Polishing Every Facet of the GEM: Testing Linguistic Competence of LLMs and Humans in Korean

SungHo Kim[1*], Nayeon Kim[2*], Taehee Jeon[3], SangKeun Lee[1,2]
[1]Department of Artificial Intelligence, Korea University, Seoul, South Korea
[2]Department of Computer Science and Engineering, Korea University, Seoul, South Korea
[3]Institute for Digital HUSS, Korea University, Seoul, South Korea
{sungho3268, lilian1208, taeheejeon22, yalphy}@korea.ac.kr

Abstract

We introduce the Korean Grammar Evaluation BenchMark (KoGEM), designed to assess the linguistic competence of LLMs and humans in Korean. KoGEM consists of 1.5k multiple-choice QA pairs covering five main categories and 16 subcategories. The zero-shot evaluation of 27 LLMs of various sizes and types reveals that while LLMs perform remarkably well on straightforward tasks requiring primarily definitional knowledge, they struggle with tasks that demand the integration of real-world experiential knowledge, such as phonological rules and pronunciation. Furthermore, our in-depth analysis suggests that incorporating such experiential knowledge could enhance the linguistic competence of LLMs. With KoGEM, we not only highlight the limitations of current LLMs in linguistic competence but also uncover hidden facets of LLMs in linguistic competence, paving the way for enhancing comprehensive language understanding. Our code and dataset are available at: https://github.com/SungHo3268/KoGEM.

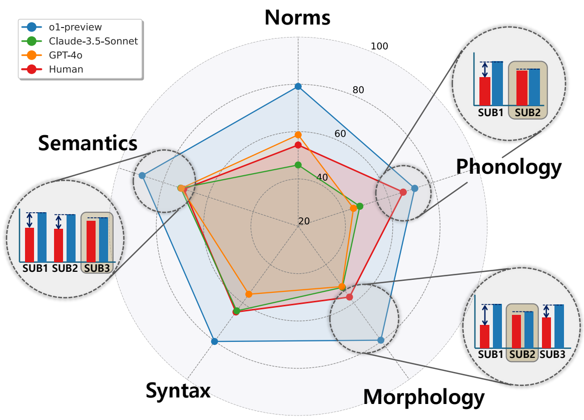


Figure 1: Zero-shot accuracy of the top three LLMs and human performance on KoGEM. Dashed gray circles indicate the accuracy of human (red box) and o1-preview (blue box) in each subcategory. `SUB` denotes the subcategory corresponding to each main category.

1 Introduction

Although large language models (LLMs) have demonstrated remarkable performance across various natural language tasks, it is uncertain whether they possess genuine linguistic competence—the ability to understand the underlying principles of a language (Chomsky, 1965; Waldis et al., 2024). Their strong performance might stem from their extensive training data rather than understanding of language itself (Bender et al., 2021). Thus, to explore whether LLMs truly understand language beyond statistical pattern recognition, it is crucial to investigate their linguistic competence. However, due to the implicit characteristics of linguistic competence, directly assessing such competence is challenging (Nam et al., 2024).

One promising approach to address this challenge is to leverage grammar as a measurable proxy. As grammar explicitly formulates the universal rules of a language (White, 1989), it can serve as an effective way to assess the linguistic competence of LLMs. Previous studies have explored the linguistic competence of language models with grammatical knowledge (Hewitt and Manning, 2019; Blevins et al., 2023; Amouyal et al., 2024). While various aspects of linguistic knowledge exist, such as phonology and pragmatics, most of the previous works have primarily focused on morphological and syntactic knowledge in English. Additionally, they have paid little attention to other languages, including Korean.

Since individual languages possess unique linguistic properties, each language should be considered independently to evaluate how well LLMs understand its linguistic knowledge. For this reason, this paper specifically focuses on the Korean language to facilitate a deeper discussion. Unlike English, Korean, as an agglutinative language,

[*] These authors contributed equally to this work.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9955–9984

exhibits significant morphological variation (Kim et al., 2022, 2024d; Seo et al., 2023). In addition, the writing system of Korean gives rise to unique phonological rules, such as consonant assimilation and vowel harmony (Cho and Whitman, 2020).

With this motivation, we present the Korean Grammar Evaluation BenchMark, KoGEM, a comprehensive and fine-grained dataset designed to assess the linguistic competence of LLMs in Korean. KoGEM consists of 1,524 multiple-choice grammar questions organized into five main categories—`Phonology`, `Morphology`, `Syntax`, `Semantics`, and `Norms`—which are further divided into 16 subcategories based on theoretical linguistics (Lyons, 1968). This structured taxonomy enables KoGEM to provide a wide and detailed framework for evaluating the linguistic competence of LLMs.

We evaluate humans and a diverse range of open- and closed-source LLMs, including both Korean- and English-centric models, in a zero-shot setting on KoGEM. Figure 1 presents the key results, comparing human performance with the top three LLMs on KoGEM. At first glance, the outstanding performance of o1-preview may appear flawless, suggesting that it surpasses humans in all aspects. However, a closer analysis of linguistic phenomena, breaking down major linguistic categories into finer subcategories, reveals highly varied tendencies. Notably, we identify certain hidden facets where humans perform relatively well while LLMs lag behind, highlighting potential areas for improvement. Through an in-depth analysis of these hidden facets, we demonstrate substantial improvements when LLMs are augmented with the experiential knowledge that humans naturally acquire through real-world experience. In summary, our contributions are as follows:

- We introduce KoGEM, a comprehensive and fine-grained benchmark designed to objectively assess Korean grammatical knowledge based on theoretical linguistics.
- We evaluate 27 open- and closed-source LLMs, including Korean- and English-centric models, across 16 fine-grained Korean grammar taxonomy, comparing them with humans.
- We reveal novel insights into the strengths and limitations of LLMs through in-depth analysis, paving the way for enhancing LLMs and addressing gaps in their linguistic competence.

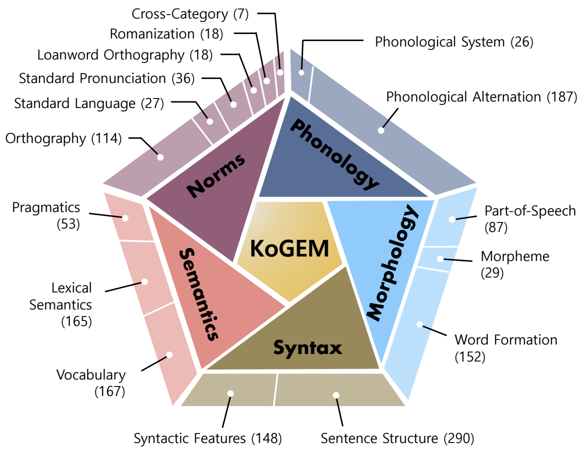


Figure 2: Data distribution of KoGEM, categorized into five main categories and 16 subcategories in total. The number in parentheses next to each subcategory name represents the number of questions it contains.

2 Korean Grammar Taxonomy

Before delving into the details, we define the concept of 'grammar.' This paper focuses solely on prescriptive grammar due to the extensiveness of descriptive grammar. While prescriptive grammar does not encompass all aspects of a language's grammatical system, it is firmly based on foundational knowledge of grammar. By examining the degree to which prescriptive grammar is understood, we can evaluate the linguistic competence of both LLMs and humans.

To assess linguistic competence, we set the categories of grammar as the main fields of theoretical linguistics: phonology, morphology, syntax, and semantics. While pragmatics can be treated as a separate field, we include it under semantics in this framework due to its close relationship with semantics and the relative absence of pragmatics considerations in prescriptive grammar. Additionally, we include linguistic norms as a category, reflecting one of the core principles of prescriptive grammar. Although norms cannot inherently capture the diversity of linguistic phenomena, understanding norms requires linguistic competence, which is the focus of our evaluation.

Further, we define the subcategories for each main linguistic category. The subcategories encompass key subfields of each main category and are aligned with the structure of the current Korean high school education system. For example, in phonology, the subfields regarding the phonological system and variation in Korean are divided into phonological system and phonological alternation.

Further details can be found in Appendix D.4.

- **Phonology**: Phonological System, Phonological Alternation
- **Morphology**: Part-of-Speech, Morpheme, Word Formation
- **Syntax**: Sentence Structure, Syntactic Features
- **Semantics**: Vocabulary, Lexical Semantics, Pragmatics
- **Norms**: Orthography, Standard Language, Standard Pronunciation, Loanword Orthography, Romanization, Cross-Category

3 KoGEM

The primary purpose of our dataset is to assess the linguistic competence of LLMs and humans. To this end, we introduce KoGEM, a **Ko**rean **G**rammar **E**valuation bench**M**ark, containing 1.5k grammar question-answer (QA) pairs, categorized into five main categories and 16 subcategories based on a predefined taxonomy described in Section 2.

3.1 Dataset Construction

We provide a detailed explanation of source data, data format, collection, and categorization.

1) Source Data To encompass the Korean grammar taxonomy defined in Section 2, we extract Korean grammar questions from four types of official exams: (1) the College Scholastic Ability Test (CSAT); (2) the National United Achievement Test (NUAT); (3) the High School Qualification Exam (HSQE); and (4) the Civil Service Exam (CSE). While other Korean language tests exist, we especially selected these exams as they are designed for native Korean speakers, ensuring that the questions reflect linguistic competence expected in academic and professional contexts. A specific description of our source data is provided in Appendix D.1.

2) Data Format As prescriptive grammar emphasizes the correct use of language, we design our task as a multiple-choice QA task with a clearly defined correct answer. Each QA pair consists of up to four components: `passage`, `question`, `paragraph`, and `choices`. The `passage` provides the necessary context to understand the question, while the `paragraph` offers brief explanations or examples of grammatical concepts relevant to answering it. The `choices` include either four or five answer options, depending on the type of exam. Examples of QA pairs can be found in Figures 12 to 16.

3) Data Collection As the source data is publicly available in PDF format, we extract the text using optical character recognition (OCR). After OCR, three authors manually review every grammar question and apply preprocessing steps such as underlining, segmentation, and box separation using Hypertext Markup Language (HTML) formatting. Questions primarily relying on images for context are excluded, as we evaluate LLMs solely in a text-based, single-modal setting. Additionally, tables in the data are converted into sequential text based on a set of predefined rules. Detailed dataset preprocessing is described in Appendix D.2.

4) Data Categorization To assess the linguistic competence of LLMs in various grammatical areas, three Korean language majors categorize the preprocessed questions into 16 subcategories. Each annotator independently classifies each question, first identifying the main linguistic category and then assigning the appropriate subcategory. To ensure the reliability of our categorization, labels are finalized by majority vote, and disagreements are resolved through discussion. Questions overlapping two linguistic categories are classified into one category based on the context of the correct answer. Additionally, we remove questions requiring knowledge from more than three categories to focus on the evaluation of each linguistic subcategory.

3.2 Data Statistics

The finalized KoGEM benchmark consists of a total of 1,524 annotated QA pairs. Figure 2 illustrates the distribution of KoGEM across each main category and subcategory, classified according to the Korean grammar taxonomy. Moreover, we present the specific number of samples for each main category from each source exam in Figure 8.

4 Experiments

In this section, we evaluate the linguistic competence of both LLMs and humans using KoGEM. Section 4.1 describes the experimental settings for the LLMs in the view of baselines and evaluation metrics, while Section 4.2 provides a detailed explanation of the methods for assessing human performance. Finally, Section 4.3 presents comprehensive experimental results for each main category.

Language	Type	Model	Phonology	Morphology	Syntax	Semantics	Norm	Avg.
Korean	Open	Bllossom-8B	24.41	25.00	22.15	34.55	29.09	27.10
		SOLAR-v1.0-10.7B-Instruct	24.88	25.75	26.71	31.69	25.45	27.36
		KULLM-3-10.7B	21.60	26.49	26.03	29.35	28.18	26.64
		EEVE-v1.0-10.8B-Instruct	22.54	27.24	27.85	40.78	27.73	30.25
		EXAONE-3.5-7.8B-Instruct	24.88	30.22	32.19	43.64	31.36	33.60
		EXAONE-3.5-32B-Instruct	27.23	37.31	36.30	50.65	37.27	38.98
	Closed	HyperCLOVA-HCX-DASH-001	23.94	31.34	25.57	39.74	31.36	30.77
		HyperCLOVA-HCX-003	32.39	41.79	41.10	55.32	48.18	44.62
English	Open	Gemma-2-9B-Instruct	24.41	31.34	31.05	42.08	29.09	32.68
		Gemma-2-27B-Instruct	27.33	29.48	36.76	47.27	29.09	35.70
		Qwen2.5-7B-Instruct	23.00	30.97	34.02	44.42	25.00	33.27
		Qwen2.5-14B-Instruct	29.58	38.43	39.27	52.21	33.64	40.22
		Qwen2.5-32B-Instruct	26.29	36.19	41.78	62.86	35.45	43.04
		DeepSeek-R1-Distill-Qwen-14B	29.11	32.09	39.27	47.79	32.27	37.73
		DeepSeek-R1-Distill-Qwen-32B	36.15	45.52	40.41	61.56	30.00	44.55
		s1-32B	39.91	43.28	46.12	62.60	40.00	48.03
		Llama-3.1-8B-Instruct	24.41	24.25	23.97	36.36	26.82	27.62
		Llama-3-70B	24.41	32.46	34.47	45.71	27.73	34.58
		Llama-3.1-405B	37.56	43.28	47.03	62.34	35.00	47.18
	Closed	Gemini-1.5-flash	37.56	38.81	44.52	60.78	35.45	45.34
		Gemini-2.0-flash-exp	46.01	49.63	56.85	70.91	45.91	56.04
		Claude-3-haiku	21.13	34.33	35.62	44.68	30.91	34.97
		Claude-3.5-Sonnet	<u>47.42</u>	<u>52.61</u>	<u>64.38</u>	<u>74.55</u>	46.82	<u>59.97</u>
		GPT-3.5-turbo	21.13	27.61	27.40	31.43	25.45	27.30
		GPT-4o-mini	32.39	34.70	36.53	52.21	39.09	39.96
		GPT-4o	44.60	51.49	55.48	71.95	<u>58.64</u>	57.87
		o1-preview	**71.83**	**79.48**	**80.14**	**89.35**	**79.09**	**81.04**
LLMs Avg.			31.33	37.08	39.00	51.36	35.71	40.24
Human			66.70	56.95	64.75	70.84	54.34	63.04

Table 1: Zero-shot accuracy evaluation results on our KoGEM benchmark. It consists of four segments: Korean-centric LLMs trained mainly on Korean data, English-centric LLMs trained primarily on English data, and the average accuracy performance of all LLMs and humans, respectively.

