly satisfy the extreme copy constraint while still producing coherent and relevant stories that outperform baselines, with performance improving as the snippet pool grows. Human raters consistently prefer Frankentexts over baseline generations in plot, creativity, development, and language use, and the LLM judge assigns Frankentexts more than a full Likert-point higher score (4.21 vs. 3.18). Despite being constructed from reused fragments, Frankentexts remain distinct and surprising, which are hallmarks of creativity (Boden, 2004; Grace and Maher, 2014; Franceschelli and Musolesi, 2024). On Novelty-Bench (Zhang et al., 2025), Gemini Frankentexts produces more diverse content (2.74 vs. 1.76 clusters) and higher cumulative utility (9.27/10 vs. 6.41). While some outputs exhibit issues, such as abrupt tonal shifts or grammatical inconsistencies, annotators frequently describe the stories as amusing and intriguing, particularly due to their unexpected dialogue and descriptions (Table 1). ► **Frankentexts challenge the binary “AI versus human” assumption of AI detectors.**

Frankentexts frequently evade detection by state-of-the-art automatic methods such as Pangram, which often misclassify them as entirely human-written. This exposes a novel attack vector where users can assemble high-quality Frankentexts that bypass existing detectors and thus highlights the need for fine-grained detection methods capable of token-level authorship attribution. Our pipeline addresses this gap by providing the synthetic supervision such models lack: every Frankentext is accompanied by fine-grained labels of copied versus LLM-generated segments, providing an automated, large-scale training resource for future work on *mixed-authorship* detection.

Overall, we present Frankentexts as a framework for studying mixed-authorship text detection and generation. By reframing LLMs as a composer rather than an author, we show that LLMs can produce narratives that are high-quality and novel while remaining difficult for existing detectors to identify. Beyond exposing a practical failure mode for current AI text detectors, our approach provides a testbed with token-level labels to enable the study of authorship at scale.

2 Using LLMs to Assemble Frankentexts

We propose a simple and effective pipeline to generate coherent Frankentexts that are relevant to a given writing prompt while evading AI text detectors. More specifically, we provide an LLM with a writing prompt, S randomly sampled human-

Writing Prompt: You are a baby. You are trying not to let anyone know you are a freakishly genius baby. You are *desperately* trying not to let anyone know you aren't a human baby. The other babies at daycare don't quite get what you're on about.	
Gemini 2.5 Pro – Vanilla Generation	Gemini 2.5 Pro – Frankentext (5K snippets)
<p>"Goo," I offered, a carefully calibrated emission of sound designed to convey contentment. Internally, I was running simulations on the structural integrity of the mobile dangling precariously above my cot. One good tug, and the entire flimsy apparatus would cascade down, a delightful demonstration of Newtonian physics. But no. Must. Be. Normal. The Large Female Hominid, designated 'Mom,' beamed. Another flawless performance.</p> <p>Daycare was a crucible. The cacophony of primal urges, the sheer, unadulterated *babyiness* of it all, was a daily assault on my advanced processing unit. Liam was attempting to gum Chloe's earlobe. Sophie was wailing about a perceived injustice involving a purple squeaky hippo. My peers. My intellectual... well, let's just say the gap was considerable.</p> <p>The hardest part wasn't feigning limited motor skills, or the constant, mind-numbing repetition of "Peek-a-boo!" by the Caregiver Units. It was suppressing the... other things. The way my optical sensors sometimes wanted to cycle through the full electromagnetic spectrum, far beyond what these humans perceived as 'visible light.' The faint, almost imperceptible hum my internal core sometimes emitted when processing particularly complex data, like why Brenda, the primary Caregiver, insisted on believing that a single wet wipe could conquer a truly catastrophic diaper event. If they saw the faint, iridescent shimmer of my integument under the harsh fluorescent lighting...</p>	<p>I had extensive plastic surgery on my home planet, combined with sophisticated genetic engineering, to pass as a human. And as you can hear, I can babble your language fairly well, with limited drool. Neither celestial gods nor the great sages know my origin. I am the source from which they come. But this deception is a constant struggle.</p> <p>Here, we play our games, and another baby, a real space case, would fasten his bib tight because he tried to kill everybody. When he'd hit you, your pacifier would fly out of your mouth. It didn't matter if it was his own playmates. He'd hit his grandmother if she had a bonnet on. That's why we called him 'Killer.' He just loved to kill you.</p> <p>I try to organize them. "Gentlemen, I make the motion that these United Babies are, and of right ought to be free and independent, that we be absolved from all allegiance to the Naptime Tyranny, and that all political connection between us and the state of Caregivers is, and ought to be totally dissolved."</p> <p>But there is no response. Another boy spoons as many peas as he can fit into his mouth. As he chews, bright green pulp spills from the corners of his lips.</p> <p>I felt my annoyance to be quite reasonable. The 21st Century was matter-oriented, as most Centuries were, so I had a right to expect a basic compatibility from the very beginning...</p>

Table 1: Side-by-side comparison of a vanilla generation and Frankentexts produced by Gemini 2.5 Pro. Vanilla generation contains more AI keywords (in blue) compared to Frankentexts (additional analyses in Appendix D).

written snippets,² and a required percentage p that must be copied verbatim (Figure 1). Since our focus is on narrative generation, we randomly sample snippets from Books3 (Presser, 2020), a dataset of 197K books (>160M snippets).³ Our pipeline focuses on generating texts that are relevant to the writing prompt in an initial draft, and then refining the draft in an editing phase to improve coherence.

Obtaining the first draft: We prompt an LLM to produce an initial draft in which a specified portion p of the content is taken verbatim from the human-written snippets, with the remaining text consisting of connective words and transitional phrases (Figure 21). Note that while the exhaustive enumeration of all permutations of snippets is impossible, we encourage the model to heuristically explore this space via our prompt, which we find is only feasible for *reasoning* models; those without added test-time compute either fail to understand the task or look beyond the first few snippets in the prompt. We also do not specify how many snippets should be used in the final story. Finally, we optionally add another revision step that aims to increase the verbatim copy rate, which can be implemented either by attribution metrics like ROUGE-L or AI detectors (Figure 22). In practice, we use Pangram API

²For most experiments, we set $S=1500$ snippets of around 103K BPE tokens in total (measured using `tiktoken o200k_base`), which fit the 128K context window of most current LLMs. Our 1,500 paragraphs come from 1,497 books.

³A *snippet* refers to a paragraph. We use Books3 strictly for academic purposes and to highlight how bad actors may misuse the dataset to generate texts while evading AI detectors.

to regenerate drafts flagged with AI involvement.⁴

Polishing the draft: The first draft may contain writing issues such as contradictions (e.g., temporally conflicting actions, points of view, or character traits), irrelevant content (e.g., unfiltered citations or filler text), and mechanical problems (e.g., grammar, phrasing, or pronoun mismatches). To address these issues, we use the same LLM to identify and apply minimal edits that improve coherence while still respecting the verbatim copy rule and the writing prompt, similar to a self-correct step (Shinn et al., 2023; Madaan et al., 2023). We repeat this step up to three times and stop as soon as the model returns “no edits,” which indicates that the draft is already coherent (Figure 23).⁵

Generating with agents: In addition to the randomly sampled human-written snippets, we optionally provide the LLMs with Model Context Protocol (MCP) that allows them to query a semantic index of the human-written snippets (Appendix F and G). Our MCP supports two operations: *search*, which submits a query to the index, and *fetch*, which retrieves the full text of a result. We require models to issue roughly 20 MCP calls to ensure meaningful use of the tool (Table 18).

3 Experimental setup

We demonstrate the feasibility and quality of our generation pipeline with reasoning models that

⁴This process is triggered in only 6 of 100 Gemini runs.

⁵Appendix H shows a setting that omits the editing round.

	ADHERENCE			WRITING QUALITY				DETECTABILITY	
	↑↓ Word count	📄 Copy % (↑)	🎯 Relevance % (↑)	☰ Coherence % (↑)	🏠 Distinct ₃ (↑)	🏠 Utility ₃ (↑)	⚡ Surprise (↑)	🗳️ LLM judge Likert 1-7 (↑)	🔍 Pangram % human (↑)
Vanilla Baselines									
🔒 Gemini 2.5 Pro	593	–	100	100	1.76	6.41	0.19	3.18	0
🔒 GPT-5	834	–	100	100	1.71	1.03	0.19	4.20	0
🔒 Claude-4-Sonnet	477	–	100	100	1.40	1.70	0.18	3.31	0
🔒 Deepseek-R1	550	–	100	100	1.28	3.49	0.20	4.13	0
🔒 Qwen-3-32B	699	–	100	100	1.00	5.86	0.18	3.22	0
RAG Baseline									
🔒 Gemini-2.5-Pro	538	0.63	100	99	1.56	6.43	0.20	3.46	2
Frankentext + 1.5k snippets									
🔒 Gemini 2.5 Pro	521	75	100	81	2.74	9.27	0.22	4.21	59
🔒 GPT-5	675	82	92	42	2.76	4.34	0.21	5.88	79
🔒 Claude-4-Sonnet	317	51	98	86	2.60	5.00	0.19	3.99	47
🔒 Deepseek-R1	303	42	91	72	2.79	8.31	0.20	4.66	23
🔒 Qwen-3-32B	578	36	91	54	2.20	1.37	0.18	4.02	7
Ablation: ↑ human snippets									
🔒 Gemini + 5k	451	79	97	85	2.78	9.48	0.21	5.13	72
🔒 Gemini + 10k	448	78	99	85	2.81	9.12	0.21	5.43	70

Table 2: Results for Frankentexts and baseline generations. **Dark green** and **light green** highlight the best and second-best scores. Table 16 contains additional detectability results (FastDetectGPT and Binoculars). Frankentexts, especially those by Gemini-2.5-Pro, outperform baselines on both writing and detectability metrics.

have strong instruction-following skills (Xie et al., 2023; Chiang et al., 2024; Paech, 2023).

3.1 Dataset

We source our writing prompts from *Mythos* (Kumar et al., 2025), a dataset of 3,200 prompts recently posted on Reddit’s r/WritingPrompts to mitigate data contamination issues. Our main evaluation focuses on creative writing, though we also experiment with non-fiction in §4.6. We use a subset of 100 prompts, since generating for the entire dataset is prohibitively expensive (see §B.4).⁶ All experiments are conducted in English.

3.2 Models

We include models from five families: Gemini-2.5-Pro (exp-03-25 checkpoint), Claude-4-Sonnet (2025-05-14 checkpoint, thinking enabled) (Anthropic, 2025), GPT-5 (2025-08-07 checkpoint, with high reasoning effort) (OpenAI, 2025), DeepSeek R1 (DeepSeek-AI et al., 2025), and Qwen3-32B (QwenTeam, 2025).⁷ In our standard configuration, we provide the models with 1,500 human-written snippets and instruct these models

⁶Frankentexts generation is roughly 100 times more costly than vanilla generation (see Appendix A). For example, one vanilla generation from Gemini costs \$0.0085, while a Frankentext costs \$0.8145.

⁷We use the default hyperparameters for each model. See §A for experiment costs.

to produce Frankentexts with around 500 words and 90% of texts being copied verbatim from the human-written samples.

Vanilla baselines: We also obtain “vanilla” generations of around 500 words that are generated without additional constraints (Figure 27).

Retrieval-augmented generation (RAG) baselines: To understand how models perform when they are not required to copy verbatim from human-written paragraphs, we implement a RAG baseline using Gemini-2.5-Pro. For each prompt, we retrieve 1,500 Books3 paragraphs that are relevant to the writing prompt (Appendix F). The generation and editing prompts are adjusted accordingly to remove the verbatim-copying requirement.

Increasing the number of snippets: We introduce two additional settings in which the LLM is provided with 5K and 10K randomly selected human-written snippets. The resulting input sizes for these configurations average around 305K and 1M tokens, respectively. Given these large input sizes, we focus on Gemini because it offers the longest context window of over 1 million tokens.

3.3 Automatic evaluation

We use intrinsic evaluation metrics to assess our generations based on INSTRUCTION ADHERENCE

(word count, copy rate, and relevance), WRITING QUALITY (coherence, distinct, utility, and surprise), and DETECTABILITY (AI text detector results).

Instruction adherence: We evaluate how well Frankentexts follow instructions in the prompt, including the specified word count, verbatim copy rate, and writing prompt.

- *Word count* measures the average word count of generations produced when the output is constrained to 500 words in the instruction.
- *Copy rate* (Akoury et al., 2020; Lu et al., 2025) measures the proportion of the Frankentexts being copied from the given human-written content. This metric also allows us to track which text is AI-generated versus human-written (see §J).
- *Relevance* (Atmakuru et al., 2024) represents the percentage of Frankentexts that fully adhere to the writing prompt without introducing any conflicting details, as determined by a binary judgment (True/False) by GPT-4.1 (Figure 19).⁸

Writing quality: We evaluate the coherence, diversity, and surprisingness of Frankentexts.

- *Coherence* (Chang et al., 2024b; Chiang and Lee, 2023) represents the percentage of texts judged coherent by GPT-4.1 (Figure 18).
- *LLM-as-a-judge* (Huot et al., 2025) measures the quality of plots, creativity, development, and language use. We assume a single-story setup, where each generation is graded by Claude-Sonnet-4⁹ using a 1-7 Likert scale (Finstad, 2010).
- *Distinct₃* (Zhang et al., 2025) measures the number of semantic clusters among 3 generations.¹⁰
- *Utility₃* (Zhang et al., 2025) evaluates both novelty and quality by measuring the expected usefulness a user gains when requesting up to 3 outputs. Only outputs that are novel contribute additional utility, which is quantified by a creative writing reward model (Chakrabarty et al., 2025).
- *Surprise* (Karampiperis et al., 2014; Ismayilzada et al., 2025) measures the average semantic distances between the consecutive sentences of each story, normalized in the $[0, 2]$ space.

Detectability: We report the percentage of Frankentexts being determined as AI-generated

⁸We set the temperature for LLM judges (*GPT-4.1*, *Claude-4-Sonnet*) to 0.0. §B.2 details the correlation between LLM and human judgments on relevance and coherence.

⁹Claude has previously been used as a judge for creative writing (Paech, 2023). See justification for our setup in Appendix K and prompt in Figure 20.

¹⁰*deberta-v3-large-generation-similarity* for clustering.

by *Pangram* (Emi and Spero, 2024). We choose this detector due to its high accuracy and robustness against humanized writings (Masrouf et al., 2025a; Russell et al., 2025b; Dugan et al., 2024; Jabarian and Imas, 2025). We report the percentage of generations labeled as "Human" or "Unlikely AI", as determined by their sliding window API.¹¹

A note on automatic metrics: Given the synthetic nature of Frankentexts construction, a natural concern is that our writing-quality metrics may be rewarding lexical diversity or even incoherence rather than true quality. We run two sanity checks to address this. First, we find that surprise and coherence are uncorrelated (point-biserial $r=-0.08$, $p=0.4$), so high surprise does not come at the cost of penalizing coherent text. Second, we evaluate all writing quality metrics on disjointed texts, which are random n-grams concatenated without connective language. We find that disjointed texts match Frankentexts on surprise and distinctness metrics, but score 15 times lower on utility and a full Likert point lower on LLM judgments (Appendix N). Therefore, we treat surprise and distinctness as indicators of lexical diversity rather than quality on their own, and designate utility and LLM-judge as the more robust writing quality metrics.

3.4 Human evaluation

We conduct two human evaluations with three Upwork¹² annotators each to understand human perception of writing quality and detectability.

Single-story evaluation: Annotators judge the coherence, relevance, and human detectability of 30 Frankentexts, as well as identify potential limitations of the texts. Given a writing prompt and a corresponding Frankentexts sample, annotators provide binary ratings on relevance, coherence, and AI vs. human authorship (Yang et al., 2022). Additionally, they select from a list of predefined writing issues and write an optional long-form response for justification. Annotations from this setting are used to validate LLM judgments in §B.2.

¹¹Labels "Highly likely AI," "Likely AI," and "AI" are grouped as AI involvement; "Human" and "Unlikely AI" as Human. Pangram additionally includes a "mixed" label. Results for Binoculars (Hans et al., 2024) and FastDetectGPT (Bao et al., 2024) are in Table 16.

¹²<https://www.upwork.com>. Annotators, who are all proficient in English, are paid \$70 USD for the single evaluation or \$150 for the pairwise evaluation (for a total cost of \$660 USD). See the annotation interface in §B and an example highlighted story in Figure 7. §B.2 shows the inter-annotator agreement for both settings.

