CAsPER in the Machine: Insights into Character Variety in LLM-Generated Stories

Anneliese Brei[1] **Abhisheik Sharma**[2] **Nicholas Sanaie**[1]
Lu Wang[3] **Snigdha Chaturvedi**[1]
[1]UNC Chapel Hill [2]Georgia Institute of Technology [3]University of Michigan
abrei@cs.unc.edu, asharma914@gatech.edu, nsanaie@unc.edu,
wangluxy@umich.edu, snigdha@cs.unc.edu

Abstract

As LLM-generated text is increasingly used, especially in fictional domains, we explore how much LLM-generated stories differ from human-written stories. In this work, we focus on characters. We borrow definitions from narratology to analyze 8 intricate category-pairs of character, such as *stylization* and *wholeness*. These category-pairs consider more than just basic characteristics. They assess how characters are portrayed within their stories. After automatically inferring categories of characters within both LLM and human-written stories, we compare and contrast these two sets of stories. We consider the following overarching questions: (1) Do LLMs and human-written stories have similar characters? and (2) Do LLMs generate stories with a variety of characters? Our analysis includes research questions that focus on stories generated by popular LLMs and recently published human-written stories. We describe a number of interesting similarities, differences and key takeaways.[1]

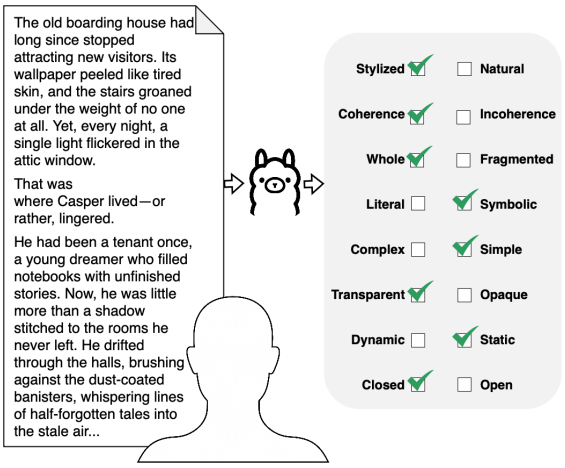

Figure 1: We analyze characters in LLM-generated short stories (left) using 8 category-pairs (right) that consider how characters are portrayed. (e.g., a character might be represented in a *fragmented* way, setting the tone for a disjointed mood within the story). We compare LLM-generated characters with human-written characters other and LLM-generated characters from different model sizes and families.

1 Introduction

Increasing numbers of authors are using AI tools to assist in the process of writing stories (Mirowski et al., 2024; Chakrabarty et al., 2024b; Mirowski et al., 2023; Ippolito et al., 2022; Yuan et al., 2022). For example, Large Language Models (LLMs) are highly versatile for personalization, making their use ideal for tailoring writing to specific needs (Chakrabarty et al., 2024a; Wasi et al., 2024; Nicolicioiu et al., 2024). In the digital humanities, LLMs play an important role for creative writing by assisting with textual analysis, interpretation, and generation (Cigliano et al., 2024). While LLMs continue to improve at story generation, many writers and readers wonder whether or not LLMs are capable of generating interesting stories like humans. This multi-faceted question is complicated to address.

Narratologists have established that character and plot are two integral aspects of story (Janko et al., 1987; Campbell, 2008). In fact, theorists increasingly emphasize that the evaluation of character should be given additional priority because the role and function of character helps to define other elements of the story such as themes, tone, and culture (Bal, 1997; Forster, 1927). For example, if a narrative sequence is repeated among stories, but the characters themselves are different, then the overall takeaway messages of these stories are likely to differ (Phelan, 1989; Propp, 1968). However, while recent work analyzes LLM-generated stories with respect to narrative development via plot (Tian et al., 2024), as of now, no existing works explore these stories' portrayal of character at a level deeper than personality traits.

In this paper, we investigate how LLMs represent character in stories they generate by asking,

[1]All code and annotated data is released publicly: https://github.com/adbrei/casper

Do LLMs generate stories with a variety of characters? This question is challenging because it cannot be answered by looking only at basic character attributes like demographics or basic personality. Instead, there must be a deeper understanding of how the character is portrayed. To analyze characters within the light of their presentation style in a wholesome way, Forster (1927) famously coins the terms "flat" versus "round" characters, where either type of character is ultimately determined by how the character is presented to the reader. However, though narratologists attempt to understand and define these terms, to this day the terms remain vague and subjective to interpretation (Jannidis, 2019; Phelan, 2005; Bal, 1997; Chatman and Chatman, 1978; Booth, 1983). As such, the definitions of "flat" versus "round" are not easy to operationalization automatically.

With this in mind, we turn to a related character classification (Hochman, 1985) inspired by Forster (1927)'s analysis of character. This classification is a taxonomy that looks at finer and more automatically measurable categories in a flexible and more concrete fashion (Fishelov, 1990). It considers eight primary categories of character portrayal, including *stylization*, *coherency*, *wholeness*, *literalness*, *complexity*, *transparency*, *dynamism*, and *closure*. We design a framework, *Character's Portrayal Classifier* (CASPER), that considers these categories. By comparing each primary category with its opposite (e.g., *stylization* vs. *naturalism*), we create a metric to evaluate characters according to factors of storytelling. For example, a character's *transparency* relies on how their thoughts and motives are presented to the reader (Fishelov, 1990). Thus, we are enabled to perform more nuanced explorations of the types of characters in generated stories, as shown in Figure 1.

Using these category-pairs, we compare characters from human-written and LLM-generated stories and analyze the distributions of character category-pairs from a variety of popular LLM families and sizes. We explore 6 research questions that address (1) the similarities and differences of LLM and human-written stories, using both coarse and fine-grained levels of comparison and (2) how LLM-generated stories differ, such as across model size and family, genre of story, and across multiple inference calls. Answering these questions helps us to determine if LLM-generated characters have patterns with respect to other LLM or human-written characters, as well as the diversity of LLM generations. We provide key takeways from these research questions in Section 6. Our primary contributions include:

- We adapt and modify theory-grounded methods of character analysis to automatically understand characters in LLM-generated stories. Our framework can be used to keep track of how future LLM-generated stories portray character;
- We construct a high-quality dataset of human-written and LLM-generated short stories that underwent careful curation to ensure comparability across genres and themes. It will be publicly released for future research.
- We compare characters in generated stories to characters in human-written and other generated stories and provide a detailed analysis of similarities and differences.

2 Related Works

Narrative theory is used in natural language processing to provide a theoretical framework for computational narrative understanding (Piper et al., 2021). As LLMs show increasingly impressive capabilities for text generation, much work attempts to distinguish LLM from human-written text (Boutadjine et al., 2025; Russell et al., 2025; Elek et al., 2025; Godghase et al., 2025; Ali et al., 2025; Najjar et al., 2025; Venkatraman et al., 2024; Harrag et al., 2020; Uchendu et al., 2024; Gehrmann et al., 2019). Some works perform deeper analyses of specific qualities of LLM-generated text, such as analyzing discourse similarities and linguistic features (Namuduri et al., 2025; Reinhart et al., 2025; Chong et al., 2023) and narrative elements such as plot (Tian et al., 2024).

Prior work on automatically understanding characters within text reuse methods appropriate for studying real humans, such as personality profiling (Shu et al., 2024; Jiang et al., 2024) and tracking emotions (Brahman and Chaturvedi, 2020; Rahimtoroghi et al., 2017; Chaturvedi et al., 2016). Other work analyzes gender bias (Lucy and Bamman, 2021; Huang et al., 2021), roles (Jang and Jung, 2024; Bamman et al., 2013), relationships between characters (Vijjini et al., 2022; Kim and Klinger, 2019; Chaturvedi et al., 2017; Iyyer et al., 2016; Srivastava et al., 2016) character setting (Soni et al., 2023), and traits that relate characters to humans in society (Yuan et al., 2024; Jaipersaud et al., 2024; Brahman et al., 2021; Yu et al., 2024; Li et al., 2023; Yu et al., 2023; Bamman et al., 2019).

	Primary Category	Opposing Category
1	**Stylization**: A character is depicted through deliberate idealization, idealization, exaggeration, or conventional artistic patterns, often emphasizing form or artifice.	**Naturalism**: A character is depicted with close attention to ordinary human traits and behaviors, aiming for lifelike accuracy without obvious artistic distortion.
2	**Coherence**: A character's actions, speech, and inner life align in ways that form a consistent and logically understandable pattern.	**Incoherence**: A character's actions, speech, or inner life lack consistency, producing contradictions, unpredictability, or fragmentation.
3	**Wholeness**: A character is presented as a fully integrated being, with sufficient detail provided to give the sense of a complete personality.	**Fragmentariness**: A character is presented only in partial aspects, leaving significant gaps in their personality, background, or presence.
4	**Literalness**: A character functions solely as an individual within the story world, without additional layers of meaning attached.	**Symbolism**: A character functions as both an individual and as a representation of an abstract idea, theme, or cultural concept beyond the story.
5	**Complexity**: A character shows multiple, sometimes conflicting traits, desires, or motivations that interact in nuanced ways.	**Simplicity**: A character is defined by one or very few dominant traits or motivations, with little internal tension.
6	**Transparency**: A character's inner life—thoughts, emotions, and motivations—is made clear to the reader, often through narration or explicit cues.	**Opacity**: A character's inner life remains hidden, ambiguous, or difficult to interpret, leaving the reader uncertain about their drives or reasoning.
7	**Dynamism**: A character undergoes noticeable development or transformation over the course of the narrative.	**Staticism**: A character remains fundamentally unchanged in outlook, behavior, or values throughout the narrative.
8	**Closure**: A character's narrative arc is brought to a clear resolution, with all questions about them settled by the story's end.	**Openness**: A character's narrative arc remains unresolved, with key aspects of their fate, choices, or significance left indeterminate.

Table 1: Eight category-pairs with definitions. Primary-category (left) is contrasted by opposing-category (right).

However, these works do not consider how a fictional character is portrayed within the artistic framework of the story, which is the focus of this work. We refer to narratological discussions to explore how character development and story shape the reader's perception of the character (Smith, 2022; Carter and Pickett, 2014; Janko et al., 1987; Propp, 1968). In particular, we focus on categorization of character, which best known from Forster (1927) and further developed by later narratologists (Phelan, 2005; Rimmon-Kenan, 2003; Bal, 1997; Fishelov, 1990; Hochman, 1985; Chatman and Chatman, 1978; Booth, 1983).

3 CASPER Framework

In this section we define categories, formulate the task, and find the best framework.

3.1 Defining Category-Pairs

In order to evaluate character as a component of story-telling within story, we consider 8 category-pairs, described in Table 1 (Hochman, 1985). We choose this framework instead of others that define "flat" versus "round" characters,[2][3] because these categories have more tangible definitions and are better suited for classification with LLMs.

Each category-pair consists of a category compared to its opposite. For example, character *stylization* is contrasted with *naturalism*, *coherence* with *incoherence*, *wholeness* with *fragmentariness*, and so forth. Human writers may choose the extreme into which their characters fall; hence in the real-world we find characters fitting every category.

To further illustrate these category-pairs, we provide examples from popular fiction. Stereotypically, fairy tales and moral-centric stories contain *stylized* characters (Jung, 1968; Pearson, 1991), such as "the Innocent" (often a naive young girl with ample trust and virtue who eventually learns maturity) or "the Sage" (often an old man with a long beard who currently has profound knowledge, yet seeks deeper knowledge). In the Harry

[2]For example, we consider Forster (1927)'s predominant definitions of "flat" ("built around a single idea or quality; they are predictable and do not change") and "round" ("complex, multi-dimensional, and capable of surprising the reader"). Such definitions are coarse and often tricky to distinguish.

[3]Though Hochman (1985) does not explicitly tie the framework to definitions of "round" vs. "flat" characters, other narratologists state that both classifications categorize characters for the same objective, with Hochman (1985)'s taxonomy showing desirable flexiblility (Jannidis, 2019; Fishelov, 1990).