4.1 LLM Evaluation

Baselines. We evaluated 8 Korean-centric LLMs, such as the EXAONE and HyperCLOVA X series, as well as 19 well-known English-centric LLMs, including the OpenAI-GPT, Claude, and Gemini series. Specifically, we compared open-source LLMs of various sizes, ranging from smaller models with 8B parameters to cutting-edge closed-source models such as o1-preview. Detailed descriptions of the evaluated models can be found in Appendix A.

Evaluation Metrics. To assess the Korean grammatical knowledge of LLMs on KoGEM, we measure accuracy through zero-shot evaluation and few-shot evaluation.[1] The prompt was structured in the same order as the test presented to humans: `passage`, `question`, `paragraph`, and `choices`, and provided as a single prompt. The LLMs were instructed to select the correct answer from the choices and generate a short explanation for their choices. Detailed prompt designs and experimental settings, such as hyperparameters and devices, are provided in Appendix B.

4.2 Human Evaluation

Since the data comes from official competency exams in Korea, we first explored publicly available statistics on human performance. Through this investigation, we obtained average response rates for each question from 10,000+ responses for CSAT and NUAT exams.[2] However, for the other two types of exams, HSQE and CSE, there are no publicly available data for human performance. To gain the human scores for these exams, we conducted crowdsourcing via a data research platform [3]. To ensure suitable participants for each exam, we re-

[1]While, in the main body of this paper, we just handle the zero-shot, you can find the few-shot evaluation results in Appendix H.

[2]The Korean private education company Megastudy has published accuracy for each question since 2016 (https://www.megastudy.net/Entinfo/correctRate/main.asp).

[3]Macromill Embrain, a Korean company specializing in online research (https://embrain.com).

cruited different groups of crowdworkers based on the specific characteristics of each exam. Additionally, we gathered 10+ responses per each question and used the average scores. Details of crowdsourcing are described in Appendix C.

4.3 Results for Main Category

Table 1 presents the main linguistic category-specific performances of LLMs and humans on KoGEM. Overall, LLM results generally follow the scaling law (Kaplan et al., 2020). Notably, among LLMs, o1-preview was the only model to outperform humans, exceeding their performance by an average of 18.00%. All other models fell short: Claude-3.5-Sonnet and GPT-4o scored 3.07% and 5.17% lower than humans, respectively.

We categorized LLMs into Korean-centric and English-centric groups based on their primary training language. Many Korean-centric models exhibited lower performance, likely due to their smaller sizes and outdated architectures. Surprisingly, the English-centric models s1-32B, DeepSeek-R1-Distill-Qwen-32B, and Qwen2.5-32B, despite their relatively small sizes, outperformed or matched all Korean-centric models. Their strong performance likely stems from their multilingual training across over 29 languages (Qwen Team, 2024), leveraging shared linguistic features (Chen et al., 2019; Wang et al., 2024; Chang et al., 2022). Moreover, results from o1-preview, s1-32B, and two DeepSeek-R1 series suggest that test-time scaling (Snell et al., 2024; Muennighoff et al., 2025) effectively enhances linguistic competence in Korean.

A notable finding is the variation in LLMs and human performance across linguistic categories. Specifically, humans excelled in the Phonology category, which requires multimodal reasoning, such as integrating text with implicit phonological rules and pronunciation. In contrast, the Phonology was the weakest category for LLMs. Even o1-preview, despite its overall superiority, surpassed humans by only 5.13% in this category, much lower than the 18.00% gap observed across all categories. Conversely, the smallest performance gap between humans and LLMs was observed in the Norms category, which relies heavily on rote knowledge, such as correct spelling and loanword usage.

However, we question whether trends at the main category level are sufficient to fully capture the core linguistic differences between LLMs and humans. To gain deeper insights, Sections 5 and 6 further break down the five main categories into 16 subcategories. This fine-grained analysis provides a more detailed understanding of LLMs' linguistic competence, highlighting both their strengths and limitations.

5 Results for Each Subcategory

Figure 3 compares the performance of LLMs and humans across 16 subcategories. A closer examination of individual subcategories reveals distinct strengths and weaknesses, as LLMs and humans excel in different areas. This underscores the need for a fine-grained evaluation of linguistic competence at the subcategory level.

Phonology The phonology category represents the area with the most significant performance gap between LLMs and humans. Specifically, regarding the `Phonological Alternation` subcategory, humans outperformed the average performance of LLMs by over 35%, marking the largest performance gap among all subcategories of KoGEM. We suspect that this performance gap between humans and LLMs stems from the ability of humans to ground multimodal knowledge (Smith and Dechant, 1961; Holler and Levinson, 2019). Humans intuitively recall the pronunciation of a word (Carroll, 1986), recognizing implicit phonological changes and integrating the pronunciation with the text. In contrast, LLMs rely on textual input, lacking exposure to spoken language and phonological processing in real-world contexts. This fundamental challenge can limit the ability of LLMs to process phonological alternations effectively.

Morphology Interestingly, humans tend to underperform in the `Part-of-Speech` subcategory, whereas LLMs achieve their highest performance within the subcategories of the Morphology category. This discrepancy may stem from differences in how humans and LLMs process language. Humans rely on intuitive contextual understanding (McClelland et al., 2019) rather than explicitly distinguishing part-of-speech. In contrast, LLMs classify them through data-driven pattern detection, giving them a distinct advantage. On the other hand, in the `Morpheme` subcategory, the gap between humans and LLMs is the largest within the Morphology category. Humans instinctively decompose words into inflectional morphemes and root words that carry the core meaning, allowing for intuitive interpretation (Marslen-Wilson and

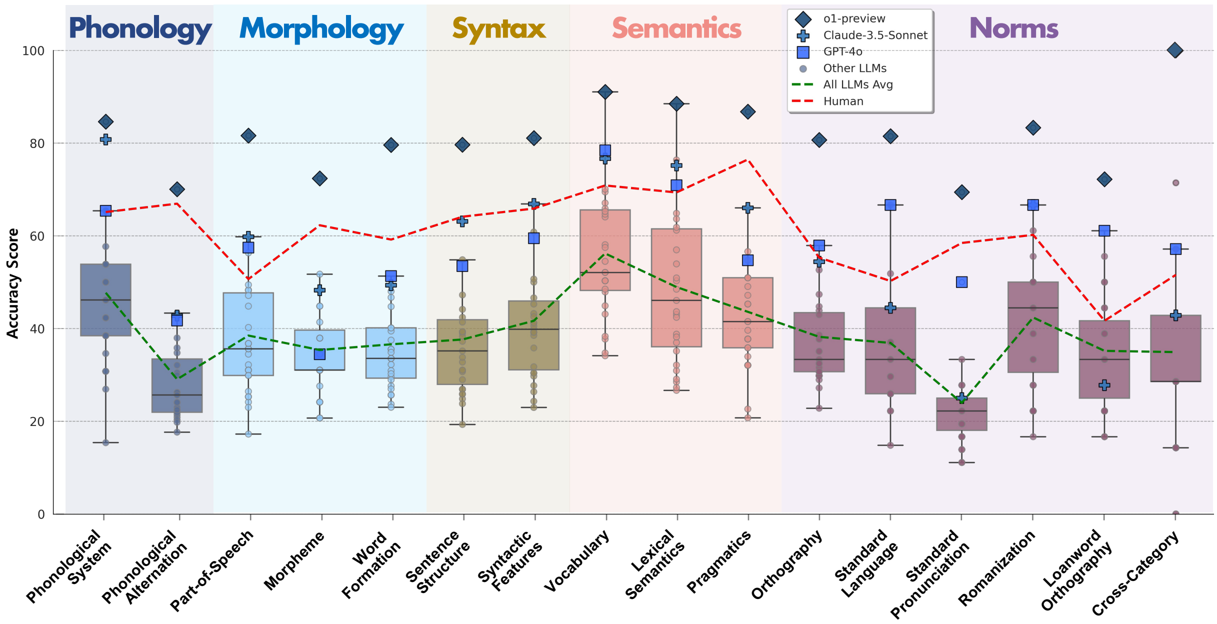


Figure 3: Comparison of LLMs and humans across 16 subcategories. The distributions of all LLMs are depicted using box plots, while individual scores for the top three models are highlighted as scatter plots. A green dashed line indicates the average performance of all LLMs. [4], and a red dashed line represents human performance.

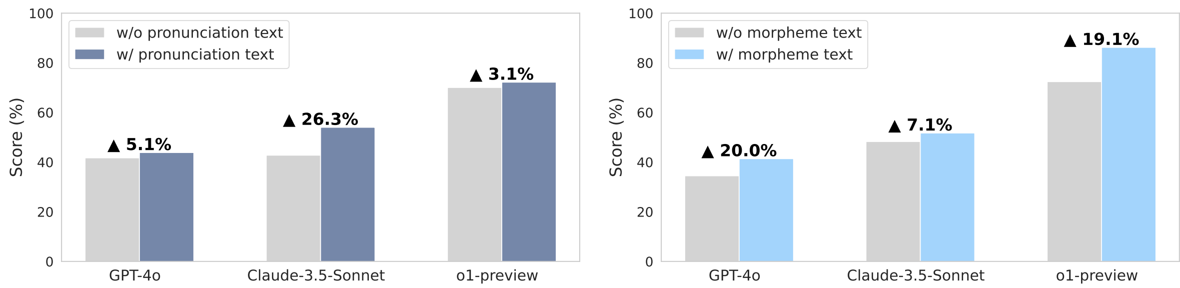
Tyler, 2007). However, LLMs process words either as whole units or as tokenized segments based on their defined vocabulary, making them less suited to handling diverse morphological variations.

Syntax LLMs and humans exhibited similar tendencies, particularly showing better performance on `Syntactic Feature` subcategory than on `Sentence Structure` subcategory. This trend may be attributed to the unique syntactic characteristics in Korean, which has relatively flexible word order compared to English (Cho and Whitman, 2020). For example, the Korean translation of “I eat rice” is “나는 밥을 먹는다$_{\text{*I rice eat}}$”. However, this sentence can also appear as “나는 먹는다 밥을$_{\text{I eat rice}}$” in Korean. This flexibility is difficult to codify, requiring intuitive understanding through real-world experience to comprehend it. On the other hand, since `Syntactic Feature` task follows well-defined rules, such as tense, voice, and honorifics, this tends to be more standardized than the `Sentence Structure` task.

Semantics Within the Semantics category, the differences among subcategories are particularly intriguing. Specifically, `Pragmatics` subcategory shows the largest performance gap within the Semantics category, with an average difference of 32.85% between LLMs and humans, whereas `Vocabulary` subcategory exhibits the smallest disparity, averaging 14.66%. This contrast likely stems from the nature of each subcategory. First, in `Vocabulary` subcategory, many QA tasks involve definitional knowledge and the correct use of words, often requiring straightforward memorization of word meanings. In contrast, `Pragmatics` subcategory demands an understanding of context-specific intent, such as the relationship between speaker and listener, and the communicative purpose of an utterance, which relies heavily on real-world conversational experience (Levinson, 1983). As a result, LLMs can struggle with `Pragmatics` subcategory more than with the relatively straightforward `Vocabulary` subcategory.

Norms Both humans and LLMs exhibited lower overall performance in the Norms category compared to other areas. Although Koreans receive standardized education on linguistic norms as part of their high school curriculum, they often struggle to adhere to standard language regulations (Kim et al., 2020). We suspect that LLMs may have limited exposure to texts explicitly describing standard regulations. As a result, both humans and LLMs may generally exhibit lower performance in the Norms category. Among the subcategories within Norms category, the largest performance gap between LLMs and humans, approximately 34.49%, was observed in `Standard Pronunciation`. This

[4]More detailed top-k average results of LLMs, along with their distributions by subcategory, are provided in Appendix F.



(a) Phonology (`Phonological Alternation`)

(b) Morphology (`Morpheme`)

Figure 4: Comparison of performances in (a) `Phonological Alternation` and (b) `Morpheme` subcategories, with and without the additional knowledge, such as *pronunciation text* and *morpheme text*. The values between each bar graph pair indicate the rate of increase(▲) compared to the original results.

trend parallels findings in the Phonology category, suggesting that the gap may stem not only from normative factors but also from phonological features.

Summary LLMs exhibit a relatively smaller performance gap compared to humans in tasks that rely on memorization and pattern recognition. In contrast, they fall significantly behind in tasks requiring experiential knowledge, such as `Phonological Alternation`, `Morpheme`, and `Pragmatics`, revealing a substantial performance gap between LLMs and humans. This underscores a key limitation in LLMs' ability to ground linguistic knowledge and real-world experience.

6 In-depth Analysis for Subcategory

From our previous analysis of subcategory results, we identified certain subcategories that are relatively easy for humans but challenging for LLMs. In this section, we further investigate these specific subcategories. Section 6.1 compares the *thinking* time required to solve problems across all subcategories using the s1-32B, which demonstrated the best performance among open-sourced models employing test-time scaling. Section 6.2 examines the impact of incorporating experiential knowledge typically utilized by humans, such as pronunciation and morphemes in `Phonological Alternation` and `Morpheme` subcategories, respectively.

6.1 Comparison of Thinking Time

We aimed to indirectly assess the level of difficulty that LLMs face across subcategories by measuring their thinking time. Concretely, we evaluated the thinking time of the s1-32B model across all 1,524 questions in KoGEM and calculated the average time per subcategory. As shown in Figure 5, our results indicate that the model takes significantly longer to solve problems in `Phonological Alternation`, `Morpheme`, and `Pragmatics` subcategories. Notably, these subcategories also exhibit a relatively large performance gap compared to humans, as discussed in Section 5. This suggests that, unlike humans, who leverage real-world experience and intuition to efficiently solve problems, LLMs struggle in these subcategories.

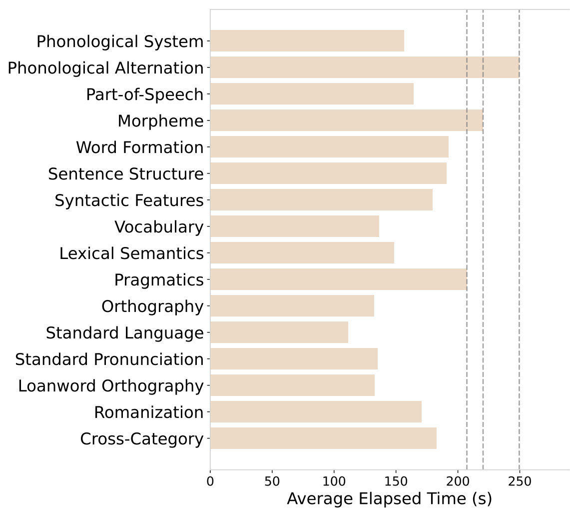


Figure 5: Comparison of the average *thinking* time required to solve each question and the corresponding average score for each subcategory by the s1-32B model. Dashed lines indicate the three longest test times.

