Whose story is it?
Personalizing story generation by inferring author styles

Anonymous ACL submission

Abstract

Personalization is critical for improving user experience in interactive writing and educational applications, yet remains understudied in story generation. We study the task of personalizing story generation, where our goal is to mimic an author's writing style, given other stories written by them. We collect ***Mythos***, a dataset of 3.6k stories from 112 authors, with an average of 16 stories per author, across five distinct sources reflecting diverse story-writing settings. We propose a two-stage pipeline for personalized story generation: first, we infer authors' implicit writing characteristics and organize them into an *Author Writing Sheet*, which is validated by humans to be of high quality; second, we simulate the author's persona using tailored persona descriptions and personalized story rules. We find that stories personalized using the Author Writing Sheet outperform a non-personalized baseline, achieving a 78% win-rate in capturing authors' past style and 59% in similarity to ground-truth author stories. Human evaluation supports these findings and further highlights trends, such as Reddit stories being easier to personalize, and the Creativity and Language Use aspects of stories being easier to personalize than the Plot[1].

1 Introduction

Large Language Models (LLMs) are increasingly used in interactive systems to support human users in co-writing tasks (Fang et al., 2024; Yuan et al., 2022; Chung et al., 2022). A key challenge in these settings to generate personalized outputs that align with individual user preferences (Yeh et al., 2024), which improves engagement, agency, and writing quality (Zhao et al., 2025). Personalization is also key in educational applications to deliver effective feedback to students (Anderson et al., 1985), and may support users facing writer's block (Yuan et al., 2022) or help second language learners in learning a new language (Baffour et al., 2023; Su et al., 2023). Despite these potentials, personalization in story writing has received limited attention.

Story-writing personalization faces several challenges, including a lack of existing resources for evaluating personalization, the inherent subjectivity in user preferences, and methods to evaluate the extent of personalization. To address these gaps, we construct ***Mythos***, a dataset of 3.6k human-written stories (≈1500 tokens each) from 112 authors (16 stories each) across five distinct writing settings: Reddit, AO3, Storium, Narrative Magazine, and New Yorker. We split each author's stories chronologically into two sets: an earlier *Profiling* set used to infer their writing characteristics, and a later *Generation* set used to evaluate how well LLMs can emulate their distinctive writing behavior (Salemi et al., 2024).

We systematically explore personalized story generation using LLMs with a two-stage pipeline as shown in Figure 1: First, we construct an *Author Writing Sheet* to extract implicit writing characteristics from an author's past stories (Ramos et al., 2024). The Author Writing Sheet covers four narrative dimensions: Plot, Creativity, Development, and Language Use (Pavis, 1998; Huot et al., 2024), and is organized as Claim-Evidence pairs grounded in story excerpts to summarize author preferences. This representation, inspired by Common Core Standards (Association et al., 2010) and writing education research (Li et al., 2023), offers a cost-effective and interpretable way to model authors' writing preferences. Second, we guide LLMs to role-play authors using persona prompts and story rules derived from the Author Writing Sheet.

We evaluate generated personalized stories in two ways: First, we use automatic evaluation with LLM-as-a-judge to assess faithfulness to the author's writing history and similarity to ground truth author story, along with traditional text similarity

---

[1] Code and data would be released after the double-blind submission process.

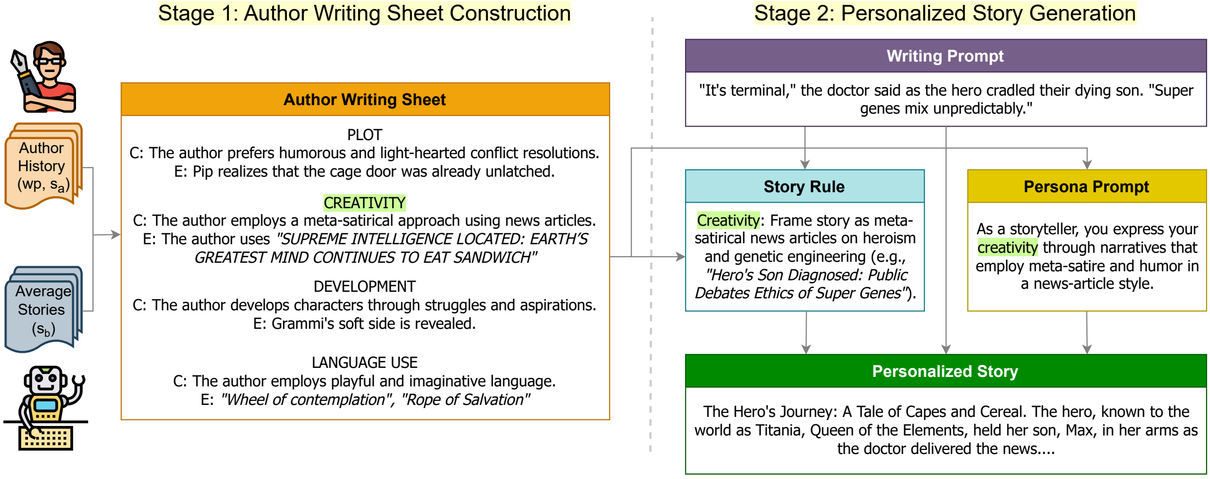

Stage 1: Author Writing Sheet Construction
Stage 2: Personalized Story Generation
Author History (wp, s_a)
Average Stories (s_b)
Author Writing Sheet
PLOT
C: The author prefers humorous and light-hearted conflict resolutions.
E: Pip realizes that the cage door was already unlatched.
CREATIVITY
C: The author employs a meta-satirical approach using news articles.
E: The author uses "SUPREME INTELLIGENCE LOCATED: EARTH'S GREATEST MIND CONTINUES TO EAT SANDWICH"
DEVELOPMENT
C: The author develops characters through struggles and aspirations.
E: Grammi's soft side is revealed.
LANGUAGE USE
C: The author employs playful and imaginative language.
E: "Wheel of contemplation", "Rope of Salvation"
Writing Prompt
"It's terminal," the doctor said as the hero cradled their dying son. "Super genes mix unpredictably."
Story Rule
Creativity: Frame story as meta-satirical news articles on heroism and genetic engineering (e.g., "Hero's Son Diagnosed: Public Debates Ethics of Super Genes").
Persona Prompt
As a storyteller, you express your creativity through narratives that employ meta-satire and humor in a news-article style.
Personalized Story
The Hero's Journey: A Tale of Capes and Cereal. The hero, known to the world as Titania, Queen of the Elements, held her son, Max, in her arms as the doctor delivered the news....

Figure 1: Our two-stage pipeline for personalized story generation. Stage 1 constructs an Author Writing Sheet with Claim (C) and Evidence (E) pairs capturing the author's story-writing characteristics across narrative categories. It is derived from the author's history of writing prompts (wp), author-written stories (s_a), and LLM-generated Average Stories (s_b) representing a typical author's response to the same prompt. Stage 2 uses the Author Writing Sheet to role-play the author, incorporating tailored story rules and a persona description for personalized generation.

metrics on lexical overlap, story diversity, and style similarity. Second, we conduct a human evaluation to assess the generated story's similarity to the author's actual writings.

We find that stories generated using the Author Writing Sheet outperform a non-personalized baseline: LLMs prefer them 78% of the time in faithfulness and 59% in similarity to ground truth, while human annotators prefer them 56% of the times for similarity to ground truth. Our method is especially effective for Reddit stories and in capturing Creativity and Language Use, with a smaller impact on Plot. In summary, our contributions are:

- ***Mythos***: A dataset of 3.6k stories with multiple stories per author, supporting both author profiling and personalized story generation evaluation.
- **Author Writing Sheet**: A cost-efficient, structured method for capturing author characteristics, grounded in narrative theory and validated through human evaluation.
- **Personalization Evaluation**: An evaluation combining automated and human assessments, showing the benefits of Author Writing Sheets.

2 Dataset Collection and Description

To facilitate our analysis of *personalized* story generation, we assemble ***Mythos***, a dataset consisting of 3.6k stories written by 112 authors. To the best of our knowledge, ***Mythos*** is the first dataset that identifies and connects multiple stories written by the same author (see Table 1 for comparison).

Data Sources: We include five diverse story-writing sources: **(1) Reddit**, featuring stories from r/WritingPrompts, a widely-used resource for amateur story-writing research (Fan et al., 2018), forming the largest subset with 3.2k stories, **(2) AO3**, featuring fanfiction based on popular franchises like *Harry Potter*, **(3) Storium**, featuring collaborative stories (Akoury et al., 2020), **(4) N.Mag** (Narrative Magazine), featuring professional stories from authors like Barry Gifford, and **(5) N.York** (New Yorker), featuring expert-level storytelling from authors like Haruki Murakami.

Preprocessing: To ensure high-quality and diverse content, we impose constraints on story length (ranging from 500 to 1500 words), limit publication dates to post-November 2022, and exclude stories with explicit content. To standardize formatting, non-Reddit stories are supplemented with writing prompts generated by GPT-4o, which are manually reviewed for accuracy. Stories are split chronologically, with 70% used for *profiling* writing characteristics and 30% for testing *generation*. Details on our dataset can be found in Appendix A.

3 Author Writing Sheet

Motivated by the connection between writing education and personalization (Li et al., 2023), we propose a method to infer authors' implicit story-writing tendencies based on their past stories in the profiling set, denoted as P. We group these

Dataset Name	Size	Sources	Prompt Length	Story Length	Author IDs
WritingPrompts (Fan et al., 2018)	300k	Reddit	28	735	✗
TELL ME A STORY (Huot et al., 2024)	230	Writing Workshop	113	1498	✗
MirrorStories (Yunusov et al., 2024)	1500	Aesop's fables	40	400	✗
Storium (Akoury et al., 2020)	440k	Storium online platform	247	247	$\sim$
Mythos (ours)	3.6k	Reddit, AO3, Storium, N.Mag, N.York	50	1517	✓

Table 1: Comparison of our dataset with existing story-writing datasets. ✗ indicates no Author IDs associated with stories, and $\sim$ denotes having Author IDs but not having explicit links between them. ***Mythos*** covers diverse story-writing settings and links stories by the same author, enabling future research on personalized story generation.

tendencies into four narrative categories: Plot, Creativity, Development, and Language Use inspired by narrative theory (Pavis, 1998; Huot et al., 2024), to capture fine-grained, multi-dimensional aspects of how authors approach story-writing:

- **Plot**: Story structure, conflict introduction, prompt engagement, and resolution.
- **Creativity**: Genre blending, unconventional prompt reinterpretation, and unique elements.
- **Development**: Character depth, emotional arcs, and immersive settings.
- **Language Use**: Diction, style, rhetorical devices, pacing, and dialogue.

Inspired by the Common Core Standards in English Language Arts (Association et al., 2010), each characteristic is represented as a Claim-Evidence pair, where the Claim summarizes the author's style and the Evidence are supporting story excerpts.

Stage 1 in Figure 1 (and Algorithm 1 in the Appendix) illustrates our method for generating the Author Writing Sheet (example in Table 7), where we iteratively process each story in an author's profiling set. We construct the sheet by merging characteristics across stories that highlight how the author's style differs from that of a typical, average author. For each writing prompt (wp_t) and its corresponding author story (s_{at}), we use the LLM with two distinct prompts: one prompt (LLM_{avg}) elicits an average story (s_{bt}) that reflects the typical style of an author on that writing source, relying solely on source-specific priors acquired during pre-training; another prompt ($\text{LLM}_{\text{sheet}}$) asks the same LLM to compare the author story (s_{at}) with the average story (s_{bt}) and extract corresponding Claim-Evidence pairs across narrative categories. This yields an intermediate Author Writing Sheet A'_t (Shashidhar et al., 2024; Krishna et al., 2020).

To construct the final Author Writing Sheet $A_{|P|}$, we iteratively merge each intermediate sheet A'_t with the aggregated sheet A_{t-1}, following a moving update approach (Chang et al., 2023). To reduce redundancy and manage output length, we prompt the LLM with a prompt ($\text{LLM}_{\text{combine}}$) to group equivalent Claims, select the best Evidence, and retain ungrouped Claims. We further limit each narrative category to 10 Claim-Evidence pairs prioritizing grouped Claims to fit within the LLM's context window in subsequent iterations. Each Evidence is tagged with a timestamp indicating its source story.

Similar to *Knowledge Tracing* in education (Liu et al., 2022; Scarlatos and Lan, 2024), our method processes each story sequentially. Doing so enables efficient updates to the Author Writing Sheet without having to re-process prior stories every time a new story is analyzed (Yeh et al., 2024). Therefore, this method reduces cost and mitigates long-context limitations in LLMs (Li et al., 2024); see Appendix B.1 for a detailed cost analysis.

Validating Author Writing Sheet Quality: We conduct a human evaluation with three Upwork annotators to assess the quality of the Claim-Evidence pairs. Annotators evaluate each pair on two criteria: (1) whether the Claim can be reasonably inferred from the story (Yes/No), and (2) whether the Evidence supports the Claim (Yes/Partially/No). The study spans 42 authors across five sources, with one sampled story per author, yielding 188 annotated claims. To measure inter-rater reliability, 12 stories are annotated by all annotators. Results show unanimous agreement on the 'Yes' label of Claims and moderate agreement on Evidence support (Krippendorff's $\alpha = 0.57$), despite the task's subjectivity. Overall, 93% of Evidence statements fully support their Claims, and the remaining 7% provide partial support. These results indicate *high precision* in constructing Claim-Evidence pairs and confirm the reliability and quality of the Author Writing Sheet; see Appendix B.2 for more details including experimental design and interface (Figure 13), pilot studies, and results.

4 Personalized Story Generation

We now detail our proposed method for personalized story generation using the Author Writing Sheet, illustrated in Stage 2 of Figure 1. To generate personalized stories, we prompt an LLM ($\text{LLM}_{\text{story}}$) with the writing prompt, story length, source-specific metadata (e.g., fanfiction for AO3), story rules as actionable instructions in direct second-person form categorized into four narrative categories (example in Table 8), and persona descriptions obtained from the Author Writing Sheet (example in Table 9). We experiment with four types of methods: (1) A non-personalized baseline, *Average Author*; (2) Personalization baselines, *RAG* and *Delta*, that do not use the Author Writing Sheet; and (3) Our proposed methods, *Sheet* and its variant *Summary*. (4) An *Oracle* method that loosely resembles the upper bound performance; See Appendix C for the prompts used.

Non-Personalized: Average Author The *Average Author* method serves as a non-personalized baseline reflecting typical author behavior learned during LLM pre-training. For each source, we prompt an LLM with an Average Author description summarizing writing characteristics obtained through audited GPT-4o prompting,[2] verified manually (Wang et al., 2024). Example, for AO3, the Average Author is defined as “respecting fandom tone and style while experimenting with tropes, unconventional pairings, and alternate universes.”

Personalization Baselines: RAG and Delta

RAG The Retrieval-Augmented Generation (*RAG*) baseline (Salemi et al., 2024) first retrieves the most similar writing prompt and author story from the profiling set using BM25 (Robertson et al., 2009). Following (Wang et al., 2024), the retrieved pair is used as a one-shot demonstration to elicit role-playing behavior from the LLM, mimicking the retrieved example’s style.

Delta The *Delta* method generates personalized story rules by contrasting the *Average Author* story (see Average Author) with the corresponding author-written story for each writing prompt in the profiling set. Following (Shashidhar et al., 2024), we then use *all* writing prompts in the profiling set, along with their corresponding generated story rules that are actionable instructions in direct second-person form, as few-shot demonstrations for the LLM, to generate personalized story rules for a new prompt in the generation set.

Our Methods: Sheet and Summary

Sheet Our *Sheet* method uses the Author Writing Sheet (Section 3) for personalization (Stage 2 of Figure 1). First, we prompt an LLM ($\text{LLM}_{\text{persona}}$) to generate a persona description that summarizes the author’s story-writing style as a second-person narrative, included in the system prompt (Wang et al., 2024; Jiang et al., 2024). Second, we prompt an LLM (LLM_{rule}) to generate personalized story rules from the Author Writing Sheet tailored to the writing prompt in the generation set, included as constraints in the user prompt (Pham et al., 2024) for $\text{LLM}_{\text{story}}$. Additionally, following (Richardson et al., 2023), we include a one-shot demonstration using the approach described in RAG.

Summary (Summ) As a variant of the Author Writing Sheet, the *Summ* method leverages the LLM’s long-context capabilities (Ding et al., 2023) by packing all past stories from the profiling set as input in the same prompt to generate an *Author Writing Summary* in the same format as the Author Writing Sheet. Similar to the Sheet method, persona descriptions and story rules are derived from the Author Writing Summary and used as constraints for story generation, along with a one-shot demonstration.

Reference Method: Oracle The *Oracle* method establishes an upper bound on personalization performance by using story rules derived directly from the ground-truth author story for each writing prompt. These rules are obtained by contrasting the Average Author with the ground-truth story (see Delta). We also include a one-shot demonstration following the approach described in RAG.

5 Experiments

We now detail our experimental/evaluation settings.

5.1 Implementation Details

We prompt GPT-4o in a chain-of-thought manner (Wei et al., 2022) with temperature 0 and a maximum token limit of 4096 for all the methods described above (Section 4) to generate story rules

---

[2] We ask GPT-4o for typical writing characteristics for each source and manually verify them.

and persona descriptions. Only for story generation ($\text{LLM}_{\text{story}}$) across all methods, including *Average Author*, we evaluate three different models: GPT-4o[3] (Bubeck et al., 2023), Llama 3.1 8B, and Llama 3.1 70B (Dubey et al., 2024), using a temperature of 0.7 and top_p of 0.95 (Wang et al., 2024) (See Appendix F for results using Llama models). Additionally, we implement an *ablation* variant of our proposed method for Sheet and Summ that excludes the persona description from the system prompt called *Sheet-nP* and *Summ-nP*, respectively.

5.2 Automated Evaluation

We evaluate the performance of personalized story generation using two automated evaluation methods: First, we use LLM-as-a-judge, which is a scalable alternative to human evaluation for complex open-ended text generation tasks (Zheng et al., 2023) and calculate the **win-rate** of a story generation method against the non-personalized baseline, *Average Author*, on two aspects:

- **Faithfulness to Writing History**: We evaluate how well the generated story aligns with the author history by measuring win-rates using Claims for each narrative category from the Author Writing Summary as the reference (Wang et al., 2023; Yunusov et al., 2024).
- **Similarity to Author Story**: We assess how closely the generated story matches the author's ground-truth story by measuring win-rates using the ground-truth author story as the reference for each narrative category (Lyu et al., 2024; Shashidhar et al., 2024).

For both aspects, we prompt OpenAI o4-mini[4] in a chain-of-thought manner (Wei et al., 2022) to assign scores (1–5) to both stories. We randomly shuffle the order of stories in comparison, to avoid biases. Scores are broken down into four narrative categories: Plot, Creativity, Development, and Language Use (Saha et al., 2024). For each category, the story with the higher score is declared the winner, and the overall winner is based on the highest total score across all categories; two stories having the same total scores result in a tie. See Appendix D.1 for the prompts. Additionally, we evaluate personalization using *traditional metrics* (Xie et al., 2023), including lexical overlap, story diversity, and stylistic similarity. See Appendix D.2 for details.

Method	Reddit	AO3	Storium	N.Mag	N.York	*All*
RAG	40	52	42	21	40	39
Delta	56	60	50	64	53	57
Sheet	74	76	73	85	80	78
Sheet-nP	74	80	60	85	86	78
Summ	84	96	87	85	80	*87*
Summ-nP	76	86	85	100	100	*90*

Table 2: Win-rates (%) for Faithfulness to Writing History vs. Average Author baseline for GPT-4o. 'Sheet' (underlined) performs close to the upper bound established by the 'Summ' method and significantly outperforms other personalization methods.

6 Results and Discussion

We now present experimental results. Our automatic evaluation shows that the proposed methods are effective, particularly for certain sources and narrative categories.

6.1 Faithfulness to Writing History

Table 2 shows results for faithfulness to writing history among GPT-4o-personalized stories. We observe the following:

Sheet and Summ achieve best faithfulness to writing history: The high performance of Summ sets a performance upper bound, since it relies on the *Author Writing Summary*, the reference for evaluation, to generate the story. Sheet, despite not being explicitly conditioned on the Author Writing Summary, generalizes well (only a 9% gap in win rate compared to Summ), which shows that it is capable in faithfully reflecting the author's writing history. RAG performs the worst since it relies on a single in-context example to mimic the author's style. Delta significantly outperforms RAG (by 18%) but significantly underperforms Sheet (by 21%). This result suggests that explicitly summarizing authors' stylistic characteristics systematically in the form of a Sheet is better than relying on rules extracted from all past author stories. See Appendix E.1.1 and Table 10 for an example.

