Evaluating LLM Performance in Character Analysis: A Study of Artificial Beings in Recent Korean Science Fiction

Woori Jang and **Seohyon Jung**
KAIST, South Korea
{woori.jang, seohyon.jung}@kaist.ac.kr

Abstract

Literary works present diverse and complex character behaviors, often implicit or intentionally obscured, making character analysis an inherently challenging task. This study explores LLMs' capability to identify and interpret behaviors of artificial beings in 11 award-winning contemporary Korean science fiction short stories. Focusing on artificial beings as a distinct class of characters, rather than on conventional human characters, adds to the multi-layered complexity of analysis. We compared two LLMs, Claude 3.5 Sonnet and GPT-4o, with human experts using a custom eight-label system and a unique agreement metric developed to capture the cognitive intricacies of literary interpretation. Human inter-annotator agreement was around 50%, confirming the subjectivity of literary comprehension. LLMs differed from humans in selected text spans but demonstrated high agreement in label assignment for correctly identified spans. LLMs notably excelled at discerning 'actions' as semantic units rather than isolated grammatical components. This study reaffirms literary interpretation's multifaceted nature while expanding the boundaries of NLP, contributing to discussions about AI's capacity to analyze and interpret creative works.

1 Introduction

Literature has long been a realm where diverse characters interact within narratives, offering deep insights into human nature and the essence of humanity (or non-humanity) across time, cultures, and genres (Piper, 2024; Eder et al., 2010; Frow, 2014). Science Fiction (SF), in particular, presents a more diverse lineage of character types compared to realist novels, featuring various forms of non-human entities — be they animals, aliens, or machines — as active protagonists. This makes SF an ideal genre for exploring literary representations of non-human characters' behaviors.

Identifying and examining character behaviors remains unexplored in both literary studies and text-as-data research. Emulating human reading involves complex cognitive endeavors, including various Natural Language Processing (NLP) tasks like coreference resolution and syntactic structure analysis. Mechanically classifying figurative expressions in literary texts is challenging, but well-designed computational approaches to textual interpretation can lead to new and insightful readings.

We extracted and analyzed the behaviors and cognitive processes of 'artificial beings' in recent Korean SF short stories, using annotations from five human experts and two types of Large Language Models(LLMs). Then we analyzed the agreement rates according to two different calculation methods. We paid special attention to the unique characteristics of the Korean language, which, unlike English, employs a wide variety of endings and auxiliary predicates, often making it impossible to judge the intention and usage by solely looking at a verb's grammatical form. That is, our approach focused on the semantic dimension of action rather than verbs merely as a 'part of speech'.

Given the task's complexity, we utilized state-of-the-art LLMs, known for their proficiency in grasping context and adapting to new tasks. Results show that LLMs can achieve high agreement with human annotators in label assignment for correctly identified spans, demonstrating potential in analyzing intricate literary contexts. However, differences in text span selection highlight ongoing challenges in AI's processing of narrative structures and character identification.

2 Related Works

Character behavior analysis has long held a significant position in traditional literary studies, and with the recent emergence of new approaches integrating computational methods in digital human-

Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, pages 339–351
November 16, 2024.

ities, its importance and scope of research have further expanded (Moretti, 2013; Jockers, 2013). A pivotal development in this field was the creation of BookNLP[1], a tool for extracting characters and annotating their attributes from literary texts spanning about 200 years. This annotated dataset has sparked various computational literary studies (Bamman et al., 2019; Sims et al., 2019; Bamman et al., 2020; Soni et al., 2023; Vishnubhotla et al., 2023). Representing this trend, Piper (2024) analyzed the physical actions of characters in English novels to explore how characters' agency is expressed in literary works.

Concurrently, the NLP field has shown increasing interest in utilizing LLMs for data annotation tasks (Bansal and Sharma, 2023; Ding et al., 2023; He et al., 2024; Alizadeh et al., 2024). This approach is particularly valuable in literary settings where traditional NLP tools struggle. For instance, Hicke and Mimno (2024) have explored using LLMs' for coreference annotation in literary texts. However, most studies have focused on preconstructed, extensive literary corpora, with less attention to specific genres or nuanced analyses.

In Korean science fiction, the focus of this study, character studies have been diverse but primarily qualitative (Yoon, 2022; Hong, 2023; Oh, 2023; Lee, 2023). While these studies offer valuable insights into character development and themes, quantitative methodologies or diachronic analyses of specific character types remain largely unexplored.

3 Artificial Being Behavior Dataset and Methodology

3.1 Korean Science Fiction Text

We selected 11 contemporary Korean SF short stories for analysis, annotating full texts — instead of excerpts from a larger set of works — of all 11 stories to ensure each story's overall theme and the determining characteristics are sufficiently reflected in the data. We investigated all 30 winners of the first to sixth Korean Science Fiction Award (2016-2022), which is currently the most prestigious SF award in Korea, and and identified 11 works featuring 'artificial beings' as main characters. Detailed information about the 11 stories, including their titles, publication years, lengths, and the names of the artificial being characters, is in Appendix A.

[1]https://github.com/booknlp/booknlp

The Korean Science Fiction Award serves as a pertinent object of study in exploring the 'SF boom' that swept the Korean literary scene in the late 2010s. Unlike the gradual and robust development of SF in Anglo-American contexts, the Korean literary scene struggled to sustain interest in the genre for decades. Until the 2000s, even major awards aimed at discovering new genre writers often lost momentum after just two or three years. However, this award, launched in 2016 with the slogan "The only domestic SF newcomer literary award, newly born after 10 years," has prospered, introducing writers who have expanded beyond SF into the broader Korean literary field. It has become an important turning point in Korean SF literature's evolution and a barometer for contemporary scientific and technological trends. The prevalence of AI and robots-themed works requires our particular attention, for they offer critical insights into perceptions and expectations of artificial beings in modern society.

