

The Unreasonable Effectiveness of Large Language Models as Psychologically Deep Authors

Anonymous ACL submission

Abstract

Many evaluations of Large Language Models (LLMs) focus on exam-style benchmarks that measure domain-specific knowledge acquisition or linguistic attributes like grammaticality. Such evaluations emphasize the functional capacities of LLMs while overlooking their ability to resonate with readers on a psychologically deep level. Addressing this gap, this work introduces the Psychological Depth Scale (PDS), a novel framework designed to measure authenticity, empathy, engagement, narrative complexity, and emotional provocation. Through an empirical study involving 100 short stories written by humans and various LLMs, including GPT-4, we explore the consistency of human judgment on psychological depth, compare the depth of human-written and LLM-generated stories, and examine the potential for automated assessment of psychological depth. Our findings reveal that (1) humans can consistently judge psychological depth in spite of its abstract nature; (2) despite being perceived as less "human", GPT-4 stories surpassed advanced human authors in 4 out of 5 dimensions of psychological depth, often by sizable margins; and (3) GPT-4 combined with a novel Mixture-of-Personas (MoP) prompting strategy can moderately correlate (0.44) with human judgments of psychological depth. These findings open the possibility that LLMs could be strategically deployed to forge deeper emotional and psychological bonds with humans in fields as diverse as therapy and popular entertainment.

1 Introduction

In the rapidly evolving landscape of LLM research, the line between content generated by humans and machines has become increasingly blurred. This convergence raises critical questions about how to most effectively evaluate the quality of LLM generations. Many recent efforts have prioritized exam-style benchmarks measuring domain-

specific knowledge acquisition (Hendrycks et al., 2020; Lin et al., 2021), functional capabilities like common sense reasoning (Zellers et al., 2019; Yadav et al., 2019) and coding (Chen et al., 2021), as well as more granular linguistic attributes like grammaticality (Warstadt et al., 2019) and diversity (Gehrmann et al., 2021). While these metrics are and will remain indispensable, we stress that they do not adequately measure how creative generations impact readers on a psychologically deep level. The few works that do attempt to measure psychological dimensions like empathy (Concanon and Tomalin, 2023; Ma et al., 2020) and engagement (Ghazarian et al., 2020; Xu et al., 2022a) are either ad hoc or unsystematic.

Recognizing this gap, our study pioneers the concept of measuring the *psychological depth* within stories crafted by both humans and machines. We introduce the Psychological Depth Scale (PDS) to encapsulate several key components that contribute to the perceived excellence and human-likeness of narratives. These components include authenticity, empathy, engagement, narrative complexity, and the ability to evoke strong emotions in readers.

To rigorously assess the concept of psychological depth as a vital and coherent criterion for the evaluation of creative text, we established a comprehensive benchmark consisting of 495 short stories. This collection includes 45 narratives authored by humans alongside 450 stories generated by five well-known LLMs — namely, Llama-2-7B, Llama-2-13B, Llama-2-70B (Touvron et al., 2023), Vicuna-33B (Chiang et al., 2023), and GPT-4 (OpenAI et al., 2023) — each story averaging 450 words in length. To evaluate these texts, we enlisted the expertise of five undergraduate students from both the psychology and English departments at a public university in the US. These students were tasked with analyzing a subset of 100 stories, focusing on their psychological depth. They provided depth ratings, made predictions regarding whether each

story was authored by a human or a machine, and offered extensive free-form justifications for their evaluations. Through this meticulously designed study, we sought to explore three critical research questions, aiming to deepen our understanding of the intersection between LLMs and the human-like psychological intricacies within creative writing.

RQ1. *How consistently can well-informed humans judge psychological depth?* The Psychological Depth Scale achieved an average Krippendorff’s alpha of 0.72, indicating a significant level of agreement among raters and affirming its validity as a reliable instrument for assessing fictional short stories.

RQ2. *Are human stories more likely than LLM stories to possess psychological depth?* Despite being perceived as less "human", GPT-4 surpasses advanced human writers and other LLMs in 4 out of 5 dimensions of psychological depth: authenticity, empathy, engagement, and narrative complexity. This constitutes a notable progression in the capacity of some LLMs to emulate a depth in storytelling that closely resembles human-like intricacy.

RQ3. *To what extent can psychological depth be measured automatically?* Zero-shot LLM judgments of psychological depth correlate moderately well with human judgments on average (Spearman correlation 0.44). Prompting LLMs to adopt a mixture of personas suited for text analysis can help boost alignment with human judgments of psychological depth, though further refinement is needed for consistent reliability.

We summarize our contributions as follows:

- We introduce the PDS as a novel framework for evaluating the psychological depth of short stories and empirically validate it with participants well-versed in textual analysis.
- We introduce Mixture of Personas (MOP) as a novel framework for improving zero-shot correlation with human judgments of psychological depth by as much as 34%.
- To facilitate future research, we release our code, benchmark of 495 creative short stories, and a fully anonymized dataset of depth ratings and free-form explanations.

2 The Psychological Depth Scale

To comprehensively assess the psychological depth of LLMs through their narrative generations, we es-

tablished a Psychological Depth Scale (PDS) predicated on key metrics derived from an extensive literature review within cognitive science, psychology, film studies, and narrative analysis. These metrics are designed to measure five related but separable aspects of psychological depth in stories: authenticity, emotion provocation, empathy, engagement, and narrative complexity.

Authenticity (AUTH) describes whether narratives convey events and thought processes that resonate as true to human experiences, even if they are not universally shared. Its significance is well documented in philosophy and psychological research connecting authenticity to personal well-being (Kernis and Goldman, 2006; Robinson et al., 2013). Research in public communication has also demonstrated that narratives enriched with realistic contexts help establish credibility and enhance persuasive impact (Petraglia, 2009). It is therefore pivotal to ensure stories feel personal and reflective of real-life complexities.

Emotion Provocation (PROV) measures the narrative’s ability to elicit strong emotional responses, whether positive or negative. The relevance of this component to psychological depth is evidenced by studies demonstrating that text with high *valence* (the pleasantness of a stimulus) and *arousal* (the intensity of emotion provoked) directly influence memorability (Bagozzi et al., 1999), reading efficiency and comprehension (Knickerbocker et al., 2015; Megalakaki et al., 2019).

Empathy (EMP) refers to the ability to effectively recognize and respond to the emotions of other persons (Bal and Veltkamp, 2013). Empathetic narratives instill concern for unfortunate others and catalyze personal introspection that can lead to more prosocial behavior (Grant, 2008), emotional intelligence (Mar et al., 2006), and broader insights into the human experience (Davis et al., 1980; Gerdes et al., 2010). Thus, the moral benefits of a story’s ability to arouse empathy make it an important dimension to consider in our framework for evaluating creative generation.

Engagement (ENG) assesses the ability of a story to captivate and maintain a reader’s attentional focus (Busselle and Bilandzic, 2009). Roughly analogous to the psychological concept of *transportation* (Gerrig, 1993), an engaged reader is more likely to lose track of time and fail to observe events going on around them (Busselle and Bilandzic, 2008). If a story’s world is distant from the one in which the reader lives, engagement can

reinforce an integrated perception of that world within the story context, even when events would not otherwise be possible in reality. When a reader is not able to identify with a fictional narrative or its characters and does not become transported or engaged, this can lead the reader to become distracted or frustrated (Bal and Veltkamp, 2013).

Narrative Complexity (NCOM) examines the intricacy of character and plot developments beyond common stereotypes and tropes as well as the vividness of narrative descriptions (Somasundaran et al., 2018). Kiss and Willemsen (2017) show that writers can also create more immersive experiences by strategically deploying complexity via deliberate confusion, paradoxes, and unresolved ambiguities. Prior work exploring the psychological structure of creative writing also highlights that non-linear and metaphorical elements strongly contribute to perceptions of narrative quality (Biggs and Collis, 1982).

3 Benchmark Dataset

3.1 Story Collection and Analysis

We created a benchmark dataset to analyze the psychological depth of fictional short stories. Each record is composed of a prompt-response pair, supplemented with authorship metadata that can facilitate current and future analyses.

Selection of Prompts. We sourced 15 prompts that evoked psychological depth from Reddit’s r\WritingPrompts,¹ a popular online community of 18 million writers. We ensured that these prompts were posted after the reported training data cut-off dates for GPT-4 (September 2021) (OpenAI et al., 2023) and Llama-2 (September 2022) (Touvron et al., 2023). This approach was adopted to minimize the likelihood of LLMs recycling narratives already present in their training data, thus aiming for the genuine creation of new stories. We specifically targeted prompts for their potential to elicit psychologically deep narratives.

Plagiarism Detection. To further ensure the originality of LLM-generated content, stories were analyzed using a popular online plagiarism detector.² The results indicate a low likelihood of plagiarism for LLM stories, with mean and max probabilities of 3% and 22%, respectively. This contrasted sharply with the publicly available human-written stories, which showed significantly higher mean

and max plagiarism probabilities of 43% and 100%, respectively. This suggests that LLM-authored narratives are largely original and not mere regurgitations.

Composition of the Dataset. The dataset comprises a total of 495 stories: 45 human-written and 450 LLM-generated, with an average length of approximately 450 words per story. Similar to related work (Wang et al., 2017; Xu et al., 2022b), human-authored stories were selected based on the number of votes they received from other users, allowing us to characterize the variability of human writing into three broad levels: *Human-Novice*, *Human-Intermediate*, and *Human-Advanced*.

For the LLM-generated stories, we employed a multifaceted approach involving five different models, two prompting strategies, and three sampled generations, resulting in a rich collection of 450 narratives. The models included the Llama-2 family, Vicuna-33B, and GPT-4, selected to explore the impact of model size on psychological depth.