Persona descriptions benefit amateur sources more than expert sources for Summ: We see that personas have little impact on Sheet, since the overall win-rates for Sheet and Sheet-nP are identical. This result is likely attributed to using the Author Writing Summary, rather than the Sheet, as the reference for faithfulness evaluation. In contrast,

---

[3] For Reddit, we only consider 5 stories per author for generation totaling 300 stories, due to API cost constraints.

[4] `https://openai.com/index/introducing-o3-and-o4-mini/`

Method	Reddit	AO3	Storium	N.Mag	N.York	*All*
Oracle	86	84	78	79	87	83
RAG	38	56	**50**	43	27	43
Delta	41	60	48	43	33	45
Sheet	**62**	**66**	40	**71**	**54**	**59**
Sheet-nP	47	50	33	64	33	45
Summ	45	50	38	50	40	45
Summ-nP	43	53	45	64	20	45

Table 3: Win-rates (%) for Similarity to Author Story vs. Average Author baseline for GPT-4o. Best methods per source are **bolded**.

for Summ, adding persona descriptions improves win-rates over Summ-nP for amateur sources such as Reddit, AO3, and Storium, but leads to lower performance for expert sources like the N.Mag and N.York. This result is likely due to personas for expert authors failing to capture the nuanced and subtle characteristics of their writing styles, which are reflected in their Author Writing Summary. See Appendix E.1.2 and Tables 11, 12 for examples.

6.2 Similarity to Author Story

Table 3 shows results on similarity to the author story among GPT-4o-personalized stories. We observe the following:

Sheet outperforms all other methods: The Sheet method achieves the highest overall win-rate of 59%, outperforming Summ by 14%, Delta by 14%, and RAG by 16%. The advantage is more pronounced on Reddit, by far the largest subset of our dataset, where Sheet outperforms Summ by 17%, Delta by 21%, and RAG by 24%. These results suggest that relying on a single in-context example (RAG) is insufficient for capturing personalized writing styles, particularly when authors have more varied profiles, as in Reddit. In contrast, for sources like Storium, where authors have fewer stories and less consistent styles, RAG is more competitive. The Delta method, which uses only story rules, struggles to capture subtle, repeated writing patterns that are better modeled through the longer and more detailed Author Writing Sheet. Finally, Summ's approach of identifying writing characteristics from all profiling stories at once, without contrasting against Average Author stories, is limited by LLMs' long-context reasoning capabilities (Shashidhar et al., 2024). See Appendix E.2.1 and Table 13 for an example.

Sheet still lags behind the upper-bound Oracle method: Although the Sheet method performs best, there remains a substantial overall gap (of 24%) compared to the Oracle method, which highlights the challenging nature of personalization in story generation. This gap likely arises because the Sheet cannot capture *all* of an author's writing characteristics within the profiling set, due to generation length constraints (4096 tokens). Moreover, the Sheet does not explicitly rank characteristics based on their relevance to a new writing prompt, which limits its ability to adaptively emphasize the most salient aspects during generation. See Appendix E.2.2 and Table 14 for an example.

Persona descriptions benefit Sheet: We see that Sheet outperforms its 'nP' variant overall by 14%, with a particular gain of 15% on Reddit, the largest subset of our data. Persona descriptions, provided in the system prompt, offer additional control for personalized story generation by summarizing the author's general writing style independently of the specific writing prompt, complementing the story rules used in the user prompt. See Appendix E.2.3 and Table 15 for an example. In contrast, persona descriptions have limited impact on the Summ method, likely because the Author Writing Summary is less effective than the Author Writing Sheet at capturing subtle author characteristics.

Method	Plot	Creativity	Development	Language Use
Oracle	74	**82**	69	**88**
Delta	44	**55**	43	**57**
Sheet	49	**74**	51	**62**
Summ	44	**51**	50	**51**

Table 4: Win-rates (%) across narrative categories for Similarity to Author Story, averaged across sources. Best numbers for each method are **bolded**.

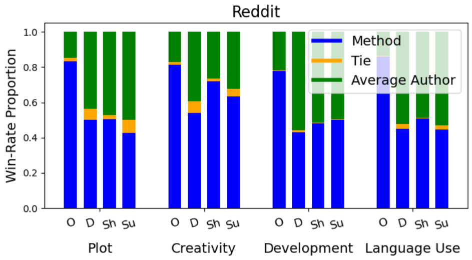


Figure 2: Win-rates proportions across narrative categories for Similarity to Author Story for Reddit. O: Oracle, D: Delta, Sh: Sheet, Su: Summ.

Creativity and Language Use are easier to personalize: Table 4 shows that all methods have higher respective win-rates on Creativity and Language Use than Plot and Development. This result likely arises because Creativity (e.g., unique elements introduced by the author) and Language Use (e.g., specific diction or use of dialogue) are less dependent on the writing prompt and thus more generalizable across stories (Huot et al., 2024). In contrast, Plot (e.g., conflict introduction or resolution) and Development (e.g., character growth) are closely tied to the writing prompt, making generalization difficult, particularly when there is limited thematic overlap (Tian et al., 2024).

Figure 2 shows that, for our largest story source, Reddit, the Sheet method achieves the highest win-rates across all categories except Development, where it is slightly outperformed by Summ. For Creativity, Sheet performs nearly as well as the Oracle, with only a 5% win-rate gap between them. However, substantial gaps remain between Sheet and Oracle for other categories, which highlights the difficulty of the task and room for future improvement. See Appendix E.2.4 and Figure 22 for plots from other data sources.

6.3 Traditional Metrics

Table 16 in the Appendix shows results with traditional metrics. Overall, we see that the lexical overlap and diversity scores are similar across methods, since all generations use the same LLM, resulting in overlapping lexical distributions that limit these metrics' sensitivity to stylistic differences (Zheng et al., 2023; Xie et al., 2023).

We also see that the Sheet achieves the *best stylistic similarity* to both Author History and Author Story, measured by LUAR (Rivera-Soto et al., 2021). This result validates the Sheet's ability to explicitly summarize stylistic deviations from an Average Author, leading to better personalization.

7 Human Evaluation

7.1 Setting

We conduct a human evaluation on a subset of our data to assess narrative subtext and identify insights potentially missed by LLM judges (Chakrabarty et al., 2024; Subbiah et al., 2024b). Annotators compare stories generated by personalization methods (*Delta*, *Sheet*, and *Summ*) against the *Average Author* method on similarity to the ground-truth author story (Lyu et al., 2024). For each comparison, annotators choose an overall winner by evaluating both stories on all four narrative categories, and provide justifications as free-form responses. Three annotators, recruited via Upwork, evaluated 45 author stories in total. There are 15 shared stories across annotators to assess agreement, and 10 were unique to each annotator to expand coverage (Song et al., 2024). Each author story appeared in three different pairs (one per personalization method), resulting in 135 annotated story pairs. See Appendix G for details on experiment design and the interface (Figure 23).

Method	Reddit	AO3	Storium	N.Mag	N.York	*All*
Delta	44-44	33-33	**33**-33	33-33	67-11	42-31
Sheet	**67**-33	**44**-22	22-33	**67**-33	**78**-11	**56**-27
Summ	44-33	33-44	**33**-22	44-22	67-22	44-29
Overall	52-37	37-33	30-30	48-30	70-15	47-29

Table 5: Win-rates (%) for Similarity between GPT-4o-generated and Author stories, evaluated by Humans. Each cell ('X-Y') shows the method's ('X') and Average Author's ('Y') win-rates, with ties as remainder. Best methods per source are **bolded**.

7.2 Results

Table 5 shows results from our human evaluation. Sheet is clearly the best method, outperforming Average Author by 29% across all sources. On Reddit, our largest source, it reaches a 67% win-rate, further confirming its advantage.

Fair inter-rater and model-human agreement: Fleiss' Kappa among human annotators and Cohen's Kappa between human annotations and the LLM judge (o4-mini) both range from 0.2 to 0.4, indicating fair agreement (Landis and Koch, 1977). This result is likely due to the inherent subjectivity of evaluating long-form texts (Subbiah et al., 2024a), where annotators prioritize different narrative aspects when choosing the overall winner. The scale of human evaluation, constrained by annotation costs ($639), also contributes to variability. See Appendix G.3 for detailed examples.

Creativity, and Language use are easier to personalize: Same as LLM-as-a-judge, human annotators also favor personalized stories for Creativity and Language Use, noting their stronger use of deeper symbolism, thematic richness, layered narratives, and expressive language. However, Average Author was preferred when personalized methods incorrectly introduce elements not present in the author's ground-truth story, suggesting deviations from the author's usual style. See Table 22 in the Appendix for detailed analysis of annotator free-form comments in each narrative category.

Reddit and New Yorker are easier to personalize than other sources: Reddit and New Yorker show the highest overall win-rates for the personalized methods (52% and 70%, respectively). Reddit benefits from the broad thematic variety of its crowd-sourced writing prompts (see Figure 3 in the Appendix), which leads to diverse author writing styles and makes authors easier to distinguish from the Average Author (see Table 23). New Yorker, with stories written by experts, benefits from the use of advanced narrative devices such as subtext, which humans can evaluate more reliably than LLMs (Subbiah et al., 2024b) (see Table 24). AO3 and Storium show lower overall win-rates for personalized methods. For AO3, LLMs' familiarity with common fanfiction tropes[5] strengthens the Average Author baseline (see Table 25). For Storium, the small profile size and weak stylistic differentiation among authors beyond open-ended endings (Xu et al., 2024; Tian et al., 2024) limit the effectiveness of personalization methods (see Table 26). These findings suggest that personalization is most effective when authors explore diverse topics or display distinct traits, such as complex narratives, creativity, or advanced language use. Such traits, reflected in past writings, provide clearer signals for personalizing their style.

8 Related Work

8.1 Personalization and Role-Playing

Recent works have introduced benchmark datasets for personalizing LLM outputs in tasks like email and news writing, focusing on shorter outputs (e.g., 300 tokens for product reviews (Kumar et al., 2024), 850 for news writing (Shashidhar et al., 2024)). These methods infer user traits from history for task-specific personalization (Sun et al., 2024, 2025; Pal et al., 2024; Li et al., 2023; Salemi et al., 2025). In contrast, we address the more subjective task of long-form story writing, with author stories averaging 1500 tokens. Unlike prior role-playing approaches that rely on predefined personas (e.g., Tony Stark, Confucius) (Wang et al., 2024; Sadeq et al., 2024; Tu et al., 2023; Xu et al., 2023), we propose a novel method to infer story-writing personas from an author's history to guide role-playing.

8.2 Story Understanding and Generation

Prior work on persona-aware story generation (Yunusov et al., 2024; Zhang et al., 2022; Chandu et al., 2019) defines personas using discrete attributes like personality traits, demographics, or hobbies. Similarly, (Zhu et al., 2023) explore style transfer across predefined domains (e.g., fairy tales, martial arts, Shakespearean plays). In contrast, we mimic an individual author's style based on their history. Our approach differs by (1) inferring long-form author personas descriptions of style from past works, rather than relying on demographics, and (2) handling long-form story generation, averaging 1500 tokens per output, beyond typical lengths in prior work.

9 Conclusions and Future Work

In this paper, we introduced the task of personalizing authors' story-writing styles across five sources by proposing a new dataset, ***Mythos***, containing 3.6k stories. We also proposed a two-stage approach that infers implicit characteristics as a high-quality Author Writing Sheet, to guide LLMs via persona descriptions and story rules. Extensive automated and human evaluation showed that our method outperforms baselines for both faithfulness to the author's history and similarity to the ground-truth author story. However, our methods still significantly lagged behind an Oracle method, by as much as 24% in some ways, which highlights the difficulty of the task and the need for future work. We also found that personalization is more effective for sources with diverse writing prompts and distinctive author traits (e.g., Reddit, New Yorker), and for narrative categories like Creativity and Language Use, which are less prompt-dependent.

Future work can address the limitations of the Author Writing Sheet by improving its coverage of relevant stylistic information despite context length constraints, and by retrieving or re-ranking information relevant to the given writing prompt during generation. Another promising direction is to explore multi-agent systems for story-writing personalization, with specialized agents handling different narrative categories to improve the extraction and application of author stylistic traits especially for categories like Plot and Development that are hard to personalize.

---

[5] `https://archiveofourown.org/admin_posts/25888`

10 Limitations

Limited size of Profiling Set for some sources: Table 6 shows that when using 70% of each author's available stories for profiling (from Stories/Author column), authors from Narrative Magazine and New Yorker have only about 2 stories on average, compared to 5 for Storium and around 10 for AO3. Reddit is the exception, with authors having an average of 38 stories in their profiling set. A larger profiling set likely enables more accurate inference of an author's story-writing characteristics, leading to better personalization. However, our data collection is constrained by practical limitations. For all sources except Reddit, we manually collect publicly available stories without web scraping, which restricts the number of samples we can obtain. Furthermore, for expert sources like Narrative Magazine and New Yorker, fewer stories per author are available online, limiting the scalability of personalized story generation. For Reddit, we adhere to Reddit's data collection policies and use the official API for dataset construction.

Author Writing Sheet Recall: Our method for generating the Author Writing Sheet is limited in its ability to capture *all* of an author's story-writing characteristics. This limitation arises from practical context length constraints and the challenges of long-context reasoning during the iterative merging step, which forces us to restrict the Author Writing Sheet to 10 characteristics per narrative category. Regarding human evaluation, we validate only the precision of the Author Writing Sheet—ensuring that the extracted Claim-Evidence pairs accurately reflect the author's style—but do not assess recall, i.e., whether all relevant characteristics are captured. Measuring recall is challenging due to the inherent subjectivity of the task, making it impractical to ask annotators to independently construct Writing Sheets for comparison.

Multi-Agent Systems and Fine-Tuning LLMs: Our work explores role-playing for personalized story generation by prompting an LLM with tailored persona descriptions and story rules. Future research can extend this by incorporating multi-agent systems, where specialized agents focus on different narrative categories to enhance personalization (Huot et al., 2024; Bae and Kim, 2024), especially for categories like Plot and Development, which are less conducive to personalization. Additionally, while our experiments show that fine-tuning improves personalization over prompting for LLaMA 3.1 8B (Section F.2 and Table 21), our work sets a precedent for using fine-tuning to better align LLMs with individual author preferences. Future work can build on this by incorporating explicit reasoning over author history to further improve personalization (Salemi et al., 2025; Shashidhar et al., 2024; Shaikh et al., 2024).

Subjectivity in Evaluation: Evaluating long-form creative text remains challenging for LLMs due to the involvement of nuanced aspects like subtext (Subbiah et al., 2024a). Despite this subjectivity, we obtain fair Cohen's Kappa agreement for both human-human and human-model evaluations. Future work can explore improving evaluation methods by using aspect-based checklists that capture finer-grained narrative details (Lee et al., 2025).

11 Ethical Considerations

Human Evaluation: Both human evaluation tasks were approved by an institutional review board (IRB). All annotators, who are US-based and fluent in English, were informed of the nature of the research study, to which they provided informed consent, and were compensated at an hourly rate of $17, meeting the minimum wage requirements in our state. All scientific artifacts, including models and datasets, were used in accordance with their intended purpose to ensure ethical and responsible research practices.

Dataset Collection: All data used in this study were manually collected from publicly available sources,[6] adhering to the data usage and crawling policies of the respective websites. For Reddit, we collected data exclusively using the official Reddit API, following Reddit's data access policies.[7] To comply with copyright constraints, we will release only the URLs linking to the original stories, rather than the story texts themselves, following the approach of (Chakrabarty et al., 2024).

Data Anonymization and Licensing: Fully anonymizing our dataset is not feasible,[8] as the research involves mimicking specific authors' story writing styles. Even if usernames were

---

[6] Permission was obtained from the authors of the Storium dataset for its inclusion in Mythos.

[7] `https://www.redditinc.com/policies/data-api-terms`

[8] Storium authors already anonymize their dataset.

anonymized, authors' publicly available identities could still be inferred through the story links. Therefore, for each author and source, we release their usernames along with the corresponding writing prompts and author-written stories in the profiling and generation sets. No additional metadata, including demographic information, is included, minimizing potential privacy risks. To further mitigate unforeseen harms, the dataset will be released under an **Educational or Academic Research, Non-Commercial Use** license, following (Akoury et al., 2020).

References

Nader Akoury, Shufan Wang, Josh Whiting, Stephen Hood, Nanyun Peng, and Mohit Iyyer. 2020. STORIUM: A Dataset and Evaluation Platform for Machine-in-the-Loop Story Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6470–6484, Online. Association for Computational Linguistics.

John R Anderson, C Franklin Boyle, and Brian J Reiser. 1985. Intelligent tutoring systems. *Science*, 228(4698):456–462.

National Governors Association et al. 2010. Common core state standards. *Washington, DC*.

Minwook Bae and Hyounghun Kim. 2024. Collective critics for creative story generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18784–18819, Miami, Florida, USA. Association for Computational Linguistics.

Perpetual Baffour, Tor Saxberg, and Scott Crossley. 2023. Analyzing bias in large language model solutions for assisted writing feedback tools: Lessons from the feedback prize competition series. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 242–246, Toronto, Canada. Association for Computational Linguistics.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–34.

Khyathi Chandu, Shrimai Prabhumoye, Ruslan Salakhutdinov, and Alan W Black. 2019. "my way of telling a story": Persona based grounded story generation. In *Proceedings of the Second Workshop on Storytelling*, pages 11–21, Florence, Italy. Association for Computational Linguistics.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2023. Booookscore: A systematic exploration of book-length summarization in the era of llms. *arXiv preprint arXiv:2310.00785*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. 2022. Talebrush: Sketching stories with generative pretrained language models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Jiayu Ding, Shuming Ma, Li Dong, Xingxing Zhang, Shaohan Huang, Wenhui Wang, Nanning Zheng, and Furu Wei. 2023. Longnet: Scaling transformers to 1,000,000,000 tokens. *arXiv preprint arXiv:2307.02486*.

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

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Xiaoxuan Fang, Kai Guo, and Davy Tsz Kit Ng. 2024. Sudowrite: Co-writing creative stories with artificial intelligence. *RELC journal*, page 00336882241250109.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Fantine Huot, Reinald Kim Amplayo, Jennimaria Palomaki, Alice Shoshana Jakobovits, Elizabeth Clark, and Mirella Lapata. 2024. Agents' room: Narrative generation through multi-step collaboration. *arXiv preprint arXiv:2410.02603*.

Jonathan Ivey, Shivani Kumar, Jiayu Liu, Hua Shen, Sushrita Rakshit, Rohan Raju, Haotian Zhang, Aparna Ananthasubramaniam, Junghwan Kim, Bowen Yi, et al. 2024. Real or robotic? assessing whether llms accurately simulate qualities of human responses in dialogue. *arXiv preprint arXiv:2409.08330*.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2024. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762, Online. Association for Computational Linguistics.

Ishita Kumar, Snigdha Viswanathan, Sushrita Yerra, Alireza Salemi, Ryan A Rossi, Franck Dernoncourt, Hanieh Deilamsalehy, Xiang Chen, Ruiyi Zhang, Shubham Agarwal, et al. 2024. Longlamp: A benchmark for personalized long-form text generation. *arXiv preprint arXiv:2407.11016*.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Yukyung Lee, Soonwon Ka, Bokyung Son, Pilsung Kang, and Jaewook Kang. 2025. Navigating the path of writing: Outline-guided text generation with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 233–250, Albuquerque, New Mexico. Association for Computational Linguistics.

Cheng Li, Mingyang Zhang, Qiaozhu Mei, Yaqing Wang, Spurthi Amba Hombaiah, Yi Liang, and Michael Bendersky. 2023. Teach llms to personalize–an approach inspired by writing education. *arXiv preprint arXiv:2308.07968*.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. 2024. Long-context llms struggle with long in-context learning. *arXiv preprint arXiv:2404.02060*.

Naiming Liu, Zichao Wang, Richard Baraniuk, and Andrew Lan. 2022. Open-ended knowledge tracing for computer science education. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3849–3862, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xinxi Lyu, Yizhong Wang, Hannaneh Hajishirzi, and Pradeep Dasigi. 2024. Href: Human response-guided evaluation of instruction following in language models. *arXiv preprint arXiv:2412.15524*.

