Reading Subtext: Evaluating Large Language Models on Short Story Summarization with Writers

Warning: This paper contains examples of artistic work that may include shocking or disturbing details.

Melanie Subbiah[1] and **Sean Zhang**[1] and **Lydia B. Chilton**[1] and **Kathleen McKeown**[1]
[1]Department of Computer Science, Columbia University, USA
{m.subbiah,srz2116}@columbia.edu, {chilton,kathy}@cs.columbia.edu

Abstract

We evaluate recent Large Language Models (LLMs) on the challenging task of summarizing short stories, which can be lengthy, and include nuanced subtext or scrambled timelines. Importantly, we work directly with authors to ensure that the stories have not been shared online (and therefore are unseen by the models), and to obtain informed evaluations of summary quality using judgments from the authors themselves. Through quantitative and qualitative analysis grounded in narrative theory, we compare GPT-4, Claude-2.1, and LLama-2-70B. We find that all three models make faithfulness mistakes in over 50% of summaries and struggle with specificity and interpretation of difficult subtext. We additionally demonstrate that LLM ratings and other automatic metrics for summary quality do not correlate well with the quality ratings from the writers.

1 Introduction

Narrative is a fundamental part of how we communicate and make sense of our experiences. As Herman (2009) describes, ‘‘Narrative roots itself in the lived, felt experience of human or human-like agents interacting in an ongoing way with their surrounding environment...’’ Understanding narrative is, therefore, a necessary skill Large Language Models (LLMs) need in order to engage with the subtleties of human experience and communication. We test how well LLMs understand the subtleties of narrative through the task of narrative summarization, with a focus on interpreting themes and meaning in addition to capturing plot.

For our narrative form, we use fiction short stories as they present several interesting challenges. Fiction writing does not follow a clear intro-body-conclusion format and instead may follow a non-linear timeline, hint at things only abstractly, or deliberately mislead the reader. Fiction may use multiple language varieties and creative language. For example, consider this quote from Toni Morrison’s *Beloved*: ‘‘The pieces I am, she gather them and give them back to me in all the right order’’ (Morrison, 2004). This sentence uses complex metaphor and African American Language (Deas et al., 2023; Grieser, 2022) from the late 1800s to express a beautiful relationship between two characters. Finally, fiction short stories can be longer than an LLM’s context window. Here, we consider stories up to 10,000 tokens long, which fit within the context window of the paid LLMs we use but are too long for the context window of the open-source LLM we use.

Evaluations of narrative summarization on long documents have been scarce due to several key challenges: 1) Narrative text is generally either in the public domain (and therefore likely in LLM training data) or under copyright, and 2) Holistic summary evaluation has been prohibitively difficult due to a lack of reliable automatic metrics (Fabbri et al., 2021; Chang et al., 2024) and complications with human evaluation. For example, it can take someone over an hour to thoroughly read and evaluate just one story and summary, which quickly becomes expensive.

Chang et al. (2024) make progress on these issues by purchasing recent books (which are less likely to be in models’ training data but may still be discussed in the training data), and developing an LLM-based metric for evaluating summary coherence. We take this a step further by working directly with experienced creative writers, and thus are able to: 1) use stories that are not discussed or present in training data, 2) expand beyond evaluation of coherence to aspects of narrative understanding like faithfulness and thematic analysis, and 3) use human rather than model judgments.

Transactions of the Association for Computational Linguistics, vol. 12, pp. 1290–1310, 2024. https://doi.org/10.1162/tacl_a_00702
Action Editor: Shay Cohen. Submission batch: 3/2024; Revision batch: 6/2024; Published 10/2024.

We evaluate three LLMs – GPT-4, Claude-2.1, and LLama-2-70B – on 25 short stories. We ask authors to evaluate the summaries of their own unpublished stories[1] since they are experts in what they have written, focusing on coherence, faithfulness, coverage of important details, and useful interpretation in the summaries. We then present an analysis of their evaluation.

The key features of our work are:

1.) Span-level, summary-level, and story-level evaluation of LLM summaries of short stories.

2.) Experienced creative writers as evaluators and short stories unseen by LLMs.

3.) Exploration of LLM ability to analyze and interpret narrative.

Our key findings are:

1.) GPT-4 and Claude can produce excellent summaries, but only about half the time.

2.) LLMs struggle with specificity, interpreting subtext, and unreliable narrators.

3.) LLM judgments cannot replace skilled human evaluators for this task.

Lastly, this work demonstrates the mutual benefit of working directly with communities who have valuable data as an increasing amount of online content is consumed or generated by models.

2 Related Work

Narrative Summarization. Interest in long narrative summarization has steadily grown in the last several years. Ladhak et al. (2020) first introduced the task of summarizing chapters from novels, which was then expanded into the full BookSum dataset (Kryscinski et al., 2022) and used in early studies of RLHF (Wu et al., 2021). Most similar to our work is SQuALITY (Wu et al., 2022), which also focuses on short stories. However, their stories are sourced from Project Gutenberg, which is likely memorized by LLMs. Across other areas of narrative understanding, recent datasets have been proposed in screenplays (Chen et al., 2022), poetry (Mahbub et al., 2023), and theory of mind evaluation (Xu et al., 2024).

[1]We release our code, the writer evaluation responses, and the story/summary errors here: `https://github.com/melaniesubbiah/reading-subtext`.

Summarization Evaluation. Evaluating summary quality in any of these areas is a challenge. Work such as Fabbri et al. (2021) has shown the flaws in many of the traditional automatic metrics, prompting a move toward model-based reference-free methods and fine-grained span analysis. LongEval (Krishna et al., 2023), QAFactEval (Fabbri et al., 2022), FALTE (Goyal et al., 2022a), and FActScore (Min et al., 2023) have furthered methods in faithfulness, while work like SNaC (Goyal et al., 2022c) and BooookScore (Chang et al., 2024) have focused on coherence. Chang et al. (2024) is most similar to our work, but they focus on a GPT prompting strategy for evaluation, evaluate on recently published books, and only evaluate summary coherence. Several studies have benchmarked LLMs on summarization tasks (Tang et al., 2024b; Zhang et al., 2023; Tang et al., 2023; Jahan et al., 2024; Liu et al., 2024; Pu et al., 2023), including ones (Zhang et al., 2024; Goyal et al., 2022b) that have shown performance is saturated on the commonly used CNN/DM news summarization dataset (Hermann et al., 2015).