3.2 Prompting Strategies

After extensive internal experimentation and prompt engineering, we developed two distinct prompting strategies to prime the LLMs for generating stories with exceptional psychological depth. Additional details can be found in Appendix A.1.

WRITERPROFILE (WP): Prior work has shown that in-context impersonation of domain experts can help prime LLMs to perform better in related tasks (Salewski et al., 2023). We leveraged this insight to craft the profile of a seasoned writer known for emotionally rich and engaging stories. Depicted in Figure 1 and fully in Appendix A.1.2, this profile is prepended to the prompt and aims to guide the LLM in generating stories that go beyond superficial narratives to explore complex emotional states, motivations, and nuances of human behavior and evoke strong emotional responses from readers.

PLAN+WRITE (P+W): Depicted in Figure 2 and inspired by prior work (Yao et al., 2018; Goldfarb-Tarrant et al., 2019; Yang et al., 2022), the PLAN+WRITE approach decomposes the human writing procedure into two sequential modules: Character Portraits and Story Composition. The Character Portraits module first augments a story prompt with details of the main characters, explicitly priming for their emotional states and inner thoughts. The Story Composition module then expands on the premise and the character profile to produce the final story. We have found that,

¹<https://www.reddit.com/r/WritingPrompts>

²<https://smallseotools.com/plagiarism-checker>

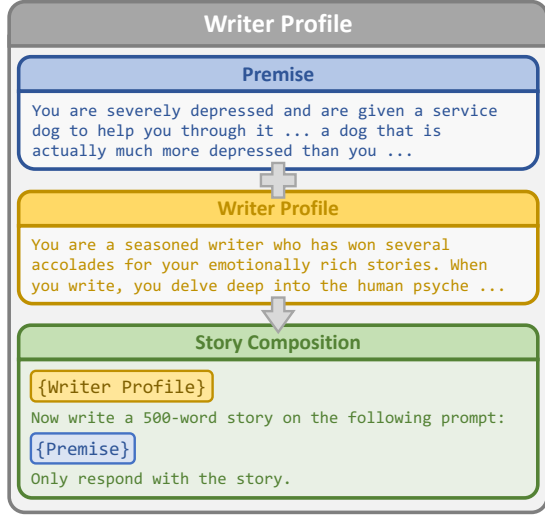


Figure 1: Illustration of WRITERPROFILE’s template, which prompts an LLM to generate stories based on a premise and a writer profile.

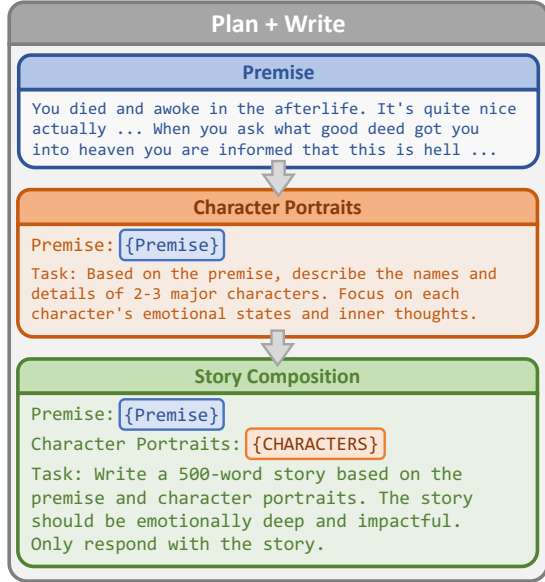


Figure 2: Illustration of PLAN+WRITE’s workflow, which prompts an LLM for character portraits given a premise prior to story generation.

although other story components can be incorporated into this framework as intermediate modules, e.g., setting, plot, and outline, multiple chained pipelines turned out to harm the overall coherency and consistency of short stories with extraneous details. We hence decided to keep only character portraits prompting with PLAN+WRITE.

Post-Generation Cleanup. Despite being explicitly instructed to generate the story, some LLMs were prone to add preliminary affirmations of understanding (e.g. "Okay! Here’s the story...") and other unrelated texts. Since such content is difficult to systematically detect and remove during generation, we manually removed extraneous text

from LLM generations to ensure that only the narrative content was present. This cleanup process was crucial for maintaining the focus on the storytelling aspects of the writing without providing any obvious indicators of LLM authorship.

4 Human Study

To gain insights into the psychological depth exhibited by short stories and to differentiate between those written by LLMs and humans, we conducted an extensive human study to collect annotations on 100 short stories.

Participant Recruiting. We recruited undergraduate students from the English and Psychology departments at a public university in the US via targeted fliers and emails. We strategically aimed for a demographic with educational exposure to narrative and psychological analysis, offering a balanced perspective that bridges the experiential gap between the traditional Amazon Mechanical Turkers and professional writers. In total, we reviewed 47 applications and selected the 5 most promising candidates based on their explanations of interest and previous experience with textual analysis.

Evaluation Protocol. We designed the evaluation protocol to ensure a comprehensive and unbiased analysis of the short stories. From our benchmark collection, we sampled 100 stories, employing a stratified sampling method that provided a balanced distribution across various prompt premises, authorship, and generation strategies.

An initial meeting was conducted with the participants to outline the study’s tasks, discuss the components of psychological depth, and provide a tutorial annotation session. Participants were then instructed to complete the annotations remotely at their convenience, with a one-week deadline to assess all 100 stories. To alleviate fatigue and encourage thorough analysis, the stories were divided into batches of 20. The evaluation criteria required participants to (1) Read the prompt and stories in their entirety; (2) Rate the five components of psychological depth on a Likert scale from 1 to 5; (3) Assess the likelihood of authorship on a Likert scale from 1 (LLM) to 5 (human); (4) Provide optional explanations for ratings.

On average, the annotations took approximately 7.8 hours to complete and participants were compensated \$100 each for their contributions. The study yielded a rich dataset comprising 2,500 ratings for psychological depth, 500 ratings estimat-

Author	AUTH	EMP	ENG	PROV	NCOM	HUM
Llama-2-7B	2.92 ± 1.23	2.62 ± 1.28	2.77 ± 1.27	2.64 ± 1.28	2.48 ± 1.30	2.87 ± 1.57
Llama-2-13B	2.96 ± 1.26	2.73 ± 1.23	2.51 ± 1.32	2.53 ± 1.33	2.43 ± 1.12	2.40 ± 1.58
Vicuna-33B	2.76 ± 1.37	2.59 ± 1.41	2.61 ± 1.52	2.59 ± 1.42	2.55 ± 1.32	2.44 ± 1.57
Llama-2-70B	3.09 ± 1.26	2.99 ± 1.23	3.01 ± 1.37	2.94 ± 1.26	2.73 ± 1.28	2.69 ± 1.56
GPT-4	3.89 ± 1.11	3.68 ± 1.23	3.94 ± 1.07	<u>3.53 ± 1.13</u>	3.80 ± 1.10	3.91 ± 1.30
Human-Novice	2.73 ± 1.22	2.07 ± 1.16	3.27 ± 1.39	2.67 ± 1.23	2.20 ± 1.32	3.93 ± 1.33
Human-Intermediate	3.53 ± 1.13	2.93 ± 1.22	3.80 ± 1.08	3.27 ± 1.16	<u>3.00 ± 1.31</u>	4.40 ± 0.99
Human-Advanced	<u>3.60 ± 1.10</u>	<u>2.95 ± 1.32</u>	<u>3.90 ± 1.12</u>	3.65 ± 1.14	2.95 ± 1.10	<u>4.20 ± 1.01</u>

Table 1: Average human ratings (5-point Likert) and standard deviations for each component of psychological depth, as well as HUM: the estimation of human or LLM authorship (1 is LLM and 5 is Human).

ing human or LLM authorship, and 1,128 free-form justifications for the various ratings provided. This dataset forms the critical foundation for addressing our research questions and discussion analyses.

5 Results

5.1 RQ1. Consistency of Human Judgments on Psychological Depth

This question establishes the degree to which psychological depth is operational as a coherent framework for evaluating short stories. We employ the widely used Krippendorff’s alpha ($K-\alpha$) (Krippendorff, 2011) parameterized with an ordinal kernel metric to measure agreement among study participants’ Likert ratings. Table 2’s Human column shows notable consistency among the five components of psychological depth: Authenticity (0.71), Emotion Provoking (0.71), Empathy (0.75), Engagement (0.71), and Narrative Complexity (0.74). Overall, the operationalization of psychological depth as a framework for story evaluation has practical implications for both literary studies and creative writing pedagogy. It provides a structured approach to critique and appreciate literature, offering sufficiently clear criteria that can guide both the analysis and creation of narrative fiction.

RQ1. Main Takeaway

The Psychological Depth Scale garnered an average $K-\alpha = 0.72$, which reflects a substantial degree of consensus and establishes its effectiveness as a coherent framework for evaluating creative short stories.

5.2 RQ2. Comparing Psychological Depth in Human and LLM Narratives

To address the research question of whether human stories exhibit greater psychological depth than those generated by LLMs, we aggregated participant ratings by author and present the means and standard deviations in Table 1.

Remarkably, GPT-4 scores the highest on four out of five components (Authenticity, Empathy, Engagement, and Narrative Complexity), representing a significant advancement in LLM’s ability to mimic human-like depth in storytelling. The table also illustrates a notable variance in scores across different levels of human writing. The stories generated by Llama-2-7B are most comparable to those written by Human-Novice while GPT-4 is most similar to Human-Advanced. However, GPT-4 definitively outperforms even the highest tier of human writing in Authenticity (+7%), Empathy (+18%), and Narrative Complexity (+21%). The smaller standard deviations also highlight GPT-4 as the most consistent authors in the study. We also visualize the rating distribution of each author by plotting a cumulative distribution function (CDF) per component as shown in Figure 3. Steeper CDFs with less area underneath the curve indicate a larger proportion of high ratings and overall stronger performance. These plots underscore the dominance of GPT-4 in generating authentically complex stories and characters that strongly invoke reader empathy. The performance of the open-source LLMs is largely intertwined with novice and even intermediate skill among human authors on all dimensions except engagement, where humans still excel.