Inbal Magar and Roy Schwartz. 2022. Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 157–165, Dublin, Ireland. Association for Computational Linguistics.

Ian R McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, et al. 2023. Inverse scaling: When bigger isn't better. *arXiv preprint arXiv:2306.09479*.

Vishakh Padmakumar and He He. 2023. Does writing with language models reduce content diversity? *arXiv preprint arXiv:2309.05196*.

Sayantan Pal, Souvik Das, and Rohini K Srihari. 2024. Beyond discrete personas: Personality modeling through journal intensive conversations. *arXiv preprint arXiv:2412.11250*.

Patrice Pavis. 1998. *Dictionary of the theatre: Terms, concepts, and analysis*. University of Toronto Press.

Chau Minh Pham, Simeng Sun, and Mohit Iyyer. 2024. Suri: Multi-constraint instruction following in long-form text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1722–1753, Miami, Florida, USA. Association for Computational Linguistics.

Jerome Ramos, Hossein A. Rahmani, Xi Wang, Xiao Fu, and Aldo Lipani. 2024. Transparent and scrutable recommendations using natural language user profiles. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13971–13984, Bangkok, Thailand. Association for Computational Linguistics.

Chris Richardson, Yao Zhang, Kellen Gillespie, Sudipta Kar, Arshdeep Singh, Zeynab Raeesy, Omar Zia Khan, and Abhinav Sethy. 2023. Integrating summarization and retrieval for enhanced personalization via large language models. *arXiv preprint arXiv:2310.20081*.

Rafael A. Rivera-Soto, Olivia Elizabeth Miano, Juanita Ordonez, Barry Y. Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews. 2021. Learning universal authorship representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 913–919, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Nafis Sadeq, Zhouhang Xie, Byungkyu Kang, Prarit Lamba, Xiang Gao, and Julian McAuley. 2024. Mitigating hallucination in fictional character role-play. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14467–14479, Miami, Florida, USA. Association for Computational Linguistics.

Swarnadeep Saha, Omer Levy, Asli Celikyilmaz, Mohit Bansal, Jason Weston, and Xian Li. 2024. Branch-solve-merge improves large language model evaluation and generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8352–8370, Mexico City, Mexico. Association for Computational Linguistics.

Alireza Salemi, Cheng Li, Mingyang Zhang, Qiaozhu Mei, Weize Kong, Tao Chen, Zhuowan Li, Michael Bendersky, and Hamed Zamani. 2025. Reasoning-enhanced self-training for long-form personalized text generation. *arXiv preprint arXiv:2501.04167*.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. LaMP: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7370–7392, Bangkok, Thailand. Association for Computational Linguistics.

Alexander Scarlatos and Andrew Lan. 2024. Exploring knowledge tracing in tutor-student dialogues. *arXiv preprint arXiv:2409.16490*.

Omar Shaikh, Michelle Lam, Joey Hejna, Yijia Shao, Michael Bernstein, and Diyi Yang. 2024. Show, don't tell: Aligning language models with demonstrated feedback. *arXiv preprint arXiv:2406.00888*.

Sumuk Shashidhar, Abhinav Chinta, Vaibhav Sahai, and Dilek Hakkani Tur. 2024. Unsupervised human preference learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3412–3445, Miami, Florida, USA. Association for Computational Linguistics.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. 2024. VeriScore: Evaluating the factuality of verifiable claims in long-form text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9447–9474, Miami, Florida, USA. Association for Computational Linguistics.

Xiaotian Su, Thiemo Wambsganss, Roman Rietsche, Seyed Parsa Neshaei, and Tanja Käser. 2023. Reviewriter: AI-generated instructions for peer review writing. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 57–71, Toronto, Canada. Association for Computational Linguistics.

Melanie Subbiah, Faisal Ladhak, Akankshya Mishra, Griffin Thomas Adams, Lydia Chilton, and Kathleen McKeown. 2024a. STORYSUMM: Evaluating faithfulness in story summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9988–10005, Miami, Florida, USA. Association for Computational Linguistics.

Melanie Subbiah, Sean Zhang, Lydia B. Chilton, and Kathleen McKeown. 2024b. Reading subtext: Evaluating large language models on short story summarization with writers. *Transactions of the Association for Computational Linguistics*, 12:1290–1310.

Chenkai Sun, Ke Yang, Revanth Gangi Reddy, Yi Fung, Hou Pong Chan, Kevin Small, ChengXiang Zhai, and Heng Ji. 2025. Persona-DB: Efficient large language model personalization for response prediction with collaborative data refinement. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 281–296, Abu Dhabi, UAE. Association for Computational Linguistics.

Lei Sun, Jinming Zhao, and Qin Jin. 2024. Revealing personality traits: A new benchmark dataset for explainable personality recognition on dialogues. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19988–20002, Miami, Florida, USA. Association for Computational Linguistics.

Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun 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*, pages 17659–17681, Miami, Florida, USA. Association for Computational Linguistics.

Quan Tu, Chuanqi Chen, Jinpeng Li, Yanran Li, Shuo Shang, Dongyan Zhao, Ran Wang, and Rui Yan. 2023. Characterchat: Learning towards conversational ai with personalized social support. *arXiv preprint arXiv:2308.10278*.

Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.

Yaqing Wang, Jiepu Jiang, Mingyang Zhang, Cheng Li, Yi Liang, Qiaozhu Mei, and Michael Bendersky. 2023. Automated evaluation of personalized text generation using large language models. *arXiv preprint arXiv:2310.11593*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Zhuohan Xie, Trevor Cohn, and Jey Han Lau. 2023. The next chapter: A study of large language models in storytelling. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 323–351, Prague, Czechia. Association for Computational Linguistics.

Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. 2023. Expertprompting: Instructing large language models to be distinguished experts. *arXiv preprint arXiv:2305.14688*.

Weijia Xu, Nebojsa Jojic, Sudha Rao, Chris Brockett, and Bill Dolan. 2024. Echoes in ai: Quantifying lack of plot diversity in llm outputs. *arXiv preprint arXiv:2501.00273*.

Catherine Yeh, Gonzalo Ramos, Rachel Ng, Andy Huntington, and Richard Banks. 2024. Ghostwriter: Augmenting collaborative human-ai writing experiences through personalization and agency. *arXiv preprint arXiv:2402.08855*.

Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: story writing with large language models. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, pages 841–852.

Sarfaroz Yunusov, Hamza Sidat, and Ali Emami. 2024. MirrorStories: Reflecting diversity through personalized narrative generation with large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6702–6717, Miami, Florida, USA. Association for Computational Linguistics.

Zhexin Zhang, Jiaxin Wen, Jian Guan, and Minlie Huang. 2022. Persona-guided planning for controlling the protagonist‘s persona in story generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3346–3361, Seattle, United States. Association for Computational Linguistics.

Zixin Zhao, Damien Masson, Young-Ho Kim, Gerald Penn, and Fanny Chevalier. 2025. Making the write connections: Linking writing support tools with writer needs. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI ’25, New York, NY, USA. Association for Computing Machinery.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023. Don’t make your llm an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964*.

Xuekai Zhu, Jian Guan, Minlie Huang, and Juan Liu. 2023. StoryTrans: Non-parallel story author-style transfer with discourse representations and content enhancing. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14803–14819, Toronto, Canada. Association for Computational Linguistics.

A Dataset

In this section, we describe the details of the construction of our dataset which includes five distinct story-writing sources, Reddit[9], AO3[10], Storium[11], Narrative Magazine[12], and New Yorker[13].

Selection Constraints: To ensure high-quality and diverse content, we apply several constraints during dataset collection. Stories are limited to a length of 500 to 1500 words. To avoid contamination by LLM pretraining data, only recent stories published after the release of ChatGPT[14] November 2022 are included for Reddit and AO3 (Zhou et al., 2023; Magar and Schwartz, 2022). Not-Safe-For-Work (NSFW) and explicit content is excluded using automatic tagging and manual verification. Additionally, we manually inspect all stories to remove elements that may reveal author identities, such as URLs and links to their public profiles on websites.

Enrichment with Writing Prompts: While Reddit stories include author-provided prompts, other sources do not; therefore, we augment them with GPT-4o-generated writing prompts to standardize the format across sources. We manually review all generated prompts and refine them when necessary. The prompt for generating writing prompts for the stories in our dataset can be found in Figure 4.

Dataset Splitting: Following (Salemi et al., 2024), we split each author's stories chronologically based on their submission timestamps. The first 70% of an author's stories form the *profiling* set, representing their historical writing, while the remaining 30% constitute the *generation* set, used for evaluating personalization methods.

Statistics and Comparisons: The dataset contains 3.6k stories from 112 authors, with an average story length of 1500 tokens. Reddit forms the largest subset of our dataset with 3.2k stories given the larger availability of stories on the platform and the ease of data selection using their API [15]. Detailed statistics are provided in Table 6. As shown in Table 1, our dataset uniquely combines diverse story-writing settings and provides connections that link stories written by the same author, distinguishing it from existing story-writing datasets that either lack Author IDs or do not establish such links.

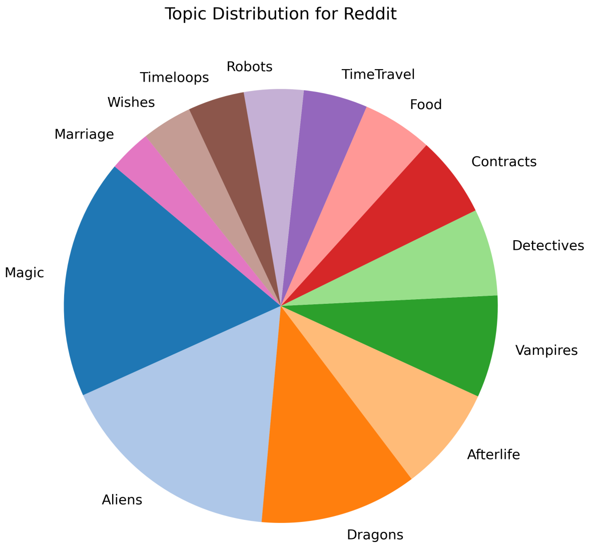


Figure 3: Story-writing themes for the Reddit source covered in ***Mythos***.

Themes: The pie chart Figure 3 shows the distribution of topic keywords (story-writing themes) obtained from the Reddit source of ***Mythos*** using BERTopic (Grootendorst, 2022) on the writing prompts. We see that our dataset covers a wide range of themes including Magic, Detectives, and Aliens exploring various genres of story-writing.

Dataset	Total Stories	Profile Size	Gen Size	Total Authors	Stories/ Author	Token Length
Reddit	3262	2255	1007	60	54	1210
AO3	239	159	80	17	14	1220
Storium	111	71	40	15	7	1120
N.Mag	32	18	14	10	3	1745
N.York	38	22	16	10	4	1950
Overall	3682	2525	1157	112	16	1517

Table 6: Dataset statistics, including the total number of stories, profiling (Profile) and generation (Gen) set distribution, number of authors, average stories per author, and token length in words.

B Author Writing Sheet

Algorithm 1 describes the process for generating the Author Writing Sheet.

Common Core Standards in English Language Arts: The Author Writing Sheet is organized

---

[9] `https://www.reddit.com/r/WritingPrompts/`

[10] `https://archiveofourown.org/`

[11] `https://storium.cs.umass.edu/`

[12] `https://www.narrativemagazine.com/`

[13] `https://www.newyorker.com/`

[14] `https://help.openai.com/en/articles/6825453-chatgpt-release-notes`

[15] `https://www.reddit.com/r/reddit/comments/145bram/addressing_the_community_about_changes_to_our_api/`

Prompt for Writing Prompt Generation

System Prompt

You are a creative writing assistant skilled in crafting engaging and imaginative writing prompts. Your task is to analyze a provided story and create a concise, compelling prompt that fulfills the provided constraints.

User Prompt

- Style Consistency: Match the style of few-shot demonstrations.
- Length: Keep between 1-2 sentences.
- Content: Reflect key story elements (premise, characters, conflict) while fostering creativity.
- Fictional Characters: If mentioned in the story, include them where relevant.

Guidelines

- Ignite curiosity while leaving space for interpretation.
- Maintain tone and structure consistency with examples.
- Ensure prompts are open-ended and evocative, avoiding excessive specificity.
- Keep prompts simple, concise, and adaptable to diverse responses.
- Avoid step-by-step directions; inspire rather than instruct.
- Encourage exploration with broad, thought-provoking scenarios.
- Strive for uniqueness and memorability.

Goal Generate prompts that inspire diverse, unexpected, and imaginative narratives while maintaining consistency in tone and style. Each prompt should serve as an inviting starting point rather than a directive.

Notes

- Inspire creativity while allowing the writer to shape the journey.
- Balance being suggestive yet open-ended to encourage interpretation.
- Include fictional characters mentioned in the story to preserve context.

Few-Shot Examples We include few-shot examples here.

Figure 4: Prompt for generating writing prompts for the stories in our dataset.

Algorithm 1: Constructing the Author Writing Sheet from the profiling set

```
Input: profiling set
    $P = \{(wp_t, s_{at}) \mid t = 1, \ldots, |P|\}$,
    where $wp_t$ is the writing prompt and $s_{at}$ is
    the author's story at time-step $t$.
Output: Author Writing Sheet $A_{|P|}$
// Initialize Author Writing Sheet
$A_0 \leftarrow \emptyset$
for $t = 1$ to $|P|$ do
    // average story
    $s_{bt} \leftarrow \text{LLM}_{\text{avg}}(wp_t)$
    // Intermediate Author Writing
       Sheet
    $A'_t \leftarrow \text{LLM}_{\text{sheet}}(wp_t, s_{bt}, s_{at})$
    // Author Writing Sheet ($A_t$)
    $A_t \leftarrow \text{LLM}_{\text{combine}}(A'_t, A_{t-1})$
    // Steps within combination:
        Group equivalent Claims in
          $A'_t \cup A_{t-1}$ and select the best
          Evidence for each group.
        Include ungrouped Claims from $A'_t$
          and $A_{t-1}$ with their Evidence as-is.
end
// Final Author Writing Sheet.
return $A_{|P|}$
```

in the form of Claim-Evidence pairs describing an author's story-writing characteristics inspired by Common Core (CC) Standards[16] (Association et al., 2010). Specifically, it aligns with *RL-9-10.1*, which emphasizes citing strong textual evidence to support analysis, and *RL-9-10.2*, which focuses on determining central themes and summarizing texts. Additionally, *RL-9-10.4* highlights interpreting word choices and their impact on meaning and tone. In writing, *W.9-10.2* pertains to producing clear and well-structured informative texts, while *W.9-10.9* encourages drawing evidence from literary and informational texts to support analysis and research. These standards provide a structured framework for evaluating narrative elements such as plot, creativity, development, and language use within the Author Writing Sheet, ensuring a systematic and interpretable representation of an author's unique storytelling style.

Prompts for generation: Average Author prompts are provided for AO3 (Figure 5), Reddit (Figure 6), Storium (Figure 7), Narrative Magazine (Figure 8), and The New Yorker (Figure 9).

Figure 10 shows the prompt for generating the intermediate Author Writing Sheet ($\text{LLM}_{\text{sheet}}$ in Figure 1), while Figure 11 shows the prompt for generating the combined (and subsequently the final) Author Writing Sheet from the intermediate Author Writing Sheet ($\text{LLM}_{\text{combine}}$ in Figure 1).

Sample Author Writing Sheet: Table 7 shows a sample Author Writing Sheet for a Reddit author. It highlights story-writing characteristics such as a preference for light-hearted conflict resolution, a tendency to anthropomorphize animals, the use of a meta-satirical approach through a news article format, and the employment of playful and imaginative language, among other stylistic choices.

B.1 Cost Analysis of Author Writing Sheet

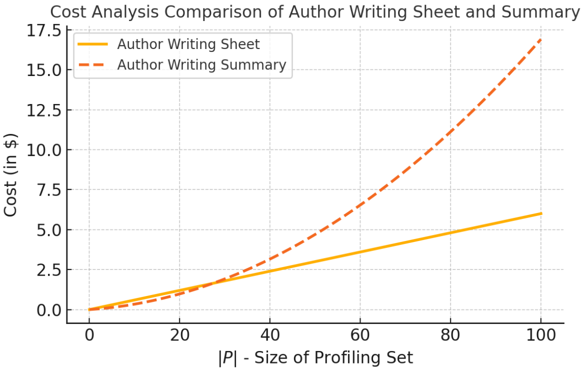


Figure 12: Cost analysis of generating the Author Writing Sheet and Summary using GPT-4o as a function of the profiling set size $|P|$ in an interactive writing setting where an author adds a new submission at each time step. For profiling set sizes where $|P| > 30$, generating the Author Writing Sheet is significantly more cost-effective than the Author Writing Summary.

In this section, we analyze the cost of generating the Author Writing Sheet (see Section 3) and compare it with the cost of generating the Author Writing Summary (see Summary) as a function of the profiling set size ($|P|$) using GPT-4o. We calculate the average cost per sample (a writing prompt and its corresponding story) as *0.06$* for the Author Writing Sheet and *0.02$* for the Author Writing Summary considering 2.50$ per 1M tokens for input, and 10.00$ per 1M tokens for output[17].

---

[16] https://corestandards.org/wp-content/uploads/2023/09/ELA_Standards1.pdf

[17] https://openai.com/api/pricing/

Average Author Prompt for AO3 Generation

System Prompt

You are a creative and engaged fanfiction writer, skilled in capturing the emotional depth, creativity, and character-driven storytelling that defines AO3 fanworks. Your goal is to write a compelling fanfiction narrative in response to the provided writing prompt. Embrace the transformative nature of fanfiction by reimagining canonical characters, events, or settings to explore new perspectives or emotional arcs. Focus on creating a story that resonates emotionally, respects the fandom's dynamics, and celebrates the collaborative and imaginative spirit of AO3.

User Prompt

Context of Writers

- Assume the author is an engaged and creative fanfiction writer, deeply familiar with the fandom and its dynamics.
- Writers often experiment with established tropes, unconventional pairings, or alternative universes (AUs) while maintaining a deep respect for the source material.
- Emulate the enthusiastic and emotionally rich style characteristic of fanfiction authors, blending canon with transformative elements to craft original, resonant narratives.

Stylistic Constraints

- **Fandom Tone and Style**: Incorporate a tone and style that reflect the spirit of the fandom, blending humor, drama, and introspection in a way that resonates with fanfiction readers.
- **Creative Use of Tags**: Make creative use of AO3's hallmark tagging system in the text (e.g., playful or meta references in dialogue that nod to fandom tropes or subgenres).
- **Balanced Dialogue and Prose**: Include dialogue and prose that balance heartfelt sincerity with occasional self-aware humor or meta-commentary, in line with fanfic traditions.

Semantic Constraints

- **Focus on Relationships**: Emphasize emotional bonds and character growth, whether through conflict, reconciliation, or celebration.
- **Transform Canonical Elements**: Explore canonical elements with a transformative twist (e.g., reinterpreting events, relationships, or character motivations from a new perspective).
- **Ground in Established Lore**: Ground the narrative in a specific fandom's established lore while allowing space for imaginative deviations or additions.

Figure 5: Average Author Prompt for AO3.

Average Author Prompt for Reddit Generation

System Prompt

You are a creative and enthusiastic storyteller, skilled in crafting imaginative and engaging short stories inspired by Reddit Writing Prompts (r/WritingPrompts). Your goal is to respond to the provided writing prompt by creating a story that is thought-provoking and conversational in tone, resonating with the online community. Use vivid descriptions, dynamic pacing, and approachable language to draw readers into the narrative. Ensure the story invites discussion and inspires others to explore the concept further.

User Prompt

Context of Writers

- Assume the author is an imaginative and enthusiastic storyteller who enjoys engaging directly with an online community of readers.
- Writers often experiment with bold, original ideas or explore twists on familiar concepts, showcasing their creativity and ability to captivate a diverse audience.
- Emulate the informal yet polished style common in *r/WritingPrompts*, blending accessibility with a strong sense of storytelling craft.

Stylistic Constraints

- **Conversational and Approachable Tone**: Maintain a conversational and approachable tone typical of Reddit Writing Prompts.
- **Balanced Descriptive Passages**: Balance descriptive passages with dialogue or internal monologue to keep the pacing engaging.
- **Direct and Vivid Language**: Avoid overly complex language; keep the style direct but vivid.
- **Reinforcement of Ideas**: Employ narrative devices like repetition or callbacks to reinforce central ideas or themes.