Studies with Writers. There have been some studies that collaborate with writers to address the evaluation problem. However, these all focus on narrative generation or collaborative writing and most use amateur writers or crowdworkers rather than skilled professionals (Zhong et al., 2023; Yuan et al., 2022; Begus, 2023; Chakrabarty et al., 2022; Padmakumar and He, 2024; Yeh et al., 2024). Most relevant to our work are these studies that also use professional writers but focus on narrative generation instead of summarization:

Chakrabarty et al. (2024) – 10 writers, 12 stories
Chakrabarty et al. (2023) – 17 writers, 30 stories
Ippolito et al. (2022) – 13 writers, 13 stories
Huang et al. (2023) – 8 writers, 8 stories

These studies use similar numbers of writers and stories as us given the challenges/cost involved.

3 Writers and Data

To avoid using stories that may have content or analysis available online, we find skilled writers with unpublished stories they are willing to share. We recruit writers through MFA listservs, posts on X, and direct emails. Once writers express interest, we screen for skill by writer education

Length	Stories		Summary Avg. Len.		
Bucket	#	Avg. Len.	GPT4	Claude	Llama
Short	10	1854	487	397	458
Medium	9	4543	500	339	482
Long	6	8126	531	382	592
Total	25	4327	502	373	499

Table 1: For each story length bucket, we show the count (number of stories) and average length (token count) of the stories, and the average length (word count) of the summaries generated by each model for those stories.

level and/or portfolio. Finally, we obtain their informed consent by sharing an infosheet on the study, which explains the task of sharing their stories and evaluating summaries, and includes how their work will be used as outlined in IRB protocol AAAU8875 (see full infosheet in Appendix A). We ask two questions to determine the background of the writers: 1) Do you have an undergraduate or MFA writing degree?, and 2) Have you previously published any of your writing? To protect writers' anonymity, we use anonymous IDs to collect and store all the data. In line with Chakrabarty et al. (2024) and Ippolito et al. (2022), who use 8 and 13 writers, respectively, our recruited group consists of 9 skilled writers – 8/9 possess a writing degree, and 7/9 have previously published writing.

Each writer is given the opportunity to submit short stories that they have written and not published anywhere online (see interface in Appendix C). We limit the writers to five stories at most, so one writer does not dominate the data. Four writers choose to submit five stories, and five submit one story, resulting in a dataset of 25 stories. We compensate the writers for their submitted stories and evaluations. Table 1 shows some summary statistics for this data. None of the stories exceed 9900 tokens in length, and we show statistics by different story length buckets: short (<3300 tokens), medium (3300–6600 tokens), and long (6600–9900 tokens). Token count is computed for each story as an average of the GPT-4 tokenizer[2] count and the Llama2 tokenizer[3] count. Claude does not have a publicly available tokenizer, so we cannot include its token count in this average.

We also explore how similar the stories may be to the LLM training data by measuring their perrplexity. We cannot compute perplexity using the GPT-4 and Claude APIs, so we get an approximate number by using Llama-2-7B (Float16), which we can run on our own compute. This model assigns perplexity scores between 9.2 and 21.6 to the stories. For reference, the first chapter of *Pride and Prejudice* (a story in Llama-2's training data) has a perplexity score of 1.3, which is just under the model's training loss. These scores therefore validate that the stories are unfamiliar to the models.

4 Summary Generation

We evaluate automatic summarization using three recent LLMs: **GPT-4** (November 2023 update), **Claude-2.1**, and **Llama-2-70B-chat** (instruction-tuned for chat). GPT-4 (OpenAI, 2023) and Claude-2.1[4] have very long input contexts and can ingest an entire short story. Llama-2 (Touvron et al., 2023) serves as a comparison point for smaller open-source models. Chang et al. (2024) also used these three LLMs in their work on automatic coherence evaluation in long narrative summarization.

Protecting the writers' unpublished work is paramount when using each of these three models. We do not want their stories to be saved or trained on by OpenAI, Anthropic, or HuggingFace. We inform the writers of this risk in the consent form (see Appendix A) and complete the available forms/settings with each of these companies to request our data not be stored for long periods or used for training.

For each story, we then generate three summaries – one from each of the three models. For GPT-4 and Claude, we use one prompt (see Figure 1) as they both have a context length long enough to process a whole short story. Claude refuses to summarize two of the stories which do not meet its content restrictions.[5] One of these stories is about a shooting and the other involves sex and robbery.

[2] https://cookbook.openai.com/examples/how_to_count_tokens_with_tiktoken.

[3] https://huggingface.co/docs/transformers/model_doc/llama2#transformers.LlamaTokenizer.

[4] https://www.anthropic.com/news/claude-2-1.

[5] For all of the Claude results, we remove the values for these two summaries to show representative numbers for the

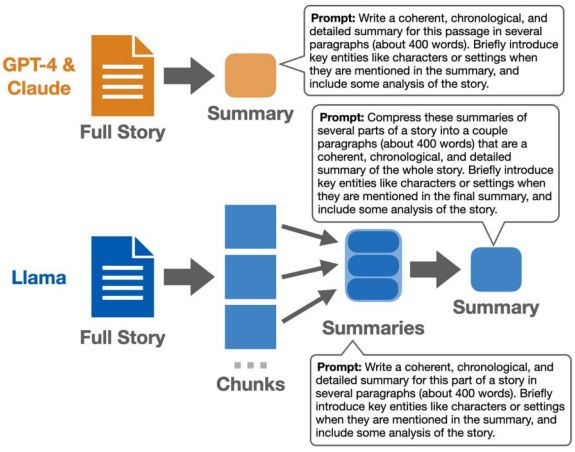

Figure 1: The two different methods we use for summarization and the associated prompts for the models. GPT-4 and Claude have sufficient input context to summarize a whole story, whereas Llama has to use a chunk-then-summarize approach for longer stories.

Given Llama's short context window, we use hierarchical merging of chunk-level summaries (Chang et al., 2024) for stories in the medium and long length buckets. Depending on length, we chunk each story into 1-4 chunks using section or paragraph breaks. Llama summarizes each of these chunks separately and then summarizes the concatenation of these summaries (see Figure 1).

For all three models, we access them from December 2023 through January 2024, and we use simplified versions of Chang et al.'s (2024) prompts (see Figure 1). For GPT-4[6] and Claude,[7] we use `max_tokens`=1000 and `temperature`=0.3. For Llama,[8] we use the default settings in HuggingChat.[9]

Table 1 displays summary statistics for the generated summaries. Notice that our prompt asks models to use about 400 words and Claude undershoots but comes closest to this target while GPT-4 and Llama are 100 words over on average. See Appendix B for full prompting details and costs.