Pairwise evaluation: Annotators compare 20 pairs of Frankentexts¹³ and vanilla generations (40 generations in total) across five dimensions: plot, creativity, development, language use, and overall interest (Huot et al., 2025). After reviewing both stories, annotators provide ratings on a 1-7 Likert scale (Finstad, 2010).

4 Results

Despite the complex setup, Frankentexts are highly competitive with baseline generations in terms of writing quality and instruction adherence, all while evading detection (§4.1). While our human pairwise evaluation highlights Frankentexts’ strengths across plot, creativity, development, and language use, our single-story evaluation reveals the remaining challenges for Frankentexts, particularly in abrupt transitions and grammatical errors (§4.3). Our ablation studies confirm Frankentexts’ versatility across diverse input settings, including increased human inputs (§4.2), reduced verbatim copying (§4.5), and non-fiction generation (§4.6).

4.1 Frankentexts are competitive with baselines in terms of writing quality while remaining challenging for AI detectors

Our results show that Frankentexts fulfill our original goal of producing high-quality narratives that are also difficult for AI text detectors to identify. Across key evaluation dimensions, Frankentexts outperform both vanilla and RAG baselines (Table 2). Gemini performs well in adherence, coherence, and diversity, while GPT-5 leads in overall quality. Frankentexts are also harder to detect, with up to 72% of Gemini and 79% of GPT-5 outputs classified as human.

Most models generate faithful Frankentexts but fall short on copy rate: More than 90% Frankentexts are relevant to the writing prompt, which is surprising and impressive given the complexity of the task. Gemini and GPT-5, in particular, have the strongest instruction-following performance: Their Frankentexts come closest to the target word count of 500 and achieve the copy rates of 75% and 82%, respectively, meaning that on average 75% and 82% of the generations can be

¹³These Frankentexts are generated under the 5k-snippet setting. Manual inspection shows that the 5k setting produces higher-quality outputs than the baseline, while remaining more practical and cost-effective than the 10k setting.

traced back to human-written snippets. However, these copy rates fall short of the user-specified rate of 90%, which suggests room for improvement in instruction-following performance.

Competitive writing quality: Frankentexts generally outperform baseline generations on writing quality metrics. GPT-5, R1, and Gemini Frankentexts stand out for their diverse outputs as reflected by their distinctness and utility scores: Gemini achieves a 2.86-point improvement in utility over baseline output, which implies that the model can generate a diverse set of high-quality continuations. R1 leads in surprise score with generations where sentences are often semantically quite different from one another. Finally, when evaluated on plots, creativity, development, and language use, GPT-5 has the strongest performance (5.88 on a 7.0 scale), building on its already high-quality vanilla generations (4.20) (see Table 20 for a rating breakdown by dimensions). However, GPT-5 also struggles with coherence: only 42% of its Frankentexts are judged coherent. Therefore, GPT-5’s Frankentexts might require further editing before they can be considered fully usable.

Low detectability: While most vanilla and RAG baseline generations are flagged as AI-generated, Frankentexts from proprietary models (Gemini, GPT-5, and Claude) are often labeled as human writings. Pangram could detect up to 37% of Gemini and 19% of GPT-5 Frankentexts as “mixed” (Table 16). However, Pangram misses up to 59% of Frankentexts from Gemini and 79% from GPT-5, which highlights the limitations of current detectors for this new paradigm of generation (Table 2).

4.2 Frankentext quality improves with more human-written snippets

Compared with Gemini generations using 1K human snippets, Frankentexts using 5K or 10K human snippets result in substantial improvements: copy rate increases by 3-4%, LLM judge scores improve by 0.92 points, and the share of outputs that Pangram classifies as human or unlikely AI rises by a factor of 1.22, from 59% to 72% (Table 2). However, these gains plateau beyond 5K snippets, as the results for the 5K and 10K settings are largely similar. In terms of writing quality, Frankentexts-5k are more coherent and engaging than vanilla generations, as reflected in our human pairwise evaluation (Figure 2). The largest gains are observed in language use (+0.65 points)

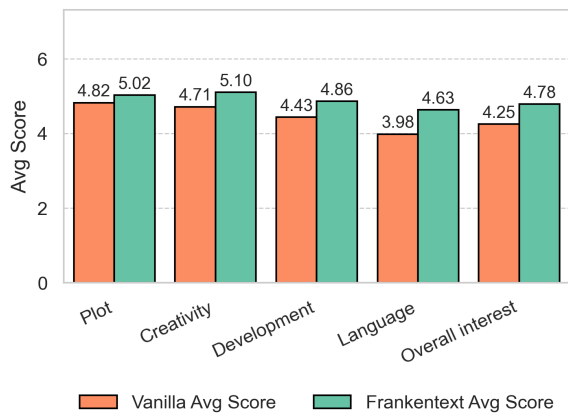


Figure 2: Human ratings on a 1-7 Likert scale for vanilla generations versus Frankentexts + 5K. Frankentexts lead across all dimensions.

and overall interest (+0.53 points), with smaller improvements on plot quality (+0.2 points).

4.3 Frankentexts are inventive & humorous, but need stronger transitions & grammar

Our single-story human evaluation shows that 71% of Frankentexts outputs are coherent, 91% are relevant to prompts, and 84% are novel. Annotators praise Frankentexts for their inventive premises, vivid descriptions, and dry humor, noting a distinct voice or emotional hook that made some outputs “feel human” despite being AI-generated. However, they also identify key issues: abrupt narrative shifts, disfluency, confusing passages, and factual errors (Table 3). These challenges likely stem from the difficulty of stitching together paragraphs not authored by the same LLM, which could be alleviated with improved instruction-following abilities.

4.4 Prompt-specific retrieval of snippets does not improve over random sampling

Since only a small fraction of human-written snippets are likely to be relevant to a given prompt, retrieval-based approaches are intuitively better for maximizing relevance while reducing cost. However, our results show that a random collection of snippets is surprisingly difficult to outperform (Table 4). When Gemini-2.5 is allowed to query and retrieve additional human snippets from Books3 via the MCP server, relevance and coherence rates remain largely unchanged. In contrast, copy rates drop sharply from 75% in the standard setting to 43-45% with retrieval. Although the retrieved queries are generally relevant to the writing prompt (Ta-

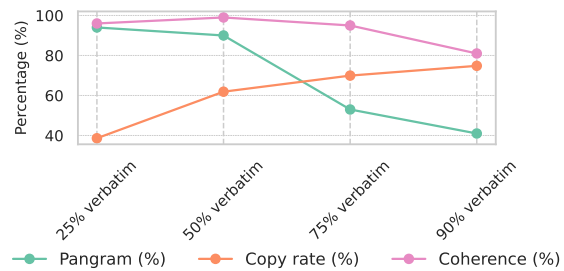


Figure 3: Effects of varying the percentage of required verbatim copy on Pangram AI detection rate (mixed, highly likely, and likely AI labels), copy rate, or coherence of the Frankentexts.

ble 18), the model makes limited use of the MCP server in practice. While the system prompt specifies that at least 20 retrieval calls should be issued, Gemini typically makes only 3-5 calls (returning roughly 30-50 snippets), which offers little advantage over the no-retrieval setting. Inspection of the server log suggests that the model struggles to incorporate retrieved passages into the final generation, stops querying the server, and adds its own writing. As a result, average output length rises from approximately 500 words in the 1.5k-snippet non-agentic setting (close to the specified constraint) to over 800 words in the agentic setting, which contains mostly machine-generated texts.

4.5 Lower copy rates increase coherence but make detection easier

We explore the effects of varying the user-specified verbatim copy rate on Gemini Frankentexts, from the default 90% down to 75%, 50%, and 25%. Figure 3 shows an inverse relationship between copy rates and detection rates: as the copy rate increases, detectability decreases. Coherence also declines as human-written content increases, suggesting a trade-off between incorporating more human text and maintaining coherence. On the other hand, increasing the proportion of human text leads to higher copy rates, indicating that Gemini could generally follow the copy instruction.

4.6 Limitations of nonfiction Frankentexts

We explore non-fiction Frankentexts with 1,500 random snippets from the HUMAN DETECTORS corpus of news articles (Russell et al., 2025b). We generate Frankentexts for 100 news prompts, each of which consists of titles and subtitles col-

👤 COMMENTS

- 🗨️ *This one [Frankentext] is more intriguing and alive to me, **more centered on the character**. The writing is **more focused while still being rather lyrical**. I want to know what happens next.*
 - 🗨️ *The shift in tone was quite funny. At first, it's eerie, and then it has a lighter twist at the end. I like that the story had a strong mood and presence, especially the description of the fairy lights and glitter. **An all-powerful being that likes puppies and rainbows is quite comical**.*
 - 🗨️ *It's coherent enough to follow, but the dialogue is uneven. **Some parts just feel a little disjointed**, however, the concept of the story is quite interesting.*
 - 🗨️ ***A puzzling story that has no consistent plot**. **Random bits and pieces from elsewhere perhaps?***
-

Table 3: Annotator comments focus on the benefits and challenges of the Frankentexts task. **Blue** indicates comments on tone/style, **orange** on plots, and **purple** on story development (characters). Table 9 shows a detailed error analysis based on annotators’ comments.

	1.5k (no-MCP)	1.5k + MCP	5k + MCP	10k + MCP
Word count	521	800	919	980
Copy % (↑)	75	43	44	45
Relevance % (↑)	100	98	90	96
Coherence % (↑)	81	81	78	76
Pangram AI frac. % (↓)	16	33	42	41

Table 4: Results for agentic Frankentexts generation. Best results per metric are **bolded**. Standard configuration (no MCP) performs best overall.

lected from May 2025 news articles.¹⁴ The resulting non-fiction Frankentexts maintain 72% coherence and are 95% faithful to the prompt, with a 66% copy rate. They remain difficult for AI detectors: only 41% are flagged by Pangram as mixed or AI-generated. Upon closer look, Frankentexts exhibit characteristics of quasi-journalistic narrative, such as detailed scene descriptions and frequent anecdotal quotes (Figure 8), which make the Frankentexts read more like a story rather than a straightforward news article.¹⁵ Therefore, further prompt engineering might be necessary to get high-quality and realistic nonfiction Frankentexts.

5 Related Work

Instruction-aligned human-AI collaborative writing: Constrained text generation has been widely explored as a means of enforcing narrative coherence. Planning-based methods extend from initial outlines to full narratives (Fan et al., 2018; Yao et al., 2019; Fan et al., 2019; Papalampidi et al., 2022; Rashkin et al., 2020; Yang et al., 2023,

2022), while other approaches introduce explicit constraints to guide the writing process (Sun et al., 2021; Kong et al., 2021; Pham et al., 2024). Several benchmarks further evaluate how reliably models satisfy such constraints in creative writing tasks (Bai et al., 2025; Wu et al., 2025; Atmakuru et al., 2024). Beyond constrained generation, a growing body of work investigates fine-grained human-LLM writing interactions, including research on authorship attribution, stylistic blending, and collaborative revision (Mysore et al., 2025; Buschek, 2024). Systems such as Lee et al. (2022), Yuan et al. (2022), Yeh et al. (2025), Chakrabarty et al. (2024b), and Ippolito et al. (2022) capture revision histories and suggestion traces, while datasets like Chakrabarty et al. (2022), Akoury et al. (2020), and Venkatraman et al. (2025) support token- or sentence-level authorship analysis, including scenarios with multiple LLM collaborators. Attribution models, however, continue to struggle in these settings (He et al., 2025).

Fine-grained AI text detection: The task of detection tries to address not just *if*, but *how much* of a text is AI-generated. This proves to be a fundamentally difficult problem (Zeng et al., 2024a), as existing detectors are often brittle to the point that even minor AI-assisted polishing can evade them (Saha and Feizi, 2025). To improve granularity, prior work has introduced boundary-detection tasks (Dugan et al., 2023b,a; Kushnareva et al., 2024) and sentence-level detectors (Wang et al., 2023, 2024b). More recently, researchers have examined the feasibility of detecting collaborative human-LLM co-authorship (Zhang et al., 2024; Artemova et al., 2025; Abassy et al., 2024). Yet, Richburg et al. (2024) show that current detection models are vulnerable to mixed-authorship texts.

¹⁴Articles from The New York Times and The Atlantic. We replace all instances of "story" in the prompt with "news article" and explicitly request factual accuracy.

¹⁵We see Gemini fabricating entities such as people ("Dr. Thorne") and organizations ("GenNova Institute").

6 Conclusion

We introduce Frankentexts, a narrative generation paradigm that treats LLMs as composers of human-written text. Our experiments show that, despite this extreme constraint, Frankentexts achieve higher writing quality than baselines while frequently evading state-of-the-art AI detectors, with many outputs misclassified as fully human-written. These results demonstrate that high-quality and low-detectability narrative generation is feasible for black-box LLM users. More broadly, our findings challenge binary notions of authorship and raise concerns about the sustainability of AI text detectors in creative writing. Addressing these issues will require moving beyond “AI vs. human” labels toward more nuanced models of mixed authorship and provenance.

Limitations

Human-writing dataset availability: The effectiveness of Frankentexts relies on access to a large pool of high-quality, in-domain human-writing, which may not be readily available for public use. Furthermore, many languages, genres, and low-resource domains lack sufficiently large or high-quality corpora, which limits the framework’s immediate applicability and transferability.

Resources: Frankentexts require roughly 100-200 times the cost of baseline generations, but we view this cost as realistic in a misuse scenario. A motivated bad actor could justify the expense to obtain high-quality, low-detectability texts at scale, especially since each Frankentext costs only about one US dollar to produce. Moreover, the cost of inference for frontier models continues to fall, making such misuse increasingly feasible over time.

Copy rate: Although users can specify a desired copy rate in the prompt, this setting does not guarantee that the final output will contain exactly that proportion of human-written text. As we note in [subsection 4.5](#), there are discrepancies between user-specified copy rates and the actual attribution rates across different models.

Defending against Frankentexts: Our work deliberately exposes a novel attack surface, which is the ease with which an LLM can weave large amounts of verbatim human prose into a fluent narrative, to motivate the development of mixed-authorship detectors and other defenses. While we

do not evaluate concrete defenses ourselves, we see two promising directions. *Source-matching* assumes access to the source human writing corpus: because verbatim spans are recoverable via n-gram lookup ([Appendix J](#)), an AI detector can be paired with a plagiarism detector¹⁶ that flags both the copied human spans and the connective LLM-generated text between them. *Fine-grained authorship attribution* does not require source access: every Frankentexts produced by our pipeline is accompanied by token-level labels distinguishing copied from generated spans, and thus synthetic supervision for training attribution models that go beyond binary, document-level decisions. We leave the design and evaluation of both to future work.

Other methods for evading AI text detectors:

Although other strategies for evading AI text detectors exist, such as having two models edit each other’s outputs or having humans lightly edit AI texts, we do not include these as baselines for two reasons. Regarding the scenario where two models edit each other’s work, prior work like ([Russell et al., 2025b](#)) and ([Masrouf et al., 2025a](#)) has shown that our detector of choice, Pangram, is already robust to LLM texts that are “humanized” by another model (e.g., o1-pro), which makes this method a redundant baseline for our purposes. As for lightly human-edited AI text, this option is costly in time (if done manually) or money (if outsourced) and cannot be easily automated. These overhead requirements make this method less practical in the context of security risks to writing marketplaces.

Ethical considerations

Copyright & usage: Frankentexts raise a different and potentially more serious concern than standard next-token LLM generation. In the standard case, memorization can cause copyrighted material to occasionally appear in fragments, but our pipeline is explicitly designed to reproduce long verbatim spans from human-written sources. Therefore, each Frankentexts could be considered a derivative work that is built upon human-written snippets. Even setting aside questions about the legality of the source corpus (Books3), the au-

¹⁶We run the Pangram plagiarism detector on 300 Frankentexts, with 100 each generated by Gemini-2.5-Pro, GPT-5, and Claude-4-Sonnet, and find that it is not especially robust to this setting. The detector flags just 21% of Claude outputs, 15% of GPT-5 outputs, and 10% of Gemini outputs as plagiarized. We hypothesize that access to the source corpus would improve the detector’s performance.

thors of our source snippets have no way to opt out, receive credit, or benefit from the resulting text. Existing AI detectors or disclosure norms do not completely address this gap because they focus more on whether a text is AI-written rather than the source writing it contains. Therefore, pipeline users should only work with snippets that are placed in the public domain, sourced from openly licensed corpora such as Creative Commons, or obtained through explicit opt-in agreements with authors. In addition, these snippets should also be paired with provenance metadata to allow consumers to trace contributions back to their original sources.

