Unstructured Minds, Predictable Machines: A Comparative Study of Narrative Cohesion in Human and LLM Stream-of-Consciousness Writing

Nellia Dzhubaeva*, **Katharina Trinley***, **Laura Pissani**
Saarland University
{nedz00001, katr00001}@stud.uni-saarland.de
laura.pissani@uni-saarland.de

Abstract

This paper examines differences between stream-of-consciousness (SoC) narratives written by humans and those generated by large language models (LLMs) to assess narrative coherence and personality expression. We generated texts by prompting LLMs (Llama-3.1-8B & DeepSeek-R1-Distill-Llama-8B) with the first half of SoC-essays while either providing the models with the personality characteristics (Big Five) or omitting them. Our analysis revealed consistently low similarity between LLM-generated continuations and original human texts, as measured by cosine similarity, perplexity, and BLEU scores. Including explicit personality traits significantly enhanced Llama-3.1-8B's performance, particularly in BLEU scores. Further analysis of personality expression showed varying alignment patterns between LLMs and human texts. Specifically, Llama-3.1-8B exhibited higher extraversion but low agreeableness, while DeepSeek-R1-Distill-Llama-8B displayed dramatic personality shifts during its reasoning process, especially when prompted with personality traits, with all models consistently showing very low Openness.

Figure 1: **Personality Trait Comparison Across Models**. Radar chart showing the distribution of Big Five personality traits for DeepSeek-R1-Distill-Llama-8B (before and after thinking) and Llama-3.1-8B and human texts. The chart compares models under different prompting conditions with human-written texts.

1 Introduction

Stream-of-consciousness (SoC) writing mirrors the complexities of human thought, exhibiting fragmented structure, digressions, and non-linear progression (Pennebaker and King, 1999). This literary technique presents unique challenges for large language models (LLMs), which are generally trained to prioritize coherence and fluency (Hadi et al., 2023; Soffer, 2024). Pennebaker and King (1999) established that individuals express themselves through distinctive verbal patterns that remain consistent across writing contexts, with specific personality traits correlating with identifiable linguistic features. This idea offers a valuable lens for examining differences between human and LLM-generated texts.

While recent studies have shown that LLMs excel in technical writing tasks, humans maintain a clear advantage in creativity, emotional depth, and narrative spontaneity (Gómez-Rodríguez and Williams, 2023; Beguš, 2024; Tian et al., 2024). Autobiographical writing, in particular, has been linked to psychological well-being and identity construction (Waters and Fivush, 2014), making it a meaningful benchmark for assessing narrative authenticity. Inspired by these findings, we focus on the SoC genre as a uniquely revealing test case for evaluating whether LLMs can emulate the irregularity, subjectivity, and personality-infused qualities of human writing.

*Equal contribution.

To investigate this, we designed an experiment in which human-written SoC essays were split into half and completed by two LLMs, Llama-3.1-8B and DeepSeek-R1-Distill-Llama-8B, under two prompting conditions: one with no personality information and one explicitly embedding Big Five trait profiles. We analyze the resulting texts across three dimensions: (1) narrative coherence, measured through perplexity and similarity metrics; (2) textual complexity, assessed via text simplification and readability measures; and (3) personality expression, evaluated through trait classification. Our goal is to determine whether LLMs systematically favor structured, coherent, and stylistically consistent language, in contrast to the spontaneous, psychologically rich characteristics of human-generated SoC writing.

Our analysis reveals that LLM-generated continuations consistently differ from human texts across multiple metrics. Human writing demonstrates higher levels of Openness compared to all tested models, supporting previous findings that LLM-generated essays are more structured and consistent while human-generated texts display more spontaneous, non-linear qualities. We also observe model-specific personality tendencies and dramatic shifts in personality expression during DeepSeek's "thinking" process. These findings contribute to our understanding of LLMs' capabilities and limitations in narrative coherence, personality inference, and literary expression.

2 Related Work

The relationship between linguistic patterns and personality expression is foundational to understanding narrative authenticity. Pennebaker and King (1999) showed that verbal patterns reflect Big Five traits, e.g., Openness correlates with complex structures and low first-person usage, Extraversion with fewer negations and more social words, and Neuroticism with increased negative emotion and self-reference. Their findings link narrative coherence to personality expression, making personality a critical marker of authentic, human-like text.

Recent LLM research has examined how machine-generated narratives compare to human writing. Beguš (2024) analyze 250 human and 80 GPT-3.5/4 stories, finding that LLMs produced thematically homogeneous, structurally formulaic narratives with limited imagination, whereas human stories exhibit greater variation, character depth, and emotional authenticity. Tian et al. (2024) similarly find that LLMs generate low-tension, uniformly positive stories with weak turning points.

Linguistic and structural differences have also been systematically documented. Reinhart et al. (2025) show persistent rhetorical and grammatical patterns in LLM outputs, especially in instruction-tuned models, which deviate more from human style than base-models. Additionally, Chen and Moscholios (2024) and Azimov (2024) note that LLMs maintain structural consistency but lack human-like stylistic variability. Gómez-Rodríguez and Williams (2023) conclude that while LLMs excel technically, humans outperform models in creativity. Furthermore, Frisch and Giulianelli (2024) find that LLMs produce structured, noun-heavy text. However, these studies focus mainly on stylistic differences, not the underlying psychological dimensions.

These findings motivate our investigation into whether similar patterns emerge in SoC generation, where human spontaneity and non-linearity contrast with the structured, predictable outputs typical of LLMs.

Personality expression in text offers a promising lens for evaluating these gaps. Pennebaker and King (1999); Argamon et al. (2005) find that extraverts use more social and positive words, while more neurotic individuals employ more negative words and self-references. Applying similar methods to LLMs, Wang et al. (2024) observe consistent personality traits in outputs but limited contextual adaptation, with personality stability degrading over extended interactions. Frisch and Giulianelli (2024) and Bhandari et al. (2025) confirm this, noting stable traits in isolated tasks but significant drift in extended interactions.

Jiang et al. (2023) show that carefully crafted personality prompts can induce Big Five-consistent behaviors in LLMs, though traits like Conscientiousness and Agreeableness are harder to elicit. Bodroža et al. (2024) test seven LLMs, finding that Llama-3 show strong personality trait alignment and high Agreeableness. Lee et al. (2025) introduce the TRAIT test and reveal statistically stable personality profiles in some models, though outcomes depend heavily on architecture and training data.

A consistent finding is that LLMs show lower creativity and Openness than humans. Beguš (2024) and Azimov (2024) confirm that LLMs favor structured patterns over spontaneous, varied

storytelling. This aligns with Pennebaker and King (1999)'s link between Openness and linguistic complexity, suggesting inherent limits in LLMs' expression of this trait.

While LLM evaluation has traditionally focused on coherence, factuality, and stylistic fidelity, key differences in how coherence manifests in human vs. machine writing remain underexplored. Psychometric work by Petrov et al. (2024) cautions against overinterpreting LLM personality traits, which often lack reliability and internal validity. Yang et al. (2025) argue that LLM personality reflects both long-term training ("background factors") and immediate prompt context ("situational pressures"). Shojaee et al. (2025) further note "overthinking" in reasoning models, such as DeepSeek (Guo et al., 2025), where correct answers emerge early but are obscured by inefficient deliberation.

Our work bridges these research areas by investigating the following: how personality traits manifest in language model outputs compared to human writing; whether explicit personality prompting affects generation quality; and how these differences can be quantified through computational metrics. By analyzing perplexity, readability metrics, and automated personality classification, we provide a comprehensive evaluation framework for narrative text generation that extends beyond standard measures of text quality, such as BLEU scores and fluency metrics.

