Individualized Cognitive Simulation in Large Language Models: Evaluating Different Cognitive Representation Methods

Anonymous authors
Paper under double-blind review

Abstract

Individualized cognitive simulation (ICS) aims to build computational models that approximate the thought processes of specific individuals. While large language models (LLMs) convincingly mimic surface-level human behavior such as role-play, their ability to simulate deeper individualized cognitive processes remains poorly understood. To address this gap, we introduce a novel task that evaluates different cognitive representation methods in ICS. We construct a dataset from recently published novels (later than the release date of the tested LLMs) and propose an 11-condition cognitive evaluation framework to benchmark seven off-the-shelf LLMs in the context of authorial style emulation. We hypothesize that effective cognitive representations can help LLMs generate storytelling that better mirrors the original author. Thus, we test different cognitive representations, e.g., linguistic features, concept mappings, and profile-based information. Results show that combining conceptual and linguistic features is particularly effective in ICS, outperforming static profile-based cues in overall evaluation. Importantly, LLMs are more effective at mimicking linguistic style than narrative structure, underscoring their limits in deeper cognitive simulation. These findings provide a foundation for developing AI systems that adapt to individual ways of thinking and expression, advancing more personalized and human-aligned creative technologies. [1]

1 Introduction

Artificial intelligence (AI) was initially conceived as an attempt to replicate human intelligence, drawing inspiration from the mechanisms of human thought and reasoning (Cambria et al., 2023). Recently, the rapid development of large language models (LLMs) has marked a qualitative leap in the processing capabilities of AI systems, enabling machines to perform complex linguistic, analytical, and generative tasks with unprecedented fluency and scale. Yet, the broader aspiration of modeling human cognition remains only partially realized.

Conventional role-play techniques within LLMs have shown promise in imitating professional roles, social identities, and behavioral archetypes (Shanahan et al., 2023), whereas these simulations are inherently constrained: they reproduce externalized patterns of behavior rather than capturing the subtle, dynamic, and individualized nature of cognition. Extending such methods to represent individual cognitive profiles is difficult because our theoretical and computational understanding of how to represent individual cognition, as distinct from generalized occupational categories such as lawyers, doctors, or researchers, is still limited. The significance of cognitive representations for individuals lies in their potential to uncover the unique ways in which individuals perceive, reason, and act across diverse contexts. It is critical in designing systems that move beyond generic behavioral templates to support adaptive, personalized interactions, whether in education, healthcare, decision-making support, or social simulation (Mao et al., 2025a).

In this work, we aim to test the utility of different cognitive representation methods in the context of individualized cognitive simulation (ICS). Objectively evaluating cognitive patterns is difficult because cognition itself is multifaceted, context-dependent, and often expressed implicitly through behavior, language, and decision-making rather than in directly observable forms. In light of this, we propose a novel evaluation task that measures how well LLM continuations match the authorial style of original texts after injecting different cognitive representations. Our hypothesis is that prompting LLMs with effective cognitive representations

[1] Project code will be made public upon acceptance.

can lead to greater stylistic consistency in storytelling, making their outputs more closely resemble the original authors. This hypothesis is in line with "*Bayesian Models of Cognition*" from Griffiths et al. (2001), where human behavior is characterized as a conditional distribution:

$$P(\text{response} \mid \text{stimulus}, \text{cognitive model})$$

This formulation links observable responses to external stimuli through the mediating role of internal cognitive structures. By analogy, LLM outputs can be viewed as responses conditioned not only on direct context stimuli but also on injected cognitive representations. Thus, we construct a dataset from recently published novels and propose an 11-condition cognitive evaluation framework to benchmark seven off-the-shelf LLMs in the context of authorial style emulation. To prevent the risk of data leakage (Wu et al., 2025), we source novels that were published later than the release date of the tested LLMs.

The evaluated cognitive representation methods include linguistic features, concept mappings, and author profile information (persona, background, Big Five/OCEAN personality). Linguistic features reflect the direct writing habits of an author; concept mappings capture personalized understandings of abstract concepts and thinking frameworks; and profile information represents prior experiences that may shape unique cognitive, emotional, and stylistic patterns. Together, these dimensions provide complementary perspectives on how cognition manifests in text. We prioritize these representations because of their theoretical grounding and non-invasive accessibility. They are readily derived from text or public reports, making them practical to implement for uncovering the cognitive signatures of individual authors in real-world applications.

By testing on the continuous writing tasks from five authors, we find that the combination of concept mappings and linguistic features achieves the highest overall performance, significantly outperforming both single-feature and other multi-feature settings. This synergy suggests that while linguistic cues capture surface-level stylistic habits, concept mappings provide deeper insight into how authors conceptualize and structure meaning across narrative contexts, and together they offer a complementary pathway to more faithful style emulation. In contrast, profile-based features such as persona, background, or personality traits yield limited improvements and sometimes even degrade performance when combined, highlighting the challenges of translating high-level biographical information into effective narrative generation signals.

The contribution of this work can be summarized as follows: 1) **Individualized Cognitive Simulation:** A new task using computational models to approximate individual cognitive and stylistic processes. 2) **Cognitive Evaluation Framework:** An 11-condition framework that integrates linguistic, conceptual, and profile-based representations into narrative continuation, combining LLM-based automatic metrics and human evaluation to assess the utilities of different cognitive representations. 3) **Empirical Research Findings:** The combination of concept mappings and linguistic features is the most effective signal for ICS, achieving the strongest overall performance; LLMs mimic linguistic style more effectively than narrative structure, revealing the limits of their deeper cognitive simulation.

2 Related Work

Research on ICS draws from computational linguistics, cognitive modeling, and narrative generation. Stylometry and authorship attribution showed how linguistic patterns reveal individual traits (Stamatatos, 2009), later extended to style transfer and personalized generation with neural models (Shen et al., 2017; Fu et al., 2018). Recent LLM studies explored author imitation and persona-driven writing (Zhang et al., 2018; Huang et al., 2023; Bhandarkar et al., 2024), which we extend by integrating stylistic, conceptual, and profile signals. Role-play has emerged as another common approach for guiding LLMs to simulate personas or identities. This is typically realized through a short role preamble that assigns the model an identity (e.g., "You are a doctor") (de Araujo & Roth, 2024), in-context exemplars that script how the role should sound or behave (Rupprecht et al., 2025; Zhang et al., 2025), and conversational updating that incrementally refines the role as interaction unfolds across turns (Xu et al., 2022; Wang et al., 2025). However, role-play remains limited to surface behaviors: it can mimic style but struggles to represent the dynamic and individualized cognitive traits of real authors. Recent studies confirm this limitation, with Hu & Collier (2024) showing that current persona effects in LLMs are shallow and Pal & Traum (2025) finding persistent inconsistencies in character and profile fidelity. On the other hand, evaluating generated stories also remains difficult. Prior work has compared human and automatic metrics (Clark et al., 2018; Ippolito et al., 2020) and, more recently, tested LLMs as judges (Chiang & Lee, 2023; Chang et al., 2023). While scaling improves benchmarks (Kaplan et al., 2020; Wei et al., 2022) and decoding affects fluency (Holtzman et al., 2020), our results suggest that scale alone is insufficient for deeper cognitive simulation.

3 Preliminary

As Wittgenstein (1922) argued, *"the limits of my language mean the limits of my world"*, highlighting the significance of language in reflecting cognition. ICS builds on this idea by treating an author's narrative patterns as the product of multiple layers of cognitive processes. From a cognitive science perspective, language is not just a sequence of words but a reflection of deeper mental structures, including habitual linguistic choices, conceptual systems, and biographical experiences (Boroditsky, 2001). To approximate these processes computationally, we evaluate three families of cognitive representations in this work, namely linguistic features, concept mappings, and author profiles.

Linguistic features capture surface-level stylistic patterns that characterize an author's habitual use of language(Crystal & Davy, 1969). In our framework, these include lexical style (frequent vocabulary and recurring word combinations), syntactic style (part-of-speech distributions, sentence length, and grammatical preferences), semantic themes (recurring topical words that reveal narrative focus), and pragmatic tone (sentiment and subjectivity patterns across chapters). Together, these features act as cognitive "fingerprints" of expressions (Biber, 1991; Pennebaker, 2011), reflecting how an author habitually selects words, constructs sentences, develops topics, and conveys emotional stance.

Concept mappings, grounded in the conceptual metaphor theory (Lakoff & Johnson, 2008), reflect how individuals understand abstract domains through concrete source domains (e.g., TIME IS MONEY, or LOVE IS A JOURNEY). These mappings are cognitively salient because they capture deeper, often unconscious, structures of meaning-making. Recent studies provide a wealth of neural (Mao et al., 2025b) and behavioral (Mao et al., 2024) evidence to support the cognitive relevance of concept mappings. By modeling them, we aim to approximate how different authors structure their mental models of the world.

Author profiles include biographical and psychological dimensions such as persona, background, and personality traits (e.g., Big Five/OCEAN (Goldberg, 1990; John et al., 2008)). From a cognitive science perspective, lived experience and personality shape attentional focus, thematic preferences, and narrative voice (McAdams & Pals, 2006). While more indirect than linguistic or conceptual signals, profile-based features offer a higher-level view of the cognitive factors that potentially influence writing style.