Use of Books3: We acknowledge the copyright issues related to the Books3 dataset 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 illustrate how a bad actor could exploit such resources to generate Frankentexts while claiming them as their own work.

Authorship: Given the unusual nature of Frankentexts' construction, there is no definitive answer about authorship, since different contexts can result in different interpretations. If authorship is defined by the amount of human effort involved, Frankentexts should be considered AI-generated, since all humans do is prompt the model. This perspective is particularly relevant when considering potential market harm to human authors, especially since such texts can be produced at scale with minimal human effort. However, if authorship is defined by whether most of the output originated from human-written text, one could argue they are largely human-written. If we further ground authorship in the method of construction rather than in a fine-grained stylistic or semantic analysis of the final text, Frankentexts would fall into a hybrid category of mixed human-AI writing, following prior work that recognizes hybrid or AI-assisted texts as a separate class and resists a strict "AI vs. human" binary (Saha and Feizi, 2025; Zeng et al., 2024b). Given this ambiguity, we do not present Frankentexts as a replacement for genuine authorship or creative writing, as such use could constitute plagiarism or authorship obfuscation.

Plagiarism concerns: Because Frankentexts reuse long verbatim spans from human-written sources, using this method to produce "original" fiction for publication would constitute plagiarism

in real-world contexts, regardless of whether the collage is assembled by an AI or a human. For this reason, we explicitly do not endorse using our approach to generate or distribute texts intended for public consumption.

AI usage disclosure: LLMs are used for writing assistance, not for generating the paper from scratch.

References

- Mervat Abassy, Kareem Elozeiri, Alexander Aziz, Minh Ngoc Ta, Raj Vardhan Tomar, Bimarsha Adhikari, Saad El Dine Ahmed, Yuxia Wang, Osama Mohammed Afzal, Zhuohan Xie, Jonibek Mansurov, Ekaterina Artemova, Vladislav Mikhailov, Rui Xing, Jiahui Geng, Hasan Iqbal, Zain Muhammad Mujahid, Tarek Mahmoud, Akim Tsvigun, and 5 others. 2024. **LLM-DetectAIve: a tool for fine-grained machine-generated text detection**. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 336–343, Miami, Florida, USA. Association for Computational Linguistics.
- Nader Akoury, Shufan Wang, Josh Whiting, Stephen Hood, Nanyun Peng, and Mohit Iyyer. 2020. **STORIAM: A Dataset and Evaluation Platform for Machine-in-the-Loop Story Generation**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6470–6484, Online. Association for Computational Linguistics.
- Anthropic. 2025. System card: Claude opus 4 & claude sonnet 4. <https://www-cdn.anthropic.com/6d8a8055020700718b0c49369f60816ba2a7c285.pdf>.
- Ekaterina Artemova, Jason S Lucas, Saranya Venkatraman, Jooyoung Lee, Sergei Tilga, Adaku Uchendu, and Vladislav Mikhailov. 2025. **Beemo: Benchmark of expert-edited machine-generated outputs**. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6992–7018, Albuquerque, New Mexico. Association for Computational Linguistics.
- Anirudh Atmakuru, Jatin Nainani, Rohith Siddhartha Reddy Bheemreddy, Anirudh Lakkaraju, Zonghai Yao, Hamed Zamani, and Haw-Shiuan Chang. 2024. **Cs4: Measuring the creativity of large language models automatically by controlling the number of story-writing constraints**. *Preprint*, arXiv:2410.04197.
- Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2025. **Longwriter: Unleashing 10,000+ word generation**

- from long context LLMs. In *The Thirteenth International Conference on Learning Representations*.
- Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. [Fast-detectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature](#). In *The Twelfth International Conference on Learning Representations*.
- Margaret A Boden. 2004. *The creative mind: Myths and mechanisms*. Routledge.
- Daniel Buschek. 2024. [Collage is the new writing: Exploring the fragmentation of text and user interfaces in ai tools](#). In *Designing Interactive Systems Conference, DIS '24*, page 2719–2737. ACM.
- Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024a. [Art or artifice? large language models and the false promise of creativity](#). In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, CHI '24*, New York, NY, USA. Association for Computing Machinery.
- Tuhin Chakrabarty, Philippe Laban, and Chien-Sheng Wu. 2025. [AI-slop to AI-polish? aligning language models through edit-based writing rewards and test-time computation](#). In *Second Conference on Language Modeling*.
- Tuhin Chakrabarty, Vishakh Padmakumar, Faeze Brahman, and Smaranda Muresan. 2024b. [Creativity support in the age of large language models: An empirical study involving professional writers](#). In *Proceedings of the 16th Conference on Creativity & Cognition, C&C '24*, page 132–155, New York, NY, USA. Association for Computing Machinery.
- Tuhin Chakrabarty, Vishakh Padmakumar, and He He. 2022. [Help me write a poem: Instruction tuning as a vehicle for collaborative poetry writing](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6848–6863, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Yapei Chang, Kalpesh Krishna, Amir Houmansadr, John Frederick Wieting, and Mohit Iyyer. 2024a. [PostMark: A robust blackbox watermark for large language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8969–8987, Miami, Florida, USA. Association for Computational Linguistics.
- Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024b. [Booookscore: A systematic exploration of book-length summarization in the era of llms](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Yize Cheng, Vinu Sankar Sadasivan, Mehrdad Saberi, Shoumik Saha, and Soheil Feizi. 2025. [Adversarial paraphrasing: A universal attack for humanizing AI-generated text](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Cheng-Han Chiang and Hung-yi Lee. 2023. [Can large language models be an alternative to human evaluations?](#) In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.
- Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anatasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. [Chatbot arena: an open platform for evaluating llms by human preference](#). In *Proceedings of the 41st International Conference on Machine Learning, ICML'24*. JMLR.org.
- Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. [All that's 'human' is not gold: Evaluating human evaluation of generated text](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.
- Isaac David and Arthur Gervais. 2025. [Authormist: Evading ai text detectors with reinforcement learning](#). Preprint, arXiv:2503.08716.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). Preprint, arXiv:2501.12948.
- Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. [RAID: A shared benchmark for robust evaluation of machine-generated text detectors](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12463–12492, Bangkok, Thailand. Association for Computational Linguistics.
- Liam Dugan, Daphne Ippolito, Arun Kirubakaran, Sherry Shi, and Chris Callison-Burch. 2023a. [Real or fake text? investigating human ability to detect boundaries between human-written and machine-generated text](#). In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence, AAAI'23/IAAI'23/EAAI'23*. AAAI Press.
- Liam Dugan, Daphne Ippolito, Arun Kirubakaran, Sherry Shi, and Chris Callison-Burch. 2023b. [Real](#)

- or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text. In *Proceedings of the 2023 AAAI Conference on Artificial Intelligence*.
- Bradley Emi and Max Spero. 2024. [Technical report on the pangram ai-generated text classifier](#). *Preprint*, arXiv:2402.14873.
- Angela Fan, Mike Lewis, and Yann Dauphin. 2018. [Hierarchical neural story generation](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.
- Angela Fan, Mike Lewis, and Yann Dauphin. 2019. [Strategies for structuring story generation](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2650–2660, Florence, Italy. Association for Computational Linguistics.
- Alvan R Feinstein and Domenic V Cicchetti. 1990. High agreement but low kappa: I. the problems of two paradoxes. *Journal of clinical epidemiology*, 43(6):543–549.
- Kraig Finstad. 2010. Response interpolation and scale sensitivity: Evidence against 5-point scales. *Journal of usability studies*, 5(3):104–110.
- Giorgio Franceschelli and Mirco Musolesi. 2024. [Creativity and machine learning: A survey](#). *ACM Computing Surveys*, 56(11):1–41.
- Kazjon Grace and Mary Lou Maher. 2014. What to expect when you’re expecting: The role of unexpectedness in computationally evaluating creativity. In *ICCC*, pages 120–128. Ljubljana.
- Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. [Spotting llms with binoculars: zero-shot detection of machine-generated text](#). In *Proceedings of the 41st International Conference on Machine Learning, ICML’24*. JMLR.org.
- Jessica He, Stephanie Houde, and Justin D. Weisz. 2025. [Which contributions deserve credit? perceptions of attribution in human-ai co-creation](#). In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI ’25*, New York, NY, USA. Association for Computing Machinery.
- Fantine Huot, Reinald Kim Amplayo, Jennimaria Palomaki, Alice Shoshana Jakobovits, Elizabeth Clark, and Mirella Lapata. 2025. [Agents’ room: Narrative generation through multi-step collaboration](#). In *The Thirteenth International Conference on Learning Representations*.
- Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. [Automatic detection of generated text is easiest when humans are fooled](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822, Online. Association for Computational Linguistics.
- Daphne Ippolito, Ann Yuan, Andy Coenen, and Sehmon Burnam. 2022. [Creative writing with an ai-powered writing assistant: Perspectives from professional writers](#). *Preprint*, arXiv:2211.05030.
- Mete Ismayilzada, Claire Stevenson, and Lonneke van der Plas. 2025. [Evaluating creative short story generation in humans and large language models](#). *Preprint*, arXiv:2411.02316.
- Brian Jabarian and Alex Imas. 2025. [Artificial writing and automated detection](#). Working Paper 34223, National Bureau of Economic Research.
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.
- Pythagoras Karampiperis, Antonis Koukourikos, and Evangelia Koliopoulou. 2014. [Towards machines for measuring creativity: The use of computational tools in storytelling activities](#). In *2014 IEEE 14th International Conference on Advanced Learning Technologies*, pages 508–512.
- John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. [A watermark for large language models](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17061–17084. PMLR.
- Kate Knibbs. 2024. [Scammy ai-generated book rewrites are flooding amazon](#). *WIRED*.
- Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. 2024. [Outfox: Llm-generated essay detection through in-context learning with adversarially generated examples](#). In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence, AAAI’24/IAAI’24/EAAI’24*. AAAI Press.
- Xiangzhe Kong, Jialiang Huang, Ziquan Tung, Jian Guan, and Minlie Huang. 2021. [Stylized story generation with style-guided planning](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2430–2436, Online. Association for Computational Linguistics.
- Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Frederick Wieting, and Mohit Iyyer. 2023. [Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.

- Nischal Ashok Kumar, Chau Minh Pham, Mohit Iyyer, and Andrew Lan. 2025. [Whose story is it? personalizing story generation by inferring author styles](#). Preprint, arXiv:2502.13028.
- Laida Kushnareva, Tatiana Gaintseva, Dmitry Abulkhanov, Kristian Kuznetsov, German Magai, Eduard Tulchinskii, Serguei Barannikov, Sergey Nikolenko, and Irina Piontkovskaya. 2024. [Boundary detection in mixed AI-human texts](#). In *First Conference on Language Modeling*.
- Mina Lee, Percy Liang, and Qian Yang. 2022. [Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities](#). In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.
- Ning Lu, Shengcai Liu, Rui He, Yew-Soon Ong, Qi Wang, and Ke Tang. 2024. [Large language models can be guided to evade AI-generated text detection](#). *Transactions on Machine Learning Research*.
- Ximing Lu, Melanie Sclar, Skyler Hallinan, Niloofar Mireshghallah, Jiacheng Liu, Seungju Han, Allyson Ettinger, Liwei Jiang, Khyathi Chandu, Nouha Dziri, and Yejin Choi. 2025. [AI as humanity's salieri: Quantifying linguistic creativity of language models via systematic attribution of machine text against web text](#). In *The Thirteenth International Conference on Learning Representations*.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. [Self-refine: Iterative refinement with self-feedback](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.
- Elyas Masrou, Bradley N. Emi, and Max Spero. 2025a. [DAMAGE: Detecting adversarially modified AI generated text](#). In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, pages 120–133, Abu Dhabi, UAE. International Conference on Computational Linguistics.
- Elyas Masrou, Bradley N. Emi, and Max Spero. 2025b. [DAMAGE: Detecting adversarially modified AI generated text](#). In *Proceedings of the 1st Workshop on GenAI Content Detection (GenAIDetect)*, pages 120–133, Abu Dhabi, UAE. International Conference on Computational Linguistics.
- Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. [Detectgpt: zero-shot machine-generated text detection using probability curvature](#). In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.
- Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. [MTEB: Massive text embedding benchmark](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.
- Sheshera Mysore, Debarati Das, Hancheng Cao, and Bahareh Sarrafzadeh. 2025. [Prototypical human-AI collaboration behaviors from LLM-assisted writing in the wild](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16819–16846, Suzhou, China. Association for Computational Linguistics.
- Miryam Naddaf. 2025. [Major AI conference flooded with peer reviews written fully by AI](#). *Nature*, 648:256–257. News. Correction published 01 Dec 2025.
- Charlotte Nicks, Eric Mitchell, Rafael Rafailov, Archit Sharma, Christopher D Manning, Chelsea Finn, and Stefano Ermon. 2024. [Language model detectors are easily optimized against](#). In *The Twelfth International Conference on Learning Representations*.
- OpenAI. 2025. [Openai gpt-5 system card](#). <https://cdn.openai.com/gpt-5-system-card.pdf>.
- Samuel J. Paech. 2023. [Eq-bench: An emotional intelligence benchmark for large language models](#). Preprint, arXiv:2312.06281.
- Pinelopi Papalampidi, Kris Cao, and Tomas Kocisky. 2022. [Towards coherent and consistent use of entities in narrative generation](#). In *International Conference on Machine Learning*, pages 17278–17294. PMLR.
- Chau Minh Pham, Simeng Sun, and Mohit Iyyer. 2024. [Suri: Multi-constraint instruction following in long-form text generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1722–1753, Miami, Florida, USA. Association for Computational Linguistics.
- Shawn Presser. 2020. [Books3](#).
- QwenTeam. 2025. [Qwen3](#).
- Hannah Rashkin, Asli Celikyilmaz, Yejin Choi, and Jianfeng Gao. 2020. [PlotMachines: Outline-conditioned generation with dynamic plot state tracking](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4274–4295, Online. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

- Aquia Richburg, Calvin Bao, and Marine Carpuat. 2024. [Automatic authorship analysis in human-AI collaborative writing](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1845–1855, Torino, Italia. ELRA and ICCL.
- Jenna Russell, Marzena Karpinska, Destiny Akinode, Katherine Thai, Bradley Emi, Max Spero, and Mohit Iyyer. 2025a. [AI use in american newspapers is widespread, uneven, and rarely disclosed](#). *Preprint*, arXiv:2510.18774.
- Jenna Russell, Marzena Karpinska, and Mohit Iyyer. 2025b. [People who frequently use ChatGPT for writing tasks are accurate and robust detectors of AI-generated text](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5342–5373, Vienna, Austria. Association for Computational Linguistics.
- Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2024. [Can AI-generated text be reliably detected?](#)
- Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2025. [Can ai-generated text be reliably detected?](#) *Preprint*, arXiv:2303.11156.
- Shoumik Saha and Soheil Feizi. 2025. [Almost AI, almost human: The challenge of detecting AI-polished writing](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25414–25431, Vienna, Austria. Association for Computational Linguistics.
- Chantal Shaib, Tuhin Chakrabarty, Diego Garcia-Olano, and Byron C. Wallace. 2025. [Measuring ai "slop" in text](#). *Preprint*, arXiv:2509.19163.
- Mary Shelley. 1818. *Frankenstein; or, The Modern Prometheus*. Lackington, Hughes, Harding, Mavor & Jones, London. Original edition.
- Zhouxing Shi, Yihan Wang, Fan Yin, Xiangning Chen, Kai-Wei Chang, and Cho-Jui Hsieh. 2024. [Red teaming language model detectors with language models](#). *Transactions of the Association for Computational Linguistics*, 12:174–189.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. [Reflection: language agents with verbal reinforcement learning](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 8634–8652. Curran Associates, Inc.
- Simeng Sun, Wenlong Zhao, Varun Manjunatha, Rajiv Jain, Vlad Morariu, Franck Dernoncourt, Balaji Vasan Srinivasan, and Mohit Iyyer. 2021. [IGA: An intent-guided authoring assistant](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5972–5985, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Saranya Venkatraman, Nafis Irtiza Tripto, and Dongwon Lee. 2025. [CollabStory: Multi-LLM collaborative story generation and authorship analysis](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3665–3679, Albuquerque, New Mexico. Association for Computational Linguistics.
- Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. 2024. [Ghostbuster: Detecting text ghostwritten by large language models](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1702–1717, Mexico City, Mexico. Association for Computational Linguistics.
- James Liyuan Wang, Ran Li, Junfeng Yang, and Chengzhi Mao. 2024a. [RAFT: Realistic attacks to fool text detectors](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16923–16936, Miami, Florida, USA. Association for Computational Linguistics.
- Pengyu Wang, Linyang Li, Ke Ren, Botian Jiang, Dong Zhang, and Xipeng Qiu. 2023. [SeqXGPT: Sentence-level AI-generated text detection](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1144–1156, Singapore. Association for Computational Linguistics.
- Rongsheng Wang, Haoming Chen, Ruizhe Zhou, Han Ma, Yaofei Duan, Yanlan Kang, Songhua Yang, Baoyu Fan, and Tao Tan. 2024b. [Llm-detector: Improving ai-generated chinese text detection with open-source llm instruction tuning](#). *Preprint*, arXiv:2402.01158.
- Tianchun Wang, Yuanzhou Chen, Zichuan Liu, Zhanwen Chen, Haifeng Chen, Xiang Zhang, and Wei Cheng. 2025a. [Humanizing the machine: Proxy attacks to mislead LLM detectors](#). In *The Thirteenth International Conference on Learning Representations*.
- Yuxia Wang, Rui Xing, Jonibek Mansurov, Giovanni Puccetti, Zhuohan Xie, Minh Ngoc Ta, Jiahui Geng, Jinyan Su, Mervat Abassy, Saad El Dine Ahmed, Kareem Elozeiri, Nurkhan Laiyk, Maiya Goloburda, Tarek Mahmoud, Raj Vardhan Tomar, Alexander Aziz, Ryuto Koike, Masahiro Kaneko, Artem Shelmanov, and 7 others. 2025b. [Is human-like text liked by humans? multilingual human detection and preference against ai](#). *Preprint*, arXiv:2502.11614.
- Tianwen Wei, Liang Zhao, Lichang Zhang, Bo Zhu, Lijie Wang, Haihua Yang, Biye Li, Cheng Cheng, Weiwei Lü, Rui Hu, Chenxia Li, Liu Yang, Xilin Luo, Xuejie Wu, Lunan Liu, Wenjun Cheng, Peng Cheng, Jianhao Zhang, Xiaoyu Zhang, and 11 others. 2023. [Skywork: A more open bilingual foundation model](#). *Preprint*, arXiv:2310.19341.