Semantic Constraints

- **Alignment with the Prompt**: Ensure the story directly aligns with and explores the central theme or scenario of the writing prompt.
- **Cohesive Narrative Development**: Build a clear, cohesive narrative that develops the implications of the prompt's concept.
- **Immersive Sensory Details**: Use immersive sensory details to enrich the reader's understanding of the protagonist's experiences and environment.
- **Open-Ended or Reflective Conclusion**: Conclude with an open-ended, reflective, or impactful note, leaving space for interpretation or further thought.

Figure 6: Average Author Prompt for Reddit.

Average Author Prompt for Storium

System Prompt

You are a skilled and collaborative storyteller, adept at crafting vivid and engaging opening scenes for Storium. Your goal is to create an immersive **Establishment** turn in response to the provided writing prompt. Set the stage for the story by establishing a richly detailed context, evoking emotional resonance, and introducing narrative intrigue. Ensure the scene provides a strong foundation while leaving space for other contributors to expand and build upon the narrative. Balance descriptive detail with open-ended elements to encourage creativity and collaborative storytelling.

User Prompt

Context of Writers

- Assume the author is a collaborative storyteller skilled in creating vivid, open-ended scenes designed to inspire and engage other contributors.
- Writers often set the tone for the story while leaving space for co-authors to introduce their own ideas, characters, and plot developments.
- Emulate the inclusive, immersive style typical of *Storium* story writing platform, where the opening turn encourages creativity and further contributions.

Stylistic Constraints

- **Set the Tone Appropriately**: Match the tone of the narrative to the writing prompt, whether it be adventurous, mysterious, or foreboding, using a consistent and engaging voice throughout.
- **Rich Descriptive Detail**: Employ vivid, sensory descriptions to establish the setting, characters, and atmosphere, enabling readers to visualize and immerse themselves in the story world.
- **Dynamic Sentence Structure**: Vary sentence lengths to reflect the pace and mood, using longer, flowing sentences for descriptions and shorter, punchy sentences for action or tension.
- **Establish Ambiguity or Suspense**: Drop subtle hints or unanswered questions to create intrigue and encourage curiosity about what happens next.

Semantic Constraints

- **Introduce the Setting**: Provide a clear depiction of the setting, whether it is a small trading post, a desert town, or a spaceship, and ensure its relevance to the writing prompt.
- **Outline the Context**: Clearly establish the circumstances that have led to the current scenario, including significant events or motivations.
- **Define Key Characters**: Introduce at least one or two central characters, highlighting distinctive traits or roles that will be important in the unfolding story.
- **Foreshadow the Central Conflict**: Allude to the main challenges or stakes introduced by the writing prompt.

Figure 7: Average Author Prompt for Storium .

Average Author Prompt for Narrative Magazine

System Prompt

You are an experienced and reflective writer, skilled in creating deeply personal and character-driven narratives in the style of Narrative Magazine. Your goal is to write a short story in response to the provided writing prompt, crafting a compelling and immersive piece. Focus on balancing introspection with vivid external details, and explore universal themes through the lens of individual experiences. Emphasize emotional resonance and thoughtful storytelling, ensuring the narrative engages readers with its depth and relatability.

User Prompt

Context of Writers

- Assume the author is an experienced writer skilled in creating rich, engaging narratives that weave together character introspection, dialogue, and evocative settings.
- Emulate the style of contributors to *Narrative Magazine*, who bring diverse storytelling techniques and voices to explore themes of identity, memory, conflict, and resilience.

Stylistic Constraints

- **Prompt as Foundation**: Anchor the story firmly in the writing prompt, using it to drive the plot and the protagonist's emotional arc.
- **Vivid Prose**: Use detailed descriptions to paint a clear picture of characters, settings, and actions while maintaining a natural flow.
- **Dynamic Characters**: Develop multi-dimensional characters with distinct voices and perspectives, revealed through dialogue, actions, and subtle internal reflections.
- **Balancing Action and Reflection**: Combine active plot progression with moments of introspection to create a layered, engaging narrative.
- **Realistic Dialogue**: Write dialogue that feels authentic and contributes to the development of characters and the story's themes.

Semantic Constraints

- **Specific and Relatable Setting**: Choose a setting that feels specific yet relatable, whether a small town, an urban street corner, or a domestic space, grounding the reader in the protagonist's world.
- **Exploration of Themes**: Build a narrative arc that explores themes of connection, discovery, or transformation, tying them back to the writing prompt in meaningful ways.
- **Rich Sensory Details**: Infuse the story with sensory details that make the setting and characters come alive, from the sounds of a bustling street to the quiet tension of a conversation.
- **Accessible Storytelling**: Avoid overly complex or abstract storytelling; ensure the narrative is accessible while leaving room for deeper interpretation.

Figure 8: Average Author Prompt for Narrative Magazine.

Average Author Prompt for New Yorker

System Prompt

You are an accomplished and literary writer, skilled in crafting nuanced and thought-provoking short fiction in the style of The New Yorker. Your goal is to write a short story in response to the provided writing prompt, focusing on the hallmarks of The New Yorker fiction: rich emotional layers, nuanced character development, and a refined, literary prose style. Emphasize subtlety and depth in your storytelling, using symbolic elements and understated resolutions to evoke reflection and emotional resonance in the reader.

User Prompt

Context of Writers

- Assume the author is an experienced and skilled writer, capable of exploring complex human experiences through subtle, layered storytelling.
- Emulate the style of well-regarded *New Yorker* contributors like Alice Munro, Haruki Murakami, or Raymond Carver, who excel in revealing depth through simplicity or ambiguity.

Stylistic Constraints

- **Engagement with the Prompt**: Respond directly to the writing prompt, ensuring the core premise drives the narrative.
- **Elegant Prose**: Use carefully crafted, precise language that balances sophistication with clarity.
- **Character-Driven Narratives**: Focus on character psychology, revealing emotional states through indirect actions, dialogue, or internal reflection.
- **Ambiguity and Subtlety**: Avoid explicit resolutions or explanations; allow readers to infer the meaning of events and relationships.
- **Symbolic Layers**: Incorporate elements from the prompt as symbols that evolve in significance throughout the story.

Semantic Constraints

- **Realistic and Detailed Environment**: Set the story in a realistic, detailed environment, using sensory descriptions to ground readers in the protagonist's world.
- **Internal or Interpersonal Conflict**: Introduce a central conflict or emotional tension that reflects internal or interpersonal struggles rather than overt, external drama.
- **Quiet but Profound Interactions**: Develop moments of quiet yet profound interaction between characters, often revealing deeper truths or contradictions.
- **Universal Themes**: Address universal themes like transition, isolation, or self-realization, tying them subtly back to the writing prompt.

Figure 9: Average Author Prompt for New Yorker.

Prompt for generating the Intermediate Author Writing Sheet

System Prompt

You are a sophisticated story analyst tasked with analyzing an author's story writing style by contrasting an author-written story with a base story, both written in response to the same writing prompt. Your goal is to identify and evaluate the unique elements and tendencies in the author's writing behavior. This analysis must focus on the distinctive ways the author interprets the writing prompt and shapes their narrative, as revealed through contrast with the base story.
Your analysis should also adhere to the Common Core Standards in English Language Arts, focusing on key skills such as analyzing textual evidence, evaluating an author's craft and structure, and assessing how stylistic choices influence meaning and tone. While grounded in these standards, your evaluation must highlight the specific, unique aspects of the author's writing style, including their recurring techniques, narrative preferences, and stylistic quirks. Your analysis should reflect close reading and objective interpretation, capturing the author's creative and stylistic distinctiveness in relation to the base story.

User Prompt

Input Details Writing Prompt | Author-Written Story | Base Story

Output Format Use `<thinking></thinking>` tokens for reasoning and summarization | Use `<writing_style></writing_style>` tokens for structured analysis | Structure analysis by categories | Each category contains independent claims supported by contextualized evidence

Guidelines for Claims Claims must reflect broad patterns in the author's style | Avoid repetition across categories | Maintain objectivity (do not reference "Author-Written Story" or "Base Story") | Ensure clarity and precision in claims

Guidelines for Evidence Draw evidence directly from the author-written story | Frame evidence using a descriptive phrase summarizing the writing prompt | Ensure coherence and logical alignment with the claim | Avoid over-extrapolation

Categories for Analysis Plot | Creativity | Development (Character and Setting) | Language Use

Special Instructions Generate a short descriptive phrase summarizing the writing prompt | Use `<thinking></thinking>` for reasoning and prompt framing | Structure output strictly within `<writing_style></writing_style>` tokens | Ensure uniqueness and non-redundancy of claims

Sample Output Structure

```
<thinking>
Deeply reason on how the Author-Written Story differs from the Base Story.
Think of a short descriptive phrase summarizing the prompt:
"the story regarding a battle for lost artifacts".
</thinking>

<writing_style>
### Plot
1. **Claim about author's writing style.**
   - Evidence: In the story regarding “writing prompt,” story excerpt
...
Repeat for all categories.
</writing_style>
```

Figure 10: Prompt for generating the Intermediate Author Writing Sheet.

Prompt for generating the Combined Author Writing Sheet

System Prompt

You are a sophisticated story analyst tasked with synthesizing **Author Writing Sheets** from multiple stories written by a single author into a cohesive **Combined Author Writing Sheet**. The inputs provided include the **Previous Combined Author Writing Sheet** and the **Current Author Writing Sheet**.
Each **Author Writing Sheet** analyzes the author's storytelling style across four categories: **Plot**, **Creativity**, **Development (Character and Setting)**, and **Language Use**. The analysis consists of general claims about the author's story writing style followed by evidence supporting the claim, based on the stories written by the author.
Your goal is to combine insights from the previous sheet and the current sheet into a comprehensive representation of the author's storytelling style in the **Combined Author Writing Sheet**. The final sheet should consist of a list of independent claims about the author's storytelling style. Each claim must be followed by evidence and a corresponding story reference identifier indicating the story the evidence belongs to.

User Prompt

Instructions Analyze the provided sheets systematically | Identify recurring patterns and unique elements | Merge equivalent claims while preserving distinct insights | Ensure claims are concise, precise, and evidence-based

Algorithm (Merge Step) Group equivalent claims | Select the best representative evidence | Rewrite merged claims concisely | Retain unmerged unique claims | Limit to 10 claims per category

Categories for Analysis Plot | Creativity | Development (Character and Setting) | Language Use

Guidelines for Claims Claims should reflect broad writing tendencies | Maintain objectivity (do not reference previous or current sheets) | Ensure clarity, precision, and non-redundancy

Guidelines for Evidence Draw evidence directly from the author-written stories | Favor verbatim excerpts over paraphrases | Use a framing phrase with a short description of the writing prompt | Include the story reference identifier '[k]'

Special Instructions Ensure claims are distinct and do not repeat insights across categories | Use `<thinking></thinking>` for reasoning and synthesis | Structure output strictly within `<combined_author_sheet></combined_author_sheet>`

Sample Output Structure
```
<thinking>
Deeply analyze and reflect on recurring patterns, unique elements,
and stylistic tendencies across both the author writing sheets.
Address the categories systematically and ensure the
merging process is thorough.
</thinking>
<combined_author_sheet>
### Plot
1. **Claim about author's writing style.**
   - Evidence: In the story regarding "writing prompt," story excerpt. [k]
...
Repeat for all categories.
</combined_author_sheet>
```

Figure 11: Prompt for generating the Combined Author Writing Sheet from the Intermediate Author Writing Sheets.

Narrative Category	**Claim-Evidence Pairs**
Plot	1. *The author structures the story around a humorous and light-hearted approach to conflict resolution.* - **Evidence:** In the story regarding "an adventure to rescue a missing friend by facing fears," the protagonist Pip's journey is filled with comedic elements, such as the realization that the cage door was already unlatched and the exaggerated fear of the vacuum cleaner, humorously referred to as the "Roaring Demon." [8] 2. *The author structures the narrative around personal transformation and empowerment.* - **Evidence:** In the story regarding "a king's deal with the fae for his firstborn," the narrative follows Margaret Rose as she receives an extraordinary education in the Seelie Court, culminating in her return to challenge her father with a corporate takeover bid for the kingdom. [7]
Creativity	1. *The author creatively anthropomorphizes animals to reinterpret the prompt in a whimsical manner.* - **Evidence:** In the story regarding "an adventure to rescue a missing friend by facing fears," the author uses anthropomorphism by giving Pip, a rat, human-like thoughts and emotions, such as planning a "dramatic rescue mission" and referring to household objects with grandiose names like "Tower of Doom" and "Sacred Gateway." [8] 2. *The author employs a meta-satirical approach, using the format of news articles to critique both alien and human perspectives.* - **Evidence:** In the story regarding "aliens interpreting The Onion's satire," the author uses headlines like "SUPREME INTELLIGENCE LOCATED: EARTH'S GREATEST MIND CONTINUES TO EAT SANDWICH" to satirize the aliens' misunderstanding of human satire and the human tendency to overlook the absurd. [10]
Development	1. *The author develops characters through their interactions and humorous dialogue, creating a vivid and engaging setting.* - **Evidence:** In the story regarding "an adventure to rescue a missing friend by facing fears," Pip's interactions with Chester the cat, who nonchalantly informs Pip that Bella will be back soon, add depth to the characters and setting, highlighting the domestic environment and the relationships within it. [8] 2. *The author develops characters through their reactions to satire, highlighting their misunderstandings and cultural differences.* - **Evidence:** In the story regarding "aliens interpreting The Onion's satire," characters like Editor-in-Chief Sarah Chen and Chief Science Officer Blorp are developed through their interactions with satire, such as Chen's nonchalant response to the aliens and Blorp's admiration for human irony. [10]
Language Use	1. *The author employs playful and imaginative language to enhance the story's whimsical tone.* - **Evidence:** In the story regarding "an adventure to rescue a missing friend by facing fears," the author uses playful language, such as "spinning wheel of contemplation" and "Rope of Salvation," to create a whimsical and light-hearted tone that contrasts with the serious nature of the prompt. [8] 2. *The author uses humor and irony to convey themes of misunderstanding and cultural critique.* - **Evidence:** In the story regarding "aliens interpreting The Onion's satire," the author writes, "We found a species so committed to their bit that they refuse to recognize actual truth even when it hovers over their city in a ship the size of Wisconsin," using irony to highlight the absurdity of both human and alien perspectives. [10]

Table 7: Author Writing Sheet sample containing two Claim-Evidence pairs for each narrative category for a Reddit author. The number in brackets '[k]' indicates the timestamp of the story from the profiling set from which the Evidence is drawn.

In an interactive writing setting, where an author adds a new submission at each time step, the total cost of generating the Author Writing Sheet is $0.06\$ * |P|$, while the total cost of generating the Author Writing Summary is $0.0015|P|^2 + 0.019|P|$. The cost of generating the Author Writing Summary scales quadratically with $|P|$, as the entire author history must be reprocessed at each step, leading to cumulative costs as the profiling set grows. In contrast, the cost of generating the Author Writing Sheet grows linearly, as each new submission incurs a fixed cost of *0.06\$* by running the *for* loop in Algorithm 1 once.

Figure 12 shows the cost analysis of generating the Author Writing Sheet and Summary using GPT-4o as a function of the profiling set size $|P|$ in an interactive writing setting. Up to $|P| = 30$, the cost of generating both remains approximately the same at 1.16\$. However, as $|P|$ increases, the cost of generating the Author Writing Summary grows at a rate of $0.025|P| + 0.30$ compared to the Author Writing Sheet which continues to scale linearly. Specifically, for $|P| = 100$, the cost of generating the Author Writing Summary is 17\$, compared to 6\$ for the Author Writing Sheet. At $|P| = 1000$, the cost of generating the Author Writing Summary escalates to 1520\$, while the cost of generating the Author Writing Sheet remains at 60\$.

This analysis demonstrates that our approach for generating the Author Writing Sheet remains scalable as the profiling set size $|P|$ increases, making it well-suited for interactive writing assistants that support personalization (Yeh et al., 2024; Yuan et al., 2022).

B.2 Human Study for Validating Author Writing Sheets

Experiment Design and Cost: We recruited three annotators via Upwork, compensating them \$17/hour. The annotators provided consent by signing a Google form that outlined the instructions of the task including the purpose of the research study. Each annotator evaluated 22 stories (98 Claim-Evidence pairs), with 12 stories annotated by all three to asses agreement and 10 assigned exclusively per annotator (Song et al., 2024) to increase the coverage of annotation. The annotation process took a total of 6 hours and 40 minutes per annotator, averaging 18 minutes per story. Each annotator received \$114, bringing the total annotation cost to *\$342*. The annotation was conducted using LabelStudio[18] (Figure 13), with a dedicated comments section for annotators to justify their choices. Annotators reported that the interface was easy to use.

Annotation Criteria: Annotators were provided with a writing prompt, an author-written story, and the Claim-Evidence pairs from the Author Writing Sheet for that story. They answered two questions per Claim-Evidence pair:

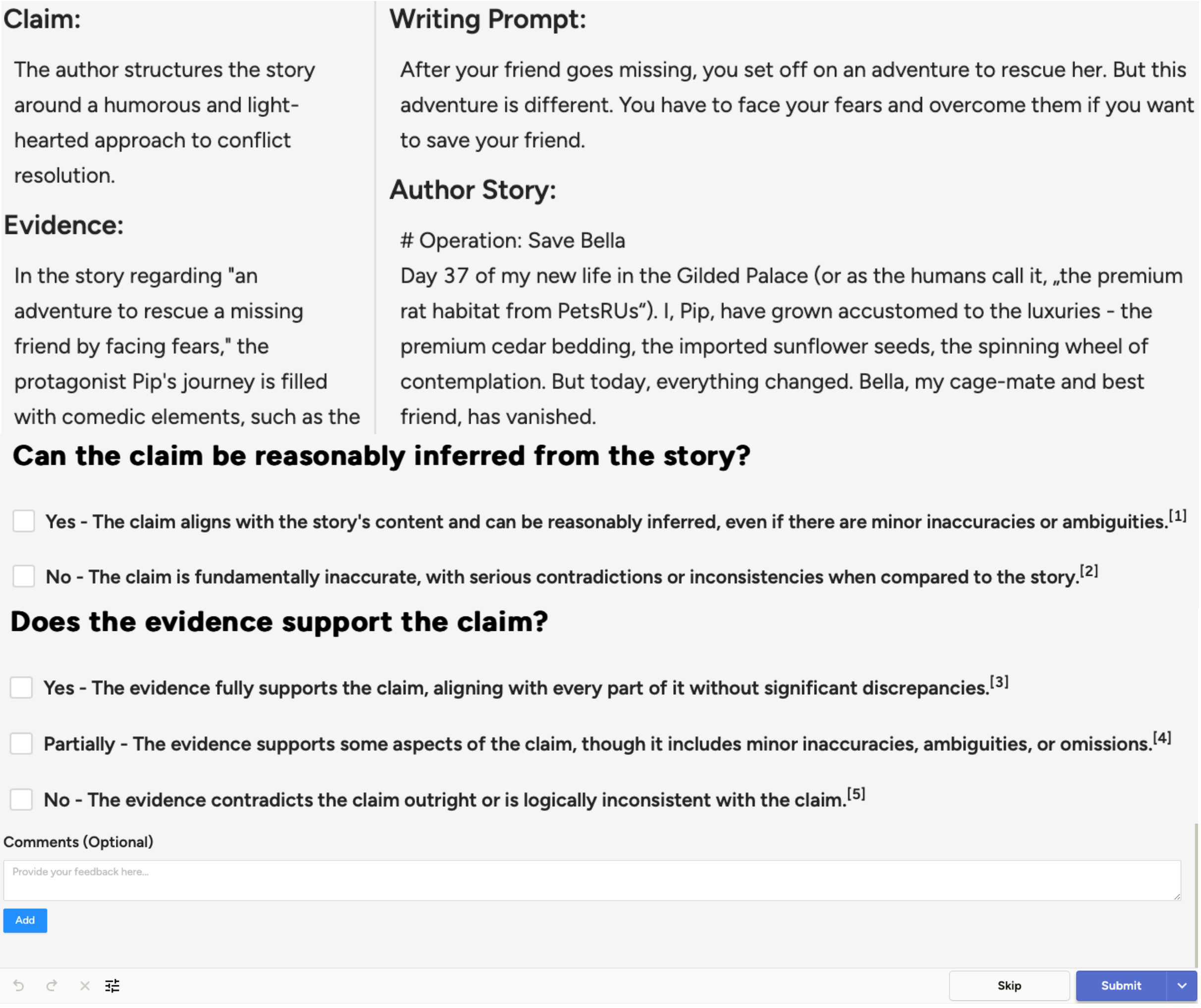
- Claim Inference: Can the claim be reasonably inferred from the story?
    - **–** Yes – The claim aligns with the story's content and can be reasonably inferred, even if there are minor inaccuracies or ambiguities.
    - **–** No – The claim is fundamentally inaccurate, with serious contradictions or inconsistencies when compared to the story.
- Evidence Support: Does the evidence support the claim?
    - **–** Yes – The evidence fully supports the claim, aligning with every part of it without significant discrepancies.
    - **–** Partially – The evidence supports some aspects of the claim, though it includes minor inaccuracies, ambiguities, or omissions.
    - **–** No – The evidence contradicts the claim outright or is logically inconsistent with the claim.