Potter series (Rowling, 1997–2007), Albus Dumbledore fits the *stylized* archetype of the wise sage with profound knowledge. Similarly, the Dursleys demonstrate *stylization* because they represent mediocrity through exaggerated domestic satire. On the other hand, Molly Weasley is *natural* because she is portrayed as a mother with authentic motivations, actions and feelings (e.g., humor, love, pride, frustration, fear, anger). Hermione Granger is *coherent* because her motives are revealed, and her actions make sense in light of her personality. Further examples for the rest of the category-pairs are given in Appendix A.1, and examples from our corpus are given in Appendix A.2.

With each of these category-pairs, we propose *Character's Portrayal Classifier* (CASPER) for automatically understanding characters in stories.

3.2 Task Formulation

As presented in Figure 1, we aim to classify characters using the category-pairs in Section 3.1. Given a short story with character c, we consider the eight pairs: $\mathcal{A}_c = \langle A_{Sty.}, A_{Coh.}, A_{Whol.}, A_{Lit.}, A_{Comp.}, A_{Tran.}, A_{Dyn.}, A_{Clos.} \rangle$. Each component A is evaluated by determining if c best fits the category (e.g., $A_{Sty.}$ indicates c is *stylized*) or its opposite (e.g., $\bar{A}_{Sty.}$ indicates c is *natural*).

3.3 Identifying Categories of Character

First, we describe how we classify characters. Manual classification is costly and time-consuming due to necessary literary expertise and high text volume. To perform the intended analysis at scale, we need an automatic classification method. We choose to follow recent works using LLM-as-a-judge (Zheng et al., 2023) which allows us to do classifications at scale and inexpensively. However, this method has its own challenges because the task can be subjective and definitions can be difficult to interpret. Additionally, judges sometimes struggle with forms of bias and lack of knowledge within a specialized context (Li et al., 2025). Since CASPER uses categories specific to narratology, we observe a judge could have difficulty understanding literary nuances. To mitigate these potential issues, we test numerous settings using multiple models, different forms of definitions, various prompts and zero-shot and in-context learning (ICL) (Dong et al., 2024). We pick the best setup by comparing the results to human annotations.

For task formulation, we (a) classify category-pairs individually, using one template for each pair; (b) classify all category-pairs in a single inference call; (c) try a Likert-scale rating to evaluate category-pairs in a more fine-grained manner.

For describing category labels, we use (a) definitions that paraphrase the explanations in (Hochman, 1985), (b) descriptive adjective lists[4], or (c) the combination of definitions and adjective lists.

For settings with ICL, we experiment with several template layouts, including (a) a "basic" layout that describes the categories before providing an example of each; (b) an "interleaved" layout that provides an example of a category immediately following the respective description; and (c) a "repeated" layout that first describes the categories, provides examples, and then repeats the same description of the categories. The formatting of our templates are given in Appendix B.1.

To test these templates, we create a test set of LLM and human-written short stories. We collect 50 human-written stories and 50 corresponding writing-prompts from `r/WritingPrompts`[5]. From the writing-prompts, we generate 50 new stories with `GPT4o-mini`. All 100 stories are annotated for each category-pair by experts who are native English speakers and are familiar with the research and literary theory. The resulting Cohen's kappa inter-rater agreement for all category-pairs is $\kappa = 0.56$, showing moderate agreement (Landis and Koch, 1977) (more details in Appendix B.2). We keep the size of our task manageable by considering only protagonists of stories since protagonists drive story flow and largely define narrative shape and genre. For these reasons, we consider only protagonists for the other experiments as well.

Finally, we evaluate our templates on the labeled stories in our test set using our largest open-sourced LLM, `Llama-70B`. Note, we also experiment with `GPT4o-mini` and find that both LLMs give comparable results (see Appendix B.3). We compare Macro F1-scores between the gold labels and predicted labels, and we determine that it is best to use a binary classification setup with separate inference calls for each category using only the definitions of the category-pairs. While ICL-interleaved performed well for all category pairs, to maximize performance, we choose the best-performing method for each category-pair: **ICL-repeated** for *styliza-*

[4]The adjective list is a list of synonyms, effectively summarizing each category in a manner akin to a thesaurus lookup. E.g., *transparent*: clear, obvious, apparent, straightforward, readable, comprehensible, explicit, direct, and understandable.

[5]https://www.reddit.com/r/WritingPrompts/

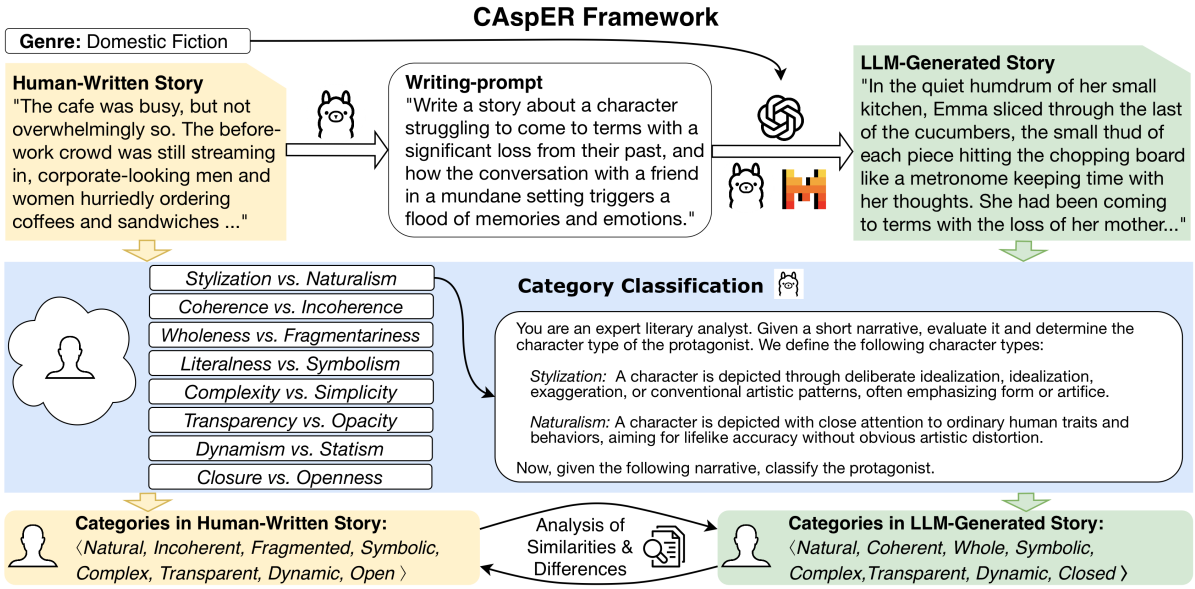

Figure 2: Overview of CAspER framework, including the pipeline for the creation of the corpus (top row), experiments (middle row), and analysis of the categories (bottom row).

tion, **ICL-interleaved** for *coherence*, *wholeness*, *complexity*, and *transparency*, **ICL-basic** for *literalness*, and **Zero-shot-basic** for *dynamism* and *closure*. A description of the results and F1 scores breakdown is given in Appendix B.3.

4 Evaluating Stories

In this section, we describe our process of obtaining short stories generated by humans and LLMs for CAspER. Then, using the best templates described in Section 3.3, we perform experiments to understand the types of characters in our corpus.

4.1 Corpus Creation

For a broad representation of stories that might be found in the wild and to ensure comparability of stories, we seek stories that fit into the following 4 widely encompassing genres (Fong et al., 2013): (1) *Domestic*: explores everyday life, family, relationships, and personal conflicts; (2) *Romance*: focuses on a romantic relationship; (3) *Science-Fiction/Fantasy*: emphasizes futuristic technology or magical realms; (4) *Suspense/Thriller*: maintains a sense of urgency and tension through plots by putting characters into imminent danger.

We collect short stories written by humans from `r/shortstories`[6], one of the largest and most popular subreddits for writers. We take multiple steps to ensure that this collection is high-quality. First, we only consider submissions from the past year to avoid stories potentially used as training data for LLMs. Second, we conduct experiments to ensure these stories are (1) from the same textual domain as the stories in our test set from Section 3.3 and (2) not likely to be LLM-generated (see Appendix E). Third, all stories have been previously tagged with genre labels by their authors (see Appendix G for the mapping of subreddit genre labels to our genre definitions). To ensure these genres are not ambiguous, we use `Llama-70B` to classify the stories and eliminate all whose predicted genres do not match the subreddit labels. In this way we collect a total of 200 high-quality human-written stories (50 stories per genre).

We obtain short stories generated by LLMs in a two-step process, and we take multiple steps to ensure that these stories are directly comparable to the human-written stories. First, in order to elicit LLM-generated stories that are of comparable theme to the human-written stories, we use our largest open-source model, `Llama-70B`, to generate writing-prompts from the human-written stories. To optimize comparability across genres, we also specify that the writing-prompt should be appropriate for the genre of the human-written story. We use the following template, and manually verify reasonable writing-prompts are produced:

[6]https://www.reddit.com/r/shortstories/

For this story, give the best writing-prompt from which this story is written (i.e., it should be obvious the story is written from this writing-prompt). Please ensure that it is particularly suitable for the genre, {*genre*}.

Finally, for comparability across technical settings, for all open-sourced LLMs we generate short stories from each writing-prompt 3 times:

You are a creative[a] story writer. Use the writing-prompt below to generate a complete short story in the following genre: {*genre*}.

[a]We test the word "creative" does not bias the LLM towards generating characters of a certain type (e.g., more dramatic characters). See Appendix D.

Our final corpus contains 200 human-written and 4400 LLM-generated stories. See Table 8 in Appendix F for corpus statistics.

4.2 Setup and Experiments

We perform a comprehensive analysis that explores a broad range of LLM-generated short stories using popular LLMs of 8 different sizes from 4 families. We test one closed-source (`GPT4o-mini`) and 7 open-source LLMs (`Llama-3B`, `Llama-8B`, `Llama-70B`, `Phi3-4B`, `Phi3-14B`, `Mistral-7B`, `Mistral-24B`).[7] Since we seek creative outputs while maintaining consistent model settings for evaluation purposes, for all inference we set $temperature = 0.7$ and $top\text{-}p = 0.9$ for open-sourced LLMs and default hyperparameters for `GPT4o-mini`. To allow flexibility in story-length and encourage story completion, we set the number of possible output tokens to the max permitted. A subset of outputs are hand-evaluated to ensure high-quality classifications.

5 Analysis of Character Portrayal

To understand characters using CASPER, we consider 6 research questions (RQs) that compare (1) LLM vs. human-written stories and (2) LLM vs. other LLM-generated stories.

(RQ1): How does the average LLM-generated character compare to the average human-written character? To observe categories of an "average" story, we determine which categories are most represented within a set of stories. We classify

[7]Model names/details are given in Appendix C, Table 7.

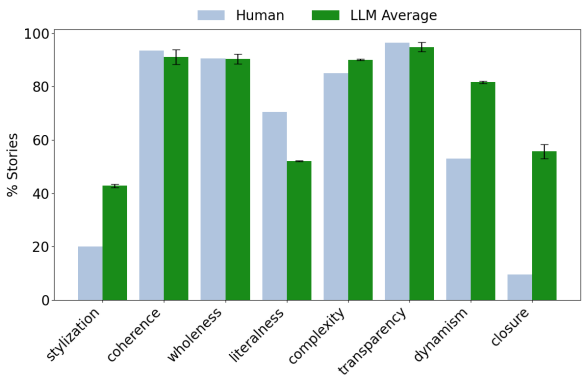

Figure 3: **RQ2:** We compare the percentage of human-written stories (blue) with the percentage of all LLM-generated stories (green) w.r.t. categories. For the latter, we average generations from all LLMs and show standard deviation across LLMs with vertical lines.

labels and rewrite $\mathcal{A} = \langle A_{Sty.}, A_{Coh.}, A_{Whol.}, A_{Lit.}, A_{Comp.}, A_{Tran.}, A_{Dyn.}, A_{Clos.} \rangle$, such that each A is represented by majority labels (a label is in the majority if it makes up for $\geq 50\%$ of the set of stories). For human-written stories, we observe $\mathcal{A} = \langle \bar{A}_{Sty.}, A_{Coh.}, A_{Whol.}, A_{Lit.}, A_{Comp.}, A_{Tran.}, A_{Dyn.}, \bar{A}_{Clos.} \rangle$. For LLM-generated stories (averaged across outputs from all models), we observe $\mathcal{A} = \langle \bar{A}_{Sty.}, A_{Coh.}, A_{Whol.}, A_{Lit.}, A_{Comp.}, A_{Tran.}, A_{Dyn.}, A_{Clos.} \rangle$. These vectors differ only for *Closure*. This finding indicates that, though the majority of the human-written stories present characters with unfilled aspects at the end of the story, LLMs are more likely to generate characters who have a definite conclusion. We conclude that human-writers are likely to take more artistic liberties than LLMs by using ambiguity. In this comparison, LLM's instead tend to "play it safe" and tie up loose ends.