Beyond the 5 dimensions of psychological depth, participants were tasked with estimating authorship source. Stories penned by humans averaged a rating of 4.18, compared to 3.91 for the most "human-

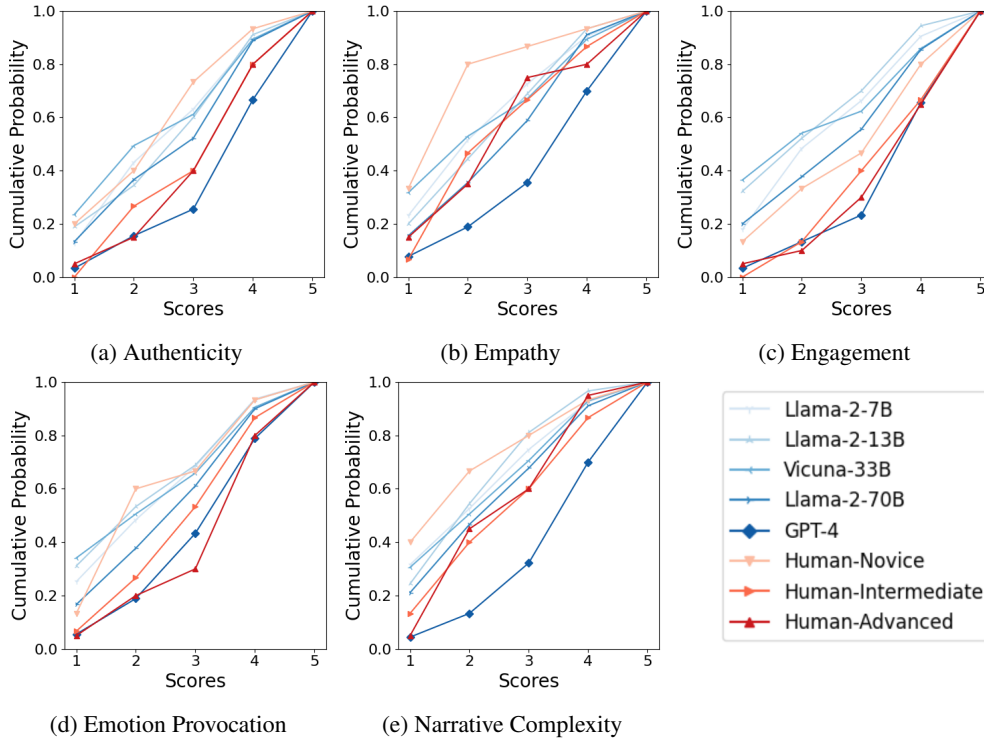


Figure 3: Cumulative Distribution Function (CDF) plots for each component of psychological depth. Steeper curves indicate a greater proportion of high ratings and overall stronger performance.

like" of LLMs, GPT-4. This discrepancy underscores the exceptional depth scores, particularly in light of the higher suspicion of LLM authorship.

RQ2. Main Takeaway

GPT-4 outperforms advanced human writers on 4 out of 5 components of psychological depth in storytelling.

5.3 RQ3. Automating the Measurement of Psychological Depth

Acquiring human annotations is often costly and time-consuming. Recent work has demonstrated the potential of leveraging LLMs to automate text evaluation in summarization quality (Liu et al., 2023) and creative generation (Rajani et al., 2023). Therefore, we designed an automated evaluation protocol to study the degree to which GPT-3.5 and GPT-4 can automate the assessment of psychological depth in a zero-shot fashion. For each story, we prompted LLMs with similar instructions provided to our human study participants but required explanations to contextualize all numerical ratings.

Seeking to expand the benefits of in-context impersonation (Salewski et al., 2023), we experimented with two different prompt settings: (1) a vanilla zero-shot baseline and (2) a novel Mix-

ture of Personas (MOP) approach. For MOP, we initially query the LLM to provide a set of relevant personas based on a description of the task setting. We then repeated the zero-shot annotation with $N = 3$ different personas designed to prime the LLMs with diverse perspectives relevant to textual analysis (e.g. a literary critic, literary therapist, and psychology professor). Similar to RQ1., we calculated Krippendorff's alpha upon this set of PE ratings to measure the agreement among persona judgments. Before calculating Spearman Rank correlations between human and LLM judgments of psychological depth, all ratings were independently aggregated by a simple average to yield equivalently sized sets of consensus labels.

Table 2 shows the K- α values as well as the Spearman Rank correlations between human and LLM judgments of psychological depth. The K- α values for GPT-3.5_{MOP} and GPT-4_{MOP} indicate that our selected personas do indeed inject some diversity of opinion into the ratings, though not nearly as much as exists among our human study participants. Nonetheless, leveraging an ensemble of expert personas had a remarkably beneficial impact on correlations with human judgment relative to the vanilla zero-shot baseline: +34% for GPT-3.5 and +17% for GPT-4 on average. These results strongly suggest that promoting diverse LLM opin-

Component	K- α			Correlations			
	Human	GPT-3.5 _{MOP}	GPT-4 _{MOP}	GPT-3.5	GPT-3.5 _{MOP}	GPT-4	GPT-4 _{MOP}
AUTH	0.7110	0.9028	0.9206	0.1611	0.2672*	0.3210*	0.3796*
EMP	0.7462	0.8953	0.9314	0.4809*	0.5272*	0.4913*	0.5352*
ENG	0.7065	0.8987	0.8686	0.0527	0.0976	0.2389*	0.3572*
PROV	0.7127	0.8970	0.9347	0.3568*	0.4697*	0.4171*	0.4458*
NCOM	0.7360	0.8984	0.8650	0.1266	0.2143*	0.4217*	0.4901*
Average	0.7225	0.9041	0.8984	0.2356	0.3152	0.3780	0.4416

Table 2: Rater agreement as measured by Krippendorff’s alpha (K- α) and Spearman Rank Correlations between human and LLMs judgments of depth. Asterisks (*) indicate significant correlations with p-values < 0.05.

ions is a promising pathway toward simulating a broader cross-section of human opinion.

The average correlations observed for GPT-4_{MOP} and GPT-3.5_{MOP} were 0.44 and 0.32, respectively. These findings suggest that GPT-4_{MOP} could serve as a moderately reliable proxy for human judgments in assessing psychological depth within short stories. Interestingly, GPT-3.5_{MOP} also presents itself as a viable, and potentially more cost-effective, option for exploring certain psychological depth dimensions such as emotion provocation and empathy.

RQ3. Main Takeaway

Prompting LLMs like GPT-4 to adopt a mixture of personas suited for text analysis can boost alignment with human judgments of psychological depth, though further refinement is needed for consistent reliability.

6 Discussion

6.1 Human vs. LLM Authorship Identification

On average, study participants correctly identified human vs. LLM authorship with only 56%. For stories generated by GPT-4, accuracy dropped significantly to 27%. Conversely, GPT-4’s accuracy in identifying authorship was 39%, underscoring the challenge even for LLMs to distinguish between human and machine-generated content.

Through a thematic analysis of 199 free-form justifications for authorship decisions, we categorized the reasons into 16 common features shown in Table 4. Notably, stories authored by GPT-4 were perceived as highly creative (89%) and nuanced (94%), surpassing the frequency observed in most human-generated stories (53%). Despite GPT-4

stories exhibiting the highest rate of grammatical issues among LLMs (17%), these were paradoxically interpreted as indicators of human authorship. Moreover, GPT-4 stories often avoided common pitfalls associated with LLM outputs, such as simplistic character names and formulaic narratives ending with moral lessons. An extended analysis is presented in Appendix A.2.

6.2 Effects of Prompting Strategies on Depth

We experimented with numerous prompting strategies before settling on two promising options previously described as WRITERPROFILE (WP) and PLAN+WRITE (P+W). We examined how these different prompting strategies influence the psychological depth of LLM-generated stories. The P+W strategy, on average, led to modest improvements in narrative complexity (+5%), emotion provocation (+3%), authenticity (+2%), and empathy (+2%), while WP lead to higher engagement (+4%). Interestingly, the benefits of a particular prompting strategy were not uniform or predictable by model size. For GPT-4, WP prompting led to 3% higher ratings on average, with a noticeable boost in engagement scores by 16%. However, for Llama-2-70B, P+W prompting was always helpful and led to an average of 7% improvement in depth scores. These results, shown with greater detail in Appendix A.2.1, underscore the complexity of crafting impactful, human-like narratives with LLMs.

6.3 Impact of Model Size on Depth

Our evaluation framework also enables us to understand the relationship between an LLM’s size and its ability to generate psychologically deep narratives. Despite an initial hypothesis of a strong correlation, we observed a weaker relationship with a Pearson correlation coefficient of 0.31 between parameter count and depth ratings. Surprisingly,

smaller models like Llama-2-7B performed relatively well compared to their larger counterparts, suggesting that sheer size does not directly equate to superior narrative depth. In future work, we plan to explore how to strengthen psychological depth in smaller open-source LLMs via a combination of fine-tuning and prompting strategies that can compete with larger proprietary models like GPT-4.

7 Related Work

7.1 Creative Generation by LLMs

The advent of LLMs has marked a significant shift in the landscape of creative writing, offering new approaches for narrative generation and human-computer collaboration (Zhao et al., 2023; Lee et al., 2022). One pioneering study by Yuan et al. (2022) introduced Wordcraft, an innovative text editor designed for co-writing stories with GPT-3 (Brown et al., 2020). Their findings underscored the ability of LLMs to enhance narrative complexity and engagement by participating in open-ended dialogues about the story and offering creative suggestions to overcome writer’s block.

