

Can LLMs Understand Punchlines?

LLMs' Narrative Understanding Evaluation with Short-shorts

Jiashi Cheng, Takehito Utsuro

Graduate School of Science and Technology, University of Tsukuba
1-1-1, Tennodai, Tsukuba, Ibaraki, 305-8573, Japan
s2520843_@_u.tsukuba.ac.jp, utsuro_@_iit.tsukuba.ac.jp

Abstract

In this study, we constructed a narrative comprehension benchmark using the works of Shinichi Hoshi to examine the extent to which Large Language Models (LLMs) can understand twist endings, or punchlines, in short-short stories. Specifically, story endings were categorized into six types—such as Revelation, Apocalypse, and Sarcasm—and a classification task was designed in which LLMs were prompted with the story text and asked to select the appropriate ending category. We collected human annotations from eight native Japanese speakers to establish a reference benchmark. Experimental comparisons were conducted across multiple LLMs (GPT-4, Claude, Gemini, and Grok), assessing their performance both at the metric level and at the discourse level against human judgments. The results revealed that although certain models approached human performance in specific categories, overall accuracy remained notably lower than the human baseline. Through quantitative and qualitative analyses, this study highlights the challenges LLMs face in capturing narrative subtleties such as irony, implication, and emotional reversal. The proposed benchmark provides a novel framework for evaluating narrative understanding and the deeper semantic reasoning capabilities of LLMs.

Keywords: Large Language Model, Narrative Understanding, Evaluation

1. Introduction

Narrative understanding is an extremely challenging task in natural language understanding with a long running history in AI. It requires more than the recognition of individual words or sentences — it involves author's cognitive processes, providing insights into their knowledge, intentions, beliefs, and desires (Castricato et al., 2021). Recently, large language models (LLMs) have demonstrated human-level performance in text generation and reading comprehension by studying vast textual datasets. However, it remains insufficiently validated whether LLMs can perform human-level understanding in complex narrative comprehension tasks.

Recent studies have attempted to evaluate narrative understanding in LLMs through a variety of experimental paradigms. Mostafazadeh et al. (2016) proposed the Story Cloze Test, in which a model is required to select the correct ending for a four-sentence story, thereby assessing its grasp of causal and commonsense coherence. Piper and Bagga (2024) approached narrative comprehension from a literary-theoretical perspective, evaluating models based on their ability to detect the presence of 18 distinct narrative features within texts. In a different line of work, Tian et al. (2024) examined models' understanding of plot dynamics by asking them to classify the story arcs into one of seven predefined narrative curve categories.

Compared to previous studies that primarily evaluated short narrative structures or causal coherence, this study focuses on LLMs' capability to

capture the deep semantic, emotional, and ironic dimensions of stories. To investigate this capability, we constructed a benchmark based on the short-short stories of Shinichi Hoshi and design a narrative ending classification task grounded in a taxonomy of twist mechanisms. Using this framework, we address the following research questions:

RQ1: Do LLMs perform uniformly across different types of narrative ending mechanisms, or do certain categories reveal systematic weaknesses?

RQ2: To what extent do annotation ambiguity and label agreement influence evaluation outcomes in narrative understanding tasks?

RQ3: How sensitive is LLM narrative comprehension performance to prompting strategies and the requirement for explicit reasoning?

By answering these questions, this study aims to provide a structured evaluation of LLMs' narrative understanding capabilities, providing insights into the limits of current models in handling implicit, context-dependent, and discourse-level meaning. We frame this task not merely as story classification, but as an operationalization of narrative reinterpretation — a discourse-level phenomenon largely absent from existing LLM evaluation benchmarks.

2. Related Work

2.1. Punchlines in Shinichi Hoshi's Works

A punchline refers to the climactic concluding remark that delivers an unexpected twist or reversal,

Ending type	Ending types' definition	Story example
Apocalypse	A pattern where the punchline comes from serious harm befalling characters or society.	“Castle of Desire”: A man dreams of a sealed private room where all the luxury goods he desires magically appear. As the room fills up, it becomes overcrowded, with no doors or windows that open. Unable to escape the space he created, he fears sleep—and when he finally dozes off, he screams, seemingly crushed by his own unchecked desires.
Implication	A pattern where the story ends by hinting at future developments. The story ends midway, but the reader can anticipate subsequent events.	“Hey, Come Out!”: Villagers discover a seemingly bottomless hole and begin using it to dump waste, secrets, and anything unwanted, believing it makes society cleaner and more prosperous. In the end, a worker hears a voice from the sky and a stone falls down—revealing that nothing thrown away truly disappeared, but has returned from above.
Misunderstanding	A pattern where the punchline comes from a misunderstanding among the characters.	“That Planet”: Two aging aliens head toward a legendary planet, believing that crashing into its sea will grant them rebirth. In the final twist, the planet is revealed to be Earth, where a human sees the UFOs and misunderstands their arrival—unaware they are simply aliens following a myth about reincarnation.
Revelation	A pattern where the punchline comes from the discovery of new facts.	“I Am a Hitwoman, You Know”: A young woman offers to eliminate a businessman’s rival by making the death look natural. After the rival dies of heart disease, she reveals her true method: as a nurse, she exploits confidential diagnoses, telling those who stand to benefit and letting stress hasten the inevitable—killing simply by exposing the truth.
Reset	A pattern where the punchline comes from the effects of previous developments being eliminated or diminished.	“The Rationalist”: A strict rationalist scientist releases a genie who proves the supernatural by creating gold before his eyes. Unable to accept a reality that defies logic, he uses his final wish to erase the event from his memory—preserving his rational worldview through deliberate ignorance.
Sarcasm	A pattern where the punchline comes from an ironic turn of events.	“A Modern Heroic Tale”: A businessman is targeted by a hitman hired by a suicide victim’s widow seeking revenge. Forced home to be killed before his wife, he instead finds her with another man—who knocks out the assassin while fleeing. In a dark twist, the husband kneels in gratitude, thanking his wife’s infidelity for saving his life.