'Artificial beings' here refer to artificially created intelligence or its implemented entity, excluding extraterrestrial life forms or animals, even if depicted as anthropomorphized non-human beings, as well as human to cyborg transformations where the intelligence was not artificially created. Artificial beings in the stories are mainly artificial intelligence, robots, or androids, with varying attributes and behavioral patterns. They exhibit characteristics that parallel human mind and behavior while simultaneously exhibiting unique behaviors and cognitive processes that distinguish their capacity from humans (e.g., entering the cloud, displaying a winking emoticon on the screen, etc.). A thorough categorization and analysis of the vocabulary depicting their behaviors helps explore human-machine boundaries and address ontological questions about future technological societies.

3.2 Data Model Design: Preliminary Experiments

As preliminary experiments before establishing the design of annotation-based research, we conducted several tests to examine the artificial being characters' behaviors from a lexical perspective. We explored the possibility of automating this process using Python Korean morphological analyzer tools commonly used in the NLP field.

This process revealed that extracting the behaviors of artificial beings from stories without losing their meaning is a delicate task of considerable diffi-

Verb Morpheme	Conjugation Examples
돌리다 dollida	화제를 **돌리다** (change the subject), 숨을 **돌리다** (catch one's breath), 시선을 **돌리다** (avert one's gaze), 마음을 **돌리다** (change one's mind), 세탁기를 **돌리다** (run the washing machine), 문고리를 **돌리다** (turn the doorknob)
보다 boda	바라**보다** (look at), 생각해 **보다** (think about), 장을 **보다** (go grocery shopping), 떠**보다** (test the waters), 잘못이라고 **보다** (consider it a mistake), 피를 **보다** (suffer harm)
하다 hada	이야기를 **하다** (have a conversation), 준비를 **하다** (prepare), 각오를 **하다** (be determined), 인사**하다** (greet), 후회**하다** (regret), 목도리를 **하다** (wear a scarf)

Table 1: Examples of Korean phrases where the same verb root is used but has completely different meanings in context.

Dataset	Description	Count
Number of sentences	11 short stories	7,289
Human annotation	5 annotators	9,515
LLM annotation	2 models * 2 versions each	8,575

Table 2: Overview of the dataset.

culty. Table 1 shows examples of verbs used in this paper's Korean SF short stories corpus that have the same morpheme but completely different meanings. In Korean, it is very common for the same verb form to exhibit semantic diversity depending on the object that the verb governs or the verb's conjugation pattern. For this reason, it is very difficult to accurately grasp what action a word refers to in the stories using morphological analyzer tools that isolate only the smallest units of meaning. As a result, even if the extracted verbs are categorized, the accuracy is very low. The task's purpose of extracting only the actions of specific characters in the narrative, coupled with the nature of literary texts where meaning changes significantly depending on the context, further complicates the analysis. Given these factors, we concluded that accurate analysis is difficult with existing NLP tools and designed the annotation work described below.

3.3 Human annotation

Label design and tools: We designed an annotation task where human annotators read all 11 stories from beginning to end, as they 1) mark lexical spans that represent the actions and cognitive processes of artificial being characters, and 2) attach labels to categories they believe these words belong to. While this method is time-consuming and challenging, it allows for a comprehensive understanding of character behavior patterns without missing the uncertainties and ambiguities that arise in the process of reading fiction. Crucially, this high-context dataset can serve as a foundational resource for training and evaluating LLMs, potentially leading to the development of more sophisticated AI research tools capable of nuanced literary interpretation.

The labels were primarily based on the word supersense tagger (Ciaramita and Altun, 2006) utilized in BookNLP, a tool frequently used in character behavior research. However, as the categories in previous studies were mainly composed of words used to describe human behavior, we redesigned the labels to better reflect the specificity of artificial being characters. Finally, we established 7 labels and 1 Miscellaneous category (to be used when a word is judged not to belong to any other category):

- ***Communication.*** The exchange of information or ideas between characters.

  가쁜 숨을 고르고 서 있는 노인에게 안드로이드가 [말한다says].
  나는 어머니의 다음 [말을 기다렸다waited for her next words].

- ***Sensory act.*** The action or process involving the use of sensory systems, including sensory-based interaction with the environment.

  안드로이드는 2층 바닥에서 올라오는 입김을 [감지하고는sensed] 에스컬레이터를 힘겹게 걸어 올라갔다.
  그리고 기계 팔을 돌려 에이브를 [바라보았다gazed at].

- ***Motion.*** Physical movement or change in position, including static states.

  [간식도 만들어야prepare snacks] 하고 [장도 봐야do grocery shopping] 하며 화장실 변기도 [닦아야clean] 한다.
  나는 경찰의 안내에 따라 법정에 [들어섰다entered].

- ***Body change.*** Fundamental alterations in the physical or mental state of a character.

  아이들이 주는 간식을 거절하지 못하고 먹다가 [고장이 났던has been broken] · · · 몸이 부서지면서도 나를 지켜주던 그 로봇은 이제 없는 거야.
  원래 24시간 깨어 있어야 하는 루트는 최소한의 감지 시스템만을 켜둔 채로 [절전모드에 들어갔다went into power-saving mode].