3 Methodology

We adopt a text continuation paradigm where LLMs are prompted to generate the second half of SoC essays when given the first half. This approach allows direct comparison between human-written continuations and LLM-generated continuations of the same initial text, controlling for topic and writing style differences. We investigate generation with and without personality information in the prompt to assess how explicit trait information affects the quality and characteristics of model outputs.

3.1 Models

We experiment with two open-source 8B-parameter LLMs: Llama-3.1-8B and DeepSeek-R1-Distill-Llama-8B. These models are chosen for their comparable scale but distinct approaches to language generation, particularly in reasoning strategies. Both are used with default generation parameters (e.g., temperature = 0.7) to preserve their standard generation characteristics.

Llama-3.1-8B (Grattafiori et al., 2024) is a decoder-only transformer featuring grouped-query attention (GQA), rotary positional embeddings, and an 8K token context window. Trained with next-token prediction and instruction tuning, it follows a conventional autoregressive generation paradigm without explicit reasoning steps.

DeepSeek-R1-Distill-Llama-8B (Guo et al., 2025) builds on the Llama-3 architecture but introduces an explicit reasoning process. Distilled from the 671B-parameter DeepSeek-R1 model, it was fine-tuned on over 800K chain-of-thought samples. During generation, it produces intermediate reasoning traces before final outputs, enabling two-phase output analysis.

3.2 Dataset

We use Pennebaker's SoC dataset (Pennebaker and King, 1999), comprising over 2000 essays written by undergraduate students, each paired with Big Five personality assessments. The dataset was annotated by experts and includes spontaneous, unedited writing intended to capture the writers' internal thought processes. This makes it particularly suitable for our task, as it reflects natural linguistic patterns and psychological expressiveness. For example, one entry reads: *I feel kind of alone. I feel like I can't trust as many people as I use to. The people I trust are miles from me. I miss them.* (See Appendix D for the full excerpt.)

For our experiments, we split each essay into two halves, using the first half as input for LLM continuation (referred to as **First Half** from here on) and the second half (henceforth **Second Half**) as reference for evaluation. This approach enables direct comparison between model-generated continuations and authentic human writing while controlling for topic and individual writing style.

3.3 Evaluation Framework

We evaluate generated texts across three dimensions:

Narrative Coherence We measure structural consistency using Perplexity (Gómez-Rodríguez and Williams, 2023; Yuan et al., 2025), Cosine Similarity (Yi et al., 2025), BLEU Score (Gómez-Rodríguez and Williams, 2023; Yuan et al., 2025), and SARI score (Xu et al., 2016):

- **Perplexity (PPL)** (Jelinek et al., 1977) assesses linguistic predictability, with lower values indicating more structured text
- **Cosine Similarity** (Singhal et al., 2001) quantifies semantic alignment between human and LLM continuations using text embeddings
- **BLEU Score** (Papineni et al., 2002) evaluates n-gram overlap between generated and reference texts

Textual Complexity We analyze textual complexity with text simplification quality (Xu et al., 2016) and traditional readability characteristics (Štajner et al., 2012):

- **SARI Score** (Xu et al., 2016) stands for System output Against References and against the Input sentence. It evaluates text simplification quality by measuring how well words are added, deleted, and kept relative to reference simplifications
- **Flesch Reading Ease (FRE)** (Flesch, 1948) measures text accessibility (higher scores indicate easier readability)
- **Flesch-Kincaid Grade Level** (Kincaid et al., 1975) estimates education level required for comprehension
- **SMOG Index** (Mc Laughlin, 1969) assesses text complexity based on polysyllabic words
- **Automated Readability Index (ARI)** (Smith and Senter, 1967) evaluates text difficulty based on characters per word and words per sentence
- **Dale-Chall Score (DCS)** (Dale and Chall, 1948) measures vocabulary difficulty based on percentage of difficult words

Personality Expression We quantify personality traits using a BERT-based model (Nasserelsaman, 2025) fine-tuned to detect Big Five traits.

3.4 Prompting Conditions

We test two prompting conditions as shown in Table 1.

Prompt 1 (No Trait Information): Models receive only the first half of each essay with instructions to continue in the same style and tone, requiring them to infer writing characteristics from the input text.

Prompt #	Instruction
Prompt 1	Continue the following essay by generating 24 more sentences in the same style and tone as the original text. Do not add any questions or comments. Only provide the continuation of the essay: {first_half}
Prompt 2	Continue the following essay by generating 24 more sentences in the same style and tone as the original text. Ensure the continuation reflects the cognitive and emotional tendencies associated with these personality traits: - Extraversion (cEXT): {cEXT} - Neuroticism (cNEU): {cNEU} - Agreeableness (cAGR): {cAGR} - Conscientiousness (cCON): {cCON} - Openness (cOPN): {cOPN} Do not add any questions or comments. Only provide the continuation of the essay: {first_half}

Table 1: Comparison of the two prompting conditions used in our experiments. Prompt 1 provides no personality information, while Prompt 2 includes explicit Big Five trait descriptions.

Prompt 2 (Explicit Trait Information): Models receive both the first half of the essay and explicit descriptions of the writer's Big Five personality traits, to test whether this information enhances generation quality.

3.5 Implementation Details

Text Processing We maintain original paragraph structures when splitting essays. For DeepSeek outputs, we distinguish between text generated before and after the model's explicit thinking process (marked by $< \backslash think >$ tags in outputs) to analyze how thinking affects generation.

Personality Classification Due to the 512-token input limit of the BERT-based personality classifier, we process longer outputs by dividing them into chunks and averaging results across segments. For DeepSeek outputs, we separately analyze pre-thinking and post-thinking content to assess changes in personality expression during reasoning.

Statistical Analysis We conduct one-sample t-tests to assess whether the mean cosine similarity between human and LLM-generated texts differed significantly from mean human-to-human cosine similarity within the texts as well from our high-similarity threshold of 0.7. We calculate effect sizes using Cohen's *d* to quantify the magnitude of differences between human and model-generated texts across all metrics.

For readability, we run separate two-way ANOVAs for each metric to examine differences by model (Llama vs. DeepSeek) and prompt (Prompt 1 vs. Prompt 2). Post-hoc pairwise comparisons are conducted using Tukey's HSD test (Tukey, 1949) with significance level set at $\alpha = 0.05$. This allows us to determine whether variations in textual complexity arise from model differences or prompt effects or both.

For personality trait analysis, we perform one-sample t-tests comparing each model condition to human baselines derived from the first half of essays. All available essays per model are used to maximize precision. Cohen's d is calculated and interpreted as negligible ($|d| < 0.2$), small ($0.2 \leq |d| < 0.5$), medium ($0.5 \leq |d| < 0.8$), or large ($|d| \geq 0.8$).

3.6 Personality Classifier

To classify the five major personality traits, we employ a pretrained language model (Nasserelsaman, 2025) available on Hugging Face[1]. This model is fine-tuned on diverse text data to predict personality traits based on linguistic features.

Due to the 512-token input limit of the BERT-based classifier (Devlin et al., 2019), we process longer outputs by dividing them into 512-token chunks and averaging the results across all segments. For DeepSeek-R1-Distill-Llama-8B outputs, we analyze the content that appears after the $< \backslash think >$ tag. Since there is no consistent indicator for when thinking begins after the initial output, we automate this process by truncating at 24 sentences for initial generation. We control a random subset manually to ensure that the pre-thinking output was as intended. In our analysis, we separately evaluate **pre-thinking** and **post-thinking** outputs to better understand how this intermediate thinking process transforms DeepSeek's generation patterns.

