Assessing GPT models' Sensitivity to Epistemic Meanings in Korean Periphrastic Construction

Yebin Lee[1], Arum Kang[2], Sanghoun Song[1]
yblee1018@korea.ac.kr, arkang@cnu.ac.kr, sanghoun@korea.ac.kr
[1]Korea University [2]Chungnam National University

Abstract

This study investigates whether large language models can process epistemic modality in Korean, where degrees of certainty are often expressed through periphrastic constructions with interrogative complementizers and epistemic predicates. Using GPT-4.1, two experiments tested the model's certainty judgments with and without contextual cues. Without context, the model consistently defaulted to a 50% certainty across different predicates, suggesting that its responses are categorical in nature. However, with context, responses became more varied and partly human-like, but still lacked the gradient sensitivity observed in human speakers. Further analysis revealed an exceptional pattern for the word *siph-* 'seem/believe', but this likely stems from the frequency and familiarity of the expression in the model's training corpus, leading to a holistic representation, rather than reflecting a genuine understanding of the semantic distinctions introduced by different interrogative complementizers. These results indicate that, while the model can respond to explicit contextual signals, it does not appear to encode the internal semantic distinctions that native speakers associate within epistemic modal meaning and Korean interrogative complementizers.

1 Introduction

What is unique about human language, compared to animal communication, is its ability to convey information beyond immediate reality. This ability enables humans to make counterfactual statements, such as lies or hypothetical scenarios. For example, look at example 1 below.

(1) a. John may be in his office.
b. John must be in his office.
c. John'll be in his office. Palmer (2001)

Example (1) illustrates epistemic modality in English. In (1a), the speaker expresses uncertainty about whether John is in his office by using the modal verb 'may'. In contrast, in (1b), the speaker conveys a firm judgment through the modal verb 'must'. Finally, in (1c), the speaker makes a judgment based on what generally happens with John. This contrast shows how epistemic modality enables speakers to modulate their degree of certainty, making communication more nuanced and context-sensitive. Not all languages encode epistemic modality morphologically, and the availability and granularity of such distinctions vary across languages.

In Korean, epistemic modality sometimes appears in periphrastic constructions, which are constructions in which multiple words function as a single grammatical unit. Notably, Korean has four interrogative complementizers: *-nci*, *-lci*, *-nka*, and *-lkka*. Each interrogative complementizers can be divided into more fine-grained elements, each carrying distinct semantic nuances. For instance, complementizers ending in *-kka* are more actively used in modalized questions, compared to those ending in *-ci*. In addition, complementizers with *-n-* typically mark realis or present meaning, whereas those with *-l-* mark irrealis or future meaning. However, when these complementizers combine with certain predicates, subtle semantic differences lead to variations in the level of epistemic commitment. Consider the example sentence below.

(2) Minci-ka nayil hakkyo-ey
Minci-NOM tomorrow school-LOC
o-nunci / o-lci / o-nunka / o-lkka
come-PRES/FUT/PRES/FUT.**whether**
kwungkumha-ta.
wonder-DECL
'I wonder whether Minci would come to the party.'

Each of the four complementizers combined with *o-* 'come' has its own distributional constraints on

the following predicates (Kang and Song, 2021). Specifically, *-nci* and *-lci* are ordinary interrogative complementizers compatible with factive or neutral predicates, while *-nka* and *-lkka* function as modalized (subjunctive) complementizers that yield weaker epistemic commitment and conjectural readings. These distinctions reveal that the complementizers encode different degrees of speaker commitment within epistemic modality. Because these semantic nuances are subtle distinctions that only native speakers can reliably perceive, determining whether a Korean-trained language model can discriminate among them is an important challenge in computational linguistics.

Therefore, based on the experimental design and findings of Kang and Song (2021), this study aims to determine whether large language models can distinguish between complementizers in periphrastic constructions and assign a similar degree of certainty to humans. We concluded that language models fail to distinguish between the different meanings of the interrogative complementizer, although they are able to capture the overall degree of certainty expressed by the sentence. This result shows that language models process the certainty as a whole, not a combination of the complementizer and the predicate.