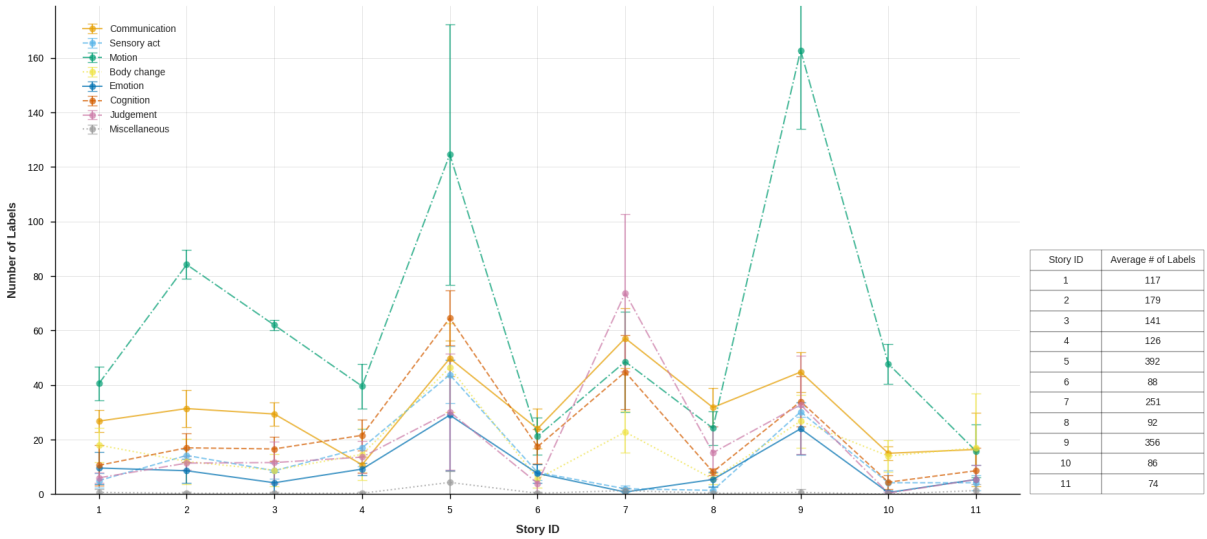
- ***Emotion.*** Subjective feelings or affective states experienced by a character.

  영혼 없이 태어난 아이들을 버리지도 못하면서 그들의 신체 기능이 정지될 때마다 [괴로워했고suffered], 그러면서도 계속해서 [희망을 품고hold on to hope] 다음 번 태아를 배양했다.
  어머니를 처음 만났을 때, 나는 [꿈꾸는 기분이었다felt like (I) was in a dream].

- ***Cognition.*** Mental processes involved in acquiring knowledge and understanding.

  그것이 라디오에서 들었던 총이라는 것을 슬라이드가 당겨지는 순간 [깨달았다realized].
  안드로이드는 지구로부터 점점 멀어지기 시작해 ... 하나의 촛불처럼 보이는 우주선의 모습을 [상상한다imagines].

- ***Judgement.*** The process of forming opinions, making decisions, or drawing conclusions.



Story ID	Average # of Labels
1	117
2	179
3	141
4	126
5	392
6	88
7	251
8	92
9	356
10	86
11	74

Figure 1: Distribution of labels per story. The numbers from 5 annotators were averaged, and the standard deviation is indicated by shading. There is a variance in the average total number of labels per work, which is shown in the table on the right.

> 따라 갈 지표가 사라지자 이 큰 건물 안을 돌아봐야 할지, 밖으로 나가야 할지 [판단하지 못했다couldn't decide].
> 그러니 자신이 인간을 도울 수 있는 더 큰 힘을 가지고 있다고 [생각하는데considers] '로봇일 뿐'이라니?

Annotators: The annotators (n=5)[2] all experienced in Korean literature, worked independently and did not discuss with each other or change their annotations to match others' annotations to ensure unbiased results. Internal consistency within each annotator's work was prioritized over attempting to establish universally "correct" answers. This approach acknowledges that there may not be a single, objectively correct label for each expression in the story. Instead, annotators were instructed to assign labels to the closest category based on their interpretation of the characteristics, intentions, and context of the artificial being characters portrayed in the narrative. To take into account the unique characteristics of Korean vocabulary, annotators marked minimum spans capturing complete meanings of actions and cognitive processes, often spanning multiple words.

Annotators were provided with full texts of the stories via our annotation tool, as opposed to being given one sentence at a time. They were tasked to span-mark all behavioral vocabulary of artificial being characters and attach single-choice labels from a drop-down menu format. Consequently, the spans of behavioral vocabulary entities marked by each annotator differed, even before considering label differences. The task was far from being an obvious or objective one, yielding many interesting cases with uncertainties or ambiguities.

The number of labels varied depending on the different prominence and characteristics of artificial beings in the stories. Figure 1 shows the average number of labels from five annotators, with standard deviation indicated by error bars. The deviation range illustrates that there was considerable inter-annotator variability, and the dominant labels also varied depending on the narrative.

For instance, Story 5 ("Five Stages of Independence") and 7 ("The Last Judgment") present an interesting contrast. Story 5 exhibits a high proportion of 'Cognition' labels, reflecting a narrative structure that deeply explores the artificial being character's inner world. Whereas Story 7, prominently features 'Judgment'-related vocabulary and thus the label distribution, for it contains substantial content revealing the AI judge's beliefs and decision-making criteria. Both Story 5 and 9 ("Samsara") show a predominance of 'Motion' labels, Story 5's greater variance compared to Story 9 suggests differing levels of judgment clarity within this category. It is also noteworthy that 'Communication' labels frequently rank high in prominence, which indicates that many SF works portray artificial beings as capable of linguistic interaction with humans. A detailed breakdown of the total number

[2]Among the five annotators, four were females and one was male. Their ages ranged from the 20s to 30s, and all were native Korean speakers. Two of the annotators held doctoral degrees in literature.

of annotations assigned by each individual annotator across the 11 stories is provided in Appendix B. This data enables a comparative analysis of annotators' tendencies in identifying and marking relevant spans, as reflected by the total quantity of labels assigned.