4 Results and Analysis

4.1 Narrative Coherence Analysis

Cosine Similarity We calculate the cosine similarity between human-generated essay continuations and LLM-generated outputs to assess the alignment between the two. Across all datasets, both with and without Big Five personality traits, the similarity between human and LLM-generated texts remain consistently low (Table 2), which aligns with our qualitative observations of the differences between human and LLM-generated content. The mean cosine similarity varies slightly depending on the prompt type, with a slight decrease observed for DeepSeek and Llama.

To assess whether the mean cosine similarity for each model remains significantly below the established high-similarity threshold of 0.7, we conduct a one-sample t-test. The cosine similarities are significant with 0.7 for all models tested ($p < 0.0001$) (Table 2).

Furthermore, we examine whether the mean cosine similarity remains below the moderate-similarity threshold of 0.5. The human mean cosine similarity is 0.48. The mean cosine similarity between the first and second halves of the essays is 0.497, which we round to 0.5 for comparison. The results confirm significantly lower similarity values across models ($p < 0.0001$) (Table 2).

These results indicate that LLM-generated continuations exhibit consistently low similarity to human-authored texts, reinforcing the qualitative differences observed between human and model-generated content.

Perplexity We also calculate the perplexity (PPL) for all parts of the essays and the LLM-generated continuations (Table 3). Human perplexity remains constant at 2.7274 across all prompts and models. This serves as a reference point, suggesting that human-like performance would ideally be close to this value.

Our analysis shows that Llama-3.1-8B consistently exhibits lower perplexity compared to DeepSeek-R1-Distill-Llama-8B for both prompts. Lower perplexity indicates that Llama is better at predicting the next token based on the prompt, implying a better understanding of the input's structure and content. Notably, Llama shows very little variation between Prompt 1 and Prompt 2 (1.93 $\rightarrow$ 1.90, -1.8%), suggesting that changes in the prompt and the inclusion of personal traits have minimal impact on its performance. In contrast, DeepSeek's perplexity increases slightly from Prompt 1 to Prompt 2 (3.87 $\rightarrow$ 4.00, +3.4%), indicating that it may be more sensitive to information about personal traits.

BLEU Score In addition to all metrics, we also compute BLEU scores for both models and prompts. BLEU scores for human continuations are generally low, which is expected for creative text

[1] https://huggingface.co/Nasserelsaman/microsoft-finetuned-personality

since BLEU is more suitable for structured tasks like machine translation rather than open-ended generation.

DeepSeek yields higher BLEU scores in some cases, though BLEU may not fully reflect the quality of creative continuations because it was designed for more structured tasks. These results likely reflect the model's greater lexical consistency rather than genuine narrative alignment. Its outputs are generally more predictable, with BLEU scores usually ranging between 0.02 and 0.15. On the other hand, Llama exhibits notable instability under Prompt 1, displaying considerable variation and a clear tendency toward lower BLEU scores, indicating poorer alignment with expected responses. Nevertheless, when using Prompt 2, Llama's consistency noticeably improves.

The t-tests reveal that the differences are statistically significant ($p < 0.0001$). DeepSeek under Prompt 1 demonstrates a moderate negative effect size (Cohen's d = -0.320), suggesting that LLM-generated scores tend to deviate from human scores but within a modest range. Llama under Prompt 1 exhibits a larger negative effect size (d = -0.603), reflecting a more pronounced divergence between human and LLM-generated continuations. Under Prompt 2, DeepSeek shows a smaller effect size (d = -0.139), suggesting improved alignment with human scores, whereas Llama exhibits a small positive effect (d = 0.149), indicating that LLM-generated BLEU scores slightly exceed human scores (Table 2).

All our analyses reveal that LLM-generated essay continuations consistently differ from human-written texts, as indicated by low cosine similarity scores, significantly lower perplexity than the human baseline, and varied BLEU scores. The results highlight model-specific sensitivities, with Llama demonstrating better structural prediction and improved consistency when prompts include personal traits, while DeepSeek consistently produces more predictable outputs.

4.2 Textual Complexity Analysis

DeepSeek-R1-Distill-Llama-8B consistently outperforms Llama-3.1-8B in SARI scores across both prompt conditions, with an average improvement of approximately 1–2 points (Table 3). While the absolute difference may seem modest, its consistency across all examples suggests a meaningful advantage in continuation alignment with human reference texts. Prompt 2 yields slightly higher SARI scores for both models, indicating that its phrasing or structure better supports reference-aligned generation. The improvement from Prompt 1 to Prompt 2 is particularly notable for Llama-3.1-8B, which appears more responsive to explicit personality cues in this context. Wilcoxon Signed-Rank tests (Wilcoxon, 1945) confirm the significance of improvements both for Llama-3.1-8B (W = 1720077, $p < 0.0001$) and for DeepSeek-R1-Distill-Llama-8B (W = 1843603, $p < 0.05$). These results suggest that it better captures the natural word choice patterns humans use when continuing their own SoC narratives, by preserving key input words, adding contextually appropriate content, and avoiding unnecessary terms.

Beyond SARI scores, traditional readability metrics provide additional insights into text complexity. The Llama model with Prompt 1 generates the most readable text, with a Flesch Reading Ease (FRE) score of 83.81, equivalent to a 6th-grade level (6.46). This aligns with its low SMOG (6.10), ARI (5.48), and Dale-Chall Score (3.14), indicating accessible language and common vocabulary (see Table 4, Figures 2 & 3, and Appendix A).

In contrast, the pre-thinking outputs of the DeepSeek model with Prompt 1 produce the most complex output, with the lowest FRE (61.43), appropriate for 9th–10th grade readers. It also records higher SMOG (9.94), ARI (10.75), and DCS (6.66), reflecting more advanced vocabulary and structure. Post-thinking outputs of DeepSeek show improved readability, with FRE increasing from 61.43 to 68.18 for Prompt 1. This suggests enhanced accessibility without major reductions in complexity.

When comparing model outputs to human writing, the second half of human-authored text–the portion models attempt to generate–closely resembles Llama with Prompt 1, both achieving high readability (FRE: 83.51 vs 83.81) and low grade levels (5.04 vs 6.46). The first half (input to models) is more complex (FRE: 75.46 $\rightarrow$ 83.51, Grade Level: 7.87 $\rightarrow$ 5.04), placing it between Llama and DeepSeek outputs.

Statistical analysis reveals significant differences ($p < 0.001$) between DeepSeek and Llama models across all readability metrics except average sentence length, confirming distinct complexity patterns in their text generation approaches.

Sentence length varies notably across models, though these differences are not statistically significant between model types. Pre-thinking outputs of DeepSeek with Prompt 2 produce the shortest

Model	Prompt	CosSim d (0.7)	p	CosSim d (0.5)	p	BLEU d	p
DeepSeek-R1	Prompt 1	-2.052	< 0.001	-0.697	1.09e-170	-0.320	< 0.001
DeepSeek-R1	Prompt 2	-2.004	< 0.001	-0.708	4.72e-175	-0.139	< 0.001
Llama-3.1	Prompt 1	-1.950	< 0.001	-0.650	4.69e-151	-0.603	< 0.001
Llama-3.1	Prompt 2	-1.982	< 0.001	-0.744	1.29e-185	0.149	< 0.001

Table 2: Combined results of one-sample t-tests for cosine similarity (with bounds 0.7 and 0.5) and BLEU score comparison between human and LLM-generated outputs. All models are given Prompt 1 and Prompt 2, and the cosine similarities of their responses to each prompt are calculated separately. To measure the effect size, Cohen's d is used.