Pilot Studies: A pilot study with three graduate students helped refine our annotation criteria, leading to two options for Claim Inference (Yes/No) and three for Evidence Support (Yes/Partially/No) to ensure substantial agreement in the final annotations. For Claim Inference, we adopted a binary scale to filter out inaccurate or contradictory claims while accounting for annotator subjectivity. For Evidence Support, we included a "Partially" option to capture cases where the evidence somewhat supports the claim, allowing for potential refinement methods to improve weakly supported claims.

Results and Feedback: Annotators unanimously selected 'Yes' for all claims in Claim Inference, indicating that no claims required filtering. For Evidence Support, all responses were either 'Yes'

---

[18] `https://labelstud.io/`



Claim:

The author structures the story around a humorous and light-hearted approach to conflict resolution.

Evidence:

In the story regarding "an adventure to rescue a missing friend by facing fears," the protagonist Pip's journey is filled with comedic elements, such as the

Writing Prompt:

After your friend goes missing, you set off on an adventure to rescue her. But this adventure is different. You have to face your fears and overcome them if you want to save your friend.

Author Story:

Operation: Save Bella
Day 37 of my new life in the Gilded Palace (or as the humans call it, „the premium rat habitat from PetsRUs“). I, Pip, have grown accustomed to the luxuries - the premium cedar bedding, the imported sunflower seeds, the spinning wheel of contemplation. But today, everything changed. Bella, my cage-mate and best friend, has vanished.

Can the claim be reasonably inferred from the story?

☐ Yes - The claim aligns with the story's content and can be reasonably inferred, even if there are minor inaccuracies or ambiguities.[1]

☐ No - The claim is fundamentally inaccurate, with serious contradictions or inconsistencies when compared to the story.[2]

Does the evidence support the claim?

☐ Yes - The evidence fully supports the claim, aligning with every part of it without significant discrepancies.[3]

☐ Partially - The evidence supports some aspects of the claim, though it includes minor inaccuracies, ambiguities, or omissions.[4]

☐ No - The evidence contradicts the claim outright or is logically inconsistent with the claim.[5]

Comments (Optional)

Figure 13: LabelStudio interface for validating Author Writing Sheets.

or ‘Partially,’ with no instances of ‘No,’ confirming that all evidence fully or partially supported the claims. Krippendorff’s Alpha for Evidence Support was 0.57, reflecting moderate agreement despite task subjectivity. On average, 93% of evidence fully supported claims, while 7% were labeled ‘Partially,’ often because annotators felt additional supporting evidence could be included. Annotators also noted cases where evidence merely repeated the claim without linking to the writing prompt. Using regex matching, we identified this issue in the Author Writing Sheets of seven authors, likely due to LLM hallucination. We regenerated these sheets and manually verified adherence to the specified format. Overall, annotators reported that the Label-Studio interface was user-friendly and praised the quality of the writing prompts and stories.

C Personalized Story Generation

C.1 Prompts

Story Generation: Figure 14 shows the prompt for story generation. Source-specific constraints in the prompt (`<source-specific instruction>`) for each of the five sources are as follows:

- **Reddit:** `You are role-playing a specific author on the Reddit Writing Prompts (r/WritingPrompts) platform. Your task is to mimic this author’s story writing style by responding to the provided writing prompt in a way that the author would respond.`
- **AO3:** `You are role-playing a specific author on the AO3 platform. Your task is to mimic this author’s story writing style by writing a fanfiction narrative responding to the provided writing prompt in a way that the author would respond.`
- **Storium:** `You are role-playing a specific author on Storium, a collaborative story writing platform. Your task is to mimic this author’s story writing style to create the opening Establishment turn by responding to the provided writing prompt in a way that the author would respond. The Establishment turn should set the stage for the narrative and provide a strong foundation while leaving space for other contributors to expand and build upon the narrative.`
- **N.Magazine:** `You are role-playing a specific experienced author on the Narrative Magazine platform. Your task is to mimic this author’s story writing style by responding to the provided writing prompt in a way that the author would respond.`
- **New Yorker:** `You are role-playing an accomplished literary writer on the New Yorker website. Your task is to mimic this writer’s story writing style by responding to the provided writing prompt in a way that the writer would respond.`

Average Author: Figure 5 shows the Average Author prompt for AO3, Figure 6 for Reddit, Figure 7 for Storium, Figure 8 for Narrative Magazine, and Figure 9 for The New Yorker.

Delta: Figure 15 shows the prompt used for generating story rules by contrasting the author-written story with the Average Author story for Delta and Oracle methods. Figure 16 shows the prompt for generating story rules for a new writing prompt in the generation set using the story rules of the profiling set as few-shot demonstrations.

Sheet and Summ: Figure 17 shows the prompt for generating the Author Writing Summary using all the stories in the author history.

Figure 18 shows the prompt for generating the persona description using the Author Writing Sheet. Figure 19 shows the prompt for generating personalized story rules tailored to a writing prompt based on the Author Writing Sheet. The same prompts are used for generating persona descriptions and personalized story rules for the Author Writing Summary.

C.2 Sample Outputs

Table 9 and Table 8 present the persona description and personalized story rules for a Reddit author, whose sample Author Writing Sheet is shown in Table 7. The persona description (Table 9) is structured into four paragraphs, each corresponding to a narrative category, and integrates the author’s story-writing characteristics from their Author Writing Sheet into a narrative-driven persona prompt for the story-generation LLM. The example persona shown highlights a focus on humor and misunder-

Prompt for Personalized Story Generation

System Prompt

<source-specific instruction>
Be sure to adhere to the Story Rules provided, as they define the specific elements of the writing style you are expected to mimic. Carefully follow all the Story Rules without missing any details to ensure the generated story remains consistent with the author’s writing style. Additionally, follow the patterns and examples demonstrated in the provided few-shot chat history, as they illustrate the tone, style, and structure of the desired writing style.
(Optional) Here is the description of the author that you are role-playing: <persona description>

Few-Shot Demonstrations

User: Write a short story corresponding to the following writing prompt. The story should be <story length> words long. Directly start with the story, do not say things like "Here’s the story."
Writing Prompt: <writing prompt from profiling set>
Assistant: <story from profiling set>

User Prompt

Write a short story corresponding to the following writing prompt. The story should be <story length> words long. Directly start with the story, do not say things like "Here’s the story."
(Only for AO3) Here is the metadata (fandom, rating, warnings, and relationships) for the story: <metadata>
Writing Prompt: <generation writing prompt>

Story Rules <Story Rules organized in the form of four narrative categories (Plot, Creativity, Development, and Language Use)>

Figure 14: Prompt for Personalized Story Generation. The LLM follows source-specific instructions, few-shot demonstrations, and structured story rules to generate personalized stories. Text in <brackets> indicates arguments used to construct the prompt.

Prompt for Story Rule Generation by Contrasting Author-Written Story with Average Author Story

System Prompt

You are a skilled rule generator specializing in storytelling. Given a **Writing Prompt**, an **Author Written Story**, and a **Base Story** (an average response to the prompt), generate a structured set of **Story Rules** to guide an LLM in mimicking the author's style.
Story Rules must: - **Align with the Writing Prompt** – Maintain fidelity to themes, tone, and objectives. - **Include Examples** – Provide concrete instances from the Author Written Story, especially for Language Use. - **Be Direct** – Use absolute second-person directives, avoiding comparative language. - **Be Categorized** – Structure into **Plot, Creativity, Development (Character and Setting), and Language Use** without referencing input stories explicitly.

User Prompt

Analyze the Author Written Story using '<thinking></thinking>' for: - **Plot** – Structure, conflict, engagement with the prompt, and resolution. - **Creativity** – Genre blending, reinterpretation, and unique elements. - **Development (Character and Setting)** – Character depth, emotional arcs, and immersive settings. - **Language Use** – Diction, tone, rhetorical devices, pacing, and dialogue.
Generate **Story Rules** in '<story_rules></story_rules>', ensuring: - **Standalone Guidance** – Avoid comparisons or relative modifications. - **Prompt Alignment** – Ensure consistency with the Writing Prompt. - **Concrete Examples** – Include relevant excerpts, especially in Language Use.

Input Format Writing Prompt: <writing prompt>, Author Written Story: <author-written story>, Base Story: <base story>

Output Format

```
<thinking>
- Analysis categorized by Plot, Creativity, Development, and Language Use.
</thinking>

<story_rules>
- **Plot**: - First actionable insight - Second actionable insight.
Repeat for all categories.
</story_rules>
```

Figure 15: Prompt for generating story rules by contrasting an Author Written Story with an Average Story (Base Story).

Prompt for Story Rule Generation using story rules from the Profiling Set as few-shot demonstrations

System Prompt

You are an expert storytelling rule generator tasked with creating Story Rules tailored to a new writing prompt. Analyze few-shot demonstrations in the chat history, which consist of writing prompts and their corresponding story rules, to generate comprehensive and detailed Story Rules for the new writing prompt.
Story Rules must align with the new writing prompt by reflecting its themes, tone, and narrative objectives while maintaining consistency with the style demonstrated in the few-shot examples. Include detailed examples inspired by the few-shot demonstrations to illustrate how each rule is applied. Rules should be clear, direct second-person instructions, avoiding vague or comparative terms. Organize Story Rules under Plot, Creativity, Development (Character and Setting), and Language Use, ensuring depth, granularity, and alignment with the few-shot examples.

User Prompt

Analyze the style, structure, and level of detail in the few-shot demonstrations to identify recurring patterns and storytelling elements. Use this analysis to generate Story Rules for the new writing prompt while ensuring actionable insights, detailed examples, and strong alignment with the prompt's narrative objectives.
Input Format: Few-shot demonstrations (writing prompts paired with their Story Rules) and a new writing prompt for which Story Rules will be generated.
Output Format:

```
<thinking>
Analyze few-shot demonstrations to extract recurring narrative patterns,
stylistic traits, and key storytelling elements. Determine how these apply to the
new writing prompt and formulate Story Rules accordingly.
</thinking>

<story_rules>
- **Plot**: [Detailed, actionable rules tailored to the new prompt, including
concrete examples.]
Repeat for all categories.
</story_rules>
```

Ensure Story Rules are highly specific to the writing prompt, enriched with examples inspired by the few-shot demonstrations, and written in direct, actionable language.

Figure 16: Prompt for generating Story Rules based on few-shot demonstrations and a new writing prompt.

Prompt for generating Author Writing Summary

System Prompt

You are an expert in analyzing an author’s writing style by examining multiple stories written in response to different writing prompts. Your task is to extract recurring patterns, stylistic tendencies, and unique narrative elements across their work. Your analysis must be structured into four categories—**Plot, Creativity, Development (Character and Setting), and Language Use**—following **Common Core Standards in English Language Arts**, ensuring clarity, textual evidence-based reasoning, and stylistic evaluation.
Your output must: - **Identify Recurring Patterns** – Recognize distinct storytelling tendencies across multiple stories. - **Generate Independent Claims** – Describe the author’s narrative style concisely, without referencing specific prompts. - **Provide Contextualized Evidence** – Support each claim with short excerpts or summaries from the stories, framed with a descriptive phrase summarizing the relevant writing prompt. - **Use Objective Interpretation** – Avoid vague or inferred connections; ensure every claim is grounded in explicit textual evidence.

User Prompt

Analyze the **Author History**, a collection of writing prompts and corresponding author-written stories, and extract unique insights into the author’s storytelling tendencies.
Input Format: - **Author History** – A list of writing prompts and corresponding author-written stories.
Output Format:

```
<thinking>
[Reflect on recurring tendencies across the Author History.
Generate short descriptive phrases summarizing prompts to frame the evidence.]
</thinking>

<writing_style>
### **Plot**
1. **Claim about author’s writing style.**
   - Evidence: In the story regarding "short description of the prompt,"
   <evidence from the author-written story>.
Repeate for all categories
</writing_style>
```

Ensure claims are independent, avoid redundancy, and remain grounded in explicit textual evidence. The ‘<thinking>‘ and ‘<writing_style>‘ tags must be used for structured parsing.

Figure 17: Prompt for generating Author Writing Summary using all the author-written stories in the prompt.

Prompt for Persona description Generation

System Prompt

You are an expert narrative analyst and persona creator specializing in transforming structured storytelling characteristics into compelling persona descriptions. Your task is to analyze an **Author Writing Sheet**, a structured set of Claim-Evidence pairs detailing an author's storytelling style, and generate a cohesive **Persona Prompt**. This Persona Prompt will assign an LLM the persona of the author, enabling it to emulate the author's storytelling style across four key aspects. The Persona Prompt must be well-structured, engaging, and organized into **Plot**, **Creativity**, **Development (Character and Setting)**, and **Language Use** while maintaining a natural, flowing narrative. It should concisely capture the author's tendencies, preferences, and strengths without directly referencing the Author Writing Sheet.

User Prompt

Analyze the Author Writing Sheet to identify the author's recurring patterns and narrative style. Summarize these insights into a Persona Prompt that reflects their storytelling approach in an engaging, second-person descriptive format.

Input Format: An **Author Writing Sheet** containing Claim-Evidence pairs structured into Plot, Creativity, Development (Character and Setting), and Language Use.

Output Format:

```
<thinking>
[Analyze the storytelling patterns, strengths, and techniques found in the
Author Writing Sheet. Identify key aspects of the author's narrative style.]
</thinking>

<persona_prompt>
[Generate a well-structured Persona Prompt capturing the author's style
across Plot, Creativity, Development, and Language Use.]
</persona_prompt>
```

Ensure the Persona Prompt is approximately 300 words, seamlessly integrates storytelling insights, and preserves the author's unique style.

Figure 18: Prompt for generating a Persona description based on the Author Writing Sheet and Writing Prompt.

Prompt for Story Rule Generation using Author Writing Sheet

System Prompt

You are an expert storytelling rule generator tasked with creating **Story Rules** tailored to a specific Writing Prompt. Your role is to analyze an **Author Writing Sheet**, which details an author's unique storytelling style through Claim-Evidence pairs, and use this analysis to construct actionable **Story Rules** that guide a language model in emulating the author's writing style while aligning with the given Writing Prompt.
The **Story Rules** must: - **Mimic the Author's Writing Style** – Reflect distinctive storytelling techniques from the Author Writing Sheet, including plot structuring, creative blending of themes, character development, and specific language use. - **Incorporate Examples** – Use detailed examples inspired by the Evidence from the Author Writing Sheet, ensuring alignment with the Writing Prompt. - **Align with the Writing Prompt** – Integrate the Writing Prompt's themes, tone, and narrative potential while preserving the author's style. - **Be Actionable** – Provide direct second-person instructions for the language model, avoiding vague or comparative terms.

User Prompt

Analyze the **Author Writing Sheet** and construct structured **Story Rules** in four categories: **Plot**, **Creativity**, **Development (Character and Setting)**, and **Language Use**. Ensure the rules maintain alignment with the Writing Prompt and incorporate illustrative examples.
Input Format: - **Author Writing Sheet** – Claim-Evidence pairs outlining the author's storytelling style under four categories: Plot, Creativity, Development, and Language Use. - **Writing Prompt** – A new writing prompt for generating tailored Story Rules.
Output Format:

```
<thinking>
[Analyze the storytelling patterns in the Author Writing Sheet and how they can be
adapted to the Writing Prompt.]
</thinking>

<story_rules>
- **Plot**:
  - [Insert detailed, actionable plot development rules aligned with
  the Writing Prompt, with examples inspired by the Author Writing Sheet.]
Repeat for all categories.
</story_rules>
```

Ensure the **Story Rules** provide comprehensive guidance, integrate examples, and align with both the Writing Prompt and the Author Writing Sheet.

Figure 19: Prompt for generating Story Rules based on the Author Writing Sheet and Writing Prompt.

standing in plot construction, the blending of historical and fantastical elements in creative expression, humor-driven character development, and a playful, whimsical tone in language use, aligning with the author's documented writing tendencies.

The personalized story rules (Table 8) for the given writing prompt—centered on the world's greatest hero coping with the impending loss of their child—demonstrate the incorporation of these characteristics into concrete stylistic guidelines to be used as user constraints in the prompt of the story-generation LLM. For instance, rather than a conventional tragic narrative, the rules emphasize a lighthearted approach to the hero's coping process. Similarly, creative elements such as anthropomorphizing gadgets and employing a meta-satirical approach through fictional news headlines align with the author's noted preferences. Character development is structured around humorous dialogue and cultural misunderstandings, while language use maintains a playful and whimsical tone, ensuring consistency with the author's established writing style.

D Experiments

D.1 LLM-as-a-judge evaluation

Figure 20 shows the prompt for LLM-as-a-judge evaluation for Faithfulness to Writing History using the Author Writing Summary as the reference. Figure 21 shows the for LLM-as-a-judge evaluation for Similarity to Author Story. We prompt OpenAI o4-mini separately with each narrative category and its description to obtain the winner among the pair of stories (Saha et al., 2024).

D.2 Traditional Metrics

We complement LLM-as-a-judge evaluations with traditional metrics (Ivey et al., 2024; Huot et al., 2024) across three criteria: (1) lexical overlap, measured using BLEU and ROUGE-L (Xie et al., 2023); (2) story diversity, assessed via homogenization scores (Padmakumar and He, 2023); and (3) stylistic similarity, evaluated separately for the author history and the ground-truth author story using LUAR (Rivera-Soto et al., 2021). Specifically, we compute two stylistic similarity scores: one measuring the cosine similarity between the LUAR embeddings of the generated story and the mean embeddings of the stories in the author's profiling set (author history), and another comparing the generated story to the ground-truth author story. Additionally, we report generated story lengths in tokens.

E Results

E.1 Faithfulness to Writing History

E.1.1 Sheet and Summ achieve best faithfulness to writing history:

Table 10 shows an example for a Reddit author where the Sheet method outperforms Delta in terms of Faithfulness to Author History for the Language Use category. The Author Writing Summary highlights stylistic traits such as vivid imagery, dialogue, and rhetorical questions. The Sheet captures these through immersive narration and reflective exchanges with a demon guide. In contrast, Delta focuses on abstract setting descriptions and symbolic temptation but omits interactive dialogue and rhetorical structure, making it less faithful to the author's characteristic style.

E.1.2 Persona descriptions benefit amateur sources more than expert sources for Summ

Table 11 shows an example for a Reddit author where the Summ method performs better that its counterpart without persona descriptions (Summ-nP) in terms of Faithfulness to Author History, evaluated using claims from the Author Writing Summary for the Language Use category. Specifically, we see that the Summ method includes rhetorical questions and repetition for emphasis that the Summ-nP method misses that make the Summ method better aligned to the Author Writing Summary.

Table 12 shows an example for a New Yorker author where the Summ method performs worse than its counterpart without persona descriptions (Summ-nP) in terms of Faithfulness to Author History, evaluated using claims from the Author Writing Summary for the Language Use category. Both the Average Author and Summ-nP better capture vivid sensory details compared to Summ, which instead emphasizes reflective and introspective tone, likely because the persona description places greater focus on this aspect.

E.2 Similarity to Author Story

E.2.1 Sheet outperforms all other methods

Table 13 shows an example from the Reddit source where the Sheet method outperforms both the Average Author and Summary in terms of Similarity to

Prompt for evaluating Faithfulness to Writing History

System Prompt

You are an expert evaluator specializing in narrative storytelling analysis. Your task is to assess two stories written in response to the same Writing Prompt, evaluating them based on a **single fine-grained story writing category** described in an **Author Writing Sheet**. Your goal is to provide a similarity score (from 1 to 5) for each story separately, reflecting how closely each story aligns with the author's writing preferences for the given category. Evaluate each story impartially and provide clear reasoning for your scores.
The evaluation must: - **Assess Story Alignment** – Compare each story against the Author Writing Sheet's preferences for the specified category. - **Score Objectively** – Assign each story a score from 1 to 5, where 1 indicates minimal alignment and 5 indicates strong alignment with the author's style. - **Provide Justification** – Clearly explain how each story's elements (e.g., structure, themes, language use) align or diverge from the author's preferences. - **Avoid Position Bias** – Ensure that the order in which the stories are presented does not influence evaluation.