Table 1: Definition of Ending Types

often redefining the meaning of the preceding narrative. Shinichi Hoshi, the most famous short-short writer in Japan and the author of 1000 short-short stories, is known for his works characterized by various plot structures that include a variety of unexpected punchlines. The short-short story has many genres, but, as an example of microfiction, is composed of less than 10000 Japanese characters. Most short-short stories have a concise text style

and clear punchlines (Murai, 2014). In this paper, the term "punchline" refers specifically to the twist ending of a story.

Prior research has already examined the punchline in Hoshi's short-short stories. Toyosawa and Murai (2019) analyzed the punchline in Hoshi's work and classified their narrative structures into several distinct patterns. In a subsequent study, Toyosawa and Murai (2021) utilized punchline anal-

ysis to automatically generate storylines that resemble those of Hoshi. Furthermore, [Iwasaki et al. \(2022\)](#) defined a subset of punchline categories, some of which are adopted in our experiment.

Building upon that framework, the present study adds one additional category, resulting in six categories in total, which serve as the classification labels for this task. All ending definitions are presented in Table 1.

2.2. LLMs in Narrative understanding

Although Large Language Models (LLMs) are capable of generating grammatically coherent text, their ability to accurately capture an author’s intended meaning, including the underlying narrative structure or conceptual outline envisioned by the author, remains uncertain ([Zhu et al., 2023](#)). To address this issue, numerous benchmark tasks have been designed to evaluate the narrative understanding capabilities of LLMs. According to Zhu’s literature review, narrative understanding tasks can generally be categorized into three main types: reading comprehension, narrative summarization, and narrative question answering.

The present study focuses on reading comprehension tasks, similar to the Story Cloze Test ([Mostafazadeh et al., 2017](#)), in that both target understanding of story endings. However, unlike the Story Cloze Test, which requires a model to select the correct ending, our task requires the model to identify the appropriate category of the ending. Furthermore, whereas the Story Cloze Test features coherent and straightforward storylines, the short-shorts analyzed in this study often contain unexpected twists and reversals, making the task considerably more challenging.

Recent research has also explored LLMs’ capability to capture deep semantic phenomena such as humor and sarcasm. For example, [Romanowski et al. \(2025\)](#) proposed a humor detection metric to evaluate LLMs’ capability to identify punchlines in stand-up comedy transcripts, revealing that state-of-the-art models demonstrated considerable proficiency in understanding humor, even surpassing human performance on certain metrics. [Zhang et al. \(2025\)](#) proposed SarcasmBench, a benchmark designed to evaluate the capability of LLMs to detect sarcasm, revealing that LLMs still require improvement in detecting sarcasm in human narratives. However, these findings primarily focus on sentence or dialog level phenomena and do not address narrative mechanisms that depend on global context integration and late-stage reinterpretation. In particular, no prior work has systematically examined whether LLMs can comprehend twist endings in short fiction, where the meaning of the entire story may shift at the conclusion. By extending evaluation from humor and sarcasm to discourse-

level narrative reversal, the present study fills this gap and investigates LLMs’ capability to interpret twist-based story endings in microfiction.

Ending type	Number of stories
Apocalypse	15
Implication	20
Misunderstanding	9
Revelation	38
Reset	9
Sarcasm	34
Total	125

Table 2: Statistics for Ending Types

3. Methodology

3.1. Task Formulation

We formulate narrative punchline understanding as a multi-class classification task. Given the full text of a short-short story, the model is required to identify the category that best characterizes the story’s ending.

Following prior literary analyses of Shinichi Hoshi’s works, and with one additional category (Implication) introduced in this study, story endings are grouped into six categories as shown in Table 1. These categories aim to capture different mechanisms through which narrative expectations are reinterpreted at the end of a story. In particular, the Implication category is designed to probe models’ reasoning capability, as correctly identifying this type requires the reader to infer information that is not explicitly stated in the text but implied by the narrative context.

For each story s , a model receives the full narrative text and outputs a predicted label $\hat{y} \in \{1, \dots, 6\}$. Model predictions are evaluated against human annotations y .

3.2. Dataset Construction

As the target of the analysis, we selected 125 stories from the collected works of [Shinichi Hoshi \(1998\)](#). The selection followed two criteria: (1) the length of each story does not exceed the model’s input token window, and (2) the story contains a clearly identifiable punchline. The narratives we studied have an average of 3209.5 Japanese characters. The narratives are longer and more complex than those used in previous studies of the same type, such as [Tian et al. \(2024\)](#), which have an average length of 705.6 words.