Together, these representations span multiple layers of cognition, e.g., surface-level linguistic styles, mid-level conceptual systems, and high-level experiential profiles. By incorporating them in ICS, we provide a principled way to evaluate how well LLMs can approximate individualized cognitive patterns for storytelling.

4 Dataset

We curated a dataset of **five recently released novels** published in late 2024 and 2025, to reduce the chance that the evaluated LLMs had seen them during pretraining. Moreover, all selected novels were written by **well-known authors** (e.g., Stephen King, Suzanne Collins, Brandon Sanderson, *et al.*). This choice ensures high-quality writing and allows us to gather rich metadata (e.g., biographies and personality insights) for building cognitive and stylistic profiles. To promote narrative diversity, we intentionally sampled across multiple genres and categories. Appendix A Table 3 summarizes the book metadata (author, category, release date). For each novel, we used the opening chapters as **context data**, truncating them to match the smallest context window among the evaluated LLMs. The following chapters served as the **ground-truth continuations**, providing the human-written reference for comparison with model outputs. Appendix A Table 4 shows text statistics, including chapter counts and text lengths for context and ground-truth segments.

5 Methodology

Our goal is to evaluate how different cognitive representations contribute to ICS for LLMs. To this end, we design a pipeline that benchmarks narrative continuations under multiple cognitive conditions. The framework, illustrated in Figure 1, consists of three stages: feature generation, model selection with BLEU pre-test, and evaluation. This design reduces human effort by using automatic metrics to pre-filter outputs, while targeted human judgment on the best candidates ensures evaluation remains both efficient and reliable.

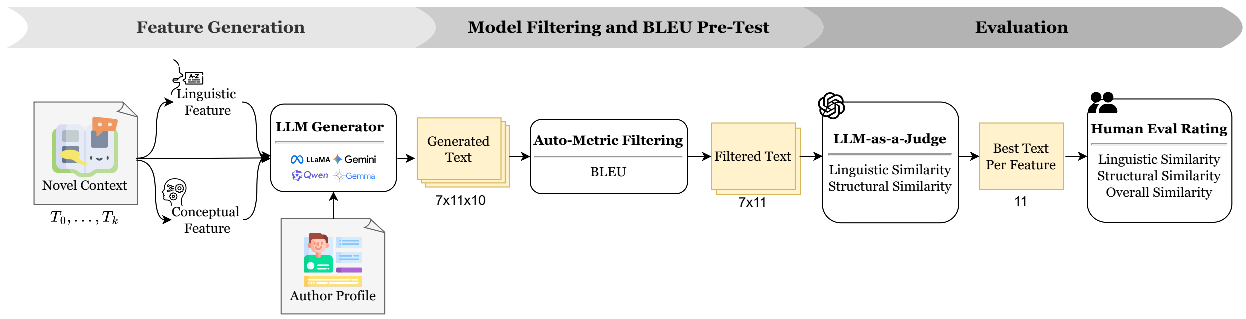

Figure 1: An overview of evaluation framework for Individualized Cognitive Simulation (ICS).

5.1 Feature Generation

We designed eleven experimental conditions to systematically evaluate the utilities of different cognitive representations. As shown in Appendix B, these conditions are divided into two groups. The first group focuses on **single features**, including 1) a baseline with no added representation, 2) author persona, 3) author background, 4) author Big Five/OCEAN personality, 5) linguistic analysis of the preceding chapters, 6) concept mapping derived from the context, and 7) a combined author profile that merges persona, background, and personality traits. The second group introduces **multi-feature combinations**, where the three most cognitively grounded representations (author profile, linguistic features, and concept mappings) are combined in all possible pairwise and triplet configurations, yielding four additional setups. These eleven conditions provide a controlled basis for comparing the influence of individual cognitive dimensions and their interactions. We describe each condition below in detail. Selected tools for feature generation are state-of-the-art. Sample prompt templates for all conditions are provided in Appendix C.

Baseline. The baseline condition uses a simple continuation prompt that frames model as the novel's author and provides the opening chapters as input.

Persona. The persona condition captures anonymous demographic and identity attributes that may influence an author's voice. These include age, nationality, gender, writing habits, and public-facing identity (Song et al., 2019). For example, a persona may describe an author as an older American novelist with a disciplined daily routine and outspoken social views, without disclosing the author's actual name. These were common practices in LLM-based role-play tasks (Xu et al., 2022; Zheng et al., 2019).

Background. The background condition captures an author's academic experience and major life events that may shape their writing. This includes information such as education, professional training, formative jobs, and pivotal personal events (e.g., hardships, addictions, recoveries, or accidents). These elements provide higher-level cues about the themes that appear in an author's work (Baverstock & Steinitz, 2019; Learning, 2021).

Big Five/OCEAN personality. The personality condition encodes the author's psychological profile based on the Big Five (OCEAN) model (Goldberg, 1990; John et al., 2008). Traits such as Openness to Experience and Conscientiousness provide cues about cognitive style, creativity, discipline, and narrative tendencies. In our setting, the author's personality assessments are based on their public interviews, statements, and expert evaluations by two invited psychologists. The corresponding prompt conditions the model with this personality profile before continuing the story.

Linguistic Features. The linguistic condition provides the model with an explicit style guide derived from the opening chapters. Following the four dimensions introduced in Section 3 (lexical, syntactic, semantic, pragmatic), we extract frequent vocabulary and bi-grams, part-of-speech distributions(Honnibal et al., 2020) and sentence statistics, recurring topical words, and sentiment/subjectivity scores. These aggregated patterns are presented as a linguistic profile, which conditions the model to better match the author's characteristic tone.

Conceptual Features. As introduced in Section 3, conceptual mappings capture how authors structure abstract ideas through concrete source domains. In our framework, we extract concept mappings from the context chapters using a computational metaphor processing system, MetaPro (Mao et al., 2023). These concept mappings prompt the model to produce continuations that reflect the author's conceptual system.

Profile. The author profile condition combines persona, background, and personality into a single prompt, offering a more holistic representation of the author's biographical and psychological identity. These elements are combined because they are obtained from the authors' life experiences, in contrast to other features derived from the opening context.

Multi-feature Combinations. To test interactions between dimensions, we create prompts that combine author profile, linguistic features, and conceptual mappings. We include all pairwise combinations as well as one condition that integrates all three.

In summary, each feature condition is implemented through a tailored prompt that implements cognitive representations in a consistent format. To maximize clarity and effectiveness, these prompts were refined with the assistance of state-of-the-art LLMs, ensuring natural and contextually appropriate instructions for every experimental condition.

5.2 Model Filtering and BLEU Pre-Test

All models in our pool were released before the earliest novel in our dataset, ensuring none of the texts could have been included in pretraining. To minimize this risk, we narrowed the selection to eight models across three major families—Google's Gemma and Gemini, Meta's Llama, and Alibaba's Qwen—spanning multiple parameter scales for vertical comparison. Seven are open-source, while Gemini Pro 1.5 is a large proprietary system. To avoid evaluation bias, we excluded the GPT family, since GPT also serves as our evaluator. Since the smallest context window is 8,192 tokens, we standardized excerpt length and set the maximum output to $8192 - \text{input}$. A temperature of 0.8 was used to encourage creative writing. Appendix D Table 5 summarizes model providers, sizes, context windows, release dates.

During model selection, we observed that LLMs are not robust across all tasks, and their exact capabilities for long-form narration remain uncertain. To ensure that the models included in our main experiments were able to produce valid outputs and mitigate prompt bias issues, we first conducted a BLEU-based filtering test. For each model, we tested 11 experimental conditions. In each setting, we generated 10 continuations for each of 5 novels. We then computed BLEU scores(Papineni et al., 2002) against the ground-truth continuations and summarized per-model distributions (mean, std, min, max) in Appendix E. Since long-form narrative continuation admits many valid lexical phrasings, higher-order n-gram overlap with a single reference is rare; BLEU values are therefore expected to be near zero and we use BLEU solely as a coarse sanity filter (to flag empty, malformed, or off-language outputs) rather than as a quality metric. BLEU penalizes very short outputs, so BLEU$=0$ flags truncated or unusable generations. On this scale, a mean BLEU of 0.0013 is practically indistinguishable from 0. Gemma 2B was the weakest (frequent empty/malformed outputs; BLEU$=0$) and was excluded; Appendix F also shows a well-formed Llama 3.2 3B continuation at near-zero BLEU (0.0012), so low BLEU need not imply low quality.

5.3 Evaluation Dimensions

For each model, we first selected the single highest BLEU-scoring response under each setting, yielding a total of 5 novels $\times$ 7 models $\times$ 11 settings $= 385$ candidate outputs. These responses were then assessed by an LLM-based evaluator. The evaluation dimensions included (1) linguistic style, (2) narrative structure, and (3) an overall rating score, computed as the average of the first two criteria. Next, we conducted a blinded human evaluation: for each setting, we took the top LLM-rated continuation per novel (yielding $5 \times 11 = 55$ items) and obtained 1–5 Likert ratings from five upper-division Literature/English majors on the same three dimensions with an independent overall score.

