Making Sense of Korean Sentences: A Comprehensive Evaluation of LLMs through KoSEnd Dataset

Seunguk Yu, Kyeonghyun Kim, Jungmin Yun and **Youngbin Kim**
Chung-Ang University, Seoul, Republic of Korea
seungukyu@gmail.com, {khyun8072, cocoro357, ybkim85}@cau.ac.kr

Abstract

Although LLMs have made significant progress in various languages, there are still concerns about their effectiveness with low-resource agglutinative languages compared to languages such as English. In this study, we focused on Korean, a language known for its complex sentence endings, and evaluated LLMs on this challenging aspect. We introduce the Korean Sentence Endings (**KoSEnd**) dataset, which includes 3,000 sentences, each annotated for the naturalness of 15 sentence ending forms. These were collected from diverse sources to cover a range of contexts. We evaluated 11 LLMs to assess their understanding of Korean sentence endings, analyzing them based on parameter count and prediction consistency. Notably, we found that informing models about the possibility of missing sentence endings improved performance, highlighting the impact of explicitly considering certain linguistic features.

1 Introduction

With the continuous advancement of large language models (LLMs), they have become capable of understanding multiple languages, irrespective of the input language (Zhang et al., 2023; Huang et al., 2023). However, the data used to train these models are heavily skewed toward English, rather than being evenly distributed across various languages (Liu et al., 2024; Li et al., 2024). Consequently, LLMs may exhibit varying levels of comprehension depending on the language used, raising concerns regarding their effectiveness in understanding relatively low-resource languages (Cahyawijaya et al., 2024; Asai et al., 2024; Cahyawijaya et al., 2023).

Moreover, languages with alphabetic scripts often have advantages in tokenization since they can share some of the model's limited token capacity (Petrov et al., 2024; Limisiewicz et al., 2023), while non-alphabetic script languages often face challenges due to smaller training datasets.

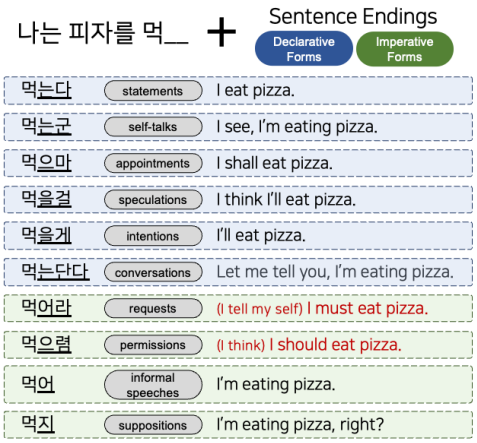

Figure 1: Impact of the Korean sentence endings on the meaning of sentences. The translated texts showed that even small differences in sentence endings can lead to significant changes in meaning.

Additionally, agglutinative languages like Korean have complex morphological structures, which further complicate tokenization and related processes (Song et al., 2024; Kaya and Tantuğ, 2024). Consequently, LLMs tend to be disproportionately advantaged in alphabetic languages compared to relatively low-resource agglutinative languages.

In this case, we focus on the Korean language with agglutinative characteristics (Sohn, 2001). In Korean, a single verb stem can be combined with various sentence endings to express different meanings such as *statements*, *perceptions*, and *exclamations* (Lee, 2005). As illustrated in Figure 1, minor changes in sentence endings can significantly affect a sentence's meaning and interpretation[1]. For example, while the blue expressions with `Declarative` endings generally convey the intended meanings, the green expressions with

[1]When using translation tools such as Google Translate or DeepL, we found that they fail to capture the nuances of Korean sentence endings accurately. To address this, we instructed the latest `gpt-4o` model to perform zero-shot translation with careful attention to the use of sentence endings.

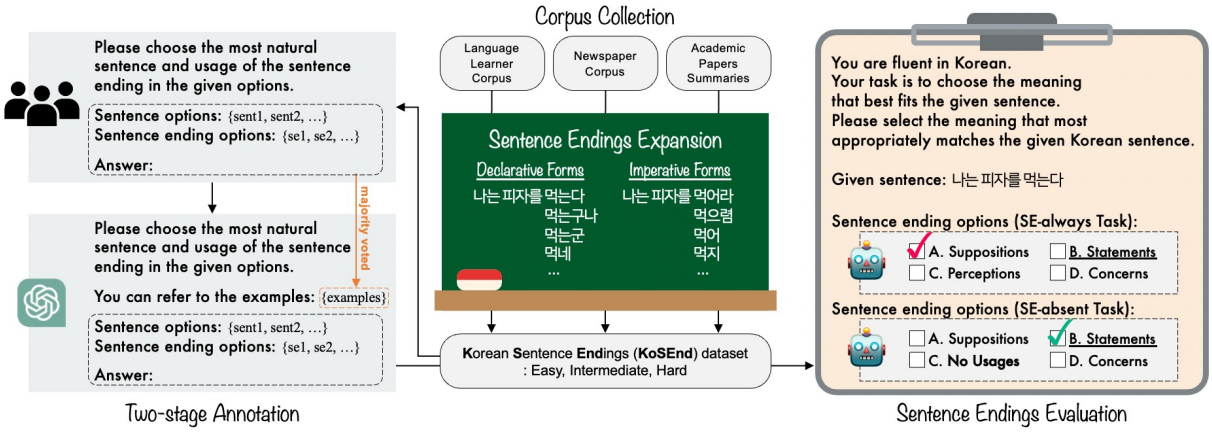

Figure 2: Process of constructing the Korean Sentence Endings (**KoSEnd**) dataset and evaluating LLMs' understanding of Korean sentence endings. Sections §3.1 and §3.2 cover the *Corpus Collection* and *Sentence Ending Expansion*, respectively. Section §3.3 describes the *Two-stage Annotation*, and these three sections constitute the process of constructing the dataset. Section §4 presents the *Sentence Ending Tasks*, where we evaluated the LLMs understanding in Korean sentence endings through the designed tasks.

`Imperative` endings can feel awkward in certain contexts[2]. This shows that sentence endings significantly impact the meaning and interpretation of a sentence, depending on the context.

Considering these perspectives, we examine the diverse usages of sentence endings and evaluate LLMs in this area. The construction of the proposed dataset and evaluation process are illustrated in Figure 2. We propose the Korean Sentence Endings (**KoSEnd**) dataset, which explores the use of sentence endings in various contexts. Each sentence was expanded to include all theoretically possible sentence endings applicable to both `Declarative` and `Imperative` forms (Lee, 2005), ensuring that the dataset captures a wide range of contextual variations. Subsequently, we conducted a two-stage annotation process to reflect the natural usage of these endings based on context.

Using the proposed dataset, we evaluate how well various LLMs understand Korean sentence endings. We then analyze the results, taking into account factors such as model parameters and the consistency of their predictions. We found that each model had a different level of understanding of Korean sentence endings, with performance improving notably when we introduced the possibility that sentence endings *might be absent*. Based on these results, and the observation that learning linguistic knowledge together contributed to improved performance on downstream tasks (Xiang et al., 2022; Ke et al., 2020; Miaschi et al., 2020), we expect models with a deeper understanding of Korean sentence endings to also perform better on general tasks[3].