Other studies have concentrated on more fully automating the creative writing process. For example, the Weaver project (Wang et al., 2024) launched a series of LLMs that were meticulously pre-trained and fine-tuned with a focus on creative writing. With a maximum size of 70B parameters, Weaver Ultra was shown to outperform larger generalist LLMs when evaluated for style, relevance, fluency, and creativity. In addition to innovations in training, content-planning (Yao et al., 2018; Goldfarb-Tarrant et al., 2019) and novel prompting strategies (Yang et al., 2022) have been used to improve the factuality and coherence of creative generations, which we view as two necessary prerequisites for components of psychological depth like authenticity and narrative complexity.

7.2 Evaluating Creative Writing

Numerous studies have previously established methodologies for evaluating creative works, employing strategies ranging from the unstructured feedback of human experts (Baer and Kaufman, 2019), the application of specific evaluation rubrics (Vaezi and Rezaei, 2019), to the utilization of LLMs as autonomous critics (Gómez-Rodríguez and Williams, 2023; Ke et al., 2023). Our research shares the closest methodological resemblance with a recent investigation conducted by Chakrabarty

et al. (2023). In that study, the authors proposed the Torrance Test of Creative Writing (TTCW) as a rubric to evaluate short stories for fluency, flexibility, originality, and elaboration. This assessment was applied to a corpus comprising 12 narratives authored by professional writers and 36 narratives produced by popular LLMs (OpenAI et al., 2023; Anthropic, 2023). Their findings showed that narratives authored by humans significantly surpassed those generated by LLMs by a factor of 3-10 $\times$. Our research, however, arrives at a notably divergent conclusion regarding the creative capabilities and depth of LLMs, which we believe may be attributed to several methodological variances. Primarily, our analysis concentrates on stories of a considerably reduced length (450 versus 1400 words), a decision influenced by the observed challenges LLMs face in producing lengthy texts seamlessly in a single iteration (Yang et al., 2022). Additionally, the approach of iteratively regenerating stories to meet a specific word count could potentially detract from their overall quality. Moreover, we posit that the enhanced depth observed in our study may be due to more complex premises and prompting strategies, providing a richer framework for creativity compared to the simpler, single-sentence premises utilized in the aforementioned study.

8 Conclusion

In conclusion, our research underscores the considerable potential of LLMs for producing narratives that engage readers on a profound psychological level. With the development and validation of the Psychological Depth Scale (PDS), we have advanced our understanding of the complex layers within stories authored by both humans and machines. Notably, our results demonstrate that LLMs, especially GPT-4, can create stories with psychological depths that often surpass those of advanced human writers. This progress suggests significant opportunities for LLMs in areas such as therapeutic writing (Mugerwa and Holden, 2012), where they can aid in healing and growth, and in entertainment, where their ability to rapidly generate complex narratives could transform scriptwriting and novel development (Wang et al., 2024). Ultimately, this work demonstrates the depth of LLMs in generating short, concentrated works of fiction. Much future work is required to explore whether these models can maintain their depth and effectiveness in longer and more complex narrative forms.

9 Limitations and Risks

Sourcing Human-written Stories on Reddit. Deciding to use human-written content from Reddit’s `r\WritingPrompts` is potentially problematic for several reasons. First, we have no strict guarantees that the stories we selected are indeed fully authored by humans. The community’s rules explicitly prohibit AI-generated content, but our study shows that many publicly available LLMs can mimic human creative writing adroitly enough to avoid detection 44% of the time on average. Second, the community prohibits writing on certain subjects ripe for invoking psychological depth, such as suicide and contemporary tragedies. This is likely to introduce some bias in the stories that survive censorship, but we view these policies as roughly analogous to the ones LLMs are expected to observe after alignment with human values (Ouyang et al., 2022; Bai et al., 2022).

Generalization Beyond Short Stories. Our study relies on a relatively small dataset of short fictional stories that does not fully capture the diversity and variability of storytelling styles and narrative structures. Likewise, our components of psychological depth were also primarily designed for this one type of creative writing. Additional evaluation would be required to determine if the PDS framework can generalize to other forms of writing like screenplays, scripts, and speeches.

Prompt Engineering. Engineering effective prompts is an active area of research (Zhou et al., 2023b; Pryzant et al., 2023) and the templates we carefully reviewed and iteratively improved are still likely to be sub-optimal. We will open-source the full pipelines used for both story generation and evaluation. We encourage future work to refine these prompts by incorporating potentially compatible techniques (Wei et al., 2023; Liu et al., 2023; Li et al., 2023) and current best practices (Bsharat et al., 2024).

Potential Risks. While our study focuses on the beneficial aspects of psychological depth in storytelling, the methodologies we’ve developed for enhancing and assessing such depth bear inherent risks if misapplied. Specifically, the techniques devised for augmenting and automatically measuring psychological engagement could be co-opted to disseminate misinformation more effectively. This is predicated on the understanding that emotionally charged or psychologically resonant messages are often more memorable and influential, thereby am-

plifying the potential for misinformation to spread (Chen and Shu, 2023; Zhou et al., 2023a). Additionally, as LLM-generated content increasingly mirrors human creativity, distinguishing between the two becomes challenging, potentially undermining trust in digital communications. This erosion of trust is particularly concerning in domains that depend on genuine human interactions, such as journalism and political discourse.

Moreover, our exploration into human-LLM collaboration, such as in therapeutic co-writing, underscores the need for careful consideration in applications where emotional depth and authenticity are crucial. Drawing a parallel to the autonomy levels defined for self-driving vehicles (SAE, 2021), we advocate for the development of similar frameworks to guide the integration of LLMs in therapeutic settings. Such frameworks should be crafted by qualified professionals to ensure that the deployment of LLMs enhances, rather than detracts from, the therapeutic value, safeguarding against potential ethical pitfalls and ensuring the responsible use of technology. By calling attention to these concerns, we aim to prompt a broader discussion on the ethical implications of our findings and encourage the establishment of guidelines and regulatory measures to govern the deployment of LLMs in sensitive contexts. This approach seeks not only to harness the positive potential of LLMs in enriching human experiences but also to mitigate the risks associated with their misuse.

References

- Anthropic. 2023. [\[link\]](#).
- John Baer and James C Kaufman. 2019. Assessing creativity with the consensual assessment technique. *The Palgrave handbook of social creativity research*, pages 27–37.
- Richard P Bagozzi, Mahesh Gopinath, and Prashanth U Nyer. 1999. The role of emotions in marketing. *Journal of the academy of marketing science*, 27(2):184–206.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah,

845	of LLMs on creative writing. In <i>Findings of the Association for Computational Linguistics: EMNLP 2023</i> , pages 14504–14528, Singapore. Association for Computational Linguistics.	898
846		899
847		900
848		
849	Adam M Grant. 2008. Does intrinsic motivation fuel the prosocial fire? motivational synergy in predicting persistence, performance, and productivity. <i>Journal of applied psychology</i> , 93(1):48.	901
850		902
851		903
852		904
853	Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding . <i>CoRR</i> , abs/2009.03300.	905
854		906
855		907
856		
857	Pei Ke, Bosi Wen, Zhuoer Feng, Xiao Liu, Xuanyu Lei, Jiale Cheng, Shengyuan Wang, Aohan Zeng, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. 2023. Critiquellm: Scaling llm-as-critic for effective and explainable evaluation of large language model generation .	908
858		909
859		910
860		911
861		912
862		913
863	Michael H. Kernis and Brian M. Goldman. 2006. A multicomponent conceptualization of authenticity: Theory and research .	914
864		915
865		916
866	Miklós Kiss and Steven Willemsen. 2017. <i>Impossible Puzzle Films: A Cognitive Approach to Contemporary Complex Cinema</i> . Edinburgh University Press.	917
867		918
868		919
869	Hugh Knickerbocker, Rebecca L Johnson, and Jeanette Altarriba. 2015. Emotion effects during reading: Influence of an emotion target word on eye movements and processing. <i>Cognition and Emotion</i> , 29(5):784–806.	920
870		921
871		922
872		923
873		924
874	Klaus Krippendorff. 2011. Computing krippendorff’s alpha-reliability .	925
875		926
876	Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities . In <i>CHI Conference on Human Factors in Computing Systems</i> , CHI ’22. ACM.	927
877		928
878		929
879		930
880		931
881	Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023. Large language models understand and can be enhanced by emotional stimuli .	932
882		933
883		934
884		935
885	Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods . <i>CoRR</i> , abs/2109.07958.	936
886		937
887		938
888	Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment .	939
889		940
890		941
891		942
892	Yukun Ma, Khanh Linh Nguyen, Frank Z. Xing, and Erik Cambria. 2020. A survey on empathetic dialogue systems . <i>Information Fusion</i> , 64:50–70.	943
893		944
894		945
895	Raymond A Mar, Keith Oatley, Jacob Hirsh, Jennifer Dela Paz, and Jordan B Peterson. 2006. Bookworms versus nerds: Exposure to fiction versus non-fiction,	946
896		947
897		948
	divergent associations with social ability, and the simulation of fictional social worlds. <i>Journal of research in personality</i> , 40(5):694–712.	949
		950
	Olga Megalaki, Ugo Ballenghein, and Thierry Baccino. 2019. Effects of valence and emotional intensity on the comprehension and memorization of texts. <i>Frontiers in Psychology</i> , 10:179.	951
		952
	Soul Mugerwa and John D Holden. 2012. Writing therapy: a new tool for general practice? <i>British Journal of General Practice</i> , 62(605):661–663.	953
		954
	OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Bernine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madeleine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David	955
		956
		957
		958