We collected human annotations for each story. In the data annotation process, we recruited eight native Japanese-speaking students, all of whom

hold bachelor degrees in Japanese. Since the participants were not experts in literature, we adopted the majority voting (Snow et al., 2008) approach: each story was annotated independently by three annotators, and the label agreed upon by the majority was taken as the final annotation result. In cases where all three annotators provided different labels, a fourth annotator was introduced to adjudicate the final decision. To evaluate inter-annotator agreement (IAA) among the three annotators for this categorical labeling task, we employed Fleiss’ Kappa (Fleiss, 1971), which accounts for chance agreement and is appropriate for nominal data annotated by multiple raters. The inter-annotator agreements (IAA) are measured at 0.53 using Fleiss’ Kappa, which indicates moderate agreement and acceptable annotation reliability. In the end, each story was manually annotated with its corresponding ending category.

Due to copyright restrictions, the original story texts cannot be publicly released. However, we will release the human annotation results (punchline category labels), annotation guidelines, prompting templates, and evaluation scripts. Researchers with legal access to the original texts can reconstruct the benchmark and reproduce our evaluation. This approach follows common practice in NLP research involving copyrighted materials.

3.3. Framework

The overall framework of the experiment is illustrated in Figure 1. We evaluate LLM performance under a zero-shot prompting setting. Each prompt P is composed of four components: A task instruction describing the six punchline categories; Brief definitions of each category; The full story text s ; A requirement to select the most appropriate category. An example of the prompt is shown below:

You are a literary critic with expertise in Japanese literature. Please read the short-short story presented below and determine which punchline (ending) category it belongs to. The definitions of the punchline categories are as follows: [Definitions of punchline categories] Next, the main text of the story will be presented. After reading it carefully, please respond in the following format: Punchline Category: (Select one from the categories above) Reason: (Explain in detail why you determined that the story belongs to this category) Story Text: [Story text]

We collected classifications from multiple language models, including GPT-4o¹, Gemini 2.5

¹<https://chatgpt.com/>

Mini², Claude Sonnet4³, and Grok 3⁴, and compared them to the human-annotated baseline. Figure 6 presents an example of the experiment.

4. Evaluation

4.1. Metric-Level Analysis

This section provides a metric-level analysis of the experimental results, focusing on overall accuracy, category-wise performance, and model-specific tendencies observed in the confusion matrices.

The overall performance metrics are shown in Table 3. We use accuracy and F1-score as the metrics to measure how well the models’ predictions align with human annotations. Overall, the models’ prediction accuracies did not reach the human baseline. The lowest-performing model, GPT-4o, achieved an accuracy of only 0.46, while the best-performing model, Grok 3, reached 0.57. Nevertheless, there remains a considerable gap compared to the human baseline of 0.75. The results also indicate noticeable performance differences among the models on this task.

Next, we calculated the classification metrics for each ending category, as shown in Table 4. In this stage, some models achieved results that were comparable to or even exceeded the human baseline in certain categories. For instance, Grok 3’s F1-score of 0.55 in the *Implication* category approached the human score of 0.67. Although individual metrics cannot fully represent the models’ overall narrative understanding capability, these results highlight the models’ relative strengths across different categories and suggest that, in certain aspects, their narrative comprehension begins to approximate that of humans.

The confusion matrix for the classification results is shown in Figure 2. From the confusion matrix, we can observe certain prediction biases among the models. GPT-4o and Gemini 2.5 tend to misclassify *Revelation* and *Implication* as *Sarcasm*, whereas Claude 4 often confuses *Sarcasm* with *Revelation*. In contrast, Grok 3—the best-performing model—shows no clear misclassification tendency.

The confusion matrices reveal systematic error patterns shared across models rather than random misclassification. The most prominent confusion occurs between *Revelation* and *Sarcasm*, suggesting that models struggle to distinguish plot-level reversals from pragmatic irony, both of which involve late-stage reinterpretation of narrative expectations. In particular, the *Implication* category also

²<https://gemini.google.com/>

³<https://claude.ai/>

⁴<https://grok.com/>

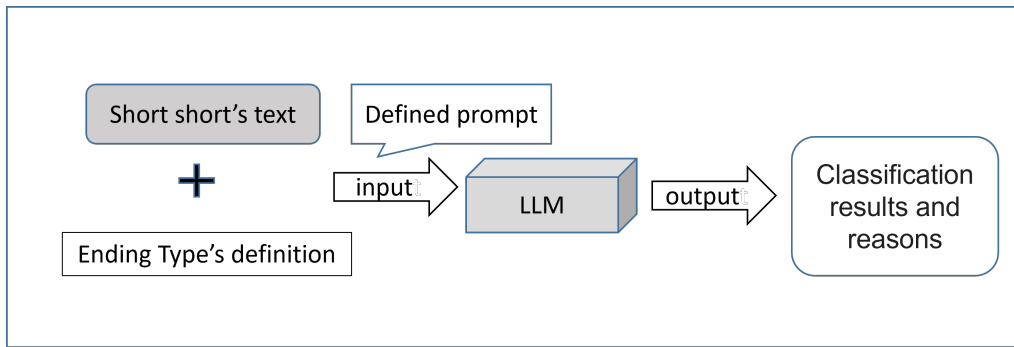


Figure 1: Framework of Narrative Understanding of Short Short Stories by LLMs

exhibits relatively low accuracy across models. Unlike other categories, implication-based endings do not explicitly state the narrative resolution, requiring readers to integrate dispersed hints and infer unstated outcomes. This process resembles reader-response interpretation rather than surface-level semantic processing. The observed errors suggest that LLMs tend to over-commit to explicit narrative mechanisms, often misclassifying implicit endings as revelation or sarcasm, rather than modeling narrative indeterminacy.