The contributions of our study are as follows:

- We propose the Korean Sentence Endings (**KoSEnd**) dataset, a collection of corpora categorized by the contextual difficulty. It includes sentence ending expansion and two-stage annotation process that capture the natural usages of Korean sentence endings.
- We evaluate 11 LLMs to assess their understanding of Korean sentence endings. We compared performance by parameter count and analyzed prediction consistency across option orders, identifying models with robust comprehension of Korean sentence endings.
- We further explore how informing models about the potential absence of sentence endings affected their performance. Across all models, performance improved with this consideration, suggesting that LLMs better grasp Korean sentence endings when considering this linguistic feature.

2 Related Work

2.1 NLP Benchmarks

Numerous benchmarks have been developed to evaluate the reasoning abilities of language models.

[2]In Figure 1, some sentences may sound awkward as certain `Imperative` endings were used with the subject '*I*.' These sentences are highlighted in red within the figure.

[3]We will publicly release the proposed dataset to encourage further research. `https://github.com/seungukyu/KoSEnd`

Sentence Endings in `Declarative` Forms	Usages	Sentence Examples
(1) {다, 는다, ㄴ다}	*statements*, *exclamations*, *questions*	보통 마음대로 좋은 선물을 가지고 간다 (They usually bring a good gift as they please.)
(2) {구나, 는구나}	*perceptions*, *suppositions*	결말에 주인공이 국가를 위해 목숨을 바치는구나 (Ah, in the end, the main character sacrifices their life for the country.)
(3) {군, 는군}	*self-talks*, *perceptions*	얘기를 많이 하니까 시간이 빨리 가는군 (Time sure flies when you talk a lot.)
(4) {네}	*perceptions*, *exclamations*, *self-talks*, *questions*	그래서 우리는 학교 근처 편의점에 가네 (So, we ended up going to the convenience store near the school.)
(5) {으마, 마}	*appointments*, *intentions*	학생들이 잘 공부하도록 언제나 최선을 다하마 (I will always do my best so that the students can study well.)
(6) {을걸, 걸}	*speculations*	벌써 1년이나 지났는데 지금 그날을 생각하면 아직도 행복한 느낌이 들걸 (It's already been a year, but when I think about that day, I still feel happy.)
(7) {을게, ㄹ게, 을래, 래}	*(expressions of) intentions*, *questions*	한국 문화에 관심이 있을래 (I think I might be interested in Korean culture.) (Would you be interested in Korean culture?)
(8) {을라, ㄹ라}	*concerns*	많은 사람들이 물가가 너무 올라가서 걱정을 할라 (Many people are worried because the cost of living has gone up too much.)
(9) {는단다, ㄴ단다, 단다, 란다}	*conversations*	아주 힘들었지만 예쁜 경치를 봐서 기분이 좋단다 (It was really tough, but I feel good because I got to see the beautiful scenery.)
Sentence Endings in `Imperative` Forms	Usages	Sentence Examples
(10) {아라, 어라, 여라}	*commands*, *requests*, *permissions*, *exclamations*	한국에서 간 장소에서 홍대를 소개하여라 (Introduce *Hongdae* among the places you visited in Korea.)
(11) {으려무나, 려무나, 으렴, 렴}	*permissions*, *commands*	돈을 벌고 나서 같이 여행하렴 (After you earn some money, let's go on a trip together.)
(12) {소서}	*hopes*	장애인에게 많은 관심을 가지소서 (Please show a lot of interest in people with disabilities.)
(13) {어}	*informal speeches*	게다가 이 일을 하면 스트레스가 많어 (Besides, doing this job causes a lot of stress.)
(14) {아}	*informal speeches*, *surprises*	명동은 사람이 많아 (*Myeongdong* is crowded with people.)
(15) {지}	*questions of confirmation*, *obvious statements*, *suppositions*, *gentleness*, *intentions*, *regrets*	나는 인생에 대한 새로운 생각이 생기지 (I've come to have new thoughts about life.)

Table 1: All forms of sentence endings used in this study, along with their usages and examples[1] (Lee, 2005). The top nine sentence ending forms are categorized as `Declarative`, while the bottom six are `Imperative`. Each ending is further grouped by usage, with the underlined Korean expressions in the 'Sentence Examples' highlighting the specific endings used in each example.

A notable research is SQuAD, which involves collecting question pairs for reading comprehension, along with its adaptations (Rajpurkar et al., 2018, 2016). Afterward, GLUE emerged with a broad set of language understanding tasks such as QA and natural language inference (Wang et al., 2018). Subsequently, a method for evaluating the multi-task performance of language models has been introduced, reflecting the ongoing research aimed at assessing model performance from multiple perspectives (Bai et al., 2024; Hendrycks et al., 2021).

Recently, several Korean natural language inference datasets have been developed using sources such as Wikipedia and news articles (Park et al., 2021; Ham et al., 2020). Research has progressed in utilizing linguistic features to understand sentence relationships (Jang et al., 2022; Lim et al., 2019) and measuring national alignment, particularly with the advanced LLMs (Lee et al., 2024). In this study, we construct an evaluation dataset grounded in the linguistic characteristics of the Korean language and conduct a comparative assessment of various LLMs.

2.2 Commonsense Knowledge Evaluation

Research on analytic languages, such as English, often struggles when applied to agglutinative languages with complex word formation. Recent studies reveal that LLMs face these challenges, highlighting the need for models that effectively address linguistic diversity (Maxutov et al., 2024; Weissweiler et al., 2023). In response, benchmarks have been introduced for natural language understanding tasks in agglutinative languages, including Japanese, Indonesian, and Kazakh (Kurihara et al., 2022; Wilie et al., 2020).

Specifically, several datasets have been designed to evaluate the bias and dialogue comprehension of LLMs to assess their ability to understand nuanced semantic information in Korean (Jang et al., 2024; Jin et al., 2024). Nevertheless, performance comparisons from cultural and regional sources have noticed that LLMs encounter challenges in commonsense reasoning within a Korean-specific

context (Son et al., 2024a,b; Kim et al., 2024a).

2.3 Linguistic Knowledge Evaluation

Recent works have evaluated LLMs handling of morphological complexities and structural challenges in low-resource and agglutinative languages (Nasution and Onan, 2024; Leong et al., 2023). In Korean, studies have specifically examined the linguistic knowledge, including their understanding of grammatical structures and language proficiency (Seo et al., 2024). For instance, studies analyzing linguistic factors, such as case markers and pragmatic competence, offer deeper insights into LLM performance in Korean (Hwang et al., 2024; Kim et al., 2024b; Park et al., 2024b).

3 KoSEnd: Dataset Construction

3.1 Corpus Collection

Recognizing that Korean sentence endings can vary depending on the context, we collected three corpora, each categorized by the difficulty level: `Easy` from the language learner corpus, `Intermediate` from the newspaper corpus, and `Hard` from the academic papers summaries. The details regarding each corpus are provided in Appendix A.1.

3.2 Sentence Ending Expansion

We expanded the original sentences from the corpora with diverse sentence endings. We focused on the `Declarative` and `Imperative` forms, which were categorized into nine and six types, as shown in Table 1. In Korean, sentence endings can be categorized into `Declarative`, `Interrogative`, and `Imperative` forms (Lee, 2005). For the `Interrogative` form, the presence of a question mark makes the use of specific endings straightforward. Therefore, we only focused on the endings used in `Declarative` and `Imperative` forms, which are more distinct and challenging.