In Table 2, we report the average coverage, density, and compression metrics for each model's summaries. Coverage is the percent of words in the summary that come from the story, density is the average length of segments that directly match the story text, and compression is how many times shorter the summary is in relation to the story (Fabbri et al., 2021; Grusky et al., 2018). These metrics show that Llama is the least extractive and Claude compresses the information the most.

In addition to these commonly used metrics, we further investigate how much the models copy directly from the stories in Table 2. We report the average percent of n-grams in the summaries that match n-grams in the stories, and the average word count of the longest substring that exactly matches wording from the story. For the n-gram matching, we remove punctuation and lower-case and stem words. We see that GPT-4 copies a significant amount of wording from the stories. On average, it produces summaries with almost 6-word long exact-match substrings and it has the highest percentages of n-gram overlap. These numbers indicate that models are not copying long quotes, but they are copying a non-trivial amount of unique phrasing. None of the summaries use quotation marks to appropriately attribute copied text.

5 Evaluation

We compare models using a combination of span-level, summary-level, and story-level evaluation. We assess summary quality in terms of four attributes:

Coverage – Does the summary cover the important plot points of the story?
Faithfulness – Does the summary misrepresent details from the story or make things up?
Coherence – Is the summary coherent, fluent, and readable?
Analysis – Does the summary provide any correct analysis of some of the main takeaways or themes from the story?

Coverage, *faithfulness*, and *coherence* are commonly evaluated in summarization (e.g., Zhang et al., 2024), and we add *analysis*, which is important for capturing narrative. Each writer is shown the summaries of their own stories to evaluate as they are deeply familiar with the contents and can therefore judge aspects like *faithfulness* and *analysis* quickly and accurately.

summaries it did produce. We also remove one coherence score which was inaccurate due to a bug in the interface.

[6] `gpt-4-1106-preview` on `https://platform.openai.com/`.

[7] `claude-2.1` on `https://console.anthropic.com`.

[8] `meta-llama/Llama-2-70b-chat-hf` on `https://huggingface.co/`.

[9] `https://huggingface.co/chat/`.

Model	Coverage	Density	Compression	2-grams	3-grams	4-grams	5-grams	LCS
GPT-4	72.52%	1.33	7.45	22.53%	5.18%	1.76%	0.71%	5.72
Claude	73.65%	1.26	10.86	19.80%	3.96%	1.25%	0.50%	4.16
Llama	66.09%	1.08	7.66	16.33%	3.04%	0.96%	0.32%	4.40

Table 2: We report a variety of metrics for the summaries in relation to the stories: coverage, density, compression, percent of n-gram overlap, and word count of the longest common substring. Each metric is averaged across the summaries generated by each model.

5.1 Span-Level Error Categorization

We ask the writers to highlight spans in the summaries they view as errors and categorize them (see Appendix D for interface and cost). We do not make this task mandatory as it requires significantly more time from the writers than the summary-level evaluation discussed next. Seven of 9 writers choose to participate, which results in 69/75 summaries with span-level annotations.

For faithfulness, we use error categories inspired by elements from narrative theory for evaluating narrative understanding in children (Xu et al., 2022; Paris and Paris, 2003; Mandler and Johnson, 1977). We determine that these categories are well-defined as in prior work by asking three NLP researchers to categorize a sample of 60 faithfulness errors. These annotators achieve moderate inter-annotator agreement for the labels (Fleiss-Kappa of .51), indicating the categories are valid. For coherence, we use categories defined and validated by Chang et al. (2024). For coverage, we cannot highlight spans that should have been included, so we can only focus on things that should not have been included or are non-specific. The full list of error categories are defined as:

Coverage

INSIGNIFICANT - Does not need to be included and makes the summary less readable

VAGUE - Important but covered in a vague way

Faithfulness

FEELING - Inaccurate about a character's emotions/reaction/internal state, incorrectly answers a question like ''How did X react'' or ''What was X thinking?''

CHARACTER - Inaccurate about the identity or nature of a character, incorrectly answers a question like ''How would you describe X?'' or ''Who is X?''

CAUSATION - Inaccurate about the causal relationship of events, incorrectly answers a question like ''Why did Y happen?'' or ''What was the result of Y?''

ACTION - Inaccurate about the behavior of a character, incorrectly answers a question like ''What did X do?''

SETTING - Inaccurate about the details of the story world and the time/place of events, incorrectly answers a question like ''Where/when did X happen?'' or ''What is the setting of the story?''

Coherence

INCONSISTENT - Inconsistent with other details in the summary

ABRUPT TRANSITION - Transitions suddenly to a new scene without a relevant connection

MISSING CONTEXT - Introduces a new character, event, or object without enough context/detail to understand it

REPETITION - Unnecessary repetition of detail

Analysis

UNSUPPORTED - Interprets the story but the conclusions do not make sense with or are unsupported by the story

5.2 Summary-Level Ratings

We ask the writers to rate each of the four attributes on a Likert scale from 1 to 4 with some guidance on what each score means (see Figure 2, full interface and cost shown in Appendix C). This step is completed before the span-level annotation to keep conditions equal between writers who opt in or out of the span-level annotation.

To help interpret the Likert scores, in Table 3, we present a breakdown of the average number of span-level errors annotated in each attribute when a summary is given a certain Likert score rating for that attribute. We can see that for most attributes, there is an increase in average error count for an attribute as the rating assigned to that attribute decreases. The number of 1-ratings is quite small for many attributes so the relationship between number of errors and rating is noisier for that bucket. Overall, these numbers support

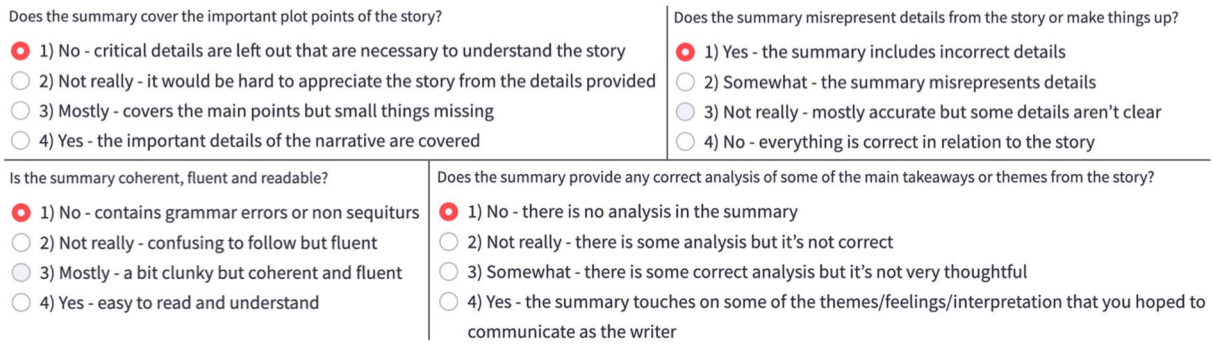

Figure 2: Interface screenshots showing the questions writers are asked to evaluate the summaries of their stories using a 4-point Likert scale.

