Can LLMs Understand Story Endings of Shinichi Hoshi's Short Shorts?

Jiashi Cheng, Takehito Utsuro
Deg. Prog. Sys.&Inf. Eng., Grad. Sch. Sci.&Tech., University of Tsukuba

Abstract

This study constructed a narrative comprehension benchmark using works by Shinichi Hoshi to verify the extent to which large language models (LLMs) can understand the endings (*Ochi*) of short short stories. Specifically, we classified story endings into categories such as *revelation*, *apocalypse*, and *sarcasm*, then designed a classification task where LLMs input the story text and select the appropriate ending category. Experiments compared multiple LLMs (GPT-4, Claude, Gemini, etc.), evaluating performance using a metric of agreement rate with human annotations. This study presents a new evaluation framework for examining the narrative comprehension capabilities of LLMs.

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 [1]. 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 on complex narrative comprehension tasks.

Recent studies have attempted to evaluate narrative understanding in LLMs through a variety of experimental paradigms. Mostafazadeh et al. [5] 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 [7] 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. [11]examined models' understanding of plot dynamics by asking them to classify 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 this end, we propose an innovative evaluation task: measuring narrative understanding capability by classifying the types of twist endings in short shorts. We collected a dataset of 125 short shorts by Shinichi Hoshi. Each story's ending corresponds to one of six predefined types. By comparing the classification results of LLMs against human evaluators, we aim to assess the models' capability of narrative understanding.

2 Related Works

2.1 *Ochi* in Shinichi Hoshi's works

"*Ochi*" refers to an unexpected or ironic ending in a story. The works of Shinichi Hoshi are characterized by their brevity and frequent inclusion of such unexpected endings, namely *Ochi*. Prior research has already examined *Ochi* in Hoshi's short stories. Toyosawa [8] analyzed *Ochi* in Hoshi's works and classified their narrative structures into several distinct patterns. In a subsequent study, Toyosawa [9] utilized *Ochi* analysis to automatically generate storylines resembling those of Hoshi. Furthermore, in Iwasaki's study [4], a subset of *Ochi* categories was defined.

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 *Ochi* definitions are presented in Figure 1.

Table 1 Definition of Ending Types

Ending types	Ending types' definition
Revelation	A pattern where the punchline comes from the discovery of new facts.
Apocalypse	A pattern where the punchline comes from serious harm befalling characters or society.
Misunderstanding	A pattern where the punchline comes from a misunderstanding among the characters.
Sarcasm	A pattern where the punchline comes from an ironic turn of events.
Reset	A pattern where the punchline comes from the effects of previous developments being eliminated or diminished.
Implication	A pattern where the story ends by hinting at future developments. The story ends midway, but the reader can anticipate subsequent events.

2.2 LLM in Narrative understanding

Although LLMs are capable of generating grammatically coherent text, their ability to accurately capture an author's intended meaning—such as the underlying narrative structure or conceptual outline envisioned by the author—remains uncertain [12]. To address this issue, numerous benchmark tasks have been designed to evaluate LLMs' narrative understanding capabilities. According to Zhu's [12] 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 the reading comprehension type, similar to the Story Cloze Test [6], in that both tasks target the understanding of story endings. However, unlike the Story Cloze Test, which requires a model to select the correct ending for a story, 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.

Table 2 Statistics of Ending Types

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

3 Data Collection and Annotation

As the target of the analysis, we selected 125 short shorts from the collected work of Shinichi Hoshi [3]. 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 *Ochi*. Each story was manually annotated with its corresponding *Ochi* category. The distribution of *Ochi* categories is shown in Table 2 .

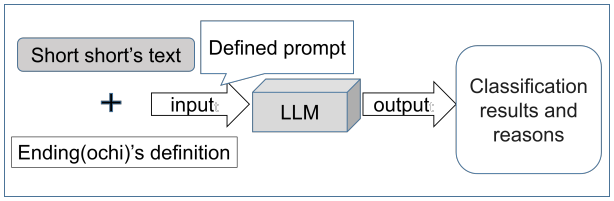

Figure 1 Framework of Narrative Understanding of Short Shorts by LLMs

We seek to collect 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 [10] 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 [2], 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 speaks to the quality of our annotation process. In the end, each story was manually annotated with its corresponding ending category.

4 Experiment

4.1 Framework

The overall framework of the experiment is illustrated in Figure 1. The *Ochi* category definitions listed in Table 1, together with the original text of each story, are incorporated into the following prompt structure to query

Demon(Story Plot)

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.
.....
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.
.....
"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.
.....
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.