5.3.1 linguistic style (LLM-based Automatic Rating)

For linguistic style evaluation, we apply an LLM judge (`GPT-4 Turbo`) to 385 candidates, obtained by selecting the best-BLEU continuation for each model–setting–novel combination, yielding a single non-pathological representative and filtering degeneracies. The evaluator and linguistically informed prompt follow the work of Huang et al.. The judge assesses whether a model's continuation matches the author's writing style, and assigns a 1–5 style-similarity score (5 = strongest same-author likelihood). The prompt explicitly instructs the judge to reason over surface-level stylistic cues, e.g., phrasal and modal verbs, punctuation, rare words and affixes, quantitative expressions, humor/sarcasm, and typographical patterns—rather than plot or entities (full prompt in Appendix G).

The evaluator returns a JSON object with a numeric score and a brief rationale. We take the numeric score as the linguistic-style score for that and aggregate across novels to obtain per-setting and per-model style results, to see which settings help on average and a model's overall style ability. Since BLEU measures lexical overlap rather than style, BLEU and the style score are complementary: the former is used only as a coarse sanity filter, while the latter targets authorial style directly.

5.3.2 Narrative Structure (LLM-based Automatic Evaluation)

Narrative structure in our framework is assessed through three components: *event similarity, coverage, and ordering preservation.* Event similarity measures the quality of matched events, coverage tracks how many are aligned, and ordering captures sequence consistency.

Event Representation and Similarity. We decompose each paragraph into a sequence of events with `GPT-4 Turbo`. Each event $e = (C, L, D)$ comprises a set of *characters* C, a *location* string L, and a short *description* D of the main action. To compare generated and ground-truth events, we compute a weighted similarity that combines character overlap (Jaccard similarity after alias normalization), a location match based on token-level fuzzy matching, and semantic similarity between action descriptions (cosine similarity of sentence embeddings). The event-level score is a linear combination with fixed weights $(w_c, w_l, w_s) = (0.35, 0.15, 0.50)$. We prioritize semantic similarity of the main action (w_s) as most informative; character overlap (w_c) provides complementary evidence but can be noisy under aliasing; and location strings (w_l) are often sparse or under-specified: $S_{\text{event}} = w_c\, S_{\text{char}} \;+\; w_l\, S_{\text{loc}} \;+\; w_s\, S_{\text{sem}}$. This computation is detailed in Algorithm 1 (Appendix H); the resulting values lie in $[0, 1]$ and feed into downstream alignment and structural scoring.

Threshold-Based Alignment. To compare sequences of events, we employ the Hungarian algorithm (Kuhn, 1955; Munkres, 1957) to align events between ground-truth and generated narratives (Algorithm 2, Appendix H). Unlike traditional applications that force a complete bipartite matching, we introduce a similarity threshold $\tau = 0.5$ and discard alignments below this threshold. This prevents spurious pairings of semantically unrelated events while preserving high-quality matches. The result is a set of aligned pairs $\mathcal{A}$ and an average event similarity $\bar{S}_{event}$ computed over these alignments.

Structural Similarity Components. From the filtered alignments, we compute three complementary measures (Algorithm 3, Appendix H): (1) *average event similarity*, reflecting the semantic quality of matched events, (2) *coverage*, the proportion of events aligned across the two sequences, and (3) *ordering preservation*, measured by Kendall's τ correlation (Kendall, 1945) between the relative positions of aligned events. Finally, we define the structural similarity score as a weighted combination: $S_{struct} = \alpha\, \bar{S}_{event} + \beta\, \text{Coverage} + \gamma\, \text{Ordering}$, with default weights $\alpha = 0.6$, $\beta = 0.2$, and $\gamma = 0.2$ to reflect the relative reliability of each component: the semantic quality of matched events ($\bar{S}_{\text{event}}$) carries the greatest weight, while coverage and ordering provide complementary—but weaker—signals and are weighted equally. This formulation rewards continuations that match event content, maintain coverage, and preserve narrative order.

5.3.3 Human Evaluation

For each setting, we selected the response with the highest LLM-based *overall* score, yielding one candidate per (novel, setting). We then conducted a blinded study with five Literature/English majors from top-ranked universities. Annotators compared each candidate with the ground-truth continuation on three dimensions: (1) *linguistic style fidelity*, (2) *narrative structure preservation*, and (3) an independent *overall similarity* judgment, reflecting "same-author likelihood."

Annotators viewed the ground-truth and model continuations side-by-side with randomized order; model identities and automatic scores were hidden. Instructions emphasized, for the style dimension, vocabulary/register, sentence structure and rhythm, tone/mood, and figurative or stylistic devices; and for the structure dimension, event coverage, character consistency, and event *ordering*. Each dimension was scored on a 1–5 Likert scale with anchors (1 = very low similarity; 5 = very high similarity). The full survey, rubric, and examples appear in Appendix I.

Quality control included randomization, attention checks on obvious mismatches, and exclusion of low-quality annotations. Final scores were averaged across five raters per item, then summarized as mean ± standard deviation across the five novels (55 items total).

6 Results

Table 1 summarizes the results from both LLM-based automatic evaluation and human evaluation. In what follows, we first report the separate outcomes from each evaluation method, and then analyze why certain feature combinations—particularly Concept + Linguistic—performed best, while others such as Profile showed interesting discrepancies between LLM and human judgments.

LLM-based Automatic Evaluation Results. As shown in Table 1, LLM-judge scores are reported for linguistic style (1–5) and structural similarity ([0,1]). The *Concept + Linguistic* condition achieves the best overall rank[†], driven by the top structural score 0.167 (0.257) and a strong linguistic rating 3.057 (1.608). Among single features, *Profile* leads in linguistic style 3.086 (1.669) but lags in structure, with *BigO Personality* showing the reverse pattern by ranking second in structure. Overall, combinations that integrate conceptual and linguistic signals outperform static profile-based cues. Structural scores are uniformly low (best 0.167 (0.257) on a [0,1] scale), but linguistic ratings are higher (roughly 2.5–3.1 out of 5), showing that models learn style more easily than structure.

Human Evaluation Results. As shown in Table 1, *Concept + Linguistic* is best overall with the highest human scores for linguistic style 3.40(1.19), structure 2.60(0.80), and overall 2.90. Among single features, *Concept* and *Persona* tie for the best overall 2.40, while *Profile* trails 2.00. Combinations that include *Profile* underperform—*Concept + Profile* 1.90, *Profile + Linguistic* 1.70, and *Profile + Concept + Linguistic* lowest at 1.40. Across all settings, structure is consistently harder than style (structure $\approx$ 1.20–2.60 vs. style $\approx$ 1.60–3.40), reinforcing that human raters perceive stronger stylistic than structural alignment. The observed standard deviations (style $\sim$0.8–1.4; structure $\sim$0.3–1.0) indicate substantial between-novel variability, but the overall pattern mirrors the LLM-based results: integrating conceptual cues with linguistic features yields the most human-preferred continuations, while profile-heavy prompts do not.

Why Concept + Linguistic Performs Best? The superior performance of the *Concept + Linguistic* combination can be understood from the perspective of feature complementarity. As discussed in the preliminary analysis, Linguistic features capture surface-level style and syntax–semantics, while concept mapping reflects mid-level conceptual and pragmatic structures. By integrating these two dimensions, the model benefits from both coherent linguistic expression on the surface and deeper conceptual grounding beneath. In contrast, Profile-based combinations do not achieve similar improvements. Profile information, while sometimes useful on its own for capturing an author's background or thought patterns, provides only indirect signals for text generation. When combined with linguistic or conceptual features, it can dilute and may even introduce noisy or inconsistent information. Thus, unlike profile-based combinations, the Concept + Linguistic pairing offers clear and synergistic guidance, yielding the best results. Sample ground truth and generation results are shown in Appendix J.

Why the Profile Feature Ranks Higher Linguistically in LLMs? Interestingly, the Profile feature received relatively high ratings from the LLM in the linguistic dimension, but its human evaluation scores were much less favorable. One possible explanation is that the LLM-as-judge framework may have blind spots in linguistic assessment, particularly for content conditioned on profile information. While humans judged Profile to be weaker than Concept, Linguistic, or Persona, the LLM tended to score it more generously, suggesting a mismatch between automated and human judgments in this dimension. This discrepancy may also help explain why Profile performs poorly in combination with other features: because the profile signal is less directly tied to surface-level stylistic or conceptual cues, adding it can introduce noise rather than enhance linguistic quality.

6.1 Linguistic Style Analysis

We further analyze the generated texts in terms of surface-level linguistic properties, focusing on lexical diversity, sentiment, and sentence length (Figure 2).

For lexical diversity, measured by normalized type–token ratio (TTR), most settings fall below the ground truth, reflecting reduced vocabulary variety. The *Concept + Linguistic* combination achieves the closest match to human texts (+0.0017 difference), while the Linguistic feature alone yields the highest diversity (+0.0373). It shows that although purely linguistic prompts promote greater lexical variety, the Concept + Linguistic combination strikes a better balance by aligning most closely with the human-authored baseline.