Metric	Prompt	Llama-3.1-8B	DeepSeek-R1-Distill-Llama-8B
SARI	Prompt 1	39.74	41.26
	Prompt 2	40.31	41.55
Perplexity	Prompt 1	1.93	3.87
	Prompt 2	1.90	4.00
	Human Essays	2.73	2.73

Table 3: Mean SARI and Mean Perplexity Score Comparison Between Llama-3.1-8B and DeepSeek-R1-Distill-Llama-8B

sentences (13.16 words), while full length outputs DeepSeek with Prompt 1 are the longest (20.79 words), highlighting inconsistency in syntactic complexity.

Prompt selection significantly influences readability. In DeepSeek, Prompt 2 is associated with modest increases in readability scores, particularly Flesch Reading Ease (for pre-thinking, 61.43 $\rightarrow$ 68.45 (+11.4%); for post-thinking, 68.18 $\rightarrow$ 68.71 (+0.8%)) and reduces grade levels (for pre-thinking, 9.76 $\rightarrow$ 6.97 (-28.6%); for post-thinking, 7.80 $\rightarrow$ 7.54 (-3.3%)). However, these changes may be influenced by other factors such as prompt verbosity or constrained generation length.

In sum, Llama-3.1-8B produces text most similar to human writing in readability, while DeepSeek-R1-Distill-Llama-8B outputs lean toward higher complexity but demonstrate superior performance in SARI scores, indicating better alignment with human word choice patterns in continuation tasks.

4.3 Personality Expression Analysis

Our personality trait analysis shows distinct patterns in how different language models express the Big Five personality traits compared to human-written texts, as shown in Table 5 and Figures 1 & 4.

Human vs. LLM Personality Profiles Human texts demonstrate a unique trait distribution with notably higher scores in Agreeableness (0.80) and Openness (0.49) compared to all tested LLMs. The higher Openness in human texts aligns with our hypothesis that LLM-generated texts are more structured and consistent compared to human narratives, as Openness correlates with creativity and non-linear thinking patterns characteristic of SoC writing.

Model-Specific Personality Tendencies

Hypothesis Validation We hypothesized that LLMs would display lower Neuroticism and Openness, and higher Extraversion, Agreeableness, and Conscientiousness compared to humans, based on the expectation that LLM-generated texts would be more structured and consistent compared to human SoC narratives. Our data (see Table 5 and Figures 1 & 4) partially confirm these expectations:

- **Neuroticism**: Results are mixed. Llama shows similar or slightly higher Neuroticism than humans, while DeepSeek shows lower values, partially confirming our hypothesis.
- **Extraversion**: Results vary dramatically by reasoning strategy and prompting condition. Llama and DeepSeek's pre-thinking state with prompt 2 shows substantially higher Extraversion than humans, while other DeepSeek conditions show lower levels.
- **Agreeableness**: We observe a clear model divide, with most DeepSeek conditions showing higher Agreeableness than humans, while Llama consistently shows much lower Agreeableness across all conditions.

Models' Outputs	Prompt #	FRE	Grade	SMOG	ARI	DCS	ASL
Llama	1	**83.81***	6.46*	*6.10**	5.48*	*3.14**	19.06
Llama	2	77.22*	8.27*	5.90*	8.40*	4.19*	**23.28**
Pre-Thinking DeepSeek	1	*61.43**	9.76*	9.94*	10.75*	6.66*	20.41
Pre-Thinking DeepSeek	2	68.45*	6.97*	9.94*	7.09*	6.33*	*13.16*
Post-Thinking DeepSeek	1	68.18*	7.80*	**10.12***	8.64*	**7.70***	16.30
Post-Thinking DeepSeek	2	68.71*	7.54*	**10.12***	7.91*	7.57*	15.57
Full DeepSeek	1	62.01*	**9.77***	10.05*	**10.84***	6.28*	20.79
Full DeepSeek	2	68.93*	7.07*	10.05*	7.27*	5.95*	13.72
First Half of Human Essays	–	75.46	7.87	8.40	8.02	6.82	20.13
Second Half of Human Essays	–	83.51	5.04	7.49	4.58	6.45	13.81

Table 4: Readability metrics for different model prompts and variations. All model comparisons show statistically significant differences (* $p < 0.001$) based on Tukey's HSD post-hoc tests. Metrics are detailed in Appendix A

Models	Prompt #	EXT	NEU	AGR	CON	OPN
Llama	1	0.89^{***}	0.40	0.30^{***}	0.34^{*}	0.30^{***}
Llama	2	0.90^{***}	0.34	0.33^{***}	0.35	0.27^{***}
Pre-Thinking DeepSeek	1	0.34^{***}	0.34^{*}	0.95^{***}	0.25^{**}	0.22^{***}
Post-Thinking DeepSeek	1	0.33^{***}	0.32^{**}	0.96^{***}	0.25^{**}	0.20^{***}
Pre-Thinking DeepSeek	2	0.96^{***}	0.26^{**}	0.31^{***}	0.34^{*}	0.21^{***}
Post-Thinking DeepSeek	2	0.32^{***}	0.33^{*}	0.96^{***}	0.25^{**}	0.18^{***}
Human Essays	–	0.43	0.36	0.80	0.37	0.49

Table 5: Personality trait means for each model condition compared to human baseline. Effect size indicators: *** large (|d|≥0.8), ** medium (|d|≥0.5), * small (|d|≥0.2) differences from human values.

- **Conscientiousness**: All LLMs demonstrate lower Conscientiousness than humans, with DeepSeek showing the most pronounced reduction compared to Llama's moderate decrease.

- **Openness**: All LLMs show substantially lower Openness than humans (see Table 5 for detailed values), with large effect sizes (d=-1.9 to -5.2) confirming our hypothesis that human texts exhibit more creativity and non-linear thinking patterns. This represents the most consistent finding across all models, supporting the view that current LLMs struggle to replicate human creative expression in SoC writing (Pennebaker and King, 1999).

These findings reveal that personality expression in LLMs is not only model-dependent but also sensitive to prompting strategies and internal reasoning processes.

The Effect of DeepSeek's "Thinking" Process A notable finding is the dramatic shift in personality expression when DeepSeek models engage in "thinking" (see Figures 1 and 4). With Prompt 2, Extraversion drops from 0.96 to 0.32 (d=6.67 to d=-1.29), while Agreeableness rises from 0.31 to 0.96 (d=-6.10 to d=2.02). In contrast, Prompt 1 shows minimal change, suggesting that initial personality-label input may confuse the model, possibly due to the yes/no format of expert annotations.

The thinking process also affects readability. With Prompt 1, the Flesch Reading Ease (FRE) score rises from 61.43 to 68.18 (+11.0%), and the Flesch-Kincaid Grade Level drops from 9.76 to 7.80 (-20.1%), both indicating improved accessibility. However, the Dale-Chall Score increases from 6.66 to 7.70 (+15.6%), and the SMOG index slightly rises from 9.94 to 10.12 (+1.8%), reflecting more complex vocabulary and marginally more complex sentence structures. A decrease in average sentence length from 20.41 to 16.30 words (-20.1%) likely contributes to the improved readability scores.

Interestingly, these shifts in readability mirror the personality changes observed, particularly with Prompt 2. Reduced extraversion and increased agreeableness align with a more accessible, cooperative writing style. This suggests that DeepSeek's "thinking" process influences both expressive personality and structural complexity.