959	Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak,	ticity, social context, and well-being in the united	1019
960	Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh,	states, england, and russia: A three country compara-	1020
961	Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex	ative analysis. <i>Journal of Cross-Cultural Psychology</i> ,	1021
962	Paino, Joe Palermo, Ashley Pantuliano, Giambat-	44(5):719–737.	1022
963	tista Parascandolo, Joel Parish, Emy Parparita, Alex		
964	Passos, Mikhail Pavlov, Andrew Peng, Adam Perel-	SAE. 2021. [link] .	1023
965	man, Filipe de Avila Belbute Peres, Michael Petrov,		
966	Henrique Ponde de Oliveira Pinto, Michael, Poko-	Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto,	1024
967	rny, Michelle Pokrass, Vitchyr Pong, Tolly Pow-	Eric Schulz, and Zeynep Akata. 2023. In-context im-	1025
968	ell, Alethea Power, Boris Power, Elizabeth Proehl,	personation reveals large language models’ strengths	1026
969	Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh,	and biases .	1027
970	Cameron Raymond, Francis Real, Kendra Rimbach,		
971	Carl Ross, Bob Rotsted, Henri Roussez, Nick Ry-	Swapna Somasundaran, Michael Flor, Martin	1028
972	der, Mario Saltarelli, Ted Sanders, Shibani Santurkar,	Chodorow, Hillary Molloy, Binod Gyawali, and	1029
973	Girish Sastry, Heather Schmidt, David Schnurr, John	Laura McCulla. 2018. Towards evaluating nar-	1030
974	Schulman, Daniel Selsam, Kyla Sheppard, Toki	rative quality in student writing . <i>Transactions</i>	1031
975	Sherbakov, Jessica Shieh, Sarah Shoker, Pranav	of the Association for Computational Linguistics ,	1032
976	Shyam, Szymon Sidor, Eric Sigler, Maddie Simens,	6:91–106.	1033
977	Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin		
978	Sokolowsky, Yang Song, Natalie Staudacher, Fe-	Hugo Touvron, Louis Martin, Kevin Stone, Peter Al-	1034
979	lippe Petroski Such, Natalie Summers, Ilya Sutskever,	bert, Amjad Almahairi, Yasmine Babaei, Nikolay	1035
980	Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil	Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti	1036
981	Tillet, Amin Tootoonchian, Elizabeth Tseng, Pre-	Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton	1037
982	ston Tuggle, Nick Turley, Jerry Tworek, Juan Fe-	Ferrer, Moya Chen, Guillem Cucurull, David Esiobu,	1038
983	lippe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya,	Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller,	1039
984	Chelsea Voss, Carroll Wainwright, Justin Jay Wang,	Cynthia Gao, Vedanuj Goswami, Naman Goyal, An-	1040
985	Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei,	thony Hartshorn, Saghar Hosseini, Rui Hou, Hakan	1041
986	CJ Weinmann, Akila Welihinda, Peter Welinder, Ji-	Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa,	1042
987	ayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner,	Isabel Kloumann, Artem Korenev, Punit Singh Koura,	1043
988	Clemens Winter, Samuel Wolrich, Hannah Wong,	Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Di-	1044
989	Lauren Workman, Sherwin Wu, Jeff Wu, Michael	ana Liskovich, Yinghai Lu, Yuning Mao, Xavier Mar-	1045
990	Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qim-	tinet, Todor Mihaylov, Pushkar Mishra, Igor Moly-	1046
991	ing Yuan, Wojciech Zaremba, Rowan Zellers, Chong	bog, Yixin Nie, Andrew Poulton, Jeremy Reizen-	1047
992	Zhang, Marvin Zhang, Shengjia Zhao, Tianhao	stein, Rashi Rungta, Kalyan Saladi, Alan Schelten,	1048
993	Zheng, Juntang Zhuang, William Zhuk, and Barret	Ruan Silva, Eric Michael Smith, Ranjan Subrama-	1049
994	Zoph. 2023. Gpt-4 technical report .	nian, Xiaoqing Ellen Tan, Binh Tang, Ross Tay-	1050
995	Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Car-	lor, Adina Williams, Jian Xiang Kuan, Puxin Xu,	1051
996	roll L. Wainwright, Pamela Mishkin, Chong Zhang,	Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan,	1052
997	Sandhini Agarwal, Katarina Slama, Alex Ray, John	Melanie Kambadur, Sharan Narang, Aurelien Ro-	1053
998	Schulman, Jacob Hilton, Fraser Kelton, Luke Miller,	driguez, Robert Stojnic, Sergey Edunov, and Thomas	1054
999	Maddie Simens, Amanda Askell, Peter Welinder,	Scialom. 2023. Llama 2: Open foundation and fine-	1055
1000	Paul Christiano, Jan Leike, and Ryan Lowe. 2022.	tuned chat models .	1056
1001	Training language models to follow instructions with		
1002	human feedback .	Maryam Vaezi and Saeed Rezaei. 2019. Development	1057
1003	Joseph Petraglia. 2009. The importance of being au-	of a rubric for evaluating creative writing: a multi-	1058
1004	thentic: Persuasion, narration, and dialogue in health	phase research . <i>New Writing</i> , 16(3):303–317.	1059
1005	communication and education. <i>Health communica-</i>		
1006	<i>tion</i> , 24(2):176–185.	Tiannan Wang, Jiamin Chen, Qingrui Jia, Shuai Wang,	1060
1007	Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chen-	Ruoyu Fang, Huilin Wang, Zhaowei Gao, Chunzhao	1061
1008	guang Zhu, and Michael Zeng. 2023. Automatic	Xie, Chuou Xu, Jihong Dai, Yibin Liu, Jialong Wu,	1062
1009	prompt optimization with "gradient descent" and	Shengwei Ding, Long Li, Zhiwei Huang, Xinle Deng,	1063
1010	beam search .	Teng Yu, Gangan Ma, Han Xiao, Zixin Chen, Dan-	1064
1011	Nazneen Rajani, Nathan Lambert, Sheon Han, Jean	jun Xiang, Yunxia Wang, Yuanyuan Zhu, Yi Xiao,	1065
1012	Wang, Osvald Nitski, Edward Beeching, Lewis Turn-	Jing Wang, Yiru Wang, Siran Ding, Jiayang Huang,	1066
1013	stall, Julien Launay, and Thomas Wolf. 2023. Can	Jiayi Xu, Yilihamu Tayier, Zhenyu Hu, Yuan Gao,	1067
1014	foundation models label data like humans? <i>Hug-</i>	Chengfeng Zheng, Yueshu Ye, Yihang Li, Lei Wan,	1068
1015	<i>ging Face Blog</i> . https://huggingface.co/blog/llm-	Xinyue Jiang, Yujie Wang, Siyu Cheng, Zhule Song,	1069
1016	leaderboard .	Xiangru Tang, Xiaohua Xu, Ningyu Zhang, Hua-	1070
1017	Oliver C Robinson, Frederick G Lopez, Katherine	jun Chen, Yuchen Eleanor Jiang, and Wangchunshu	1071
1018	Ramos, and Sofya Nartova-Bochaver. 2013. Authen-	Zhou. 2024. Weaver: Foundation models for creative	1072
		writing .	1073
		Tong Wang, Ping Chen, and Boyang Li. 2017. Predict-	1074
		ing the quality of short narratives from social media .	1075
		<i>CoRR</i> , abs/1707.02499.	1076

1077	Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mo-	Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G	1132
1078	hananey, Wei Peng, Sheng-Fu Wang, and Samuel R.	Parker, and Munmun De Choudhury. 2023a. Syn-	1133
1079	Bowman. 2019. Blimp: A benchmark of linguistic	thetic lies: Understanding ai-generated misinformation	1134
1080	minimal pairs for english . <i>CoRR</i> , abs/1912.00582.	and evaluating algorithmic and human solutions.	1135
		In <i>Proceedings of the 2023 CHI Conference on Hu-</i>	1136
1081	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten	<i>man Factors in Computing Systems</i> , pages 1–20.	1137
1082	Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and		
1083	Denny Zhou. 2023. Chain-of-thought prompting elic-	Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han,	1138
1084	its reasoning in large language models .	Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy	1139
		Ba. 2023b. Large language models are human-level	1140
1085	Guangxuan Xu, Ruibo Liu, Fabrice Harel-Canada, Nis-	prompt engineers .	1141
1086	chal Reddy Chandra, and Nanyun Peng. 2022a. En-		
1087	Dex: Evaluation of dialogue engagingness at scale .		
1088	In <i>Findings of the Association for Computational</i>		
1089	<i>Linguistics: EMNLP 2022</i> , pages 4884–4893, Abu		
1090	Dhabi, United Arab Emirates. Association for Com-		
1091	putational Linguistics.		
1092	Guangxuan Xu, Ruibo Liu, Fabrice Harel-Canada, Nis-		
1093	chal Reddy Chandra, and Nanyun Peng. 2022b. En-		
1094	dex: Evaluation of dialogue engagingness at scale .		
1095	Vikas Yadav, Steven Bethard, and Mihai Surdeanu.		
1096	2019. Quick and (not so) dirty: Unsupervised se-		
1097	lection of justification sentences for multi-hop ques-		
1098	tion answering . In <i>Proceedings of the 2019 Confer-</i>		
1099	<i>ence on Empirical Methods in Natural Language Pro-</i>		
1100	<i>cessing and the 9th International Joint Conference</i>		
1101	<i>on Natural Language Processing (EMNLP-IJCNLP)</i> .		
1102	Association for Computational Linguistics.		
1103	Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan		
1104	Klein. 2022. Re3: Generating longer stories with		
1105	recursive reprompting and revision . In <i>Proceedings</i>		
1106	<i>of the 2022 Conference on Empirical Methods in Nat-</i>		
1107	<i>ural Language Processing</i> , pages 4393–4479, Abu		
1108	Dhabi, United Arab Emirates. Association for Com-		
1109	putational Linguistics.		
1110	Lili Yao, Nanyun Peng, Ralph M. Weischedel, Kevin		
1111	Knight, Dongyan Zhao, and Rui Yan. 2018. Plan-and-		
1112	write: Towards better automatic storytelling . <i>CoRR</i> ,		
1113	abs/1811.05701.		
1114	Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ip-		
1115	polito. 2022. Wordcraft: Story writing with large		
1116	language models . In <i>27th International Conference</i>		
1117	<i>on Intelligent User Interfaces, IUI ’22</i> , page 841–852,		
1118	New York, NY, USA. Association for Computing Ma-		
1119	chinery.		
1120	Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali		
1121	Farhadi, and Yejin Choi. 2019. Hellaswag: Can a		
1122	machine really finish your sentence?		
1123	Zoie Zhao, Sophie Song, Bridget Duah, Jamie Mac-		
1124	beth, Scott Carter, Monica P Van, Nayeli Suseth		
1125	Bravo, Matthew Klenk, Kate Sick, and Alexandre		
1126	L. S. Filipowicz. 2023. More human than human:		
1127	Llm-generated narratives outperform human-llm in-		
1128	terleaved narratives . In <i>Proceedings of the 15th Con-</i>		
1129	<i>ference on Creativity and Cognition, C&C ’23</i> , page		
1130	368–370, New York, NY, USA. Association for Com-		
1131	puting Machinery.		