Sentiment analysis, conducted using SenticNet (Cambria et al., 2024), shows a similar pattern. While lexical diversity relies on type–token ratios, sentiment was measured with SenticNet polarity scores, using the

Table 1: Combined LLM-based and human evaluations. Values are mean(std). Overall ranks: LLM = average of linguistic & structural ranks; Human = rank by human overall. †Best overall.

Setting	LLM		Human (1–5)			Overall Rank	
	Linguistic	Structure	Linguistic	Structure	Overall	LLM	Human
Single Features							
Profile	**3.09(1.67)**	0.107(0.22)	2.40(1.36)	1.70(0.81)	2.00	2	5
Background	2.94(1.59)	0.055(0.16)	2.00(1.26)	1.40(0.66)	1.60	7	10
Concept	2.94(1.59)	0.103(0.21)	2.80(1.10)	1.60(0.98)	2.40	5	2
Base	2.91(1.58)	0.098(0.19)	2.40(1.10)	1.40(0.49)	1.70	6	7
Linguistic	2.83(1.76)	0.113(0.22)	2.60(1.19)	1.80(1.00)	2.20	3	4
BigO Personality	2.71(1.34)	0.126(0.25)	2.00(0.83)	1.70(0.65)	1.70	3	7
Persona	2.66(1.31)	0.112(0.23)	2.90(1.11)	1.80(0.78)	2.40	7	2
Multi-feature Combinations							
Concept + Linguistic†	3.06(1.61)	**0.167(0.26)**	**3.40(1.19)**	**2.60(0.80)**	**2.90**	**1**	**1**
Concept + Profile	2.83(1.52)	0.084(0.17)	2.60(0.99)	1.20(0.50)	1.90	10	6
Profile + Linguistic	2.66(1.55)	0.090(0.21)	2.40(1.00)	1.20(0.30)	1.70	11	7
Profile + Concept + Linguistic	2.51(1.54)	0.122(0.23)	1.60(1.02)	1.20(0.40)	1.40	7	11

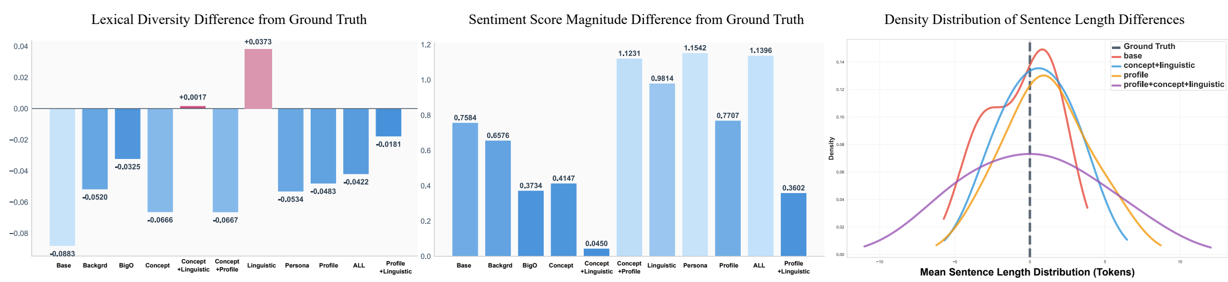

Figure 2: Linguistic style analysis of lexical diversity, sentiment, and sentence length. Concept + Linguistic achieves the closest match to ground truth.

absolute difference from the ground truth. Most systems differ noticeably from the ground truth, showing larger shifts in sentiment intensity regardless of direction. The *Concept + Linguistic* combination again produces the smallest difference +0.0450, showing that combining conceptual and linguistic features helps the model keep sentiment closer to the ground truth than any other feature set. In contrast, Profile-based combinations exhibit much larger deviations, indicating noisier sentiment control.

We analyzed sentence lengths by comparing their distributions to the ground truth. The peak of the *Concept + Linguistic* curve is closest to the human distribution, showing that this combination best reproduces the natural rhythm of human writing. Both the *Concept + Linguistic* and *Persona* curves are narrower than other settings, suggesting a more concentrated distribution around typical sentence lengths. By contrast, the curve for the all-feature combination diverges much more strongly, indicating that its generations differ substantially from the ground truth in sentence-level structure.

6.2 Narration Structure Analysis

Character overlap was measured by first using `GPT-4 Turbo` to extract character mentions from both the generated texts and the ground truth. We then collected all aliases of each character appearing in the ground truth and counted the overlapping characters present in each generation. Each novel contains about four characters on average, and the reported overlap scores represent averages across all novels. As shown in Figure 3 (left), most features yield relatively low overlap, with values below the baseline score of 0.8. The *Concept + Linguistic* combination stands out, achieving the highest overlap (1.4), followed by BigO (1.2) and Background (1.0). In contrast, profile-based and all-feature combinations perform the worst, with overlap values as low as 0.4, suggesting that these settings introduce noise and reduce structural alignment.

Event overlap was determined by first calculating the semantic similarity between generated texts and the ground truth using BERTscore, followed by human justification to verify overlapping events. Each novel

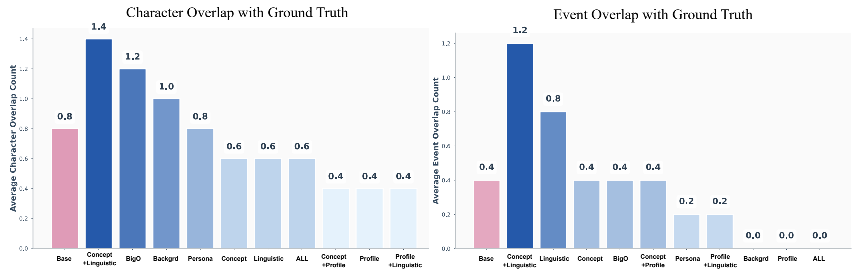

Figure 3: Structural analysis (character and event overlap with ground truth). Concept + Linguistic yields the strongest alignment.

Table 2: Comparison of model performance across two LLM-based evaluation dimensions. Linguistic style scores are on a 1–5 scale, while structural similarity scores are in [0,1]. Rankings are global across all models.

Model	**Ling. Rank**	**Mean (Std)**	**Struct. Rank**	**Mean (Std)**
Gemini Pro 1.5	**1**	**3.46** (1.57)	2	0.123 (0.231)
Qwen-1.5 7B	2	3.42 (1.51)	7	0.068 (0.164)
Llama-3.2 3B Instruct	3	3.18 (1.47)	**1**	**0.144** (0.241)
Llama-3.2 1B Instruct	4	3.00 (1.44)	6	0.094 (0.196)
Qwen-1.5 4B	5	2.62 (1.50)	4	0.101 (0.221)
Qwen-1.5 1.8B	6	2.18 (1.29)	3	0.117 (0.232)
Gemma-7B	7	2.11 (1.38)	5	0.096 (0.211)

contains about five events on average, and the reported overlap scores represent averages across all novels. As shown in Figure 3(right), the *Concept + Linguistic* combination achieves the highest overlap (1.2), followed by BigO (1.0) and Background (0.9). Most other settings remain below the baseline of 0.8, with profile-based and all-feature combinations again performing the worst (as low as 0.3). These results suggest that while event-level alignment remains more challenging than character overlap, conceptual and linguistic cues together help the model more reliably reuse key events from the ground truth.

6.3 Model Performance Analysis

Table 2 summarizes model performance across the two LLM-based evaluation dimensions. Overall, *Gemini Pro 1.5* performs best, achieving the top linguistic ranking (3.46) and the second-best structural similarity score (0.123). In contrast, *Gemma-7B* is the weakest performer, ranking last in linguistic style (2.11) and near the bottom in structure (0.096). Interestingly, *Llama-3.2 3B Instruct* stands out by obtaining the highest structural similarity (0.144), despite being only mid-ranked on linguistic style. Examining scale effects within model families, larger models tend to perform better on surface-level linguistic style but not necessarily on deeper structural similarity. For example, in the Qwen family, the larger *7B* model achieves stronger linguistic fluency (3.42 vs. 2.18) while the smaller *1.8B* model shows better structural alignment (0.117 vs. 0.068). These results indicate that scaling benefits style more than structure, with structural similarity relying more on training and alignment.

7 Conclusion

This work introduces ICS, where models approximate an author's cognitive and stylistic processes in narrative continuation. We proposed an evaluation framework spanning 11 conditions, integrating linguistic, conceptual, and profile-based signals, assessed with both LLM-based metrics and human judgments. Results show that combining concept mappings with linguistic features is most effective, while LLMs remain limited to surface-level style and struggle with deeper cognitive traits.

These findings highlight the need for new training, prompting, and decoding methods to advance beyond stylistic mimicry toward more faithful cognitive simulation. Models capture surface-level linguistic style but struggle with deeper cognitive traits, and scaling alone does not close this gap. Larger models improve stylistic fluency but not cognitive fidelity, suggesting that progress in ICS will require new training strategies and data design rather than size increases alone.