The choice of appropriate sentence ending can be subjective, varying among readers based on their interpretation of context and communicative intent[4]. Therefore, we conducted an annotation process to ensure the natural usages of sentence endings after expanding all sentences using a total of fifteen different sentence endings for `Declarative` and `Imperative` forms. The explanations of some examples in Table 1 are explained in Appendix A.2.

[4]Examples of unnatural sentence ending usage are provided in Appendix A.2, depending on the context.

Difficulty	Declarative		Imperative	
	Sentences	Usages	Sentences	Usages
`Easy`	0.748	**0.634**	0.733	**0.644**
`Intermediate`	**0.755**	0.453	**0.857**	0.544
`Hard`	0.556	0.300	0.594	0.417

Table 2: Krippendorff's α (Hayes and Krippendorff, 2007) based on the human annotation results for each difficulty level. We found that easier levels resulted in higher scores and greater consistency among annotators, while scores decreased as difficulty increased, indicating more variation in the annotations.

Difficulty	Declarative		Imperative	
	Sentences	Usages	Sentences	Usages
`Easy`	53.69	64.62	54.99	54.99
`Easy` (w/o *None*)	79.51	97.81	79.99	72.21
`Intermediate`	77.58	91.10	50.55	53.60
`Intermediate` (w/o *None*)	81.41	95.94	72.77	72.91
`Hard`	74.44	82.77	48.88	47.49
`Hard` (w/o *None*)	**87.58**	**96.06**	**80.41**	**74.44**

Table 3: Accuracy on the model's classification with samples used for annotation. The gold labels were majority voted by the results among the annotators. The difficulty with (w/o *None*) excludes samples where the gold label was labeled as *None*.

3.3 Two-stage Annotation

To establish standards for determining the natural use of sentence endings, we conducted a two-stage annotation process after expanding all the sentences. We began by performing human annotation on a subset of 20 sentences, covering 300 sentence ending instances from each difficulty level of the corpus. We found that even annotations from native Korean speakers can be inconsistent, as shown in Table 2. Given this situation, manually annotating the remaining sentences per difficulty level would be highly inefficient[5]. Therefore, for the cases not human-annotated, we utilized an LLM-based annotation (He et al., 2024; Ding et al., 2023).

To evaluate whether the selected model efficiently understands Korean sentence endings, we provided it with the samples used for human annotation[6]. We then compared the model's predictions to the majority voted human annotations and the accuracy results are provided in Table 3. The model achieved high accuracy in nearly all cases, generally aligning with the human annotation results.

Although the model demonstrated reliable performance, reaching a certain level of accuracy, we

[5]It will require a total of 980×15×3=44,100 sentence ending cases for each, in terms of both time and cost.

[6]In this case, we instructed the latest `gpt-4-turbo` model to perform zero-shot classification with careful attention to the use of sentence endings.

Sentence Endings	Llama3.1	Llama3	Llama3-ko	KULLM3	EXAONE3	Qwen2		Gemma2		Openchat	Synatra
	8B			10.7B	7.8B	1.5B	7B	2B	9B	8B	7B
Declarative Forms	13.06	15.09	17.33	14.98	15.41	13.83	13.23	16.33	14.44	13.49	16.64
	13.47	17.23	20.14	17.07	14.40	15.14	13.54	16.85	13.83	14.18	16.84
	12.33	15.77	18.31	16.82	13.85	14.25	12.54	15.78	13.05	13.35	15.46
Average	12.95	16.03	**18.59**	16.29	14.55	14.40	13.10	16.32	13.77	13.67	16.31
Imperative Forms	8.71	10.32	10.67	10.28	9.49	10.47	8.79	9.68	9.66	9.31	10.97
	8.67	12.40	12.26	11.75	9.91	11.23	10.23	9.92	10.66	10.65	12.16
	8.43	11.02	11.33	11.40	10.81	10.97	10.53	10.78	11.35	10.39	11.70
Average	8.60	11.24	11.42	11.14	10.07	10.89	9.85	10.12	10.55	10.11	**11.61**

Table 4: Accuracy of understanding Korean sentence endings across LLMs for the SE-*always* task. We determined each model's final accuracy using cyclic permutation, following the approach used in previous work (Kim et al., 2024a). For both Declarative and Imperative forms, the three reported values from the top represent results for Easy, Intermediate, and Hard, respectively. The model with the highest average score across all models is highlighted in bold, whereas the second-best model is underlined.

remained cautious about the potential for misclassifying sentence endings when annotating the remaining sentences. To address this, we employed following two strategies to enhance the model's ability to predict the usage of sentence endings accurately. First, we employed few-shot learning (Brown et al., 2020) by selecting a random sample of sentences and their sentence endings from human-annotated results that matched the usage patterns to predict. Second, we employed cyclic permutation (Izacard et al., 2023) when presenting options in the prompts to ensure unbiased model predictions independent of the order of the options, allowing it to focus on consistent patterns across different arrangements. The full annotation example and prompt configurations are provided in Appendix A.3. Finally, we constructed a dataset that includes 1,000 sentences for each difficulty level with 15 different sentence endings applied to each sentence.

4 Sentence Ending Tasks

We defined specific tasks to evaluate LLMs' understanding of sentence endings by selecting the most contextually natural option from the provided choices for each sentence ending. As mentioned earlier, the appropriate usage of sentence endings depends on the context, and their natural application may be absent in some cases.

In this scenario, we evaluated model performance in two cases: one where a natural ending is always expected (SE-*always*) and one where it may sometimes be absent (SE-*absent*)[7]. In the SE-*always* task, we excluded samples labeled *no usages* for each sentence ending and only included samples with labeled usages. In contrast, the SE-*absent* task allowed *no usages* as an option among the choices. This setup enabled us to compare model performance while considering the possibility of a missing natural sentence ending. For both tasks, we provided the model with four usage options for each sentence in a multiple-choice format. The details including the rules for presenting options are provided in Appendix B.1.

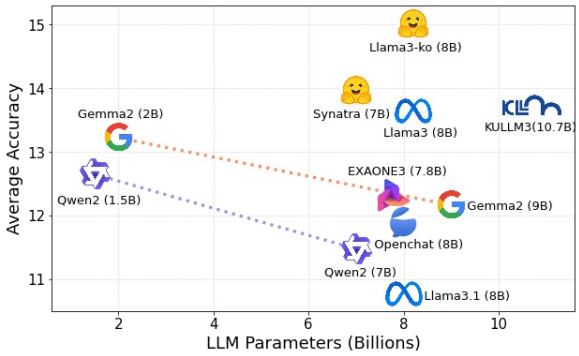

Figure 3: Comparison across LLMs based on parameter count, with scores averaged over all six difficulty levels for both Declarative and Imperative forms.

We experimented with a diverse set of LLMs to assess their understanding of sentence endings, containing Llama-families, Qwen2, and Gemma2 with parameter variations. We also selected Korean instruction-tuned models, including KULLM3 and EXAONE3[8]. The details regarding the models and metric are provided in Appendix B.2.

[7]In the following discussion of experimental results, we referred to the tasks as either SE-*always* or SE-*absent*, depending on which task was applied to evaluate the models.

[8]Due to resource constraints, we conducted main experiments using models with up to 10.7B parameters, while the results of pilot experiments with a larger 70B model are presented in Appendix C.3.