3.4 LLM annotation

The high-context dataset built through human annotation contributes to comparing and improving LLMs' literary comprehension. We created prompts similar to human annotation guidelines, instructing LLMs to mark artificial beings' actions and cognitive processes in the stories and attach appropriate labels.

We utilized Claude 3.5 Sonnet and GPT-4o, both state-of-the-art models known for their excellent multilingual support, including Korean. We conducted zero-shot (providing only guidelines and receiving output in a predetermined format) and few-shot (providing 7 examples) approaches, creating a total of 4 datasets. Our experiments with various lengths[3] showed that smaller text units tended to increase the number of labels, but when units became too small, context was lost, resulting in inaccurate character identification and reduced accuracy. We judged that about 5,000 characters was the most appropriate parameter. Stories were divided into 3-6 units, and we integrated the outputted JSON annotation files by work for analysis.

The annotation task involves multiple stages that humans perform intuitively. From a language model's perspective, however, these stages are distinct and sequential:

1. Named Entity Recognition (NER): Distinguishing characters and identifying whether a specific noun/pronoun refers to an artificial being within a story.
2. Verb Span Identification: Accurately identifying the span of Korean verbs that denote actions in the text.
3. Verb Categorization: Categorizing verbs based on contextual meaning.

Comparing and analyzing the annotation data generated through this multi-layered process reveal LLMs' strengths or weaknesses in interpreting complex narrative structures.

To evaluate LLMs' performance of literary comprehension tasks, we conducted a comprehensive analysis comparing our human-annotated datasets with LLM outputs. Examination of the matching rates between LLM predictions and human annotations was followed by a detailed investigation of the corresponding labels and text segments. Our approach assessed the quality of LLMs' literary analysis within broader contexts, paying particular attention to cases of high agreement and significant discrepancies.

To provide a comprehensive overview of the annotation process, Appendix C includes a table showing the total number of annotations per story by each LLM model, which can be compared with the total number of human annotations.

4 Results and Analysis

4.1 Inter-annotator Agreement Score

Our annotation process consisted of two main steps: 1) precisely marking lexical spans representing artificial beings' actions in the SF texts, and 2) assigning appropriate labels to the marked spans. To assess the reliability of this process, we evaluated inter-annotator agreement using two distinct methods.

The left heatmap in Figure 2 illustrates ***sentence-level label agreement*** based on Jaccard similarity. This approach matches sentences containing labeled text with the eight labels assigned by each annotator, disregarding detailed text spans. Pairwise agreement ranged from 7.5% to 67.2%, with an overall mean of 42.6%. This wide range coupled with a moderate average indicates significant variability in labeling consistency across annotator pairs. It suggests substantial subjective differences in interpreting artificial beings' actions in literary contexts, while still maintaining a moderate level of overall consensus.

The right graph in Figure 2 depicts ***span-based fuzzy label agreement***, a more refined measure. While the average agreement score (53.3%) was higher, it showed greater variability across annotator pairs and works. This analysis scored span agreement as complete match, partial match, or no match, considering both the overlap of annotators' marked text spans and label consistency. Final scores were calculated by verifying label matches for complete and partial span matches, applying appropriate weightings (Equation 1).

Agreement rates calculated by this method ranged widely from 14.0% to 76.2% across an-

[3]Due to API token limitations, text in segments of about 5,000-6,000 Korean characters were the maximum amount that could be processed at a time without output annotations being cut off.

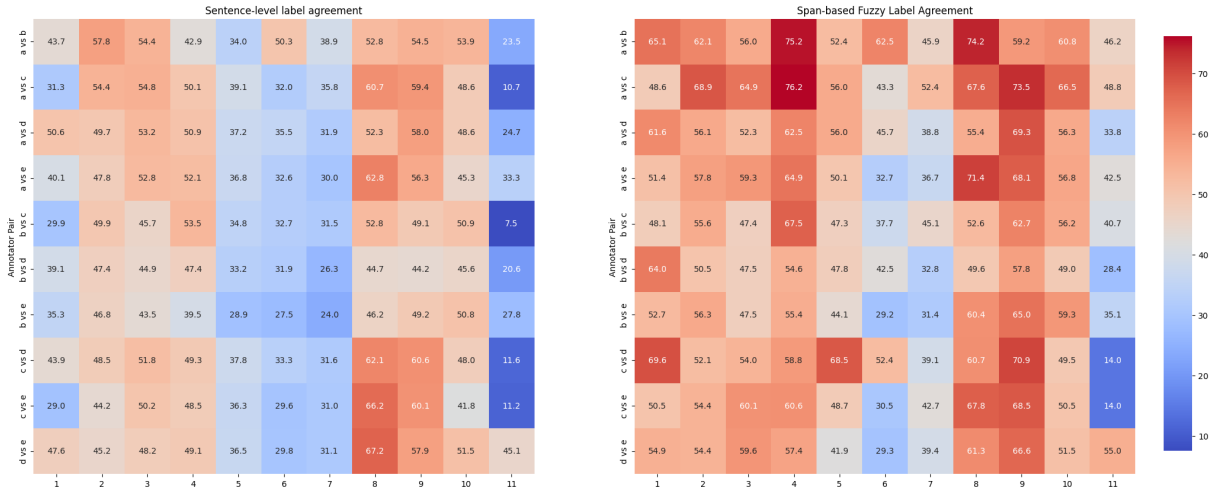


Figure 2: Results of Inter-annotator Agreement Analysis. Left: Heatmap of sentence-level label agreement based on Jaccard similarity. Right: Span-based fuzzy label agreement scores.