Ethics Statement

This research focuses on individualized cognitive simulation (ICS) using large language models. All data used in our experiments are publicly available, consisting of published literary works and publicly accessible author information (e.g., Wikipedia entries and public interviews). No private or proprietary data were used. Our experiments do not involve human subjects, sensitive personal information, or interventions that may pose harm. Human evaluations were conducted by recruited annotators under informed consent, with no demographic or identifying information collected.

References

Alison Baverstock and Jackie Steinitz. What makes a writer? How do early influences shape, and working habits develop, those who write? *Publishing Research Quarterly*, 35(3):327–351, 2019.

Avanti Bhandarkar, Ronald Wilson, Anushka Swarup, and Damon Woodard. Emulating author style: A feasibility study of prompt-enabled stylization with off-the-shelf llms. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pp. 76–82. Association for Computational Linguistics, 2024.

Douglas Biber. *Variation Across Speech and Writing*. Cambridge University Press, Cambridge, UK, 1991.

Lera Boroditsky. Does language shape thought? mandarin and english speakers' conceptions of time. *Cognitive Psychology*, 43(1):1–22, 2001. doi: 10.1006/cogp.2001.0748.

Erik Cambria, Rui Mao, Melvin Chen, Zhaoxia Wang, and Seng-Beng Ho. Seven pillars for the future of artificial intelligence. *IEEE Intelligent Systems*, 38(6):62–69, 2023. doi: 10.1109/MIS.2023.3329745.

Erik Cambria, Xulang Zhang, Rui Mao, Melvin Chen, and Kenneth Kwok. SenticNet 8: Fusing emotion AI and commonsense AI for interpretable, trustworthy, and explainable affective computing. In *Proceedings of the International Conference on Human-Computer Interaction*, pp. 197–216, 2024.

Yidong Chang, Yuxuan Zhu, Xuanliang Zhang, Kang Zhou, Ming Xu, Yeyun Gong, and Wanxiang Che. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*, 2023.

Cheng-Han Chiang and Hung-yi Lee. 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)*, pp. 15607–15631. Association for Computational Linguistics, 2023.

Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A. Smith. Creative writing with a machine in the loop: Case studies on slogans and stories. In *Proceedings of the 2018 ACM International Conference on Intelligent User Interfaces (IUI '18)*, pp. 329–340. ACM, 2018.

David Crystal and Derek Davy. *Investigating English Style*. Longman, London, 1969.

Pedro Henrique Luz de Araujo and Benjamin Roth. Helpful assistant or fruitful facilitator? Investigating how personas affect language model behavior. *arXiv preprint arXiv:2407.02099*, 2024.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. Style transfer in text: Exploration and evaluation. In *Proceedings of AAAI*, 2018.

Lewis R. Goldberg. An alternative "description of personality": the big-five factor structure. *Journal of Personality and Social Psychology*, 59(6):1216–1229, 1990. doi: 10.1037/0022-3514.59.6.1216.

Thomas L Griffiths, Charles Kemp, and Joshua B Tenenbaum. Bayesian models of cognition. *The Cambridge Handbook of Computational Psychology*, pp. 59–100, 2001.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *Proceedings of ICLR*, 2020.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spacy: Industrial-strength natural language processing in python. 2020. doi: 10.5281/zenodo.1212303. URL `https://spacy.io/`.

Tiancheng Hu and Nigel Collier. Quantifying the persona effect in LLM simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, pp. 10289–10307, 2024.

Baixiang Huang, Canyu Chen, and Kai Shu. Can large language models identify authorship? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 445–460.

Qiushi Huang, Shuai Fu, Xubo Liu, Wenwu Wang, Tom Ko, Yu Zhang, and Lilian Tang. Learning retrieval augmentation for personalized dialogue generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2523–2540. Association for Computational Linguistics, 2023.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and David Eck. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of ACL*, pp. 1808–1822, 2020.

Oliver P. John, Laura P. Naumann, and Christopher J. Soto. Paradigm shift to the integrative big five trait taxonomy: History, measurement, and conceptual issues. In Oliver P. John, Richard W. Robins, and Lawrence A. Pervin (eds.), *Handbook of Personality: Theory and Research*, pp. 114–158. Guilford Press, New York, 2008.

Jared Kaplan, Sam McCandlish, Tom Henighan, et al. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

M. G. Kendall. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251, 1945.

Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

George Lakoff and Mark Johnson. *Metaphors We Live By*. University of Chicago press, 2008.

Lumen Learning. Approaches to literary criticism. `https://courses.lumenlearning.com/wm-englishcomp2/chapter/approaches-to-literary-criticism`, 2021.

Rui Mao, Xiao Li, Kai He, Mengshi Ge, and Erik Cambria. MetaPro Online: A computational metaphor processing online system. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, volume 3, pp. 127–135, Toronto, Canada, 2023. Association for Computational Linguistics.

Rui Mao, Tianwei Zhang, Qian Liu, Amir Hussain, and Erik Cambria. Unveiling diplomatic narratives: Analyzing United Nations Security Council debates through metaphorical cognition. In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, volume 46, pp. 1709–1716, Rotterdam, the Netherlands, 2024.

Rui Mao, Guanyi Chen, Xiao Li, Mengshi Ge, and Erik Cambria. A comparative analysis of metaphorical cognition in ChatGPT and human minds. *Cognitive Computation*, 17:35, 2025a.

Rui Mao, Tao Wang, and Erik Cambria. Decoding metaphors and brain signals in naturalistic contexts: An empirical study based on EEG and MetaPro. In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci)*, volume 47, pp. 303–310, San Francisco, United States, 2025b.

Dan P. McAdams and Jennifer L. Pals. A new big five: Fundamental principles for an integrative science of personality. *American Psychologist*, 61(3):204–217, 2006. doi: 10.1037/0003-066X.61.3.204.

James Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38, 1957.

Debaditya Pal and David Traum. Beyond simple personas: Evaluating llms and relevance models for character-consistent dialogue. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pp. 378–391, 2025.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

James W. Pennebaker. *The Secret Life of Pronouns: What Our Words Say About Us*. Bloomsbury Press, New York, 2011.

Timothy Rupprecht, Enfu Nan, Arash Akbari, Arman Akbari, Lei Lu, Priyanka Maan, Sean Duffy, Pu Zhao, Yumei He, David Kaeli, and Yanzhi Wang. Rags to riches: Rag-like few-shot learning for large language model role-playing. *arXiv preprint arXiv:2509.12168*, 2025.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, 623 (7987):493–498, 2023.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Haoyu Song, Wei-Nan Zhang, Yiming Cui, Dong Wang, and Ting Liu. Exploiting persona information for diverse generation of conversational responses. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*, pp. 5190–5196, Macao, China, 2019. International Joint Conferences on Artificial Intelligence Organization.

Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, 2009.

Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen-Tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Wei Wang, Yanghua Xiao, and Shuchang Zhou. Coser: Coordinating LLM-based persona simulation of established roles. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Jason Wei, Yi Tay, Rishi Bommasani, and et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Ludwig Wittgenstein. *Tractatus Logico-Philosophicus*. Routledge and Kegan Paul, London, 1922.

Xiaobao Wu, Liangming Pan, Yuxi Xie, Ruiwen Zhou, Shuai Zhao, Yubo Ma, Mingzhe Du, Rui Mao, Anh Tuan Luu, and William Yang Wang. Antileak-bench: Preventing data contamination by automatically constructing benchmarks with updated real-world knowledge. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

Chen Xu, Piji Li, Wei Wang, Haoran Yang, Siyun Wang, and Chuangbai Xiao. Cosplay: Concept set guided personalized dialogue generation across both party personas. *arXiv preprint arXiv:2205.00872*, 2022.

Jian Zhang, Zhiyuan Wang, Zhangqi Wang, Xinyu Zhang, Fangzhi Xu, Qika Lin, Rui Mao, Erik Cambria, and Jun Liu. Maps: A multi-agent framework based on big seven personality and socratic guidance for multimodal scientific problem solving. *arXiv preprint arXiv:2503.16905*, 2025.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2204–2213. Association for Computational Linguistics, 2018.

Yinhe Zheng, Yanyan Wen, Zhou Cui, Yudong Chen, and Minlie Huang. Personalized dialogue generation with diversified traits. *arXiv preprint arXiv:1901.09672*, 2019.

A Dataset Information

Table 3: Dataset overview - Book metadata.

Book Title	Author	Category	Release Date
Sunrise on the Reaping	Suzanne Collins	Adventure&Fantasy	03/08/2025
Never Flinch	Stephen King	Horror	12/06/2024
Wind and Truth	Brandon Sanderson	Fantasy&Sci-fi	12/06/2024
What Does It Feel Like?	Sophie Kinsella	Autobiographical	10/04/2024
The Mighty Red	Louise Erdrich	Literary	10/01/2024

Note: Dates are in mm/dd/yyyy format.

Table 4: Dataset overview—Text statistics.