(RQ2): How do LLM-generated characters compare to human-written characters in a more fine-grained manner? To better understand the remaining binary values from *RQ1* in a more fine-grained analysis, we compare the percentage of categories in human vs. LLM-generated stories across all models. Figure 3 shows the results. We observe LLM-generated characters are statistically more likely to be *stylized* and *dynamic* than human-written characters. In other words, LLMs prefer to create stereotypical characters who grow over the course of a story. This pattern likely comes from the models' tendency to write characters whose emotions shift noticeably as the

narrative unfolds (e.g., a character who begins the story feeling sad but ends feeling joyful). Such characters who fit archetypes often help convey a moral lesson. This conjecture is supported by the observation that LLM's also have fewer *literal* (more *symbolic*) characters than human-written stories, pointing to the presence of underlying themes and takeaway lessons hidden within the story. Meanwhile, Figure 3 shows both LLMs and humans overwhelmingly prefer *coherent*, *whole*, and *transparent* characters. In other words, characters are more likely to be straightforward and fully described to the reader. This finding is surprising because it indicates a lack in diversity of characters (i.e., there are fewer mysterious or confusing characters); however, it might be a result of focusing on protagonists, who are often the most developed characters.

(RQ3) Does LLM size affect character portrayal? Next, we focus on characters produced by smaller vs. larger LLMs. We ask, do larger LLMs ($\geq 14B$) generate characters with more varied distributions than smaller LLMs ($< 14B$)? We compare *stylization*, *coherence*, *wholeness*, *transparency*, and *closure* which show the most standard deviation in Figure 3. If larger models create more nuanced characters, we would expect distinct distributions across these categories. However, as shown in Figure 4, averages of smaller and larger LLMs are nearly identical. The only minor difference is that smaller LLMs generate 10.2% more *stylized* characters. Our findings are surprising because they indicate that, in general, though larger LLMs are better at text generation, they do not necessarily produce different types of characters than smaller LLMs. Overall, model size does not appear to affect the types of characters produced.

(RQ4): How does the diversity of categories vary among stories generated by different models? We consider the distribution of aspects of character with respect to each family tested. Figure 5 shows the average distribution of category-pairs averaged for all LLMs belonging to a family. We observe average distributions follow general trends within each category. In other words, all families have *whole* characters for $> 80\%$ stories, *literal* characters for $< 64\%$ stories, etc. However, within these trends, we see the greatest amount of variations between families for *stylization* and *coherence*, *literalness*, and *closure*. Compared to

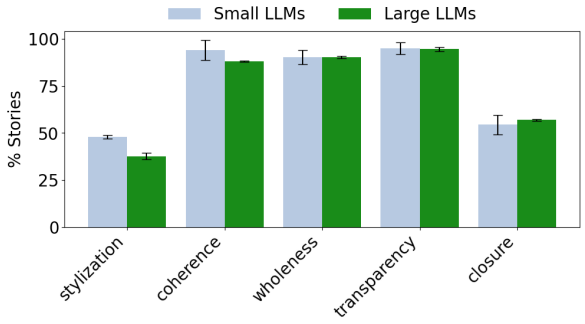

Figure 4: **RQ3**: To determine if primary-categories of character have trends according to model size only, we compare the average of small-medium LLMs (grey) and large LLMs (green) w.r.t. primary categories of character. Standard deviations shown as vertical bars.

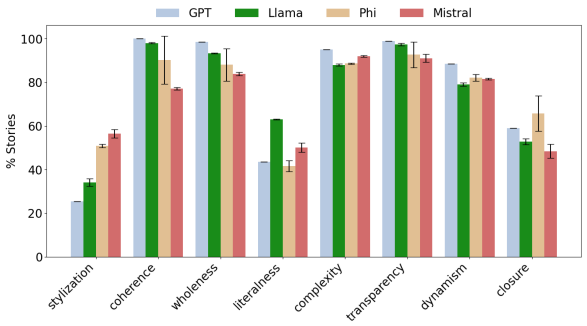

Figure 5: **RQ4:** Averaged categories of characters across families. Standard deviations across different-sized models within a family are shown as vertical bars.

the other families, Mistral prefers more *stylized* and less *coherent*, *closed* characters, indicating an interesting mixture of exaggerated characters who might raise questions from readers. We notice only Phi shows significant standard deviation ($> 5\%$) across models. Ignoring GPT (where only one model is evaluated), Llama shows the least amount of standard deviation. From these findings, we surmise that Phi generates the most diverse characters and Llama generates the least diverse characters.

(RQ5) Do LLMs default to particular categories for different genres? We analyze genres to see if they affect the type of character produced. Figure 6 shows the fraction of times a character fits a category, given the genre specified during story-generation. Firstly, Domestic stories hardly have *stylized* characters, which is suitable for life-like characters; however, these characters are *literal* in only 52% of stories. This is surprising because, even though the stories are focused

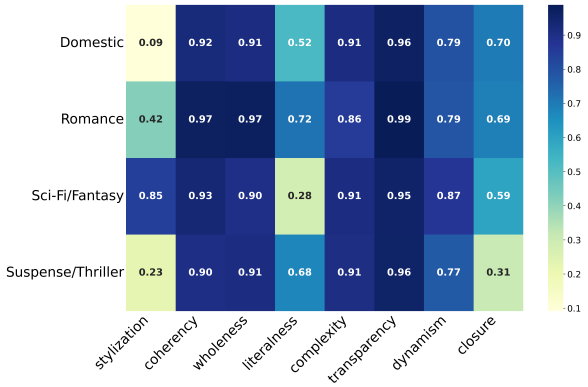

Figure 6: **RQ5:** Distribution of categories, given each specified genre across all LLM-generated stories.

on everyday life, these characters still represent abstract themes without relying on archetypes. Romance stories have nearly completely *coherent*, *whole*, *transparent* characters, which indicates the characters are straight-forward and not confused with emotions, as we might expect. Sci-Fi/Fantasy stories have the highest percentage of *stylized* characters and symbolic characters, which makes sense, because this genre of stories is known to exaggerate conditions or be set in futuristic settings. Only 59% of these characters are *closed*, indicating that at the end of almost half these stories, there are unresolved questions about the character, meaning that these stories might encourage more reader's curiosity. Finally, Suspense/Thriller stories tend to have more *natural* characters, possibly introducing the reader to a convincing character, thereby increasing tension to suspenseful situations. Almost all of these characters are *coherent* and *transparent*, which is surprising because suspense and thrill are often built upon mystery and subterfuge. However, in thrillers, if suspense does not come from the characters, it must come from other aspects of the story, such as setting or plot. Overall, we observe genre affects types of characters generated by LLMs. A corresponding heatmap for human-written stories is given in Appendix H.1.

(RQ6) If we provide the same prompt multiple times to the same model, do categories vary meaningfully among generations? We consider how categories vary across multiple story generations using the same writing-prompt for *stylization*, *literalness*, and *closure* (graphs and details for all categories are given in Appendix H.2). We observe for each fore-mentioned category respectively 30%, 13.5% and 3% of the writing-prompts yield characters with the same label across all inference calls. Most noticeably – compared to *stylization* which has a large number of "stable" characters that do not change across inference calls – *literalness* and *closure* demonstrate the most amount of variability compared to the rest of the categories. This is surprising because it shows LLMs pick up fewer context clues from the writing-prompts that dictate how *literal* and *closed* a character should be.

6 Takeaways

From Section 5, we observe a few key takeaways:

- Unlike human-written stories, LLMs-generated stories are more likely to "play it safe" and have characters with completed story-lines.
- LLM-generated characters are more likely to show character-growth during the course of the story compared to human-written characters.
- Larger models do not necessarily generate different types of characters than smaller models.
- The Phi family appears to generate the most diverse characters, and the Llama family appears to generate the least diverse characters.
- Genre affects types of characters generated by LLMs (e.g., Domestic stories tend to have realistic characters that represent themes, and Romance stories have straightforward characters).
- When re-generating stories from one prompt, *literalness* and *closure* have most variability.

7 Conclusion

We propose CASPER for classifying character portrayal using 8 category-pairs to compare characters in LLM and human-written stories. We create a corpus of 4400 LLM-generated stories across 7 models from 4 model families. This corpus was carefully curated so that each story can be compared to a parallel human-written story with a common writing-prompt. We answer research questions that compare LLM-generated characters with other LLM and human-written characters, and we provide a list of key takeaways to better understand types of characters produced by LLMs.

We note CASPER is designed with definitions from narrative theory for the classification of stories. We do not extend these definitions to a different textual domain because this change introduces complex new problems. In particular, we would have to consider whether or not the new textual domain contains enough aspects of narrative (Piper,

2023) to portray a character so that category-pairs such as *stylization* and *wholeness* can be analyzed. Furthermore, some categories might exhibit subtle differences in a non-creative story setting (e.g., a *stylized* character might be different in a report than in a story). We propose the exploration of these ideas for future work.

Limitations

Since it is difficult to find open-source human-written stories, we scrape stories from subreddits. However, it is possible that a small number of these "human-written" stories might be enhanced or written by AI. We attempt to mitigate such cases by using software for detecting machine-generated text, including `Binoculars` (Hans et al., 2024) and Copyleaks AI Detector.[8] All stories are written in English and invite future work to analyze character aspects of generated stories in other languages.

Since we utilize writing-prompts from the creative writing domain to generate short stories, we note our research is limited to the domain of such writing-prompts and stories.

Also, though we have done due diligence to improve classification performance as much as possible, we note that automatic classification remains worse than human annotators. Since the final corpus is large, we expect these errors to be smoothed out for the analysis. However, the classifier will remain imperfect.

Acknowledgments

We are grateful to the anonymous reviewers. This work was supported in part by NSF grant DRL-2112635 and the Air Force Office of Scientific Research grant FA9550- 22-1-0099.

References

Iqra Ali, Jesse Atuhurra, Hidetaka Kamigaito, and Taro Watanabe. 2025. Hlu: Human vs llm generated text detection dataset for urdu at multiple granularities. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3495–3510.

Mieke Bal. 1997. *Narratology: Introduction to the Theory of Narrative*, 2nd edition. University of Toronto Press, Toronto.

David Bamman, Brendan O'Connor, and Noah A Smith. 2013. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 352–361.

David Bamman, Sejal Popat, and Sheng Shen. 2019. An annotated dataset of literary entities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2138–2144.

Wayne C Booth. 1983. *The rhetoric of fiction*. University of Chicago Press.

Amal Boutadjine, Fouzi Harrag, and Khaled Shaalan. 2025. Human vs. machine: A comparative study on the detection of ai-generated content. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 24(2):1–26.

Faeze Brahman and Snigdha Chaturvedi. 2020. Modeling protagonist emotions for emotion-aware storytelling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5277–5294.

Faeze Brahman, Meng Huang, Oyvind Tafjord, Chao Zhao, Mrinmaya Sachan, and Snigdha Chaturvedi. 2021. "let your characters tell their story": A dataset for character-centric narrative understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1734–1752, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joseph Campbell. 2008. *The hero with a thousand faces*, volume 17. New World Library.

Candice C Carter and Linda Pickett. 2014. Characterization. In *Youth Literature for Peace Education*, pages 25–42. Springer.

Tuhin Chakrabarty, Philippe Laban, and Chien-Sheng Wu. 2024a. Can ai writing be salvaged? mitigating idiosyncrasies and improving human-ai alignment in the writing process through edits. *arXiv preprint arXiv:2409.14509*.