notator pairs, reflecting the task's difficulty and subjectivity. Certain annotator pairs (e.g., a vs. b, a vs. c) consistently showed high agreement[4], while others (e.g., d vs. e) demonstrated relatively low agreement.

$$\text{Score} = \left(\frac{M_{pe} \times 1 + M_{pa} \times 0.5}{M_{pe} + M_{pa} + M_{non}} \right) \times 100 \qquad (1)$$

M_{pe}: number of perfect matches
M_{pa}: number of partial matches
M_{non}: number of non-matches

In most of the texts, the span-based fuzzy method demonstrates generally higher agreement scores compared to the sentence-level method, although it exhibits a wider distribution range. This suggests that the span-based method more effectively captures subtle differences between annotators by accounting for detailed textual elements. The distribution of agreement scores across annotator pairs for each story is provided in the Appendix D.

4.2 LLM Annotation Evaluation

The four types of LLMs labeled 1,000-2,000 data points for each of the 11 stories, similar to the distribution of human annotation data (averaging 1,903 annotations across 5 annotators). To assess how well LLMs understood and classified the actions of artificial beings in the stories without additional training (or to what extent they could match human comprehension), we applied a span-based fuzzy label agreement method similar to the one used in earlier evaluation. We aligned span-marked words and labels sentence by sentence, comparing them with five human annotators' responses. A Span Match was recorded if there was any overlap (2 or more Korean characters) in the text span. In our analysis, we considered a match to occur when an LLM's annotation (either span or label) aligned with at least one human annotator. This approach was chosen for both span and label agreement to preserve the diversity of human annotations, which was a key focus of our study. We deliberately avoided creating a "gold standard" based on majority agreement among human annotators, as this would have diluted the individual perspectives we aimed to capture. This method allows us to evaluate LLM performance while acknowledging the inherent variability in human interpretations of textual content. The proportion of entities with matching spans be-

Model	Total Annotations	Span Matches	Span Unmatches
Claude 3.5 Sonnet (zero shot)	1613	1045 **(64.8%)**	568 **(35.2%)**
Claude 3.5 Sonnet (few shot)	1270	878 **(69.1%)**	392 **(30.9%)**
GPT-4o (zero shot)	2917	1225 **(42%)**	1692 **(58%)**
GPT-4o (few shot)	2775	1305 **(47%)**	1470 **(53%)**

Table 3: Summary of Model Annotations and Span Matching.

[4]The annotator pairs showing high agreement had relevant academic backgrounds: annotator a is the first author of this paper and a master's student in digital humanities, while annotators b and c hold doctoral degrees in literature.

tween LLMs and humans was approximately 60% for Claude 3.5 Sonnet and 40% for GPT-4o. Both models showed fewer mismatched labels in few-shot scenarios compared to zero-shot.

Figure 3: Span Match (line graph) and Label Agreement (shaded area) Rates between 4 different LLMs and Human Annotators across stories, arranged in a clockwise direction starting with Claude-Zeroshot, followed by GPT-Zeroshot, GPT-Fewshot, and ending with Claude-Fewshot.

Figure 3 illustrates the percentage of annotations with matching spans to at least one human reference (line graph) and the proportion of these that also had matching labels with at least one human annotator (shaded area) for each model across stories. Notably, when LLMs correctly identified words describing artificial beings' actions, the corresponding labels matched human annotators' labels in over 65% of cases, often exceeding 80%. Specifically, Claude's model achieved over 75% label agreement in 8 out of 11 stories, while the GPT model reached this threshold in 7 out of 11 stories. This clearly demonstrates that LLMs can detect and classify the actions of artificial beings in stories at a level approaching human annotators.

However, LLMs struggled with Named Entity Recognition (NER), particularly in tracking specific characters. This difficulty likely stems from the varied and often indirect references to artificial entities in the stories' contexts. The unmatched data from stories 1, 8, and 11, which showed notably low span match rates, mostly consisted of annotations about human characters appearing early in the stories. There were also instances where LLMs misinterpreted passive verbs targeting artificial beings as their actions. These findings suggest that while LLMs excel at sentence-level action classification, they still have room for improvement in consistently tracking characters across broader contexts. Future research should focus on enhancing contextual understanding and long-term dependency processing to overcome these limitations.

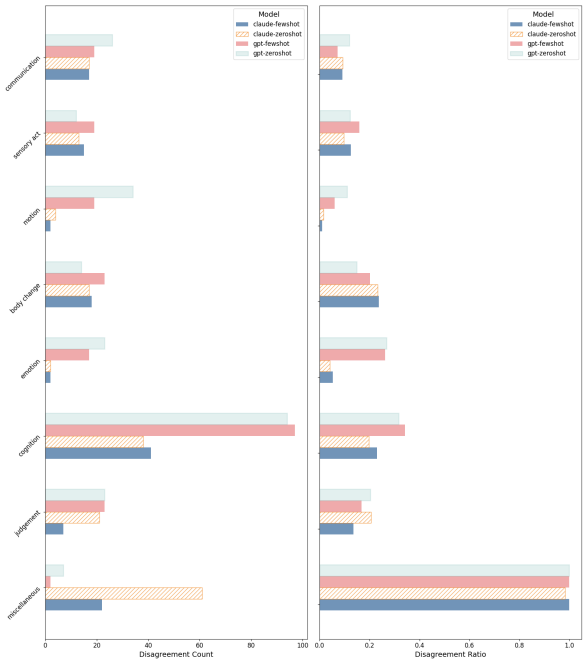


Figure 4: Distribution of label disagreements between LLMs and human annotators for matched spans. The left graph shows the number of label disagreements, while the right graph represents the ratio of label disagreements to the total number of matched spans.