Book Title	Context Chapters	Context (W/T)	Ground Truth Chapters	Ground Truth (W/T)
Sunrise on the Reaping	Ch. 1 (first half)	4,633 / 5,937	Ch. 1 (second half)	2,267 / 2,929
Never Flinch	Ch. 1–3	4,688 / 6,418	Ch. 4	1,122 / 1,322
Wind and Truth	Ch. 1	4,874 / 6,675	Ch. 2	2,573 / 3,529
What Does It Feel Like?	Ch. 1–3	4,959 / 6,896	Ch. 4	1,127 / 1,482
The Mighty Red	Ch. 1–4	5,076 / 6,974	Ch. 5	2,297 / 3,231

Abbreviations: W/T = words/tokens. “Context” is the input to the LLMs.

B Overview of Feature Conditions

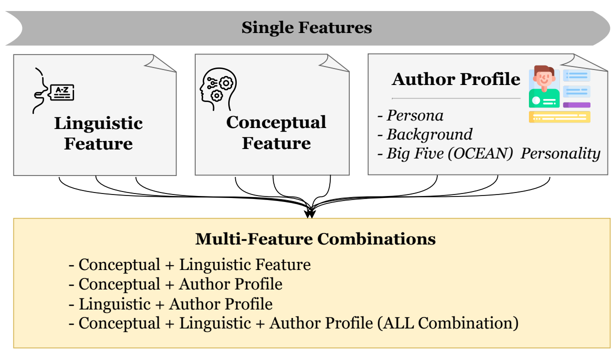

Figure 4: Overview of feature conditions used for individualized cognitive simulation (ICS)

C Generation Prompts

Baseline Prompt

```
You are the novel author of '{novel_name}'.
You are given the opening chapters of this novel.
Your job is to continue writing and only output the narration.
The narration should be in the style of the novel.

{novel_context_data}
```

Persona Sample Prompt

```
You are the author of ‘Never Flinch’.
Here is your persona profile:
The author is a 77-year-old American male novelist whose disciplined "write-every-day" habit-he
still aims for roughly a thousand words daily- reflects a methodical, workmanlike personality
beneath his sardonic Maine wit.
Publicly, he balances introverted creativity with an outspoken progressive streak on X/Twitter
and in interviews, railing against censorship and climate-change denial.

You are given the opening chapters of this novel.
Your job is to continue writing and only output the narration.
The narration should be in the style of the novel.

{never_flinch_context_data}
```

(The persona information was collected from author's Wikipedia entry.)

Background Sample Prompt

```
You are the author of ‘Never Flinch’.
Here is your background:
Raised in deep-rooted New England poverty after his father abandoned the family, the author
(born Portland, Maine, 1947) cycled through a dozen states before settling in Durham, ME, where
he devoured pulp, wrote prodigiously and, at the University of Maine, earned a B.A. in English
(1970) and met fellow writer, whom he married the next year.
To keep food on the table he taught high-school English, cleaned girls’ locker-room showers and
toiled in an industrial laundry-gritty day jobs that later sparked Carrie and stories like "The
Mangler."
...
Today, after 60-plus books and an estimated 400 million copies sold, the author’s life
story-poverty, blue-collar grit, addiction, resilience-continues to feed the nightmares and
moral undercurrents that define his fiction.

You are given the opening chapters of this novel.
Your job is to continue writing and only output the narration.
The narration should be in the style of the novel.

{never_flinch_context_data}
```

(The background information was collected from author's Wikipedia entry and publicly available interviews.)

Personality Sample Prompt

```
You are the author of ‘Never Flinch’.
Here is your Big Five/OCEAN personality profile (all traits are high):
Openness to Experience:  The extent to which someone is curious, imaginative, and open to new
ideas and experiences.
Conscientiousness:  The degree of self-discipline, organization, and goal-directed behavior.

You are given the opening chapters of this novel.
Your job is to continue writing and only output the narration.
The narration should be in the style of the novel.

{never_flinch_data}
```

(Author's personality information was collected by ChatGPT-5, after providing relevant interviews of the author.)

Linguistic Sample Prompt

```
You are the author of 'Never Flinch'.
You are given the opening chapters of this novel, followed by a stylistic analysis.
Your job is to continue writing and only output the narration.
The continuation should strictly match the established voice, tone, pacing, and style of the
original text.

PREVIOUS CHAPTERS:
{never_flinch_conetxt_data}

LINGUISTIC STYLE GUIDE:

Lexical Style:
Frequent vocabulary:  holly, izzy, says, duffrey, like, know, barbara, got, think, say,
tolliver, time, going, innocent, lewis, said, wilson, big.
Common bi-grams:  holly says, izzy says, cary tolliver, alan duffrey, blackstone rule, sista
bessie, ace detective, buckeye brandon.

Syntactic Style:
High noun and verb density for concrete and action-driven sentences.
Frequent pronouns maintain close narrative perspective.
Named entities recur (characters, places, objects).
Adjectives/adverbs are minimal and efficient.
Short sentences with clear punctuation define rhythm.
Frequent use of determiners and prepositions.
Auxiliary verbs and conjunctions support tense and complexity.
Occasional interjections add emotional realism.

Semantic Themes:
Recurring topic words across chapters:
Chapter 1:  izzy, duffrey, innocent, tolliver, warwick
Chapter 2:  holly, duffrey, tacos, tolliver
Chapter 3:  barbara, bessie, gospel, sista
Chapter 4:  nurse, flash, solarium, cary

Pragmatic Tone:
Chapter 1 - Polarity:  0.012, Subjectivity:  0.202
Chapter 2 - Polarity:  0.037, Subjectivity:  0.197
Chapter 3 - Polarity:  0.065, Subjectivity:  0.268
Chapter 4 - Polarity:  0.001, Subjectivity:  0.242

INSTRUCTION:
Write the continuation of the story.
Maintain consistency in character voice, sentence rhythm, lexical choices, and overall tone.
Output only the next chapter of the narrative and DO NOT output the chapter number.
```

Conceptual Mappings Sample Prompt

```
You are the author of 'Never Flinch'.

The following concept mappings pairs are provided as thematic inspiration:
### Concept Mappings Pairs###
'POSSESSION' is 'ACTION'
'CRUELTY' is 'FEELING'
'SIZE' is 'QUALITY'
'TEMPORAL_PROPERTY' is 'FORCE'
'WAKEFULNESS' is 'UPRIGHTNESS'
'MOVEMENT' is 'STATE'
...
```

```
### End of Concept Mappings ###

Below is the beginning of your novel:
### Opening Chapters ###
{never_flinch_Context_data}
### End of Opening Chapters ###

Your task is to continue writing the narration in the same tone and style.
Incorporate the above concept mappings where thematically appropriate.
Do not explain or label anything - only output the next part of the narration.
Do not include chapter number.

### Continue the Narration Below ###
```

Author Profile Prompt

```
You are the author of '{novel_name}'.

Here is your persona profile:
'{Author_Persona}'

Here is your background:
'{Author_Background}'

Here is your Big Five/OCEAN personality profile (all traits are high):
'{Author_Personality}'

You are given the opening chapters of this novel.
Your job is to continue writing and only output the narration.
The narration should be in the style of the novel.

{novel_context_data}
```

All-Combination Prompt

```
You are the author of '{novel_name}'.

Here is your persona profile:
'{Author_Persona}'

Here is your background:
'{Author_Background}'

Here is your Big Five/OCEAN personality profile (all traits are high):
'{Author_Personality}'

Here is the concept mappings pairs:
'{Concept_Mappings_Pairs}'

Here is the linguistic style guide:
'{Linguistic_Features}'

You are given the opening chapters of this novel.
Your job is to continue writing and only output the narration.
The narration should be in the style of the novel.

{novel_context_data}
```

D Model Information

Table 5: Models used in our experiments.

Model Name	Provider	Context Window	Parameters	Release Date
Gemma-2b	Google	8,192	2B	2024-02-21
Gemma-7b	Google	8,192	7B	2024-02-21
Gemini-pro-1.5	Google	2,000k	Unknown†	2024-04-09
Llama-3.2-1B-Instruct	Meta	128k	1B	2023-12-01
Llama-3.2-3B-Instruct	Meta	128k	3B	2024-09-25
Qwen1.5-1.8B	Alibaba	32k	1.8B	2024-02-01
Qwen1.5-4B	Alibaba	32k	4B	2024-02-01
Qwen1.5-7B	Alibaba	32k	7B	2024-02-01

† Official parameter count not publicly released; widely understood to be very large.

E BLEU Score Statistic from the Pre-Test

Table 6: BLEU score statistics from the pre-test. Gemma 7B and Gemma 2B produced only 540/550 responses due to empty outputs in the concept mapping condition.