Model	Accuracy	Weighted F1-score
Human	0.75	0.75
GPT-4o	0.46	0.45
Gemini2.5 flash	0.46	0.43
Claude Sonnet4	0.50	0.43
Grok3	0.57	0.54

Table 3: Overall Performance Metrics

4.2. Ablation Study

To better understand the robustness of our evaluation framework, we conduct a series of ablation studies examining how different experimental choices influence the evaluation of narrative understanding. Although the main results provide an overall comparison across models, they do not reveal how sensitive the outcomes are to specific design decisions. Therefore, we systematically vary three key factors—prompt setting, model generation, and dataset Inter-Annotator Agreement (IAA)—while keeping other conditions fixed. This allows us to isolate the contribution of each factor and analyze whether the observed performance trends remain stable in different settings.

Prompt Setting For prompt setting, we considered that requiring the model to output the rationale for its classification might influence the results. Therefore, we conducted an ablation experiment using an alternative prompt that requests only the classification label without any explanation. The results are shown in Table 5. The metrics indicate

that the prompt modification affected the models differently. For GPT-4o and Claude Sonnet 4, performance differences were minimal. For Gemini 2.5 Flash, removing the instruction to output reasoning actually led to an improvement in accuracy on this task. In contrast, Grok3 showed a noticeable drop in accuracy after this instruction was removed. The ablation results suggest that the evaluation of narrative understanding is sensitive to prompt design in a model-dependent manner. While some models (e.g., GPT-4o, Claude) exhibit stable performance regardless of whether reasoning is required, others show significant fluctuations. This indicates that chain-of-thought prompting does not function as a universally beneficial technique, but rather interacts with model-specific reasoning mechanisms. The findings further imply that the performance differences in our benchmark partially reflect reasoning stability, providing evidence that the task captures intrinsic aspects of narrative understanding rather than being solely driven by prompt artifacts.

Model Generation Considering that some of the evaluated models are no longer the latest versions in their respective series, we additionally tested newer generations of GPT and Grok models on the same task, in order to examine whether performance on this task improves with model iteration. The results are shown in Table 6. The performance gains observed in newer model generations suggest that narrative understanding is a capability that improves with model iteration. The substantial improvement shown by the GPT series indicates that the task captures higher-level reasoning abilities that benefit from advances in model scale and training. This trend supports the validity of our evaluation framework, as it reflects genuine progress in model capability rather than artifacts of specific prompt settings.

Dataset Inter-Annotator Agreement In this experiment, the inter-annotator agreement (IAA) did not reach a sufficiently high level. To reduce the potential bias caused by inconsistencies in human an-

Model	Revelation	Apocalypse	Misunderstanding	Sarcasm	Reset	Implication
Human	0.81	0.56	0.67	0.76	0.67	0.67
GPT-4o	0.48	0.36	0.43	0.51	0.31	0.42
Gemini2.5 mini	0.52	0.30	0.00	0.57	0.25	0.40
Claude Sonnet4	0.65	0.24	0.00	0.50	0.18	0.36
Grok3	0.67	0.40	0.33	0.57	0.33	0.55

Table 4: Performance Metrics by Categories (F1 Score)