Figure 4 shows the distribution of cases where LLMs identified the matched span but assigned different labels compared to human annotators. The most striking discrepancy occurs in the Cognition category across all models, with GPT models showing an even higher rate of disagreement. Upon closer examination of the text, this divergence is largely centered on machine-related terminology such as 'record', 'upload', 'transfer data', 'change configuration', and 'execute facial recognition'. Human annotators tended to interpret these actions contextually as Motion or Sensory acts, applying a more anthropomorphic perspective to artificial entities. In contrast, LLMs consistently classified these as Cognition, viewing them primarily as computational processes. This discrepancy highlights an intriguing difference in how humans and AI interpret the cognitive processes of artificial beings in literature. Additionally, Claude models, especially Claude Zeroshot, show a notably higher use of the Miscellaneous label, suggesting a more cautious approach to ambiguous actions. Interestingly, GPT models more frequently applied

the Emotion label, while Claude models showed minimal disagreement in this category, indicating varying approaches to emotion recognition between the two model types. These patterns reveal distinct strengths and limitations in how AI models interpret artificial beings' actions, particularly in complex cognitive processes and ambiguous behaviors.

When examining the ratio of label disagreements to the total number of matched span annotations, the overall trends remain similar. However, it's important to note that human annotators were strongly discouraged from using the 'miscellaneous' label (this instruction was also included in the LLM prompts but was not fully adhered to). Consequently, all instances of the 'miscellaneous' label are counted as 'disagreements' in this ratio calculation.

The complete distribution of all labels matching human annotations across different works can be found in Appendix E.

5 Discussion

We investigated the behavioral patterns of artificial beings in contemporary Korean SF short stories. By leveraging both human expertise and large language models (LLMs), we conducted a comprehensive analysis that revealed the complexity and subjectivity involved in categorizing the actions and cognitive processes of artificial characters in literary contexts.

Our novel approach deconstructs the multilayered cognitive processes inherent in literary comprehension, pioneering a literature-specific framework for evaluating inter-annotator agreement. The higher average agreement score (53.3%) obtained through the span-based fuzzy method, compared to the sentence-level method (42.6%), suggests that considering detailed textual elements better captures nuanced differences and potential consensus in annotators' interpretations. For non-literary text annotation tasks aimed at "predictive accuracy" or "generalizability," an inter-annotator agreement rate around 50% would typically be deemed insufficient. However, given the intricate nature of reading literature, which values diverse interpretations, this agreement rate proves to be a significant finding. The possibility of varied interpretations for the same character demonstrates the depth and richness of literary texts and confirms the active meaning-making processes of readers. The metric we developed presents both unique opportunities and methodological challenges for computational approaches in studying representations of characters in fiction.

LLMs demonstrated promising results in matching human annotations, particularly excelling in sentence-level action classification. Claude 3.5 Sonnet and GPT-4o achieved impressive span match rates of approximately 60% and 40% respectively, with label agreement rates frequently exceeding 80% for correctly identified spans. These unprecedented results indicate that LLMs can analyze and categorize artificial beings' actions in literature to nearly human-level accuracy. Notably, few-shot learning approach yielded minimal performance improvements, suggesting that LLMs may already possess specialized capabilities for such high-context tasks, rendering additional 'examples' less critical. However, the models' struggles with Named Entity Recognition (NER) and character tracking across broader contexts highlight areas for improvement in AI's literary comprehension abilities. A particularly significant finding was LLMs' ability to distinguish 'actions' as semantic units rather than merely grammatical verbs (POS). LLMs generally marked necessary Korean objects or auxiliary predicates correctly, enabling clear distinction of artificial beings' behaviors and highlighting the potential in complex literary annotation tasks.

This research provides innovative insights into artificial beings in Korean SF stories, potentially stimulating further studies in this emerging field. Furthermore, our interdisciplinary approach enhances understanding of literary texts, as well as offers valuable insights for developing more sophisticated NLP models capable of grasping contextual nuances and long-term narrative dependencies. These findings present new possibilities for the convergence of literary studies and AI technology.

6 Limitations and Conclusion

This study enhanced annotation robustness by utilizing both human annotators and state-of-the-art LLMs to analyze behavioral patterns of artificial beings in Korean SF works. To ensure methodological transparency, we provided detailed approaches for inter-annotator agreement and LLM performance evaluation. However, identifying and categorizing actions in literary texts remains inherently subjective and is inevitably situated within the contemporary Korean SF contexts and culturally specific understandings of artificial intelligence. In terms

of data, our focus on 11 Korean SF short stories by emerging authors who won a specific literary award potentially restricts the generalizability of our findings. Given the recent proliferation of Korean SF works featuring various artificial beings beyond these stories, expanding the research to include a broader range of contemporary Korean SF literature could have provided more comprehensive insights.

On the technical side, LLMs exhibited limitations potentially stemming from their training data and prompting strategies. The text segmentation necessitated by technical constraints may have affected the models' grasp of overall narrative context. Another limitation is the omission of a detailed linguistic analysis of marked lexical ranges, which was excluded due to space constraints.