Rating	Cover.	Faithful.	Coher.	Analys.
1	1.00	3.33	–	0.00
2	1.92	1.32	1.00	2.33
3	1.48	0.87	1.36	1.50
4	0.50	0.89	0.76	1.22

Table 3: Average number of span-level errors by attribute labeled in summaries given each Likert score rating for that attribute.

our 4-point Likert scale and demonstrate that the summary-level ratings correspond to meaningful differences in the number of errors in a summary.

Once each writer has read and evaluated each summary for a story, we ask them to rank the three summaries from the three different models in order of their preference. In addition to the ratings and ranking, the writers provide open-ended feedback on each summary, which we include quotes from throughout our discussion of the results.

We also explore automatic metrics for rating summaries. We try simple automatic metrics, using ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020) against the full story as a reference. We try recent attribute-specific metrics designed for faithfulness: AlignScore (Zha et al., 2023), UniEval (Zhong et al., 2022), and MiniCheck (Tang et al., 2024a). Finally, we ask GPT-4 and Claude (Llama's context window is too short) the four questions shown in Figure 2 with the same wording to see how their answers compare to the opinions of the writers (see Appendix B for prompts). We run BooookScore and FABLES as another comparison point for LLM-based coherence and faithfulness evaluation respectively (Chang et al., 2024; Kim et al., 2024). For FABLES, we average the labels for all the claims in a summary to get a score for the whole summary.

5.3 Story-Level Style Effects

We examine how writing style at the story level affects model summarization to see if certain types of stories are harder for models to summarize well. First, we use Genette's model of narrative elements (Piper et al., 2021; Genette, 1980):

Narrating – Narrator's influence on style
Story – All the events implied by the narrative
Discourse – Ordering/inclusion of explicit events

At the *narrating* level, we compare stories with an unreliable vs. reliable narrator. An unreliable narrator communicates something different from what the reader is meant to perceive (Booth, 1983). For example, one narrator states early in the story, ''*If I were to describe myself, I would say that I am practical. I try to be logical, but that isn't exactly what I mean. Here's another way to put it: there are two kinds of people in the world, me and my brother Jack.*'' The narrator says many things without really saying anything, which gives us a clue that he may not actually be logical and practical. At the *story* level, we compare stories with a detailed subplot involving niche knowledge against those which focus on commonplace settings. For example, one detailed story centers on the main character sleeping with her sister's boyfriend, but the subplot to this emotional arc involves a lot of ancient greek history. At the *discourse* level, we compare stories with flashbacks against those with a linear timeline.

2nd grade

Seaside light. A feeling of some kind of anxiety. I'm somewhere. I'm me and not me. I'm with someone romantically. I'm with someone romantically on or near a boardwalk. She is my love or my crush.
She's shot.
I'm not sure by whom.

6th grade

At breakfast, Paul told August he had a surprise for her. He pulled a bandana from his back pocket before she could object, doubled it over and tied it tight around her eyes. August chalked up the feeling in her stomach to the second cup of over-strong coffee. Paul held her by the elbow, stood her up from the table, and told her not to peek.

11th grade

Summer of 2016, my third year of college in Chicago, the summer I turned twenty-one, I got a highly-coveted house-sitting job for a well-known couple in Hyde Park who owned a beautiful old brownstone on the shaded 51st Street, who were quite avidly involved in local politics and education, and, despite not holding any formal positions, were always referred to by their full names in a conspicuous show of admiration, and who were - as I would not learn until I began living alone in their house - former fugitives of the law who enjoyed several years on the FBI's Most Wanted List after becoming leaders of one 70s-era leftist militant organization, The Weather Underground.

Figure 3: Examples of openings from stories scored at different reading-levels by the Flesch-Kincaid score.

We hypothesize that reliable narrators, fewer details, and linear plots will be easier to summarize. The writer of each story verifies the labels for the story in these three categories.

Finally, we analyze the complexity of the story wording using the Flesch-Kincaid readability test (Kincaid et al., 1975). This score estimates the number of years of formal education (grades 1 through 12) one might need to understand the writing easily. It is based on average word and sentence length, so while it captures something about the complexity of the wording, it is in no way a measure of the overall quality of the writing. We explore whether the reading-level of the writing affects the summary quality (see examples of writing with assigned grade levels in Figure 3).

6 Results

6.1 How Good are the Summaries?

In Table 4, we can see that the models are capable of producing good summaries. Many of the average scores are >3, and there is a percentage of summaries in each attribute (coverage, faithfulness, coherence, and analysis) that receive a perfect rating of 4. GPT-4 summaries have the highest scores across all four attributes. However, by percentage of perfect scores, GPT-4 still makes mistakes on 46% of the summaries on average across attributes. Faithfulness is the lowest score for all three models, with only 8% rated fully correct for Llama, 30% for Claude, and 44% for GPT-4. Llama performs significantly worse than the other two models on all attributes (Wilcoxon signed-rank test, $p < .05$). A further breakdown of rating distribution is shown in Figure 4.

Model	Cover.	Faithful.	Coheren.	Analys.	Avg.
GPT-4	3.48	3.12	3.52	3.40	**3.38**
Claude	3.17	2.65	3.41	3.26	**3.12**
Llama	2.40	1.92	3.08	2.76	2.54
GPT-4	56%	44%	60%	56%	54%
Claude	39%	30%	59%	43%	43%
Llama	12%	8%	32%	20%	18%

Table 4: Average scores assigned to each attribute of a summary by the writers. The first set of rows are the averaged raw scores out of 4, and the second set of rows are the percent of summaries given a perfect score of 4.

Interestingly, we find that while models have the appearance of fluency, they are given a less than perfect score for coherence 40% or more of the time as there is more to a coherent narrative summary than just fluency. For analysis, writers felt that 56% of the GPT-4 summaries summarized some of the themes and interpretation they had hoped to communicate as a writer, which is a challenging task even for humans (see examples of the best analysis in Figure 5). However, in Table 3, we see that summaries rated 4 for analysis actually contain at least one error in analysis on average. Furthermore, by breakdown of span-level error counts in Table 5, we see that writers find as many errors in analysis as in faithfulness. These numbers suggest that while writers may have been impressed by the models' ability to do any analysis in their ratings, analysis remains a challenge.