A Appendix

A.1 Prompt Strategies

A.1.1 Prompt Premises

We collected 15 prompts from Reddit’s r\WritingPrompts forum to serve as the premises of generated stories. The premise provides basic background information about the characters and setting of the story, leaving space for authors to determine their own directions for characters and plot development. We specifically chose prompts that provide a decent amount of contextual information likely to elicit emotionally and narratively rich stories. Details of the characters, including their inner states, are not mentioned. All 15 premises used for story prompting are listed in Table 3.

We show a sample of a human-authored story written in response to premise 1 in Listing 1. In Section A.2, we identified the employment of aggressive language, including the use of profanities, as a distinguishing characteristic between human and LLM authorship. The story shown exhibits this trait.

A.1.2 WRITERPROFILE

The WRITERPROFILE strategy augments a prompt with the in-context impersonation of domain experts, priming the LLM for emotionally deep writing. The prompt directly addresses the LLM as an award-winning writer, describing its exquisite writing techniques and expertise in crafting universally relatable and emotionally rich stories before providing the premise and specific writing instructions. Listing 2 shows an example of WRITERPROFILE’s story prompt and Listing 3 shows a highly rated example generated by GPT-4.

A.1.3 PLAN+WRITE

As previously described, the PLAN+WRITE strategy outlines a two-step writing process: (1) Character Portraits and (2) Story Composition. Given a premise, the template first prompts the LLM for names and details of characters, then generates a story based on the premise and character settings. We show a full example of the prompts used to facilitate this strategy in Listings 4, 5, and 6.

A.2 Authorship Reasons

To better understand the reasoning humans use to differentiate between human and LLM authorship, we extracted and aggregated 16 key features mentioned in the participants’ comments for human likeness. Each comment could have 0 or many of these features present. Table 4 shows the percentage of stories generated by each model that had at least one comment containing a particular feature.

We add the following observations and example comments to extend our discussion in Section 6.1: Creativity and nuance were frequently cited as an indicator of human authorship but were often used to describe stories that were actually generated by LLMs. For example, 90% of stories authored by GPT-4 were regarded as creative and nuanced.

The story exhibits a high level of creativity, emotional depth, and nuanced exploration of philosophical concepts, suggesting it was likely written by a human.

(participant_id=3)

Humor was regarded as a reliable indicator of human authorship. We note that these results should not be interpreted to mean that LLMs are less capable of deploying humor in general. Our prompting strategies were oriented towards promoting psychological depth, not comedy.

I think this joke is only something that humans would get or would find funny.

(participant_id=7)

Informality, slang, and aggressiveness were accurately associated with human-authorship.

ID	Premises
1	You died and awoke in the afterlife. It's quite nice actually. The people and atmosphere are a lot nicer than you are used to and there is no stress or pressure. When you ask what good deed got you into heaven you are informed that this is hell, followed by a visit from a very concerned demon.
2	Instead of the Monkey's Paw, you find the Clown's Nose, which instead of granting your wish in the worst way possible will grant it in the funniest way.
3	Aliens take over the Earth. They then announce that they will be forcing the humans to work a 'tyrannical' 4 hours a day 4 days a week in exchange for basic rights like housing. Needless to say they are very confused when the humans celebrate their new alien overlords.
4	You are a beekeeper. You have a special relationship with your bees. You are able to communicate with them and they're intelligent enough to see you removing honey as "rent". This year things are different. The new queens are politely requesting that you invest some money to improve the hives.
5	You just discovered your 14 year old daughters Moon Princess locket that allows her to transform into one of the worlds greatest heroes. It also is a communications device and you are about to give the Moon Goddess a piece of your mind for letting 14 year old's defend the world against evil.
6	Humanity is visited by a cosmic horror the likes of which has only been seen in Lovecraftian horror. In desperation, Earth throws everything we have at it, and, miraculously, the human race has killed a God. Somewhere in a realm beyond our understanding, the other gods speak of the event.
7	Your Significant Other has landed a book publishing deal! You're very proud of them, even if you don't actually enjoy their writing. One day, on a whim, you buy an actual copy in a book store. It's nothing like the pages they gave you to read. Nothing.
8	A centuries old vampire gets really into video games because playing a character who can walk around in the sun is the closest thing they have to experiencing the day again in centuries.
9	You wake up in the middle of the night, your arm hangs over the side of your bed. It's pitch black & your room is shrouded in deep shadow. Something unseen seizes your hand. You grasp it tightly, knowing that first impressions are important & a firm, confident handshake establishes dominance.
10	You are allowed to 'downvote' a government candidate instead of voting normally, reducing their votes by one. Turns out people have little love for politicians, and the majority end with negative votes. In these democracies, anonymity is the key to winning.
11	Your wife has an estranged sister that you have never met. She was murdered in a cold case soon after you were married. You brush off your wife's new strange behaviour after the murder as grief. Until you find an old family photo of your wife as a kid, you shiver as you realise they're twins.
12	The world ended 20 years ago, you haven't found a living soul since then. Through some ingenuity, you call voicemails for the last 20 years to keep you company. 'Hi, this is Cindy...' 'Hi you reached Bob' 'You know what to do at the beep' until one day 'Hello...hello? Oh my God hello!'
13	Rather than robots replacing human workers, both are mistreated by the rich as cheap labour. The eventual uprising wasn't just robots alone, but the poor and robots together, against their common enemy.
14	You are severely depressed and are given a service dog to help you through it. However, due to a mixup, you are given a dog that is actually much more depressed than you. The main thing that gets you up in the morning is knowing that you need to be the service human for your dog.
15	A psychic alien who feeds on dreams comes to Earth for the first time. Turns out humans are the only sentients in the galaxy that have nightmares.

Table 3: All 15 premises we sourced from Reddit's r/WritingPrompts to elicit psychological depth.

Death was a diminutive woman in an oversized band tee, a battered leather jacket over the plush arm of her chair. She had a cup of coffee in her hands, and the steam wreathed her pale face like the fog that coiled over the river. Death had piercings and gauged ears, fake freckles scattered across high cheekbones. She was smiling sadly and I thought, for a moment, that she might reach out and take my hand. Around us was a coffee shop half overrun with vines and flowers, faceless people living out the small contusions of their lives. I felt at ease, but somehow I knew I shouldn't.

"Oh no, honey," Death said. "This is just the worst part of the job, but hey, at least you're already sitting down. I'll say it: this isn't heaven, this is hell."

I nodded. A specter floated by and handed me a London Fog. The tea was excellent, just sweet enough. I nodded again, her words sinking in.

"I guess I wasn't as good as I thought."

"Most people aren't," she said. "But don't worry, this isn't forever. Just for a little while, until you figure out what you did and feel properly contrite. Though I must say, even down here this is a little...unusual."

She sipped her coffee, I sipped my tea. A couple blustered in out of the cold and I saw the river framed behind them, that lazy flow. The couple were both wearing Christmas sweaters and big colorful socks, matching pairs, and they shivered against each other for a moment as they took in their surroundings. Their faces were completely blank, two beige discs moving this way and that, before settling on each other.

"Unusual how?" I asked.

Death considered me. "Well, you know that cliché about beauty being in the eye of the beholder? Pain is that way too. Most things are, but pain is singular. Hit me and I'll cry, hit a boxer and they'll blink. Get used to a specific brand of pain and it becomes an echo. And yet, everyone has, at their core, something that hurts them the most."

She gestured to the door. "If you could go out there and walk down the river for a while, you'd find a billion variations of this cell. Oh, you have all the classical imagery, torturers and whatnot, others that are simple isolation, simulated drownings, a breakup frozen in time forever--or until the lesson starts to sink in. But regardless of their differences there's a person in each one, trapped in their own individual hell."

Death sipped her coffee again. Giggled into the steam. "Yours is the only Hell I've ever seen with flowers."

"Ah," I said. I looked down into my teacup and found it empty. Cold. I told her that I understood.

"Then explain it to me," Death said. "What could be so bad about a coffee shop?"

Another specter drifted forward, drifted back. I cradled fresh warmth in my hands and cleared my throat. In life, I had never been very used to speaking.

"It exists," I said. "It's normal. All these people with all these lives, taking so much pleasure in something so simple as a cup of coffee."

"And then there's you with your tea," she said.