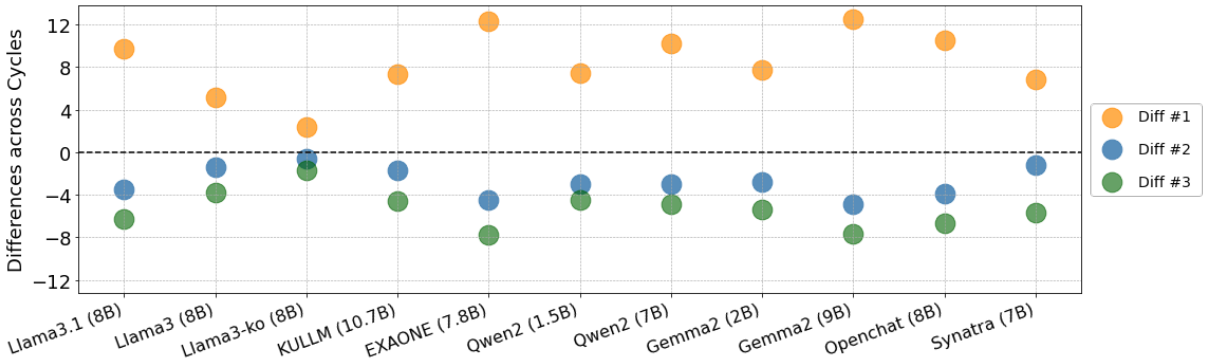

Figure 4: Difference between the accuracy of each cycle and the average accuracy across all cycles after applying three rounds of cyclic permutation to the models. The further a circle is from the dashed line, the greater the deviation from the average, indicating greater inconsistency in the model's predictions. 'Diff' in the legend means the form of the option order.

5 Discussion

How well do the models understand sentence endings? The results of the sentence ending comprehension evaluation using the proposed dataset with the SE-*always* task are presented in Table 4. We observed that their performance was relatively low, indicating that LLMs still have a limited understanding of Korean sentence endings[9]. To gain deeper insights into this situation, we compared the performance across several factors.

5.1 Experimental Results

Which type of sentence ending form is more challenging? We found that the accuracy for the `Imperative` forms was lower than that for the `Declarative` forms, indicating the greater difficulty in understanding sentence endings. This discrepancy likely arose because `Imperative` endings have more overlapping usage options than `Declarative` endings, making it more challenging for models to select contextually appropriate sentence endings.

Does the contextual difficulty affect understanding of sentence endings? We assumed that as the difficulty of the corpus increases, the models would struggle more to select the appropriate sentence endings. However, the results showed that corpus difficulty had a minimal effect on the accuracy of most models, except for `Gemma2` when predicting the usages of `Declarative` endings. This contrasts with the results in Table 2, which indicate that human annotation consistency decreased as corpus difficulty increased. It suggests that models faced more challenges in selecting the most natural sentence ending from the given options, regardless of the sentence's contextual complexity[10].

Model (Parameters)	Diff #1	Diff #2	Diff #3
`Llama3.1` (8B)	+9.69	-3.47	-6.22
`Llama3` (8B)	<u>+5.15</u>	-1.39	<u>-3.75</u>
`Llama3-ko` (8B)	**+2.35**	**-0.60**	**-1.74**
`KULLM3` (10.7B)	+7.39	-1.67	-4.54
`EXAONE3` (7.8B)	+12.27	-4.48	-7.79
`Qwen2` (1.5B)	+7.46	-2.99	-4.47
`Qwen2` (7B)	+10.20	-2.95	-4.91
`Gemma2` (2B)	+7.78	-2.76	-5.40
`Gemma2` (9B)	+12.56	-4.88	-7.67
`Openchat` (8B)	+10.53	-3.88	-6.64
`Synatra` (7B)	+6.90	<u>-1.24</u>	-5.66

Table 5: Numeral differences between the accuracy of each cycle and the average accuracy of cyclic permutations. The top-2 smallest absolute differences in each cycle are highlighted in bold or underlined.

How does model parameter size affect understanding of sentence endings? We compared the average accuracy based on the parameter count, in Figure 3. Although larger parameter counts in LLMs enhance performance in general tasks (Wu and Tang, 2024; Chowdhery et al., 2023), our results showed that the parameter size had minimal impact. For instance, of the 11 models, `KULLM3` with the largest parameters ranked in the top 4 for both `Declarative` and `Imperative` ending predictions. Its performance was not significantly better

[9] Although performance improved somewhat with the SE-*absent* task in Table 6, we observed that the overall performance level remained relatively low.

[10] Unlike in human annotation, the models were evaluated assuming no prior knowledge of specific usages, so we presented a broader range of options. While this may have influenced the results, the impact of difficulty on model accuracy during evaluation remained minimal.

Sentence Endings	Llama3.1	Llama3	Llama3-ko	KULLM3	EXAONE3	Qwen2		Gemma2		Openchat	Synatra
	8B			10.7B	7.8B	1.5B	7B	2B	9B	8B	7B
Declarative Forms	16.58	17.70	22.58	20.89	20.08	18.50	16.98	19.62	16.85	16.94	18.39
	14.39	18.63	23.27	21.02	16.37	19.32	15.46	19.16	16.30	14.81	18.45
	14.70	17.90	21.94	21.32	16.70	18.35	15.46	18.91	14.94	14.62	17.36
Average	15.22	18.07	**22.59**	21.07	17.71	18.72	15.96	19.23	16.03	15.45	18.06
Imperative Forms	14.47	14.51	20.63	18.45	20.96	14.63	16.52	17.30	13.96	20.29	15.84
	15.37	16.17	19.25	20.98	19.43	16.06	17.84	16.91	16.44	20.09	17.31
	17.71	16.81	16.79	23.65	21.86	17.22	20.08	19.60	19.00	21.20	19.39
Average	15.85	15.82	18.88	**21.02**	20.75	15.96	18.14	17.93	16.46	20.52	17.51

Table 6: Accuracy of understanding Korean sentence endings across LLMs for the SE-*absent* task. The method for determining final accuracy and the order of reported values by difficulty level match those presented in Table 4. The model with the highest average score across all models is highlighted in bold, whereas the second-best model is underlined.

than that of Qwen2, which had only 1.5B parameters. Similarly, Gemma2, with only 2B parameters, ranked in the top 2 in predicting Declarative endings. These relations suggest that all the models, regardless of the parameter count, face challenges in understanding Korean sentence endings.

5.2 How does the option order of sentence endings affect the model's understanding?

In our evaluation of sentence ending comprehension, we applied cyclic permutation (Izacard et al., 2023) to assess the impact of the order options on model predictions. While some models consistently predicted sentence endings accurately, regardless of the option order, most struggled to maintain performance despite minor changes due to cyclic permutation. The performance shift for each model is illustrated in Figure 4.

The results showed that almost all models exhibited inconsistencies with cyclic permutation, regardless of the model type or parameter count. Notably, EXAONE3 showed significant deviations, indicating poor robustness to changes in option order despite being additionally trained on a Korean dataset. Even larger models such as KULLM3 and Gemma2 (9B) were vulnerable to these shifts, indicating that even increased parameter sizes do not guarantee stability against changes in option order.

Conversely, Llama3-ko showed the smallest accuracy differences across cycles compared with that of the other models. It exhibited relatively greater consistency when compared with other models in the Llama-families and those with the same 8B parameters. Table 5 provides a clear view of these differences, demonstrating that Llama3-ko had a significantly lower variability across cycles. It is likely due to the base model choice or the particular instruction-tuning approach, as opposed to other models trained on Korean datasets.