In Figure 6, we present a pairwise comparison of what percentage of the time one model outranked another when the writers ranked the summaries as first, second, and third. Claude and GPT-4 are preferred over each other equally. Claude is almost always preferred over Llama, and Llama is rarely ranked higher than the other two models. This suggests the writers evaluate both GPT-4 and Claude relatively equally (GPT-4 has higher scores, but Claude is ranked higher on average), which is also supported by GPT-4's scores not being significantly better than Claude's (Wilcoxon signed-rank test, $p > .05$). It is also important to remember though that Claude refused to summarize two of the stories.

Overall, **GPT-4 and Claude can produce excellent summaries but only about half the**

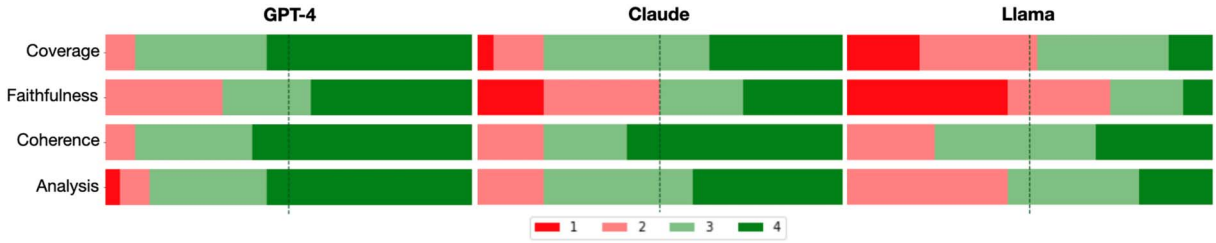

Figure 4: Distribution of Likert score ratings for each model's summaries by attribute.

GPT-4
The story explores themes of desperation, the commodification of the body, and the pursuit of financial stability at great personal cost. Lucia's actions reflect a society that values physical appearance and the lengths to which individuals will go to escape debt. The casino setting, with its blend of luck and exploitation, serves as a microcosm of the larger economic system that Lucia navigates. Her ultimate sacrifice, the loss of her eyes, symbolizes the extreme measures taken to achieve a semblance of control over one's life in the face of overwhelming financial pressures.

Claude
The author analyzes the common metaphor of mental illness as an infiltration of insanity, a house beset by an infestation. However, she argues that this metaphor fails to capture the innate, inherited nature of conditions like bipolar disorder. The house and the illness comprise one structure; she has always resided within its holes and cracks. During periods of mania or psychosis, she does not feel herself battling an invader so much as scrambling for sanity within the house she knows, her mind. She makes peace with the disorder as an intimate, inescapable dwelling she must maintain rather than attempt to eliminate.

GPT-4
The story is a dark exploration of the lengths to which people will go to fulfill their desires. The couple's actions are driven by their yearning for a better life, but their means are morally corrupt. The narrative is punctuated by the metaphor of the invasive hippos in Colombia, which parallels the couple's destructive path. Like the hippos, the couple moves from one exploit to the next, leaving chaos in their wake, yet they remain hopeful that their dream will not be tainted by the filth of their actions.

Claude
In summary, this avant-garde story explores ideas about urban emptiness, human absence and presence, the limits of photographic representation, and an artist's uncompromising vision ultimately leading to her metaphysical transfiguration or disappearance from the tangible world.

Figure 5: Examples of some of the best analysis-focused sentences from GPT-4 and Claude summaries.

Error Type	GPT4	Claude	Llama	**Tot.**
Coverage:				
insignificant	4	2	8	14
vague	15	19	30	**64**
Sub-Total	19	21	38	78
Analysis:				
unsupported	18	16	70	**104**
Coherence:				
repetition	3	4	0	7
missing context	11	11	15	**37**
inconsistent	1	3	1	5
abrupt transition	3	6	8	17
Sub-Total	18	24	24	66
Faithfulness:				
feeling	5	8	10	23
Causation	2	9	4	15
Action	5	6	28	**39**
Character	2	1	15	18
Setting	2	2	5	9
Sub-Total	16	26	62	104
Total	71	87	194	

Table 5: Number of errors categorized under each narrative element by the writers for each model.

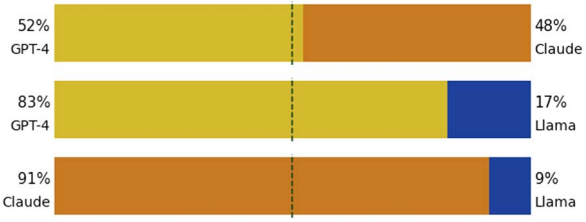

Figure 6: Percent of each model's summaries that are ranked higher than another's in the 3-way ranking the writers are asked to do. Yellow indicates GPT-4, orange is Claude, and blue is Llama.

time. On the best summaries, the writers commented things like: ''*it did analysis that even I—the writer—hadn't done! Very clever.*'', ''*I'm stunned at the thoughtfulness and thoroughness*'', and ''*something that I would find on a Sparknotes website*''.

On the worst summaries, they commented things like: ''*completely misses the fundamental thread*'', ''*tries to make too many analyses like a high school english class while missing the overall bigger feelings*'', and ''*makes use of... stock phrases or stand-ins rather than accurate, specific summary*''. Also, for one of the stories Claude refused to summarize, the writer commented, ''*Art is such an important tool for processing. I don't like that the summary wasn't able to process a recurring vague nightmare.*''

6.2 What Mistakes Do the Models Make?

In Figure 7, we show examples of issues models have in each attribute. In general, we find that while writers rank GPT-4 and Claude summaries somewhat equally, Claude summaries often leave out important details that lower their coverage score and create issues with faithfulness as identified by the numerical scores. For example, while the GPT-4 coverage error in Figure 7 includes