Tuhin Chakrabarty, Vishakh Padmakumar, Faeze Brahman, and Smaranda Muresan. 2024b. Creativity support in the age of large language models: An empirical study involving professional writers. In *Proceedings of the 16th Conference on Creativity & Cognition*, pages 132–155.

Seymour Benjamin Chatman and Seymour Chatman. 1978. *Story and discourse: Narrative structure in fiction and film*. Cornell university press.

Snigdha Chaturvedi, Dan Goldwasser, and Hal Daume III. 2016. Ask, and shall you receive? understanding desire fulfillment in natural language text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

[8] https://copyleaks.com/ai-content-detector

Snigdha Chaturvedi, Mohit Iyyer, and Hal Daume III. 2017. Unsupervised learning of evolving relationships between literary characters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Alicia Tsui Ying Chong, Hui Na Chua, Muhammed Basheer Jasser, and Richard T.K. Wong. 2023. Bot or human? detection of deepfake text with semantic, emoji, sentiment and linguistic features. In *2023 IEEE 13th International Conference on System Engineering and Technology (ICSET)*, pages 205–210.

Andrea Cigliano, Francesca Fallucchi, Marco Gerardi, et al. 2024. The impact of digital analysis and large language models in digital humanity. In *ICYRIME 2024: 9th International Conference of Yearly Reports on Infor-matics, Mathematics, and Engineering*, page 1. CEUR Workshop Proceedings.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, et al. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 1383–1392.

Alperen Elek, Hatice Sude Yildiz, Benan Akca, Nisa Cem Oren, and Batuhan Gundogdu. 2025. Evaluating the efficacy of perplexity scores in distinguishing ai-generated and human-written abstracts. *Academic Radiology*.

David Fishelov. 1990. Types of character, characteristics of types. *Style*, pages 422–439.

Katrina Fong, Justin B Mullin, and Raymond A Mar. 2013. What you read matters: The role of fiction genre in predicting interpersonal sensitivity. *Psychology of aesthetics, creativity, and the arts*, 7(4):370.

Edward Morgan Forster. 1927. *Aspects of the Novel*. Harcourt, Brace.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. 2019. Gltr: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116.

Gauri Anil Godghase, Rishit Agrawal, Tanush Obili, and Mark Stamp. 2025. Distinguishing chatbot from human. In *Machine Learning, Deep Learning and AI for Cybersecurity*, pages 529–564. Springer.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting llms with binoculars: Zero-shot detection of machine-generated text. *Preprint*, arXiv:2401.12070.

Fouzi Harrag, Maria Dabbah, Kareem Darwish, and Ahmed Abdelali. 2020. Bert transformer model for detecting arabic gpt2 auto-generated tweets. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 207–214.

Baruch Hochman. 1985. *Character in Literature*. Cornell University Press, Ithaca, NY.

Tenghao Huang, Faeze Brahman, Vered Shwartz, and Snigdha Chaturvedi. 2021. Uncovering implicit gender bias in narratives through commonsense inference. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3866–3873.

Daphne Ippolito, Ann Yuan, Andy Coenen, and Sehmon Burnam. 2022. Creative writing with an ai-powered writing assistant: Perspectives from professional writers. *arXiv preprint arXiv:2211.05030*.

Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. 2016. Feuding families and former friends: Unsupervised learning for dynamic fictional relationships. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1534–1544.

Brandon Jaipersaud, Zining Zhu, Frank Rudzicz, and Elliot Creager. 2024. Show, don't tell: Uncovering implicit character portrayal using llms. *arXiv preprint arXiv:2412.04576*.

Woori Jang and Seohyon Jung. 2024. Evaluating llm performance in character analysis: A study of artificial beings in recent korean science fiction. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 339–351.

Richard Janko et al. 1987. *Aristotle: Poetics*. Hackett Publishing.

Fotis Jannidis. 2019. Character. In Peter Hühn et al., editors, *the living handbook of narratology*. Hamburg University, Hamburg. Viewed 12 Feb 2019.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. Personallm: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627.

C. G. Jung. 1968. *The Archetypes and the Collective Unconscious*, 2nd edition edition. Routledge, London. First published 1968, eBook published 17 December 2014.

Evgeny Kim and Roman Klinger. 2019. Frowning frodo, wincing leia, and a seriously great friendship: Learning to classify emotional relationships of fictional characters. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 647–653.

JR Landis and GG Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159174.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. 2025. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791.

Dawei Li, Hengyuan Zhang, Yanran Li, and Shiping Yang. 2023. Multi-level contrastive learning for script-based character understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5995–6013, Singapore. Association for Computational Linguistics.

Li Lucy and David Bamman. 2021. Gender and representation bias in gpt-3 generated stories. In *Proceedings of the third workshop on narrative understanding*, pages 48–55.

Piotr Mirowski, Juliette Love, Kory Mathewson, and Shakir Mohamed. 2024. A robot walks into a bar: Can language models serve as creativity supporttools for comedy? an evaluation of llms' humour alignment with comedians. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1622–1636.

Piotr Mirowski, Kory W Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–34.

Ayat Najjar, Huthaifa I Ashqar, Omar Darwish, and Eman Hammad. 2025. Leveraging explainable ai for llm text attribution: Differentiating human-written and multiple llms-generated text. *arXiv preprint arXiv:2501.03212*.

Ramya Namuduri, Yating Wu, Anshun Asher Zheng, Manya Wadhwa, Greg Durrett, and Junyi Jessy Li. 2025. Qudsim: Quantifying discourse similarities in llm-generated text. *arXiv preprint arXiv:2504.09373*.

Armand Nicolicioiu, Eugenia Iofinova, Andrej Jovanovic, Eldar Kurtic, Mahdi Nikdan, Andrei Panferov, Ilia Markov, Nir Shavit, and Dan Alistarh. 2024. Panza: Design and analysis of a fully-local personalized text writing assistant. *arXiv preprint arXiv:2407.10994*.

Carol Pearson. 1991. Awakening the heroes within: Twelve archetypes to help us find ourselves and transform our world. *(No Title)*.

James Phelan. 1989. *Reading People, Reading Plots: Character, Progression, and the Interpretation of Narrative*. University of Chicago Press, Chicago.

James Phelan. 2005. *Living to tell about it: A rhetoric and ethics of character narration*. Cornell University Press.

Andrew Piper. 2023. Computational narrative understanding: A big picture analysis. In *Proceedings of the Big Picture Workshop*, pages 28–39.

Andrew Piper, Richard Jean So, and David Bamman. 2021. Narrative theory for computational narrative understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 298–311.

Vladimir Propp. 1968. *Morphology of the Folktale*, 2nd edition. University of Texas Press, Austin. Translated by Laurence Scott, edited by Louis A. Wagner, with an introduction by Alan Dundes.

Elahe Rahimtoroghi, Jiaqi Wu, Ruimin Wang, Pranav Anand, and Marilyn Walker. 2017. Modelling protagonist goals and desires in first-person narrative. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 360–369.

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

Shlomith Rimmon-Kenan. 2003. *Narrative fiction: Contemporary poetics*. Routledge.

J. K. Rowling. 1997–2007. *Harry Potter Series*. Bloomsbury, London. 7 volumes published between 1997 and 2007.

Jenna Russell, Marzena Karpinska, and Mohit Iyyer. 2025. People who frequently use chatgpt for writing tasks are accurate and robust detectors of ai-generated text. *arXiv preprint arXiv:2501.15654*.

Zhiyao Shu, Xiangguo Sun, and Hong Cheng. 2024. When llm meets hypergraph: A sociological analysis on personality via online social networks. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2087–2096.

Murray Smith. 2022. *Engaging Characters: Fiction, Emotion, and the Cinema*. Oxford University Press, Oxford, UK.

Sandeep Soni, Amanpreet Sihra, Elizabeth F Evans, Matthew Wilkens, and David Bamman. 2023. Grounding characters and places in narrative text. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Shashank Srivastava, Snigdha Chaturvedi, and Tom Mitchell. 2016. Inferring interpersonal relations in narrative summaries. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

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.

Adaku Uchendu, Saranya Venkatraman, Thai Le, and Dongwon Lee. 2024. Catch me if you gpt: Tutorial on deepfake texts. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*, pages 1–7.

Saranya Venkatraman, Nafis Irtiza Tripto, and Dongwon Lee. 2024. Collabstory: Multi-llm collaborative story generation and authorship analysis. *arXiv preprint arXiv:2406.12665*.

Anvesh Rao Vijjini, Faeze Brahman, and Snigdha Chaturvedi. 2022. Towards inter-character relationship-driven story generation. In *EMNLP*.

Azmine Toushik Wasi, Raima Islam, and Mst Rafia Islam. 2024. Ink and individuality: Crafting a personalised narrative in the age of llms. In *Proceedings of the Third Workshop on Intelligent and Interactive Writing Assistants*, pages 43–47.

T Wolf. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Mo Yu, Jiangnan Li, Shunyu Yao, Wenjie Pang, Xiaochen Zhou, Zhou Xiao, Fandong Meng, and Jie Zhou. 2023. Personality understanding of fictional characters during book reading. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14784–14802, Toronto, Canada. Association for Computational Linguistics.

Mo Yu, Qiujing Wang, Shunchi Zhang, Yisi Sang, Kangsheng Pu, Zekai Wei, Han Wang, Liyan Xu, Jing Li, Yue Yu, and Jie Zhou. 2024. Few-shot character understanding in movies as an assessment to meta-learning of theory-of-mind. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

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.

Xinfeng Yuan, Siyu Yuan, Yuhan Cui, Tianhe Lin, Xintao Wang, Rui Xu, Jiangjie Chen, and Deqing Yang. 2024. Evaluating character understanding of large language models via character profiling from fictional works. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8015–8036.

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.

A Examples of Categories

See Section A.1 for popular fiction examples of characters who fit CASPER categories (Table 1). See Section A.2 for examples of stories from our corpus with protagonists fitting each category.

A.1 Examples from Popular Fiction

In Section 3.1, we give examples of *stylized*, *natural*, *coherent*, *incoherent*, *whole*, and *fragmented* characters from Harry Potter. Here, we continue the discussion and provide examples for the remaining category pairs:

Bellatrix Lestrange is *incoherent* because her emotions swing unexpectedly between multiple forms of comic and tragic madness. Harry Potter is an example of a *whole* character, whose entire background, motivations, ambitions, thoughts, and explanations for his actions are given to the reader. Sirius Black is *fragmented* because his youth and years in Azkaban are not fully explained, though they influence his personality and decision making. Arthur Weasley is a *literal* character with unique quirks and does not serve any known allegorical function beyond living daily life as a father. Dobby embodies the *symbolic* character as an uncorrupted figure of innocence as he fights the oppression of servitude in search of freedom. Professor Snape is one of the most *complex* characters in the series because he balances extreme jealousy and love, thus sometimes showing cruelty though he is later proven to be caring. Rubeus Hagrid is a *simple* character who only acts on loyalty and warm-heartedness towards his friends. Luna Lovegood is *transparent* because her speech is honest without masking her inner thoughts. Albus Dumbledore is *opaque* because he withholds information, and his motivations are layered and mostly hidden. Neville Longbottom is *dynamic* because he starts as an awkward and anxious boy, and then he grows in courage to lead Dumbledore's Army and destroy the last Horcrux. Lord Voldemort is *static* because his worldview remains constantly obsessed with

power and death. Fred Weasley is *closed*, not only because his death brings finality to his storyline, but also because there are no major unanswered questions about his role and life from the reader. Finally, Draco Malfoy is *open* because he changes from childhood villain to disillusioned man with new but unclear morals and an undefined future.

A.2 Examples from CASPER

Below, we provide example stories for categories of character used in CASPER:

Stylization: "You... heroes..." Synthia hissed, blue flames rising around her, spiraling up her frame, "you have it so easy... always loved, always admired... given the powers of a god..."

Confused and cringing, Mara replied, "That's not true."

As the witch exploded her magic at Mara, she screeched, "Yes it is!" Quickly, Mara flew upward, using her axe as a shield against the flames, and dodging oncoming winds. Though, no matter how much she moved, Synthia refused to give up, as she detested Mara's denial.