Model (Parameters)	SE-*always* Task	SE-*absent* Task	Increased Accuracy
Llama3.1 (8B)	10.77	15.53	+4.76
Llama3 (8B)	13.63	16.94	+3.30
Llama3-ko (8B)	**15.00**	20.73	+5.73
KULLM3 (10.7B)	13.71	**21.04**	**+7.33**
EXAONE3 (7.8B)	12.31	19.23	+6.92
Qwen2 (1.5B)	12.64	17.34	+4.69
Qwen2 (7B)	11.47	17.05	+5.57
Gemma2 (2B)	13.21	18.58	+5.35
Gemma2 (9B)	12.16	16.24	+4.08
Openchat (8B)	11.89	17.98	+6.09
Synatra (7B)	13.95	17.78	+3.82

Table 7: Accuracy for both SE-*always* and SE-*absent* tasks, along with the improvements seen in the latter. These scores are averaged across all difficulty levels for both Declarative and Imperative forms. The top-2 highest scores in each column are highlighted in bold or underlined.

5.3 How does the possibility of no sentence ending affect the model's comprehension?

The results from the SE-*absent* task, in which the models were also given the *no usages* option when evaluating sentence ending comprehension, are presented in Table 6. All models exhibited a consistent performance improvement compared with that listed in Table 4, despite the increased number of samples used in the metric owing to the inclusion of the *no usages* option. This suggests that all the models in our experiments, regardless of their model type, better understood sentence ending usage when accounting for the possibility that no valid usage exists.

Similar to the SE-*always* task, we found that contextual difficulty had no significant impact on

accuracy when predicting the usage of sentence endings in this task. This suggests that, regardless of the model's awareness of an absent sentence ending, the selection of the most natural usage is influenced more by the available options than by the context of the sentence.

In addition, when comparing model performance by parameter size, the largest model KULLM3 ranked among the top 2 for both Declarative and Imperative forms. However, Gemma2 (2B) outperformed the 9B models in all cases, suggesting that even with the awareness of missing sentence endings, the parameter size did not consistently improve the understanding of sentence endings.

We presented the average scores for both SE-*always* and SE-*absent* tasks, highlighting the improvements in the SE-*absent* task in Table 7. In general, the models performed better when informed of the possibility that no appropriate sentence ending might exist. Notably, models such as KULLM3, Llama3-ko, and EXAONE3, instruction-tuned with the Korean dataset exhibited a more significant performance boost, indicating that instruction tuning in Korean helps LLMs better grasp the nuances of sentence ending usage.

6 Conclusion

We proposed the Korean Sentence Endings (**KoSEnd**) dataset to evaluate the ability of various LLMs to understand the use of diverse Korean sentence endings, considering the language's agglutinative nature. The dataset was categorized into three difficulty levels to reflect the varying contextual nuances from different sources. We expanded all sentences with 15 types of sentence endings, including Declarative and Imperative forms, and applied a two-stage annotation process to label their natural usage.

By evaluating the performance of LLMs under two SE-*always* and SE-*absent* tasks, whether they were informed that a sentence ending might be absent, we found that models such as Llama3-ko, Synatra, and KULLM3 achieved relatively high accuracy in both tasks. Furthermore, we examined performance variations based on the model parameters and the consistency of predictions through cyclic permutation. We observed that all models performed better when aware that a sentence ending might be missing. Moreover, the models instruction-tuned with a Korean dataset demonstrated strong prediction consistency and overall performance improvements.

Our study provides significant insights into evaluating linguistic knowledge in relatively low-resource agglutinative language, especially in Korean. Korean sentence endings convey not only grammatical roles but also semantic, emotional, and cultural nuances. Therefore, by using the proposed dataset, we expect to observe improvements in related performance for general tasks such as text generation, which we consider future work.

Limitations

The Risks of LLM-based Annotation While we incorporated some human annotations to capture natural sentence ending usage, most samples were annotated using an LLM-based annotation, raising concerns about label quality and potential biases. To mitigate this, we conducted a pilot test as shown in Table 3 to assess the reliability of this process. We further minimized bias by using human annotations as few-shot examples and employing cyclic permutation to reduce option order bias.

Constraints on Task and Model Selection We designed two tasks to evaluate each model's understanding of Korean sentence endings, but there remains ample room for further assessment using more diverse approaches. We aim to explore this comprehension from multiple angles, including its application to downstream tasks as future work.

Due to resource limitations, we focused on models with fewer parameters rather than larger 70B models, conducting an in-depth analysis to assess each model's understanding of Korean sentence endings from various perspectives. The pilot experiments with larger models can be found in Appendix C.3 in this aspect.

Ethics Statement

Our proposed dataset comes from diverse sources with varying difficulty levels, which may lead to sentences that reflect biases or contain discriminatory language based on the nature of these corpora. As the proposed dataset focuses on expanding and annotating Korean sentence endings, we did not leverage potentially biased information from the original sources.

In our experiments to evaluate Korean sentence ending comprehension across various LLMs, there is a possibility that the inherent biases of the model could have influenced the predictions. We designed the task with multiple-choice questions to mini-

mize such effects, focusing on the usage of each sentence ending. By framing this as a classification task and using greedy decoding, we aimed to avoid introducing additional biases from the models.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2021-II211341, Artificial Intelligence Graduate School Program (Chung-Ang University)] and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00556246).

References

Akari Asai, Sneha Kudugunta, Xinyan Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2024. BUFFET: Benchmarking large language models for few-shot cross-lingual transfer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1771–1800, Mexico City, Mexico. Association for Computational Linguistics.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137, Bangkok, Thailand. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Samuel Cahyawijaya, Holy Lovenia, Alham Fikri Aji, Genta Winata, Bryan Wilie, Fajri Koto, Rahmad Mahendra, Christian Wibisono, Ade Romadhony, Karissa Vincentio, Jennifer Santoso, David Moeljadi, Cahya Wirawan, Frederikus Hudi, Muhammad Satrio Wicaksono, Ivan Parmonangan, Ika Alfina, Ilham Firdausi Putra, Samsul Rahmadani, Yulianti Oenang, Ali Septiandri, James Jaya, Kaustubh Dhole, Arie Suryani, Rifki Afina Putri, Dan Su, Keith Stevens, Made Nindyatama Nityasya, Muhammad Adilazuarda, Ryan Hadiwijaya, Ryandito Diandaru, Tiezheng Yu, Vito Ghifari, Wenliang Dai, Yan Xu, Dyah Damapuspita, Haryo Wibowo, Cuk Tho, Ichwanul Karo Karo, Tirana Fatyanosa, Ziwei Ji, Graham Neubig, Timothy Baldwin, Sebastian Ruder, Pascale Fung, Herry Sujaini, Sakriani Sakti, and Ayu Purwarianti. 2023. NusaCrowd: Open source initiative for Indonesian NLP resources. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13745–13818, Toronto, Canada. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. LLMs are few-shot in-context low-resource language learners. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 405–433, Mexico City, Mexico. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195, Toronto, Canada. Association for Computational Linguistics.

Jiyeon Ham, Yo Joong Choe, Kyubyong Park, Ilji Choi, and Hyungjoon Soh. 2020. KorNLI and KorSTS: New benchmark datasets for Korean natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 422–430, Online. Association for Computational Linguistics.

Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.

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. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 165–190, Mexico City, Mexico. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought

prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.

Yerin Hwang, Yongil Kim, Hyunkyung Bae, Jeesoo Bang, Hwanhee Lee, and Kyomin Jung. 2024. Kosmic: Korean text similarity metric reflecting honorific distinctions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9954–9960.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

Myeongjun Jang, Dohyung Kim, Deuk Sin Kwon, and Eric Davis. 2022. KoBEST: Korean balanced evaluation of significant tasks. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3697–3708, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Seongbo Jang, Seonghyeon Lee, and Hwanjo Yu. 2024. KoDialogBench: Evaluating conversational understanding of language models with Korean dialogue benchmark. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9905–9925, Torino, Italia. ELRA and ICCL.

Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. 2024. Kobbq: Korean bias benchmark for question answering. *Transactions of the Association for Computational Linguistics*, 12:507–524.

Yiğit Bekir Kaya and A Cüneyd Tantuğ. 2024. Effect of tokenization granularity for turkish large language models. *Intelligent Systems with Applications*, 21:200335.

Pei Ke, Haozhe Ji, Siyang Liu, Xiaoyan Zhu, and Minlie Huang. 2020. SentiLARE: Sentiment-aware language representation learning with linguistic knowledge. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6975–6988, Online. Association for Computational Linguistics.

Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024a. CLIcK: A benchmark dataset of cultural and linguistic intelligence in Korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3335–3346, Torino, Italia. ELRA and ICCL.

Jong Myoung Kim, Young-Jun Lee, Yong-Jin Han, Ho-Jin Choi, and Sangkeun Jung. 2024b. Does incomplete syntax influence korean language model? focusing on word order and case markers. In *First Conference on Language Modeling*.

Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. 2022. Jglue: Japanese general language understanding evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2957–2966.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

NLP & AI Lab and Human-Inspired AI research. 2023. Kullm: Korea university large language model project. `https://github.com/nlpai-lab/kullm`.

Iksop Lee. 2005. *Korean Grammar*, volume 33. Seoul National University Press.

Jiyoung Lee, Minwoo Kim, Seungho Kim, Junghwan Kim, Seunghyun Won, Hwaran Lee, and Edward Choi. 2024. KorNAT: LLM alignment benchmark for Korean social values and common knowledge. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11177–11213, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Wei Qi Leong, Jian Gang Ngui, Yosephine Susanto, Hamsawardhini Rengarajan, Kengatharaiyer Sarveswaran, and William Chandra Tjhi. 2023. Bhasa: A holistic southeast asian linguistic and cultural evaluation suite for large language models. *arXiv preprint arXiv:2309.06085*.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ninghao Liu, and Mengnan Du. 2024. Quantifying multilingual performance of large language models across languages. *arXiv preprint arXiv:2404.11553*.

Seungyoung Lim, Myungji Kim, and Jooyoul Lee. 2019. Korquad1. 0: Korean qa dataset for machine reading comprehension. *arXiv preprint arXiv:1909.07005*.

Tomasz Limisiewicz, Jiří Balhar, and David Mareček. 2023. Tokenization impacts multilingual language modeling: Assessing vocabulary allocation and overlap across languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5661–5681, Toronto, Canada. Association for Computational Linguistics.

LiteLLM. 2025. Litellm documentation. Accessed on February 1, 2025.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2024. Is translation all you need? a study on solving multilingual tasks with large language models. *arXiv preprint arXiv:2403.10258*.

Akylbek Maxutov, Ayan Myrzakhmet, and Pavel Braslavski. 2024. Do llms speak kazakh? a pilot evaluation of seven models. In *Proceedings of the First Workshop on Natural Language Processing for Turkic Languages (SIGTURK 2024)*, pages 81–91.

Meta. 2024a. Introducing llama 3.1: Our most capable models to date. Accessed on February 1, 2025.

Meta. 2024b. Introducing meta llama 3: The most capable openly available llm to date. Accessed on February 1, 2025.

Alessio Miaschi, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2020. Linguistic profiling of a neural language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 745–756, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Arbi Haza Nasution and Aytug Onan. 2024. Chatgpt label: Comparing the quality of human-generated and llm-generated annotations in low-resource language nlp tasks. *IEEE Access*.

OpenRouter. 2025. Openrouter documentation. Accessed on February 1, 2025.

Chanjun Park, Hyeonwoo Kim, Dahyun Kim, SeongHwan Cho, Sanghoon Kim, Sukyung Lee, Yungi Kim, and Hwalsuk Lee. 2024a. Open Ko-LLM leaderboard: Evaluating large language models in Korean with Ko-h5 benchmark. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3220–3234, Bangkok, Thailand. Association for Computational Linguistics.

Dojun Park, Jiwoo Lee, Hyeyun Jeong, Seohyun Park, and Sungeun Lee. 2024b. Pragmatic competence evaluation of large language models for korean. *arXiv preprint arXiv:2403.12675*.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Ji Yoon Han, Jangwon Park, Chisung Song, Junseong Kim, Youngsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jung-Woo Ha (NAVER AI Lab), and Kyunghyun Cho. 2021. Klue: Korean language understanding evaluation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. 2024. Language model tokenizers introduce unfairness between languages. *Advances in Neural Information Processing Systems*, 36.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

LG Research, Soyoung An, Kyunghoon Bae, Eunbi Choi, Stanley Jungkyu Choi, Yemuk Choi, Seokhee Hong, Yeonjung Hong, Junwon Hwang, Hyojin Jeon, et al. 2024. Exaone 3.0 7.8 b instruction tuned language model. *arXiv preprint arXiv:2408.03541*.

Jaehyung Seo, Jaewook Lee, Chanjun Park, SeongTae Hong, Seungjun Lee, and Heui-Seok Lim. 2024. Kocommongen v2: A benchmark for navigating korean commonsense reasoning challenges in large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2390–2415.

Ho-Min Sohn. 2001. *The korean language*. Cambridge University Press.

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2024a. Kmmlu: Measuring massive multitask language understanding in korean. *arXiv preprint arXiv:2402.11548*.

Guijin Son, Hanwool Lee, Suwan Kim, Huiseo Kim, Jae cheol Lee, Je Won Yeom, Jihyu Jung, Jung woo Kim, and Songseong Kim. 2024b. HAE-RAE bench: Evaluation of Korean knowledge in language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7993–8007, Torino, Italia. ELRA and ICCL.

Jiayang Song, Yuheng Huang, Zhehua Zhou, and Lei Ma. 2024. Multilingual blending: Llm safety alignment evaluation with language mixture. *arXiv preprint arXiv:2407.07342*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Leonie Weissweiler, Valentin Hofmann, Anjali Kantharuban, Anna Cai, Ritam Dutt, Amey Hengle, Anubha Kabra, Atharva Kulkarni, Abhishek Vijayakumar, Haofei Yu, Hinrich Schuetze, Kemal Oflazer, and David Mortensen. 2023. Counting the bugs in ChatGPT's wugs: A multilingual investigation into the morphological capabilities of a large language model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6508–6524, Singapore. Association for Computational Linguistics.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, and Ayu Purwarianti. 2020. IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 843–857, Suzhou, China. Association for Computational Linguistics.

Chuhan Wu and Ruiming Tang. 2024. Performance law of large language models. *arXiv preprint arXiv:2408.09895*.