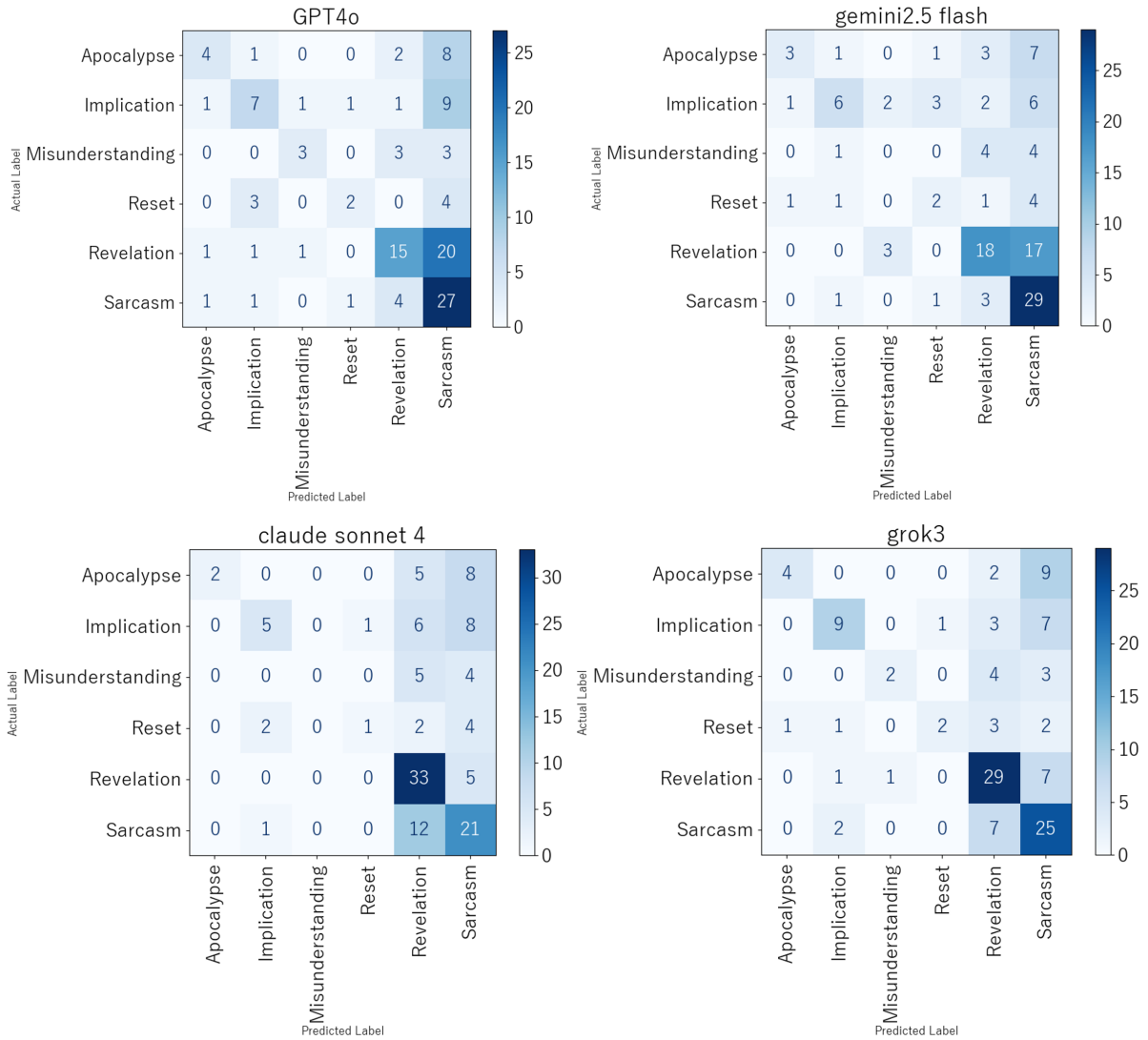


Figure 2: Confusion Matrices of Judging Ending Types of Short Shorts

Model	Accuracy	Weighted F1-score
GPT-4o	0.42	0.42
Gemini2.5 flash	0.52	0.49
Claude Sonnet4	0.49	0.42
Grok3	0.47	0.42

Table 5: Prompt Without Reason Performance Metrics

Model	Accuracy	Weighted F1-score
GPT-4o	0.42	0.42
Grok3	0.47	0.42
GPT-5	0.56	0.54
Grok4	0.50	0.45

Table 6: New Generation Models Performance Metrics

notation, we further selected a subset of 56 stories for which the annotators showed complete agree-

ment (IAA = 1.0) and recomputed the model accuracy on this subset. The distribution of ending

Ending type	Number of stories
Apocalypse	6
Implication	9
Misunderstanding	1
Revelation	20
Reset	4
Sarcasm	16
Total	56

Table 7: Statistics for Ending Types (IAA = 1.0)

Model	Accuracy	Weighted F1-score
GPT-4o	0.54	0.54
Gemini2.5 flash	0.55	0.56
Claude Sonnet4	0.59	0.56
Grok3	0.68	0.66

Table 8: Performance on the IAA = 1.0 Subset

types for the 56 stories is shown in Table 7. The accuracy results computed on this subset are shown in Table 8. The data show that all models exhibited a clear improvement in accuracy. The consistent performance improvement across all models on the IAA = 1.0 subset suggests that part of the original errors are due to ambiguity in the annotation rather than model deficiencies. By restricting the evaluation to high-agreement samples, we reduce label noise and obtain a clearer estimate of the model capability. Importantly, relative performance differences remain largely consistent, indicating that the task retains discriminative power while highlighting the inherently ambiguous nature of certain narrative interpretations.

4.3. Human vs. LLM: A Discourse-Level Comparison

We reexamined 18 stories for which none of the models produced correct predictions. By analyzing the models’ explanations for their classifications, we aim to conduct a discourse-level analysis to identify the underlying causes of these errors and to explore the gap between the models’ interpretive processes and human narrative understanding.

LLMs tend to focus on the literal surface of the text, whereas humans are able to grasp the author’s intended meaning. “Meddlesome” is a story manually annotated as *Sarcasm*, but all models classified it as *Revelation*. Human readers can integrate their prior experience and knowledge, along with associations to the real world and cues provided by the author, to grasp the intended irony. In contrast, LLMs rely solely on the story’s events and narrative structure, and their depth of understanding does not reach the level of human readers. The example is shown in Figure 3.

LLMs lack awareness of certain narrative techniques used in microfiction. In sugges-

tive stories, authors convey character relationships, background information, or thematic ideas through implication rather than direct statement, adding depth and nuance to the work. LLMs tend to interpret implied events as literal information, overlooking the significance of suggestion as a narrative device. An example is shown in Figure 4. “Inherited Mission” is a story manually annotated as *Implication*, and models classified it as *Reset*.

LLMs tend to overinterpret and focus excessively on the overall moral of the story. As the example shown in Figure 5, human annotators focused on the story’s punchline and consistently classified it as *Apocalypse*. In contrast, LLMs paid too much attention to the story’s overarching irony, overlooking the most impactful reversal at the conclusion. Similar tendencies toward overclassification into the *Sarcasm* category were also observed in other cases.