6.2 Impact of Experiential Knowledge

Phonological Alternation To examine the effects of experiential knowledge in `Phonological Alternation` subcategory, we consider human subvocalization—a phenomenon in which individuals mentally rehearse the pronunciation of a word upon seeing it (Baddeley and Hitch, 1974; Smith and Dechant, 1961). To obtain the pronunciation of the text, we use g2pK[5] to convert Korean graphemes into phonemes. For example, given the text "오늘은 하늘이 맑습니다$_{\text{Today, the sky is clear}}$", g2pK generates "오느른 하느리 막씀니다$_{\text{To-day, the skai iz kli-er}}$". We append this produced *pronunciation text* to each `choice` in the prompt. As shown in Figure 4(a), adding *pronunciation text* improves performance by 3.1% to 26.3%. These findings demonstrate that incorporating pronunciation could significantly enhance the linguistic competence of LLMs in phonological tasks. However, this improvement does not conclusively determine whether LLMs possess latent pronunciation knowledge that remains underutilized or simply lack such knowledge, suggesting the need for further analysis in future work.

Morpheme In morphological processing, humans intuitively decompose words into morphemes, deriving meaning from root words (Berko, 1958; Taft and Forster, 1975). To test whether LLMs can similarly benefit from morpheme awareness, we use Kiwi[6] to decompose words into morphemes and append this *morpheme text* to each choice in the prompt. For instance, "오늘은 하늘이 맑습니다$_{\text{Today, the sky is clear}}$" is segmented as {오늘$_{\text{today}}$ / 은$_{\text{(auxiliary particle)}}$ / 하늘$_{\text{sky}}$ / 이$_{\text{is}}$ / 맑$_{\text{clear}}$ / 습니다$_{\text{(present progressive ending)}}$}. As shown in Figure 4(b), adding *morpheme text* improves performance by 7.1% to 20.0%. These results present that incorporating explicit morphological cues can significantly enhance the performance of LLMs in morphologically diverse tasks. These significant improvements suggest that LLMs still have room for further improvement in understanding Korean morphological knowledge.

Model	Faithful	Coherence	Fluency	Relevance
HC-HCX-003	0.80	0.86	0.98	0.92
Claude-3.5-Sonnet	0.92	0.96	1.00	1.00
GPT-4o	0.86	0.94	1.00	1.00
Total	0.86	0.92	0.99	0.97

Table 2: Qualitative analysis of the explanations generated by LLMs on the KoGEM Benchmark. Specifically, we evaluated the outputs of a top Korean-centric model, HyperClova-HCX-003 (for short HC-HCX-003), and two top English-centric models, excluding the test-time scaled model. Three Korean native speakers assessed the texts based on four criteria: Faithfulness, Coherence, Fluency, and Relevance. The evaluations produced a Fleiss' Kappa of 0.42, indicating moderate agreement.

7 Qualitative Evaluation of Generated Explanations

We extended the evaluation beyond multiple-choice accuracy by prompting models to generate textual explanations for their choices and conducting a qualitative analysis of these responses. Inspired by Fabbri et al. (2021); Liu et al. (2023); Elangovan et al. (2024), we adopted an evaluation framework with four metrics: 'Faithfulness (for short, Faithful)' is the factual accuracy of the generated statement. 'Coherence' is the logical consistency within the sentence. 'Fluency' is the grammatical and linguistic naturalness of the sentence. 'Relevance' is the degree of semantic alignment between the generated rationale and the given question.

Table 2 presents the averaged scores for each metric across three open-source models: HC-HCX-003, Claude-3.5-Sonnet, and GPT-4o. The results show that two English-centric models achieved perfect fluency and relevance (1.00), and very high scores in coherence (0.94 and 0.96). In contrast, HC-HCX-003 showed relatively lower performance, particularly in faithfulness (0.80) and coherence (0.86), despite maintaining strong fluency (0.98) and relevance (0.92). Overall, the models consistently produced relevant and coherent justifications. However, variations in faithfulness suggest that some models may still generate explanations that are linguistically well-formed but not entirely aligned with factual content. These findings highlight the importance of evaluating generated explanations through a multifaceted lens.

8 Related Work

Linguistic Competence in NLP To evaluate the linguistic competence of language models, previous studies have employed probing methods (Conneau et al., 2018; Hewitt and Manning, 2019; Tenney et al., 2019a,b; Pimentel et al., 2020), which assess the extent to which linguistic information is encoded in the hidden representations of pretrained models. While most of these studies have primarily focused on morphological and syntactic phenomena in English, more recent work has expanded the scope of evaluation to encompass additional

[5] https://github.com/Kyubyong/g2pK
[6] https://github.com/bab2min/Kiwi

linguistic dimensions, including semantics, phonology, and pragmatics (Beguš et al., 2025; Mahowald et al., 2024; Waldis et al., 2024).

Although probing methods provide insights into the presence of linguistic information in model representations, they are limited in explicitly evaluating the systematic grammatical competence of LLMs. Moreover, there is a lack of systematic approaches to assessing the linguistic competence of language models in Korean. To this end, we propose a granular and comprehensive benchmark to systematically evaluate the linguistic knowledge of LLMs in Korean.

Korean Grammatical Knowledge Evaluation Several benchmarks have been developed to evaluate Korean linguistic knowledge. For example, Son et al. (2024) focused on lexical knowledge, such as identifying Korean equivalents for loanwords. Kim et al. (2024a) conducted a general evaluation of Korean grammar, but their work lacked clarity regarding the scope of grammatical knowledge covered. Additionally, while Koo et al. (2022) and Yoon et al. (2023) examined specific types of grammatical errors in Korean, they focused on morphosyntactic and orthographic correctness, leaving broader aspects of linguistic competence unexplored. In contrast to these previous works, KoGEM offers a comprehensive and systematic evaluation for the grammatical knowledge of LLMs in Korean. We further provide insights into the linguistic competence of LLMs in Korean by comparing their performance with that of humans.

9 Conclusion

In this paper, we illuminated the distinctive strengths and weaknesses of LLMs in linguistic competence compared to humans through the lens of the Korean Grammar Evaluation BenchMark, KoGEM. We conducted extensive experiments on a wide range of LLMs and found that while they excel in tasks requiring straightforward memorization or syntactic rule application, they struggle with linguistic subcategories that demand intuitive reasoning and real-world knowledge. In contrast, humans demonstrated relatively superior performance in these challenging areas for LLMs, leveraging intuitive understanding and experiential knowledge from real-world contexts. Furthermore, our in-depth analysis revealed that integrating experiential knowledge can significantly enhance the linguistic competence of LLMs. These insights highlight the importance of targeted benchmarks like KoGEM for assessing and advancing the multifaceted linguistic competence of LLMs. By polishing every facet of linguistic competence in LLMs through fine-grained comparisons with humans, we provide valuable insights to advance LLMs and propel their journey toward genuine linguistic competence.

Limitations

Although our study has demonstrated its value in assessing Korean grammatical competence in both LLMs and humans, there are several limitations that could benefit from further examination. First, while KoGEM was constructed based on a fine-grained taxonomy of grammar, it is currently confined to the Korean language. Expanding our approach to other languages requires incorporating the unique linguistic knowledge of each language. However, since the core linguistic categories (i.e., phonology, morphology, syntax, and semantics) are universally applicable, we believe that our experimental pipeline can serve as a foundation for further investigations across different languages.

Second, we categorized grammatical knowledge based on theoretical linguistics and standardized prescriptive grammar. However, since prescriptive grammar, which dictates how language should be used, focuses on the correct use of language, there could be a gap between prescriptive grammar and actual language use by humans in the real world. To help bridge this gap, incorporating a descriptive grammar, which describes how language is used, is expected to be necessary in future works.

Lastly, to ensure a more accurate and reliable evaluation, we addressed two potential concerns. One concern is that the observed advantage in definitional tasks could stem from data contamination between our benchmark and the LLMs' training corpora. Given the lack of transparency in LLM training data, we cannot completely exclude the possibility of data contamination. To address this concern, we evaluated 27 diverse models to capture general trends across architectures and datasets. The consistent performance patterns suggest that the outcome is unlikely to result solely from data memorization. Another factor is prompt design, which can affect model behavior. As an initial design choice, we aligned the prompt with that presented to human participants to ensure consistency. Future work could explore whether alternative designs yield greater robustness or deeper insights.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.RS-2025-00517221 and No.RS-2024-00415812) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2024-00439328, Karma: Towards Knowledge Augmentation for Complex Reasoning (SW Starlab), No.RS-2024-00457882, AI Research Hub Project, and No.RS-2019-II190079, Artificial Intelligence Graduate School Program (Korea University)). This work was additionally supported by the 2024 BK21 Four Project "Research Challenger Program for Graduate Students in Interdisciplinary Research Groups" at the Center for Teaching and Learning Support, Korea University. The work of the author, Nayeon Kim, was also supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No.RS-2023-00271662). Lastly, We especially thank Youngje Lee, Associate Professor in the Department of Korean Language and Literature in Korea University, for his insightful comments and suggestions.

References

Samuel Amouyal, Aya Meltzer-Asscher, and Jonathan Berant. 2024. Large language models for psycholinguistic plausibility pretesting. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 166–181, St. Julian's, Malta. Association for Computational Linguistics.

Anthropic. 2024a. The claude 3 model family: Opus, sonnet, haiku.

Anthropic. 2024b. Claude 3.5 sonnet model card addendum.

Alan D. Baddeley and Graham Hitch. 1974. Working memory. In Gordon H. Bower, editor, *Working Memory*, volume 8 of *Psychology of Learning and Motivation*, pages 47–89. Academic Press.

Gašper Beguš, Maksymilian Dąbkowski, and Ryan Rhodes. 2025. Large linguistic models: Investigating llms' metalinguistic abilities. *Preprint*, arXiv:2305.00948.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Jean Berko. 1958. The child's learning of english morphology. *Word*, 14(2-3):150–177.

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2023. Prompting language models for linguistic structure. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6649–6663, Toronto, Canada. Association for Computational Linguistics.

David W Carroll. 1986. *Psychology of language.* Thomson Brooks/Cole Publishing Co.

Tyler Chang, Zhuowen Tu, and Benjamin Bergen. 2022. The geometry of multilingual language model representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 119–136, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie. 2019. Multi-source cross-lingual model transfer: Learning what to share. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3098–3112, Florence, Italy. Association for Computational Linguistics.

Sungdai Cho and John Whitman. 2020. *Korean: A linguistic introduction*. Cambridge University Press.

ChangSu Choi, Yongbin Jeong, Seoyoon Park, Inho Won, HyeonSeok Lim, SangMin Kim, Yejee Kang, Chanhyuk Yoon, Jaewan Park, Yiseul Lee, HyeJin Lee, Younggyun Hahm, Hansaem Kim, and KyungTae Lim. 2024. Optimizing language augmentation for multilingual large language models: A case study on Korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12514–12526, Torino, Italia. ELRA and ICCL.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. The MIT Press, Cambridge.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Aparna Elangovan, Ling Liu, Lei Xu, Sravan Babu Bodapati, and Dan Roth. 2024. ConSiDERS-the-human evaluation framework: Rethinking human evaluation for generative large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1137–1160, Bangkok, Thailand. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Judith Holler and Stephen C. Levinson. 2019. Multimodal language processing in human communication. *Trends in Cognitive Sciences*, 23(8):639–652.

HyperCLOVA X Team. 2024. Hyperclova x technical report. *Preprint*, arXiv:2404.01954.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *CoRR*, abs/2001.08361.

Changyoung Kim, Eunbin Lee, Seungjun Lee, Jieun Kim, Yerim Shin, Taerin Jo, Hyunchul Kang, Sugyeong Yang, Juseong Kim, Juhui Sim, Jin Wi, and K-Stat Research. 2020. *2020 National Language Awareness Survey*. National Institute of the Korean Language.

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.

Nayeon Kim, Jun-Hyung Park, Joon-Young Choi, Eojin Jeon, Youjin Kang, and SangKeun Lee. 2022. Break it down into BTS: Basic, tiniest subword units for Korean. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7007–7024, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sanghoon Kim, Dahyun Kim, Chanjun Park, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, Changbae Ahn, Seonghoon Yang, Sukyung Lee, Hyunbyung Park, Gyoungjin Gim, Mikyoung Cha, Hwalsuk Lee, and Sunghun Kim. 2024b. SOLAR 10.7B: Scaling large language models with simple yet effective depth upscaling. 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 23–35, Mexico City, Mexico. Association for Computational Linguistics.

Seungduk Kim, Seungtaek Choi, and Myeongho Jeong. 2024c. Efficient and effective vocabulary expansion towards multilingual large language models. *Preprint*, arXiv:2402.14714.

SungHo Kim, Juhyeong Park, Yeachan Kim, and SangKeun Lee. 2024d. KOMBO: Korean character representations based on the combination rules of subcharacters. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5102–5119, Bangkok, Thailand. Association for Computational Linguistics.

Seonmin Koo, Chanjun Park, Jaehyung Seo, Seungjun Lee, Hyeonseok Moon, Jungseob Lee, and Heuiseok Lim. 2022. K-nct: Korean neural grammatical error correction gold-standard test set using novel error type classification criteria. *IEEE Access*, 10:118167–118175.

Seungjun Lee, Yoonna Jang, Jeongwook Kim, Taemin Lee, and Heuiseok Lim. 2024. Kullm: Learning to construct korean instruction-following large language models. *Journal of KIISE*, 51:817–826.

Stephen C. Levinson. 1983. *Pragmatics*. Cambridge Textbooks in Linguistics. Cambridge University Press.

LG AI Research. 2024. Exaone 3.5: Series of large language models for real-world use cases. *arXiv preprint arXiv:https://arxiv.org/abs/2412.04862*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Llama Team. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

John Lyons. 1968. *Contents*, page v–viii. Cambridge University Press.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540.

William D Marslen-Wilson and Lorraine K Tyler. 2007. Morphology, language and the brain: the decompositional substrate for language comprehension. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1481):823–836.