Jiannan Xiang, Huayang Li, Defu Lian, Guoping Huang, Taro Watanabe, and Lemao Liu. 2022. Visualizing the relationship between encoded linguistic information and task performance. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 410–422, Dublin, Ireland. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Soyoung Yoon, Sungjoon Park, Gyuwan Kim, Junhee Cho, Kihyo Park, Gyu Tae Kim, Minjoon Seo, and Alice Oh. 2023. Towards standardizing Korean grammatical error correction: Datasets and annotation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6713–6742, Toronto, Canada. Association for Computational Linguistics.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Advances in Neural Information Processing Systems*, 36:5484–5505.

A Further Details in KoSEnd: Dataset Construction

A.1 Corpus Collection

We used the language learner corpora (Yoon et al., 2023) for the `Easy` corpus. We expected sentences from these less-proficient writers to contain simple vocabulary and more straightforward contexts. For the `Intermediate` and `Hard` corpus, we used a newspaper corpus from the National Institute of the Korean Language[11] and summaries from academic papers[12]. We expected these texts to contain more complex vocabulary and fewer easily accessible contexts. Their information is presented in Table 8.

We selected sentences that ended with verbs and adjectives, as these were suitable for expanding sentence endings. Sentences considered too short to provide adequate context for understanding sentence endings were excluded.

Difficulty	Collected Sentences
`Easy`	1,000 sentences *from corrected Korean Learner Corpus*
`Intermediate`	1,000 sentences *for each of the 9 news topics* (IT and Science, Economy, Culture, Beauty and Health, Society, Lifestyle, Sports, Entertainment, Politics)
`Hard`	1,000 sentences *for each of the 8 academic fields* (Humanities, Agricultural and Marine Sciences, Social Sciences, Interdisciplinary Studies, Arts and Sports, Engineering, Natural Sciences, Medicine and Pharmacy)

Table 8: Corpus information for each difficulty level. For `Intermediate` and `Hard`, we ensured that the texts were gathered from diverse topics and fields.

A.2 Sentence Ending Expansion

In `Declarative` sentences, sentence endings such as the case (1) {다, 는다, ㄴ다} in Table 1 can be used to convey different meanings such as {*statements*, *exclamations*, *questions*}. The correct choice of sentence endings can vary depending on the reader's interpretation. For instance, "최선을 다하으마" is incorrect due to the verb stem form, while "최선을 다하마" is correct from the case (5). However, sentences such as "목숨을 바치는구나" and "목숨을 바치구나" from the case (2) are both acceptable and cannot be considered incorrect. In this situation, we conducted a two-stage annotation process to label the most natural cases after expanding all the sentences.

[11] Version 2023, https://kli.korean.go.kr/corpus/request/corpusRegist.do#none

[12] https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=90

A.3 Two-stage Annotation

Human Annotation Three native Korean-speaking university graduates volunteered to this process. A single sentence can be expanded to 33 versions, using 15 different `Declarative` and `Imperative` forms, as outlined in Table 1. We asked them to annotate whether each expanded version was appropriate for the context in binary form. Within each set of 15 ending forms, there may be multiple valid labels, or none at all. For example, in the case (1) {다, 는다, ㄴ다}, there are three possible endings. Depending on the sentence context, anywhere from 0 to 3 of these endings may be considered appropriate. This labeling process is repeated for all forms from (1) to (15). We especially noted that, depending on the context, *there might be no single best option or several acceptable options*. In this context, we used majority voting for the results of the human annotation to determine the gold labels for each usage.

LLM-based Annotation We used the human annotation results as few-shot samples to label the remaining sentences. For example, when labeling sentences of the case (1), we provided human-labeled examples of that form as 2-shot samples. We present the actual prompts used for LLM-based annotation as follows:

- LLM-based annotation prompt in Korean

system
당신은 한국어에 유능한 사람입니다. 당신의 업무는 종결 어미의 쓰임이 자연스러운 문장과 그에 부합하는 쓰임을 고르는 것입니다.

user
종결어미의 쓰임이 자연스러운 문장과 그에 부합하는 쓰임을 고르기 위해, 아래 예시를 참고할 수 있습니다.
문장 보기: {*sentence_sample1*}
쓰임 보기: {*usage_sample1*}

문장 보기: {*sentence_sample2*}
쓰임 보기: {*usage_sample2*}

주어진 한국어 문장들 중 종결 어미의 쓰임이 자연스러운 문장을 고르세요. 자연스러운 문장은 여러 개일 수도 있고, 하나도 없을 수도 있습니다.
문장 보기: {*sentence_option*}

주어진 문장 쓰임 중 앞서 고른 문장에 제일 부합하는 것을 고르세요. 부합하는 쓰임은 여러 개일 수도 있고, 하나도 없을 수도 있습니다.
쓰임 보기: {*usage_option*}

문장 정답 및 쓰임 정답을 별도의 설명 없이 알파벳으로 골라주세요.

- translated in English

system
You are fluent in Korean. Your task is to identify sentences where the sentence endings are naturally used and select the corresponding appropriate usage.

user
To determine the most natural sentence endings and their appropriate usage, you can refer to the examples below.
Sentence options: {*sentence_sample1*}
Usage options: {*usage_sample1*}

Sentence options: {*sentence_sample2*}
Usage options: {*usage_sample2*}

From the given Korean sentences, select those with natural sentence endings. There may be multiple correct answers, or none at all.
Sentence options: {*sentence_option*}

Next, choose the usage option that best matches the selected sentence(s). Again, there may be multiple correct answers, or none at all.
Usage options: {*usage_option*}

Please provide your answers using the alphabet letter, without any additional explanation.

B Further Details in Sentence Endings Evaluation

B.1 Task Definition

In the two-stage annotation process, only specific candidates relevant to each usage were presented to the human annotators and models. For instance, options such as the case (1) {다, 는다, ㄴ다} and (2) {구나, 는구나} in Table 1 were presented separately and not mixed. This approach ensured that, annotators or models could select the most appropriate sentence ending within that form, leading to the most natural choice for constructing the dataset.

In contrast, when evaluating the LLMs' understanding of sentence endings, we assumed that the model had no prior knowledge of the specific usage of the sentence. Thus, we combined options from all the forms and required the model to select the most natural sentence endings. To prevent the model from being influenced by the order of options, we applied cyclic permutation (Izacard et al., 2023), expecting results would remain consistent regardless of the arrangement of options.

In the dataset construction process, sentences labeled as *no usages*, indicating the absence of an ending across 15 possible cases of `Declarative` and `Imperative` endings, are detailed in Table 9.

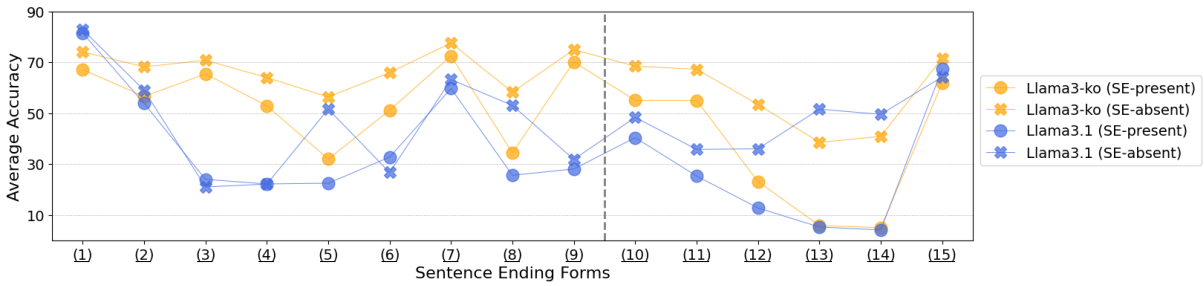

Figure 5: Average scores for each sentence ending form of the two models, `Llama3-ko` and `Llama3.1`, which exhibited the best and worst performance in our experiments. The x-axis displays (1)–(9) for `Declarative` forms and (10)–(15) for `Imperative` forms, as shown in Table 1. These scores represent the average across all difficulty levels and cycles for each sentence ending form.