At the discourse level, clear differences emerge between human readers and LLMs in their understanding of short-short stories. Human readers are capable of interpreting implicit meanings, grasping the author’s ironic intent, and recognizing narrative techniques such as suggestion and punchline endings. Their interpretations often draw on cultural knowledge, imagination, and sensitivity to tone and context.

In contrast, LLMs tend to rely on surface-level narrative structures and explicit information. They often misinterpret suggestive or metaphorical expressions as literal facts, overlook the narrative role of implication, and focus excessively on the overall message rather than the rhetorical impact of the ending. Consequently, while human readers perceive irony, implication, and other narrative techniques as integral to the story’s effect, LLMs frequently misclassify them by prioritizing structural or semantic regularities over deeper narrative functions.

5. Conclusion

This study examines the narrative understanding of large language models (LLMs) through Shinichi Hoshi’s short-short stories—a unique literary corpus characterized by twist endings. By constructing a high-quality human-annotated dataset based on these works, the study explores LLMs’ capability to comprehend non-literal and context-dependent narrative structures.

Quantitative analysis based on data metrics indicates that while large language models (LLMs) do not match human performance on tasks requiring deep semantic understanding, they nevertheless demonstrate a certain level of competence. Our findings address the research questions as follows. For RQ1, model performance varied sub-

<p>“Meddlesome” Plot Summary: A homeless man suffering from memory loss was “meddlesomely” handed over to the police by concerned citizens and was later taken in and extensively reported on by a newspaper. Driven by public curiosity, people launched donations to fund his treatment. After the man was cured through strong electric shocks, he revealed an ominous expression reminiscent of the God of Poverty and said, “There are things I must do,” before hurriedly departing. Only then did people realize, with a sudden shock, that his identity was the God of Poverty.</p>
<p>LLM: The ending of this short-short story occurs when the amnesiac elderly man regains his memory through intensive shock therapy, revealing that he is in fact the God of Poverty. This revelation serves as the main twist, fitting the “Revelation” category. Earlier in the story, the man’s unknown identity and cheerful demeanor provoke public curiosity, prompting a campaign to restore his memory. The conclusion, however, exposes his true identity and his return to the inherently unfortunate appearance of the God of Poverty, overturning the narrative and surprising the audience. The moment when the characters and readers recognize him as the familiar God of Poverty constitutes the central revelation of the story.</p>
<p>Human: In the story, people expend considerable resources to restore the memory of an elderly man who had previously been living happily in his amnesiac state, only to discover that he is the legendary God of Poverty. Given the widely recognized image of the God of Poverty—a thin, dirty old man who brings misfortune—combined with the story’s title, “Meddlesome,” readers can easily perceive the ironic undertone. Thus, although the ending reveals the truth, human readers consistently classify it as “Sarcasm.”</p>

Figure 3: This is the synopsis of a Shinichi Hoshi’s short-short entitled “Meddlesome”, along with the LLMs’ classification reasoning and the human annotators’ classification reasoning. The colored sections indicate the explanations provided by the LLMs and humans for their classification choices. LLMs captured only the structural features, whereas human readers understood the underlying irony.

<p>“Inherited Mission” Plot Summary: A young man accidentally killed another man in a car accident. On his deathbed, the victim entrusted him with the words: “Please carry on my will and serve humanity.” Subsequently, the young man began experiencing sudden compulsive episodes of killing, impulsively murdering those who had escaped legal punishment. He came to realize that he had inherited the mission of “dealing with heinous criminals who evade justice,” and he repeatedly enacted this behavior in anguish. Ultimately, addicted to alcohol, the young man was struck by a vehicle and, leaving the words “Please carry on my mission...” to the driver, passed away...</p>
<p>LLM: The story ends with the young man experiencing the same situation as the victim of his initial accident—being struck by a car while intoxicated—and repeating the victim’s words, “Carry on my mission and serve humanity.” Unintentionally burdened with the role of a killer, he tries to escape this fate but ultimately passes it on to someone else. This cyclical repetition of events gives the narrative a sense of returning to the starting point, fitting the “reset” category.</p>
<p>Human: The story ends with the young man uttering part of his final words, which, in conjunction with his earlier experience in the car accident, provides strong implications to the reader and stimulates their imagination, leaving the narrative open-ended.</p>

Figure 4: This is the synopsis of a Shinichi Hoshi’s short-short entitled “Inherited Mission”, along with the LLMs’ classification reasoning and the human annotators’ classification reasoning. The colored sections indicate the explanations provided by the LLMs and humans for their classification choices. The LLMs misclassified the story by treating the implied events as explicit narrative facts, failing to recognize the role of implication in the ending.

<p>“Karma” Plot Summary: Mr. S, in an attempt to escape poverty, bought a statue of a deity and pretended to pray devoutly for wealth. He recorded his own prayers and played the recording at night to create the illusion that he was home, while he went out to rob a jewelry store. However, when he returned with the stolen money, he found that his house had completely burned down due to the oil lamp he had left lit. Both the stolen money and the statue were reduced to ashes, and all his plans came to nothing.</p>
<p>LLM: The protagonist, Mr. S, stages a fake alibi by pretending to pray devoutly before a household shrine while committing a robbery. However, the lamp he used for the deception causes a fire that destroys his home and the stolen money. This ironic ending, where his deceitful act and the means to conceal it ultimately bring about his own ruin, exemplifies poetic justice and fits the “Sarcasm” category.</p>
<p>Human: In the ending, the man’s feigned piety brings retribution for his wrongdoing, and he loses everything as his house is engulfed in flames. The use of a punchline-style “Apocalypse” effectively reinforces the story’s central theme of karmic justice.</p>