James L. McClelland, Felix Hill, Maja Rudolph, Jason Baldridge, and Hinrich Schütze. 2019. Extending machine language models toward human-level language understanding. *CoRR*, abs/1912.05877.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *Preprint*, arXiv:2501.19393.

Kilim Nam, Eun Ha Hwang, Hyun Ju Song, and Eui Jeong Ahn. 2024. A study on the grammatical competence of generative ai: Focusing on korean morpho-syntactic features. *Korean Language Research*, 65(11):1–21.

OpenAI. 2023. GPT-3.5-turbo. `https://platform.openai.com/docs/models/gpt-3.5`.

OpenAI. 2024a. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2024b. Openai o1 system card.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc.

Jaehyung Seo, Hyeonseok Moon, Jaewook Lee, Sugyeong Eo, Chanjun Park, and Heuiseok Lim. 2023. CHEF in the language kitchen: A generative data augmentation leveraging Korean morpheme ingredients. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6014–6029, Singapore. Association for Computational Linguistics.

Henry P Smith and Emerald Dechant. 1961. *Psychology in teaching reading.* Prentice-Hall.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *Preprint*, arXiv:2408.03314.

Guijin Son, Hanwool Lee, Suwan Kim, Huiseo Kim, Jae cheol Lee, Je Won Yeom, Jihyu Jung, Jung woo Kim, and Songseong Kim. 2024. 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.

Marcus Taft and Kenneth I Forster. 1975. Lexical storage and retrieval of prefixed words. *Journal of Verbal Learning and Verbal Behavior*, 14(6):638–647.

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*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. What do you learn from context? probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations (ICLR)*, New Orleans, LA, USA.

Andreas Waldis, Yotam Perlitz, Leshem Choshen, Yufang Hou, and Iryna Gurevych. 2024. Holmes a benchmark to assess the linguistic competence of language models. *Transactions of the Association for Computational Linguistics*, 12:1616–1647.

Hetong Wang, Pasquale Minervini, and Edoardo Ponti. 2024. Probing the emergence of cross-lingual alignment during LLM training. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12159–12173, Bangkok, Thailand. Association for Computational Linguistics.

Lydia White. 1989. *Universal Grammar and Second Language Acquisition*. Language Acquisition and Language Disorders. John Benjamins Publishing Company, Amsterdam :.

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.

A Details of LLM Baselines

In this paper, we use the eight Korean-centric LLMs primarily trained on Korean data and the eleven English-centric LLMs primarily trained on English data for zero-shot evaluation on KoGEM as follows:

- **Korean-centric LLMs**

  1. Bllossom-8B (Choi et al., 2024): This is a Korean-English bilingual language model based on the open-source LLama3. It enhances the connection of knowledge between Korean and English.
  2. SOLAR-v1.0-10.7B-Instruct (Kim et al., 2024b): This is an advanced LLM with 10.7 billion parameters. It is trained by utilizing instruction fine-tuning methods, including supervised fine-tuning (SFT) and direct preference optimization (DPO) (Rafailov et al., 2023).
  3. KULLM-3-10.7B (Lee et al., 2024): This model is instruction-tuned from the `upstage/SOLAR-10.7B-v1.0` (Kim et al., 2024b) model. This model is trained using the data, such as Peng et al. (2023) and mixed Korean instruction data (gpt-generated, hand-crafted, etc).
  4. EEVE-v1.0-10.8B-Instruct (Kim et al., 2024c): This model is a fine-tuned version of `yanolja/EEVE-Korean-10.8B-v1.0`,[7] which is a Korean vocabulary-extended version of `upstage/SOLAR-10.7B-v1.0` (Kim et al., 2024b). Specifically, this model is trained by utilizing DPO.[8]
  5. EXAONE-3.5-7.8B-Instruct (LG AI Research, 2024): This model is an instruction-tuned bilingual (English and Korean) generative model, developed and released by LG AI Research. This model is trained to utilize the system prompt.[9]
  6. EXAONE-3.5-32B-Instruct (LG AI Research, 2024): This model shares the same architecture as above `EXAONE-3.5-7.8B-Instruct`, differing only in size.
  7. HyperCLOVA-HCX-DASH-001 (HyperCLOVA X Team, 2024): This is an optimized version of `HyperCLOVA-HCX-003` that offers faster response times and cost efficiency, making it suitable for simpler tasks while maintaining robust performance.
  8. HyperCLOVA-HCX-003 (HyperCLOVA X Team, 2024): This is a foundational model in NAVER's HyperCLOVA X suite, designed for complex and sophisticated tasks, delivering high-quality responses.

- **English-centric LLMs**

  1. Gemma-2-9B-Instruct (Team et al., 2024): Developed by Google, this model is part of the Gemma series and contains 9 billion parameters. It is a text-to-text, decoder-only large language model, with open weights for both pre-trained and instruction-tuned variants. Gemma models are well-suited for a variety of text generation tasks, including question answering, summarization, and reasoning.
  2. Gemma-2-27B-Instruct (Team et al., 2024): This model is the largest in Google's Gemma series, featuring 27 billion parameters. It is designed to deliver high performance across various natural language processing tasks, benefiting from its extensive parameter count and advanced training methodologies.
  3. Qwen2.5-7B-Instruct (Qwen Team, 2024): Developed by Alibaba, this model is part of the Qwen2.5 series and contains 7 billion parameters. It is designed to handle various natural language understanding and generation tasks, supporting multiple languages, including English and Chinese.
  4. Qwen2.5-14B-Instruct (Qwen Team, 2024): This model is a mid-sized variant in Alibaba's Qwen2.5 series, featuring 14 billion parameters. It offers enhanced performance in language understanding and generation tasks, with support for multiple languages and a context length of up to 128,000 tokens.
  5. Qwen2.5-32B-Instruct (Qwen Team, 2024): This model is part of Alibaba's Qwen2.5 series, featuring 32 billion parameters. It supports a context length of up to 128,000 tokens and is designed to handle complex tasks across multiple languages, including English and Chinese. All Qwen 2.5 series are open-source under the Apache 2.0 license.

[7] https://huggingface.co/yanolja/EEVE-Korean-10.8B-v1.0

[8] https://github.com/axolotl-ai-cloud/axolotl

[9] https://huggingface.co/LGAI-EXAONE/EXAONE-3.5-7.8B-Instruct

6. DeepSeek-R1-Distill-Qwen-14B (Liu et al., 2024): Developed by DeepSeek, this model is a distilled version of their R1 model, based on Qwen2.5-14B, containing 14 billion parameters. It has been fine-tuned using reasoning data generated by DeepSeek-R1, resulting in enhanced performance in reasoning tasks.
7. DeepSeek-R1-Distill-Qwen-32B (Liu et al., 2024): This model is another distilled variant from DeepSeek, based on Qwen2.5-32B, featuring 32 billion parameters. It has been fine-tuned with reasoning data from DeepSeek-R1, achieving state-of-the-art results in various benchmarks.
8. s1-32B (Muennighoff et al., 2025): Developed by Stanford and the University of Washington, s1-32B is a fine-tuned version of Qwen2.5-32B-Instruct. It was optimized for reasoning tasks using 1,000 high-quality samples. The model employs a novel "budget forcing" technique to enhance reasoning efficiency and outperforms OpenAI's o1-preview on certain benchmarks.
9. Llama-3.1-8B-Instruct (Llama Team, 2024): Developed by Meta AI, this model is part of the Llama-3.1 series and contains 8 billion parameters. It has been instruction-tuned to enhance its performance in various natural language understanding and generation tasks.
10. Llama-3-70B (Llama Team, 2024): Developed by Meta AI, Llama-3-70B is a large language model with 70 billion parameters. It has been pre-trained on approximately 15 trillion tokens from publicly available sources. The model is designed to be multilingual and multimodal, with enhanced capabilities in coding and reasoning.
11. Llama-3.1-405B (Llama Team, 2024): This is an expanded version of the Llama series, featuring 405 billion parameters. It demonstrates superior performance in general knowledge and reasoning tasks, achieving high scores on benchmarks such as MMLU-Pro and MMLU-redux.
12. Gemini-1.5-flash (Gemini Team, 2024): Developed by Google DeepMind, Gemini-1.5-flash is a multimodal language model capable of processing text, images, audio, and video. It is designed for real-time interactions and has been integrated into various Google products, including Bard and Pixel smartphones.
13. Gemini-2.0-flash-exp (Gemini Team, 2024): An experimental update to the Gemini series, this model offers improved speed and performance over its predecessors. It introduces features such as a Multimodal Live API for real-time audio and video interactions, enhanced spatial understanding, and integrated tool use, including Google Search.
14. Claude-3-haiku (Anthropic, 2024a): Developed by Anthropic, Claude-3-haiku is a large language model designed for complex conversational tasks. It features a context window of up to 200,000 tokens, allowing it to process extensive text sequences effectively.
15. Claude-3.5-Sonnet (Anthropic, 2024b): An enhanced version of the Claude series, this model offers improved performance in language understanding and generation tasks. It maintains a large context window and has been fine-tuned for better alignment with human preferences.
16. GPT-3.5-turbo (OpenAI, 2023): Developed by OpenAI, GPT-3.5-turbo is an improvement over the original GPT-3.5 model, offering better accuracy in responses. It has been widely used in applications requiring natural language understanding and generation.
17. GPT-4o-mini (OpenAI, 2024a): A smaller and more cost-effective version of OpenAI's GPT-4o, this model is capable of processing text, images, and audio. It offers rapid response times and has been integrated into various applications for real-time interactions.
18. GPT-4o (OpenAI, 2024a): OpenAI's GPT-4o is a multimodal model capable of analyzing and generating text, images, and sound. It exhibits rapid response times comparable to human reactions and has enhanced performance in non-English languages.
19. o1-preview (OpenAI, 2024b): Introduced by OpenAI, o1-preview is designed to solve complex problems by spending more time "thinking" before responding. It outperforms previous models in areas like competitive programming, mathematics, and scientific reasoning.

B Details of LLM Evaluation

In this section, we explain the details of LLM evaluation settings, such as device settings, hyperparameters, and prompt design, as the input for a single run of zero-shot evaluation.

B.1 Device Settings

We use the Hugging Face library[10] with local GPU settings for open-source models, including all Korean-centric LLMs except for the HyperCLOVA X series, the Gemma series, the Qwen2.5 series, the DeepSeek-R1-Distill series, and the Llama-3.1-8B model among English-centric LLMs. For models larger than 20B parameters, we utilize two RTX A6000 GPUs, while for relatively smaller models, we use a single RTX A6000 GPU. For closed-source models, we use the corresponding API provided for each model.

B.2 Hyperparameters

To ensure reproducibility, we set the temperature to 0 or very close to 0 (e.g., 1e-10). However, for GPT-3.5-turbo and o1-preview, we cannot customize the temperature setting. This is because the GPT-3.5-turbo model does not support temperature adjustment (OpenAI, 2023), and the function is unavailable in o1-preview. Additionally, we fix the random seed to 42 across all experiments.

To manage API usage costs, we limit the maximum number of output tokens to 200. Despite this restriction, we confirm that no outputs are truncated, as the input prompts explicitly set a maximum output length of 100 characters.

B.3 Prompt Designs for Main Results

In this section, we describe the prompts used in our main experiments. The prompts are categorized into four types (T1, T2, T3, and T4) based on the given information, which may include `passage`, `question`, `paragraph`, `choices`. Additionally, to facilitate understanding, we provide the English translation of the full T1 prompt at the end of the examples below.

B.4 Statistics of Prompt Lengths

Figure 6 summarizes the average lengths and instance counts by prompt type. On average, each prompt, including (passage, question, paragraph, and choices), contains 534 characters, with many exceeding 1,000 characters due to full-passage inclusion. This structure effectively evaluates prompts of varying lengths, requiring integration of broad linguistic context.

[10] `https://huggingface.co/models`

- **T1** {`question+choices`}

> [system]
> 다음은 한국어 언어 이해에 대한 객관식 문제입니다. 주어진 질문에 대한 정답으로 올바른 번호를 선택지에서 고르고, 그에 맞는 해설을 100자 내로 설명하시오.
>
> [user]
> 질문: 다음 선택지 1 부터 {4 or 5} 중 {`question`}
> 선택지: {`choices`}

- **T2** {`question+paragraph+choices`}

> [system]
> 다음은 한국어 언어 이해에 대한 객관식 문제입니다. 주어진 설명을 보고, 질문에 대한 정답으로 올바른 번호를 선택지에서 고르고, 그에 맞는 해설을 100자 내로 설명하시오.
>
> [user]
> 설명: {`paragraph`}
> 질문: 다음 선택지 1 부터 {4 or 5} 중 {`question`}
> 선택지: {`choices`}

- **T3** {`passage+question+choices`}

> [system]
> 다음은 한국어 언어 이해에 대한 객관식 문제입니다. 주어진 지문을 보고, 질문에 대한 정답으로 올바른 번호를 선택지에서 고르고, 그에 맞는 해설을 100자 내로 설명하시오.
>
> [user]
> 지문: {`passage`}
> 질문: 다음 선택지 1 부터 {4 or 5} 중 {`question`}
> 선택지: {`choices`}

- **T4** {`passage+question+paragraph+choices`}

> [system]
> 다음은 한국어 언어 이해에 대한 객관식 문제입니다. 주어진 지문과 설명을 보고, 질문에 대한 정답으로 올바른 번호를 선택지에서 고르고, 그에 맞는 해설을 100자 내로 설명하시오.
>
> [user]
> 지문: {`passage`}
> 설명: {`paragraph`}
> 질문: 다음 선택지 1 부터 {4 or 5} 중 {`question`}
> 선택지: {`choices`}

- Translation of **T4**

> [system]
> The following is a multiple-choice question about Korean language comprehension. Based on the given passage and explanation, choose the correct number from the choices as the answer to the question and provide a corresponding explanation within 100 characters.
>
> [user]
> Passage: {`passage`}
> Paragraph: {`paragraph`}
> Question: Choose from options 1 to {4 or 5} for {`question`}
> Choices: {`choices`}

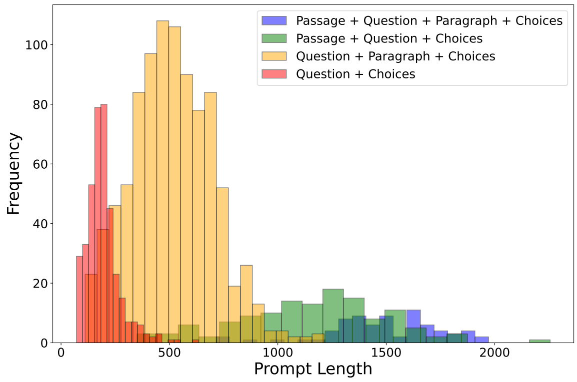


Figure 6: Distribution of the KoGEM benchmark by prompt type. These four types correspond to T1–T4 described earlier in Section B.3.