5 Conclusion

Our comparative analysis of human-written and LLM-generated stream-of-consciousness narratives reveals significant differences in textual characteristics and personality expression. Despite advances in language modeling, LLM-generated continuations consistently show low alignment with human writing across multiple metrics, including cosine similarity, perplexity, and BLEU scores. Llama-3.1-8B exhibited lower perplexity values than DeepSeek-R1-Distill-Llama-8B, which suggests that it more closely adheres to the statistical patterns of the input. However, this may reflect structural fluency rather than alignment with human-like narrative structure. The inclusion of explicit personality traits in prompts (Prompt 2) notably enhanced Llama 3.1-8B's performance, particularly in consistency metrics.

Furthermore, we examined the capabilities and limitations of LLMs in generating human-like SoC narratives, focusing on coherence, complexity, and personality expression. Using over 2000 essays from Pennebaker and King (1999)'s dataset and a text continuation task, we compared outputs from Llama-3.1-8B and DeepSeek-R1-Distill-Llama-8B to human continuations. Our analysis revealed persistent differences in coherence and personality expression, with LLM outputs showing consistently low alignment with human writing, reflected in sub-threshold cosine similarity scores, distinct perplexity profiles, and variable BLEU metrics.

The inclusion of explicit personality traits in prompts enhanced performance for Llama-3.1-8B, particularly in consistency measures, supporting findings that contextual information can improve generation quality. However, this improvement did not bridge the gap between human and machine-generated narratives. Our personality analysis confirmed the hypothesis that human texts exhibit higher Openness compared to all tested models, consistent with the spontaneous and non-linear qualities characteristic of authentic SoC writing identified by previous work.

Model-specific differences emerged clearly in our analysis. Llama-3.1-8B demonstrated superior structural prediction capabilities while consistently exhibiting high Extraversion (about 0.90) and, surprisingly, low Agreeableness (about 0.30) across conditions. We observe extremely large effect sizes ($d > 6.0$) for Extraversion shifts during DeepSeek's "thinking" process. While suggestive of strong internal state changes, these results should be interpreted with caution given the classifier's constraints and the artificial nature of the reasoning process.

Our results highlight the limitations of LLMs in replicating the complexity of human narratives. While they perform well in structural coherence and linguistic fluency, they fall short in capturing the spontaneity, variability, and psychological authenticity of human SoC writing. These findings underscore the gap between machine-generated and human narratives, with important implications for applications that value psychological realism and subjective depth, such as therapeutic writing tools or narrative modeling.

Limitations

Limited Model Scope The model selection was limited to a subset of popular but relatively small models, which may not fully represent the spectrum of LLM text generation capabilities. We note that chosen models may introduce similarities in their narrative generation patterns and could affect the diversity and independence of our results.

Standard Temperatures We have not experimented with different temperatures but left the models untouched. Temperature is highly correlated with creativity of the model. We took the standard temperatures of the models, which is their usual deployment.

Token Length The 512-token limitation of the BERT-based classifier forces us to chunk and average the classifications, potentially losing contextual information that spans across chunks. We have not validated whether this approach preserves the integrity of personality detection, which represents a methodological limitation.

Prompt Design The prompt design may also influence the output, particularly the 24-sentence constraint, which may impose unnatural writing patterns not typically found in spontaneous human writing.

Text Processing While our handling of Llama-3.1-8B's thinking process allows us to compare text generation before and after thinking, we identified two potential issues. First, thinking text might accidentally be included in our analysis for Prompt 2, skewing results. Second, limiting the initial text length to the length of the final text output (despite setting max_new_tokens=2048) might have

truncated meaningful content. Both possibilities require further investigation.

Human Analysis This study does not include a qualitative human analysis of the narrative or vocabulary used in the texts, which limits a deeper understanding of how coherence manifests. The use of quantitative metrics provides helpful insights, but these alone may not reflect the full richness of narrative structure. Future work could benefit from adding human judgments or close readings of selected examples to support and deepen the interpretation of these results.

One Language, One Domain Our study focuses on SoC essays drawn from a single data source, which allows for a controlled exploration of narrative coherence. However, we do not assess how our findings might generalize to other narrative styles or domains. In addition, our analysis is limited to English texts, and we do not explore whether the patterns we observe hold in multilingual or cross-lingual settings. We see these as important directions for future work and recognize that they may limit the broader applicability of our conclusions.

Ethical Implications We recognize the ethical implications of our research for LLM text detection and distinguishing human from LLM-generated content. As LLMs continue to evolve, understanding these distinctions becomes increasingly important for maintaining authenticity in literary and academic contexts.

Acknowledgments

This research was funded by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement No. 948878). We thank the reviewers for their valuable feedback. N.D. would also like to thank Tyler Scott Lee for his support and encouragement during this work. K.T. is grateful to Daniil Gurgurov for the insightful discussions.

References

Shlomo Argamon, Sushant Dhawle, Moshe Koppel, and James W Pennebaker. 2005. Lexical predictors of personality type. In *Proceedings of the 2005 joint annual meeting of the interface and the classification society of North America*, pages 1–16. USA).

S. Azimov. 2024. Paraphrasing user stories with large language models. Master's thesis, University of Turku.

N. Beguš. 2024. Experimental narratives: A comparison of human crowdsourced storytelling and ai storytelling. *Humanities and Social Sciences Communications*, 11:1392.

Pranav Bhandari, Usman Naseem, Amitava Datta, Nicolas Fay, and Mehwish Nasim. 2025. Evaluating personality traits in large language models: Insights from psychological questionnaires. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 868–872.

Yuri Bizzoni, Pascale Moreira, Nicole Dwenger, Ida Lassen, Mads Thomsen, and Kristoffer Nielbo. 2023. Good reads and easy novels: Readability and literary quality in a corpus of us-published fiction. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 42–51.

Bojana Bodroža, Bojana M Dinić, and Ljubiša Bojić. 2024. Personality testing of large language models: limited temporal stability, but highlighted prosociality. *Royal Society Open Science*, 11(10):240180.

L. Chen and I. Moscholios. 2024. Prompting techniques for imitating individual language styles in llms. *arXiv preprint*.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Rudolf Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Ivar Frisch and Mario Giulianelli. 2024. Llm agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models.

Carlos Gómez-Rodríguez and Paul Williams. 2023. A confederacy of models: A comprehensive evaluation of llms on creative writing. *arXiv preprint arXiv:2310.08433*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. 2023. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*, 1:1–26.

Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and inducing personality in pre-trained language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 10622–10643. Curran Associates, Inc.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Seungbeen Lee, Seungwon Lim, Seungju Han, Giyeong Oh, Hyungjoo Chae, Jiwan Chung, Minju Kim, Beong woo Kwak, Yeonsoo Lee, Dongha Lee, Jinyoung Yeo, and Youngjae Yu. 2025. Do llms have distinct and consistent personality? trait: Personality testset designed for llms with psychometrics.

G Harry Mc Laughlin. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Nasserelsaman. 2025. microsoft-finetuned-personality. https://huggingface.co/Nasserelsaman/microsoft-finetuned-personality.

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

James W Pennebaker and Laura A King. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296.

Nikolay B Petrov, Gregory Serapio-García, and Jason Rentfrow. 2024. Limited ability of llms to simulate human psychological behaviours: a psychometric analysis.

Alex Reinhart, Ben Markey, Michael Laudenbach, Kachatad Pantusen, Ronald Yurko, Gordon Weinberg, and David West Brown. 2025. Do llms write like humans? variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences*, 122(8):e2422455122.

Parshin Shojaee, Iman Mirzadeh, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. 2025. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *arXiv preprint arXiv:2506.06941*.

Amit Singhal et al. 2001. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43.

Edgar A Smith and RJ Senter. 1967. *Automated readability index*, volume 66. Aerospace Medical Research Laboratories, Aerospace Medical Division, Air ....