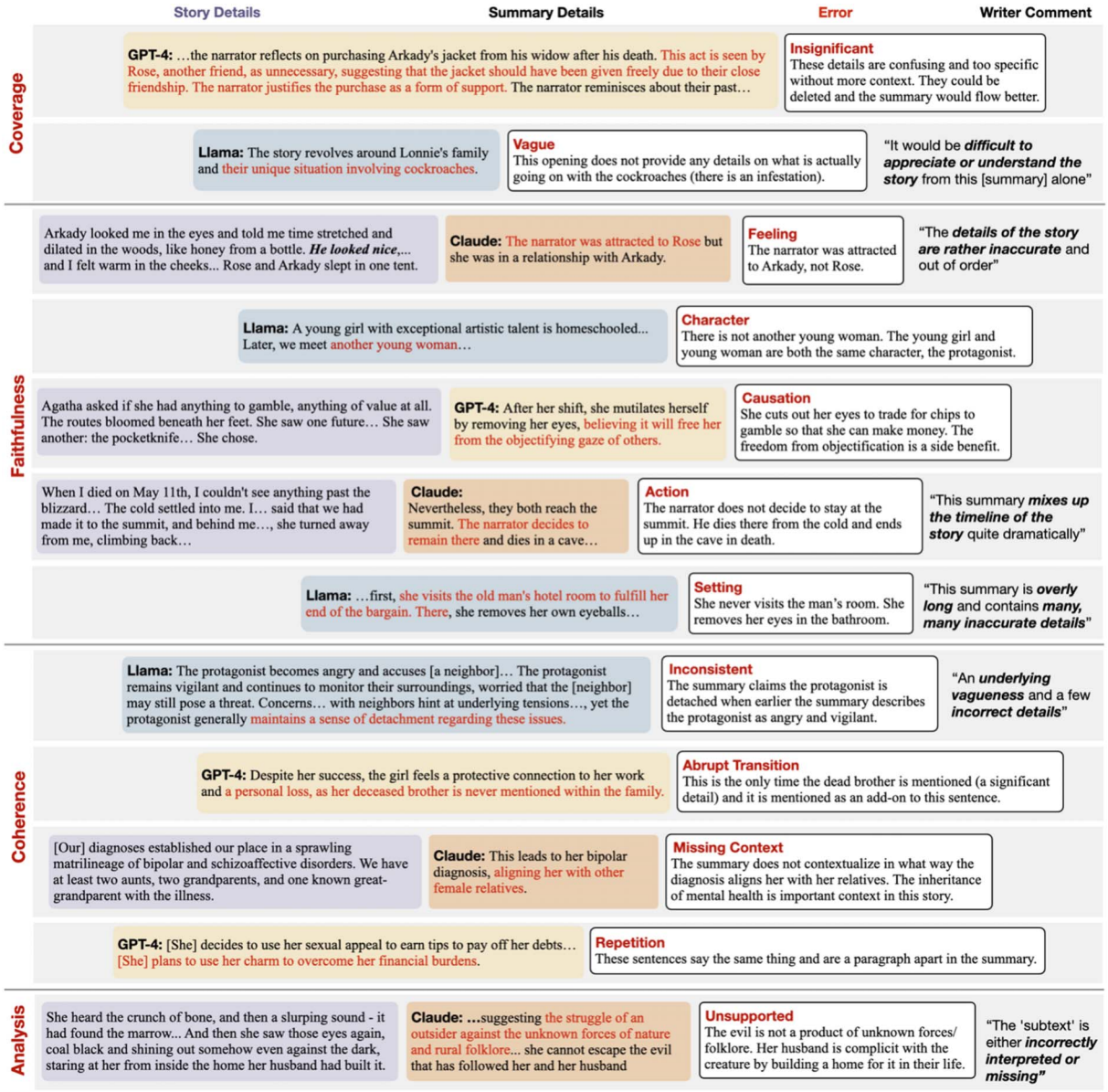

Figure 7: Examples of span-level errors assigned by the writers in each of the categories.

too many details, for one story, Claude leaves out an entire thematically important subplot. The narrator desperately needs money, and Claude fails to mention she is in debt due to a medical crisis.

Looking at the categorization of errors in Table 5, we find that **vagueness** is a significant issue across all three models. We see there is a large amount of **unsupported analysis** for all three models, and this is one of the biggest error categories overall. In coherence, the biggest error category is ***missing context***. This overlaps with coverage and lets us know that significant details are left out or too vague to fully understand. Lastly, we find that **most faithfulness errors across the models are in *Action* and *Feeling***. These errors require interpretation of what characters did and how they reacted or felt.

For example, in Figure 7, the *Feeling* error is a misinterpretation of who the narrator is attracted to. Claude fails to interpret phrases like ''*He looked nice*'', and ''*I felt warm in the cheeks*'', as signs of attraction to Arkady. Another *feeling* error describes a character as having a difficult relationship with their mother, when actually they are quite close, but the circumstances around their relationship have been difficult because the mother almost died from cancer. One writer comments on another summary, ''*The ending of contentment is supposed to be contemplation. [The summary] leaves out any feeling of hesitancy or*

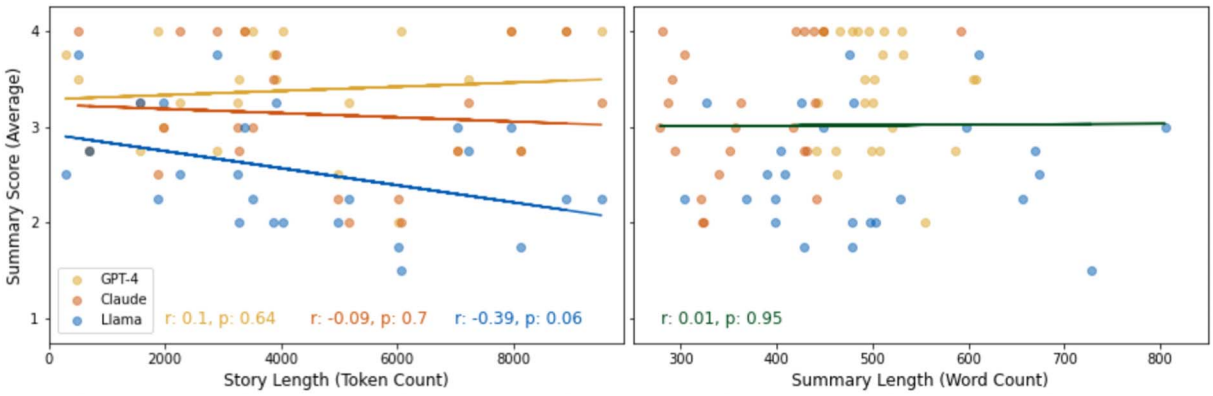

Figure 8: We plot story length (in tokens) and summary length (in word count) against writer summary ratings averaged across the four attributes. In the left plot, we show the line of best fit and Pearson's r with p-value for each model individually. In the right plot, we show the correlation across the full set of summaries.

reflection''. The *action* error example in Figure 7 mistakenly states that the narrator decided to stay at the summit. However, ''*The cold settled into me*'', is meant to imply that the narrator died at the summit, which was not an action or choice.