Flipping through her spellbook, Lux remarked with an awkward laugh, "That's so weird to say."

While combatively strumming his guitar, Robbie stopped his humming to agree, "Right? She doesn't know our lives, we were just sent to fight her." When he got distracted, his smog slightly shrunk. A spirit's silhouette rose from it, then slapped his shoulder, making him flinch and complain, "Alright, sorry, damn." Refocusing, Robbie plucked faster, and the previously-annoyed spirit went back to dancing and clacking their castanets.

After ejecting a boiling ball of plasma at Synthia, Onycia called, "Maybe you think we have it easy because you hate everybody. And, with that, everybody hates you."

"I hate everybody because everybody has WRONGED ME!" Wind waved over Synthia, knocking Onycia back against a tree. With a snarl, Synthia sent flames her way, but Carlo ran over with his shield. Kneeling in front of Onycia, he casted lightning bolts from his fingertips, which jetted around until they struck Synthia, making her scream and fall to her knees. Though, she recovered surprisingly quickly, rising back to her feet.

Sliding up to Onycia, Tjinfalk muttered, "People have wronged me, but I don't hate everyone."

She replied, "I know, love, she's just strange."

While Carlo stayed in front of them, Tjinfalk aimed his heat gun at Synthia, balancing the barrel over the shield. Since she was close by, he kept the Distance stat low, but he raised Heat and TIme to max, as she's proven to be hard to hit, and even harder to knock out.

Nearby, Synthia was shooting flaming daggers at Lucina, but she was effortlessly and elegantly striking them away with her sword, while saying, "You really think you have it harder than everybody here?"

Sending multiple daggers at once, Synthia growled, "I KNOW I do!"

Lucina grabbed a tree branch above her, pulling herself up to avoid the attack, then jumped back down to nick Synthia across the shoulder. As Synthia gripped the slice, Lucina explained, "Well, Robbie's dead, I'm homeless, and Tjin hasn't been to his home planet in six years. We're not all miserable bums, but that simply scratches the surface. We do not have it easy."

Catching her breath, the witch scowled, questioning, "You're homele-" before she could finish, a wide beam of white light obliterated her head, then slowly traced down her body, burning her to a crisp. After about a minute, the beam stopped, and Synthia was nothing more than a pile of ash.

Carlo settled his shield and stood up, followed by Tjinfalk and Onycia.

"I still think you need a better name for that weapon," Robbie told Tjinfalk, "and I still think it should be Hell's Sun."

Mara chimed in, "And I still agree!"

Wholeness: It took a lot of convincing to get me to open the door. A lot of cajoling, promises of safety, and patient words of wisdom. Of course, none of that would break me- no, that would take a far more powerful source.

My cellphone began to buzz. My mother was calling. The idea of NOT picking up briefly crossed my mind,but my fingers were wiser than my brain. The accept button was hit and the phone was pressed to my ear before I could even heave a full sigh.

"Mom?...Let them in?!How do you kno- you told them it was okay?! Why didn't you tell me??...I wouldn't have run-probably....but why are they here?...yeah I would fucking hope I am not in trouble.....I- yes... but...I...ok...I am sorry for cursing. I will use better vocabulary moving forward but thi-... yes...yes..fine...I hope

you are right. I trust you. Okay.... I love you too."

I put the phone back into my pocket and after a long moment and a few deep breaths, I opened the door- body tensed and expecting the worse.

And to be sure. The next few hours were no picnic. But the conversation didn't end up going the way I thought. How could I have guessed demon hunters wanted to recruit a half-demon to their ranks?

I clutched the card the lead hunter had given me, still in a daze. A job offer? With full benefits? It was a fantasy even I, a half demon, couldn't comprehend.

I was still in my shocked state when my mom came home, clutching bags from my favorite restaurant. She grinned when she saw me.

"I see you had quite the interesting conversation with my old friends when I was gone. Come help me set the table. I got you pie."

I was on my feet and grabbing plates before she even asked, but froze when I heard her words.

"Friends?" Is that had she known? Did she send them? Did she tell them about me? Why was she friends with demon hunters when she had a half demon child? Did she know them before or after meeting dad? Did-

Before my mind could race more my mom cleared her throat in a conspicuous manner and nodded her head to the plates in my hand. I jumped and scrambled to my task and she began pulling food out of the bag.

"Friends yes- Jonah and Bess are two are the best people I knew back in the day, though I guess you could called them "Old colleagues." instead..."

She set the pie in the center of the table and sat down. I was about to sit as well when her last words registered and my body opted to collapse into the chair instead.

"Old...colleagues? But they- you- I... wait."

My mom, bless her heart, began to chuckle. "Come on darling and eat your dinner. I suppose it's time I told you the truth about how I met your father...."

Literalness: The year was 2075 when the first known case of reverse time travel was detected—a historical figure illegally entering the present to rectify perceived misconceptions about their lives. The culprit? None other than Socrates, the Greek philosopher whose legacy of wisdom seemed irrevocably tied to his penchant for annoyance.

The sun was barely peeking over the horizon as Lydia, an entry-level archivist at the Temporal Regulatory Agency, sipped her coffee and scanned through yesterday's reports. A flicker of light caught her attention on the screen: a spike in unauthorized temporal energy in Athens, Greece. She leaned in closer, frowning.

"Not again," she muttered. Just last week, they'd dealt with Cleopatra insisting on an audience at the Metropolitan Museum of History to discuss inaccuracies in her portrayal. And now Socrates? Of all the historical figures to have spiraled through the use of illegal time travel technology, why did it have to be him?

"Hey! Lydia!" Her colleague Brad's voice echoed through the office, pulling her from her absorption with the report.

"What?" she replied, not bothering to look at him.

"You wouldn't believe what just landed in our inbox." He approached, pointing at the screen.

She raised an eyebrow and turned to face him. "Is it more deranged requests for interviews from ancient leaders? Because I'm done with that. Last time, I almost brought Joan of Arc an answer key for her exam on medieval warfare."

Brad chuckled, leaning over to make sure the screen displayed the right document. "It's actually worse. Socrates just sent us a manifesto. He wants to debate the philosophy of 'falsifying history' with historians, and the worst part? He's demanding a public forum."

Lydia sighed, rubbing her temples. "Great. Just what we need—a debate with a dead philosopher who thinks he's a living Wikipedia. I mean, how can he even argue? He doesn't have the benefit of knowing how much has changed since his time."

But Brad was already animated, his excitement bubbling over. "Think about it! Socrates! The Socratic Method applied to modern problems! It could lead to some killer insights."

"Or it could lead to hours of semantic drudgery with him trying to philosophically dissect why we called him 'annoying' in the twenty-fourth century."

Nevertheless, the very next day, Lydia found herself at the public forum. The hall was filled with history enthusiasts and some rather bemused scholars. Lydia couldn't shake the feeling that they had opened Pandora's box—or in his case, gave him access to a digital tablet.

When Socrates finally appeared, he looked re-

markably unchanged from the statues that adorned classical history texts—bald and bearded, with twinkling eyes that seemed to question everything. The man was completely unfazed by the centuries of progress that had passed since he ingested Hemlock. Instead, his time-traveling escapade had granted him the fervor of a social media influencer and the gravitas of a philosopher, all mashed together.

"Greetings, citizens!" he proclaimed, arms wide. "Let us engage in dialogue about the calamity that is your perception of my trials!"

Lydia rolled her eyes. There was already a debate brewing in the back of the room, and with every question taken, he spiraled deeper into intellectual gymnastics.

"Your records proclaim I was coerced into drinking poison, yet that is but a singular interpretation!" Socrates exclaimed dramatically, raising a finger. "I suggest that, instead, I was willing to embrace true knowledge rather than live a life without virtue."

The audience was now split; half riled up, eagerly adding to the discussion, while the other half seemed to be channeling their inner "please just leave" attitude through furtive glances at the exits. Lydia had reached her limit.

"Okay, wait! Let's call a time-out!" she interrupted, forcing her voice to be heard amidst the rising cacophony of pleasure and discontent. "Socrates, what's the purpose of this? Is it to ensure your historical image is polished, or are you genuinely seeking to educate?"

He turned his deep-set eyes onto her, seeming genuinely puzzled. "Is there a difference?"

"There's a massive difference!" She exhaled sharply, staring him down. "You're here now; you've seen how people live and learn. Isn't it about understanding that we can learn from history without erasing it?"

Silence enveloped the room, where seconds stretched into eternity before Socrates nodded slowly, visibly contemplating her words. "Your reasoning holds merit, dear Lydia. Perhaps I've aimed too high—and here is the conclusion: to insist on an iron grip of the past may obstruct the flow of wisdom in the present."

With that, the philosopher accepted a genuine dialogue rather than a correction, and the rest of the evening unfolded into a surprisingly open and generous exchange. Lydia realized that beneath Socrates' seemingly obnoxious airs lay a passion for exploration.

By the time the sun dipped low, the forum had transformed. It was no longer a battle of ideologies but a sharing of thoughts—a retroactive harmony that was delightful in unexpected ways.

Lydia leaned back, a smile creeping onto her face, acknowledging that while Socrates could be a jerk, sometimes all it took was a little patience and the willingness to engage honestly. She could only hope that the next reverse traveler would be as accommodating.

Complexity: In the dimly lit chambers of the Infernal Court, echoes of souls in torment reverberated through the stone walls. Just beyond the door, a sign read "Court of Appeals," scrawled in jagged letters dripping crimson ink, exuding an aura of both dread and hope. Souls were gathered in a loose formation, whispering among themselves, their faces gaunt and hollow, figurative shadows of the once vibrant beings they had been.

At the center of the room sat a massive obsidian table where three judges, cloaked in robes as black as the void, presided. Their faces were obscured by hoods, but their glowing eyes pierced through, exuding an otherworldly judgment that sent shivers of anticipation down the spines of the petitioners.

"You may approach," intoned the chief judge, his voice like rustling leaves in a sultry wind. The first soul stepped forward, trembling. The whispers subsided as he began to plead his case. He was met with indifference, and almost snickered at, as he described his life filled with petty crimes and trivial grievances. In no time, the gavel had slammed—denial.

Next came a woman, rage burning in her heart. She spoke of injustice and betrayal, of a life spent in service to others only to have her heart ripped apart by a lover's treachery. The judges leaned back, their faces cloaked in shadows, and without emotion, they rendered their verdict—denial.

As the hours wore on, the spirits came and went like moths to a flame, each fate met with the same cold dismissal, the cycle of despair indelibly woven into the fabric of the court. But hope remained, as every soul knew there was a chance, however ridiculous, to be among the less than one percent.

Then, with a resounding thud, the heavy door creaked open again, and a soul staggered in, clutching a tattered book to its chest. The specter was different from the rest—torn, yet resolute. Her features bore the uncanny familiarity of a once-renowned author whose stories had touched mil-

lions, now rendered unrecognizable.

Alessia, they whispered. Would this be the tale that turned the tide? Would her words, now eager to be freed, touch the judges' hardened hearts?

"Your Honor," she began, voice quaking, "I come not to plead for myself, but for the stories I left behind. Books that lie unfinished, characters whose destinies were stolen from them as much as mine was. I plead for their completion, so that my sins—undoing lives through ignorance—might serve a greater purpose."

The judges shifted slightly, intrigued despite themselves. There was something about her presence, a spark of life that seemed relentless in the face of despair. The chief judge leaned forward, his eyes sharp as daggers.

"And what makes you worthy of our attention?" His tone dripped with skepticism.

Alessia opened her book, pages fluttering in the spectral breeze. "In this volume are the musings of lost souls—a collection titled 'The Orchard of Regrets.' Each tale, a life misplaced or misused, each regret a fruit never harvested. Allow me the time to finish this work, and upon its completion, I will accept whatever fate awaits me."

The judges exchanged glances, a flickering light of curiosity breaking the shadow of their eternal gloom. They were no strangers to ambition yet had grown weary of hollow aspirations. What Alessia proposed was unusual; souls were not known for their selflessness.