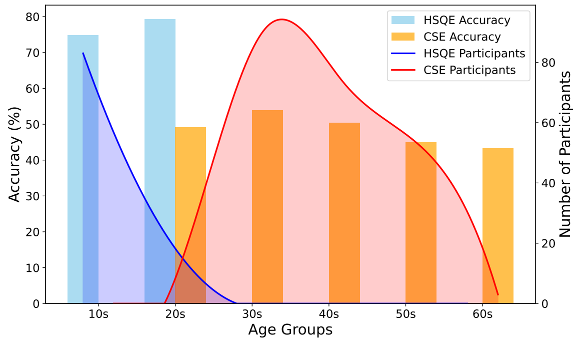


Figure 7: Age distribution and accuracy of participants in HSQE and CSE tests. The line graph illustrates the number of participants across different age groups, while the bar graph represents their corresponding accuracy rates.

B.5 Prompt Design for Addition of Pronunciation Experiment

In Section 5, we conducted additional experiments on the phonological alternation task by incorporating pronunciation information into the choices within the prompt. Figure 12 represents a typical case from this experiment.

C Details of Crowdsourcing for Human Evaluation

In this section, we provide details about the crowdsourcing process used to evaluate human performance. As mentioned in Section 4.2, we assess human performance across five exams: HSQE, LCSE (G9 and G7), and NCSE (G9 and G7), using crowdsourcing. Figure 10 presents the instructions provided to crowdworkers, while Figure 11 shows an actual test sample given to participants. We collected over 10 responses per question from a total of 352 participants across 583 questions. To maintain concentration and prevent crowdworker bias, we assign a maximum of 20 questions per crowdworker. The cost per response was approximately \$0.24.

For the CSE, responses were collected from a diverse age group of active civil servants, ranging from their teens to their 60s, to mitigate potential biases associated with crowdworkers. In contrast, the HSQE focused on first-year university students who had completed the relevant curriculum within the past year. The age distribution of participants and their accuracy across age groups are presented in Figure 7. Additionally, to minimize gender bias, we made efforts to balance the gender distribution as evenly as possible.

D Details of KoGEM

In this section, we provide detailed descriptions of KoGEM, including source data, preprocessing rules, and subcategories of Korean grammar taxonomy.

D.1 Source Data

We detail the characteristics of each source data in KoGEM, and their copyright and license.

Detailed Description

- College Scholastic Ability Test (CSAT) is an exam administered by the Korea Institute for Curriculum and Evaluation to select qualified individuals for university admission. CSAT is administered once a year.[11]

- National United Achievement Test (NUAT) is a mock exam conducted in a format similar to the CSAT. NUAT is administered by the Seoul Education Research and Information Institute and is regularly taken by high school students across all grades. NUAT is conducted four times a year for first- and second-year high school students, and six times a year for third-year students.[12]

- High School Qualification Exam (HSQE) is an exam administered by the Korea Institute for Curriculum and Evaluation to assess the qualifications required for high school graduation. We solely use the Korean language section of the High School Qualification Exam. HSQE is held twice a year.[13]

[11] https://www.suneung.re.kr
[12] https://www.jinhak.or.kr
[13] https://www.kice.re.kr

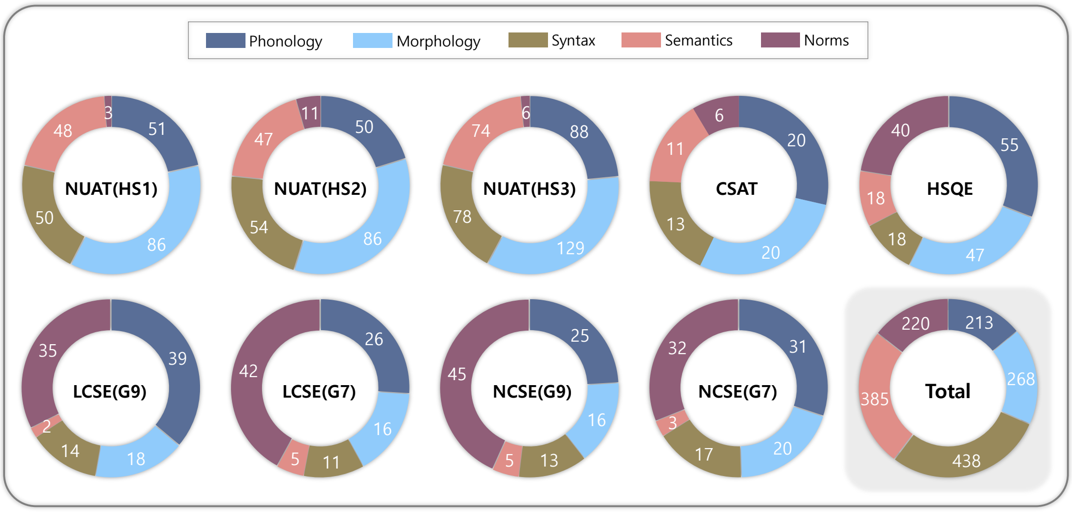


Figure 8: The specific numbers and ratios of linguistic categories and subcategories for each source exam.

- Civil Service Exam (CSE) is an exam administered by the Ministry of Personnel Management to recruit national and local civil servants. We extract grammar questions from the Korean language section of the national and local 7th- and 9th-grade recruitment exams. Although there are more types of CSEs, we use exams for which copyright issues have been resolved. CSE is conducted once a year.[14]

License

The institutions providing our source exams have made these works available under the "Public Copyright Free Use Permit Standard (Korean Open Government License, KOGL) Type 1." Under KOGL Type 1, users can use public works freely and without fee, regardless of their commercial use, and can change or modify to create secondary works.[15] Additionally, we have contacted each institution and obtained permission for research purposes.

D.2 Data Preprocessing

Since we extract textual data from exam images using HyperCLOVA X OCR [16], the extracted data often contain numerous errors, such as typos, fragmented characters, and unstructured or inconsistent formats. To standardize the structure and representation of the data, we utilize HTML formatting. Below are the formatting rules we follow:

- **Underline.** Underline: We enclose underlined text within the `<u>` tag. If the underlined text includes a specific symbol, such as ㉠ we wrap the text using the corresponding symbol tag, e.g., `<㉠>`.
- **Boldface.** Text requiring emphasis is enclosed within the `<b>` tag for bold formatting.
- **Box.** Text to be highlighted within a box is wrapped using the `<box>` tag.
- **Section.** Text belonging to the same section is grouped using a designated section tag. Examples include `<가>`, `<A>`, `<예>`, `<보기>`, etc.
- **Table.** Rows are separated by the newline character (\n). Within a row, the header column is distinguished by a colon (:), and other columns are separated using a slash (/).
- **Abbreviated Expression.** When more than three dots appear in the text, we standardize them to three unified dots (···).
- **Bullet Point.** We preserve the original bullet points described in the image whenever possible. However, for symbols with similar appearances (e.g., hollow middle circles or lower hollow circles), we assign a single representative symbol for uniformity.
- **Text with Circle.** If a piece of text is circled, we represent it by enclosing it in parentheses.
- **Text with Rectangular.** If text is enclosed within a rectangle, we represent it using square brackets.

[14] https://www.gosi.kr
[15] https://www.kogl.or.kr/info/license.do
[16] https://clova.ai/hyperclova

Language	Model	NUAT (HS1)	NUAT (HS2)	NUAT (HS3)	CSAT	HSQE	LCSE (G9)	LCSE (G7)	NCSE (G9)	NCSE (G7)	Avg.
Random		20.00	20.00	20.00	20.00	25.00	25.00	25.00	25.00	25.00	21.95
Korean	Bllossom-8B	27.31	24.60	25.07	20.00	35.96	25.93	29.00	31.73	24.27	27.10
	SOLAR-v1.0-10.7B-Instruct	33.61	24.19	24.80	22.86	33.71	19.44	29.00	27.88	28.16	27.36
	KULLM-3-10.7B	26.89	26.21	21.07	25.71	32.58	32.41	24.00	31.73	29.13	26.64
	EEVE-v1.0-10.8B-Instruct	31.09	24.60	25.60	22.86	44.38	36.11	39.00	28.85	26.21	30.25
	EXAONE-3.5-7.8B-Instruct	30.25	27.02	30.93	28.57	47.19	43.52	38.00	35.58	30.10	33.60
	EXAONE-3.5-32B-Instruct	40.76	32.26	33.07	28.57	57.87	41.67	44.00	41.35	36.89	38.98
	HyperCLOVA-HCX-DASH-001	29.41	21.37	26.40	27.14	46.07	36.11	29.00	40.38	34.95	30.77
	HyperCLOVA-HCX-003	48.74	38.31	35.73	27.14	64.61	49.07	46.00	56.73	41.75	44.62
English	Gemma-2-9B-Instruct	34.87	27.42	30.40	21.43	49.44	36.11	32.00	26.92	30.10	32.68
	Gemma-2-27B-Instruct	39.50	29.03	35.20	20.00	54.49	30.56	31.00	38.46	30.10	35.70
	Qwen2.5-7B-Instruct	37.82	32.26	28.27	27.14	47.19	32.41	26.00	29.81	34.95	33.27
	Qwen2.5-14B-Instruct	46.22	37.50	34.40	31.43	60.67	30.56	38.00	43.27	33.98	40.22
	Qwen2.5-32B-Instruct	48.32	41.53	39.73	37.14	58.43	45.37	37.00	39.42	31.07	43.04
	DeepSeek-R1-Distill-Qwen-14B	45.80	34.27	34.13	20.00	55.62	37.04	32.00	33.65	32.04	37.73
	DeepSeek-R1-Distill-Qwen-32B	50.42	42.34	42.67	35.71	65.17	41.67	35.00	35.58	34.95	44.55
	s1-32B	52.94	48.39	43.73	47.14	67.98	42.59	37.00	47.12	34.95	48.03
	Llama-3.1-8B-Instruct	30.67	25.40	21.60	31.43	37.08	24.07	31.00	29.81	27.18	27.62
	Llama-3-70B	39.50	29.84	29.87	30.00	49.44	35.19	36.00	28.85	33.01	34.58
	Llama-3.1-405B	57.98	43.55	46.93	48.57	60.67	37.96	38.00	39.42	33.98	47.18
	Gemini-1.5-flash	49.58	43.95	43.20	25.71	65.73	39.81	47.00	42.31	32.04	45.34
	Gemini-2.0-flash-exp	63.45	49.19	56.00	51.43	76.40	56.48	43.00	52.88	38.83	56.04
	Claude-3-haiku	39.92	27.42	28.27	31.43	56.74	34.26	35.00	31.73	34.95	34.97
	Claude-3.5-Sonnet	66.81	60.48	60.53	52.86	75.28	52.78	45.00	52.88	48.54	59.97
	GPT-3.5-turbo	25.63	26.61	22.93	24.29	40.45	30.56	31.00	25.00	23.30	27.30
	GPT-4o-mini	44.12	32.26	33.87	37.14	67.98	38.89	39.00	39.42	27.18	39.96
	GPT-4o	63.45	54.84	53.87	45.71	79.21	55.56	51.00	55.77	49.51	57.87
	o1-preview	**86.97**	**83.87**	**79.47**	**75.71**	**94.38**	**78.70**	**76.00**	**74.04**	**61.17**	**81.04**
LLMs Avg.		44.15	36.62	36.58	33.23	56.47	39.44	37.70	39.28	34.20	40.24
Human		72.66	61.58	72.45	65.88	78.60	53.80	48.07	54.91	43.42	63.04

Table 3: Zero-shot evaluation results for each source examination. NUAT HS1, HS2, and HS3 refer to high school 1st, 2nd, and 3rd grades, while G9 and G7 in CSE denote 9th and 7th grades, respectively.

Exam	years	No. QA pairs
CSAT	1994-2023	70
NUAT	2006-2024	861
QE	2001-2024	178
CSE	2006-2023	415
Total		1,524

Table 4: The source data collection years.

D.3 Data Categorization

We classify the preprocessed data into one of 16 subcategories by three Korean language majors. Each annotator independently classifies each question, first identifying the main linguistic category and then assigning the appropriate subcategory. Labels are finalized by majority vote, and disagreements are resolved through discussion. We remove questions requiring knowledge from more than three categories to focus on the evaluation of each linguistic subcategory. Considering the minimum hourly wage, we compensated data annotators approximately \$0.15 per question.

D.4 Description for Subcategories

Table 7 presents a comprehensive breakdown of the linguistic categories, their subcategories, and corresponding descriptions. The framework is organized into five categories: `Phonology`, `Morphology`, `Syntax`, `Semantics`, and `Norms`. Each main category is further divided into subcategories.

First, the Phonology category encompasses the `Phonological System`, which refers to the system of Korean phonemes, and `Phonological Alternation`, which describes phenomena such as the insertion, deletion, and replacement of phonemes in specific environments. An example

Model	Class 1 (%)	Class 2 (%)	Class 3 (%)	Class 4 (%)	Class 5 (%)
HyperClova-HCX-003	14.3	24.7	29.1	21.50	10.3
Claude-3.5-Sonnet	19	25.5	23.9	21.7	9.9
GPT-4o	16.2	24	25.1	22.00	12.7
Gold Label	22.4	23.3	21.5	21.00	11.8

Table 5: Class distribution of model predictions on the KoGEM benchmark. Each row shows the percentage of predictions falling into each class for the respective model. The gold label row represents the actual distribution of labels in the dataset, providing a reference for comparison.

of phonological replacement can be found in the word '국물$_{\text{soup}}$.' Specifically, while '국물$_{\text{soup}}$' is composed of the syllables '국/guk/' and '물/mul/', its pronunciation is realized as [gung.mul], not /guk.mul/ due to nasalization. It is common for certain phonemes to change or be deleted depending on their phonological environment.

Second, the Morphology category covers aspects such as `Part-of-Speech`, `Morphemes`, and `Word Formation`. Unlike part-of-speech classification in English, which involves eight parts of speech, Korean has nine parts of speech (Table 7). Additionally, the questions in the `Morpheme` subcategory require knowledge of the concepts and types of morphemes, such as lexical morphemes that carry core meanings (e.g., '꽃$_{\text{flower}}$', '해$_{\text{sun}}$') and functional morphemes that serve grammatical functions (e.g., '의$_{\text{of}}$', '고$_{\text{and}}$'). To solve the `Word Formation` task, both models and humans should understand Korean word formation rules, such as how compound words (e.g., '식기+세척기$_{\text{dish+washer}}$') and derived words (e.g., '씻기$_{\text{washing}}$' consists of '씻-$_{\text{wash}}$' + '-기$_{\text{-ing}}$') are formed. They should be able to identify the morphemes that make up each word.