The rest of this paper is organized as follows. Section 2 presents the theoretical and computational background of the study. The theoretical background focuses on epistemic modality and the subjunctive mood in Korean periphrastic constructions, while the computational background addresses the concept of certainty in language models. Section 3 details the experimental setup, including the dataset, model, and procedure. Section 4 presents the results of all experiments, along with a discussion. Finally, Section 5 concludes with a summary and a discussion of the study's limitations.

2 Background

2.1 Linguistic Background: Epistemic Modality

Epistemic modality refers to the speaker's attitude toward the reality or likelihood of a given proposition, expressing how strongly the speaker believes in its truth. In Korean, this is conveyed through various linguistic devices, including verbal endings such as *-keyss-* 'will', periphrastic constructions like *-(u)l get* 'may', and adverbs such as *ama* 'maybe' and *cheoldae* 'never' (Son, 2016). These expressions not only signal the degree of certainty or inference but also reflect the source and nature of the speaker's knowledge, whether it is derived from direct observation, memory, reasoning, or hearsay. Epistemic modality thus functions as both a semantic and pragmatic system, providing insight into how Korean speakers evaluate and encode their access to information.

Additionally, epistemic modality plays a key role in Korean interrogatives through so-called modalized questions. Recent studies such as Kang and Yoon (2019) show that sentence-final particles such as *-lkka* and *-nka* function not only as interrogative markers but also as indicators of epistemic uncertainty and emotional involvement. For instance, the question *o-ass-ulkka* 'might come'? not only seeks information but also conveys the speaker's speculation or concern. These constructions demonstrate that modality and interrogativity are tightly integrated at the clause level in Korean, forming a complex system that encodes both grammatical and affective meanings.

Another key theoretical notion underlying Korean epistemic constructions is nonveridical equilibrium (Kang and Yoon, 2020). Nonveridical equilibrium refers to a semantic state in which a proposition is evaluated as neither fully true nor fully false within the speaker's epistemic space, allowing both p and $\neg p$ worlds to remain accessible. This equilibrium characterizes linguistic environments where the speaker's commitment is suspended, such as conjecture, doubt, or emotional speculation. Specifically, complementizers like *-nka* and *-lkka* instantiate this nonveridical equilibrium by encoding a weakened truth commitment and balancing the likelihood of opposing possibilities. Thus, these complementizers not only mark interrogativity but also convey a specific epistemic stance of suspended commitment, distinguishing them from indicative forms like *-nci* or *-lci*, which presuppose veridical or factive evaluation.

To empirically verify these theoretical claims, Kang and Song (2021) conducted a collostructional analysis using the Sejong Corpus (National Institute of the Korean Language, 2009) and two follow-up experiments: an acceptability judgment task and a context-sensitive evaluation. Their results demonstrated that *-lkka* and *-nka* exhibited high collostructional strength and acceptability when paired with nonveridical predicates such as *siph-* 'seem' or 'believe', *kekcengsulep-* 'worried', and *molu-* 'do not

know', with corresponding lower certainty judgments (10–50%). In contrast, *-nci* and *-lci* appeared more frequently with veridical predicates like *al-* 'to know' and *hwaksinha-* 'be certain', and were associated with stronger commitment. These findings provide quantitative support for the hypothesis that the Korean subjunctive is marked at the complementizer level and that these markers carry systematic semantic effects in relation to certainty.

Building on this prior work, the present study extends the experimental design of Kang and Song (2021) by applying it to a large language model, especially for GPT series. Rather than relying solely on theoretical constructs, this research uses their empirical predicate-complementizer pairings and replicates their acceptability and context-based judgment tasks under comparable conditions. This allows for a direct comparison between human and model behavior, providing a more grounded assessment of how LLMs interpret degrees of certainty in structurally complex Korean constructions.

2.2 Computational Background: Language Model Uncertainty

For language models, certainty refers to the ability to express the degree of confidence in their judgments based on internal knowledge. This involves more than simply providing correct answers; it also encompasses distinguishing between known and unknown information, and using appropriate linguistic cues to convey uncertainty in ambiguous or underspecified contexts. Such capacity is essential in high-stakes domains like medical consultations and legal reasoning, where reliable communication is crucial.