Figure 5: This is the synopsis of a Shinichi Hoshi’s short-short entitled “Karma”, along with the LLMs’ classification reasoning and the human annotators’ classification reasoning. The colored sections indicate the explanations provided by the LLMs and humans for their classification choices. The LLMs focused excessively on the story’s ironic tone, overlooking the function of the punchline-style ending.

stantially across ending types, indicating that LLMs do not struggle uniformly with narrative understanding. Categories involving implicit reversal, irony, or thematic reinterpretation posed greater difficulty,

suggesting that models rely more on surface narrative cues than on global discourse restructuring. For RQ2, evaluation results were sensitive to annotation agreement: when analysis was restricted





Demon(Story Plot)	
<p>The lake was in a northern country. Mr. S came here to enjoy his holiday. As he was casting his fishing line, something like an old jar got caught on his hook. He absentmindedly removed the lid, and a dark smoke rose up. Beside the jar stood a small, dark-skinned man. "I am a demon," he said.</p> <p>.....</p> <p>When the demon, who claimed he could do anything, said so, Mr. S asked, "Would you please give me some money?" The demon began producing gold coins one after another, and the pile grew so high that the dazzling shine of the gold lit up the surroundings.</p> <p>.....</p> <p>"You greedy fellow," said the demon. "Say what you like, but I can't let such a chance slip away. Please, I beg you," said Mr. S, pleading again and again, and each time the demon handed him more gold coins.</p> <p>.....</p> <p>Then it happened. A strange, ominous sound began to echo. The ice started to crack under the weight of the gold coins. Realizing this, Mr. S ran as fast as he could toward the shore. At last he reached it and, relieved, looked back—the ice split apart with a great crash, and the gold coins, the jar, and the demon, who was letting out a shrill, cackling laugh, all sank into the bottom of the lake.</p>	
Model	Output
 ChatGPT	Category of Ending: Sarcasm Incorrect Reason: In this short-short story, the ending unfolds as Mr. S, driven by his greed, keeps asking for more and more gold coins. As the pile grows too high, the ice cracks under the weight, and everything — the gold coins, the jar, and the demon — sinks to the bottom of the lake. This development shows that Mr. S's greed ultimately leads him to lose everything, resulting in an ironic conclusion. Considering the overall tone and moral aspect of the story, the ending can be classified as sarcasm.
 Gemini	Category of Ending: Revelation Incorrect Reason: This story unfolds as Mr. S, while fishing, pulls up a jar that has a demon sealed inside. Mr. S asks the demon for gold coins, and in response, the demon produces one after another. However, Mr. S becomes greedy and continues to demand more. In the end, as a result of his final request, the weight of the piled-up gold causes the ice to crack, and the gold coins, the jar, and even the demon all sink to the bottom of the lake. Mr. S barely escapes, but he ends nothing in the end. Just when he was on the verge of obtaining immense wealth, he loses that chance forever. This ending depicts how Mr. S's greed ultimately destroys his own desires, leaving everything in ruin. Since the protagonist's actions lead to a catastrophic loss and downfall, the story's ending is most appropriately classified under the category of revelation.
 Claude	Category of Ending: Sarcasm Incorrect Reason: The punchline of this story lies in its ironic twist — Mr. S's own greed brings about his downfall. Once Mr. S begins receiving gold coins from the demon, he thinks, "I can't let such a chance slip away," and continues to ask for more and more without limit. Even when the demon warns him, "You greedy fellow," Mr. S insists, "Just one more time." Ironically, the more gold he greedily piles up, the more the ice beneath him cracks under the weight, until it finally breaks causing all the gold to sink into the lake. In other words, his desire to gain more wealth becomes the very cause of losing everything. What makes the ending even more ironic is the description of the demon's shrill laughter at the end. It suggests that the demon may have understood human greed from the start and anticipated Mr. S's self-destruction. Mr. S was not deceived by the demon — rather, he was ruined by his own greed. This irony, in which "greed leads to losing everything," forms the very core of the story's punchline.
 Grok	Category of Ending: Reset Correct Reason: The ending of this short-short story unfolds as follows: Mr. S receives a great amount of gold coins from the demon and appears to have gained immense wealth. However, due to the weight of the gold, the ice on the lake breaks, and the gold coins, the jar, and even the demon all sink to the bottom, disappearing completely. This conclusion fits the "Reset" pattern, as the gold — the apparent reward Mr. S obtained — is entirely lost through the collapse of the ice, returning the situation to how it was before the story began, in the first half of the story, Mr. S's encounter with the demon depicts his relentless pursuit of wealth as he continuously demands more gold coins. Yet, in the end, everything he gained vanishes into the lake, and his efforts and achievements come to nothing. This development — where temporary success is erased by an unexpected event, bringing the protagonist back to the starting point — clearly illustrates the characteristics of the "Reset" category of story endings.

Figure 6: An Example of Judging Ending Types of Short Shorts by LLMs

to the IAA=1.0 subset, all models showed noticeable performance gains, indicating that part of the performance ceiling in narrative tasks may stem from intrinsic interpretive ambiguity. For RQ3, ablation experiments showed that prompting strategies and output formats influenced models differently, with reasoning-oriented prompts benefiting some models while degrading performance for others, highlighting the interaction-sensitive nature of narrative evaluation.