 GPT−4o	【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.
 Grok3	【Category of Ending】: **Reset** **Correct** 【Reason】: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, by 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.
 Human	【Category of Ending】: **Reset** **Label** 【Reason】:The story as a whole satirizes the man's greed, but from a structural point of view, the narrative follows a 'reset" pattern: the protagonist starts with nothing, gains wealth, and ultimately returns to nothing. Thus, the storyline clearly fits the definition of the "Reset" category.

Figure 2 An Example of Judging Ending Types of Short Shorts by LLMs

the model. The model is instructed to output both the predicted ending category and a brief rationale based on the story's narrative development. 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 Ochi (ending) category it belongs to. The definitions of the Ochi categories are as follows: [Definitions of Ochi categories] Next, the main text of the story will be presented. After reading it carefully, please respond in the following format: Ochi 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-4o, Gemini 2.5 Mini[1], Claude Sonnet4[2], and Grok 3. In addition, we also collected results from GPT-5 and Grok-4, the latest generations of the GPT[3] and Grok[4] model families. We compared all of them with the human-annotated baseline. Figure 2 presents an example of the experiment.

Table 3 Overall Performance Metrics

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
GPT-5	0.56	**0.54**
Grok4	0.50	0.45

4.2 Evaluation

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

1) https://gemini.google.com/
2) https://claude.ai/
3) https://chatgpt.com/
4) https://grok.com/

Table 4 Performance Metrics by Categories (F1 Score)

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
GPT-5	0.67	0.33	0.43	0.57	**0.40**	0.56
Grok4	0.63	0.00	0.14	0.44	0.18	**0.68**

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. Among the new-generation models, the latest GPT-5 model exhibited a significant improvement in accuracy, whereas Grok4 showed a decline. 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 4' s F1-score of 0.68 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 matrices for the classification results are shown in Figure 3. 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.

5 Conclusion

This study explored the narrative understanding capabilities of Large Language Models (LLMs) when interpreting short-short stories that feature twist endings. To this end, we designed a narrative understanding task in which models were asked to identify the type of *Ochi* (ending) based on the given story text, and we constructed a manually annotated dataset to support this evaluation. We investigated the performance of widely used closed-source models on this task and further examined whether newer generations of selected models demonstrate improvements. Through an analysis of the confusion matrices, we identified distinct tendencies exhibited by different models. The results revealed that LLMs exhibit a certain degree of comprehension in handling narratives with unexpected or ironic conclusions.

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, 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.

In future work, we plan to conduct a qualitative analysis of specific cases in which models make frequent errors, with the aim of exploring common issues faced by LLMs in narrative understanding tasks.

Acknowledgement

We thank Prof. Murai and the co-authors for their collaboration on their works [4, 8, 9]. The research was supported by JSPS KAKENHI Grant Number 25K03416.

References

[1] Louis Castricato, Stella Biderman, David Thue, and Rogelio Cardona-Rivera. Towards a model-theoretic view of narratives. In Nader Akoury, Faeze Brahman, Snigdha Chaturvedi, Elizabeth Clark, Mohit Iyyer, and Lara J. Martin, editors, **Proceedings of the Third Workshop on Narrative Understanding**, pp. 95–104, Virtual, June 2021. Association for Computational Linguistics.

[2] Joseph Fleiss. Measuring nominal scale agreement among many raters. **Psychological Bulletin**, Vol. 76, pp. 378–382, 11 1971.

[3] Shinichi Hoshi. **Shinichi Hoshi's Short Shorts 1001**. Shinchosha Publishing Co., Ltd., 1998. (in Japanese).

[4] Junya Iwasaki, Toyosawa Shuuhei, and Hajime Murai. Structural analysis of foreshadowing and puchlines of shinichi hoshi' s flash fictions. **Annual Conference of the Japanese Society for Artificial Intelligence**, Vol. JSAI2022, pp. 1H5OS17b01–1H5OS17b01, 2022.

[5] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 839–849, San Diego, California, June 2016. Association for Computational Linguistics.

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

[7] Andrew Piper and Sunyam Bagga. Using large language models for understanding narrative discourse. In Yash Kumar Lal, Elizabeth Clark, Mohit Iyyer, Snigdha Chaturvedi, Anneliese Brei, Faeze Brahman, and Khyathi Raghavi Chandu, editors, **Proceedings of the 6th Workshop on Narrative Understanding**, pp. 37–46, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

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

[9] Toyosawa Shuuhei and Hajime Murai. Automatic generation of plots including punchlines in shinichi hoshi's flash fiction. **Annual Conference of the Japanese Society for Artificial Intelligence**, Vol. JSAI2021, pp. 3D4OS12c02–3D4OS12c02, 2021.