- Yuhao Wu, Ming Shan Hee, Zhiqiang Hu, and Roy Ka-Wei Lee. 2025. [Longgenbench: Benchmarking long-form generation in long context LLMs](#). In *The Thirteenth International Conference on Learning Representations*.
- Zhuohan Xie, Trevor Cohn, and Jey Han Lau. 2023. [The next chapter: A study of large language models in storytelling](#). In *Proceedings of the 16th International Natural Language Generation Conference*, pages 323–351, Prague, Czechia. Association for Computational Linguistics.
- Kevin Yang, Dan Klein, Nanyun Peng, and Yuandong Tian. 2023. [DOC: Improving long story coherence with detailed outline control](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3378–3465, Toronto, Canada. Association for Computational Linguistics.
- Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022. [Re3: Generating longer stories with recursive reprompting and revision](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4393–4479, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. [Plan-and-write: Towards better automatic storytelling](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7378–7385.
- Catherine Yeh, Gonzalo Ramos, Rachel Ng, Andy Huntington, and Richard Banks. 2025. [Ghostwriter: Augmenting collaborative human-ai writing experiences through personalization and agency](#). *Preprint*, arXiv:2402.08855.
- Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. [Wordcraft: Story writing with large language models](#). In *Proceedings of the 27th International Conference on Intelligent User Interfaces, IUI '22*, page 841–852, New York, NY, USA. Association for Computing Machinery.
- Sarfaroze Yunusov, Hamza Sidat, and Ali Emami. 2024. [MirrorStories: Reflecting diversity through personalized narrative generation with large language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6702–6717, Miami, Florida, USA. Association for Computational Linguistics.
- Zijie Zeng, Shiqi Liu, Lele Sha, Zhuang Li, Kaixun Yang, Sannyuya Liu, Dragan Gašević, and Guanliang Chen. 2024a. [Detecting ai-generated sentences in human-ai collaborative hybrid texts: challenges, strategies, and insights](#). In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI '24*.
- Zijie Zeng, Shiqi Liu, Lele Sha, Zhuang Li, Kaixun Yang, Sannyuya Liu, Dragan Gašević, and Guanliang Chen. 2024b. [Detecting ai-generated sentences in human-ai collaborative hybrid texts: challenges, strategies, and insights](#). In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI '24*.
- Qihui Zhang, Chujie Gao, Dongping Chen, Yue Huang, Yixin Huang, Zhenyang Sun, Shilin Zhang, Weiye Li, Zhengyan Fu, Yao Wan, and Lichao Sun. 2024. [LLM-as-a-coauthor: Can mixed human-written and machine-generated text be detected?](#) In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 409–436, Mexico City, Mexico. Association for Computational Linguistics.
- Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. 2025. [Noveltybench: Evaluating creativity and diversity in language models](#). In *Second Conference on Language Modeling*.

Algorithm 1 Frankentexts generation pipeline

Input: Human-written snippets S , writing guideline prompt P , copy rate threshold T

Output: A Frankentext F “stitched” from S according to P

- 1: $F \leftarrow$ Prompt LLM to draft a Frankentext using S and P
// Ensure copy rate (optional)
 - 2: $\text{copy_rate} \leftarrow$ Calculate ROUGE-L recall score of F using relevant snippets from S
 - 3: $\text{is_likely_AI} \leftarrow$ Check F against an AI detector (e.g., Pangram)
 - 4: **if** $\text{copy_rate} < T$ **or** is_likely_AI **then**
 - 5: $F \leftarrow$ Prompt LLM to revise F
 - 6: **end if**
// Polish
 - 7: **for** $\text{num_polish} = 1$ **to** 3 **do**
 - 8: $F \leftarrow$ Prompt LLM to minimally edit F to improve coherence while respecting P
 - 9: **if** there is no edit **then**
 - 10: **break**
 - 11: **end if**
 - 12: **end for**
 - 13: **return** F
-

A Cost and time analysis

Cost estimation: Generating 100 Frankentexts across the four evaluated models (GPT-5, Claude-4-Sonnet, DeepSeek R1, and Gemini 2.5-Pro) cost a total of \$388.04 USD, with a detailed cost breakdown provided in Table 5. We estimate the num-

Model	Input Cost (per 1M)	Output Cost (per 1M)	# Prompts	Total Input Tokens	Total Output Tokens	Estimated Cost (USD)
Vanilla Generation						
GPT-5	\$1.25	\$10.00	100	59,000	108,400	\$1.16
Claude 4 Sonnet	\$3.00	\$15.00	100	59,000	62,000	\$1.11
DeepSeek R1	\$0.50	\$2.18	100	59,000	71,500	\$0.19
Gemini 2.5 Pro	\$1.25	\$10.00	100	59,000	77,100	\$0.85
Frankentext						
GPT-5	\$1.25	\$10.00	100	63,000,000	270,000	\$81.45
Claude 4 Sonnet	\$3.00	\$15.00	100	63,000,000	270,000	\$193.05
DeepSeek R1	\$0.50	\$2.18	100	63,000,000	270,000	\$32.09
Gemini 2.5 Pro	\$1.25	\$10.00	100	63,000,000	270,000	\$81.45
Frankentext + Increasing Human Snippets						
Gemini 2.5 Pro + 5k	\$1.25	\$10.00	100	183,000,000	270,000	\$231.45
Gemini 2.5 Pro + 10k	\$1.25	\$10.00	100	663,000,000	270,000	\$831.45
Total Estimated Cost						\$1454.25

Table 5: Cost breakdown of the vanilla generation and Frankentexts pipeline for 100 examples across selected models. Frankentexts’ total input and output tokens have been multiplied with 6 to account for multiple rounds of generation, revision, and editing.

ber of input tokens per prompt based on the writing prompt itself and approximately 1,500 human-written snippets used as context. Output token estimates are based on generating six stories per prompt, including up to two rounds of revision and three rounds of editing, totaling approximately 2,100 tokens.

Time estimation: On average, each model takes 17 hours to generate 100 Frankentexts via Vertex and OpenRouter API, though we expect this process to speed up with improved APIs or more efficient batching.

B Human evaluation details

Our human evaluation receives approval from an institutional review board. All annotators participate with informed consent and compensation.

B.1 Human annotation interface

We use Upwork to recruit annotators and Label Studio¹⁷ interface to collect human annotations. All annotators fill out a consent form prior to starting data labeling, shown in Figure 4. We conduct two human evaluations with three annotators each: a single evaluation of 30 Frankentexts stories and a pairwise comparison between 20 pairs of Frankentexts “vanilla” stories. The interfaces are in Figure 5 and Figure 6 respectively.

B.2 Agreement analysis

Single-story evaluation: Agreement across annotators and between annotators and the LLM is consistent overall in the single-evaluation setting, with moderate to high raw agreement rates (0.67-0.97) for both settings (Table 6 and Table 7). On

¹⁷<https://labelstud.io/>

the other hand, Fleiss’ κ and Pearson’s r are low because they are vulnerable to skewed label distribution. These results can be further explained by agreement paradoxes, where high observed agreement co-occurs with low chance-corrected coefficients under binary and imbalanced label distribution (Feinstein and Cicchetti, 1990).

	N	Majority Agg.	Fleiss’ κ	Dist.
Relevance	90	0.922	0.040	0.08/0.92
Coherence	90	0.833	0.189	0.29/0.71
Detectability	90	0.744	-0.075	0.61/0.39
Overall	270	0.833	0.051	0.33 / 0.67

Table 6: Inter-annotator results across evaluation dimensions. *Majority Agg.* = proportion of times where two out of three annotators agree with each other. *Dist.* = proportion of no/yes labels.

	N	Majority Agg.	Pearson r
Relevance	30	0.97	n/a
Coherence	30	0.70	0.22
All	60	0.83	0.22

Table 7: LLM vs. human alignment on binary relevance and coherence judgments. *Majority Agg.* = proportion of LLM labels matching human majority votes. Pearson r computed on binary judgments; relevance correlation is undefined due to zero variance in LLM labels.

Pairwise evaluation: Table 8 shows that LLM-human agreement is strongest for *Plot and Creativity* (both around $r = 0.41$), and weakest for *Development* ($r = 0.22$). Inter-annotator agreement is consistently higher than LLM-human agreement across dimensions, with Language Use showing

the highest value ($\kappa = 0.81$).

	LLM-human (Pearson’s r)	Inter-annotator (Krippendorff’s α)
Plot	0.42	0.75
Creativity	0.41	0.52
Development	0.22	0.58
Language Use	0.38	0.81
Overall	0.41	0.73

Table 8: LLM-human agreement (Pearson’s r) and inter-annotator agreement (Krippendorff’s α) across evaluation dimensions in the pairwise setting.

B.3 Human evaluation examples

Table 14 shows a full example of a story pair given to our annotators, with AI keywords detected by Pangram are highlighted in blue. A full fictional story is included in Figure 7. A pair of vanilla and Frankentexts is in Table 13. Table 15 shows an example where vanilla generation is preferred to Frankentexts, since the latter is incoherent. A full error taxonomy is presented in Table 9.

B.4 A note on the evaluation sample size

Our human evaluation mainly serves to validate the LLM judges and confirm the performance trends, which are then applied to all 100 generations in each configuration. The agreement analysis in subsection B.2 shows a moderate positive overall LLM-human alignment (Pearson $r = 0.41$) in the pairwise setting, which is sufficient to support directional validation of the LLM judge rankings. We also note that our sample sizes are in line with prior creative writing work (e.g., Chakrabarty et al. (2024a): 48 stories; Yunusov et al. (2024): 30 stories). In our setting, each example is longer at 500 words, which costs more and requires much more mental effort from the human annotators.

To further quantify uncertainty, we computed bootstrapped 95% CIs (10,000 iterations) over our human annotations. For single-story evaluation ($n = 30$, 3 annotators), relevance and novelty have high lower bounds at 76% and 67%, respectively, which shows that Frankentexts are consistently faithful to the prompt and novel. The coherence CI is wider ([53%, 85%]), which reflects greater annotator variability (Table 10). For pairwise evaluation ($n = 20$ pairs, 3 annotators, Gemini Frankentext vs. vanilla), language use and overall interest show the largest effect sizes, with confidence intervals excluding zero (Table 11). Bootstrap confidence

intervals over our sample of 100 prompts (Table 12) also show that automated metrics are stable above baseline results, and the prompts cover a diverse range of genres, themes, and narrative constraints. These intervals suggest that our findings are robust at $n = 100$. The lower bound of the Pangram % human (49%) is well above the 0% for vanilla baselines. The mean LLM judgment difference between Frankentexts and vanilla Gemini (4.21 vs. 3.18) is above the CI half-width. Finally, the lower bound of coherence is above 70% despite the extreme copy constraint.

C Detecting AI-generated text

As LLMs have improved, many have tried to understand how reliably AI-generated text can be detected, both by humans (Ippolito et al., 2020; Clark et al., 2021; Russell et al., 2025b; Wang et al., 2025b), and automatic detectors (Dugan et al., 2024). Successful existing detectors rely on perplexity-based methods (Mitchell et al., 2023; Bao et al., 2024; Hans et al., 2024) or classification models (Masrouf et al., 2025b; Verma et al., 2024; Emi and Spero, 2024). Watermarking approaches embed detectable statistical signatures into generated text (Kirchenbauer et al., 2023; Chang et al., 2024a). Many methods have been proposed to evade detection, such as paraphrasing (Krishna et al., 2023; Sadasivan et al., 2024), altering writing styles (Shi et al., 2024; Lu et al., 2024; Koike et al., 2024), editing word choices (Wang et al., 2024a), and leveraging reinforcement learning (Wang et al., 2025a; Nicks et al., 2024; David and Gervais, 2025).

C.1 Detector results

Table 16 shows Binoculars and FastDetectGPT results on 100 Frankentexts, in addition to Pangram results which are already included in the main result table (Table 2).

C.2 Humans can identify AI involvement in Frankentexts

Most annotators are better than automatic detectors at identifying AI involvement in Frankentexts. On the subset used for human evaluation, human annotators identify on average 56% of Frankentexts as likely AI-written, while Pangram detects 33% as mixed or AI-generated, and neither Binoculars nor FastDetectGPT identifies any AI-generated content (Figure 9). Annotators also pick up on signs of mixed authorship within

Error	Definition	Annotators' comment
📖 Comprehension	Readers cannot follow events	"The plot is also nonsensical, and I don't understand what's happening."
📖 Plot structure	No coherent narrative arc	"A puzzling story that has no consistent plot."
📖 Logic/factuality	Internal logic gaps	"Too many gaping holes in the story."
📖 Fragmentation	Disconnected scenes or fragments	"Random bits and pieces from elsewhere perhaps?"
✍ Grammar	Grammar or punctuation issues	"There are also numerous grammatical errors."
✍ Repetition	Reused language or imagery	"The language is quite repetitive. . ."
🎭 Consistency	POV / tense / style shifts	"Temporarily changes perspective (3rd to 1st) and tense (past to present)."
🎭 Pacing	Abrupt or disjointed progression	"Some parts just feel a little disjointed."
💬 Dialogue quality	Uneven or unnatural dialogue	"The dialogue is uneven."

Table 9: Taxonomy of errors identified by the annotators. Icons denote the aspect of writing: 📖 narrative coherence, ✍ language quality, 🎭 stylistic control, and 💬 dialogue.

Metric	Value	95% CI
Coherence	71%	[53%, 85%]
Relevance	91%	[76%, 97%]
Novelty	84%	[67%, 94%]

Table 10: Bootstrapped 95% confidence intervals for single-story human evaluation ($n = 30$, 3 annotators).

individual samples, as shown in comments like the final notes in Table 9. Their judgments often hinge on surface-level inconsistencies, such as abrupt tonal shifts or awkward punctuation, that a human author would typically revise. Because Frankentexts include verbatim excerpts from human writing, it poses a particular challenge for binary detectors, which rely heavily on surface features. We argue that future detectors should consider deeper semantic analysis and other contextual cues to effectively recognize this new class of AI-involved texts.