In recent surveys, certainty in language models has been characterized as encompassing two interrelated aspects: the model's internal confidence in its output and the explicit linguistic expressions it employs to signal that confidence. Both dimensions are central to understanding how models handle epistemic judgments. The importance of certainty lies not only in reducing hallucinations by preventing overconfident yet inaccurate statements, but also in revealing how closely model behavior aligns with human cognitive processes.

Recent studies have pointed out structural limitations in how language models express or evaluate certainty. Suzgun et al. (2024) assessed models such as GPT-4, Claude-3, and LLaMA-3 and found that, while these models perform well on fact-based questions (Such as a question that starts with "Is it true that ..."), their accuracy drops significantly on tasks involving falsehoods or beliefs. This issue is particularly severe when models are asked to affirm beliefs that contradict factual information or are expressed in the first person (Such as a question that starts with "I believe ..., Do I believe ... ?"). Extending this line of inquiry, Li et al. (2025) investigated how models distinguish among fact, fiction, and forecast, and evaluated their use of evidence-based certainty expressions. Despite the fluency of their outputs, models often failed to select expressions that matched the strength of evidence or epistemic commitment, indicating a limited understanding of epistemic modality.

Language models thus lack a coherent internal mechanism to determine what they know and what they do not. As a result, they frequently make confident statements regardless of the information's actual reliability. Yona et al. (2024) introduced the concept of faithful response uncertainty, which quantifies the mismatch between a model's internal confidence and the decisiveness of its verbal output. Their findings show that models often express high certainty even when internally uncertain, potentially misleading users. Similarly, Krause et al. (2023), in a multilingual QA setting, found that GPT-3.5 tended to output high-confidence responses regardless of accuracy, especially in low-resource languages, revealing a persistent overconfidence bias.

To address this issue, various methods have been proposed to estimate or calibrate uncertainty in LLMs, including probability-based scoring, uncertainty quantification, and prompt-based self-assessment. However, most are limited to numerical scores or token-level probabilities, and linguistically grounded approaches that examine how certainty is expressed in natural language remain underexplored. Xia et al. (2025) provide a comprehensive survey of uncertainty estimation methods and highlight that LLMs still struggle to communicate what they do or do not know. Moreover, most prior work has focused on English, leaving a gap in understanding how models handle epistemic reasoning in languages with distinct grammatical systems for encoding certainty. Before applying advanced calibration methods, it is therefore necessary to first determine whether a model can recognize and appropriately use the linguistic markers of certainty in a given language. This study addresses that question by examining whether a language model can identify and produce the complementizer–predicate

combinations that encode degrees of certainty in Korean.

3 Method

3.1 Data

The dataset used in this study consists of full Korean sentences constructed using six epistemic predicates and four complementizers. As described in Section 2.1, epistemic modality in Korean is conveyed through periphrastic constructions that pair a complementizer with an epistemic predicate. However, as noted earlier, not all predicate–complementizer combinations are grammatically acceptable, as their compatibility is constrained by the semantic class of the predicate. For example, the complementizer *-nci*, which presupposes a strong belief that the event has occurred, cannot combine with counterfactual predicates such as *siph* 'believe'.

Acceptable combinations were selected based on Kang and Song (2021), who evaluated each construction's acceptability using a 5-point Likert scale and then converted the scores to Z-scores. To quantify the relative acceptability of each construction, we applied the cumulative distribution function (CDF) of the standard normal distribution. The resulting values range from 0 to 100, representing the percentile rank of each construction. To ensure that sentence acceptability would not confound the experimental results, we included only those combinations whose CDF-based percentile exceeded 50%. Notably, the predicate *sayangakha-* 'think' was selected to represent the 90% certainty condition, based on its high acceptability. For the *uysimsulep-* 'doubt', although it was not included in the acceptability test of Kang and Song (2021), it was included in the present study because its semantic property naturally conveys a very low degree of epistemic certainty.

Because the outputs of language models are inherently stochastic, it is difficult to generalize from a single trial. To address this, we artificially constructed a sufficient number of examples for robust evaluation. Specifically, 100 sentence pairs were handcrafted for each of the 16 valid predicate-complementizer combinations, using different verbs and nouns, yielding a total of 1,600 sentences. The complete list of selected predicates and complementizers is provided below (Table 1), and a single representative sentence pair is included in Appendix A.