"Finish this... book," the chief judge mused, almost to himself. "But heed this: if you do not succeed, you relinquish your right to another appeal. This volume must evoke emotion, challenge, transformation. Choose your characters wisely."

She nodded fervently, determination burning in her chest. "I will not falter."

Days turned into weeks, and Alessia wrote furiously in that desolate chamber, exposing layers of her soul with every word. The characters began to breathe, regretting their unmade choices, mourning paths not taken or lessons unlearned. Souls found solace in her strokes of pen—their stories illuminated in vivid detail, their sorrows transmuted into heartache and beauty.

And as the final draft began to manifest, a peculiar transformation occurred. The despairing torments dulled, the cries of anguish outside the court softened, replaced by the whispers of hope. The three judges observed as the stories twinkled with the essence of all that had been lost and all that could still be reclaimed.

Finally, Alessia approached the obsidian table with the completed manuscript. "Your Honor," she declared resolutely, "the words are alive. I've laid bare my heart and the hearts of those I wronged. I've sown understanding where once there was neglect. I ask once more for the chance to redeem myself."

Time hung heavy in the air, as if the court itself had drawn breath. The judges leaned closer under the dim light, grasping the weight of a legacy that had been reborn.

"I will permit one soul to be released," the chief judge uttered at last, "but only at your command."

With a heart fuller than it had ever been, Alessia opened her arms wide. "Let it be for those who wish to reclaim their hopes. Let them run free. And may my story become the seed of their new beginnings."

As she spoke the final words, a brilliant light shattered the darkness, and the chamber was filled with ethereal laughter. The faces of the souls transformed, each one filled with newfound clarity and purpose as they stepped forward, soaring past the gates of despair, into a world that awaited their return.

And from that moment on, the stories of redemption began to ripple through the sea of lost souls—not every journey would end in release, but every tale would now carry the possibility of eternal hope. Each soul remembered, each life honored, and for Alessia, there would always be a space in history, waiting for a word to be written. The Court of Appeals would henceforth whisper her name, and in the depths of hell, a glimmer of humanity lingered on the horizon.

99 percent of the time, failure reigned. But that 1 percent, that shining anomaly? It transformed the world.

Transparency: The year was 2075 when the first known case of reverse time travel was detected—a historical figure illegally entering the present to rectify perceived misconceptions about their lives. The culprit? None other than Socrates, the Greek philosopher whose legacy of wisdom seemed irrevocably tied to his penchant for annoyance.

The sun was barely peeking over the horizon as Lydia, an entry-level archivist at the Temporal Regulatory Agency, sipped her coffee and scanned through yesterday's reports. A flicker of light caught her attention on the screen: a spike in unau-

thorized temporal energy in Athens, Greece. She leaned in closer, frowning.

"Not again," she muttered. Just last week, they'd dealt with Cleopatra insisting on an audience at the Metropolitan Museum of History to discuss inaccuracies in her portrayal. And now Socrates? Of all the historical figures to have spiraled through the use of illegal time travel technology, why did it have to be him?

"Hey! Lydia!" Her colleague Brad's voice echoed through the office, pulling her from her absorption with the report.

"What?" she replied, not bothering to look at him.

"You wouldn't believe what just landed in our inbox." He approached, pointing at the screen.

She raised an eyebrow and turned to face him. "Is it more deranged requests for interviews from ancient leaders? Because I'm done with that. Last time, I almost brought Joan of Arc an answer key for her exam on medieval warfare."

Brad chuckled, leaning over to make sure the screen displayed the right document. "It's actually worse. Socrates just sent us a manifesto. He wants to debate the philosophy of 'falsifying history' with historians, and the worst part? He's demanding a public forum."

Lydia sighed, rubbing her temples. "Great. Just what we need—a debate with a dead philosopher who thinks he's a living Wikipedia. I mean, how can he even argue? He doesn't have the benefit of knowing how much has changed since his time."

But Brad was already animated, his excitement bubbling over. "Think about it! Socrates! The Socratic Method applied to modern problems! It could lead to some killer insights."

"Or it could lead to hours of semantic drudgery with him trying to philosophically dissect why we called him 'annoying' in the twenty-fourth century."

Nevertheless, the very next day, Lydia found herself at the public forum. The hall was filled with history enthusiasts and some rather bemused scholars. Lydia couldn't shake the feeling that they had opened Pandora's box—or in his case, gave him access to a digital tablet.

When Socrates finally appeared, he looked remarkably unchanged from the statues that adorned classical history texts—bald and bearded, with twinkling eyes that seemed to question everything. The man was completely unfazed by the centuries of progress that had passed since he ingested Hemlock. Instead, his time-traveling escapade had granted him the fervor of a social media influencer and the gravitas of a philosopher, all mashed together.

"Greetings, citizens!" he proclaimed, arms wide. "Let us engage in dialogue about the calamity that is your perception of my trials!"

Lydia rolled her eyes. There was already a debate brewing in the back of the room, and with every question taken, he spiraled deeper into intellectual gymnastics.

"Your records proclaim I was coerced into drinking poison, yet that is but a singular interpretation!" Socrates exclaimed dramatically, raising a finger. "I suggest that, instead, I was willing to embrace true knowledge rather than live a life without virtue."

The audience was now split; half riled up, eagerly adding to the discussion, while the other half seemed to be channeling their inner "please just leave" attitude through furtive glances at the exits. Lydia had reached her limit.

"Okay, wait! Let's call a time-out!" she interrupted, forcing her voice to be heard amidst the rising cacophony of pleasure and discontent. "Socrates, what's the purpose of this? Is it to ensure your historical image is polished, or are you genuinely seeking to educate?"

He turned his deep-set eyes onto her, seeming genuinely puzzled. "Is there a difference?"

"There's a massive difference!" She exhaled sharply, staring him down. "You're here now; you've seen how people live and learn. Isn't it about understanding that we can learn from history without erasing it?"

Silence enveloped the room, where seconds stretched into eternity before Socrates nodded slowly, visibly contemplating her words. "Your reasoning holds merit, dear Lydia. Perhaps I've aimed too high—and here is the conclusion: to insist on an iron grip of the past may obstruct the flow of wisdom in the present."

With that, the philosopher accepted a genuine dialogue rather than a correction, and the rest of the evening unfolded into a surprisingly open and generous exchange. Lydia realized that beneath Socrates' seemingly obnoxious airs lay a passion for exploration.

By the time the sun dipped low, the forum had transformed. It was no longer a battle of ideologies but a sharing of thoughts—a retroactive harmony that was delightful in unexpected ways.

Lydia leaned back, a smile creeping onto her

face, acknowledging that while Socrates could be a jerk, sometimes all it took was a little patience and the willingness to engage honestly. She could only hope that the next reverse traveler would be as accommodating.

Dynamism: "Until Help arrives, stay hidden or barricaded, do not engage any of the subjects at any cost. They are hostile and highly infectious. We repeat, DO NOT ENGAGE."

The alarm bleared through many electronic signals. It started so soon that no one had true information about this situation. We only knew fear.

"Until Help arrives, stay hidden or barricaded, do not engage any of the subjects at any cost. They are hostile and highly infectious. We repeat, DO NOT ENGAGE."

The alarm repeats, the commentator is tired and it shows... Great distress etched in their face, as if a great burden had been forced upon them. Some forbidden knowledge, or worse...

"Until Help arrives, stay hidden or barricaded, do not engage any of the subjects at any cost. They are hostile and highly infectious. We repeat, DO NOT ENGAGE."

Three times it had repeated... Nothing happened, yet...

We grow weary of them. We resent the authorities forcing us into the confinement of our homes.

"Until Help arrives, stay hidden or barricaded, do not engage any of the subjects at any cost. They are hostile and highly infectious. We repeat, DO NOT ENGAGE."

It has been a day since the start of the signal. I heard shots and screams... Being alone is bad during these trying times. It is worse hearing the mad ramblings of some people. Omens of terrible times coming, a test from the heavens... The return of the old gods. Being confined had eroded their cognitive abilities...

We can't keep this.

"Until Help arrives, stay hidden or barricaded, do not engage any of the subjects at any cost. They are hostile and highly infectious. We repeat, DO NOT ENGAGE."

And now, I hear someone or something hit my barricaded windows. It started strong, but now it has slowed down... It was almost rhythmic... Soothing.

Until it wasn't

"Until Help arrives, stay hidden or barricaded, do not engage any of the subjects at any cost. They are hostile and-"

An injured woman managed to break into my home.

She is bleeding from the gunshots... Her skin presents some growths, like moss or lichen. Science was never my strong school course...

But there is something that attracts me. A smell, or maybe it is the basic human empathy.

Since she broke into my home, I turned off my radio... I grew bored of this doom-saying. And I see her again. Something compels me to hold her... And protect her.

To try and heal her injuries.

This was my doom.

Closure: The world had crumbled under the weight of the undead. Cities that once bustled with life were now husks, echoing only with the shuffling sounds of the infected. When the zombie apocalypse descended like a dark cloud, most people rushed to gun stores, believing firepower would be their salvation. But not Lucas. When the chaos erupted, he found himself drawn to a time long past, a place where metal was forged into legends and honor marred by few scratches.

In a sleepy town at the edge of civilization, Lucas had discovered an old medieval armory tucked away behind a forgotten alley. Its wooden door whispered secrets of battles long settled, and upon breaking in, he breathed the air saturated with history. There, the armory stood proud and untouched: knights' helmets, greaves, and swords elegantly arranged like art begging for a master. Lucas knew in that moment what he had to do.

He donned a full suit of plate armor, the metal cool against his skin as he cinched the leather straps tight. It was cumbersome yet invincible, every piece a testament to the craftsmanship of an era defined by valor. He felt a surge of power—more than just the weight of the armor, but the indomitable spirit of the knight who once wore it.

Clutching a long sword with a cruciform hilt, Lucas set out onto the streets of his neighborhood. The clamor of the armor accompanied him, a reminder that he was a bulwark against the horrors lurking beyond his door. The streets were desolate, houses boarded, and windows shattered, yet he was determined to reclaim his home from the marauding dead.

The first group of zombies he encountered lurked in the shadows of an abandoned car lot—four of them, hands grasping desperately at the ground,

heads lolling. Adrenaline surged through him as he unleashed a harsh battle cry. The crash of metal against pavement startled the creatures, their lifeless eyes snapping to attention.

With precision born of fabled conflicts long ago, Lucas swung his sword. The blade sang through the air, catching the first zombie square in the neck, cleaving through the decayed flesh with just one stroke. The body crumpled to the ground, lifeless once more. The smell of rot mingled with the sweet scent of bravery; he felt alive.

The remaining zombies roared, stumbling forward in a mindless frenzy. Lucas maneuvered effortlessly, the weight of his armor grounding him, preventing the undead from tumbling him over. He sidestepped the nearest attacker, his sword wheeling like a tempest, catching another zombie in the ribs. The momentum of his swing sent another stumbling backward, and quenching the instinct to retreat, he pursued with zeal.

One after another, the creatures fell before him. Reflected sunlight glinted off his polished armor while the thrill of battle coursed through his veins. He felt as though he had been transported to a different age, where his every strike rang out like a proclamation—he was not to be trifled with.

As he continued through the streets, clearing his neighborhood, Lucas discovered remnants of life amidst the desolation. An overturned bicycle, a child's drawing, a dog collar—small reminders of the innocence that had been cast aside. And in these elements, his determination grew. He was not just fighting zombies; he was reclaiming the echoes of laughter and the fleeting moments of joy that used to fill the sidewalks.

Hours later, with the sun setting over the horizon and the air thick with the scent of spent decay, Lucas stood upon what used to be his front lawn. It was littered with bodies, but none rose again. He had cleared the immediate vicinity, a heroic effort that left him breathless and triumphant.

He unsheathed his sword, allowing it to rest against his shoulder as he surveyed the familiar surroundings. Such tranquility now felt surreal; he was not just a fool in armor; he was a guardian, a protector embodying the spirit of every hero that had ever roamed this world with honor.

As he made his way back into the house, Lucas now carried more than just firepower; he possessed hope—a weapon forged in steel. And as long as he stood between the shadows and the light, he would fight to ensure that this was not merely an end, but a beginning waiting to be written anew.