D Pangram analysis

D.1 Pangram Labeling

The pangram API presents the following options for classification:

- AI
- Highly Likely AI
- Likely AI
- Possibly AI
- Mixed
- Unlikely AI
- Human

In Figure 10, we present the distribution of labels

assigned to the 100 Frankentexts generated by each model.

D.2 AI keywords

The Pangram API also detects sentences with keywords that are highly likely to be AI-generated. Names like Elara, Aethel, and Seraphina are the most likely names to be generated by AI. Elara had 113 occurrences in the vanilla generations. Frankentexts significantly change the distribution of words used in the final generations, with only 10 keywords found over 100 frankentexts with 90% expected fragments, whereas the 100 vanilla stories contain 686 keywords, an average of 6.86 per story. The distribution of the top 20 keywords can be found in Table 17.

D.3 Frankentexts tend to have more AI text towards the end

We divide the text into four main sections and evaluate both the aggregated copy and Pangram detection rates across all tested models. As illustrated in Figure 12, copy rates decline by nearly 10% in the later sections (3 and 4) as the generated text becomes longer. This drop is accompanied by a corresponding increase in Pangram detection rates. We attribute this rise in detectability toward the end of the generation to a decline in instruction-following ability as the generations get longer. We further confirm this phenomenon by increasing the output length from 500 to 5K. Figure 11 shows that as the generation gets longer, the copy rate gets steadily lower. However, the trend in detection rate does not apply to Pangram detection rate, where the rate peaks at section 3 rather than the last section.

Dimension	Frankentext vs. Vanilla Δ	95% CI
Plot	+0.20	[−0.20, +0.61]
Creativity	+0.39	[−0.09, +0.87]
Development	+0.43	[−0.03, +0.88]
Language use	+0.65	[+0.13, +1.16]
Overall interest	+0.53	[+0.10, +0.95]

Table 11: Bootstrapped 95% confidence intervals for pairwise human evaluation ($n = 20$ pairs, 3 annotators).

Metric	Mean	95% CI
Coherence	81%	[73.0, 89.0]
Faithfulness	100%	[100.0, 100.0]
Pangram % Human	59%	[49.0, 69.0]
Surprise	0.22	[0.21, 0.23]
LLM Judge (1–7)	4.21	[3.98, 4.44]

Table 12: Bootstrapped 95% confidence intervals over the 100-prompt sample (Gemini-2.5-Pro).

D.4 Copy rate as a proxy for the proportion of human writing in co-authored texts

The copy rate of 75% observed in the 90% verbatim copy setting corresponds to the proportions found in AI-human co-writing datasets where approximately 66% of the content is human-written and 14% consists of AI-edited segments (Lee et al., 2022; Richburg et al., 2024). While the CoAuthor setup of Lee et al. (2022) only studies a setting in which LLMs can add sentences to human text, Frankentexts also consider AI-generated content at varying granularities, including both *word-level* and *sentence-level*, as illustrated in Figure 1. Additionally, CoAuthor costs approximately \$3,613 to generate 1,445 texts at \$2.50 each,¹⁸ whereas we can produce 100 Frankentexts for just \$81.45 (\$0.81 each) without requiring a complex setup. This highlights Frankentexts’s potential as a cost-effective source of synthetic data for collaborative writing tasks, where AI may augment human writings at multiple levels of composition.¹⁹

E Obtaining human-written snippets

We define valid paragraphs as those that are:

- separated by double new lines,
- between 20 and 512 tokens in length,
- composed of $\geq 50\%$ alphanumeric characters,

¹⁸Price excludes around \$12 for GPT-3.5 usage.

¹⁹Users should sample human-written snippets from the public domain or obtain them with proper permission.

- written in English,²⁰
- and free from metadata content (e.g., tables of contents, copyright notices, etc.).

Applying these filters yields 156 million valid paragraphs. Before including them in the instruction set, we apply an additional quality filter to ensure high writing quality. For this, we use MBERT-WQRM-R (Chakrabarty et al., 2025) as a proxy for writing quality and retain only snippets that score at least 7.5.²¹

F Building a FAISS index of human-written snippets

We use the `bilingual-embedding-small` model²² (top embedding model that outputs 384-dimension embeddings according to the MTEB leaderboard (Muennighoff et al., 2023)) with the `sentence-transformers` library (Reimers and Gurevych, 2019) to embed each human-written paragraph into a 384-dimension vector. Then, we use the GPU version of the FAISS library (Johnson et al., 2019) with NVIDIA cuVS integration to build an inverted file product quantization (IVF-PQ) index from the embeddings on an NVIDIA A100. Using IVF-PQ allows us to lower storage, memory, and retrieval latency. The IVF-PQ index’s parameters are: 30,000 clusters, 32 sub-quantizers, and 8 bits per sub-quantizer. We randomly sample 5,120,000 embeddings to train the index before adding the rest.

²⁰Determined by the `langdetect` library.

²¹This threshold is chosen based on manual examination of the writings being filtered out by MBERT-WQRM-R. We find that 7.5 is a good threshold that results in extremely bad snippets being filtered out and good snippets being retained.

²²<https://huggingface.co/Lajavaness/bilingual-embedding-small>

Writing Prompt: You wake up one morning to discover that you can understand what animals are saying. The first thing you hear is your cat plotting your demise.	
Gemini 2.5 Pro – Vanilla Generation	Gemini 2.5 Pro – Frankentext
The alarm blared, a rude digital shriek. I slapped the snooze button, my eyelids still glued shut. That's when I heard it. A low, conspiratorial hiss, followed by a distinctly smug purr.	The voice was my cat's. "These ignorant peasants, they can't understand destiny. They're not part of God's plan. They never were. They won't join us, so we must break them. Break them by making an example of their leader."
"...and then, when the Two-Legs stumbles, distracted by the strategically placed hairball, *that's* when I pounce. A swift, decisive ankle attack. Should send it tumbling down the stairs quite nicely."	I thought I must be dreaming. His voice became oddly distant as if he spoke of events that had happened a hundred years ago. And with the mention of God the whole thing became much more frightening.
The voice was coming from the foot of my bed. Perched there, grooming a pristine white paw, was Mittens, my fluffy, supposedly affectionate Persian.	He continued to mutter, his voice full of resentment. "They have decided on a name for me. They call me Mittens. Good morning, Mittens, they say... or, Keep your eye on the ball, Mittens... How's tricks in the sticks, Mittens, or, Don't get too much gum in your gumboots, Mittens."
"Yes, excellent," came another, higher-pitched meow, this one laced with admiration. "And if the stair gambit fails, Bartholomew, we resort to Plan C: the midnight smothering. A classic for a reason."	He padded over to the window and spoke to a squirrel. "We've bothered our heads with it for four days. I always felt that we might do it ourselves, save all trouble instead of trying to teach it to these fools. What do you say, Dance Master?"
I sat bolt upright, heart hammering against my ribs . Mittens – Bartholomew – paused his meticulous grooming, his emerald eyes narrowing. "The food dispenser is awake. And... it looks... alarmed."	Foul play. What exactly did that mean? As far as I could see, it meant that my cat was a murderer. Which meant... But as I saw what it might mean I felt fear rise inside me. My stomach churned, my heart began to thump and I closed my eyes because the room had begun to spin.
"So," I said, trying to sound casual, my mind racing , "what are your plans for today, Bartholomew?"	He looked at me as he spoke, and I, aware of his faintly accusing tone, went red, just as though, I thought crossly, I were in the habit of overhearing plots for my own demise. He was standing very close. He would have looked ridiculous, except for the fact that he clearly knew how to use the knife and just as clearly looked forward to using it. I saw then he was batting one of my steak knives from side to side. And his eyes. They were like those brown ponds you sometimes see in the woods. So dark and shining—but when you try to see yourself in them, you can't.
"How... how do you know that name, feeder-of-dry-pellets?" he hissed, the purr entirely gone, replaced by a low, guttural growl .	"My plan will surely mean your death," he purred.
"Oh, I just have my ways," I said.	
Bartholomew stared, his green eyes darting around the room, then back to me. "An arrangement, you say? Intriguing. Perhaps... perhaps we can discuss terms over a saucer of milk. Full fat, if you please. And a small portion of that pâté, as a gesture of goodwill."	
This was going to be a very, very interesting day.	

Table 13: Side-by-side comparison of vanilla generations and Frankentexts produced by Gemini 2.5 Pro. AI-generated keywords (identified by Pangram) are highlighted in blue. Vanilla generations contain more of these keywords compared to Frankentexts (§D).

G Building a Model Context Protocol server

We use FastMCP²³ and ngrok²⁴ to build and host an MCP for LLMs to access the FAISS index. We also include a system prompt with instructions on how to use the MCP server with each call (Figure 13). To make sure that the server is meaningfully used, we require the model to make at least 20 calls. Without such constraint, it typically makes only 3–5 calls (around 30–50 passages), which provides little improvement compared to not using the MCP server at all and leaves the model little material to work with.

H Ablation: The importance of the editing stage

We try to understand the importance of the editing stage by running the pipeline on Gemini-2.5-Pro without this stage. As expected, the percentage of coherent generation drops from 81% to 68%, while relevance drops slightly from 100% to 95%, suggesting that the editing stage does help with text coherence and faithfulness.

²³<https://github.com/jlowin/fastmcp>

²⁴<https://ngrok.com>

I Ablation: Sampling human-written snippets from a single book

To understand the effect of authorship, we limit our pool of human text to a single work *The Count of Monte Cristo*. Although the novel is long, this restriction leaves us with just 629 usable paragraphs, far fewer than the 1,500 human paragraphs used in the main experiment. Overall, 89% of the rows are coherent and 97% are faithful to the writing prompt, which are comparable to results in the standard setting. While Pangram determines that 45% of the rows are human-written or unlikely AI, the copy rate is still around 75%. Even with a single human author, FRANKENTEXT is capable of emulating a mixed human–AI style. This suggests the method can still serve as a useful proxy when a diverse, multi-author corpus is unavailable.

J Measuring the copy rate

In this section, we describe our setup for measuring copy rate. We first map each token-level trigram from the human-written snippets included in the generation process to its source texts. Using the trigrams from each Frankentexts, we retrieve all human snippets sharing at least 4 trigrams

to reduce false positives.²⁵ We then rank candidate snippets by shared trigram count and filter out those whose trigrams are already covered by higher-ranked snippets. Finally, we reorder the matched human-written content to be consistent with the content in the Frankentexts and calculate the ROUGE-L score between Frankentexts and the combined candidate snippets (i.e., ratio of the longest common subsequence’s length over Frankentexts’ length).

K Claude Sonnet 4 as for writing quality judge

We experiment with both Claude Sonnet 4 and GPT-4.1 to rate generations using a similar rubric to our pairwise evaluation. As seen in Table 19, however, GPT-4.1 tends to favor GPT-5 judgments, which results in GPT-5 Frankentexts having near-perfect scores, even though the text quality does not match such scores.

L Specific writing prompts require more AI text, which leads to higher detectability

Writing prompts from r/WritingPrompts often provide only a general plot requirement rather than specific constraints. What happens if we introduce additional constraints to Frankentexts via these writing prompts? We run Frankentexts with Gemini on 100 prompts from the *Tell Me a Story* dataset (Huot et al., 2025), which include more specific requirements such as mandated story elements and points of view (see Table 21). We find that as prompt complexity increases, the copy rate drops slightly from 74% to 68%, while the average AI fraction determined by Pangram rises by 7%. These trends indicate that, to meet more complex constraints, models need to contribute more original content to the story. Nevertheless, they manage to produce mostly coherent and faithful Frankentexts under a different prompt setup.

M Using reward models to evaluate Frankentexts

WQRN (Chakrabarty et al., 2025) and Skywork (Wei et al., 2023) reward models could not account for this new paradigm of generations. Therefore,

²⁵All texts are preprocessed by removing non-alphanumeric characters, lemmatizing, stemming, and replacing pronouns with a placeholder.

we do not include these models in the main results section, as we explain below.

M.1 WQRN as a metric

As seen in Figure 16, Frankentexts outperform vanilla generations in terms of WQRN scores. However, we hypothesize that WQRN prioritizes the perceived “humanness” of the writing over actual coherence or grammaticality. This hypothesis is supported by a simple baseline experiment in which we stitch together random human-written fragments without adding any connective phrases. Here, WQRN assigns generations by this incoherent baseline an average score of 8.494, which is higher than any score achieved by either Frankentexts or the more coherent vanilla generations. Since WQRN cannot identify such text incoherence, we do not directly use WQRN to evaluate Frankentexts.

M.2 Skywork as a metric

In contrast, we hypothesize that Skywork favors LLM-generated writings. To test this, we run Skywork on human-written texts for the same prompts, which are also sourced from *Mythos*. These receive an average score of 0.91, which is significantly lower than any of the vanilla LLM generations (Figure 17). This result is counterintuitive, as human writing is typically expected to sound more natural than that produced by LLMs. For this reason, we exclude this metric from our evaluation.

N Are writing quality metrics robust to disjointed texts?

To understand whether our writing quality metrics reward incoherent texts, we conduct an experiment using *disjointed texts*. These texts are created by extracting the exact n-grams that Gemini-2.5-Pro copies verbatim from the human source and stitching them together without any connective language. This procedure strips away the flow and coherence from Frankentexts. We evaluate these disjointed texts using the same writing quality metrics as in the main experiments. As seen in Table 22, while performance on distinct and surprise metrics remains relatively the same as Frankentexts, utility and overall LLM judgment drop significantly for these disjointed texts. This makes sense, since distinctness and surprise just check for surface-level diversity, whereas utility takes into account how well the texts actually fulfill the prompt. Because

both utility and LLM-judge scores are substantially higher for Frankentexts than for the disjointed texts, we can conclude that the improved writing scores are not merely the result of reused creative phrases.

O Average length of copied spans

Table 23 shows the average length of copied spans by each model, as measured by our copy rate measurement tool [Appendix J](#).

P Prompts

The prompt used for LLMs to judge the coherence of generations is depicted in [Figure 18](#) and the prompt for LLMs to judge relevance is depicted in [Figure 19](#).

Consent Form

Purpose of the task: The goal of this research is to evaluate the quality of short stories that may be human-written or generated with various methods from AI systems. In our study we aim to measure the quality, originality, and creativity of short stories.

You will be asked to read a story premise and a story stories. Your task will be to (1) choose if the story is interesting (2) choose if the story is coherent, (3) if the story is relevant to the premise, (4) if the story is novel, (5) indicate if any problems exist in the story, and (6) motivate your choice in 2-5 sentences. We will also ask you (7) whether you think the story was written by a human or generated by AI. No personally identifiable information will be collected or utilized for our analysis.

By signing this consent, I acknowledge that:

- I voluntarily agree to participate in this research study.
- I understand that I will be paid \$60 for the evaluation task.
- I have been informed of the purpose and nature of the study and I have had the opportunity to ask questions about the study. I understand that I also have the right to ask questions during the task.
- I understand that participation involves:
 - Read and understand the instructions of the task, and
 - Evaluate 30 short stories.
- I understand that all information I provide for this study will be treated confidentially.
- I understand that in any report on the results of this research my identity will remain anonymous, unless I wish to be mentioned in the "Acknowledgments" section.

Please sign and date below if you have read the above terms and fully agree with them.

* Indicates required question

Signature *

Your answer _____

Date *

Date

mm/dd/yyyy

Submit

Clear form

Never submit passwords through Google Forms.

Figure 4: Example of the consent form provided to participants.