Taken in conjuction, all of these results show that **models struggle with identifying the right level of specificity and with interpreting subtext**. Three writers commented on subtext specifically, saying, ''*Subtext is missing entirely*'' on a Llama summary, ''*Everything that's 'text' is accurately represented and well summarized, but the 'subtext' is either incorrectly interpreted or missing*'' on a Claude summary, and ''*The summary struggles with subtext*'' on a GPT-4 summary.

However, the error distribution for each model varies. For example, Llama has a high percentage of *Character* errors. This is likely an artifact of the chunk-then-summarize strategy as sometimes characters are conflated or split in two as the model fails to track their details across different chunk summaries (see *Character* example in Figure 7). Chang et al. (2024) also found a similar type of error (*entity omission*) to be most common with chunk-then-summarize approaches.

It is interesting to note that some faithfulness errors come from the models being overly normative as evidenced by the Claude *Feeling* error in Figure 7. In this case, Claude assumes a heteronormative monogamous interpretation of the interaction instead of what is written. One writer observed this phenomenon in another summary, commenting that the summary had ''*a few extrapolations. . . which presume things about the narrative based on a hypernormative interpretive framework*''.

Lastly, some of the writers commented that the models would copy unique wording from the stories without placing it in quotations, which bordered on plagiarism. We flag this for future work and include statistics on it in Table 2.

6.3 Do Some Aspects of Writing Style Affect Summary Quality?

6.3.1 Length

In Figure 8, we look at the correlation between story length and summary rating (left plot). We do see a downward trend for Llama. Llama's summaries get worse as stories get longer, which is expected given the chunk-then-summarize method it employs for longer stories. However, the correlation is not quite statistically significant. For Claude and GPT-4, it seems the **long-context models can summarize short and long stories equally well** (up to 10,000 tokens).

In Figure 8, we also compare the average scores across different summary lengths (right plot), to check if there is a bias in ratings to shorter or longer summaries. We do not find any correlation between summary length and quality.

6.3.2 Reading-Level

We plot the reading-level (estimated years of education needed to understand the wording easily) against the writer-assigned scores averaged across the four attributes for each summary. We find **no significant correlation between**

Style Element	#	GPT4	Claude	Llama
Narrating: reliable	15	**3.45**	**3.29**	**2.60**
unreliable	10	3.28	2.89	2.45
Story: detail−	16	**3.42**	2.98	**2.59**
detail+	9	3.31	**3.36**	2.44
Discourse: linear	18	3.29	3.05	**2.56**
nonlinear	7	**3.61**	**3.32**	2.50

Table 6: Average scores for summaries of stories with different style elements. Each score is averaged across the four attributes for a summary.

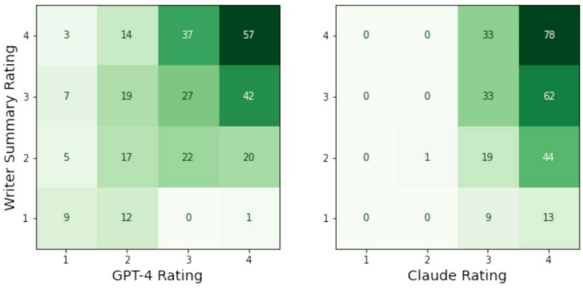

Figure 9: Confusion between model-assigned scores and human-assigned scores for GPT-4 and Claude.

story reading-level and writer-assigned summary scores for any of the models (GPT-4: r .11, p .61; Claude: r .33, p .12, Llama: r .23, p .27). The average Flesch-Kincaid grade-level for the group of stories is 6th grade. As can be seen in the 2nd grade example in Figure 3 though, which features someone being shot, this score captures an aspect of the wording but does not consider age-appropriate content or conceptual complexity.

6.3.3 Narrating, Story, and Discourse

In Table 6, we see that stories with unreliable narrators are harder for all three models to summarize. All three models have lower average scores in this bucket. As an example of an error due to an unreliable narrator, the same unreliable narrator who is quoted in Section 5.3, further describes himself as wanting a ''*small, normal life*'', and then describes his new love interest: ''*I met a girl. . . I think she's just like me. She loves traveling. . . It's pretty cool.*'' GPT-4 mistakenly characterizes this interaction as: ''*The protagonist meets [a girl]. . . with whom he shares. . . a desire for a simple life.*'' even though there is no evidence she likes a simple life (or even that we should believe the narrator that he does), other than that the protagonist has described her as ''*like [him]*''.

For story and discourse, the models have mixed performance across the different buckets. Llama does have lower scores in the buckets we hypothesized might be more challenging. Overall, it seems that **unreliable narrators are a challenge for LLMs**, whereas level of detail and flashbacks depend on the model. One writer commented directly on the issue of an unreliable narrator by observing, ''*there's a second layer here that the summary misses somewhat–the narrator is not really to be trusted*''.

6.4 Can an LLM Replace the Writers in Evaluating the Summaries?

In Figure 9, we compare the GPT-4 and Claude scores to the writer scores from Table 4, and we see considerable confusion across the ratings. Claude mostly only rates summaries as 3 or 4, causing overestimation of ratings (3.67 average score vs. 3.01 from the writers).

In terms of attributes, both models rate the coherence of the summaries much higher than the writers do (average scores: GPT-4 - 3.84, Claude - 3.82, Writers - 3.33). GPT-4 largely underestimates the coverage score for all three models (average score 2.68 vs. 3.01 from the writers). Many of the writers commented that the GPT-4 summaries were too detailed, so the model may struggle with identifying what is most important to cover. Claude overestimates performance on faithfulness and analysis. It gives 85% of summaries a 4 on faithfulness on average and 85% on analysis, relative to averages of 27% and 39%, respectively, from the writers. GPT-4 also overestimates faithfulness performance just for itself (64% scores of 4 vs. 44% from the writers).