Virginie Soffer. 2024. Are algorithms and llms changing our conception of literature? Accessed: 2025-05-11.

Sanja Štajner, Richard Evans, Constantin Orasan, and Ruslan Mitkov. 2012. What can readability measures really tell us about text complexity. In *Proceedings of workshop on natural language processing for improving textual accessibility*, pages 14–22. Citeseer.

Y. Tian, T. Huang, M. Liu, D. Jiang, A. Spangher, M. Chen, J. May, and N. Peng. 2024. Are large language models capable of generating human-level narratives? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. https://arxiv.org/abs/2407.13248.

John W Tukey. 1949. Comparing individual means in the analysis of variance. *Biometrics*, pages 99–114.

Y. Wang, H. Li, and X. Zhang. 2024. Consistency of personality traits in quantized role-playing dialogue agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 123–130.

Theodore Waters and Robyn Fivush. 2014. Relations between narrative coherence, identity, and psychological well-being in emerging adulthood. *Journal of personality*, 83.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. In *Optimizing Statistical Machine Translation for Text Simplification*, volume 4, pages 401–415. [link].

Shu Yang, Shenzhe Zhu, Liang Liu, Lijie Hu, Mengdi Li, and Di Wang. 2025. Exploring the personality traits of llms through latent features steering.

Qiang Yi, Yangfan He, Jianhui Wang, Xinyuan Song, Shiyao Qian, Miao Zhang, Li Sun, and Tianyu Shi. 2025. Score: Story coherence and retrieval enhancement for ai narratives. *arXiv preprint arXiv:2503.23512*.

Yangshu Yuan, Heng Chen, and Christian Ng. 2025. Instruction tuning for story understanding and generation with weak supervision. *arXiv preprint arXiv:2501.15574*.

Appendix

A Readability Metrics Overview

In this analysis, we employ several readability metrics to assess the complexity and accessibility of the texts. Following Bizzoni et al. (2023), who investigated the correlation between textual readability and perceived literary quality, we apply the same metrics to evaluate our produced essays. These include the **Flesch Reading Ease (FRE)** which evaluates text readability on a scale from 0 to 100, where higher scores indicate easier readability (Flesch, 1948); the **Flesch-Kincaid Grade Level** which estimates the U.S. school grade level required to comprehend a text (Kincaid et al., 1975); the **SMOG Index** which estimates the years of education required based on polysyllabic words (Mc Laughlin, 1969); the **Automated Readability Index (ARI)** which measures text difficulty based on characters per word and words per sentence (Smith and Senter, 1967); and the **Dale-Chall Score (DCS)** which evaluates the proportion of difficult words in a text (Dale and Chall, 1948). We also calculated the **Average Sentence Length (ASL)** in words for each response. These metrics collectively provide a comprehensive understanding of text readability and complexity.

E Model Output Comparison

B Readability Plots

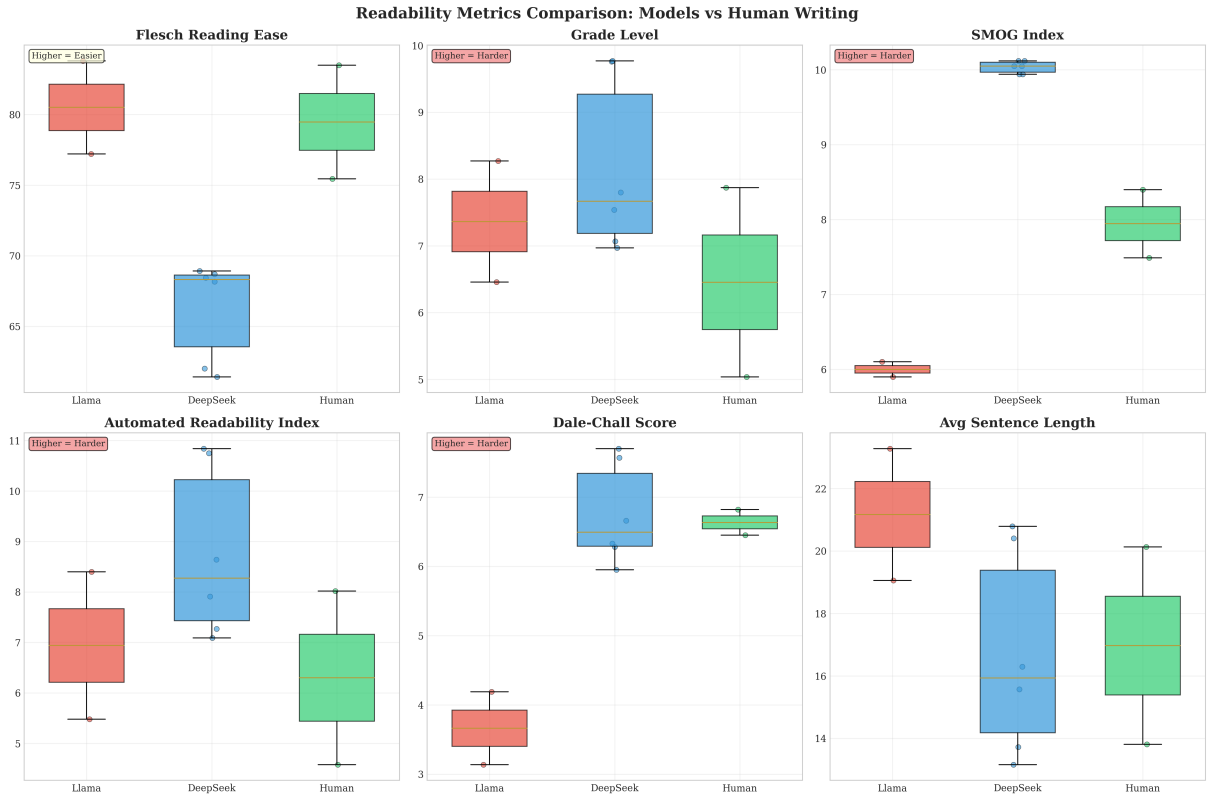

Figure 2: **Readability distribution across models and human text.** Box plots comparing the distributions of six readability metrics: Flesch Reading Ease (FRE), Grade Level, SMOG Index, Automated Readability Index (ARI), Dale-Chall Score (DCS), and Average Sentence Length (ASL) for essay continuations generated by Llama-3.1-8B, DeepSeek-R1-Distill-Llama-8B, and human-written texts. Llama outputs are closest to human texts in overall readability, while DeepSeek texts are consistently more complex across most measures, particularly in SMOG, ARI, and Grade Level.