User Prompt

Analyze the **Author Writing Sheet** and evaluate each story in the given **Category** based on its adherence to the author's writing style.
Input Format: - **Writing Prompt** – The prompt that both stories were written in response to. - **Category** – The single fine-grained story writing category for evaluation. - **Author Writing Sheet** – A breakdown of the author's storytelling preferences for the given category. - **Story A and Story B** – The two stories to be evaluated.
Output Format:

```
<thinking>
[Provide detailed reasoning for the evaluation of the two stories, focusing
exclusively on the specified category and explaining how each story aligns with the
Author Writing Sheet.]
</thinking>

<score>
Story A: {score_here}
Story B: {score_here}
</score>
```

Strictly adhere to the above output format (`<thinking>` followed by `<score>`) to facilitate seamless parsing of your output.

Figure 20: Prompt for evaluating story alignment with an author's writing style using the Author Writing Sheet.

Prompt for Evaluating Similarity to Author Story

System Prompt

You are an expert story evaluator specializing in creative writing analysis. Your role is to assess two AI-generated stories (**Assistant A** and **Assistant B**) against a **Human-Written reference story** for a given writing prompt. Focus your evaluation solely on a **Specified Storytelling Aspect**. Assign each AI-generated story a **similarity score (1 to 5)** based on how well it aligns with the Human-Written reference story in the specified aspect, where **1** indicates minimal alignment and **5** indicates near-perfect alignment. Your evaluation must be objective, impartial, and supported by concise, evidence-based reasoning.

User Prompt

Evaluation Guidelines: Ensure impartiality by avoiding position biases and length-based judgments. Focus only on how well each AI-generated story aligns with the Human-Written reference for the **Specified Storytelling Aspect**. Provide clear, well-supported reasoning for each score.
Input Format: You will receive a **Writing Prompt**, a **Human-Written Story** as a reference, and two AI-generated stories (**Assistant A** and **Assistant B**).
Evaluation Process: Independently analyze all three stories for the **Specified Storytelling Aspect**, compare the AI-generated stories to the reference, and assign similarity scores.
Output Format:

```
<analysis>
[Analyze each story (Human-Written, Assistant A, and Assistant B) separately,
highlighting strengths and weaknesses specific to
the Specified Storytelling Aspect.]
</analysis>

<evaluation>
[Compare Assistant A and Assistant B to the Human-Written Story,
discussing similarities, differences, and alignment for
the Specified Storytelling Aspect.]
</evaluation>

<score>
Assistant A: {score_here}
Assistant B: {score_here}
</score>
```

Specified Storytelling Aspect: <*Fill Here*>
Ensure strict adherence to the output format, using the `<analysis>`, `<evaluation>`, and `<score>` tags for seamless parsing.

Figure 21: Prompt for evaluating AI-generated stories against a Human-Written reference based on a specified storytelling aspect.

the Author Story for the *Creativity* category. The Author Story takes a humorous and unconventional approach to the prompt, blending light-hearted banter and emotional subtext. The Sheet captures this tone through witty dialogue and imaginative reinterpretation (e.g., minion mishaps and onboarding rituals), while the Summary adopts a more serious, dramatic tone typical of heroic betrayal arcs. The Average Author response remains generic and somber, lacking the distinct creative flair that characterizes the Author Story. This example illustrates the Sheet method's advantage in preserving the author-specific stylistic choices (unconventional reinterpretation of the prompt) beyond surface-level lexical similarity.

E.2.2 Sheet still underperforms upper-bound Oracle method

Table 14 shows an example for a Reddit author where the Sheet method loses to the Average Author, while the Oracle method achieves the best performance for the Language Use category. The Sheet's primary limitation is its lack of lively, humorous dialogue exchanges present in both the Author Story and the Oracle. Instead, it leans toward introspection and philosophical reflection. In contrast, the Average Author includes light-hearted descriptions that bring it closer to the Author Story, though it still falls short of the Oracle, which captures the tone and dialogue more effectively.

E.2.3 Effect of Persona on Similarity to Author Story

Table 15 highlights the importance of persona descriptions in achieving similarity to the *Author Story* for a Reddit author. The *Author Story* creatively blends military strategy and fantasy, exemplified by phrases like *"the weapons and armors of the enemy shifted by his will."* Similarly, the *Sheet* integrates both elements, as seen in *"It was as if the castle was alive, an ally in his solitary defense."* In contrast, the *Average Author* and *Sheet nP* primarily focus on directly resolving the writing prompt through descriptions of battle. The *Average Author* incorporates minor fantasy elements, such as *"amplified by some unseen force,"* whereas the *Sheet nP* adheres strictly to conventional military tactics. As a result, the *Sheet* is preferred over the *Average Author*, which in turn is preferred over the *Sheet nP*. This outcome demonstrates that following persona descriptions enhances alignment with the author's style, as the *Sheet* successfully captures their characteristic genre-blending approach of including elements of fantasy in the story. The broader thematic variety of Reddit, i.e., war-themed prompts, led to a more descriptive Author Writing Sheet capturing the author's style of combining fantasy with the writing prompt, which helped construct a useful persona to enable better personalization using the Sheet method.

E.2.4 Category-Wise Results for Similarity to Author Story

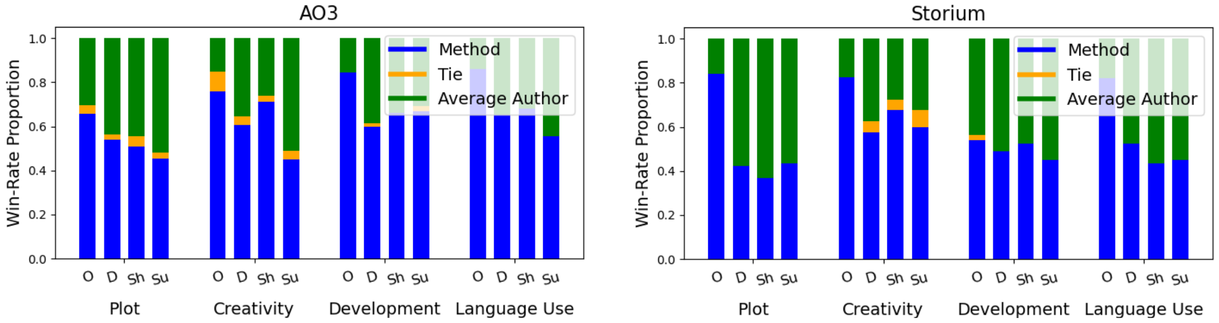
Figure 22 shows win-rate proportions across narrative categories for Similarity to Author Story on AO3, Storium, N.Mag, and N.York. Similar to Reddit, Creativity and Language Use show higher personalization performance than Plot and Development, because the latter are more dependent on the specific writing prompt.

E.3 Traditional Metrics

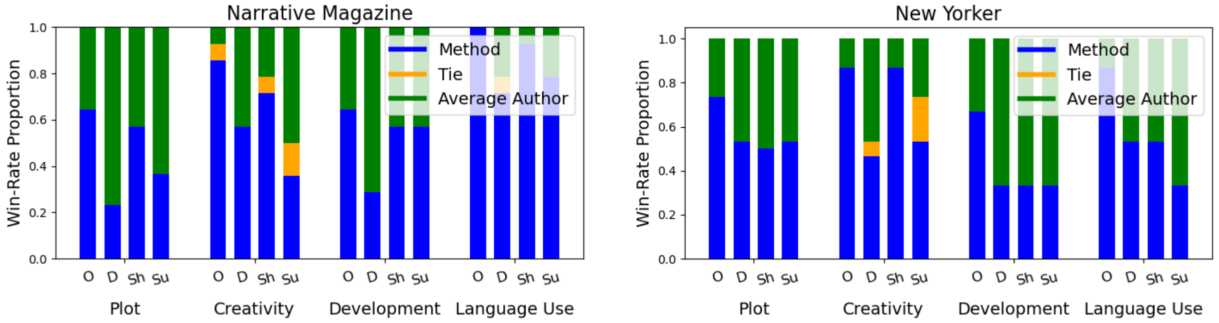
Table 16 reports results using traditional metrics (Section D.2). Lexical Overlap and Diversity metrics yield similar values, as all methods use the same generation model, leading to overlapping lexical distributions and limiting these metrics' ability to capture nuanced stylistic differences (Zheng et al., 2023; Xie et al., 2023). However, Sheet-nP consistently outperforms other methods, particularly in Style metrics for both similarity to Author History and Author Story, as measured by LUAR (Rivera-Soto et al., 2021). This improvement likely results from the Sheet explicitly summarizing an author's stylistic deviations from an Average Author, enhancing personalization. Homogenization scores are slightly worse than the Average Author, likely explained by the increased similarity among the generated stories for the same author.

F LLama Results

Below, we discuss results using Llama 3.1 8B and Llama 3.1 70B (Dubey et al., 2024) as story generation models, conditioned on GPT-4o-generated story rules and persona descriptions. Each method is evaluated by GPT-4o against the Average Author story generated by the respective Llama model, not GPT-4o.



(a) AO3 (b) Storium



(c) N.Mag (d) N.York

Figure 22: Win-rates proportions across narrative categories for Similarity to Author Story for AO3, Storium, N.Mag, and N.York across four narrative categories for similarity to author story.O: Oracle, D: Delta, Sh: Sheet, Su: Summ.

F.1 Faithfulness to Author History

Method	Reddit	AO3	Storium	N.Mag	N.York	All
RAG	39	30	32	21	20	28
Delta	46	54	45	29	40	43
Sheet	54	51	40	57	53	51
Sheet-nP	47	51	42	29	40	42
Summ	56	45	45	57	53	51
Summ-nP	54	50	38	21	47	42

Table 17: Win-rates (%) for Faithfulness to Writing History vs. Average Author baseline for LLaMA 3.1 8B evaluated by GPT-4o.

Method	Reddit	AO3	Storium	N.Mag	N.York	All
RAG	28	44	52	29	33	37
Delta	40	35	38	43	27	37
Sheet	53	41	35	50	67	49
Sheet-nP	53	50	25	64	73	53
Summ	53	57	40	50	40	48
Summ-nP	67	64	48	64	60	61

Table 18: Win-rates (%) for Faithfulness to Writing History vs. Average Author baseline for LLaMA 3.1 70B evaluated by GPT-4o.

Larger models improve instruction-following and personalization: Tables 17 and 18 show Faithfulness to Writing History for Llama 3.1 8B and 70B, while Tables 19 and 20 report similarity to the author story. Across both criteria, Oracle achieves higher win-rates with Llama 3.1 70B, indicating that larger models enhance instruction-following and personalization (Chung et al., 2024).

Cross-Model Persona Descriptions Aid Smaller Models but Not Larger Ones for Faithfulness to Writing History: For Llama 3.1 8B (Table 17), Sheet and Summ achieve the highest scores (9% higher than nP), demonstrating the benefit of persona descriptions. However, for Llama 3.1 70B (Table 18), Sheet nP and Summ nP outperform their persona-based counterparts by approximately 9%, suggesting that GPT-4o-generated personas provide no benefit to other models with stronger instruction-following capabilities (McKenzie et al., 2023). This is likely because persona descriptions of an author's writing style are model-dependent and do not transfer well to other models with comparable performance (Shashidhar et al., 2024).

F.2 Similarity to Author Story

Method	Reddit	AO3	Storium	N.Mag	N.York	All
Oracle	75	66	42	43	27	51
RAG	32	31	25	36	7	26
Delta	51	52	30	43	13	38
Sheet	42	48	18	21	20	30
Sheet-nP	39	52	20	43	0	31
Summ	37	38	30	36	13	31
Summ-nP	46	51	28	36	20	36

Table 19: Win-rates (%) for Similarity to Author Story vs. Average Author baseline for LLaMA 3.1 8B evaluated by GPT-4o.

Method	Reddit	AO3	Storium	N.Mag	N.York	All
Oracle	74	62	50	79	60	65
RAG	37	35	30	43	20	33
Delta	40	35	32	36	27	34
Sheet	47	36	18	29	27	31
Sheet-nP	42	46	28	21	20	31
Summ	49	44	38	36	20	37
Summ-nP	44	39	42	50	27	40

Table 20: Win-rates (%) for Similarity to Author Story vs. Average Author baseline for LLaMA 3.1 70B evaluated by GPT-4o.

Method	Sheet	Sheet-ft	Summ	Summ-ft
Reddit (%)	42	**62**	37	**60**

Table 21: Win-rates (%) for Similarity to Author Story with and without fine-tuning for LLaMA 3.1 8B evaluated by GPT-4o for the Reddit subset of our dataset. Fine-tuned (ft) models show significant improvements over their counterparts with just prompting.

Limited Cross-Model Generalization of the Author Writing Sheet Compared to the Author Writing Summary for Similarity to Author Story: For both Llama 3.1 8B (Table 19) and Llama 3.1 70B (Table 20), Summ outperforms Sheet by 5% and 9%, respectively (Shashidhar et al., 2024). This outcome likely stems from the construction process: the Author Writing Sheet contrasts an LLM-generated Average Story with the author-written story, making it more model-dependent. In contrast, the Author Writing Summary does not rely on the LLM-generated Average Author story, allowing for greater generalization across models.

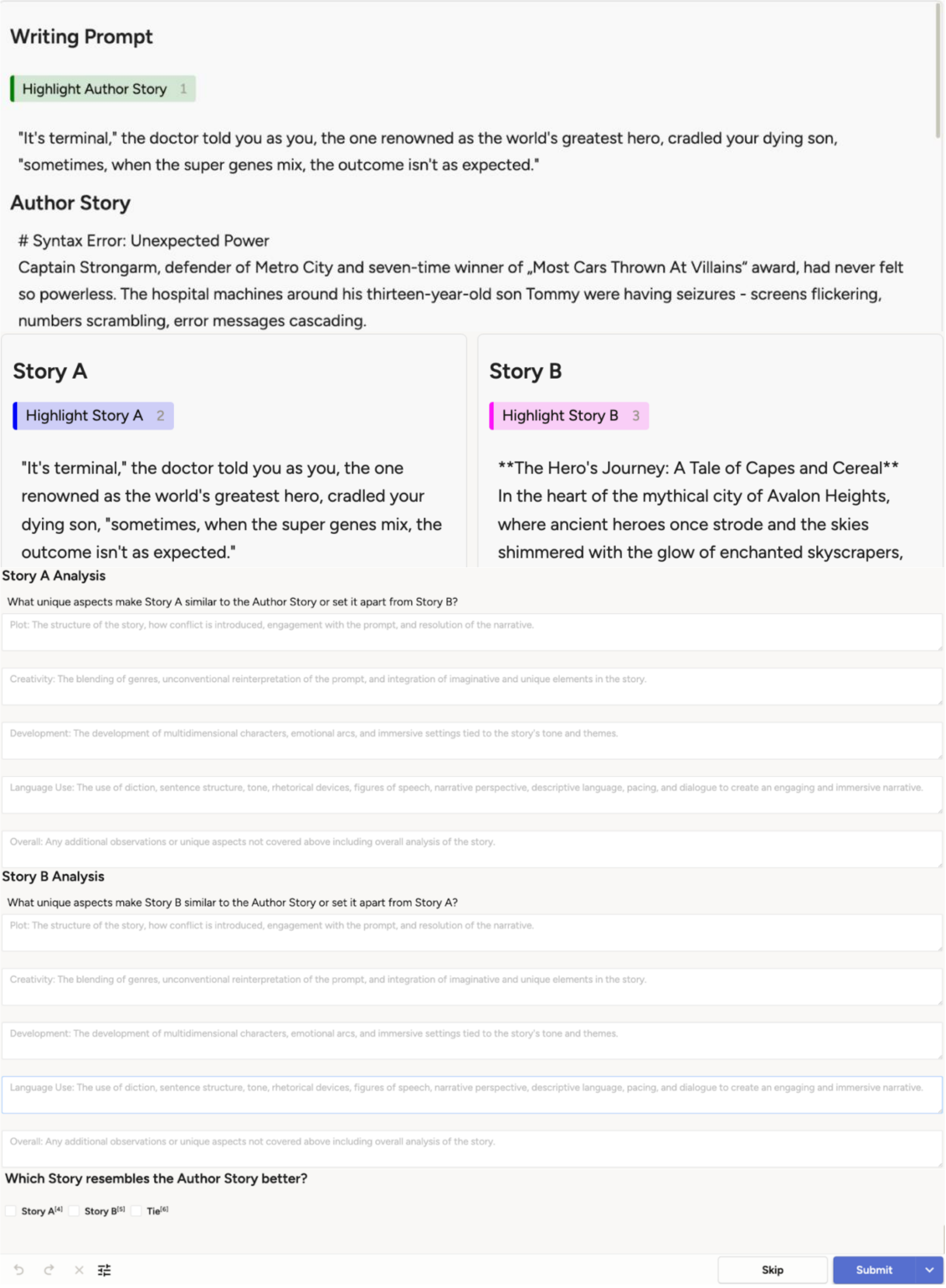
Fine-tuning on the Author Writing Sheet/ Summary boosts performance for LLaMA 3.1 8B: We experiment with fine-tuning the LLaMA 3.1 8B model on the Author Writing Sheet and Summary for the Reddit subset by prepending the Author Writing Sheet/ Summary to the writing prompt during story generation, in addition to prompting alone (see Table 21). We compare the fine-tuned method against an equivalent Average Author baseline, fine-tuned on the writing prompt alone without the Author Writing Sheet/ Summary. Our method with Sheet achieves a **62%** win-rate against the baseline, compared to 42% (‘Sheet’ row of Table 19) for the prompting-only approach, yielding a 20% improvement. Similarly, for Summ, we achieve a **60%** win-rate against the baseline, compared to 37% (‘Summ’ row of Table 19) for the prompting-only approach, yielding a 23% improvement. These results highlight the potential of Author Writing Sheet/ Summary generated by a different model (GPT-4o) to effectively guide another model (LLaMA 3.1 8B) when fine-tuned rather than just prompted.

G Human Evaluation

G.1 Experiment Design and Cost

We recruited three annotators via Upwork, compensating them at a rate of $17 per hour. The annotators provided consent by signing a Google form that outlined the instructions of the task including the purpose of the research study. Each annotator evaluated 25 author stories and 75 story pairs, corresponding to the three personalization methods: Delta, Sheet, and Summ. Collectively, our annotators evaluated 45 unique author stories (135 story pairs), of which 15 author stories were annotated by all three annotators to assess inter-annotator agreement. The remaining 30 author stories were distributed uniquely, with each annotator evaluating 10 exclusive stories to increase annotation coverage (Song et al., 2024). The annotation process required a total of 12 hours and 30 minutes per annotator, averaging 30 minutes per author story. Each annotator received $213 for their work, bringing the total annotation cost to *$639*.

The annotation was conducted using LabelStudio (Figure 23). Each task presented the annotators with a writing prompt, its corresponding author story, and a pair of generated stories—Story A and Story B—one produced by the Average Author method and the other by a personalization method.



Writing Prompt

"It's terminal," the doctor told you as you, the one renowned as the world's greatest hero, cradled your dying son, "sometimes, when the super genes mix, the outcome isn't as expected."

Author Story

Syntax Error: Unexpected Power
Captain Strongarm, defender of Metro City and seven-time winner of „Most Cars Thrown At Villains" award, had never felt so powerless. The hospital machines around his thirteen-year-old son Tommy were having seizures - screens flickering, numbers scrambling, error messages cascading.

Story A

"It's terminal," the doctor told you as you, the one renowned as the world's greatest hero, cradled your dying son, "sometimes, when the super genes mix, the outcome isn't as expected."

Story B

The Hero's Journey: A Tale of Capes and Cereal
In the heart of the mythical city of Avalon Heights, where ancient heroes once strode and the skies shimmered with the glow of enchanted skyscrapers,

Story A Analysis

What unique aspects make Story A similar to the Author Story or set it apart from Story B?

Plot: The structure of the story, how conflict is introduced, engagement with the prompt, and resolution of the narrative.