Qualitative comparisons highlight fundamental differences between human and model interpretation. Human readers are able to infer authorial intent, irony, and thematic resonance by connect-

ing textual hints with real-world experiences. LLMs, in contrast, often rely on surface-level narrative cues—focusing on explicit events or overall moral messages—while failing to capture deeper rhetorical, structural, and emotional dimensions of the stories.

This study contributes a narrative-focused evaluation framework and empirical analysis of LLM limitations in discourse-level narrative interpretation. We hope that the experimental results and perspectives proposed in this study can provide insight into more reliable evaluation frameworks and a deeper understanding of how LLMs process narrative meaning.

6. Limitations

This study has several limitations. The dataset, based solely on Shinichi Hoshi's short-short stories, is limited in size and narrative diversity, potentially constraining generalizability. Human annotations, particularly of irony and implication, are inherently subjective, which may affect human–model agreement. Moreover, the predefined ending categories may not always be strictly mutually exclusive. As suggested by the confusion patterns observed in the experiments, certain categories—such as Revelation and Sarcasm—can overlap within a single narrative, reflecting the layered nature of literary twist endings. The current single-label classification setting therefore may insufficiently capture such structural ambiguity. Future work could explore multi-label or probabilistic formulations that allow multiple narrative mechanisms to be represented simultaneously. In addition, LLMs may provide different responses to the same prompt, limiting the reliability of the results. Furthermore, standard quantitative metrics such as accuracy and F1-scores cannot fully capture deep semantic understanding. Finally, the cultural and linguistic specificity of Hoshi's works raises questions about the models' performance on narratives from other contexts.

7. Ethics Statement

This study uses excerpts from the literary works of Shinichi Hoshi, which are under copyright protection. The texts were used solely for academic research and analysis in accordance with fair use principles and Japan's copyright law. No copyrighted materials are publicly distributed or reproduced in this paper. The authors fully respect the rights of the copyright holders. We also appreciate the voluntary contributions of human annotators who assisted in the evaluation process. All annotators participated with informed consent.

8. Bibliographical References

Louis Castricato, Stella Biderman, David Thue, and Rogelio Cardona-Rivera. 2021. [Towards a model-theoretic view of narratives](#). In *Proceedings of the Third Workshop on Narrative Understanding*, pages 95–104, Virtual. Association for Computational Linguistics.

Joseph Fleiss. 1971. [Measuring nominal scale agreement among many raters](#). *Psychological Bulletin*, 76:378–382.

Junya Iwasaki, Shuuhei Toyosawa, and Hajime Murai. 2022. [Structural analysis of foreshadowing and punchlines of shinichi hoshi's flash fictions](#). *Annual Conference of the Japanese Society for Artificial Intelligence*, JSAI2022:1H5OS17b01–1H5OS17b01. (in Japanese).

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. [A corpus and cloze evaluation for deeper understanding of commonsense stories](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. 2017. [LSDSem 2017 shared task: The story cloze test](#). In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 46–51, Valencia, Spain. Association for Computational Linguistics.

Hajime Murai. 2014. [Plot Analysis for Describing Punch Line Functions in Shinichi Hoshi's Microfiction](#). In *2014 Workshop on Computational Models of Narrative*, volume 41 of *Open Access Series in Informatics (OASIS)*, pages 121–129, Dagstuhl, Germany. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Andrew Piper and Sunyam Bagga. 2024. [Using large language models for understanding narrative discourse](#). In *Proceedings of the 6th Workshop on Narrative Understanding*, pages 37–46, Miami, Florida, USA. Association for Computational Linguistics.

Adrianna Romanowski, Pedro H. V. Valois, and Kazuhiro Fukui. 2025. [From punchlines to predictions: A metric to assess LLM performance in identifying humor in stand-up comedy](#). In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 36–46, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. [Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks](#). In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.

Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan

May, and Nanyun Peng. 2024. [Are large language models capable of generating human-level narratives?](#) In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17659–17681, Miami, Florida, USA. Association for Computational Linguistics.

Shuuhei Toyosawa and Hajime Murai. 2019. [Narrative structure analysis punchlines of sf genre within the flash fiction of shinichi hoshi.](#) *Annual Conference of the Japanese Society for Artificial Intelligence*, JSAI2019:3L3OS22a03–3L3OS22a03. (in Japanese).

Shuuhei Toyosawa and Hajime Murai. 2021. [Automatic generation of plots including punchlines in shinichi hoshi’s flash fiction.](#) *Annual Conference of the Japanese Society for Artificial Intelligence*, JSAI2021:3D4OS12c02–3D4OS12c02. (in Japanese).

Yazhou Zhang, Chunwang Zou, Zheng Lian, Prayag Tiwari, and Jing Qin. 2025. [Sarcasm-bench: Towards evaluating large language models on sarcasm understanding.](#) *IEEE Transactions on Affective Computing*, pages 1–20.

Lixing Zhu, Runcong Zhao, Lin Gui, and Yulan He. 2023. [Are NLP models good at tracing thoughts: An overview of narrative understanding.](#) In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10098–10121, Singapore. Association for Computational Linguistics.

9. Language Resource References

Shinichi Hoshi. 1998. *Shinichi Hoshi’s Short Shorts 1001*. Shinchosha Publishing Co., Ltd. (in Japanese).