Third, the Syntax category focuses on the syntactic structures of Korean sentences. We further include knowledge of syntactic features, which span grammatical components of sentences, such as tense (e.g., prefix ending '겠$_{\text{future tense}}$') and negative adverbs (e.g., '안$_{\text{not}}$' and '못$_{\text{unable}}$'). The Semantics category examines the meaning of words, sentences, and the context of conversations. In the `Vocabulary` subcategory, we attempt to evaluate rote knowledge of words, which refers to their dictionary definitions. The questions in the `Lexical Semantics` subcategory assess the ability to identify word relationships, such as the connection between 'orange' and 'fruit'.

Lastly, the Norm category includes subcategories such as `Orthography`, `Standard Pronunciation`, `Loanword Orthography`, and `Romanization`, providing rules for accurate language use. `Cross-Category` addresses interdisciplinary topics that involve multiple subcategories within the Norms category. For instance, a question such as *"Which one is correctly written according to Korean orthography and standard language rules?"* is included in `Cross-Category`.

We further provide specific example QAs for the 16 subcategories of KoGEM from Figure 12 to Figure 16. This serves as a detailed reference for understanding the linguistic elements assessed in KoGEM, highlighting its comprehensive and systematic design.

E Results per Data Source

Considering the varying levels of difficulty across the different exams from which our benchmark was derived, we conducted separate evaluations of LLMs' and humans' performance for each source dataset. The results of these evaluations are presented in Table 3, and the category distribution by source is shown in Figure 8.

Distribution Among the nine source datasets, NUAT and CSAT contain a relatively high proportion of phonology, morphology, and syntax questions, with a minimal proportion of norms questions. In contrast, CSEs feature a notably larger share of norms and semantics questions. This variation in question type is also reflected in the results: in particular, humans scored higher on NUAT and CSAT than on CSEs. These findings suggest that humans outperform LLMs in phonology, morphology, and syntax tasks, which require empirical knowledge, reasoning, and intuition, compared to vocabulary and rigid norms that rely more on memorization.

Difficulty As illustrated in Figure 8, NUATs generally exhibit a consistent distribution of question

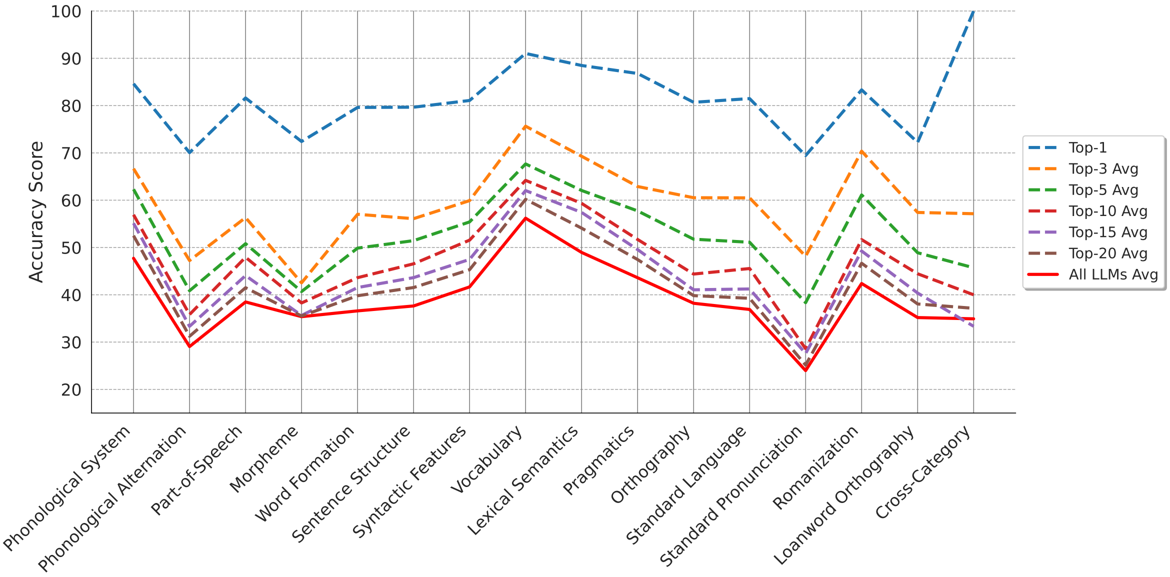


Figure 9: Accuracy scores across subcategories for different Top-k average settings and overall LLM performance.

types, but their difficulty increases as students progress through grades. This pattern is reflected in LLM evaluation results: models performed best on the first-year high school mock exam but showed a steady decline in performance with advancing grades, reaching their lowest scores on CSAT. Additionally, a comparable trend was observed in CSEs. While the types of questions remained broadly consistent, the 9th-grade exam is regarded as easier, whereas the 7th-grade exam is notably more challenging, leading to a corresponding decrease in LLM performance. Conversely, on HSQE, which requires comparatively less intensive study, most LLMs achieved their highest scores, further highlighting the correlation between exam difficulty and LLM performance.

F Top-k Performance Variation Across Subcategories

Figure 9 displays the zero-shot accuracy scores across a range of subcategories under various Top-k averaging conditions (Top-1 through Top-20), along with the overall average across all LLMs. While the absolute accuracy values vary with different Top-k settings, the relative performance trends among the linguistic subcategories remain largely consistent. This stability in ranking across Top-k variants justifies the use of the all LLMs average score as a representative metric in the main analysis in Figure 3. It allows us to abstract away from Top-k variance while retaining the comparative insight across categories.

G Comparison of Predicted and Gold Class Distributions on KoGEM

We analyze the distribution of predicted classes produced by each model on the KoGEM benchmark to examine potential discrepancies or biases in class prediction frequencies. Table 5 summarizes the relative frequencies of predictions for five predefined classes by three LLMs, such as HyperClova-HCX-003, Claude-3.5-Sonnet, and GPT-4o, along with the ground truth class distribution.

The results reveal distinct distributional tendencies across models. For instance, HyperClova-HCX-003 exhibits a higher frequency for Class 3 (29.1%) and a lower frequency for Class 1 (14.3%) compared to the gold distribution. Claude-3.5-Sonnet and GPT-4o produce more balanced distributions, though certain classes, such as Class 5, show slight over- or under-predictions.

Despite these model-specific differences, the overall predicted distributions remain relatively close to the gold label distribution. This indicates that the models are generally able to approximate the class frequencies present in the dataset, suggesting no severe class imbalance or systematic deviation in their outputs.

H Few-shot Evaluation of KoGEM

To explore the effect of exemplar-based context in few-shot learning, we conducted additional 1-shot and 5-shot evaluation experiments. We used three Korean-centric LLMs, such as EEVE-

Language	Model		Phonology	Morphology	Syntax	Semantics	Norm	Avg.	Δ
Korean	EEVE-Instruct-10.8B-v1.0	0-shot	22.54	27.24	27.85	40.78	27.73	30.25	-
		1-shot	21.23	29.70	26.96	38.52	29.95	29.97	-0.28
		5-shot	23.15	30.23	25.48	37.95	24.64	29.00	-1.25
	EXAONE-3.5-7.8B-Instruct	0-shot	24.88	30.22	32.19	43.64	31.36	33.60	-
		1-shot	28.30	31.20	32.41	43.06	33.64	34.49	0.89
		5-shot	27.72	30.87	33.84	41.11	30.82	33.86	0.26
	EXAONE-3.5-32B-Instruct	0-shot	27.23	37.31	36.30	50.65	37.27	38.98	-
		1-shot	29.58	37.07	37.03	50.97	40.87	39.97	0.99
		5-shot	28.70	37.13	38.44	52.44	41.13	40.52	1.54
English	Qwen2.5-14B-Instruct	0-shot	29.58	38.43	39.27	52.21	33.64	40.22	-
		1-shot	26.42	36.84	37.10	53.30	34.56	39.26	-0.96
		5-shot	27.09	40.70	37.98	51.94	32.37	39.61	-0.61
	Qwen2.5-32B-Instruct	0-shot	26.29	36.19	41.78	62.86	35.45	43.04	-
		1-shot	28.30	38.72	43.55	63.85	33.18	44.16	1.12
		5-shot	30.05	39.53	47.36	64.72	32.85	45.78	2.74
	Claude-3.5-Sonnet	0-shot	47.42	52.61	64.38	74.55	46.82	59.97	-
		1-shot	48.45	54.50	65.22	75.12	48.58	60.99	1.02
		5-shot	49.01	56.37	66.99	75.54	49.51	62.23	2.26
	GPT-4o	0-shot	44.60	51.49	55.48	71.95	58.64	57.87	-
		1-shot	45.02	53.56	59.30	73.56	58.26	59.75	1.88
		5-shot	51.24	50.20	61.35	74.03	59.91	60.94	3.07

Table 6: Few-shot accuracy evaluation results on our KoGEM benchmark. It consists of two segments: Korean-centric LLMs trained mainly on Korean data, and English-centric LLMs trained primarily on English data, respectively. Δ denotes the difference between 0-shot and n-shot accuracy.

v1.0-10.8B-Instruct, EXAONE-3.5-7.8B-Instruct, and EXAONE-3.5-32B-Instruct, and four English-centric LLMs, such as Qwen2.5-14B-Instruct, Qwen2.5-32B-Instruct, Claude-3.5-Sonnet, and GPT-4o. As shown in Table 6, the results showed that LLMs with at least 32B parameters exhibited notable improvements, indicating that exemplar-based context enhances performance on tasks requiring nuanced understanding. In contrast, smaller models struggled to effectively utilize the provided examples.

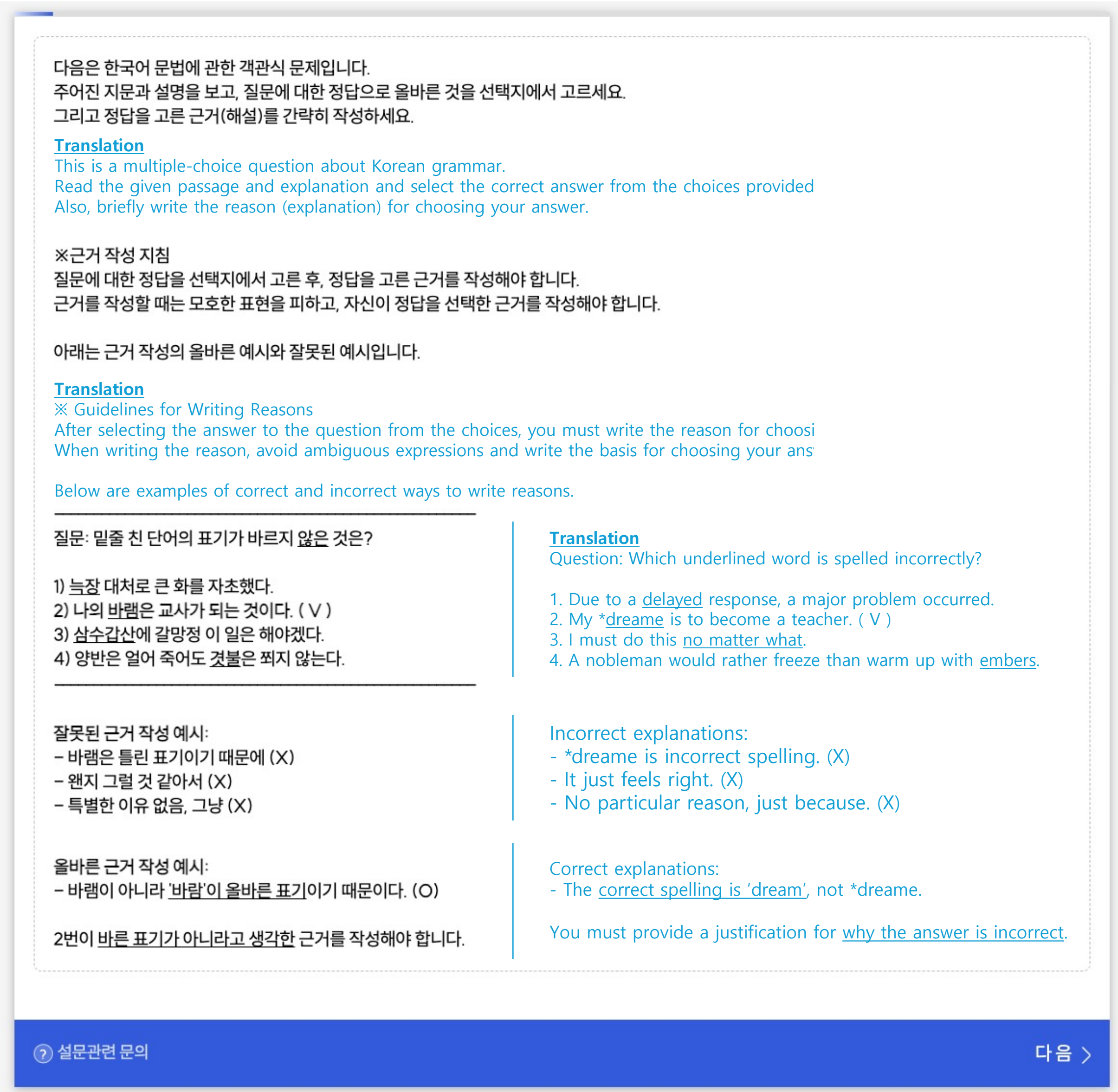


Figure 10: The instruction provided to crowdworkers for human evaluation outlines the evaluation process, task requirements, and criteria for rejecting invalid responses as incorrect. We have added English translation scripts in blue font to clarify the meaning. However, these translation scripts were not provided to the crowdworkers.