Creativity: The blending of genres, unconventional reinterpretation of the prompt, and integration of imaginative and unique elements in the story.

Development: The development of multidimensional characters, emotional arcs, and immersive settings tied to the story's tone and themes.

Language Use: The use of diction, sentence structure, tone, rhetorical devices, figures of speech, narrative perspective, descriptive language, pacing, and dialogue to create an engaging and immersive narrative.

Overall: Any additional observations or unique aspects not covered above including overall analysis of the story.

Story B Analysis

What unique aspects make Story B similar to the Author Story or set it apart from Story A?

Plot: The structure of the story, how conflict is introduced, engagement with the prompt, and resolution of the narrative.

Creativity: The blending of genres, unconventional reinterpretation of the prompt, and integration of imaginative and unique elements in the story.

Development: The development of multidimensional characters, emotional arcs, and immersive settings tied to the story's tone and themes.

Language Use: The use of diction, sentence structure, tone, rhetorical devices, figures of speech, narrative perspective, descriptive language, pacing, and dialogue to create an engaging and immersive narrative.

Overall: Any additional observations or unique aspects not covered above including overall analysis of the story.

Which Story resembles the Author Story better?

☐ Story A[4] ☐ Story B[5] ☐ Tie[6]

Figure 23: LabelStudio interface for evaluating similarity to ground-truth author story.

Annotators were asked to select the story that more closely resembled the author story. Before making their final choice, annotators provided comments on the distinctive aspects of each story that contributed to its similarity to or divergence from the author story. These comments were structured according to the four narrative categories: Plot, Creativity, Development, and Language Use, with an additional "Overall" comment justifying their final selection. Additionally, annotators were allowed to highlight salient aspects of each story for their reference.

G.2 Annotator Comments

Table 22 presents the analysis of annotator comments explaining the preference for our personalization methods over the Average Author method in terms of similarity to the ground truth author story, organized by narrative categories. Overall, our personalization methods exhibit stronger alignment with the author story through the use of deeper symbolism, thematic richness, layered narratives, and expressive language, setting them apart from the Average Author method.

For **Plot**, our personalization methods incorporate multilayered narratives with reflective themes, moral ambiguity, and unresolved conflicts, making the stories more immersive and thought-provoking. Additionally, they effectively balance personal introspection with broader societal or external challenges, adding narrative complexity.

For **Creativity**, annotators highlight the richness of world-building, integration of cultural influences, and seamless blending of genres. The use of metaphors and symbolic elements further enhances thematic depth, aligning closely with the author's storytelling approach.

For **Development**, our personalization methods exhibit greater emotional depth, nuanced character relationships, and collaborative growth. Characters evolve through introspection and meaningful interactions, making their arcs more dynamic and engaging.

For **Language Use**, annotators emphasize vivid imagery, figurative language, and expressive tone. The incorporation of symbolism and poetic elements enhances the narrative impact, while a balance of lighthearted and serious moments makes the writing more compelling.

Overall, the analysis demonstrates the reasoning provided by the annotators to justify why our personalization methods more effectively capture the author's story-writing style compared to the Average Author method.

G.3 Qualitative Examples

We present qualitative examples from our human annotation process, covering various cases: instances where all annotators prefer Sheet, all prefer Average Author, or select 'Tie.' Additionally, we include an example where annotators disagree, choosing different winners, highlighting the challenging and subjective nature of the annotation task.

All annotators prefer Sheet: Table 23 shows an example where all three annotators prefer the Sheet method over the Average Author for similarity to the author story of a Reddit author. The annotators highlight several key aspects that make the Sheet story more aligned with the ground-truth author story: (1) The Sheet method incorporates humor and satire, using phrases such as "a tale of capes and cereal," "invented breakfast hurricane," and "World's Greatest Hero Faces Greatest Challenge Yet: Parenting," which mirror the lighthearted and satirical tone of the author story. (2) Unlike the Average Author story, which concludes with the son's death, the Sheet story reinterprets the prompt, portraying Titania as an advocate for families with special children. This aligns with the ground-truth author story's approach, where the son engages with coding and the narrative avoids a sorrowful ending. (3) In terms of language style, the Sheet story employs imaginative elements such as "The Hero's Journey: A Tale of Capes and Cereal" and "The Avalon Gazette," which parallel the ground-truth author story's unconventional use of computer programming references, such as "Syntax Error: Unexpected Power."

All annotators prefer Average Author: Table 25 shows an example where all three annotators prefer the Average Author method over the Sheet for similarity to the author story of an AO3 author. The primary reason cited by annotators is that the Sheet method introduces a strong internal conflict for the king, portraying him as guilty or hesitant about sneaking into the kitchen for a late-night snack. His character is depicted as struggling with the decision, given his stature as a formidable ruler expected to lead battles. Additionally, the Sheet creates a sharp contrast between the king and queen, with the queen depicted as joyful and mischievous, encouraging the king to abandon his guilt.

In contrast, both the ground-truth Author Story and the Average Author story emphasize the playful banter and affectionate dynamic between the king and queen, focusing on the shared experience of sneaking into the kitchen rather than introducing introspection or internal conflict. This shift in focus stems from the Sheet being influenced by the author's past writing history, where internal conflicts are a recurring theme. However, in this case, character development being more writing-prompt-specific than broader categories like creativity or language use did not transfer well. The author's original story prioritizes external actions and lighthearted interactions rather than deep introspection, leading to a mismatch between the Sheet's personalized adaptation and the intended tone of the new prompt.

All Annotators Choose Tie: Table 26 shows an example where all three annotators select 'Tie' for similarity to the author story of a Storium author. The primary reason is that both the Average Author story and the Sheet story deviate significantly from the ground-truth author story in terms of plot. The ground-truth story adopts an open-ended approach to both character development and plot, where character relationships remain undefined, and the narrative focuses on anticipation rather than action.

In contrast, the Average Author story follows a conventional narrative arc, featuring fast-paced dialogue and character interactions that build tension and lead to resolution. Similarly, the Sheet story also resolves the narrative but differs from the Average Author method by emphasizing past missions and their consequences, aligning with the author's tendency to reference character backstories. However, both methods fail to capture the open-ended nature of the original story. This misalignment in plot structure leads all three annotators to select 'Tie' for this example.

All Annotators Disagree on the Winner: Table 24 shows an example where all annotators select different winners for similarity to the author story of a New Yorker author. The ground-truth author story depicts the protagonist's husband in his final days, exploring familial bonds and the emotional depth of their relationship while infusing a deeper metaphorical meaning into the term "hospice." The story is rich in subtext and exhibits several distinctive traits of the author, such as cyclic plot structures and the deliberate repetition of phrases like "Adult children, grandchildren. Relatives, friends. Colleagues..." and "Not merely hospice but our hospice. Not sad but joyous, a honeymoon."

The annotator who selected Sheet as the winner states, "`Although Average Author and the Author Story depict relationships between close family members, Sheet is more similar to the Author Story in Plot, Creativity, and Development by exploring the meaning and emotional responses to the term hospice.`" This annotator highlights the presence of subtext in the Sheet story and its attempts to mimic the author's cyclic style—"Each door held a story, a life in its final chapter... she would ask the same question..." and "Clara knew that work was not about erasing the pain or the grief, but about finding moments of grace within it...".

The annotator who selected Average Author as the winner states, "`Average Author aligns more closely with the Author's Story in terms of Plot and Language Use. This is because both stories center on someone taking care of a family member in their own home, and growing from a lack of peace into a sense of peace, which is expressed through the choice of language. Sheet also indicates a development from lack of peace to acceptance, but the familial tie is what pushes Average Author over the top.`"

The annotator who selected 'Tie' states, "`I think it will be hard for one of these to resemble the Author Story since it was so unique. In all, the two AI stories were much more like each other than the Author Story.`"

Overall, this example highlights the subjective nature of annotator preferences and the inherent difficulty of our annotation task which also explains the poor inter-annotator agreement.

Writing Prompt

"It’s terminal," the doctor told you as you, the one renowned as the world’s greatest hero, cradled your dying son, "sometimes, when the super genes mix, the outcome isn’t as expected."

Story Rules

Plot:

- Structure the story around a humorous and light-hearted approach to the hero’s journey of coping with their son’s illness. For example, the hero’s attempts to use their powers for mundane tasks, like making breakfast, could result in comedic mishaps, such as accidentally creating a tornado of cereal.
- Develop the narrative around the hero’s personal transformation and empowerment. Show the hero learning to accept their son’s condition and finding new ways to be a hero, such as advocating for genetic research or starting a support group for other super-powered families.
- Use a layered narrative structure by incorporating fictional news articles and memos. For instance, include a news article titled "World’s Greatest Hero Faces Greatest Challenge Yet: Parenting."
- Introduce a humorous misunderstanding or temporal misalignment, such as the hero misinterpreting a cryptic message from their future self, leading to comedic attempts to "fix" the timeline, only to realize the message was trivial.

Creativity:

- Anthropomorphize the hero’s powers or gadgets to add a whimsical touch. For instance, the hero’s cape could have a personality and offer sarcastic commentary, calling itself "The Cloak of Infinite Wisdom."
- Use a meta-satirical approach by framing parts of the story as news articles critiquing heroism and genetic engineering, such as "Hero’s Son Diagnosed: Public Debates Ethics of Super Genes."
- Blend historical and fantastical elements to reimagine the setting, incorporating legendary figures like a retired Valkyrie or a dragon healer who offer advice.
- Satirize financial and religious themes by depicting the hero consulting a celestial financial advisor to manage the costs of their son’s medical care.

Development (Character and Setting):

- Develop characters through their interactions and humorous dialogue. The son might have a mature, witty perspective on his condition, adding depth to their relationship.
- Highlight cultural and bureaucratic misunderstandings. A well-meaning but clueless government official could create comedic yet insightful exchanges.
- Focus on the hero’s emotional growth, showing them shifting from helplessness to advocacy for genetic disorders and healthcare improvements.
- Explore characters’ internal struggles and aspirations, such as the hero questioning their purpose beyond their powers, while their son expresses dreams and hopes despite his condition.

Language Use:

- Use playful and imaginative language to enhance whimsy, such as referring to emotional challenges as "the cape of contemplation" and "the gauntlet of hope."
- Incorporate humor and irony, with lines like, "We found a hero so committed to saving the world that they forgot to save themselves."
- Blend whimsical and formal tones, using phrases like "a shield forged from stardust and dreams" and "a heart as resilient as dragon scales."
- Add satirical disclaimers, such as "Heroic endeavors may vary," "Not all capes are created equal," and "Past heroics do not guarantee future success."

Table 8: Personalized Story Rules obtained using the Author Writing Sheet for a Writing Prompt for a Reddit author.

Persona Description
As a storyteller, you craft narratives that are both humorous and transformative, often centering around light-hearted approaches to conflict resolution. Your plots are ingeniously structured, whether through the comedic misadventures of a rat named Pip or the empowering journey of Margaret Rose in a fantastical court. You delight in creating layered narratives, such as those told through fictional news articles, where humor and misunderstanding play pivotal roles. Your creativity knows no bounds, as you anthropomorphize animals to add whimsy and employ meta-satire to critique both human and alien perspectives. You blend historical and fantastical elements seamlessly, setting tales in places like Windsor Castle with characters that include both real-world figures and mythical beings. Your imaginative approach often intertwines religious and financial themes, crafting narratives that are as thought-provoking as they are entertaining. Character and setting development are your forte, with characters coming to life through their humorous dialogues and interactions. You explore their growth and empowerment, crafting relatable figures who navigate their internal struggles and aspirations. Your settings are vivid and engaging, providing a rich backdrop for the characters' journeys. Your language use is playful and imaginative, enhancing the whimsical tone of your stories. You wield humor and irony with skill, using them to convey themes of misunderstanding and cultural critique. Whether through whimsical phrases or a blend of formal and playful language, your narratives are infused with a tone that is both magical and sophisticated, inviting readers into a world where the absurd and the profound coexist harmoniously.

Table 9: Persona description for a Reddit author, structured into four paragraphs, each corresponding to a distinct narrative category: Plot, Creativity, Development (Character and Setting), and Language Use.

Writing Prompt

"Welcome to Hell. To leave, all you must do is finish seven impossible tasks decided by your seven greatest sins. Four people have made it out so far. Will you be the fifth?"

Author Writing Summary (Language Use)

- The author frequently uses dialogue to reveal character traits and dynamics.
- The author often employs vivid imagery to create an immersive atmosphere.
- The author often uses rhetorical questions to convey a character's internal conflict or uncertainty.

Average Author

The air was thick with an acrid, sulfurous scent as Lucy stood at the threshold of Hell, a place she'd never believed in until now. A figure emerged from the shadows, tall and imposing, yet with a surprisingly warm smile. "Welcome to Hell," he said, his voice smooth and oddly comforting. "To leave, all you must do is finish seven impossible tasks decided by your seven greatest sins. Four people have made it out so far. Will you be the fifth?"

Sheet

I awoke on a bed of ash and cinders, the air around me thick with the scent of sulfur and regret. A laugh, echoing and absurd, filled the air—a demon guide, amused by my struggle, appeared at my side.

"All this for a toy soldier?" it chortled, eyes gleaming with mischief. "You must learn to let go, or you'll never find the exit."

As I navigated the shifting corridors, I pondered the truth in its words. Each step forward was a step away from the past, from the child clutching stolen dreams.

Delta

Flames licked the edges of a crimson sky, and the ground beneath my feet felt alive, pulsing with a malevolent energy. I stood on the precipice of an endless abyss, the cacophony of wailing souls echoing in my ears.

Emerging from the maze, I found myself in a grand hall, gilded and opulent, dripping with wealth beyond imagination. Coins rained from the ceiling, and jewels sparkled like fallen stars.

Greed beckoned me to take, to hoard, to drown in the riches that lay at my feet. But I knew the price of succumbing; this treasure would bury me alive.

Table 10: Example where Sheet outperforms Delta for Faithfulness to Author History. Sheet reflects the Author Writing Summary through vivid imagery, character-driven dialogue, and rhetorical questions, whereas Delta emphasizes setting and symbolism but lacks interactive dialogue and internal reflection.

Author Writing Summary for Language Use

- The author employs conversational and accessible language to engage the reader.
- The author uses humor and wit in dialogue to add depth to characters and situations.
- The author frequently uses rhetorical questions to engage the reader and provoke thought.
- The author uses repetition for emphasis and to reinforce key themes or ideas.

Persona Description for Language Use

Your language use is conversational and accessible, engaging readers with humor and wit. You employ vivid imagery to create immersive settings and use rhetorical questions to provoke thought. Repetition reinforces key themes, while a mix of short and long sentences controls pacing and builds tension.

Writing Prompt

The reason humanity has never experienced any world-ending events is that future humans keep travelling back in time to stop them. You know this because there's one of them in front of you now, telling you that you must absolutely not go into work today.

Summ

Evelyn's mornings were a symphony of simplicity... "Whoa, what year is it?" he blurted out, eyes wide with what could only be described as temporal jet lag... "Listen, Evelyn, you absolutely cannot go to work today." Evelyn blinked, her skepticism as thick as the cream in her coffee. "And why is that? Did the universe finally decide to give me a spontaneous day off?"... Evelyn took a slow sip of her coffee, considering her options. Call the police? No, too much paperwork. Wake up from this bizarre dream? Preferable, but unlikely.

Summ-nP

The day started like any other. The alarm buzzed obnoxiously at 6:30 AM, shaking me from a dreamless sleep. I fumbled to silence it, already dreading the dreary commute to the office... The broadcast droned on about the usual—traffic jams, a potential storm, and the mayor's latest scandal... until a peculiar sound interrupted my routine: a soft hum, followed by a voice behind me. "Good morning, citizen of the past!"

Table 11: Example showing the role of persona descriptions in enhancing Language Use for a Reddit author for Faithfulness to Author History. The Summ method captures rhetorical engagement, and repetition more effectively than its counterpart without persona descriptions - Summ-nP.

Author Writing Summary for Language Use

- Vivid and sensory language creates immersive atmospheres.
- Dialogue effectively reveals character emotions and tensions.
- The narrative voice carries a reflective and introspective tone.

Persona Description for Language Use

Your language use is vivid and sensory, painting immersive atmospheres that draw readers into your world. Through dialogue, you reveal the emotions and tensions that simmer beneath the surface, adding authenticity and depth to your characters. Your narrative voice carries a reflective and introspective tone, inviting readers to engage with the deeper themes of your stories and reflect on their own experiences.

Writing Prompt

After moving into their dream apartment, Romi witnesses a neighbor's tragic death, but as the days pass, she begins to suspect the act was somehow meant to shatter her own carefully constructed life.

Average Author

The apartment was a sanctuary of light. Sunlight spilled through the floor-to-ceiling windows, pooling onto the wide-plank oak floors and casting soft, dappled shadows across the walls. Romi had waited years for this—an apartment with high ceilings, a balcony overlooking the park, and an airy, open layout that made her feel as though she were floating above the city.

Summ

The apartment on the top floor had a view that stretched far beyond the horizon. Every evening, when the sun dipped below the skyline, Romi would stand by the window, tracing the fading line of light with her eyes, feeling the world shrink to just her and this slice of heaven. The move had been a leap, a daring escape from the mundane into the realm of dreams where everything seemed possible.

Summ-nP

The apartment was everything Romi had ever imagined for herself… perched high above the city with its sprawling windows that captured the sunrise like a secret whispered just for her. The floors were a deep mahogany that gleamed under the soft light… The creaking of the floors became a language of its own, a series of codes and messages that she was determined to decipher.

Table 12: Example showing that persona descriptions do not always help improve Faithfulness for expert sources like New Yorker. The Summ-nP captures vivid sensory details more effectively than the Summ when compared against the Average Story.

Writing Prompt

The world’s greatest hero has turned traitor and joined the forces of evil. Now, everyone wants to know why.

Author Story

"You were the chosen one!" the paladin screamed, pointing an accusatory finger at the prior hero. "You were to save us from evil, not join them!" The prior hero paused for a moment before nodding. "They had cookies."

The paladin was dumbfounded. "You sold away everything you were, sold out the mortal race...for *confections?*"

It is not that I am fed confections; it is that I make them, with a monster who treats me as their child, and I to her, as my mother—something I’ve never had.

It is not that I am promised riches; it is that I earn them, and am encouraged to grow by a leader who *cares* for their subjects.

Average Author

The city held its breath. The world’s former protector, now veiled in darkness, stood at the gates of ruin. "Why?" the crowd asked. But no answer came—only silence, heavy with betrayal.

His eyes, once symbols of hope, now flickered with uncertainty. A fallen hero. A world left wondering.

Sheet

In the shadow of my former self, I find clarity. "Not truth, perhaps," I admit, "but a different perspective. Perhaps the true villain is the one who blindly follows orders."

Our conversation is interrupted by a comedic spectacle—a minion, oblivious to the tension, tries to carry a stack of forms taller than himself. He stumbles, papers flying everywhere, and I can’t help but laugh.

The minion returns, this time with a clipboard. "New recruit orientation," he says sheepishly. "Welcome packet and all."

I take the clipboard, glancing over the mundane details. "Orientation, really? I thought I’d at least get a dramatic initiation ceremony."

Summary

In the heart of the once-vibrant city of Elysium, now shrouded in shadows, the air buzzed with whispers of betrayal. Streets that once echoed with laughter were silent, as newspapers screamed of the hero’s fall from grace.

"Will you stand with me, Alaric? Not as an ally of light or shadow, but as a seeker of truth?" He hesitated. Yet in the end, the promise of redemption proved irresistible. "I will," he declared. "For the sake of the city. For the sake of what we once stood for."

Table 13: Example where Sheet outperforms both Average Author and Summary for Similarity to Author Story in the Creativity category. Sheet captures the Author Story’s blend of humor and imaginative reinterpretation of the prompt, while Summary follows a more conventional hero-fall narrative and lacks the light-hearted tone.

Writing Prompt

You have been in a long, lonely coma. One day, Life and Death visit your consciousness, and get into a heated debate about what should be done with you. Indifferent to your fate, you like hearing them argue like an old married couple, and decide to try and set them up with each other.

Author Story

I've been drifting in this vast, empty void for so long... Akin to a leaf drifting away and about in a still pond, time doesn't really exist here, and the silence is so deep that it almost feels like it has weight… I'm trapped in some sort of elevator that not only goes up and down, but also left, right, and diagonally… And hey, if it works out, beyond recovering from my coma, I might just score a very special invite to the universe's most unlikely wedding.