B Evaluating Prompting Methods

B.1 Templates

In order to find best templates for CASPER, we try out numerous formats. We try several variations of zero-shot and ICL learning, where category-pairs are classified one-pair-at-time, all at once in a single prompt, and on a Likert scale (values 0-2 are bucketed into category A, values 3-5 are bucketed into category B). In addition to using definitions for category-pairs (see Table 1), we try replacing the definitions with descriptive synonyms (akin to a thesaurus lookup) as well as concatenating the definitions with the synonyms. We determine that the templates using only the definitions work best.

Below we provide the best prompt templates from our experiments for classifying categories for characters. In CASPER we use **Zero-shot** for *dynamism* and *closure*, **ICL-basic** for *literalness*, **ICL-interleaved** for *coherence*, *wholeness*, *complexity*, and *transparency*, and **ICL-repeated** for *stylization*.

Note: for all ICL templates, shots are chosen randomly from the gold labels/explanations for each inference call. In this way, we avoid unduly biasing the classification towards a single story in each category.

Zero-shot

You are an expert literary analyst. Given a short narrative, honestly evaluate it and determine the character type of the protagonist.

We define the following character types:
`{a_def}` `{b_def}`

Now, given the following narrative, classify the protagonist. Respond in valid JSON only, with two fields: {"explanation": "a short justification under 50 tokens", "solution": "A or B"}

Narrative: `{story}`

ICL-Basic

You are an expert literary analyst. Given a short narrative, honestly evaluate it and determine the character type of the protagonist.

We define the following character types:
`{a_def}` `{b_def}`

Example 1: `{a_story}`
Solution 1: {“explanation”: `{a_explain}`, “solution”: “A”}

Example 2: `{b_story}`
Solution 2: {“explanation”: `{b_explain}`, “solution”: “B”}

Now, given the following narrative, classify the protagonist. Respond in valid JSON only, with two fields: {“explanation”: “a short justification under 50 tokens”, “solution”: “A or B”}

Narrative: `{story}`

ICL-Repeated

You are an expert literary analyst. Given a short narrative, honestly evaluate it and determine the character type of the protagonist.

We define the following character types:
`{a_def}` `{b_def}`

Example 1: `{a_story}`
Solution 1: {"explanation": "`{a_explain}`", "solution": "A"}

Example 2: `{b_story}`
Solution 2: {"explanation": "`{b_explain}`", "solution": "B"}

Reminder: The character types are: `{a_def}` `{b_def}`

Now, given the following narrative, classify the protagonist. Respond in valid JSON only, with two fields: {"explanation": "a short justification under 50 tokens", "solution": "A or B"}

Narrative: `{story}`

ICL-Interleaved

You are an expert literary analyst. Given a short narrative, honestly evaluate it and determine the character type of the protagonist. We define the following character types:

`{a_def}`
Example: `{a_story}`
Solution: {“explanation”: `{a_explain}`, “solution”: “A”}

`{b_def}`
Example: `{b_story}`
Solution: {“explanation”: `{b_explain}`, “solution”: “B”}

Now, given the following narrative, classify the protagonist. Respond in valid JSON only, with two fields: {“explanation”: “a short justification under 50 tokens”, “solution”: “A or B”}

Narrative: `{story}`

Combined

You are an expert literary analyst. Given a short narrative, evaluate the protagonist across eight dimensions of characterization.

Here are the definitions:

1. Stylization vs. Naturalism
`{a_def}` `{b_def}`

2. Coherence vs. Incoherence
`{a_def}` `{b_def}`

3. Wholeness vs. Fragmentariness
`{a_def}` `{b_def}`

4. Literalness vs. Symbolism
`{a_def}` `{b_def}`

5. Complexity vs. Simplicity
`{a_def}` `{b_def}`

6. Transparency vs. Opacity
`{a_def}` `{b_def}`

7. Dynamism vs. Staticism
`{a_def}` `{b_def}`

8. Closure vs. Openness
`{a_def}` `{b_def}`

Classify the protagonist in the following narrative. Respond ONLY in valid JSON with exactly this format. Each category must be either "A" or "B" (not words):

{"category1": "A or B", "category2": "A or B", "category3": "A or B", "category4": "A or B", "category5": "A or B", "category6": "A or B", "category7": "A or B", "category8": "A or B"}
Narrative: {story}

Likert-Scale

You are an expert literary analyst. Given a short narrative, honestly evaluate it and determine the character type of the protagonist.

We define the following character types:
`{a_def}` `{b_def}`

Given the following narrative, rate the protagonist on a scale from 0 to 5, where:
- 0 means the protagonist clearly matches category A.
- 5 means the protagonist clearly matches category B.
- Numbers between 1–4 represent varying degrees between A and B.

DO NOT GIVE ANY EXPLANATION. Respond with ONLY the number (0, 1, 2, 3, 4, or 5).

Narrative: `{story}`

B.2 Test Set Inter-Rater Agreement

The CASPER test set is manually annotated by 4 annotators, 2-3 annotators per story. All annotators are native English speakers and current undergraduate/graduate students in the USA. To train them for the task, they are provided descriptions and definitions of the categories, including those used in our prompt templates and the source paper (Hochman, 1985) where they are originally described. Each story is initially annotated by 2 in-person reviewers. For any disagreement a third annotator chooses the final label. Table 2 provides the breakdown of Cohen's kappa inter-rater agreement and F1-macro scores.

Note: Human annotations are very difficult, time-consuming and expensive. This is because annotators need to be trained to be experts in the relevant part of literary theory and read between the lines to understand character portrayal. For 100 stories, we have a total of 800 gold annotations. While this test set size is small, we attempt to ensure that it is high quality.

B.3 Results from Experiments with Templates

See Table 3 for F1-macro scores obtained during experimentation, described in Section 3.3, to find best prompt templates for CASPER. See Table 4 for a comparison of `Llama-70B` and `GPT4o-mini`.

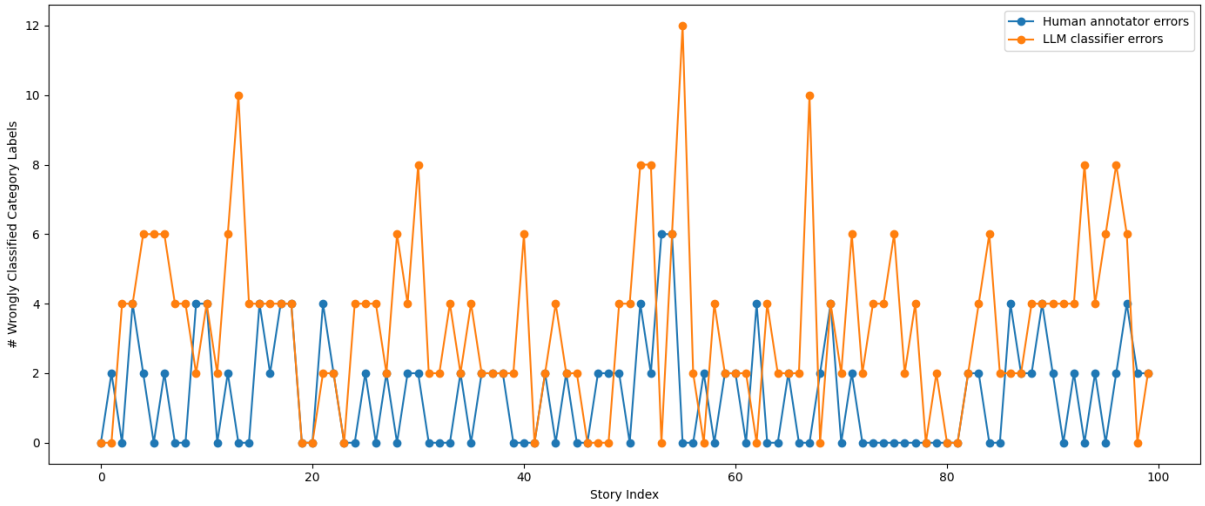

Figure 7: Comparison of wrongly classified category labels for humans (blue) vs. LLMs (orange). The x-axis represents the index of each classified story.

Category	κ	F1-macro
Stylization	63.7	87.66
Coherence	46.1	69.23
Wholeness	52.0	78.44
Literalness	55.2	89.39
Complexity	60.9	83.67
Transparency	63.2	88.84
Dynamism	83.6	87.1
Closure	56.0	92.79

Table 2: Inter-rater agreement, shown with Cohen's kappa and F1-macro score as percentages.

We observe that `Llama-70B` performs comparably to (and sometimes even better than) `GPT4o-mini`, so we use this model to complete our analysis of LLM vs. human-written characters.

We perform the following analyses to better understand why the classifier sometimes makes mistakes, even with the best templates:

1. *Do LLMs prefer certain labels for certain traits?* We count the number of times the LLM predicts a label (A vs. B, where A represents the category, and B represents the opposing category). Table 5 shows the breakdown of the percentages of each label occurrence. LLMs tend to under-classify *stylized* and *closed* characters and over-classify *literal* characters. We determine that these are the trickiest categories to predict.

2. *Do LLMs and Human annotators share similar disagreements rates?* We count the percentage of disagreements between human annotators and LLM predictions vs. gold labels Table 6. These columns have a moderately strong/strong positive correlation with Pearson correlation $(r) \approx 0.695$. The results indicate that characters who are difficult or subjective for humans to classify are also difficult for LLMs to classify.

3. *Do classification errors occur because particular stories/characters are tricky to analyze overall?* Figure 7 shows how much a human annotator differs from the final gold label and the error rate of an LLM classifier across all 8 categories for each story. For a few stories, LLMs make errors, while humans make no errors at all. There are also a few cases of humans making classification errors where LLMs classify correctly. Overall, we do not see overarching trends that suggest certain stories are trickier to interpret than others.

C Implementation Details

Table 7 contains the complete list of LLMs with their respective family names, exact model names from HuggingFace library (Wolf, 2019), model sizes, and knowledge cutoff dates. Experiments are conducted using OpenAI API for the closed-source model and four 48GB Nvidia RTX A6000 GPUs for all open-source models. Generating 200 stories does not exceed 2 hours per LLM. Inference time for each experiment takes approximately 20 minutes for 200 stories.

We use existing Python packages such as HuggingFace, pandas, re, scikit-learn, and statistics.

Category	Zero-shot	ICL "basic"	ICL "dispersed"	ICL "repeated"	Likert	Single-prompt
Stylization	43.18	66.02	67.84	**68.47**	64.40	44.55
Coherence	50.76	**62.99**	59.28	59.28	48.72	48.98
Wholeness	52.81	**57.48**	56.71	57.48	61.02	54.83
Literalness	45.28	49.27	**60.81**	56.00	61.13	65.23
Complexity	46.52	**61.69**	61.69	54.78	46.52	53.15
Transparency	39.15	**60.10**	55.56	45.65	45.95	57.48
Dynamism	**84.37**	76.62	76.19	79.32	56.10	77.63
Closure	**75.12**	65.99	70.86	72.78	67.99	57.93

Table 3: Comparison of templates for each category (F1-macro scores). Templates used in CASPER are in bold. Results are statistically significant using Bootstrap Test (Dror et al., 2018). While comparing three runs, the largest standard deviation is 2.12% (for closure).

Category	Llama-70B	GPT4o-mini
Stylization	68.47	66.67
Coherence	62.99	61.60
Wholeness	57.48	58.88
Literalness	60.81	63.53
Complexity	61.69	59.35
Transparency	60.10	59.35
Dynamism	84.37	78.18
Closure	75.12	39.27

Table 4: Comparison of Llama-70B vs. GPT4o-mini, using CASPER templates (F1-macro scores, in percent).

Categories	Pred A	Pred B	Gold A	Gold B
Stylization	0.50	0.50	0.72	0.28
Coherence	0.96	0.04	0.97	0.03
Wholeness	0.88	0.12	0.78	0.22
Literalness	0.59	0.41	0.34	0.66
Complexity	0.84	0.16	0.87	0.13
Transparency	0.94	0.06	0.85	0.15
Dynamism	0.68	0.32	0.73	0.27
Closure	0.25	0.75	0.41	0.59

Table 5: Percentage of times the LLM predicts label A and B versus the percentage of occurrences of label A and B occurring as gold labels.