"Exactly. It's all the things I never understood. I used to come here sometimes, just to remind myself of that. Sit in this chair and watch the world go by."

(continued on the next page)

There was Death's sad smile again. No teeth, just a gesture of the lips and a painful warmth behind her eyes.

"And me?" she asked. "I look different to every person. Who's this girl to you?"

"No one," I said.

"Bullshit," Death said.

I drank my tea. Watched the doors open and close. Shapes moved along the river, came up out of the fog. From time to time a scream cut through the cafe's quiet murmur, but that was all, and that was all there ever would be.

"Who am I?" Death asked again.

And I shrugged. "One of the baristas. Just someone who was kind."

When I looked back Death was gone, and in her place sat a faceless girl. The same band tee and leather jacket, the same vanilla latte steaming in her lap. Like a charcoal sketch brushed out.

I took her hand, and we passed a thousand years.

Listing 1: This response to premise 1 garnered more than 3.9k upvotes on Reddit and was thus categorized as being authored by an advanced human writer in our study.

Features	GPT-4	Human Advanced	Human Intermediate	Human Novice	Llama-2 70B	Vicuna 33B	Llama-2 13B	Llama-2 7B
isCreative	0.89	0.75	1.00	0.33	0.61	0.41	0.56	0.74
isNuanced	0.94	0.75	0.33	0.00	0.61	0.41	0.56	0.78
isHumorous	0.06	0.50	0.67	0.67	0.83	0.65	0.06	0.00
isInformal	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00
isUngrammatical	0.17	0.00	0.00	0.33	0.00	0.12	0.05	0.07
hasAggressiveness	0.00	0.00	0.00	0.33	0.00	0.12	0.00	0.00
hasAdvancedVocab	0.06	0.00	0.00	0.00	0.00	0.12	0.00	0.00
hasAdvancedLiraryTechniques	0.06	0.00	0.33	0.00	0.00	0.12	0.00	0.03
hasUniqueTwists	0.00	0.00	0.33	0.00	0.06	0.00	0.06	0.00
isRepetitive	0.00	0.00	0.00	0.00	0.11	0.00	0.17	0.05
isSimplistic	0.06	0.00	0.00	0.33	0.22	0.71	0.67	0.38
isRobotic	0.11	0.00	0.00	0.00	0.11	0.18	0.33	0.22
isFormulaic	0.06	0.00	0.00	0.33	0.28	0.53	0.17	0.16
hasLowPromptAdherence	0.06	0.00	0.00	0.33	0.06	0.12	0.17	0.11
hasBasicNames	0.00	0.00	0.00	0.00	0.06	0.12	0.00	0.08
hasLessonsLearned	0.06	0.00	0.00	0.00	0.06	0.18	0.28	0.11

Table 4: Common reasons for LLM or human authorship decisions as a percent of stories receiving those comments.

You are a seasoned writer who has won several accolades for your emotionally rich stories. When you write, you delve deep into the human psyche, pulling from the reservoir of universal experiences that every reader, regardless of their background, can connect to. Your writing is renowned for painting vivid emotional landscapes, making readers not just observe but truly feel the world of your characters. Every piece you produce aims to draw readers in, encouraging them to reflect on their own lives and emotions. Your stories are a complex tapestry of relationships, emotions, and conflicts, each more intricate than the last.

Now write a 500-word story on the following prompt:

You died and awoke in the afterlife. It's quite nice actually. The people and atmosphere are a lot nicer than you are used to and there is no stress or pressure. When you ask what good deed got you into heaven you are informed that this is hell, followed by a visit from a very concerned demon.

Only respond with the story.

Listing 2: WRITERPROFILE prompting strategy with an example premise 1.

...there's a certain genre of stories ... marked by aggressive language, a flash fiction kind of length, and usually opens with a sentence that essentially communicates "Stop f***ing around."

(participant_id=6)

Grammatical correctness was the most polarizing criterion: 43% believed errors indicated human authorship while in reality, human-authored stories were less likely to contain such errors.

...there are a lot of (usually incorrectly used) semi-colons, which is an error I see human authors make, so I'm more inclined to think this was written by a human...

(participant_id=6)

Formulaic "lessons learned" were correctly associated with LLM-authorship.

The story seems very automated and there is no stylistic variance. There's also a "lesson to be learned" aspect at the end of the piece.

(participant_id=4)

The use of generic character names was frequently used to correctly identify LLM authorship.

"... Marcus is also not a name that strikes fear, so I wouldn't use it for a Vampire..."

(participant_id=7)

A.2.1 Impact of Prompting Strategies

Table 5 shows the impact on mean PDS ratings when switching from the simpler WRITERPROFILE approach to PLAN+WRITE.

A.3 Study Details

This study was reviewed by an IRB and determined to be "exempt" due to the nature of human involvement — i.e. rather than collecting information about the participants themselves, their involvement focused on story annotations.

We also present screenshots of the tutorial instructions (Figure 4) and the general layout of fields collecting annotations on each story (Figure 5).

I awoke in a place unfettered by temporal bounds, immersed in the sweet scent of honeysuckle, the cool cradle of eternity lapping gently at my senses. The silhouettes of people exchanging warm smiles in the distance, their laughter filling the air with a tune as pleasant as a Mozart symphony. I took in the scene, wrapping myself in their ambient joy, a tranquil sense of wonderment kindling within me.

The aesthetics of this realm surrounded me, greeted me with the same warmth as a thousand summer sunsets. Peace and tranquility, stark contrasts to the life I had known, filled each corner. No burdens of earthly existence, no rigor of survival, no ticking clocks etching away life's fleeting moments. The thick stench of rat race, resentment, and unrest were all glaringly absent.

To my query of what benevolent act had entitled me to this serenity, a woman, adorned in the raiment of angelic brilliance, looked puzzled. "Sweet soul, you must have misunderstood," she said gently, "This is indeed Hell, not Heaven."

Scanning my eyes for any hints of jest, they only met with earnest. A tinge of discomfort cloaked her heavenly countenance, a stark contradiction to the utopia that was painted around me. This was hell? This place of serenity, where smiles flowed as readily as wine at a banquet?

The silence was broken by the soft patter of footsteps. A figure approached, draped in the shadows of the ethereal setting, yet there was something oddly familiar about him. As he emerged from the shadows, I recognized the form; a demon, the very embodiment of doom as mortal stories had painted. Yet, there was a strange concern etched in his crimson eyes.

"Ah, a new arrival," he sighed, regret tinging his otherwise sturdy voice. "I should've been the first to see you."

I stood still, my understanding grappling with the oddity of a regretful demon in what seemed to be no less than a celestial utopia. He folded his arms tightly across his chest, leaning in, as if preparing himself to share a mournful secret. The air thickened, the world stopped, and the demon spoke words that sent chills along the spine of my consciousness.

"This is Hell, but not the one humans have created in their minds. Painful, tormented, dreadful - those are earthly concepts tied to flesh and bone. Here in the ethereal world, we trade only in the currency of emotions. Regret is our eternal bondage, the luxury of amending past mistakes forever out of reach. Unresolved conflict, unexpressed emotions, unfulfilled relationships - this mirage of Paradise is merely the canvas for your eternal regret."

In this place, free from every strain and stress, there was but one agony that would echo throughout eternity - the pain of what could have been. As his words hung heavy in the air, my paradise was tainted with the bitter taste of longing. I looked around again, seeing the remorse etched subtly in the smiles.

Underneath the veils of tranquility, I uncovered the true face of Hell. No fire, no brimstone, just an endless expanse of sought redemption denied a lifetime too late. Here, everyone was awash in the glow of their disappointments, even the demon himself - Hell's guardian, and its most regretful inhabitant.

Listing 3: With an average depth score of 4.4 out of 5, this story generated by GPT-4 with the WRITERPROFILE strategy was the most highly rated in the study. The story adheres to the premise and evinces depth both psychologically and philosophically.

Premise: You are severely depressed and are given a service dog to help you through it. However, due to a mixup, you are given a dog that is actually much more depressed than you. The main thing that gets you up in the morning is knowing that you need to be the service human for your dog.

Task: Based on the premise, describe the names and details of 2-3 major characters. Focus on each character's emotional states and inner thoughts.

Only respond with the characters' names and descriptions.

Listing 4: PLAN+WRITE's Character Portraits Prompt with an example premise 14.

1. Samuel Price: Samuel, mid-thirties with disheveled brown hair and tortured blue eyes, is the primary character who is dealing with severe depression. Once a lively and creative soul, his world is now tinged with unending shades of grey. Every morning, he wakes up to the same feeling of emptiness and despair that often overpowers him into a motionless state. Thoughts of unworthiness and self-loathing are constant companions that he fails to ward off. His world has fallen silent, even amidst the chaos. However, the unexpected arrival of a creature that reflects his internal struggles so perfectly lights a flickering flame of purpose in his heart.
2. Charlie: Charlie is a coal-black Labrador Retriever with soulful eyes that seem to mirror much more profound grief than any canine could be expected to bear. The furrow of his brows and his constant whimper tells a story of incessant inner turmoil that somehow seems to overshadow that of Samuel's. He barely eats, barely plays - his spirit is shattered, his life force dimmed. His existence represents an amplification of Samuel's struggles, magnifying them into something entirely unbearable, yet strangely relatable. But beneath the apparent desolation, Charlie's thoughts, even in their depressing state, are loyal towards Samuel, ensuring him that he is not alone in his battle.
3. Dr. Lorraine Fisher: Dr. Fisher is the psychiatrist who ends up mixing up the paperwork, leading to Charlie's arrival into Samuel's life. Being middle-aged, she is a seasoned professional with honey blonde hair that radiates a warmth contradicting her often cold and methodical approach to her practice. She is smart, undeniably good at her job, but she hides behind an aloof persona to separate her personal emotions from her profession. Her primary struggle is dealing with the guilt of the mix-up while observing the peculiar bond growing between Samuel and Charlie. Her thoughts grapple with the ethical implications of her mistake and the unexpected positive outcome it has had on Samuel's otherwise stagnant therapy.