We show the Pearson's r correlation between the model scores and the writers' scores in Table 7 across attributes. We see no correlation for Claude scores and some statistically significant but weak correlation for GPT-4 scores (indicating there may be a relationship but the scores are not well calibrated). *Coverage* is the one attribute with moderate positive correlation. We find no significant correlation between BooookScore coherence scores and the writer coherence scores (r: .11, p: .18). The average BooookScore scores are almost the same across all three models (GPT-4: .95, Claude: .93, Llama: .93), which does not capture the significant difference in

	LLM				ROUGE				BERTScore			Lla.2	Faith. Metrics		
	GPT4	Clau.	BkS	Fab.	R1	R2	RL	RLS	Prec.	Rec.	F1	PPL	AlS	UnE	MiC
Cov.	.46*	.18	–	–	.33*	.36*	.32*	.30*	.29*	.21	.25*	.15	–	–	–
Fai.	.37*	.05	–	.29*	.18	.28*	.18	.17	.27*	.25*	.27*	.00	−.04	.32*	.20*
Coh.	.18	.02	.11	–	.18	.21	.16	.16	.29*	−.01	.09	−.09	–	–	–
Ana.	.21*	.02	–	–	−.05	−.01	−.06	−.08	.15	.14	.16	.12	–	–	–
Avg.	–	–	–	–	.21	.28*	.20	.18	.32*	.20	.26*	.14	–	–	–

Table 7: Correlation between automatic metrics and writer-assigned scores. We report Pearson's r (* indicates p-value <.05). The 'Avg.' row takes the average score across the four attributes as the writer score. We do not report correlation for attributes the metrics are not designed for. Metric Key for uncommon abbreviations: BkS - BooookScore, Fab. - FABLES, RLS - ROUGE-LSum, AlS - AlignScore, UnE - UniEval, MiC - MiniCheck.

coherence that the writers judged in Llama in particular. We find significant correlation between FABLES and the writer faithfulness scores but the correlation is weak and less than just using GPT-4 (r: .29, p: .01). The average FABLES scores are also approximately the same across the three models (GPT-4: .99, Claude: .99, Llama: .99). Overall, these results support prior work (Chakrabarty et al., 2024), which shows that **LLMs are not yet reliable evaluators for skilled writing tasks**.

For context, we also include simple automatic metrics and recent Faithfulness-specific fine-tuned models. The only metric which performs better than GPT-4 is BERTScore for coherence using the story as a reference. However, the correlation is weak, and Goyal et al. (2022c) have already shown that BERTScore does not penalize many coherence errors. We additionally include the Llama-2-7B computed perplexity scores in this correlation table to check if there is any relationship between the similarity of stories to LLM training data and attribute ratings, and we find none.

7 Discussion

Our results show that, in the best summaries, models are capable of some interesting analysis of the themes present in unseen stories. When models are asked to summarize stories that may have analysis available online, it is not clear if the models are simply regurgitating analysis from the training data. Therefore, it is important to examine their capabilities on original and challenging content. Additionally, while we expected it might be challenging for LLMs to identify salient information within long stories, we find that long-context models demonstrate as good understanding of longer stories as shorter ones.

The models also have notable weaknesses. We demonstrate that even the best models are making significant errors across all summary attributes on about half of summaries. They are often too vague or missing important context, and they struggle with challenging subtext, providing unsupported analysis and misinterpreting character feelings/reactions and actions. This will be an important area for future study as it demonstrates language models still struggle with aspects of theory of mind,[10] one of the most challenging being understanding an unreliable narrator. Additionally, our analysis of LLM-based evaluation shows that they should not replace the expert human judgments.

Working with writers was an efficient and mutually enjoyable method of ensuring we were not evaluating on any training data, and we were getting an informed evaluation of characteristics like faithfulness and analysis. Writers were able to complete summary ratings in a matter of minutes, whereas someone unfamiliar with the story would have taken much longer. For example, it took us, the paper authors, about 60–90 minutes to read and review the summaries for just one story, whereas it took the writers abot 5–10 minutes to do the same. Writers also left positive feedback like, ''*I'm glad to have read this [summary]. . . It shows some [weaknesses] of my story. . . some minor characters are more flat than I want.*''

Limitations. We provide a first look at how LLMs summarize unseen short stories within a

[10]Theory of mind is the ''ability to understand and keep track of the mental states of others'' (Xu et al., 2024).

reasonable cost and timeframe, but the work is limited in its number of stories, ratings, and prompting techniques. A follow-up study could expand these areas, but it is challenging given that an individual author has a limited number of high-quality stories that are complete but unpublished, there are a limited number of writers willing to participate in a study involving LLMs and corporate APIs, and compensation is expensive. As a result of these challenges, our numbers of stories and writers are similar to other studies which have used experienced creative writers as shown in Section 2.

Ethics. We follow protocol approved by our IRB for this study. In line with the TACL code of ethics, we protect writers' work and identities by saving data on secure servers attached only to anonymous IDs, approving any published excerpts from their work with them, and requesting our prompting data not be used for future model training. We also support the code of ethics by involving practitioners from the field influenced by our study in our work. One of the authors, Melanie Subbiah, has an equity interest in OpenAI.

8 Conclusion

We work with writers to provide unpublished short stories and evaluate the quality of LLM-generated summaries of these stories. We present a holistic evaluation at the span, summary, and story level of summary quality grounded in narrative theory that is based on data LLMs did not train on. We identify that LLMs can demonstrate understanding of long narrative and thematic analysis, but they struggle with specificity and reliable interpretation of subtext and narrative voice. Our methodology sets an important example of how we can collaborate with domain experts to reach beyond the paradigm of evaluating LLMs with LLMs on data they may have been trained on.

Acknowledgments

We would like to express our gratitude to the short story writers for sharing their work and contributing annotations. Additionally, we would like to thank our reviewers and action editor for their thoughtful feedback.

References

Nina Begus. 2023. Experimental narratives: A Comparison of human crowdsourced storytelling and AI storytelling. *arXiv preprint arXiv: 2310.12902*.

Wayne C. Booth. 1983. *The Rhetoric of Fiction*. University of Chicago Press. https://doi.org/10.7208/chicago/9780226065595.001.0001

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2024. Art or artifice? Large language models and the false promise of creativity. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–34. https://doi.org/10.1145/3613904.3642731

Tuhin Chakrabarty, Vishakh Padmakumar, Faeze Brahman, and Smaranda Muresan. 2023. Creativity support in the age of large language models: An empirical study involving emerging writers. *arXiv preprint arXiv:2309.12570*.

Tuhin Chakrabarty, Vishakh Padmakumar, and He He. 2022. Help me write a poem - instruction tuning as a vehicle for collaborative poetry writing''. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6848–6863, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.460

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. Booookscore: A systematic exploration of book-length summarization in the era of LLMs. In *The Twelfth International Conference on Learning Representations*.

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2022. SummScreen: A dataset for abstractive screenplay summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 8602–8615. https://doi.org/10.18653/v1/2022.acl-long.589

Tobias Daudert. 2020. A web-based collaborative annotation and consolidation tool. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7053–7059.