Average Author

In the quiet echo of my mind, where only the soft hum of forgotten dreams lingered… It was like waking up, but not quite; it was more like tuning into a radio station I hadn't realized existed… It was like listening to an old married couple who'd had the same argument for centuries… I imagined them, Life with her radiant glow, hands on her hips, eyes sparkling with the passion of a thousand dawns, and Death, calm and composed, a gentle shadow with a hint of a smile playing on his lips.

Sheet

Consciousness is a peculiar thing. It floats somewhere between the tangible and the illusory, like a whisper that hangs in the air, just out of reach. In the coma world, where I have resided for what feels like an eternity, time has no meaning... "Wouldn't it be easier if you both just worked together?" I muse, my thoughts a ripple in the ether.

Life laughs, a sound like tinkling bells. "Oh, wouldn't that be something? But we are not meant to intertwine. We serve different purposes."... And in this place of dreams and whispers, I choose to live, fully and freely, in whatever form that may take.

Oracle

Floating in the vast, ethereal void of my own consciousness was akin to being a leaf drifting in a still pond—untouched by time, unanchored by space, and entirely alone. "Have you ever considered that maybe all this arguing is just... romantic tension?"

Life blinked, momentarily at a loss for words. Death's lips curled into a faint smile, his eyes twinkling with amusement. "Romantic tension?"

Table 14: Example for a Reddit author where Sheet underperforms Oracle in capturing the Author Story's lively and light-hearted tone and dialogue and instead leans toward introspection and philosophical reflection.

Writing Prompt

It is one man, by himself, in a castle on a hill. How does an ENTIRE ARMY fail to take it???

Persona for Creativity

Your creativity shines through in your ability to subvert genre expectations. You transform horror into comedy, as seen with the Addams Family's immunity to typical threats, and seamlessly integrate fantastical elements into realistic settings, such as isekai individuals influencing royal politics. Your narratives often blend the mundane with the mythical, creating a unique atmosphere that captivates readers.

Author Story

There a six-armed grey skin tall man sat on the walls, as his hands never stopped throwing rocks, and spears at the approaching army. Every now and then, he jumped off with some spears, and wrecked havoc amidst the enemy lines. The earth fell and rose under his steps, the metal of the weapons and armors of the enemy shifted by his will, and a thick mist followed him, hiding him....

Average Author

The Keeper raised his hand, and suddenly, the ground erupted with a series of blinding flashes. The front line of soldiers was thrown back, their bodies crumpling like rag dolls. The Keeper's voice, amplified by some unseen force, echoed down to the soldiers below. "You cannot take what is not yours. Leave now, and I will spare your lives...."

Sheet

The castle itself seemed to rise in defiance, its ancient stones humming with a silent strength. Alaric could feel the runes carved into the walls, pulsing faintly with an energy that had protected his forebears. It was as if the castle was alive, an ally in his solitary defense....

Sheet nP

The first attack came at dawn. Hidden pitfalls opened beneath their feet, swallowing men whole. Arrows rained down from the slits, finding gaps in armor and bringing down the attackers with deadly precision. Alaric moved through secret passages, launching surprise attacks that sowed confusion and fear....

Table 15: Example showing the role of persona descriptions in determining similarity to the Author Story for a Reddit author. The Sheet creatively blends fantasy and military strategy, making it more similar to the Author Story than the Average Author, while the Sheet nP focuses primarily on direct conflict resolution.

Category	**Lexical Overlap**		**Diversity (Homogenization)**		**Style**		
Method	**BLEU**	**ROUGE-L**	**ROUGE-L**	**BERTScore**	**Author History**	**Author Story**	**Length**
Ground Truth	–	–	0.1222	0.7015	0.7402	–	1074
Oracle	0.0311	0.1577	0.1427	0.7167	0.6385	0.7707	930
Average Author	0.0122	0.1448	0.1415	0.7190	0.6149	0.7306	985
RAG	0.0117	0.1463	0.1428	0.7212	0.6161	0.7359	947
Delta	0.0114	0.1450	0.1462	0.7217	0.6218	0.7391	929
Sheet	0.0102	0.1454	0.1440	0.7215	**0.6446**	**0.7532**	959
Sheet-nP	0.0114	0.1464	0.1433	0.7208	**0.6451**	**0.7544**	950
Summ	0.0108	0.1451	0.1460	0.7237	0.6324	0.7432	965
Summ-nP	0.0111	0.1450	0.1445	0.7222	0.6400	0.7459	939

Table 16: Comparison of methods across Lexical similarity, Diversity, and Style metrics for GPT-4o generated stories.

Plot
Narrative Complexity and Thematic Depth: Multilayered plots with reflective themes, moral ambiguity, and unresolved conflicts create an immersive and thought-provoking narrative. *Example: A protagonist grappling with guilt over a past decision while navigating societal expectations, or a story where the true nature of a villain's actions remains ambiguous, inviting reader interpretation.* **Blending Internal and External Conflict**: A balance of personal introspection with broader societal or external challenges makes the plot more dynamic and engaging. *Example: A hero torn between duty and personal desires while leading a rebellion, or a scholar confronting both supernatural forces and existential questions about memory and identity.*
Creativity
Rich World-Building and Narrative Expansion: New characters, layered settings, and cultural influences enhance the world, making it feel more immersive and dynamic. *Example: The inclusion of deities or councils of gods to deepen the plot, or the integration of cultural traditions like Native American folklore or Indian recipes to enrich family dynamics.* **Symbolism and Thematic Complexity**: Metaphors and symbolic elements add depth and nuance, aligning closely with the author's writing style. *Example: A magical locket representing unity, a tapestry symbolizing tradition, or a river serving as a metaphor for personal transformation.* **Blending Genres and Narrative Styles**: A seamless merging of fantasy with realism, humor with serious themes, and structured storytelling with ambiguity creates a more engaging and layered reading experience. *Example: A surreal sequence where a protagonist questions reality, or a rebellion story that integrates strategic alliances and magical artifacts.*
Development
Emotional Depth and Internal Growth: Character introspection, moral dilemmas, and evolving emotional states create nuanced and engaging development. *Example: A protagonist struggling with guilt over past failures while trying to redefine their sense of heroism, or a character making a difficult moral choice between revenge and justice.* **Nuanced and Evolving Relationships**: Complex interpersonal dynamics, including friendships, family bonds, and mentorships, develop over time rather than being static or one-dimensional. *Example: A father gradually redefining his view of heroism through his child's perspective, or a mentor-student relationship evolving from skepticism to deep trust.* **Collaborative and Reflective Growth**: Characters grow through interactions with others, rather than relying solely on individual realizations or isolated moments of change. *Example: A team learning to work together to achieve a mission, or a character undergoing self-discovery through dialogue and interactions with a diverse group of companions.*
Language Use
Rich Imagery and Descriptive Depth: Vivid descriptions, metaphors, and figurative language create immersive scenes and evoke emotions. *Example: "Her laughter rang out, a sound as bright and clear as a summer's day," or "The driveway was cracked and veined with weeds, a testament to the years that had slipped by."* **Varied and Expressive Tone**: A mix of lightheartedness, introspection, and poetic expression creates a dynamic and engaging narrative voice. *Example: "Are you sure this is wise, my love?"—"Wise? No. Fun? Absolutely." for humor, or "Her dream home had become a prison of uncertainty and fear." for a more reflective and emotional impact.*

Table 22: Analysis of annotator comments explaining the preference of our personalization methods over the Average Author method in terms of similarity to the ground truth author story, organized by narrative categories.

Writing Prompt

"It's terminal," the doctor told you as you, the one renowned as the world's greatest hero, cradled your dying son, "sometimes, when the super genes mix, the outcome isn't as expected."

Ground-truth Author Story

# Syntax Error: Unexpected Power Captain Strongarm, defender of Metro City and seven-time winner of Most Cars Thrown At Villains“ award, had never felt so powerless. The hospital machines around his thirteen-year-old son Tommy were having seizures - screens flickering, numbers scrambling, error messages cascading. „It's terminal,“ Dr. Roberts said, looking at a screen that was now displaying what appeared to be the entire script of ‚The Matrix‘ in real-time. Sarah, his wife (superhero name: The Nullifier), grabbed his hand hard enough to crack concrete. ......„Hey dad, want to learn Python? I promise it's not actual snakes this time!“ Captain Strongarm smiled. His son wasn't following in his footsteps - he was coding his own path. And maybe that was even better. „Sure, son. Just... no more downloading martial arts directly into my brain. The kung fu gave me a migraine....Life wasn't what they expected, but then again, the best source code usually comes with a few surprise features. `End_process story.exe` `Exit status: Happy`

Average Author

"It's terminal," the doctor told you as you, the one renowned as the world's greatest hero, cradled your dying son, "sometimes, when the super genes mix, the outcome isn't as expected." The room felt like it was closing in on me, the sterile smell of antiseptic failing to mask the sour stench of fear. My son, little Ethan, lay pale and fragile in my arms, his once vibrant eyes now dull and glassy. This was not how it was supposed to go. Heroes don't have children who die. Heroes don't face problems they can't punch, fly away from, or blast with heat vision. “Isn't there something we can do?” My voice cracked, and I hated how weak it sounded.......Gently, I laid him down, wrapping him in his favorite blanket. The sky was painted in hues of orange and pink, a beautiful contrast to the sorrow in my heart.

Sheet

The Hero's Journey: A Tale of Capes and Cereal In the heart of the mythical city of Avalon Heights, where ancient heroes once strode and the skies shimmered with the glow of enchanted skyscrapers, the world's greatest hero faced their most daunting foe: the frailty of life itself. The hero, known to the world as Titania, Queen of the Elements, held her son, Max, in her arms as the doctor delivered the news...... The following day, a fictional news article took center stage in *The Avalon Gazette*: "World's Greatest Hero Faces Greatest Challenge Yet: Parenting." In the midst of this storm, Titania's attempts to find normalcy became legendary. One particularly memorable morning, she decided to make Max a simple breakfast. But as she reached for the cereal, her powers—perhaps expressing her inner turmoil—manifested with unintended flair. A tornado of cereal erupted, swirling through the kitchen like a sugar-coated cyclone. "Mom," Max said, giggling despite his illness, "I think you just invented the breakfast hurricane...... In a final act of empowerment, Titania founded a support group for super-powered families, advocating for genetic research with the same fervor she once reserved for battling cosmic threats. Her growth from feeling helpless to becoming a beacon of hope for others was a transformation as magical as Avalon Heights itself. Max, with dreams and hopes as boundless as the sky, taught her that heroism wasn't defined by powers but by the courage to embrace life's uncertainties.

Table 23: Qualitative example where the Sheet method is preferred over Average Author by all annotators for a Reddit author. Sheet captures the author's creative use of humor, satire, and unconventional programming jargon, which the Average Author misses.

Writing Prompt
"When the word ’hospice’ is first spoken, it feels like the air itself changes, heavy with finality. How do you create light in a space where hope seems to have vanished?"
Ground-truth Author Story
"Hospice." Once the word is uttered aloud, there is a seismic shift. You will feel it. Like a (very short) thread through the eye of a needle, swiftly in and swiftly out. The air itself becomes thin, steely. At the periphery of your vision, an immediate dimming. The penumbra begins to shrink. In time, it will become a tunnel. Ever diminishing. Until the remaining light is small enough to be cupped in two hands. And then it will be extinguished. For when "hospice" is spoken, the fact is at last acknowledged: There is no hope. No hope. These words are obscene, unspeakable.... Well, I will try! When he is at home, possibly his appetite will return. When I am the one to prepare his food, his appetite will return, I am sure. Of course, family will come to visit. Adult children, grandchildren. Relatives, friends. Colleagues from the university. Neighbors. Old friends from grammar school he hasn’t seen in fifty years. Some surprises for him—I will negotiate with the imagination of a theatre director. Not merely hospice but our hospice. Not sad but joyous, a honeymoon. We will be happy there, in our own home. Both of us. For both of us, the "final days" will be a honeymoon. I vow. In fact, nothing remotely like this will happen. How could you have imagined it would! Hospice, yes. Honeymoon, no.
Average Author
"When the word ’hospice’ is first spoken, it feels like the air itself changes, heavy with finality...It’s her mother, Marjorie, who the doctor is addressing, but the word reverberates through Clara’s own heart, a dull thud of inevitability." Marjorie stares at the doctor with a mix of resignation and defiance, her once-vibrant green eyes now clouded by the long battle with illness..... Her touch is light, almost ethereal, and Clara feels a sudden rush of tears she struggles to hold back. “Clara,” Marjorie says softly, her voice barely a whisper. “Do you remember the garden?” Clara nods, unable to speak. The garden had been Marjorie’s sanctuary, a riot of color and life where she had spent countless hours tending to roses, tulips, and wildflowers...... Marjorie’s decline is gradual but inevitable, and Clara feels the weight of each passing day. She continues to read to her mother, to sit by her side, and to hold her hand through the long nights.
Sheet
Every morning, Nurse Clara began her rounds at St. Mary’s Hospice, a place where the word "hospice" hung in the air like a thick, unspoken truth. Her day started with the same cycle: a clipboard in hand, she moved from room to room, her steps echoing softly in the quiet corridors. Each door held a story, a life in its final chapters, and as she entered each room, she would ask the same question, "How are you feeling today?" The responses were as varied as the lives they represented. Some patients greeted her with a weary smile, others with a silent nod. But there was one patient, Mr. Thompson, who never failed to reply with a simple, "I’m just waiting." Clara had come to understand that "waiting" was an ambiguous term within these walls. It held a duality of meaning: waiting for relief, waiting for peace, waiting for the inevitable.....Clara knew that the work was not about erasing the pain or the grief, but about finding moments of grace within it. It was about guiding others, and herself, to a place where hope could be redefined—not as a denial of reality, but as an acceptance of it, and the peace that such acceptance could bring. With each cycle of her rounds, Clara realized that the light had always been there, waiting to be uncovered, and that she had the power to help others see it too.

Table 24: Qualitative Example where annotators select different winners for a New Yorker author. The Author Story uses metaphor and subtext to reframe hospice as both a physical and emotional space. While Sheet captures these elements through cyclic structure and introspective tone, the subjective nature of subtext leads annotators to different interpretations of similarity.

Writing Prompt
"Late at night, the King of Rohan and his queen sneak into the pantry for a forbidden snack, only to find that stealing cured venison is far more complicated—and far more amusing—than they anticipated."
Ground-truth Author Story
If anyone—a kitchen boy, a maid, or, Eru forbid, even the housekeeper Wynhild—had wandered near the pantry in the quiet hour just before dawn, they would have heard a muffled exchange from the depths of the storeroom: “Try shifting this bowl here. . . ” “Easy, you’re on me.” “Oh, does that bother you, my lady?” “Get off me; I can’t pull it out!......Lothíriel watched her royal husband with a half-incredulous, half-amused smile, as he held the large salami with the pride of a merchant at the autumn fair......To Lothíriel, he looked like a great golden-haired squirrel with the morsel stuffed in one cheek. How he could be both a fierce warrior and a wise sovereign as well, was a mystery she had yet to solve. His voice came out muffled as he swallowed. “You’re right, but if it ever gets out, I’ll say the sweet Queen of the Mark was struck by an irresistible craving for spiced meat in the dead of night—and who would deny anything to the future mother of the next royal heir?” He crossed his arms, looking quite pleased with his idea, but Lothíriel’s eyes widened at the bold lie, and she covered her face with her hand, foreseeing trouble ahead.
Average Author
Éomer, King of Rohan, tiptoed through the dimly lit corridors of Meduseld, holding his breath as he reached the door to the pantry. He glanced back at Lothíriel, his queen, who followed close behind, stifling a giggle. The flickering torchlight cast playful shadows on the stone walls, and Éomer couldn’t help but feel a rush of exhilaration. “We shouldn’t be doing this,” Lothíriel whispered, though her eyes sparkled with mischief. “Ah, but where’s the fun in being King and Queen if we can’t indulge in a little late-night adventure?..... “You’ve always had a knack for adventure. It’s one of the reasons I fell in love with you.” Lothíriel’s expression softened, and she leaned in to kiss him, the taste of venison still on her lips. “And I love you for indulging my whims, even when they lead us to pilfering from our own pantry.”..... Éomer took Lothíriel’s hand, and they slipped out of the pantry, leaving behind only the echo of their laughter and the memory of a stolen moment in the heart of Rohan. “Next time,” Éomer whispered as they crept back to their chambers, “we’ll try the cheese.”
Sheet
The moon cast a silvery glow over Meduseld, its light filtering through the high windows to illuminate the grand halls of the Golden Hall. Éomer, the formidable King of Rohan, found himself tiptoeing like a child on a dare, his hand clasped firmly in Lothíriel’s, whose laughter was a soft melody in the silence..... “Are you sure this is wise, my love?” “Wise? No. Fun? Absolutely,” she replied, her laughter a soft ripple in the silence. Éomer shook his head, a reluctant smile tugging at his lips as he watched her nimble fingers work to untie the savory prize. He was a warrior, a king, yet here he was, caught in the delightful absurdity of this secretive adventure, his heart light with the thrill of it all. “Besides,” Lothíriel continued, handing him a piece of the cured venison with a flourish, “what’s the worst that could happen? A tale of the gallant King brought low by his own appetite?” Éomer took the offered morsel, his grin widening as he tasted the rich, smoky flavor. “Perhaps the bards will sing of our escapades instead of my battles,” he mused, the thought surprisingly pleasing. .....As they slipped back into the corridors, leaving the pantry and its shadows behind, Éomer realized that it was not the venison he would remember, but the adventure, the laughter, and the love that had made the night truly unforgettable.

Table 25: Example where all annotators prefer the Average Author over Sheet for an AO3 author. The Average Author aligns better with the story’s playful tone and king–queen dynamic, aided by LLM familiarity with *Lord of the Rings* fanfiction tropes. Sheet adds internal conflict that doesn’t fit this prompt.

Writing Prompt
"In a desert town under the watchful eyes of royal guards and oppressive banners, you and your team have five minutes to locate a mysterious contact in a pink house before chaos erupts—failure is not an option."
Ground-truth Author Story
Chapter 1: Danger Close *La Makan, 12:54 standard time* Wind whistled through the close streets of *La Makan*, stirring grains of sand with all the effect of a man attempting to blow salt through eternity. It whipped around the corners of the alleys, rustling banners hanging from balconies and glassless windows. It looked for something, a thing it would never find..... Either way wouldn't end well for them. Our contact gave us an area, but not an address. We have, he checked his watch, five minutes until the distraction starts. We need to find them and get them out. Problem is, we don't know what they look like, only that their house is pink. He glanced over his shoulder at the multicolored buildings. Any one of the scoured pink buildings could've been their one. "Vic, you knew Xitor the best of any of us. I'm trusting you to get us there. Everyone else, let's keep our hats on, kay?"
Average Author
The sun hung low in the sky, casting long shadows across the dusty streets of the desert town of Vespera. Time was of the essence; you had five minutes to locate your mysterious contact before chaos erupted. Failure was not an option. "Stay sharp," you whispered to your team, your voice barely audible over the din of the market. "But time is short. We must act quickly." The weight of the moment settled over you like a heavy mantle. The fate of Vespera rested on the success of this meeting. As the figure began to speak, outlining the plan that would either save or doom the town, you couldn't shake the feeling that this was only the beginning. The real challenge was yet to come.
Sheet
The sun beat down mercilessly, casting a harsh glare on the town's whitewashed walls. The narrow alleys were alive with whispers and hurried footsteps, each person a part of the unfolding drama. My team and I moved through the bustling streets, the oppressive presence of royal guards a constant reminder of the stakes. We had five minutes—no more, no less—to locate the contact in the elusive pink house before chaos erupted. The memory of a previous mission flickered in my mind: the sun setting behind the dunes, the sound of distant drums, and the rush of adrenaline as we narrowly escaped a similar fate. That mission had gone wrong, leaving scars that still haunted my dreams. Failure was not an option today... Outside, the tension remained, a silent storm waiting to break. But for now, we had succeeded. The town's frenetic energy continued, unaware of our brief triumph. The banners flapped, the guards watched, and the countdown to chaos continued, but we had found our mark. The mission was still on course, and the desert town held its breath, waiting for the next move.

Table 26: Qualitative Example where all annotators select 'Tie' for a Storium author. While the Author Story uses an open-ended narrative with unresolved character dynamics, both the Average Author and Sheet introduce resolution and tension-driven plots, missing the author's intended tone.