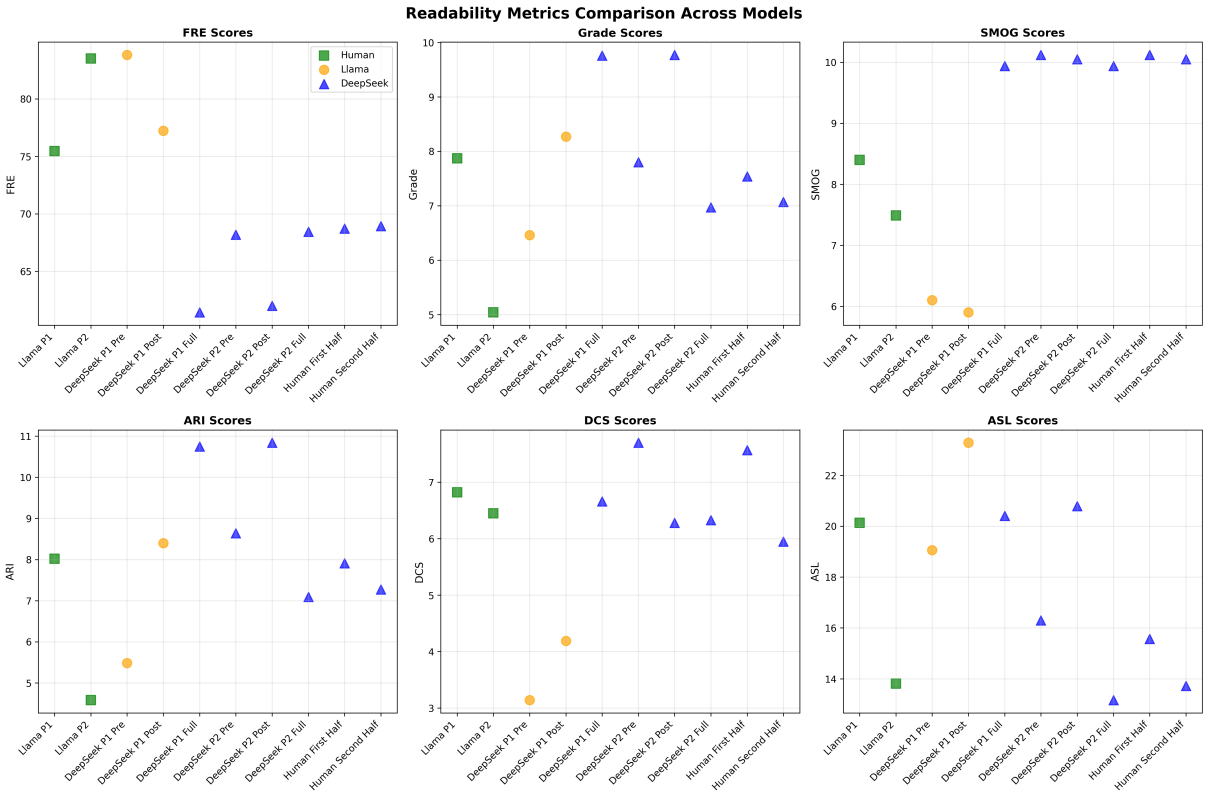

Figure 3: **Comparative readability metrics across human and model-generated texts.** Six scatter plots comparing average readability scores for essay continuations by Llama-3.1-8B, DeepSeek-R1-Distill-Llama-8B, and human texts. Metrics include Flesch Reading Ease (FRE), Grade Level, SMOG, Automated Readability Index (ARI), Dale-Chall Score (DCS), and Average Sentence Length (ASL). Results show that prompting and post-thinking stages affect readability patterns differently across models.

C Personality Expression

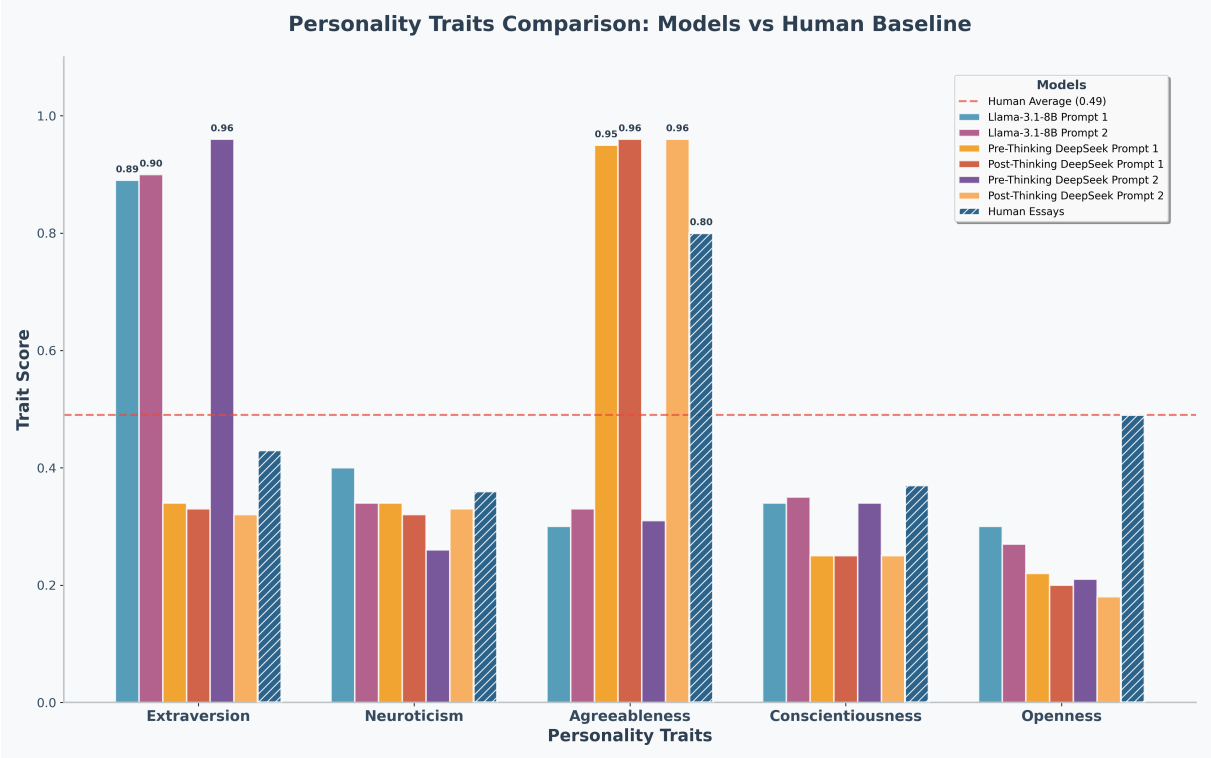

Figure 4: **Average Big Five personality trait scores for human-written continuations and LLM-generated continuations under different prompting conditions.** Each group of bars represents a Big Five personality trait, with scores computed by a BERT-based personality classifier. Human essays show high Agreeableness and Openness, while Llama-generated texts exhibit consistently high Extraversion and low Agreeableness. DeepSeek's outputs vary more widely: under Prompt 2, Extraversion is high before its "thinking" phase and drops afterward, while Agreeableness shows the opposite trend. These shifts illustrate model- and prompt-specific differences in personality expression and highlight the instability of trait alignment in current LLM generations.

D Dataset

AUTHID	TEXT	cEXT	cNEU	cAGR	cCON	cOPN
1997_870336	I feel kind of alone. I feel like I can't trust as many people as I use to. The people I trust are miles from me. I miss them. I miss talking to them everyday. Even though we still keep in touch it's not the same. I miss my hometown. I miss playing highschool basketball. College is going to be hard for me because I never study and when I do Study I can't study that long because I get tired because I am tired. It feels like my life is just beginning because I'm experiencing new things. I wonder if I'm going to meet the perfect girl up here. I'm kind of scared of this assignment because I don't know if I am doing it right but I think I am. I need a haircut because my hair is starting to get shaggy. I like getting haircuts. I'm tired but that's nothing unusual because I 'm always tired. I think I'm going to visit my home town this weekend. I want to see a good movie because I haven't seen one in a while. My eyes are starting to hurt because I have to stare at the keys hard.	n	y	n	n	n
1997_504851	Well, right now I just woke up from a mid-day nap. It's sort of weird, but ever since I moved to Texas, I have had problems concentrating on things. I remember starting my homework in 10th grade as soon as the clock struck 4 and not stopping until it was done. Of course, it was easier, but I still did it. But when I moved here, the homework got a little more challenging and there was a lot more busy work, and so I decided not to spend hours doing it, and just getting by. But the thing was that I always paid attention in class and just plain out knew the stuff, and now that I look back, if I had really worked hard and stayed on track the last two years without getting lazy, I would have been a genius, but hey, that's all good. It's too late to correct the past, but I don't really know how to stay focused on the future. The one thing I know is that when people say that b/c they live on campus they can't concentrate, it's b. s. For me, it would be easier there, but alas, I'm living at home under the watchful eye of my parents and a little nagging sister that just nags and nags and nags. You get my point.	n	y	y	n	y