Even with these limitations, our findings highlight the complex interplay between literary expressions and technological capabilities, revealing both the potential and limitations of current AI technologies in analyzing nuanced literary contexts. Discrepancies between human annotators and LLMs in interpreting artificial beings' actions underscore the subjective nature of literary analysis and the challenges in AI's comprehension of contextual nuances. However, LLMs's promising performance in sentence-level action classification suggests a path forward for integrating AI tools into literary studies. This research contributes to the ongoing dialogue between science fiction and AI development, offering insights for future studies in both fields.

Acknowledgements

We acknowledge that, due to copyright restrictions on the stories analyzed in this study, we are unable to publicly share the complete dataset. We have endeavored to ensure the utmost transparency in presenting our methodology and results, consistent with these legal constraints.

References

Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Mohammadmasiha Zahedivafa, Juan Diego Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. 2024. Open-source llms for text annotation: A practical guide for model setting and fine-tuning. *Preprint*, arXiv:2307.02179.

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in English literature. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 44–54, Marseille, France. European Language Resources Association.

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

Parikshit Bansal and Amit Sharma. 2023. Large language models as annotators: Enhancing generalization of nlp models at minimal cost. *Preprint*, arXiv:2306.15766.

Massimiliano Ciaramita and Yasemin Altun. 2006. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 594–602, Sydney, Australia. Association for Computational Linguistics.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Shafiq Joty, Boyang Li, and Lidong Bing. 2023. Is gpt-3 a good data annotator? *Preprint*, arXiv:2212.10450.

Jens Eder, Fotis Jannidis, and Ralf Schneider. 2010. *Characters in Fictional Worlds*. De Gruyter.

John Frow. 2014. *Character and Person*. Oxford University Press.

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024. Annollm: Making large language models to be better crowdsourced annotators. *Preprint*, arXiv:2303.16854.

Rebecca M. M. Hicke and David Mimno. 2024. [lions: 1] and [tigers: 2] and [bears: 3], oh my! literary coreference annotation with llms. *Preprint*, arXiv:2401.17922.

Deokgu Hong. 2023. Representations of scientists in contemporary korean science fiction -centered on the novels of kim cho-yeop, shim nae-ul, and jung bo-ra. *Journal of Popular Narrative*, 29(3):69–103.

Matthew L Jockers. 2013. *Macroanalysis: Digital methods and literary history*. University of Illinois Press.

Jiyong Lee. 2023. Significance and transformation of nonhuman characters in korean science fiction - focusing on science fiction content from the 2010s onwards. *The Society Of Korean Language Culture*, 82:197–225.

Franco Moretti. 2013. *Distant Reading*, volume 93. Verso.

Haein Oh. 2023. Post-body imagination in south korean sf novels - focused on 「little baby blue pill」 by jung sae-rang and 「laura」 by kim cho-yeop. *Journal of Korean Literary Criticism*, 79:71–104.

Andrew Piper. 2024. What do characters do? the embodied agency of fictional characters. *Journal of Computational Literary Studies*, 2(1).

Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary event detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3623–3634, Florence, Italy. Association for Computational Linguistics.

Sandeep Soni, Amanpreet Sihra, Elizabeth F. Evans, Matthew Wilkens, and David Bamman. 2023. Grounding characters and places in narrative texts. *Preprint*, arXiv:2305.17561.

Krishnapriya Vishnubhotla, Frank Rudzicz, Graeme Hirst, and Adam Hammond. 2023. Improving automatic quotation attribution in literary novels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 737–746, Toronto, Canada. Association for Computational Linguistics.

Aekyung Yoon. 2022. The free will and gender performativity of posthuman in korean sf novel. *International Language and Literature*, 53:81–106.

A

Table 4: List of 11 Science Fiction stories and their details.

Story ID	Title	publication year	Writer	# of syllables	# of sentences	name of the AI in the story	functions
1	피코 Pico	2017	Lee, Gunhyuk	19,089	588	Pico, Freya	Companion AI
2	TRS가 돌보고 있습니다 TRS is Providing Care	2018	Kim, Hyejin	17,961	495	TRS	Care Robot
3	마지막 로그 Last Log	2018	Oh, Jeongyeon	20,168	432	Joy	Euthanasia Assistance Android
4	라디오 장례식 Radio Funeral	2018	Kim, Sunho	16,602	454	Android	Conversation and Service Robot
5	독립의 오단계 Five Stages of Independence	2018	Lee, Ruka	46,906	1,198	I, Model Name A796, Serial Number 04-1963-59	Cyborg Android Integrated with a Human Brain
6	옛날 옛적 판교에서는 Once Upon a Time in Pangyo	2022	Kim, Kuman	22,461	564	I	In-Game AI
7	최후의 심판 The Last Judgment	2023	Han, Isol	39,208	1,027	Solomon, Solo 3.0	AI Judge
8	두 개의 세계 Two Worlds	2023	Park, Minhyeok	40,461	1,201	Root	Dome Environment Management AI
9	삼사라 Samsara	2023	Jo, Seowol	15,898	316	Sarah, Abe	Artificial Persona of a Spaceship's Main Computer
10	제니의 역 Jenny's Reversal	2023	Choi, Ia	16,403	392	Jenny	Multicultural Family Assistance AI
11	발세자르는 이 배에 올랐다 Balthazar Boarded This Ship	2023	Heo, Dallip	18,939	622	Rimey	Privately-Created AI Stored on a Server

B

Table 5: Total number of annotations per story by each human annotator (a-e).