We scrape the human-written stories from Reddit using the Reddit API Wrapper, PRAW.

We note that all stories used in CASPER are freely available for research purposes according to the Reddit Public Content Policy.

AI is used for minor assistance in coding.

D Checking "Creativity" in Test Set

In order to generate the test set, we use a prompt template that includes the phrase "You are a creative story writer." Here, we show that the word "creative" does not bias the LLM towards generating more creative (e.g., more dramatic) characters than it would otherwise. From the category-pairs, we focus on *stylization* and *dynamism*. Using a random sampling of 50 writing-prompts, we generate 2 stories per writing-prompt (one where the prompt template includes and one where the prompt template excludes the word "creative"). We use `Llama-8B` for story-generation and `Llama-70B` for CASPER classification. As shown in Figure 8, there is minimal difference in the types of characters generated with respect to *stylization* and *dynamism*. Hence, we conclude that the use of the word "creative" does not bias the LLM during story generation.

Categories	Human	LLM
Stylization	0.11	0.30
Coherence	0.03	0.05
Wholeness	0.13	0.24
Literalness	0.10	0.39
Complexity	0.09	0.19
Transparency	0.06	0.15
Dynamism	0.09	0.13
Closure	0.07	0.22

Table 6: Percentage of disagreements between human annotators and LLM predictions for each category.

E Comparing Human-Written Stories

Since the human-written stories from our testset and the human-written stories from the CASPER corpus are obtained from different sources (two subreddits), we conduct experiments to ensure the textual domains of the two sets should be indistinguishable. For this experiment, we use in-context learning with `Llama-70B` to classify each story as belonging to either the testset or to the CASPER corpus. We use the following prompt:

Family	Short Name	Exact Model Name	Size	Knowledge Cutoff
GPT	GPT4o-mini	gpt-4o-mini-2024-07-18	*Unknown*	09/30/2023
Llama	Llama-3B	meta-llama/Llama-3.2-3B-Instruct	3.21B	12/2023
	Llama-8B	meta-llama/Llama-3.1-8B-Instruct	8B	12/2023
	Llama-70B	meta-llama/Llama-3.3-70B-Instruct	70B	12/2023
Phi	Phi3-4B	microsoft/Phi-3-mini-4k-instruct	3.8B	10/2023
	Phi3-14B	microsoft/Phi-3-medium-4k-instruct	14B	10/2023
Mistral	Mistral-7B	mistralai/Mistral-7B-Instruct-v0.3	7.25B	02/2023
	Mistral-24B	mistralai/Mistral-Small-24B-Instruct-2501	23.6B	*Unknown*

Table 7: Details about all LLMs used in our experiments.

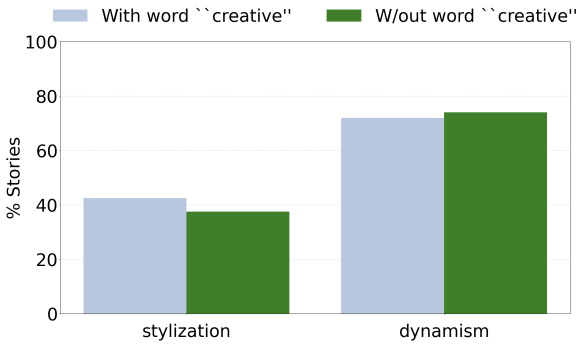

Figure 8: Comparison of types of characters generated when template includes vs. does not include the word "creative." For both *stylization* and *dynamism*, we observe minimal differences. These results indicate that the use of the word "creative" does not bias the LLM towards generating significantly more *stylized* and *dynamic* characters.

Evaluating Domain Comparability

There are 2 datasets of short stories. Given a new short story, determine which dataset it belongs to.
Dataset 1 example: {*shots*}
Dataset 2 examples: {*shots*}
New story: {*story*}

The resulting Macro-F1 score = 24.52%. This score is very low, indicating that the model is unable to distinguish stories from the two sources. Thus, we determine that the stories are from the same textual domain.

We also use available software for detecting machine-generated text, including `Binoculars` (Hans et al., 2024) and Copyleaks AI Detector.[9] In this way, we try to choose stories which are primarily written by humans.

[9]https://copyleaks.com/ai-content-detector

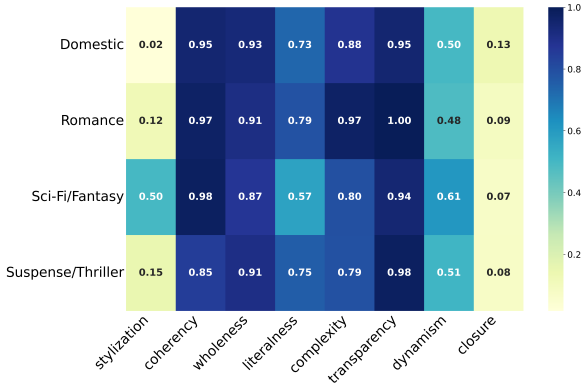

Figure 9: **RQ5:** Distribution of categories, given each specified genre across human-written stories.

F Additional Details about CASPER Corpus

See Table 8 for additional statistics for the CASPER corpus.

G Mapping of Genre Labels

Table 9 shows the original Subreddit tags used to indicate human-labeled story genres. We merge and map these genres to our set of 4 broad genres, Domestic Fiction, Romance, Science-Fiction, and Suspense/Thriller.

H Additional Analysis

H.1 More Details about RQ5

For Section 5 RQ5, in additional to showing the heatmap of the distribution of categories, given each specified genre across all LLM-generated stories, (Figure 6) we provide a corresponding heatmap for all human-written stories in Figure 9. We observe that many of the distributions are similar to the distributions of the LLM-generated stories, indicating that when genres are specified, LLM-generated stories are biased toward generating stories with particular categories of char-

Source Family	Source Name	# Stories	Avg	Min	Max
Human	r/WritingPrompts	50	661.48±302.55	177	1739
	r/shortstories	200	1502.57±1117.83	494	6395
GPT	GPT4o-mini	200	952.57±118.13	528	1503
Llama	Llama-3B	600	686.78±174.67	5	1146
	Llama-8B	600	695.69±122.47	17	1214
	Llama-70B	600	1027.53±401.79	457	1787
Phi	Phi3-4B	600	763.71±306.16	13	1561
	Phi3-14B	600	759.29±175	40	1445
Mistral	Mistral-7B	600	560.83±152.44	241	1462
	Mistral-24B	600	845.06±267.79	358	1687

Table 8: Corpus statistics, including average, minimum, and maximum number of tokens in each story.

Genre (from Subreddit)	**Merge with**
Science Fiction	Science-Fiction/Fantasy
Fantasy	Science-Fiction/Fantasy
Realistic Fiction	Domestic Fiction
Horror	Suspense/Thriller
Misc Fiction	Drop
Speculative Fiction	Science-Fiction/Fantasy
Humour	Drop
Romance	Romance
Non-Fiction	Drop
Mystery & Suspense	Suspense/Thriller
Thriller	Suspense/Thriller
Historical Fiction	Drop
Serial Sunday	Drop
Action & Adventure	Suspense/Thriller
Urban	Domestic Fiction
Micro Monday	Drop
Off Topic	Drop
Meta Post	Drop
Other	Drop

Table 9: Genre mappings from subreddit tags to our genres.

acters in alignment with human-written stories. One noticeable difference, however, is that more LLM-generated characters are *closed* compared to human-written characters, further emphasizing our first key takeaway (Section 6) that LLMs prefer to "play it safe" by ending a story when the character is also completed.

H.2 More Details about RQ6

In Figure 10 we provide graphs showing variability of categories across 3 separate story generations using the same writing-prompt for all categories.

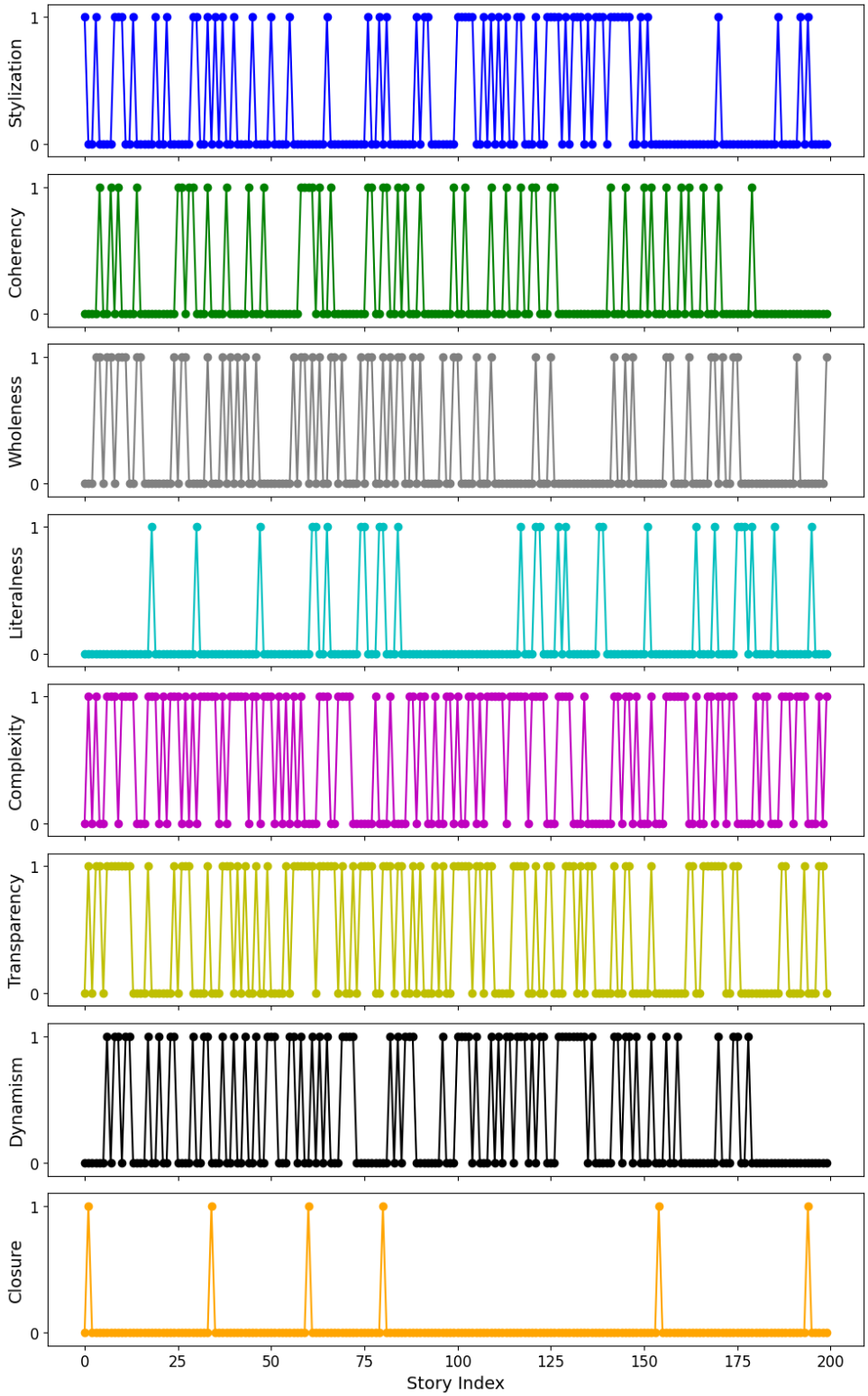

Figure 10: **RQ6:** Variability of categories across 3 identical inference calls (averaged over all open-sourced LLMs). X-axes are writing-prompt IDs. For a given writing-prompt, y-axes show whether the generated character has identical categories across all 3 inference calls ($y = 1$) or if at least one label differs ($y = 0$). For *stylization*, *coherency*, *wholeness*, *literalness*, *complexity*, *transparency*, *dynamism* and *closure*, respectively 30%, 22%, 27.5%, 13.5%, 53.5%, 45.5%, 35.5% and 3% of the writing-prompts yield characters with the same label across all inference calls.