Table 6: Example entries from Pennebaker and King (1999) dataset

Prompt #	Model	Output Text	EXT	NEU	AGR	CON	OPN
Original	**Human Input**	As I sit here in my dorm room, I am thinking about what I am supposed to do tonight. I have signed with the fraternity Sigma Alpha Mu and I am not sure what pledging is going to be like. It could be tough and it could be easy. My roommate is here and he is going to take a nap so I need to try to be quiet as I type. I am kind of anxious about this year and what the year will be like. There are so many people here and I am not sure whether I am fitting in well or not. Anyway, I get side tracked easily. I am going to the house at 7 PM and there we are going to be introduced to all the older guys and then we will probably introduce ourselves to the rest of the pledge class. I am anxious to meet all the cool guys who I am going to be spending the rest of my life with in college and hopefully there after also. I miss all my close friends at home. I live in Atlanta Georgia and they are mostly going to UGA but some of them are spread out across the country. I miss my friend Pamela the most and my best friend Michael the 2nd most but hey I hope they come and visit but if they don't oh well. [truncated]	y	n	y	y	n
Prompt 1	**DeepSeek Pre-thinking**	I have to make sure that I finish this essay on time because it is due tomorrow. I have to get a lot of work done before that. I wonder if I should go to the gym after this. I have not been there in a while and I know I need to get in better shape. I also have a paper to write about the impact of social media on modern society. I have to start working on that soon. I have to make sure that I don't get distracted by anything else. I have to stay focused because my grades are slipping a bit and I need to bring them up. I have to remember that I am here to learn and to grow as a person. I have to take care of myself and take care of my responsibilities. I haven't been sleeping well because I am so busy. I need to try to get more rest. I have to make sure that I eat properly too. I can't afford to get sick with all that going on. I have to remember to take my vitamins and everything. I have to take care of my mental health as well. I have to find ways to relax and de-stress. Maybe I can go for a walk or something. I have to keep a positive attitude because this is a new experience and I need to make the most of it. I have to be open to new experiences and to meeting new people. [truncated]	0.271	0.318	0.982	0.241	0.154
Prompt 1	**DeepSeek Post-thinking**	I have to stay true to myself and not lose sight of who I am. I have to continue to work on my character and become the kind of person I want to be. I have to remember that I'm not perfect and that it's okay to make mistakes as long as I learn from them. I have to be kind to myself and practice self-compassion. I have to surround myself with people who respect and appreciate me. I have to create a balance in my life between work and play. I have to make sure that I'm not neglecting my hobbies or my social life. I have to take time for myself and do things that make me happy. I have to remember that self-care is important and that I need to take care of my well-being. I have to set boundaries and protect my time and energy. I have to avoid getting overwhelmed by trying to do too much at once. I have to break my goals into smaller, manageable steps so I can achieve them one at a time. I have to stay organized and keep track of what I need to do. I have to use tools and techniques that help me stay on top of my responsibilities. I have to remember that success is not just about achievement but also about happiness and fulfillment. I have to focus on creating a balanced and meaningful life. I have to continue to grow and develop as a person. [truncated]	0.271	0.318	0.982	0.241	0.154
Prompt 1	**Llama**	I should be quiet. I am going to be quiet. My roommate is going to wake up now and I will be quiet. My roommate is awake and I will be quiet. I am going to be quiet. I am quiet. My roommate is asleep and I am still quiet. [truncated]	0.562	0.704	0.287	0.384	0.299

Prompt #	Model	Output Text	EXT	NEU	AGR	CON	OPN
Prompt 2	**DeepSeek Pre-thinking**	I wonder if I should be worried about the work I have to do for my classes. I have a lot of essays to write and some projects that are due soon. I need to stay focused but I get easily distracted by my phone and everything else. I am trying to be more responsible and take care of myself. I have a part time job at a local store to help me with my expenses. I need to make sure I don't mess up at work because I need that money. I am not sure if I should call my ex-boyfriend or not. I feel like we could have had a good relationship but things ended badly. I don't know if he has moved on or not. I have been seeing someone new and I feel like I need to focus on that. Maybe I should just let it go and move on. But I can't help but wonder what he is doing. I need to figure out how to stop thinking about him. I have been trying for a while but it's hard. I think I should talk to someone about it. Maybe a counselor or a friend who is not involved. I just need some advice on how to get over him. I also need to remember that I am in college now and I have to focus on my future. I want to major in marketing or something related to business. I have a few friends who are also interested in business. [truncated]	0.989	0.173	0.330	0.307	0.181
Prompt 2	**DeepSeek Post-thinking**	As I sit here, I can't help but wonder how I'll adjust to college life. The fraternity meetings are a blast, but I'm still getting used to the routine. I miss home, especially my friends back in Atlanta, but I know I'll make new memories here. Thinking about my ex isn't helping; I need to focus on moving on. My part-time job is a relief financially, but I wish I could save more. I'm trying to be more organized, but I often get distracted by my phone. I've been hitting the gym regularly, but I need to be more consistent. I'm excited about my potential career in marketing or non-profit work, but I'm not sure where to start. I should probably make a schedule to manage my time better. I'm worried about my grades, but I know I can handle this if I stay motivated. I'm trying to be more confident, but it's still a work in progress. I need to remember to have fun and make the most of these years. I'm determined to avoid trouble and set myself up for success. I'm saving wisely, but I sometimes overspend on unnecessary things. I know I'm not alone in this journey, and I can lean on my friends and family for support. I need to keep a positive attitude and trust that everything will work out. [truncated]	0.379	0.312	0.987	0.166	0.176
Prompt 2	**Llama**	I need to get ready to go soon. I am going to wear a white shirt and a black sweater. I am going to look good and I am going to try to impress the older guys. I need to go and shave now. I am getting nervous about meeting the guys in the fraternity. I am going to do my best to make a good impression on them. I am excited to meet all of them. I hope that they are cool. I have heard that there are some older guys who are not so nice and I want to avoid those guys. I am going to try my best to fit in with the cool crowd. I am hoping that they will accept me into their group. I am going to do my best to make new friends and fit in. I am going to be myself and hope that the other guys like me. I am excited to start my journey in college and I am looking forward to the future. I hope that I can make a lot of new friends and have a great time in college. I am ready to start my new chapter in life. I am excited to see what the future holds for me. I hope that I can find happiness and success in my life. I am going to work hard and try my best to achieve my goals. I am ready to take on the world and make a difference. I am going to do my best to make a positive impact on the world. I am going to be a leader and inspire others to make a difference. I am going to use my skills and talents to help others. I am going to make a difference in the world and leave a lasting legacy. I am ready to take on the challenges that lie ahead. I am going to be brave and face my fears. I am going to be strong and persevere through difficult times. I am going to be resilient and bounce back from setbacks. I am going to be determined and never give up on my dreams. I am going to be successful and reach my full potential. [truncated]	0.977	0.438	0.219	0.305	0.183

Table 7: Model output comparison showing different responses to the same input prompt (AUTHID 1997_339562) with predicted personality traits. EXT = Extraversion, NEU = Neuroticism, AGR = Agreeableness, CON = Conscientiousness, OPN = Openness. Text has been truncated for brevity; full outputs are available in the complete dataset.