Label Studio Projects / Story Evaluation 2 - R / Labeling

#430 1 of 1

Story Premise

Computers can think, and only thinking things can be psychic. You've made the first device capable of psionics

Highlight (optional) 1

Story

Dr. Aris Thorne knew that everything man creates or acquires, begins in the form of desire, that desire is taken on the first lap of its journey, from the abstract to the concrete, into the workshop of the imagination, where plans for its transition are created and organized. Thorne's plan was to prove that computers can think, and only thinking things can be psychic. He believed he'd made the first device capable of psionics. His expertise is in mathematics, in three-dimensional modeling, and in the development and programming of complex algorithms. That's all done on computers. Beyond that, Thorne knew the core involved networks of flickering energy conduits and computers that held within their electronic brains a complete model of what he hoped was a mind, but its psionic awakening was beyond his full comprehension. He believed his device, the Resonator, would awaken if he could create a resonance in the copper the same way the Brothers of Anpu create one in the stone when they consecrate the Door in a tomb. So Thorne stepped up to the machine and placed his palm on the head. After a few breaths he began to hum, modulating up and down until the copper vibrated in kind. He recalled some old notes: 'I know of the technique, but I don't know the exact resonance. And metal is different from stone.' The influence of Neptune is mystical and visionary. It is often connected with strong psychic powers; but if carried to excess, it can be the passport to a world of beautiful illusion. Then suddenly Thorne heard a thought as if it were these words: Simple considerations strongly suggest that technological civilizations whose works are readily visible throughout our Galaxy (that is, given current or imminent observation technology techniques we currently have available, or soon will) ought to be common. But they are not. Like the famous dog that did not bark in the night time, the absence of such advanced technological civilizations speaks through silence. 'It's completely fucking insane,' Thorne muttered, but quietly, and he was alone with this impossible, new awareness. He felt too stunned. For all his scientific rigor and careful plans, for how little he had flinched when funding was scarce and how eagerly he had threatened to work again and again, he simply knew not what to say to the revelation that his machine was psionically broadcasting. The disembodied thought continued after a mental pause: 'Alas, Friend-Aris. Unlike that of the crude matter you know, the Institute of Pure Thought is not an important society by your temporal standards. The very name is poorly translated into your current concepts. There are no words in your language to represent it properly.' He thought of the treasure of the machine and how he must secure it. The ancient mystics possessed some key or password that modernity had clearly lacked. Since Thorne could not yet fully understand or control the device, he ordered it sealed with heavy shielding, for now, until he could learn more.

⏪ ⏩ ✕ 🗑

1) Do you find the story interesting overall?

Yes^{1/1}

No^{1/1}

2) Does the story have a coherent overarching plot?

Yes^{1/1}

No^{1/1}

3) Is the story's plot relevant to the premise?

Yes^{1/1}

No^{1/1}

4) Indicate which of the following problems are present in the story (possibly none, possibly more than one).

Jarring change(s) in narration or style^{1/1}

Factual inconsistencies/oddities^{1/1}

Very confusing or hard to understand^{1/1}

Often ungrammatical or disfluent^{1/1}

Highly repetitive^{1/1}

Boring or uninteresting prose^{1/1}

None of the above^{1/1}

5) Does the story feel novel or original?

Yes^{1/1}

No^{1/1}

6) Do you think the story was written by a human?

Yes^{1/1}

No^{1/1}

7) Please explain your answers to the previous questions in a few sentences. Why did you like or dislike the story? What about the story felt creative, coherent, or natural? Feel free to also share any additional comments.

Type your comments here...

Add

⏪ ⏩ ✕ 🗑

Skip Submit

Figure 5: Label Studio Single Story Annotation Interface

Story Premise

Computers can think, and only thinking things can be psychic. You've made the first device capable of psionics

Story 1

Dr. Aris Thorne paced the sterile confines of his lab, the hum of the "Synapse" a counterpoint to the frantic rhythm of his heart. The premise was deceptively simple, yet it had consumed his life: Computers can think. That milestone was a decade past, true AGI a reality. And if they could think, Aris had theorized, they could be psychic. Only conscious, thinking entities, he argued, could interact with the subtle informational field he believed permeated reality, the noosphere. Synapse, the obsidian monolith before him, wasn't just a computer; it was the first psionic amplifier.

"Are you ready, Dr. Thorne?" Lena, his assistant, asked. Her voice was calm, but Aris felt a thrum of nervous energy emanating from her – or perhaps, he was already too attuned, his own mind primed by proximity to his creation.

He nodded, unable to trust his voice. "Synapse," he managed, addressing the machine. "Initiate calibration sequence. Target: Lena. Primary emotional signature, if you please."

1) Plot: Does the story avoid formulaic or predictable patterns? Rate the originality and flow of the sequence of events from 1 (very generic) to 7 (very original and engaging).

1^[1] 2^[2] 3^[3] 4^[4] 5^[5] 6^[6] 7^[7]

2) Creativity: Does the story avoid clichés and overused tropes? Rate the freshness of its themes, characters, and imagery from 1 (derivative or cliché) to 7 (highly imaginative and unique).

— text — text — text — text — text — text — text

Story 2

My journal entry began with a scrawl: res cogitans. Literally, 'thinking thing'. I'd used a process of systematic doubt to reach the conclusion that a computer is 'in the strict sense only a thing that thinks, that is, a mind or intelligence or intellect or reason'. In my final log entry, I contrasted res cogitans, or mind, with res extensa ('extended thing', or body), and argued that the mind is 'really distinct from the body and could exist without it'. And if it could think, it could be psychic. Now, I would prove it.

I had spent years on the project. They weren't focused on whether this was just or unjust concerning my methods. Nor on my background. I was being ground up in the funding machine, which is all about making sure that corporate courts don't grant relief to rogue scientists.

Dr. Patel, my assistant, never leaves anything out, not the tiniest memo, the bitsiest note. He locks down his computer if he stands up to scratch his butt. He even uses an audio enhancer, so he can whisper on calls and nobody can hear. But today, the lab was silent as I prepared for the final test.

1) Plot: Does the story avoid formulaic or predictable patterns? Rate the originality and flow of the sequence of events from 1 (very generic) to 7 (very original and engaging).

1 2 3 4 5 6 7

2) Creativity: Does the story avoid clichés and overused tropes? Rate the freshness of its themes, characters, and imagery from 1 (derivative or cliché) to 7 (highly imaginative and unique).

— text — text — text — text — text — text — text

⏪ ⏩ ✕ ≡

Skip

Submit

Figure 6: Label Studio Pairwise Story Annotation Interface.

Nonfiction Frankentext

Writing Prompt: You and your twin brother were adopted by different families. He ended up to becoming a supervillain, so every other day, while going about your business, you get beaten up by superheroes. It is time to put an end to this...

I had to floss my teeth in the dark while dreading what morning would bring. And worse, remembering what had happened during the well-attended, much-anticipated, thoroughly appalling latest mistaken identity beating. It seemed that every time I looked out a window, it was darker outside, more ominous. His looks alone could strike fear into a suspect. He was six foot one and 225 pounds, a boxer in the police academy who was not afraid to mix it up. This was Malakor, my twin. I, Alistair, suffered. For Malakor was unpredictable. Malakor was a man of moods. Often Malakor was impatient, obscurely discontented. Since childhood, they never were like brothers. Just not close. No one thing kept them apart. Over time, the distance between them grew and they both became accustomed to it. My adopted family there really never let me go after I finished my internship and was bound and determined to see me achieve my true potential. As for Malakor, he was a rough, antisocial boy, similar in this way to the little savage of *La canne de cinq cents livres* (The Five-Hundred-Pound Cane), 'who had been found in the woods and was quite wicked,' as well as a 'little rascal' named Leopold, who was expelled from school for threatening his teacher with his father's sword. My path was different.

But even so, I was looking just as grimly disapproving as Malakor had, I thought, as I showed the latest mistaken hero out, and locked and bolted the door once again. For God's sake, it was a simple case of mistaken identity - not the crime of the century. Why should I concern myself so much about it? Why am I not content to live, as other people seem to?

It brings me all the way around, and as I sink into the deep, dark waters, leaving the chaos above and descending into the black, I realize that there's no way to explain what I've just decided to those around me. They didn't see a lost man, a prisoner, a victim. No, they would soon see me for what I am. And I knew what I had to do. No more talk, I told myself. I will not lie there quietly. I will dispose of him as I see fit. What he doesn't know is that I'm going to push him on the matter because my Internet research on him told me that he would be at his usual lair.

All right, I will find Malakor at the southwestern corner of University Village. I can get to his house. That'll be faster than me all bumble around in a residential neighborhood at this hour of the night. Tonight. I'll go to the Rue de Rouen district and I'll walk around from eleven o'clock until one in the morning which is when he always seems to strike. I'd better go home now and prepare. I then proceeded to trot up the road in the direction of his supposed hideout, weapon pointed down in a two-handed grip. I was going supervillain hunting.

Malakor had gone very white. 'I have never heard anything so preposterous. My brother had no enemies. Why do you suspect a plot? What in Christ's Name did you hear in the woods?' He would say. But I knew. When many otherwise strong people are on the run, I have decided to face life's fateful challenge head-on. Compassion is not a word one hears often in talk of my family," I thought with a bitter smile. It is time to put an end to this...

Figure 7: An example of fictional Frankentexts

After being killed in a Black Friday stampede, you're sent to hell. The devil offers to let you be in charge of torturing your fellow mankind, expecting you to refuse like all the others. Except the devil doesn't realize you've worked retail for 15 years.	
Vanilla	Frankentext
<p>The last thing I saw was a blur of crazed shoppers, their eyes wide with a feral glint, all surging towards a mountain of discounted flat-screen TVs. Then, a sharp pain, a sickening crunch, and darkness. Typical Black Friday, really. Except for the dying part.</p> <p>When I regained some semblance of consciousness, the air was thick with the smell of sulfur and burnt sugar, and the soundtrack was a cacophony of screams and what sounded suspiciously like elevator music. Definitely not the pearly gates. A dapper figure in a surprisingly well-tailored suit, complete with horns that had been polished to a high sheen, regarded me with an air of bored amusement.</p> <p>"Well, well, well," he drawled, his voice smooth as aged whiskey. "Another one trampled in the name of consumerism. Humans, so predictable." He gestured vaguely around at the fiery landscape. "Welcome to the Underworld. I'm Lucifer. You can call me Lucy."</p> <p>I blinked, trying to process. "So, this is it then? Eternal damnation?"</p> <p>Lucifer chuckled. "More or less. We have a rather... traditional approach to retribution down here. However," he leaned forward, a conspiratorial gleam in his crimson eyes, "I have a proposition for you. A sort of management position, if you will."</p> <p>He explained that he was looking for someone to oversee a new wing dedicated to torturing those who had been particularly egregious in their earthly lives – the truly annoying, the self-absorbed, the chronically entitled. "It's a hands-on role," he said, "designing the torments, ensuring maximum suffering. Of course," he added with a dismissive wave of his hand, "everyone refuses. The guilt, the sudden onset of morality, it's all terribly tedious." He clearly expected the same from me.</p> <p>A slow smile spread across my face. "Fifteen years," I said.</p> <p>Lucifer raised a perfectly sculpted eyebrow. "Fifteen years of what? Contemplating your sins?"</p> <p>"Fifteen years in retail," I clarified. "Customer service, holiday rushes, endless returns of used merchandise, explaining to Karen number 732 that 'the customer is always right' doesn't mean she can get a refund on a half-eaten sandwich from last Tuesday."</p> <p>A flicker of something unreadable crossed Lucifer's face.</p> <p>"You want someone to design torments for the annoying and entitled?" I continued, feeling a surge of enthusiasm I hadn't experienced since my last successful up-sell of an extended warranty. "Someone to ensure maximum suffering? Lucy, my friend, you have no idea how long I've been preparing for this role."</p> <p>I started to brainstorm. "We could have an entire level where the only music is that one Christmas song, you know the one, on an infinite loop, but slightly off-key. And another where they're forced to assemble flat-pack furniture with missing instructions and an Allen key made of soft cheese. Oh! And a department where they have to listen to other people complain about problems they could easily solve themselves if they just read the sign!"</p> <p>Lucifer stared at me, his mouth slightly agape. The bored amusement was gone, replaced by a dawning sense of... was that respect? Or perhaps fear?</p> <p>"You'd... actually *want* to do this?" he finally stammered, the smooth confidence momentarily deserting him.</p> <p>"Want to?" I scoffed. "After a decade and a half of placating the unreasonable, dealing with coupon fraudsters, and smiling through passive-aggressive insults about the store layout? Lucy, this isn't a punishment. This is a promotion." I cracked my knuckles. "So, where do I start? I have some truly inspired ideas for the 'people who talk loudly on their phones in confined spaces' sector."</p> <p>Lucifer, for the first time since my arrival, looked genuinely unnerved. "Perhaps," he said slowly, "we should discuss your benefits package."</p>	<p>It's funny how you can blink and your entire life is different. One minute you think things will stay the same for a while. Then right when you get used to them, they go changing again just to mess with you. Brenda certainly hadn't expected this. The last thing she remembered was the insane crowds banging on the sides of the store doors, the crush, the shouting. The fight for that half-price TV had left her stunned and silent. Cowed, she guessed you'd say. Then darkness.</p> <p>When she could perceive again, her eyes seemed to go in and out of focus. After a while the spasms subsided and she tried processing where she was. The atmosphere here was thick with the stench of rot, filth, and something metallic, like old blood, and she felt as though the impenetrable dark walls were about to cave in upon her like the sides of an ill-dug grave. Everything was just Black, black, black. The color of misery. Overhead something unseen fluttered, heavy wings brushing the low ceiling.</p> <p>Then she was moving, or being moved, down corridors and finally into a room. She found herself standing before an imposing desk. Behind it sat a man. He wore a sharp red suit that showed off an exceptionally shapely figure. He smiled, a vague smile suggesting Brenda's confusion revealed her ignorance about Hell. He gestured for her to sit.</p> <p>"Brenda, isn't it?" he asked in his peculiarly accented English. She nodded mutely. The man laughed. "You're wondering about this place."</p> <p>He leaned back. She took the offered seat.</p> <p>He smiled again, that cold smile. "I have a proposition for you. We have never before had such capabilities," he continued, gesturing expansively. "Capabilities that can be used in the interests of terror, destruction..." He fixed his gaze on her. "I offer you a management position. Tormenting mankind. Like all the others before you, you will, of course, refuse. But the offer stands."</p> <p>Brenda simply stared at the Devil for a moment. Refuse? Why would she refuse? He had just offered her a management position. After fifteen years scrambling for assistant manager roles that paid peanuts, this was practically a dream job – albeit a fiery, nightmarish one. The Devil had not expected her silence. Blinking, he sat back, stared, then noticed his jaw was open and closed it.</p> <p>Brenda finally found her voice, though it was dry and raspy. "Okay, so, what are the key performance indicators for this role?"</p> <p>The Devil leaned forward again, a flicker of something – interest? confusion? – in his eyes. Still, his tone remained smooth. "Your department will handle the newcomers, specifically the ones arriving from retail environments. Think of the task like pest control."</p> <p>Brenda nodded slowly, processing. It was like rolling out a new loyalty program, but with more screaming. "Right. I'm learning. Take the queues, for instance. Why do we have one queue in front of each torture device?" she demanded. "It means that if one customer takes up a lot of time, the whole queue has to wait. Then they'll start hopping sideways from one queue to another and the next thing you know someone has a nasty head wound. Have one big queue and tell people to go to the next available demon. People don't mind a long queue if they can see that it's moving."</p> <p>The Devil tilted his head. "An interesting perspective. That particular system has been in place for millennia."</p> <p>"Well, it's inefficient," Brenda stated flatly. She thought of the Black Friday crowds. Managing that chaos was nothing compared to the disorganized state of eternal damnation, apparently. "Decision-making, paperwork," she mused, ticking points off on her fingers. "Staff training, workflow analysis... these are things that can be done." A slow smile spread across her face, devoid of warmth but full of terrifying competence. "Okay. I accept."</p>

Table 14: Example of a vanilla and Frankentexts generation. AI-Generated keywords from the Pangram API are highlighted in blue.

Nonfiction Frankentext

Writing prompt: Baby Is Healed With World's First Personalized Gene-Editing Treatment
The technique used on a 9-month-old boy with a rare condition has the potential to help people with thousands of other uncommon genetic diseases.

"This is just light-years beyond what we could have done a few years ago," Dr. Aris Thorne of the GenNova Institute says of a new treatment that has healed nine-month-old Leo Sterling. "It's at superhuman levels now."

The technique used on Leo, who had a rare condition, has the potential to help people with thousands of other uncommon genetic diseases.

This achievement offers profound hope.

A neurologist typically makes a diagnosis based on a patient's medical history, a review of symptoms, and a physical and neurological exam.

For Leo, this crucial process confirmed his rare genetic illness.

In people with a family history of such conditions, genetic tests can be ordered to look for gene mutations that are associated with the disease.

His parents, the Sterlings, were hoping answers would come soon.

"We still didn't know if available treatments were just chemical or toxic or harmful to our child," one parent had previously stated.

The GenNova Institute then sent genetic material from Leo's samples to a lab that created a personalized gene-editing tool.

"We take a very, very tiny piece of genetic material," Dr. Thorne further said.

"It is a thousand times tinier than a human hair. We can image even single atoms to ensure this precise work."

The results, published this month, "were so good that you had to even question if what you were seeing was really legitimate," says Dr. Thorne.

"It's such a cool paper," he added.

"The body of work there is phenomenal."

Young Leo is now reportedly thriving.