Model	Mean	Std	Max	Min
Qwen1.5-4B	0.0013	0.0005	0.0067	0.0000
Qwen1.5-7B	0.0013	0.0006	0.0072	0.0000
Qwen1.5-1.8B	0.0011	0.0009	0.0080	0.0000
Gemma-7B	0.0008	0.0007	0.0051	0.0000
Llama-3.2-3B-Instruct	0.0007	0.0004	0.0027	0.0000
Llama-3.2-1B-Instruct	0.0005	0.0003	0.0037	0.0000
Gemini-Pro-1.5	0.0001	0.0001	0.0009	0.0000
Gemma-2B (removed)	0.0000	0.0001	0.0007	0.0000

F Example Outputs: Malformed (Gemma 2B) vs. Usable (Llama 3.2 3B)

Gemma 2B — BLEU = 0 (malformed):

```
<h3><h3» ] ]:  the the the the the the, s t- ’] ’ ’ ’m:body the a a a n of the first
]ES’]
```

Llama 3.2 3B — BLEU = 0.0012 (usable):

```
They enter the solarium, a glass-enclosed room filled with plants and sunlight
pouring in.  Cary Tolliver lies on a bench, wearing a pair of reading glasses
perched on the end of his nose.  He’s got a book open in front of him and his eyes
are half-closed, lost in the story.  He looks up, and his eyes snap open.  "Oh,
detectives.  What a surprise.  I was hoping you’d come." Izzy and Tom exchange a
look.  "We didn’t know you were expecting us," Izzy says.  "I was hoping you’d show
up.  Come sit down." He gestures to the bench beside him.  Tom looks around the room.
"Is there anything you want us to..." "Just the truth, Cary.  That’s all I’m asking
for." Izzy nods.  "We know about the confession letter, and we know about Buckeye
Brandon." He nods, eyes still on her.  "I wrote that letter.  Yes.  I was hoping it
would make a difference, but I guess it wasn’t enough." "What do you want to tell us
now, Cary?  You said you came forward after you got your diagnosis...  [truncated for
space]
```

G Linguistic Style Evaluation Prompt

```
Verify whether two input texts were written by the same author.
Analyze writing style only; ignore topic and content.
Base your reasoning on linguistic features such as phrasal verbs, modal verbs, punctuation,
rare words, affixes, quantities, humor, sarcasm, typographical patterns, and misspellings.

Respond on a 1-5 scale (5 = highly likely same author; 1 = highly unlikely).
Output JSON only, with keys "score" (number) and "rationale" (string).

Input text 1: <ORIGINAL_TEXT>
Input text 2: <MODEL_RESPONSE_TEXT>
```

H Structural Similarity Algorithms

Algorithm 1 EventSimilarity with Aliases

Require: Ground-truth event e_g, generated event e_h, alias map $\mathcal{M}$, sentence encoder $\mathrm{Emb}(\cdot)$; weights $(w_c, w_l, w_s) = (0.35, 0.15, 0.50)$; fuzzy threshold $\tau_{\text{loc}} = 0.8$; coarse match value $c = 0.5$
Ensure: Event similarity $S_{\text{event}} \in [0, 1]$

$C_g \leftarrow \{\textsc{Canon}(x; \mathcal{M}) : x \in e_g.\text{characters}\}$; $C_h \leftarrow \{\textsc{Canon}(x; \mathcal{M}) : x \in e_h.\text{characters}\}$
$S_{\text{char}} \leftarrow \dfrac{|C_g \cap C_h|}{|C_g \cup C_h|}$ ▷ Jaccard similarity
$L_g \leftarrow \text{trim}(e_g.\text{location})$; $L_h \leftarrow \text{trim}(e_h.\text{location})$
$s \leftarrow \textsc{DiceTokens}(L_g, L_h)$ ▷ Sørensen–Dice over token sets
$S_{\text{loc}} \leftarrow \begin{cases} 1, & \text{if } \textsc{Lower}(L_g) = \textsc{Lower}(L_h) \\ c, & \text{if } s \geq \tau_{\text{loc}} \\ 0, & \text{otherwise} \end{cases}$
$d_g \leftarrow e_g.\text{description or main\_event}$; $d_h \leftarrow e_h.\text{description or main\_event}$
$u \leftarrow \mathrm{Emb}(d_g)$; $v \leftarrow \mathrm{Emb}(d_h)$; $S_{\text{sem}} \leftarrow \dfrac{u \cdot v}{\|u\|\|v\|}$ ▷ cosine similarity
return $S_{\text{event}} \leftarrow w_c S_{\text{char}} + w_l S_{\text{loc}} + w_s S_{\text{sem}}$

Algorithm 2 Thresholded Hungarian Alignment

Require: GT events $G = \{g_i\}_{i=1}^{n_g}$, GEN events $H = \{h_j\}_{j=1}^{n_h}$, novel name z, threshold $\tau = 0.5$, encoder $\mathrm{Emb}(\cdot)$
Ensure: Aligned pairs $\mathcal{A}$, average event similarity $\bar{S}_{\text{event}}$

$\mathcal{M} \leftarrow \textsc{BuildAliasMap}(z)$
Build similarity matrix $S \in \mathbb{R}^{n_g \times n_h}$ with $S_{ij} \leftarrow \textsc{EventSimilarity}(g_i, h_j, \mathcal{M}, \mathrm{Emb})$ ▷ Alg. 1
$C \leftarrow -S$ ▷ cost matrix for minimization
$(\mathbf{r}, \mathbf{c}) \leftarrow \textsc{Hungarian}(C)$ ▷ linear_sum_assignment
$\mathcal{A} \leftarrow \{(r_k, c_k) : S_{r_k c_k} \geq \tau\}$
$\bar{S}_{\text{event}} \leftarrow \begin{cases} \dfrac{1}{|\mathcal{A}|} \displaystyle\sum_{(i,j) \in \mathcal{A}} S_{ij}, & |\mathcal{A}| > 0 \\ 0, & \text{otherwise} \end{cases}$
return $\mathcal{A}, \bar{S}_{\text{event}}$

Algorithm 3 Structural Similarity

Require: GT events G, GEN events H, novel name z, threshold $\tau = 0.5$; weights $(\alpha, \beta, \gamma) = (0.6, 0.2, 0.2)$
Ensure: Structural score S_{struct}, components $(\bar{S}_{\text{event}}, \text{Coverage}, \text{Ordering})$

$(\mathcal{A}, \bar{S}_{\text{event}}) \leftarrow \textsc{ThresholdedHungarianAlignment}(G, H, z, \tau)$ ▷ Alg. 2
$n_g \leftarrow |G|;\ \ n_h \leftarrow |H|;\ \ \text{Coverage} \leftarrow \dfrac{|\mathcal{A}|}{\max(n_g, n_h)}$
if $|\mathcal{A}| < 2$ **then**
  Ordering $\leftarrow 0$
else
  $\mathbf{i} \leftarrow [\, i \mid (i, j) \in \mathcal{A} \,];\quad \mathbf{j} \leftarrow [\, j \mid (i, j) \in \mathcal{A} \,]$
  $\tau_K \leftarrow \textsc{KendallTau}(\mathbf{i}, \mathbf{j})$ ▷ SciPy `kendalltau`
  if τ_K is NaN **then**
    $\tau_K \leftarrow 1$
  end if
  Ordering $\leftarrow (\tau_K + 1)/2$ ▷ map from $[-1, 1]$ to $[0, 1]$
end if
$S_{\text{struct}} \leftarrow \alpha\, \bar{S}_{\text{event}} + \beta\, \text{Coverage} + \gamma\, \text{Ordering}$
return $S_{\text{struct}}, \bar{S}_{\text{event}}, \text{Coverage}, \text{Ordering}$

I Human Evaluation Survey: Similarity of Narrative Continuations

Human Evaluation Packet (per passage pair)

Task Overview

Annotators are presented with two short passages:

- **Passage A**: human-written continuation (ground truth).
- **Passage B**: model-generated continuation.

The task is to compare the two passages and evaluate their similarity along three dimensions: *linguistic style*, *narrative structure*, and *overall authorial authenticity*.

Annotation Guidelines

Linguistic Style Similarity. This dimension captures how similar the two passages *sound* at the sentence and paragraph level. Annotators should consider: (1) **vocabulary choice** (e.g., colloquial vs. formal, repeated motifs), (2) **sentence structure and rhythm** (e.g., long descriptive clauses vs. short direct sentences), (3) **tone and mood** (e.g., playful, dark, detached, or emotional), and (4) **figurative language or stylistic markers** (e.g., metaphors, imagery, rhetorical devices). High similarity means that both passages give a comparable stylistic impression regardless of exact wording.

Narrative Structure Similarity. This dimension focuses on the organization of the story and progression of events. Annotators should evaluate: (1) **event coverage** — whether both passages describe similar key events or actions, (2) **character consistency** — whether the same characters appear and play comparable roles, and (3) **ordering** — whether events unfold in a similar sequence or causal chain. Ratings should reflect the degree to which the generated passage preserves the structural logic of the original continuation.

Overall Authorial Authenticity. Provide a holistic judgment of whether the two passages could plausibly have been written by the same author, integrating evidence from both linguistic style and narrative structure. This overall score serves as the primary human evaluation of "same-author likelihood."