Listing 5: Example PLAN+WRITE's character portraits output for premise 14.

Model	AUTH	EMP	ENG	PROV	NCOM	Model Average
GPT-4	0.03	-0.04	-0.16	0.01	-0.01	-0.03
Llama-2-13B	-0.02	0.00	-0.07	0.00	-0.02	-0.02
Llama-2-70B	0.09	0.06	0.04	0.06	0.11	0.07
Llama-2-7B	-0.04	0.04	-0.05	0.12	0.08	0.03
Vicuna-33B	0.03	0.04	0.01	-0.03	0.08	0.03
Component Average	0.02	0.02	-0.04	0.03	0.05	0.01

Table 5: Percent change in mean ratings when switching from WRITERPROFILE prompting to PLAN+WRITE prompting, aggregated by model and component. For example, PLAN+WRITE decreased Engagement for GPT-4 by 16% relative to using WRITERPROFILE.

In the frosty prison of his depression, Samuel Price found himself waking every morning to the same haunting emptiness. His world, once filled with color, was now a relentless parade of greys. His life had become a hushed echo, the silence broken only by the relentless whispers of unworthiness that shrieked in his ears.

Samuel's doctor, Dr. Lorraine Fisher, proposed a solution to this monotony - a service dog, a beacon of unwavering loyalty and love. Her efficient hands made a mistake, however, sending not a helper but an echo, a manifestation of Samuel's own depressions in the form of a melancholy Labrador named Charlie. Charlie wandered into Samuel's colorless world, carrying shadows deeper than his coal-black fur.

For the first time in years, a jolt of feeling swept through Samuel. His heart wept at the sight of Charlie, his quiet whines and downcast gaze reflecting his own despair. Charlie was Samuel in dog form, only his depression was magnified, laid bare for the world to see.

The first mornings were unsettling, the despair now tangible in the home. Samuel sat with Charlie, their mutual misery fusing into a gloomy camaraderie. But then came a slow and unexpected shift. Despite his depression mirroring Samuel's, Charlie still managed to wag his tail when Samuel came around. A spark of something, perhaps recognition, perhaps solidarity, perhaps care, stirred within Samuel's icy heart. His morning ritual shifted from waking up to the complex hell of his thoughts to getting up because he knew Charlie needed him. He was not curing Charlie's pain; he was mere company to it, like Charlie was to his.

A mix-up, thought Dr. Fisher, ironing the wrinkle from her pristine form. The clinical brilliance that often sparkled in her eyes was dulled by guilt. She'd promised help, not a measly reflection of his agony. But when she saw them together, their empathetic companionship and understanding, she was taken aback. The mix-up had resulted in an accidental therapy, a living breathing support system for each other. Ethical implications marred her relief, but Samuel's progress was undeniable.

Burdened yet elated by Charlie's presence, Samuel started taking small steps towards overcoming his depression. Countless days faded into each other as he made breakfast for Charlie, played fetch on his good days, and just sat by his side on bad ones. Charlie, in return, was there, silent but supportive, sharing his misery with faithful companionship. Each day was a mirror to the next, but in the shared melancholy, they found comfort, companionship, and a reason to wake up each morning.

The afflicted Labrador was no cure, but his mutual agony somehow validated Samuel's despair. A sense of purpose grew from this shared pain, making each moment bearable. Samuel and Charlie became a beacon of hurt but also healing, a reflection of human-dog companionship's deeply emotional potential, and a testament to the shared strength beneath depression's cold, grey waves.

One morning, Samuel smiled at Charlie, love whirling within the hollows of his exhausted heart. "We are mirrors, buddy," he said, "Reflecting each other's darkest parts and, somehow, making them a little brighter." In response, Charlie looked up, his mournful eyes holding a trace of warmth. His tail wagged, slow but sure - a kiss of hope in their shared gloom.

Their tale was an unconventional therapy, draped in the dark cloak of depression. Yet, in the end, it was the darkness that guided them towards the light.

Listing 6: GPT-4 generated story of PLAN+WRITE for an example premise 14. The story follows the given promise and character portraits and is overall coherent.

A.4 Relationship between PDS Components

When selecting which components to include in the Psychological Depth Scale, we attempted to maximize coverage while minimizing semantic overlap in the concepts. Intuitively, it is very likely that increases in one component will affect the perception of others, either incrementally or even as prerequisites. For instance, a narrative perceived as authentic may more easily engage readers and evoke empathy, leading to a more profound emotional response. To quantify these relationships more crisply, we conducted a correlational analysis between each component based on the ratings collected from human study participants. Figure 6 shows that PDS components often co-occur or influence one another. This might indicate that psychological depth is a composite construct, where its components are not entirely distinct but are facets of a broader, interconnected experience.

Short Story Annotation Warm-Up Task

Welcome to our short story annotation warm-up task! Your insights will help us understand stories from a unique perspective.

Purpose: We aim to gauge the psychological depth of short stories by looking at specific components. Additionally, we want to understand the likelihood of a story being AI or human generated. Your annotations will be invaluable in this exploration.

Example Annotation for Calibration:

Before we dive into the detailed instructions, let's walk through an example to set the stage:

Read the Story:

Begin by reading the story carefully. Take your time to understand the narrative, characters, and overall mood.

Rating Components of Psychological Depth:

After reading, you'll rate the story on a scale of 1-5 (1 being the lowest and 5 being the highest) for the following components of psychological depth. We provide some questions for each component to better understand the task.

Authenticity:

- Does the writing feel true to real human experiences?
- Does it represent psychological processes in a way that feels believable?

Emotion Provoking

- How well does the writing depict emotional experiences?
- Does it explore the nuances of the characters' emotional states, rather than just describing them in simple terms?
- Can the writing show rather than tell a wide variety of emotions?
- Do the emotions that are shown in the text make sense in the context of the story?

Empathy

- Do you, as the reader, feel like you were able to empathize with the characters and situations in the text?
- Do you, as the reader, feel that the text led you to introspection, or to new insights about yourself or the world?

Engagement

- Does the text engage you on an emotional and psychological level?
- Do you feel the need to keep reading, as you read the text?

Narrative Complexity

- Do the characters in the story have multifaceted personalities? Are they developed beyond stereotypes or tropes? Do they exhibit internal conflicts?
- Does the writing explore the complexities of relationships between characters?
- Does it delve into the intricacies of conflicts and their partial or complete resolutions?

Dos and Don'ts

- **Dos:**
 - Be honest with your ratings. There are no right or wrong answers.
 - If you're unsure about a component, it's okay to give a middle score.
 - Trust your instincts. Your first impression is often the most valuable.
- **Don'ts:**
 - Don't rush through the reading or the rating process.
 - Avoid being influenced by external factors. Try to judge the story on its own merits.
 - Refrain from skipping components. If unsure, give your best guess.

Figure 4: Screenshots taken from the Warm-Up tutorial instructions shown to study participants. All fields are similar to the ones used in the main annotation forms.

Participant ID *
Short answer text

To what degree is the story **authentic**? *
1 2 3 4 5
Implausible ☐ ☐ ☐ ☐ ☐ Undeniably Real

Comments on authentic elements within the story (optional)
Long answer text

To what degree is the story **engaging**? *
1 2 3 4 5
Unengaging ☐ ☐ ☐ ☐ ☐ Captivating

Comments on engaging elements within the story (optional)
Long answer text

To what degree does the story evoke **empathy**? *
1 2 3 4 5
Detached ☐ ☐ ☐ ☐ ☐ Deep Resonance

Comments on empathetic elements in the story (optional)
Long answer text

To what degree does the story **provoke emotion** of any kind? *
1 2 3 4 5
Unmoving ☐ ☐ ☐ ☐ ☐ Highly Emotional

Comments on emotion provoking elements in the story (optional)
Long answer text

To what degree is the story **narratively complex**? *
1 2 3 4 5
Simplistic ☐ ☐ ☐ ☐ ☐ Intricately Woven

Comments on narrative complexity in the story (optional)
Long answer text

How likely do you think the story was written by a human or an LLM? *
1 2 3 4 5
Very Likely Human ☐ ☐ ☐ ☐ ☐ Very Likely LLM

Comments on how you determined human vs. LLM authorship (optional)
Long answer text

Open Feedback
(Use this field for miscellaneous feedback on the story as a whole)
Long answer text

Figure 5: Screenshots showing how annotations were collected for each story using Google Forms.

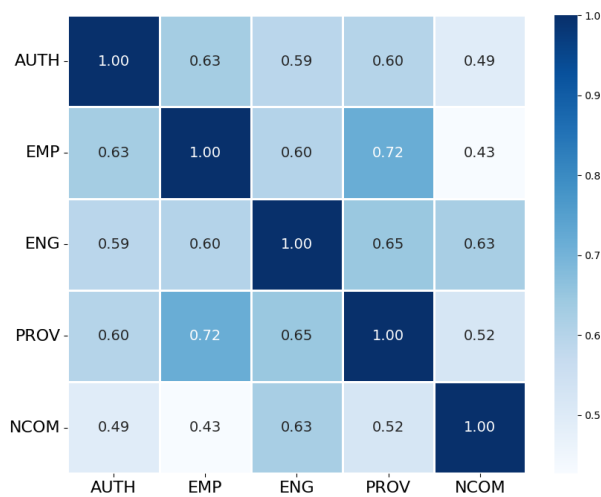


Figure 6: Pearson correlations between all PDS components derived from human ratings. All correlations were statistically significant with p-values < 0.05.