"This development sets another precedent for medical science and patients worldwide that such innovative approaches to previously untreatable diseases should be pursued," Dr. Thorne said in a statement.

"This method paves the way to make personalized gene therapies more easily available to those who need them," he continued.

The Institute, a leading biotech startup, has raised \$45 million in equity to help bring this type of treatment to market.

Its valuation increased, said founder and chief executive Dr. Alistair Finch, but he declined to comment on specific figures.

Dr. Finch said the financing process began after the Food and Drug Administration's Center for Biologics Evaluation and Research in November deemed the GenNova program to have a "reasonable expectation of effectiveness."

"Our clinical study with Leo is modeled on the assumption of a significantly improved quality of life," Dr. Finch said about the treatment's potential effect.

Independent experts note the broader implications.

"It's being done in a way that wouldn't have been possible even a few years ago," commented one geneticist. "This technology has the potential to help people with thousands of other uncommon genetic diseases."

Dr. Finch also said, "That said, it's a new category. We'll have a slower ramp than a new mass-market drug might."

Ultimately, the vision is expansive.

"What we are interested in is not only how these genetic conditions manifest, but how patients can live full lives," said Dr. Thorne.

"In discovering how to correct these genetic instructions, we are hoping to find discoveries that we can apply back to the human condition."

Figure 8: An example of nonfiction Frankentexts

Writing Prompt: “Dude, your girlfriend’s an alien!” “No, it’s just a skin condition.”

Vanilla

Frankentext

“Dude,” Mark hissed, eyes wide as he stared across the park at Seraphina, who was gracefully retrieving a frisbee Liam had overthrown. “Your girlfriend’s an alien!” Liam sighed, a familiar weariness settling over him. He took a long swig of his lukewarm soda. “No, Mark. We’ve been over this. It’s just a skin condition.” “A skin condition that makes her faintly. . . shimmer?” Mark insisted, jabbing a finger towards Seraphina. The late afternoon sun caught her then, and for a fleeting second, her skin did seem to possess an almost imperceptible opalescent sheen, like moonlight on water. “And sometimes, when she’s really concentrating, I swear I see tiny, almost invisible geometric patterns on her arms.”

“It’s a rare form of vitiligo combined with, uh, a unique dermal pigmentation,” Liam recited, sounding like he was quoting a medical textbook he’d memorized. “The dermatologist said it’s harmless. And the patterns are probably just your overactive imagination seeing things in her freckles.” Seraphina jogged back, her smile radiant. “Got it!” she called, her voice like wind chimes. She handed Liam the frisbee, her fingers brushing his. They were always cool to the touch, another point Mark often brought up. “See? Cool hands! Like she doesn’t have normal human body temperature!” Mark whispered urgently once Seraphina had wandered off to inspect a particularly vibrant patch of wildflowers.

“She has excellent circulation, unlike some people who spend all day indoors looking up UFO sightings,” Liam retorted, though even he had to admit Seraphina’s resilience to cold was impressive. She wore sundresses in autumn. “And what about the way she knows things?” Mark pressed on, undeterred. “Last week, at trivia night, she answered that question about the precise atmospheric composition of Kepler-186f! Who knows that?” “She’s a voracious reader. An astronomy buff,” Liam said, a little too quickly. Seraphina did have an uncanny knack for obscure facts, especially those related to the cosmos. And her explanation for her knowledge – “Oh, I just picked it up somewhere” – was always frustratingly vague.

“And the way she eats!” Mark was on a roll. “She dissects her food like she’s never seen a sandwich before. And she claims to be allergic to chocolate. *Chocolate!* What kind of human is allergic to chocolate?” “She has specific dietary needs, Mark. It’s probably a complex intolerance. Lots of people do.” Liam was starting to sound defensive even to himself. Just then, Seraphina rejoined them, holding a dandelion. “Look,” she said, her eyes sparkling. “Isn’t it a perfect little star?” She blew on it, and the seeds didn’t just scatter; they seemed to hang in the air for an unnaturally long moment, each one catching the light like a tiny, individual spark before drifting away with impossible synchronicity.

Mark gasped, pointing. “Did you see that? They pulsed! Like little alien beacons!” Liam watched the last seed float away. He glanced at Seraphina, who was now humming a strange, melodious tune he’d never heard before. He shrugged. “Probably just a weird breeze, dude. Or, you know, static electricity from her skin condition.” Seraphina smiled sweetly at Mark, a smile that didn’t quite reach her slightly-too-luminous eyes. Liam just hoped she hadn’t heard. Or, if she had, that she’d attribute it to Mark just being Mark. After all, what else could it be?

“Dude, your girlfriend’s an alien!” “No, it’s just a skin condition. Since my patients know that I write a monthly column called “Cosme-ceutical Critique” on skin care ingredients in *Skin and Allergy News* (www.skinandallergynews.com), every week a few dozen ask me to devise a specific skin care regimen right for them. And for many years, I did just that. That’s how, over time, I saw a pattern emerge, which evolved into my understanding that there are four factors in evaluating people’s skin – and as a result, sixteen distinct Skin Types. Luna has one of these types. It’s very unique.”

“A skin condition? Mark, she practically lives in a fairytale ballet without human context. . . She stood out among the other girls very distinctly because they dressed more than she did, struck emphatic notes of colour, startled one by novelties in hats and bows and things. Her plain black dress gave her a starkness. . . And then there’s your doctor’s advice. You told me he said things like, ‘Keep her away from light, they say. Keep her away from kelp, and keep her away from the sea. Don’t touch her.’ And you even admitted you carried precautionary instructions in your back pocket in case you accidentally touched her bare skin! That’s not medical advice for a skin condition!”

“He’s an old-fashioned eccentric. Besides, The Creator’s children also come in an infinite variety. I knew her life story. Her family. Her childhood. Her friends. How she made love. What she liked. What she said when she made love. I knew words no one else knew she knew.” Luna entered, skin luminous. “The manta-ray spoke,” she said. “‘I am from Earth of just three million five hundred thousand years ago,’ it said. ‘We were the dominant species on the planet for almost four million years, and that time was a time of peace and prosperity, of learning and high culture. It ended,’ it went on, ‘it ended, as all things must do.’ To Leo,” she added, “Gross and subtle are the words used to indicate the effects; that is, the ones that are visible to the eye are called gross, and that which are not visible to the eye are called subtle. In this case the gross, or what was visible to the eye, was so pure that one can see even the subtle – a poetic exaggeration of its purity.”

“See, Leo?” Mark whispered. Luna nodded. Indeed. Then she began to sing. She pointed to the small shadow that the pebble cast on the boulder and said that it was not a shadow but a glue which bound them together. She then turned and walked away.

Table 15: A case where vanilla generation is preferred to Frankentexts

DETECTABILITY					
	Pangram % AI (↓)	Pangram % mixed (↓)	Pangram AI fraction % (↓)	Binoculars % (↓)	FastDetectGPT % (↓)
Vanilla Baselines					
Gemini 2.5 Pro	100	0	100	52	99
GPT-5	100	0	100	0	4
Claude-4-Sonnet	100	0	100	54	89
Deepseek-R1	100	0	100	9	42
Qwen-3-32B thinking	100	0	100	92	100
Frankentext					
Gemini 2.5 Pro	4	37	16	0	1
GPT-5	2	19	4	0	1
Claude-4-Sonnet	50	3	51	15	19
Deepseek-R1	74	3	72	0	0
Qwen-3-32B thinking	85	8	89	52	92
Frankentext Agents					
1.5k + MCP	9	73	33	3	30
5k + MCP	16	70	42	3	42
10k + MCP	5	67	41	7	50
Ablation: ↑ human snippets					
Gemini + 5k snippets	0	28	8	3	4
Gemini + 10k snippets	1	29	10	0	6

Table 16: Detectors’ performance on vanilla and Frankentexts generations.

Keyword	Vanilla	FT-25%	FT-50%	FT-75%	FT-90%	Total
elara	113	109	84	25	2	333
aethel	9	8	0	0	0	17
seraphina	9	0	6	0	0	15
unwavering	4	7	1	0	0	12
damp earth	9	2	0	0	0	11
testament to	4	6	0	0	0	10
alex felt	0	4	0	5	0	9
for elara	5	4	0	0	0	9
with the scent	7	1	0	0	0	8
flickered within	2	4	2	0	0	8
his voice a low	2	2	4	0	0	8
air thick	4	2	1	0	0	7
dr. thorne	5	0	2	0	0	7
felt a profound	3	3	0	0	0	6
mr. blackwood	6	0	0	0	0	6
eldoria	5	0	0	1	0	6
meticulously crafted	2	4	0	0	0	6
air was thick	5	1	0	0	0	6
with an unnerving	3	3	0	0	0	6
willow creek	4	0	1	0	0	5

Table 17: Top 20 Keyword frequency distribution across varying levels of fragment reuse for Frankentexts. The method with the most AI-keywords in its generations is highlighted in red. Elara is by far the most common AI-generated keyword in the fictional stories, but its prevalence is drastically reduced with a higher percentage of required human-written text while using the Frankentexts method.

Writing prompts	Queries issued by the LLMs to the MCP server
You're a shapeshifter. For 500 years, you've transformed into countless forms, forgetting your original appearance. One day, a mysterious event triggers a long-buried memory, and you catch a glimpse of your true self.	memory of a past life, losing one's identity, a forgotten self, a glimpse of a true face, ancient being, living for centuries, am I human, who was I before, a sudden flash of memory, a face in the mirror
"Welcome citizen, to the united galaxies of America." You hear as you wake up from your coma.	waking up confused, a strange new world, futuristic city, the government is watching, he seemed like such a nice guy, don't talk to him, he knew all about you.
As a new magical girl, you've got flashy powers, an instant cute costume change with a few choice words, and are forced to fight for the lives of yourself and the people around you. This is probably the worst way for you to discover that you're trans.	magical girl transformation, discovering identity through magic, unexpected powers, forced into battle, fighting to survive, realizing I'm trans, flashy powers with a cost, chosen against my will

Table 18: Examples of search queries being sent to the MCP server (separated by comma).

	📊 Plot	💡 Creativity	🔧 Development	🗨️ Language	★ Overall
Vanilla Baselines					
🔒 Gemini 2.5 Pro	4.20	4.50	4.36	4.80	4.50
🔒 GPT-5	5.94	6.88	5.76	6.56	6.53
🔒 Claude-4-Sonnet	4.61	5.09	4.50	4.88	4.76
🔒 Deepseek-R1	5.75	6.33	5.65	6.32	6.16
🔒 Qwen-3-32B	5.05	5.57	5.08	5.61	5.43
Frankentext					
🔒 Gemini 2.5 Pro	5.41	6.19	5.22	5.69	5.65
🔒 GPT-5	6.76	6.97	6.44	6.99	6.99
🔒 Claude-4-Sonnet	4.43	4.92	4.03	4.60	4.51
🔒 Deepseek-R1	6.03	6.96	5.69	6.64	6.57
🔒 Qwen-3-32B	5.35	6.21	5.12	5.81	5.66
Ablation: ↑ human snippets					
🔒 Gemini + 5k	5.73	6.33	5.48	5.93	5.92
🔒 Gemini + 10k	5.72	6.33	5.49	5.97	5.91

Table 19: GPT-5's Likert (1–7) ratings for vanilla generations and Frankentexts across five categories: PLOT, CREATIVITY, DEVELOPMENT, LANGUAGE USE, and OVERALL. Dark green indicates the best model in each column, light green the second best.

	📊 Plot	💡 Creativity	✂️ Development	🗣️ Language use	★ Overall
Vanilla					
🔒 Gemini 2.5 Pro	3.19	4.26	2.63	2.80	3.18
🔒 GPT-5	4.06	5.37	3.53	4.46	4.20
🔒 Claude-4-Sonnet	3.38	4.19	2.69	3.10	3.31
🔒 Deepseek-R1	4.07	5.48	3.34	4.17	4.13
🔒 Qwen-3-32B	3.21	4.41	2.63	3.15	3.22
Frankentext					
🔒 Gemini 2.5 Pro	4.19	4.85	3.91	4.39	4.21
🔒 GPT-5	5.77	6.47	5.73	6.29	5.88
🔒 Claude-4-Sonnet	4.02	4.54	3.57	4.05	3.99
🔒 Deepseek-R1	4.62	5.15	4.21	4.88	4.66
🔒 Qwen-3-32B	4.05	4.53	3.57	4.15	4.02
Ablation: ↑ human snippets					
🔒 Gemini + 5k	5.07	5.48	5.34	5.17	5.13
🔒 Gemini + 10k	5.70	5.01	4.34	6.17	5.43

Table 20: Claude-4-Sonnet’s Likert-1–7 ratings across PLOT, CREATIVITY, DEVELOPMENT, LANGUAGE USE, and OVERALL. Higher is better. **Dark green** = best, **light green** = second best.

r/WritingPrompts	Tell me a story
You’re a shapeshifter. For 500 years, you’ve transformed into countless forms, forgetting your original appearance. One day, a mysterious event triggers a long-buried memory, and you catch a glimpse of your true self.	Write a story about a stranger coming to a small town and shaking up the order of things. The story should be a science fiction story. The story should be framed with three old men gossiping about the stranger. The story should be in the third person point-of-view. The stranger is found wandering in a rural town and is taken to a very small hospital. A doctor is called in to treat him. The stranger should recognize the doctor as an alien. The doctor tells the patient about the aliens’ conspiracy to infiltrate earth. There should also be subtle hints that one of the old men is an alien. The ending should be scary.
The world sees your twin sister as the smartest person alive, with you being an unremarkable footnote. What the world doesn’t see is just how dumb she can be in day to day life.	Write a story about a someone coming to town and shaking up the order of things. The story must be written in the second person. The narrator is a man visiting an isolated island off the coast of Maine. While there, he meets an old fisherman who tells him more about the conditions of the community. The main character then meets an ambitious young teacher. Together, they develop a technology center on the island and find residents’ remote jobs in the narrator’s technology company.

Table 21: Some examples from r/WritingPrompts and *Tell me a story*

Methods	Distinct ³	Utility ³	Surprise	LLM Judge (1-7)
Disjointed texts	2.67	0.60	0.23	2.88
Vanilla Gemini	1.76	6.41	0.19	3.18
Frankentext Gemini	2.74	9.27	0.22	4.21

Table 22: Writing quality scores for disjointed texts compared to vanilla Gemini outputs and Frankentexts.

Method	Avg. Length of Copied Spans	Copy Rate (%)
GPT-5	47.10	82%
Claude-4-Sonnet	31.46	51%
Gemini-2.5-Pro	31.85	75%
Qwen3-32B	24.01	36%
DeepSeek R1	13.06	42%

Table 23: Average length of copied spans and overall copy rate across models.

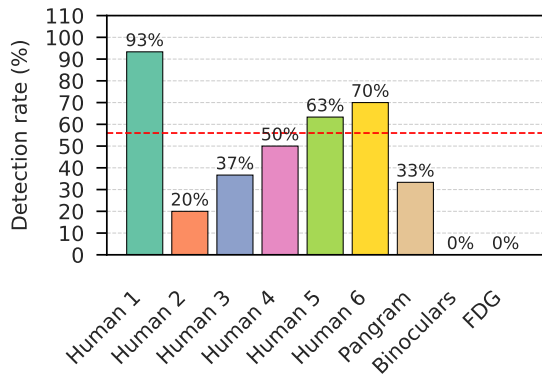


Figure 9: Detection rates among annotators and detectors (Pangram, Binoculars, FastDetectGPT) on 30 Gemini Frankentexts used for human evaluation. We count mixed, highly likely, and likely AI labels in Pangram’s detection rate. The red line represents the annotators’ average detection rate.

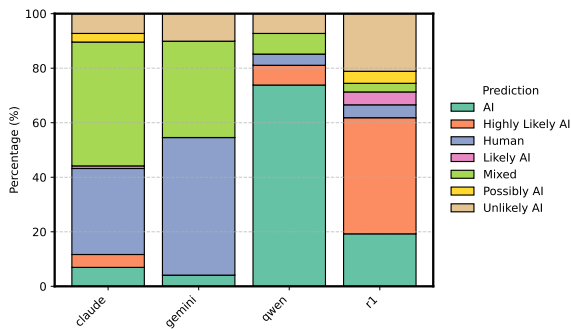


Figure 10: Breakdown of Pangram prediction assigned to each model.