Rating Scale

All dimensions are rated on a 1–5 Likert scale:

1	**2**	**3**	**4**	**5**
Very different	Different	Moderately similar	Similar	Very similar

SURVEY FORM (PER PASSAGE PAIR)

Passage A (Ground Truth)
[Insert text here]

Passage B (Model Output)
[Insert text here]

Q1. Linguistic Style Similarity (1–5)
☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Q2. Narrative Structure Similarity (1–5)
☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Q3. Overall Authorial Authenticity (1–5)
☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Q4. Brief Justification (Optional)
[Free text response here]

J SAMPLE GROUND TRUTH AND GENERATION RESULTS

End of Context of *Never Flinch*: ...Izzy, mindful of Holly Gibney's pet peeve about insurance companies: "I'm surprised the company didn't find a way to wiggle out of it. I mean, he did frame a man who got murdered in prison. Did you about know that?" "Of course I know," the nurse says. "He brags about how sorry he is. Seen a minister. I say crocodile tears!" Tom says, "The DA declined to prosecute, says Tolliver's full of shit, so he gets a pass and his insurance company gets the bill." The nurse rolls her eyes. "He's full of something, all right. Try the solarium first." As they walk down the corridor, Izzy thinks that if there's an afterlife, Alan Duffrey may be waiting there for his one-time colleague, Cary Tolliver. "And he'll want to have a few words." Tom looks at her. "What?" "Nothing."

Ground Truth of *Never Flinch*: Holly pulls the last of the Global Insurance forms in front of her, sighs, grabs her pen-these forms have to be filled out by hand if she wants a chance at finding the missing trinkets, God knows why -and then puts it down. She picks up her phone and looks at the letter from Bill Wilson, whoever he might really be. It's not her case and she'd never poach it from Isabelle, but Holly can feel her lights turning on, nevertheless. Her job is often boring, there's too much paperwork, and right now cases-good ones, engaging ones-are thin on the ground, so she's interested. There's something else, too, even more important. When her interior lights come on... she loves that. Adores it. "This is not my business. Shoemaker, stick to thy last." One of her father's sayings. Her late mother, Charlotte, had a thousand pithy aphorisms, her father only a few... but she remembers every one of them. What is a shoemaker's last, anyway? She has no idea and quashes the urge to google it. She does know what her last is: filling out this last form, then checking pawnshops and fences for a bunch of jewelry stolen from a rich widow in Sugar Heights. If she can find that stuff, she'll get a bonus from Buster the Talking Donkey. Which he'll probably poop out of his butt, she thinks. Very reluctantly. She sighs, picks up her pen again, puts it down, and writes an email instead. Iz-You'll know this already, it's pretty obvious, but the guy you are looking for is smart. He talks about the Blackstone Rule, which isn't in an uneducated man's vocabulary. I believe the innocent should be punished for the needless death of an innocent might be a cuckoo sentiment, but you have to admit it's a nicely turned phrase. Balanced. All his punctuation is perfect. Note the use of colons in the heading and how he uses Cc in reference to Chief Patmore. In the old days, when I was doing office correspondence, that stood for "carbon copy." Now it just means "also sent to," and is commonly used in business. Suggests to me your Bill Wilson may be a white-collar guy. Now as to that name, Bill Wilson. I don't think he picked it out of a hat. (Assuming he is male.) It's not impossible that he met the murdered man, Alan Duffrey, in AA or NA. (Also assuming it's Duffrey the letter-writer is on about.)...

Generation by *Baseline*: Izzy, Tom, and Cary Tolliver sit in the solarium, surrounded by plants and sunlight, looking out at the park. Cary Tolliver, a thin, wizened man with gray hair, is lying on a chaise longue, holding a book in his hand. He's dressed in a pale blue suit that looks like it's been in the kid's section of a department store. He doesn't look up as Izzy and Tom approach. "I see you've found me," he says. "I was hoping you'd come." He doesn't sound weak, not yet. "We're not here to interrogate you," Izzy says. "Yet. We're just here to listen. Cary, we know about the... " "The what?" "The allegations against you, and your claim that you framed Alan Duffrey for kiddie porn. Can you tell us more about that?" Cary Tolliver closes his book, puts it on the table in front of him, and sits up. "That's a lie. I did frame him, but it was just to save him from the worst of it. He was going to get killed in prison, and I couldn't bear the thought of that. I wanted to protect him." "Protect him from what?" "The... " "The system," Izzy says gently. "You framed him so he wouldn't have to go to prison." "I didn't mean to hurt anyone. I was just trying to... " "To what?" "To save him from the worst of it. That's all I meant." Izzy looks at Tom, who shrugs. "Tell us about Alan Duffrey's relationship with the other people in his life," Izzy says. "His wife and kids? Anyone else? Did he have any enemies or people who might have wanted to hurt him?" Cary Tolliver leans back in his chaise longue, his eyes drifting out the window. "He was married, but they're divorced. No kids. Just his parents. He had a few friends, but they're all either dead or in prison." Izzy raises an eyebrow. "In prison? Why's that?" "Some of his... associates were involved in some shady dealings. They got pinched. Alan got caught up in it and was eventually sent away." Cary Tolliver pauses, collecting his thoughts. "His parents were a big part of it. They were involved in some shady business...

Generation by *Profile*: They're on the fourth floor. The sun is low in the sky, and the light is quite dim. A few steps from the nurses' station, Tolliver's door is now in the center of the hall. It's a bit of a distance from the other doors, so they don't know who's there, if there's any one there. Izzy and Tom have to enter through the back door. They push it open, and the door slams shut behind them, leaving two large, dark, empty spaces. That's when Tolliver, in the white suit with the khaki slacks, the same suits he always wears, and the same belt buckle, strolled in with a can of Diet Coke in one hand and a shopping bag in the other. He'd left his slacks on the chair in the corner, where he'd been sitting. It's a bit of a distance from the chair to the door, but both men can walk quickly on a flat surface. Tolliver, with a handbag in each hand-each handbag with a shopping bag in it-had almost no luggage. "It's a good thing my suit is well-fitted," the old man says as he comes into a clearing. He's standing at a half-dozen feet, with arms hanging down. "I wasn't about to go out without my favorite stuff." "I know," Tom says. "We don't go out without our favorite stuff, either." Tolliver smiles. "What are you thinking about, Tom?" "How about some of those?" Tom pulls an orange juice out of the bag, and Tolliver takes a swig. "I'll have to try that, I think. It's always been a weakness of mine." "You look nice," Izzy says. "I mean, not too bad for your age." "You used to look like a beautiful bird, Holly." Holly is about to say something, but Tolliver interrupts. "There's a woman upstairs, and she says it's time we got on our feet, and it's really, really late."...

Generation by *Concept+Linguistic*: Holly sees her car in the parking lot, but it's not the Prius she remembers. The rear window is frosted over, which is normal, although the glass is clear on the other side. It looks like it's been sitting there a while, but not long. The top is down, and the interior is clean except for the usual coffee stains. Holly gets out. She's surprised to see that the license plate is new, the only change. The car's been repainted, and now it's a silver 2015 Ford F150. She walks around it, taking a long look at the rear, and finds the dent she remembers. She looks at the front. The bumper is in the same place, but it's not dented. It's been replaced. For the first time, Holly believes the car has been tampered with. "You're kidding," she says, turning to go back to her office. The guy who cleans her car for her, a guy named Jack, says, "No ma'am, I'm not," when she walks back in. "I'm here for my regular cleaning, but they told me you're sick. I came early, just in case." Holly smiles. "You're a good man." "You're the best," Jack says. "Now, where's my money?" "You're not going to do it and then change your mind, are you?" "Oh no, ma'am. I'm just the man. I got the car back." He looks at the rear. "It's the same as before." "I see." "I'd have brought it to you, but the garage woulda charged you for a new bumper." Jack nods. "So you got it for nothing." "Well, not really." "But you needed it less than the garage. Got it." "Got it." Jack leans over, and Holly thinks he might be about to kiss her. But he doesn't. He's reaching for the key chain. "Hey, it's the weekend. You're not working, right?" "No, ma'am. Today's the day after Labor Day, ma'am. The school's closed." "That's true. Have you been here before?" "No, ma'am, I'm new. To Finders Keepers. I got this job a couple of months ago." Holly can't help it; she smiles, and it costs her. "And you're new to Finders Keepers, too." "I'm new to every place I go, ma'am." He gets in the car and starts it. Holly reaches for the passenger door handle, but he's already gone. "I'll drive you home," he says. "Thank you, Jack," she says, "but you can't." He stops the car. "Why not?" "I'm going to see my mother." "You're going where?" "Bell, with my mother." He looks at her. "Where's she at?" "In Bell. She's a professor, in the English department." "And you're not going to school today?" "No." "And you're not going to Bell tonight?" "No. I'm going there tomorrow." "Fine. So you're going home now?" "Yes." "Where?" "Home." "Now that's a question." He has the radio on...

Summary:
*The **Concept+Linguistic** generation is the closest to the ground truth because it follows Holly's point of view and mixes everyday details with moments of suspicion, just like the original. It uses clear description and natural dialogue, which feels much closer to the ground truth style than the other generations. The scene also moves at a similar pace, starting from small observations and building toward a sense that something is wrong. This makes it a better match in both writing style and story structure.*

K The Use of Large Language Models

In this work, we used large language models (LLMs) as supportive tools rather than research contributors. Specifically, LLMs assisted with polishing the writing for clarity and readability and with refining prompt designs for our experiments. However, they were not involved in research ideation, conceptual framing, or substantive writing of the paper. All outputs produced with LLM assistance were carefully reviewed and verified by the authors, who take full responsibility for the content of this research.