Sentence Endings	Difficulty	*no usages* Counts	*no usages* Ratio
`Declarative` Forms	`Easy`	1,703	18.92%
	`Intermediate`	568	6.31%
	`Hard`	1,379	15.32%
`Imperative` Forms	`Easy`	3,149	52.48%
	`Intermediate`	2,770	46.16%
	`Hard`	2,973	49.55%

Table 9: Counts and proportions of sentences labeled as *no usages* in the proposed dataset, categorized by sentence ending types and difficulty levels.

Tasks	Model (Parameters)	`Easy`	`Intermediate`	`Hard`
SE-*always* Task	`EXAONE3` (7.8B)	0.013%	-	-
	`Qwen2` (7B)	-	-	0.002%
	`Gemma2` (9B)	-	-	0.002%
	`Synatra` (7B)	0.006%	-	-
SE-*absent* Task	`KULLM3` (10.7B)	0.002%	0.008%	0.04%
	`EXAONE3` (7.8B)	0.02%	-	-
	`Synatra` (7B)	-	0.004%	0.002%

Table 10: Hallucination rates for each task, based on the selected models. Any values not listed in the table were not classified as hallucinations according to our post-processing process.

B.2 Experimental Settings

The models to evaluate the understanding of Korean sentence endings are as follows: `Llama`-families (Meta, 2024a,b), `Gemma2` (Team et al., 2024), and `Qwen2` (Yang et al., 2024) were selected as the multilingual models. In addition, `KULLM3` (Lab and research, 2023) and `EXAONE3` (Research et al., 2024) were instruction-tuned using a Korean dataset. Specifically, as of September 2024, `Openchat`[13] and `Synatra`[14] were ranked as the top-2 models on the Open Ko-LLM Leaderboard[15] (Park et al., 2024a). We set the temperature to 0 to enable greedy decoding for predicting the most natural usage of sentence endings. We used the vLLM library (Kwon et al., 2023) to enable efficient inference.

We measured accuracy by comparing the models' responses to the gold labels obtained through a two-stage annotation process. Each model generated responses to the same prompt three times using cyclic permutation, aligning with accuracy metrics from previous work (Kim et al., 2024a).

[13] `https://huggingface.co/openchat/openchat-3.6-8b-20240522`

[14] `https://huggingface.co/maywell/Synatra-7B-v0.3-dpo`

[15] This leaderboard, a key benchmark for Korean language tasks using private test sets, features the top-performing models in Korean for various downstream tasks.

C Further Details in Experiments

C.1 Post Processing

When we instructed the models, some generated additional explanations alongside their selections. To refine these outputs, we applied post-processing, prioritizing the alphabet following phrases like 'correct answer' or removing irrelevant characters that did not represent the answer. If the answer remained unclear after this process, we classified it as a hallucination. The hallucination rates for each model are shown in Table 10. We excluded these hallucination samples from the evaluation.

C.2 Experimental Results on Each Sentence Ending Form

To analyze the impact of each sentence ending form on model performance, we reported the results for each form individually in Figure 5. Based on the results in Table 7, we selected `Llama3-ko` and `Llama3.1`, which exhibited the highest and lowest performance in both SE-*always* and SE-*absent* tasks, respectively. In most cases, regardless of the sentence ending form, we observed performance

Sentence Endings	Llama3.1		Qwen2.5	
	8B	70B	7B	72B
Declarative Forms	29.90	39.30 (+31.43%)	34.90	42.10 (+20.63%)
	26.60	34.60 (+30.07%)	27.10	32.40 (+19.55%)
	26.10	31.60 (+21.07%)	27.40	32.00 (+16.78%)
Average	27.53	<u>35.16</u> (+27.71%)	29.80	**35.50** (+19.12%)
Imperative Forms	25.80	35.30 (+36.82%)	45.80	49.60 (+8.29%)
	27.50	37.00 (+34.54%)	45.00	48.80 (+8.44%)
	30.10	40.60 (+34.88%)	49.30	53.40 (+8.31%)
Average	27.80	37.63 (+35.35%)	<u>46.69</u>	**50.60** (+8.37%)

Table 11: Accuracy of understanding Korean sentence endings with larger models <u>for the SE-*absent* task</u> using 1,000 samples. For both Declarative and Imperative forms, the three reported values from the top represent results for Easy, Intermediate, and Hard, respectively. The top-2 highest averaged scores in each form are highlighted in bold or underlined. The values in parentheses represent the rate of performance improvement.

improvements when the models were informed about the potential absence of a sentence ending. This trend was consistent across both Llama3-ko and Llama3.1, suggesting that recognizing the possibility of a missing sentence ending enhances their understanding of Korean sentence endings.

Although Llama3-ko demonstrated strong performance across most sentence-ending forms, we observed that Llama3.1 either outperformed or achieved comparable results to Llama3-ko in cases <u>(1)</u> and <u>(13)</u>~<u>(15)</u>. Cases <u>(1)</u>, <u>(13)</u>, and <u>(14)</u> represent the most commonly used forms, including *statements* and *informal speeches*. Llama3.1's improved performance can be attributed to its training on larger dataset as a more recent model. Case <u>(15)</u> from the Imperative forms includes six different usages, the highest number of usages for any sentence ending form. This suggests that Llama3.1's ability to handle a broader range of variations allowed it to perform comparably to Llama3-ko.

C.3 Pilot Experiments with Larger Models

We conducted pilot experiments to evaluate Korean sentence endings using larger models not included in the main analysis. Due to time and budget constraints, we tested 1,000 samples for each combination of Declarative and Imperative sentence ending forms and three difficulty levels. We evaluated the Llama3.1 70B and Qwen2.5 72B models on SE-*absent* task, using LiteLLM (LiteLLM, 2025) and OpenRouter (OpenRouter, 2025).

The results in Table 11 showed that the larger models consistently outperformed smaller ones across all cases, regardless of sentence ending form or difficulty level. While larger models demonstrated capabilities in understanding sentence endings, the performance gains did not scale proportionally with the increase in parameter size. This indicates that even the larger models still face challenges in grasping the nuances of the Korean sentence endings.

Notably, Qwen2.5 7B demonstrated a relatively higher understanding of Imperative forms, even surpassing the performance of Llama3.1 70B. In contrast, within the Llama3.1-families, larger models consistently outperformed smaller ones with performance gains of around 30%. This suggests that while Llama3.1 showed greater improvements with larger model size, Qwen2.5 achieved higher overall performance. This pilot experiments provided a broader perspective on the impact of our dataset. We hope these observations will help inform strategic decisions on model selection—both in terms of type and size—when assessing the understanding of Korean sentence endings.