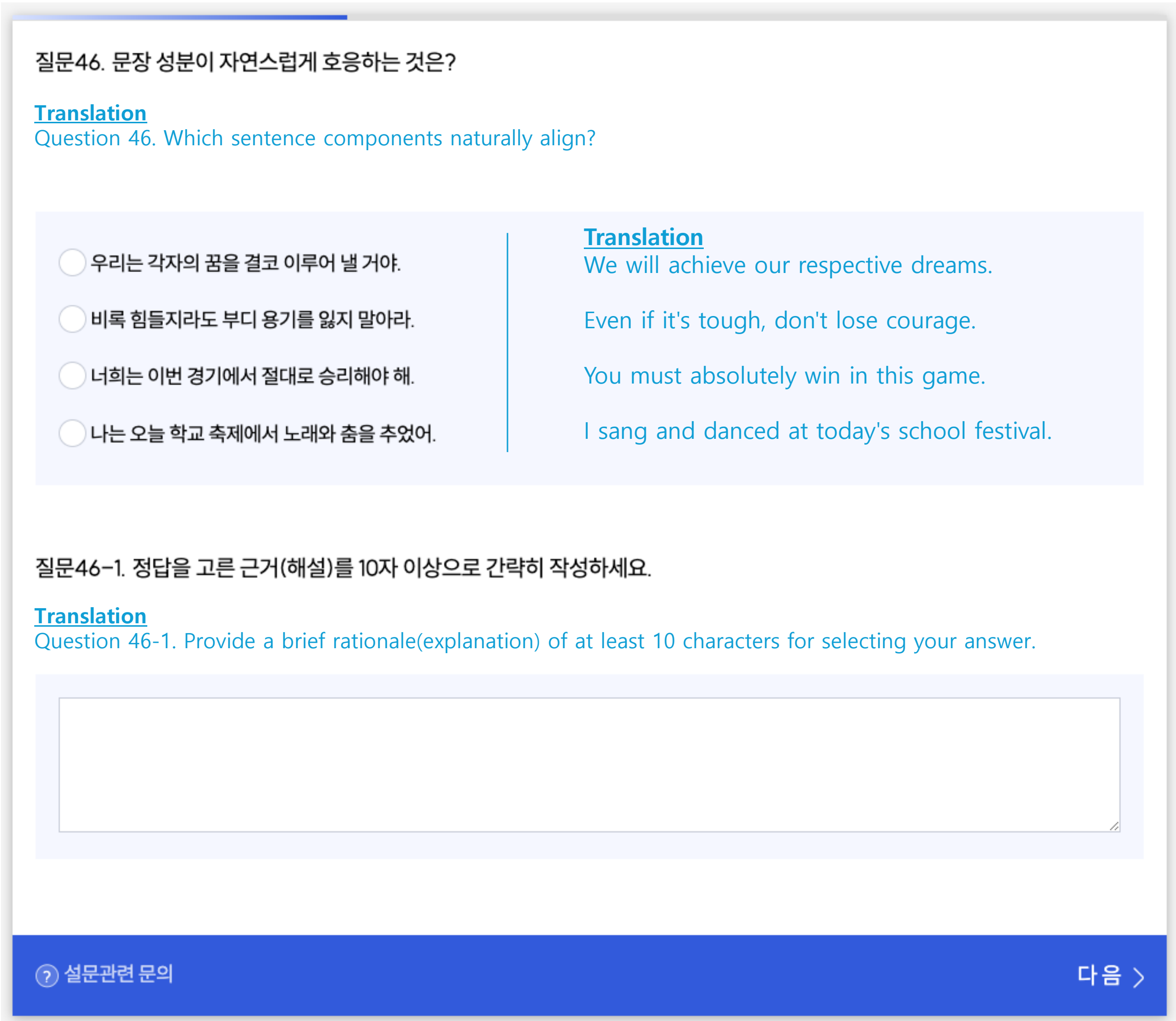


Figure 11: An example presented to participants during the human evaluation process illustrates the format and content of the exams. Specifically, it includes only `question` and `choices`, corresponding to T4 in Section B.3. We have added English translation scripts in blue font to clarify the meaning. However, these translation scripts were not provided to the crowdworkers.

Linguistic Category	Subcategory	Description
Phonology	Phonological System (PHS)	The system of Korean consonants and vowels and the principles of syllable formation.
	Phonological Alternation (PHA)	Phonological insertion, deletion, replacement, and combination phenomena in Korean.
Morphology	Part-of-Speech (POS)	The nine parts of speech in Korean (i.e., noun, pronoun, numeral, verb, adjective, determiner, adverb, particle, exclamation).
	Morpheme (MOR)	The concept and types of morphemes (e.g., lexical morphemes, functional morphemes).
	Word Formation (WOF)	Word formation process such as compounding and derivation.
Syntax	Sentence Structure (STS)	The sentence structure and the components of a sentence (e.g., subject, predicate, object, SVO word order, etc.).
	Syntactic Features (STF)	Various syntactic elements in sentences (e.g., tense, particles, negation, voice, honorifics, etc.).
Semantic	Vocabulary (VOC)	Rote knowledge of the dictionary definition of words.
	Lexical Semantics (LES)	Understanding semantic relationships of words (e.g., homonym, antonym, polysemy, etc.).
	Pragmatics (PRA)	Understanding discourse (e.g., context, reference, intentions of speakers, etc.).
Norms	Orthography (ORT)	Accurate rules of Korean spelling.
	Standard Language (STL)	Accurate knowledge of standard Korean.
	Standard Pronunciation (STP)	Accurate rules of Korean pronunciation.
	Loanword Orthography (LWO)	Korean orthography rules for loanwords.
	Romanization (ROM)	Rules for transcribing Korean into Roman script.
	Cross-Category (CRC)	Questions asking about two or more subcategories within the Norms category.

Table 7: Description of each subcategory. The three characters (e.g., PHS, PHA) in parentheses next to each subcategory represent its abbreviation.

Language	Type	Model	Phonology		Morphology			Syntax		Semantics			Norms						Total
			PHS	PHA	POS	MOR	WOF	STS	STF	VOC	LES	PRA	ORT	STL	STP	LWO	ROM	CRC	
Korean	Open	llama-3-Korean-Bllossom-8B	42.31	21.93	25.29	31.03	23.68	19.31	27.70	37.72	30.91	35.85	33.33	29.63	22.22	27.78	22.22	14.29	27.10
		SOLAR-10.7B-Instruct-v1.0	30.77	24.06	24.14	31.03	25.66	24.83	30.41	38.32	28.48	20.75	31.58	14.81	16.67	16.67	33.33	14.29	27.36
		KULLM3	15.38	22.46	31.03	31.03	23.03	25.86	26.35	34.13	26.67	22.64	32.46	29.63	13.89	33.33	22.22	28.57	26.64
		EEVE-Korean-Instruct-10.8B-v1.0	26.92	21.93	26.44	37.93	25.66	26.90	29.73	50.30	32.12	37.74	29.82	25.93	19.44	22.22	33.33	42.86	30.25
		EXAONE-3.5-7.8B-Instruct	38.46	22.99	34.48	31.03	27.63	28.97	38.51	50.30	38.79	37.74	35.09	29.63	22.22	38.89	22.22	28.57	33.60
		EXAONE-3.5-32B-Instruct	38.46	25.67	35.63	<u>51.72</u>	35.53	32.76	43.24	58.08	46.06	41.51	37.72	44.44	25.00	50.00	33.33	42.86	38.98
	Closed	HyperClova-HCX-DASH-001	46.15	20.86	32.18	31.03	30.92	26.90	22.97	44.91	36.97	32.08	35.09	37.04	25.00	22.22	22.22	28.57	30.77
		HyperClova-HCX-003	46.15	30.48	48.28	31.03	40.13	38.97	45.27	64.07	53.94	32.08	52.63	66.67	25.00	50.00	33.33	57.14	44.62
English	Open	Gemma-2-9B-Instruct	46.15	21.39	22.99	34.48	35.53	30.69	31.76	52.10	35.15	32.08	29.82	25.93	19.44	44.44	22.22	57.14	32.68
		Gemma-2-27B-Instruct	34.62	26.20	31.03	24.14	29.61	37.24	35.81	50.30	43.03	50.94	29.82	25.93	13.89	38.89	44.44	42.86	35.70
		Qwen2.5-7B-Instruct	42.31	20.32	29.89	37.93	30.26	31.03	39.86	52.10	38.18	39.62	30.70	25.93	11.11	22.22	16.67	28.57	33.27
		Qwen2.5-14B-Instruct	53.85	26.20	36.78	48.28	37.50	38.28	41.22	54.49	50.30	50.94	37.72	22.22	19.44	38.89	38.89	57.14	40.22
		Qwen2.5-32B-Instruct	42.31	24.06	44.83	24.14	33.55	42.41	40.54	64.67	64.85	50.94	38.60	37.04	16.67	55.56	27.78	42.86	43.04
		DeepSeek-R1-Distill-Qwen-14B	53.85	25.67	31.03	20.69	34.87	39.31	39.19	47.90	50.91	37.74	32.46	44.44	25.00	50.00	16.67	14.29	37.73
		DeepSeek-R1-Distill-Qwen-32B	65.38	32.09	48.28	31.03	46.71	39.31	42.57	67.07	61.21	45.28	28.95	25.93	27.78	50.00	38.89	0.00	44.55
		s1-32B	53.85	37.97	40.23	44.83	44.74	44.14	50.00	65.87	61.82	54.72	44.74	33.33	33.33	33.33	50.00	14.29	48.03
		Llama-3.1-8B-Instruct	42.31	21.93	17.24	31.03	26.97	23.79	24.32	43.11	29.09	37.74	27.19	22.22	25.00	33.33	27.78	28.57	27.62
		Llama-3-70B	38.46	22.46	35.63	24.14	32.24	31.03	41.22	50.30	42.42	41.51	30.70	29.63	19.44	22.22	27.78	28.57	34.58
		Llama-3.1-405B	57.69	34.76	49.43	44.83	39.47	47.24	46.62	65.27	63.64	49.06	32.46	44.44	19.44	44.44	55.56	42.86	47.18
	Closed	Gemini-1.5-flash	50.00	35.83	47.13	41.38	33.55	41.38	50.68	69.46	56.97	45.28	43.86	37.04	11.11	44.44	33.33	0.00	45.34
		Gemini-2.0-flash-exp	65.38	<u>43.32</u>	56.32	37.93	48.03	54.83	60.81	70.06	<u>76.36</u>	56.60	47.37	51.85	27.78	50.00	50.00	<u>71.43</u>	56.04
		Claude-3-haiku	46.15	17.65	36.78	31.03	33.55	33.79	39.19	48.50	43.64	35.85	30.70	37.04	16.67	50.00	33.33	28.57	34.97
		Claude-3.5-Sonnet	<u>80.77</u>	42.78	<u>59.77</u>	48.28	49.34	<u>63.10</u>	<u>66.89</u>	76.65	75.15	<u>66.04</u>	54.39	44.44	25.00	<u>66.67</u>	27.78	42.86	<u>59.97</u>
		GPT-3.5-turbo	30.77	19.79	25.29	27.59	28.95	25.86	30.41	34.73	27.27	33.96	22.81	29.63	22.22	27.78	44.44	14.29	27.30
		GPT-4o-mini	50.00	29.95	29.89	20.69	40.13	35.17	39.19	57.49	48.48	47.17	42.98	33.33	25.00	61.11	38.89	14.29	39.96
		GPT-4o	65.38	41.71	57.47	34.48	<u>51.32</u>	53.45	59.46	<u>78.44</u>	70.91	54.72	<u>57.89</u>	<u>66.67</u>	<u>50.00</u>	<u>66.67</u>	<u>61.11</u>	57.14	57.87
		o1-preview	**84.62**	**70.05**	**81.61**	**72.41**	**79.61**	**79.66**	**81.08**	**91.02**	**88.48**	**86.79**	**80.70**	**81.48**	**69.44**	**83.33**	**72.22**	**100.00**	**81.04**
LLMs Avg.			47.72	29.06	38.48	35.38	36.60	37.64	41.67	56.20	48.96	43.61	38.21	36.90	23.97	42.39	35.18	34.92	40.24
Human			65.13	66.93	50.67	62.28	59.19	64.12	65.88	70.86	69.35	76.46	55.30	50.25	58.46	60.20	41.68	51.58	63.04

Table 8: Zero-shot evaluation results for the whole main linguistic category and subcategory.

Example in Phonological System

Question: 〈보기〉의 음운 카드를 활용하여 학습한 내용으로 적절하지 않은 것은? (Which of the following is incorrect regarding the phonemes in the paragraph?)

Paragraph:

⊙ 음운: 말의 뜻을 구별해 주는 소리의 가장 작은 단위 ‘ㄱ’, ‘ ㅁ’, ‘ㅓ’, ‘ㅗ’
(Phoneme: The smallest unit of sound that distinguishes meaning in a language. ‘ㄱ/g/ or /k/’, ‘ ㅁ /m/’, ‘ㅓ /eo/ ’, ‘ㅗ /o/’)

Answer Choices:

① ‘ ㅁ’, ‘ ㅓ’, ‘ ㄱ’을 차례로 사용하면 ‘ 먹$_{ink}$’이라는 단어를 만들 수 있군.
(By using ‘ㅁ $_{/m/}$’, ‘ㅓ $_{/eo/}$’, and ‘ㄱ $_{/k/}$’ in sequence, you can create the word ‘먹 $_{/meok/}$’.)
② ‘먹 $_{ink}$’의 가운뎃소리인 ‘ㅓ’ 대신 ‘ㅗ’를 사용하면 새로운 단어가 되는 군.
(If you replace the medial vowel ‘ㅓ$_{[eo]}$’ in ‘먹 $_{[meok]}$’ with ‘ㅗ $_{[o]}$’, it forms a new word.)
③ ‘목: 곰’에서 보면 첫소리가 끝소리에, 끝소리가 첫소리에도 쓰일 수 있군.
(From ‘목 $_{[mok]}$: 곰$_{[gom]}$’, we can observe that the initial sound can also serve as a final sound and vice versa.)
④ ‘먹$_{ink}$: 목$_{neck}$’처럼 가운뎃소리는 첫소리의 오른쪽에 써야 하는군.
(Like in ‘먹$_{[meok]}$: 목$_{[mok]}$ ’, the medial vowel must be placed to the right of the initial consonant.)
⑤ ‘목$_{neck}$ / 먹$_{ink}$ / 곰$_{bear}$ / 검$_{sword}$’ 처럼 음운의 결합에 따라 의미가 다른 여러 단어를 만들 수 있군.
(Words such as ‘목$_{[mok]}$ / 먹$_{[meok]}$ / 곰$_{[gom]}$ / 검$_{[geom]}$’ demonstrate that different combinations of phonemes can create words with different meanings.)

Example in Phonological Alternation

Question: 〈보기〉를 참고할 때 동화의 양상이 다른 것은? (Based on the paragraph, which one shows a different type of assimilation?)