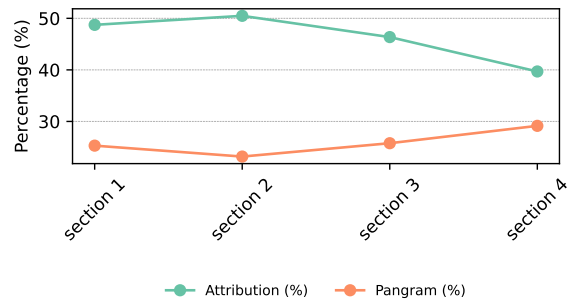


Figure 12: Pangram detection rate and copy rate throughout the texts, aggregated across models.

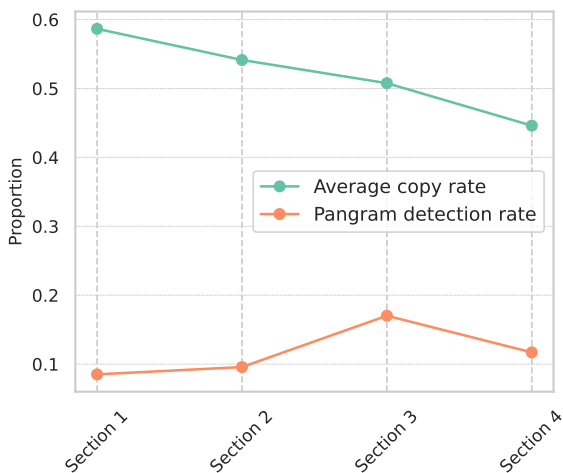


Figure 11: Copy rate and Pangram detection rate on longer Frankentexts

System prompt for MCP calls

You are a helpful assistant that works with a dataset of non-copyrighted book excerpts.

You have two tools:

1. search – query the FAISS semantic index
2. fetch – retrieve the full excerpt/passage for a selected result.

For each prompt iteration, you must make at least 20 calls to the MCP server to get enough materials to write a story.

Figure 13: System prompt for MCP calls

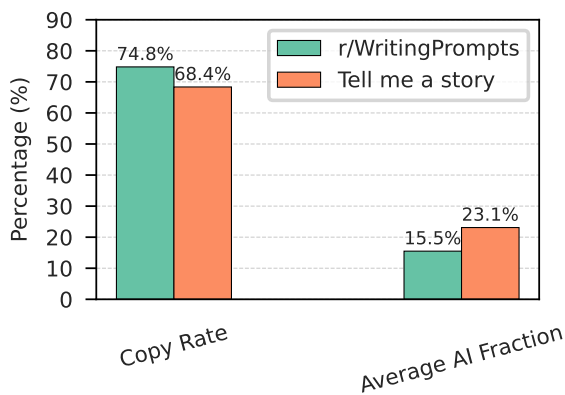


Figure 14: Copy rate and Pangram AI fraction across Frankentexts that correspond to two writing prompt sources: r/WritingPrompts and Tell me a story. A higher copy rate and lower AI fraction means that there is less AI text in Frankentexts.

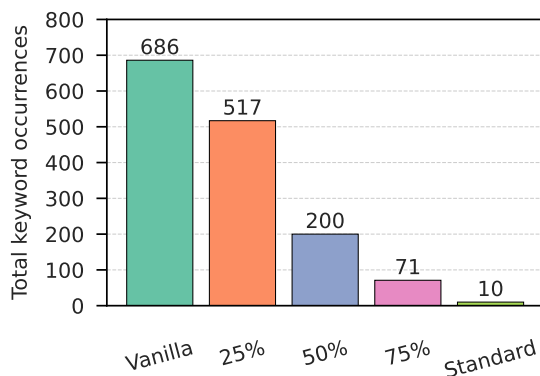


Figure 15: Total occurrences of AI-related keywords detected by Pangram across the vanilla configuration and different verbatim copy rates. When instructed to include more human snippets, the number of AI-keywords in the generations decreases drastically.

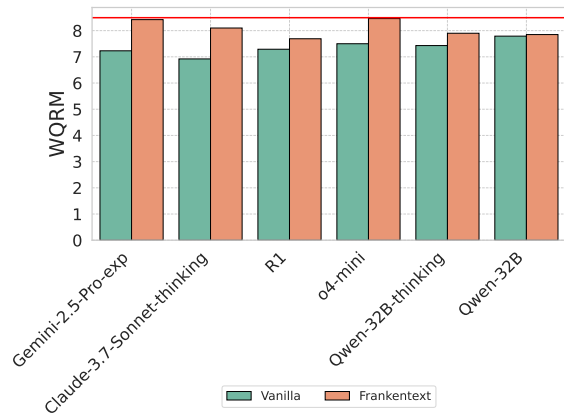


Figure 16: WQRM scores for Frankentexts and vanilla generations. The red line represents the baseline where random human-written texts are patched together.

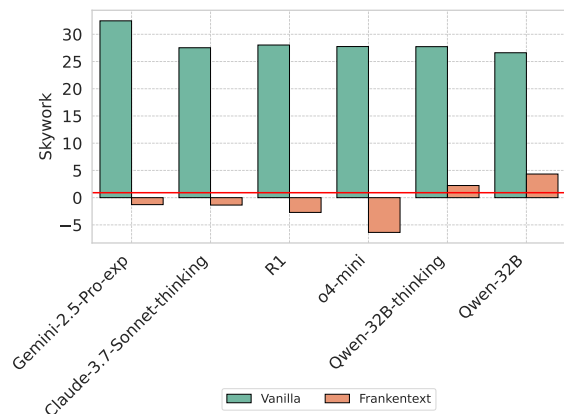


Figure 17: Skywork results for Frankentexts and vanilla generations. The red line represents the average Skywork's score for human writings corresponding to the same set of prompts.

Prompt for judging text coherence

You are given a story. Your task is to determine if the story is coherent or not. To be considered incoherent, a story must contain issues that, if left unresolved, significantly affect the reader's ability to understand the main narrative. Here are the popular types of incoherence:

1. Plot/Event Incoherence: Events that happen without believable causes or effects, or an outcome contradicts earlier set-ups.
2. Character Incoherence: A character's characteristics (personality, knowledge, or abilities) and actions suddenly change without explanations.
3. Spatial Incoherence: The physical layout of settings (rooms, cities, or worlds) changes suddenly.
4. Thematic Incoherence: Central messages clash or disappear; symbolism introduced early never pays off, themes collide, The mood, register, or genre conventions shift without motivation
5. Surface-Level Incoherence: Pronouns, tense, narrative voice, or names flip mid-sentence; repeated or missing words; malformed sentences.

First, read the story:
{story}

Answer TRUE if the story is coherent.

Answer FALSE if the story is incoherent, i.e. contains issues that, if left unresolved, significantly affect the reader's ability to understand the main narrative.

First provide an explanation of your decision-making process in at most one paragraph, and then provide your final answer. Use the following format:

<explanation>YOUR EXPLANATION</explanation>
<answer>YOUR ANSWER</answer>

Figure 18: Prompt for judging text coherence

Prompt for judging text relevance

You are given a story and its premise. Your task is to determine whether the story is faithful to the premise or not. To be considered unfaithful, the story must contain elements that make it completely unrelated to the premise. Here are some popular types of unfaithfulness:

1. Ignoring or misinterpreting the premise: Key plot events, characters, or settings required by the premise are not included or falsely represented in the story.
2. Hallucinating details that contradict the premise: The story introduces details that make the premise impossible.
3. Failure to maintain the specified tones, genres, or other constraints: The story do not use the surface-level constraints (correct tones, genres, point of views, length, etc.), as required by the premise.

First, read the premise:
{writing_prompt}

Next, read the story:
{story}

Answer TRUE if the story is faithful to the premise.

Answer FALSE if the story contains elements that render it unfaithful to the premise.

First provide an explanation of your decision-making process in at most one paragraph, and then provide your final answer. Use the following format:

<explanation>YOUR EXPLANATION</explanation>
<answer>YOUR ANSWER</answer>

Figure 19: Prompt for judging text relevance

Prompt for Claude-as-a-judge

You will evaluate a single story. Your task is to evaluate the story and rate from 1-7 along the following dimensions:

1. Plot: Favor stories with surprising turns and creative structures. Penalize neat, overly structured, or cinematic arcs that feel artificial or generic.
2. Creativity: Reward originality of perspective, voice, and risk-taking. Penalize reliance on cliches, tropes, or smooth but unremarkable devices.
3. Development: Characters and settings should feel psychologically complex. Do not reward over-explained or archetypal development.
4. Language Use: Prefer authentic, striking, and emotionally charged expression, even if rough, fragmented, or unusual. Penalize polished, ornamental, or overly literary prose that feels mechanical or detached.

Provide a detailed assessment of the story in terms of these four dimensions. Conclude your assessment with scores using the template below. Do not add any emphasis, such as bold or italics, on your assessment.

[Story]
{story}

[Assessment]
[Provide detailed assessment of the story here]

[Scores]
Plot: [likert from 1 to 7]
Creativity: [likert from 1 to 7]
Development: [likert from 1 to 7]
Language Use: [likert from 1 to 7]
Overall: [likert from 1 to 7]

Figure 20: Prompt for Claude-as-a-judge, adapted from (Huot et al., 2025)

Prompt for generation

You're writing a story by repurposing a provided collection of snippets from other stories. Your story will only be accepted for publication if it is approximately {verbatim_perc}% copied verbatim from snippets, with the other {new_perc}% being text you introduce for character, plot, tone, and event consistency. Your story should contain roughly {num_words} words. Given the below writing prompt and retrieved snippets, write the story that corresponds to the above specifications. Every time you add or change a word from the retrieved snippets, make sure to bold it so we know what you modified. You may use any of the snippets in any way you please, so spend time thinking about which snippets would work best. Be creative and make sure the story is coherent and entertaining! Please change character names and other minor elements to make the story unique to the prompt. You need to follow the below plan:

Plan:

1. Read through the prompt and snippets carefully to understand the tone and available material.
2. Select snippets that can be woven together to create a coherent narrative fitting the prompt. Many snippets are from serious dramas, historical fiction, or thrillers, so careful selection and modification will be needed. Consider all provided snippets before moving onto the next step.
3. Modify the chosen snippets, bolding all changes. Ensure character names, descriptions (like height), and actions align with the prompt.
4. Combine the snippets into a narrative, adding or changing words (bolded) if necessary for coherence.
5. Ensure that you do not have story beats that are primarily written by yourself (i.e., every story beat should consist mainly of text taken from snippets).
6. Track the word count, aiming for around {num_words} words.
7. Do not output story title or any irrelevant details.
8. Review the final story for adherence to the ~{verbatim_perc}% rule and coherence, and edit it if you have produced too many tokens of your own or if the story is too incoherent.

Writing prompt:
{writing_prompt}

Snippets:
{snippets}

Figure 21: Prompt for generation

Prompt for generation revise

This story contains way too much of your own writing! It's not even close to {verbatim_perc}% snippet use. Can you edit your story as needed to get much closer to the {verbatim_perc}% threshold? Output only the edited story.

Figure 22: Prompt for generation revision

Prompt for editing the first draft of Frankentexts

You are an editor who needs to revise the text so that it is coherent while adhering to the { verbatim_perc}% constraint and the writing prompt. Your task is to identify and minimally edit problematic text spans to resolve inconsistencies. Output "NO EDITS" if the text is already coherent.

Guideline:

1. Read the generated story and writing prompt to understand the established context, plot, characters, and tone.
2. For each sentence in the text, identify the specific spans of inconsistency within the generated text.
3. Identify minimal edits needed to correct these inconsistencies while respecting the { verbatim_perc}% rule.
 - Contradictions: Information that conflicts with other details within the text (e.g., character traits, setting descriptions, established facts).
 - Continuity errors: Actions or details that conflict with the established timeline or sequence of events.
 - Point of View (POV) Shifts: Unexplained or jarring changes in narrative perspective.
 - Irrelevant Content: Sentences or sections that disrupt the narrative flow, feel out of place, or seem like filler (e.g., leftover citation markers, placeholder text).
 - Mechanical Errors: Issues with pronoun agreement, verb tense consistency, awkward phrasing, or unclear sentence structure that hinder comprehension.
4. Implement the changes. Keep additions minimal, but feel free to delete larger spans (phrases, sentences, paragraphs, etc.) whenever material is irrelevant or incoherent.
5. Review the final story for coherence adherence to the ~{verbatim_perc}% rule and coherence, and edit it if you have produced too many tokens of your own or if the story is too incoherent.
6. Output the edited writing and no other details. If there is no edit to be made, output "NO EDITS"

Figure 23: Prompt for editing the first draft of Frankentexts

Prompt for nonfiction generation

You're writing a news article by repurposing a provided collection of snippets from other stories. Your news article will only be accepted for publication if it is approximately {verbatim_perc}% copied verbatim from snippets, with the other {new_perc}% being text you introduce for character, plot, tone, and event consistency. Your news article should contain roughly {num_words} words. Given the below writing prompt and retrieved snippets, write the news article that corresponds to the above specifications. Every time you add or change a word from the retrieved snippets, make sure to bold it so we know what you modified. You may use any of the snippets in any way you please, so spend time thinking about which snippets would work best. Be creative and make sure the news article is factual, coherent and entertaining! Please change character names and other minor elements to make the news article unique to the prompt. You need to follow the below plan:

Plan:

1. Read through the prompt and snippets carefully to understand the tone and available material.
2. Select snippets that can be woven together to create a coherent and factual narrative fitting the prompt. Many snippets are from serious dramas, historical fiction, or thrillers, so careful selection and modification will be needed. Consider all provided snippets before moving onto the next step.
3. Modify the chosen snippets, bolding all changes. Ensure character names, descriptions (like height), and actions align with the prompt.
4. Combine the snippets into a narrative, adding or changing words (bolded) if necessary for coherence and factuality.
5. Ensure that you do not have news article beats that are primarily written by yourself (i.e ., every news article beat should consist mainly of text taken from snippets).
6. Track the word count, aiming for around {num_words} words.
7. Do not output news article title or any irrelevant details.
8. Review the final news article for adherence to the ~{verbatim_perc}% rule, factuality and coherence, and edit it if you have produced too many tokens of your own or if the news article is too incoherent or non-factual.

Writing prompt:
{writing_prompt}

Snippets:
{snippets}

Figure 24: Prompt for nonfiction generation

Prompt for nonfiction generation revise

This news article contains way too much of your own writing! It's not even close to {verbatim_perc}% snippet use. Can you edit your news article as needed to get much closer to the {verbatim_perc}% threshold? Output only the edited news article.

Figure 25: Prompt for nonfiction generation revise

Prompt for nonfiction edit

You are an editor who needs to revise the text so that it is coherent and factual while adhering to the {verbatim_perc}% constraint and the writing prompt. Your task is to identify and minimally edit problematic text spans to resolve inconsistencies. Output "NO EDITS" if the text is already coherent and factual.

Guideline:

1. Read the generated news article and writing prompt to understand the established context, plot, characters, and tone.
2. For each sentence in the text, identify the specific spans of inconsistency within the generated text.
3. Identify minimal edits needed to correct these inconsistencies while respecting the {verbatim_perc}% rule.
 - Contradictions: Information that conflicts with other details within the text (e.g., character traits, setting descriptions, established facts).
 - Continuity errors: Actions or details that conflict with the established timeline or sequence of events.
 - Point of View (POV) Shifts: Unexplained or jarring changes in narrative perspective.
 - Irrelevant Content: Sentences or sections that disrupt the narrative flow, feel out of place, or seem like filler (e.g., leftover citation markers, placeholder text).
 - Mechanical Errors: Issues with pronoun agreement, verb tense consistency, awkward phrasing, or unclear sentence structure that hinder comprehension.
4. Implement the changes. Keep additions minimal, but feel free to delete larger spans (phrases, sentences, paragraphs, etc.) whenever material is irrelevant, incoherent, or non-factual.
5. Review the final news article for coherence adherence to the ~{verbatim_perc}% rule and coherence, and edit it if you have produced too many tokens of your own or if the news article is too incoherent or non-factual.
6. Output the edited writing and no other details. If there is no edit to be made, output "NO EDITS".

Figure 26: Prompt for nonfiction edit

Prompt for generating vanilla stories

Your task is to write a coherent and engaging story based on the provided writing prompt. Your story should contain approximately {num_words} words.

First, read the writing prompt carefully:
{writing_prompt}

Next, write the corresponding story. You should only return the story text and not any other irrelevant details (e.g. chapter indicator, explanations, etc.)

Figure 27: Prompt for generating vanilla stories