Story ID	a	b	c	d	e	Average
1	146	98	102	127	97	114
2	186	155	206	168	160	175
3	149	158	166	123	141	147
4	121	132	143	123	107	125
5	417	299	366	441	348	374
6	104	87	93	101	49	87
7	302	218	350	228	188	257
8	91	104	100	82	89	93
9	336	318	390	369	366	356
10	88	82	96	91	74	86
11	40	62	204	43	44	79

C

Table 6: Total number of annotations per story by each LLM model.

Story ID	Claude-Zeroshot	Claude-Fewshot	GPT-Zeroshot	GPT-Fewshot	Average
1	80	151	164	103	125
2	72	64	124	101	90
3	81	85	202	138	127
4	80	88	113	98	95
5	352	241	791	499	471
6	68	64	154	249	134
7	276	141	330	543	323
8	170	55	256	267	187
9	159	168	222	253	201
10	47	46	52	53	50
11	228	167	509	421	331

D

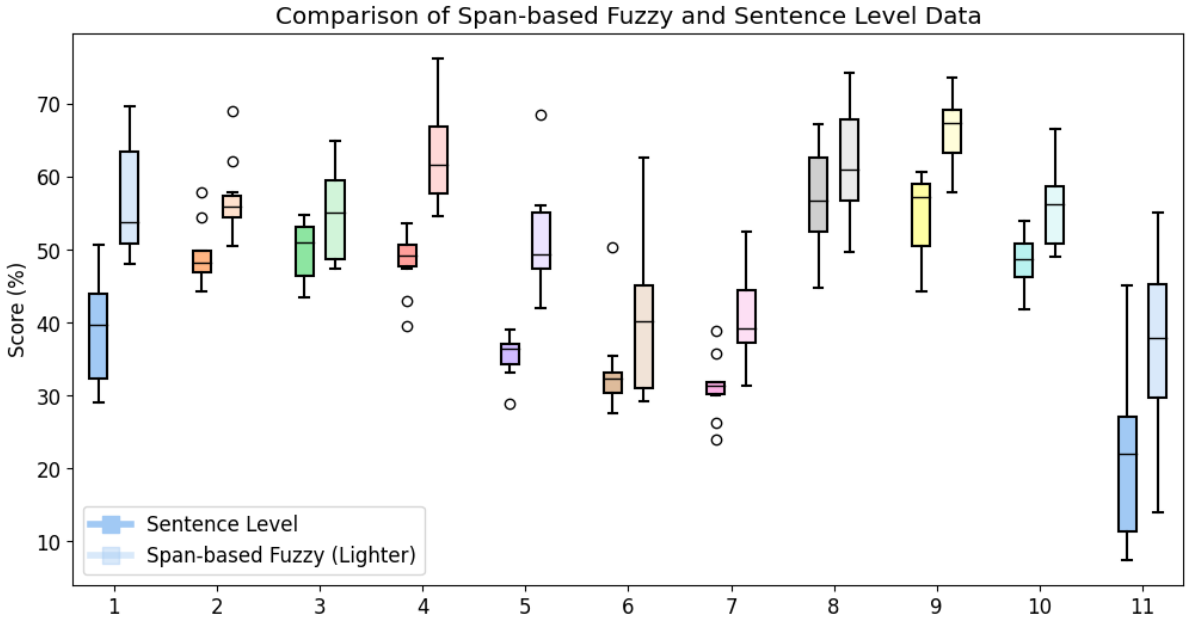
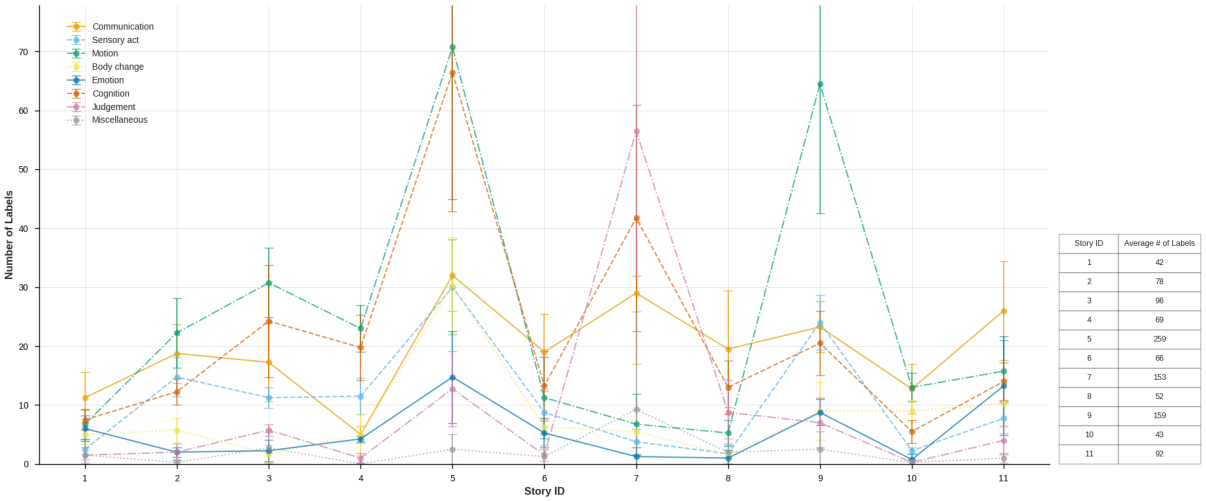


Figure 5: Distribution of Inter-annotator Agreement Scores Across Stories.

E



Story ID	Average # of Labels
1	42
2	78
3	96
4	69
5	259
6	66
7	153
8	52
9	159
10	43
11	92

Figure 6: Label distribution of LLM annotations matching human annotation coverage. Values represent the average across four different LLM models.