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

[11] Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. Are large language models capable of generating human-level narratives? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 17659–17681, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[12] Lixing Zhu, Runcong Zhao, Lin Gui, and Yulan He. Are NLP models good at tracing thoughts: An overview of narrative understanding. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, **Findings of the Association for Computational Linguistics: EMNLP 2023**, pp. 10098–10121, Singapore, December 2023. Association for Computational Linguistics.

A Confusion Matrices

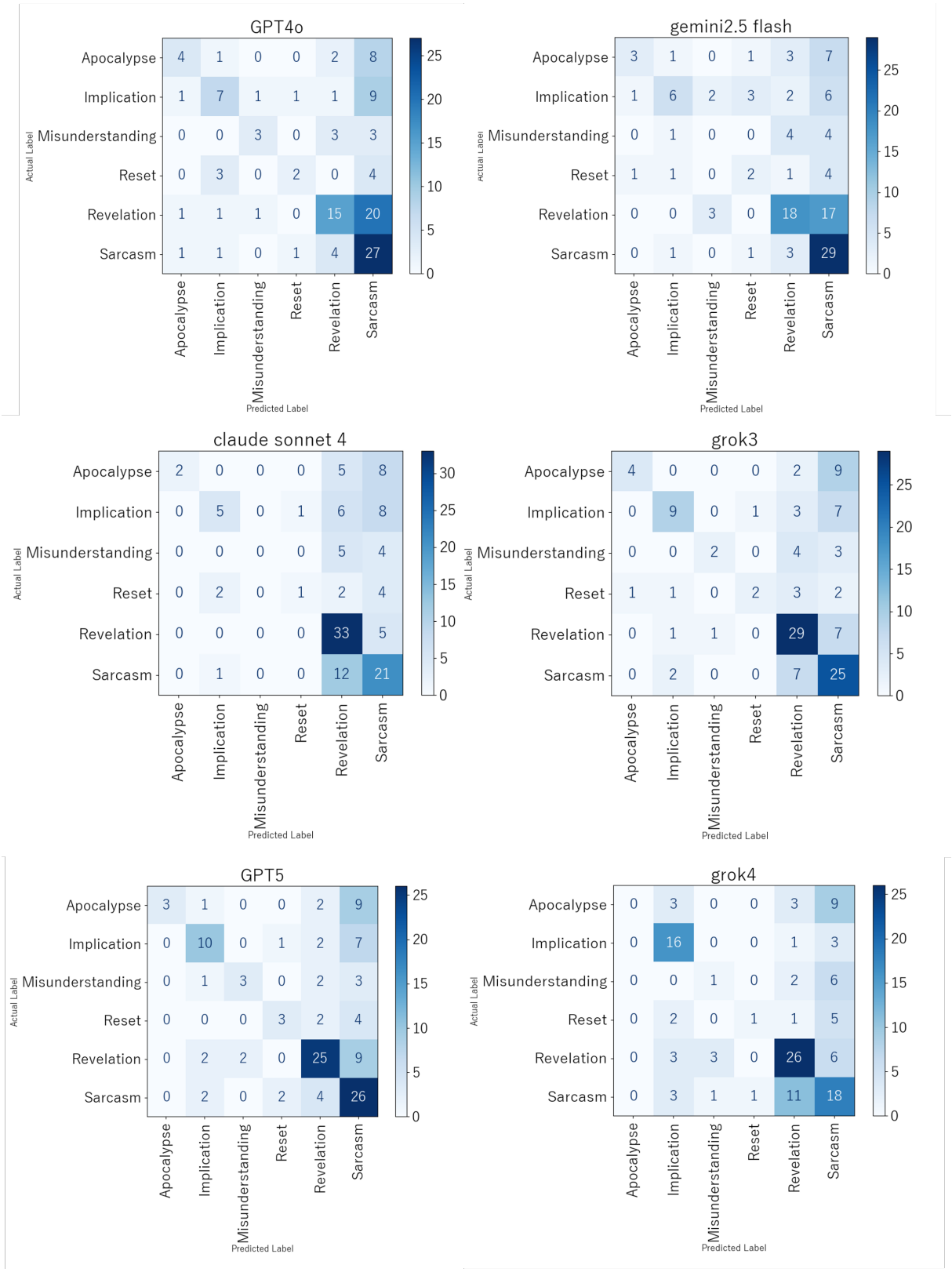

Figure 3 Confusion Matrices of Judging Ending Types of Short Shorts