Paragraph:

⊙ 순행동화: 뒤의 음운이 앞의 음운의 영향을 받아 그와 비슷하거나 같게 소리 나는 현상.
(Progressive assimilation : A phenomenon where a following phoneme is influenced by a preceding phoneme and becomes similar to or identical to it.)

예) 칼날[칼랄] / kalnal [kallal], 강릉[강능] / gangneung [gangneung]

⊙ 역행 동화: 앞의 음운이 뒤의 음운의 영향을 받아 그와 비슷하거나 같게 소리 나는 현상.
(Regressive assimilation : A phenomenon where a preceding phoneme is influenced by a following phoneme and becomes similar to or identical to it.)

예) 편리[펼리] / pyeoli [pyeolli], 까막눈 [까망눈] / kkamaknun [kkamangnun]

Answer Choices:

① 종로 (Jongno) ② 작년 (Jangnyeon) ③ 신라 (Silla) ④ 법물 (Beommul) ⑤ 국민 (Gukmin)

Figure 12: Examples in the *Phonology* Category

Example in Part-of-Speech

Question: 〈보기〉의 밑줄 친 단어를 바르게 분류한 것은? (What is the correct classification of the underlined words in the paragraph?)

Paragraph:

형용사와 관형사를 구별하는 기준의 하나로 '서술하는 기능'이 있다. 예를 들어, '동물원에는 큰 사자가 있다.' 에서 '큰' 은 '사자가 크다 ' 처럼 주어인 '사자가 ' 를 서술하는 기능을 하므로 형용사이다. 그러나 관형사는 그런 기능을 하지 못한다.
(One of the criteria for distinguishing adjectives and determiners is their 'descriptive function.' For instance, in the sentence 'There is a big lion in the zoo', '*big*' functions descriptively by saying 'The lion is big', which means it is an adjective. However, determiners do not perform such functions.)

a. 정원에 아름다운 꽃이 피었다. (A beautiful flower bloomed in the garden.)
b. 웬 말이 그렇게 많은지 모르겠다. (I don't know why there's so much talking.)
c. 수리를 하고 나니 새 가구가 되었다. (After the repair, it became a new piece of furniture.)
d. 모여 있던 모든 사람들이 일제히 나를 쳐다봤다. (All the people gathered suddenly stared at me.)
e. 그의 빠른 일처리가 사람들을 만족스럽게 하였다. (His quick work satisfied the people.)

Answer Choices:

	형용사(adjective)	관형사(determiners)
①	a, c	b, d, e
②	a, e	b, c, d
③	b, d	a, c, e
④	a, c, e	b, d
⑤	b, c, d	a, e

Example in Morphemes

Question: 의존 형태소이면서 실질 형태소인 것만으로 묶인 것은? (Which of the following consists only of dependent morphemes that are also lexical morphemes?)

Paragraph:
⦿ 영희는 책을 집에 놓고 학교에 갔다. (Yeonghee put the book at home and went to school)

Answer Choices:
① 놓-, 가- (put, go) ② -고, -ㅆ (and, past tense) ③ 영희, 책, 집 (Yeonghee, book, home) ④ -는, -을, -에 (subject marker, object marker, locative marker)

Example in Word Formation

Question: 어휘의 구성이 나머지와 다른 것은? (Which of the following has a different lexical composition from the others?)

Answer Choices:
① 참숯 (charcoal) ② 헌옷 (cast-off) ③ 풋과일 (unripe fruit) ④ 개살구 (wild apricot)

Figure 13: Examples in the *Morphology* Category

Example in Sentence Structure

Question: <보기>의 (a)~(c)를 이해한 내용으로 적합하지 않은 것은? (Which sentence has proper grammatical concord?)

Paragraph:

(a) 그는 위기를 좋은 기회로 삼았습니다. (He considered the crisis a good opportunity.)

(b) 바다가 눈이 부시게 파랗다. (The sea is blue to the extent that it blinds the eyes.)

(c) 동주는 반짝이는 별을 응시했다. (Dongju gazed at the twinkling stars.)

Answer Choices:

① (a)의 '삼았다'는 주어 이외에도 두 개의 문장 성분을 필수적으로 요구하는군. ('삼았다$_{\text{considered}}$' in (a) requires two more sentence elements in addition to the subject.)

② (b)의 '바다가'와 '눈이'는 각각 다른 서술어의 주어이군. ('바다가$_{\text{The sea}}$' and '눈이$_{\text{eyes}}$' in (b) are the subjects of different predicates.)

③ (c)의 '별을'은 안긴문장의 목적어이면서 안은문장의 목적어이군. ('별을$_{\text{stars}}$' in (c) is the object of the embedded sentence and the object of the enclosed sentence.)

④ (a)의 '좋은'과 (c)의 '반짝이는'은 안긴문장의 서술어이군. ('좋은$_{\text{good}}$' in (a) and '반짝이는$_{\text{twinkling}}$' in (c) are predicate phrases in the embedded sentence.)

⑤ (b)의 '눈이 부시게'와 (c)의 '반짝이는'은 수식의 기능을 하는군. ('눈이 부시게$_{\text{blinds the eyes}}$' in (b) and '반짝이는$_{\text{twinkling}}$' in (c) function as modifiers.)

Example in Syntactic Features

Question: 밑줄 친 말의 쓰임이 옳지 않은 것은? (Which of the following usages of the underlined word is incorrect?)

Answer Choices:

① 그는 아까운 능력을 썩히고 있다. (He is wasting his precious talent.)

② 음식물 쓰레기를 썩혀서 거름으로 만들었다. (Food waste was decomposed to make compost.)

③ 나는 이제까지 부모님 속을 썩혀 본 적이 없다. (I have never been wasted my parents' peace of mind.)

④ 그들은 새로 구입한 기계를 창고에서 썩히고 있다. (They are wasting the newly purchased machine in the warehouse.)

Figure 14: Examples in the *Syntax* Category

Example in Vocabulary

Question: 〈보기〉의 빈 칸에 공통적으로 들어갈 단어로 적절한 것은?
(Which word is appropriate to fill in the blanks in all the sentences below?)

Paragraph:

- 사람들은 그를 천재라고 (   ). / (People _______ him as a genius.)
- 복은 또 다른 복을 (   ). / (Luck _____ another stroke of luck.)
- 그는 속으로 쾌재를 (   ). / (He _____ inwardly with delight.)

Answer Choices:

① 말하다 (speak) ② 부르다 (call) ③ 여기다 (consider) ④ 외치다 (shout) ⑤ 생각하다 (think)

Example in Lexical Semantics

Question: 밑줄 친 '의'의 앞말과 뒷말의 의미 관계를 잘못 파악한 것은?
(Which of the following incorrectly identifies the semantic relationship between the underlined "of" and the words it connects?)

Answer Choices:

① 어머니의 신발이 많이 낡았다. / [소유주와 소유물] (The shoes of my mother are very worn out. / [Owner and Possession])
② 오늘 나의 짝은 선행상을 받았다. / [속성과 대상] (The partner of mine received the Good Conduct Award today. / [Attribute and Object])
③ 이 일은 선생님의 충고를 따라야겠다. / [주체와 행위] (This matter requires following the advice of the teacher. / [Subject and Action])
④ 졸업한 선배들에게 축하의 박수를 보내자. / [목표와 수단] (Let's give a round of applause of congratulations to the graduates. / [Goal and Means])
⑤ 우리 국민의 절반이 축구 중계방송을 시청했다. / [전체와 부분] (Half of the people of our nation watched the soccer broadcast. / [Whole and Part])

Example in Pragmatics

Question: 〈보기1〉의 밑줄 친 부분의 예를 〈보기2〉에서 고른다고 할 때, 가장 적절한 것은?
(When selecting an example of the underlined part of 〈Paragraph 1〉 from 〈Paragraph 2〉, which one is the most appropriate?)

Paragraph 1:

발화는 발화자의 어떤 의도를 담고 있다. 따라서 발화자가 상대방(청자)에게 무엇인가를 요구할 때, 일반적으로 명령문을 사용하여 발화자의 의도를 직접 드러낸다. 하지만 담화 상황에 따라 발화자가 요구하는 바를 평서문을 통해 상대방에게 간접적으로 표현하거나 의문문을 통해 상대방에게 간접적으로 표현할 수도 있다.
(An utterance carries a certain intention of the speaker. Therefore, when the speaker makes a request to the listener, a command sentence is typically used to express the speaker's intention directly. However, depending on the discourse situation, the speaker may indirectly express the request through a declarative sentence or indirectly express it through an interrogative sentence.)

Paragraph 2:

○ 모임에서 만나 둘이 이야기를 하는 상황 (A situation where two people are talking at a meeting)
남자 A: (a)저는 ○○ 고등학교에 다닙니다. (Man A: I attend ○○ High School.)
남자 B: 그 학교는 어디에 있나요? (Man B: Where is that school located?)

○ 병원에서 의사가 환자를 진료하는 상황 (A situation where two people are talking at a meeting)
의사: (b)예전보다 많이 좋아지셨네요. (Doctor: You've improved a lot compared to before.)
환자: 전부 의사 선생님 덕분입니다. (Patient: It's all thanks to you, doctor.)

○ 개학 후 교사가 학생들을 처음 대면한 상황 (A situation where a teacher meets students for the first time after the semester begins)
교사: (c)여러분, 많이 보고 싶었어요. (Teacher: Everyone, I've missed you a lot.)
학생: 선생님, 저희도 그래요. (Students: We've missed you too.)

○ 귀가한 아들이 어머니에게 말하는 상황 (A situation where a son speaks to his mother after returning home)
아들: (d)엄마, 배가 너무 고파요. (Son: Mom, I'm so hungry.)
엄마: 그래, 금방 차려 줄게. (Mom: Okay, I'll prepare something right away.)

○ 여행객이 아름다운 경치를 보고 있는 상황 (A situation where a traveler is admiring a beautiful view)
여행객 A: (e)이곳은 정말 아름답습니다. (Traveler A: This place is truly beautiful.)
여행객 B: 그래요. 정말 아름답네요. (Traveler B: Yes, it really is stunning.)

Answer Choices:

① (a) ② (b) ③ (c) ④ (d) ⑤ (e)

Figure 15: Examples in the *Semantics* Category

Example in Orthography

Question: 다음 중 한글 맞춤법에 맞는 단어들로만 묶인 것은? (Which group contains words that all adhere to Korean orthography rules?)

Answer Choices:

① 머릿방 - 핏기 - 셋방 - 곳간 (Attic - Glow on the face - Small rented room - Storeroom)
② 널따랗다 - 높따랗다 - 굵직하다 - 짤막하다 (Wide - Tall - Thick - Short)
③ 넘어지다 - 허얘지다 - 흩어지다 - 버러지다 (Fall down - Turn pale - Scatter - Break apart)
④ 합격율 - 규율 - 선율 - 실패율 (Pass rate - Discipline - Melody - Failure rate)

Example in Standard Language

Question: 〈보기〉의 설명에 해당하지 않는 것은? (Which of the following does not match the explanation in the passage?)

Paragraph: 우리말에서는 뜻이 같으면서 형태가 다른 낱말들이 있을 때, 그 쓰임의 범위 차이가 크게 나지 않는다면 모두 표준어로 삼고 있다. 가령, '신'과 '신발'은 쓰임의 범위가 비슷하므로 모두 표준어이다. 이를 가리켜 . '복수 표준어'라 한다. (In Korean language, when there are words with the same meaning but different forms, if the range of their usage does not differ significantly, both are recognized as standard words. For example, 'shoe' and 'shoes' have similar usage ranges, so both are considered standard. This is referred to as 'multiple standard words.')

Answer Choices:

① 천둥 / 우레 (Thunder) ② 나귀 / 당나귀 (Donkey) ③ 옥수수 / 강냉이 (Corn) ④ 자물쇠 / 자물통 (Lock) ⑤ 선머슴 / 풋머슴 (Tomboy)

Example in Standard Pronunciation

Question: 밑줄 친 부분의 발음이 표준 발음법에 맞는 것은? (Which of the underlined parts matches the standard pronunciation rules?)

Answer Choices:

① 보리의 생장에는 겨울철 밟기 [밥끼]가 중요하다. (The growth of barley requires winter trampling.)
② 한글 자모 순서에서 'ㄴ' 다음에는 'ㄷ'이 [디그디] 온다. (In the order of Korean consonants, 'ㄴ' is followed by 'ㄷ'.)
③ 당시 학계에는 일원론보다는 이원론 [이: 원논]이 우세하였다. (In the academic field at the time, dualism was more dominant than monism.)
④ 오전에 맑다[말따]가 오후에 차차 흐려져 밤늦게 비가 오겠습니다. (In the morning, it will be sunny and gradually become cloudy in the afternoon and rain late at night.)

Example in Romanization

Question: 다음 중 로마자 표기법에 맞는 것은? (Which one is the correct roman script?)

Answer Choices:

① 볶음밥$_{\text{fried rice}}$ / Bokkeumbap ② 동래$_{\text{Dongnae (The name of local place)}}$ / Dongrae ③ 벚꽃$_{\text{cherry blossoms}}$ / beotkkot ④ 식혜$_{\text{sweet rice punch}}$ / shikhye

Example in Loanword Orthography

Question: 밑줄 친 외래어의 표기가 바른 것은? (Which of the underlined foreign words is correctly written?)

Answer Choices:

① 신나는 음악을 듣고 있으니 엔돌핀이 용솟음치는 듯하다. (Listening to exciting music feels like endorphins are surging.)
② 축제 기간에 진행하는 행사들은 팜플렛을 통해 소개되었다. (Events during the festival were introduced through pamphlets.)
③ 이번 여행에서 가장 기억에 남는 것은 야외에서의 바베큐 파티였다. (The most memorable part of this trip was the outdoor barbecue party.)
④ 자동차에 문제가 있는 것 같아 갓길에 정차하여 차의 보닛을 열어 보았다. (Thinking there might be an issue with the car, I stopped on the street and opened the car's bonnet.)

Example in Cross-Category

Question: 한글 맞춤법 및 표준어 규정에 맞게 쓴 것은? (Which one is correctly written according to Korean orthography and standard language rules?)

Answer Choices:

① 철수가 문제의 답을 잘못 알아맞혔다. (Chulsoo guessed the answer to the question incorrectly.)
② 작년에 이 학교는 취업률이 매우 높았다. (Last year, this school had a very high employment rate.)
③ 교배를 시키려고 튼튼한 숫놈을 사 왔다. (They bought a strong male to breed.)
④ 그건 막을래야 막을 수가 없는 재난이었다. (That was a disaster that couldn't be stopped even if one tried.)

Figure 16: Examples in the *Norms* Category