	-nci	*-nka*	*-lci*	*-lkka*
kwungkumha- 'wonder'	**85.5**	**69.3**	**78.3**	**73.6**
kekcengsulep- 'worried'	**69.8**	49.4	**74.6**	**76.0**
molu- 'not know'	**63.6**	**52.9**	**68.9**	**69.9**
siph- 'seem', 'believe'	39.3	**53.0**	**59.7**	**74.1**

Table 1: Acceptability of the Periphrastic Constructions

3.2 Experiment

An experiment in this study was conducted following Kang and Song (2021). In their study, the experiment was conducted in two ways: first, to identify acceptable combinations within the periphrastic construction, and second, to determine the degree of certainty each combination conveys. The first experiment was an acceptability task in which participants judged the acceptability of given sentences. The second experiment was context-based, in which participants answered yes/no questions about whether a sentence was appropriate for a given interpretation. The main framework of the present study follows the latter approach, which is further divided into two specific subtypes.

The present study aims to achieve two goals. First, to determine whether the presence of contextual information influences the model's certainty judgments. Second, to assess whether the language model exhibits patterns similar to those reported in Kang and Song (2021). The experiment was designed analogously to the human study, consisting of two settings: a context-free task and a context-based task. In the context-free task, the model was prompted to provide an answer reflecting its level of commitment (10%, 50%, or 90%) without any supporting context. In the context-based task, the model was presented with specific discourse contexts and asked to respond with yes or no, indicating whether a given probability level of commitment (10%, 50%, or 90%) was appropriate for the situation. Both tasks require the model to evaluate the certainty of a sentence, but differ in whether they include contextual cues.

In the context-free task, the model is presented with a sentence in isolation and asked to choose a certainty level among 10%, 50%, or 90%. In contrast, the context-based task provides the model with three components: a question, a sentence, and a context. The question takes the form of a yes or no interrogative, such as "Does the following sentence accurately reflect the speaker's thoughts in response to the interlocutor's question?" The sentence is a declarative containing a complementizer–predicate combination that fits the given ques-

tion and context. The context provides information necessary to evaluate the plausibility of the sentence, indicating one of three likelihood levels (10%, 50%, or 90%). This task differs from the context-free one in that it provides an explicit epistemic frame for interpretation. Consequently, the model's response format also changes: while the context-free task requires selecting among scalar degrees of certainty, the context-based task asks for a binary judgment ("yes" or "no") on whether the sentence is appropriate given the contextual information.

3.3 Model

The language models used in this experiment are the OpenAI GPT-4 series: GPT-4.1, GPT-4o, and GPT-4.1 mini(Achiam et al., 2023). These models were selected not to compare their performance, but to evaluate whether state-of-the-art language models can distinguish different degrees of certainty based on Korean linguistic cues. All test sentences were delivered via API calls with a base temperature of 0, and no prior conversational context or inference state was provided. Each request included a shared system prompt: *"Your role is to evaluate the certainty of the following sentences as a native Korean speaker."* The maximum token length for each response was set to 50. For the sake of space and clarity, this paper reports only the results from GPT-4.1.

4 Result

4.1 Task1 & Task2 Comparison

From Table 2, we can see that the results from Task 1 and Task 2 are different in a certain way. In particular, when comparing the overall distributions between context-free and context-based tasks, a clear contrast emerges. In the absence of contextual signals, the model exhibited an overwhelming preference for certainty 50% in almost all items, suggesting either epistemic default or indecision. However, once context was introduced, responses became more varied and better aligned with human patterns. The presence of contextual information appears to activate a dormant sensitivity in the model, enabling it to shift away from the 50% default and select 10% or 90% when appropriate.

Notably, predicates such as *uysimsulep-* 'doubt' and *sayngkakha-* 'think' demonstrated striking improvements in alignment with human expectations. In the case of *uysimsulep-* 'doubt', the model began selecting 10% with much greater frequency in the context-based task, mirroring human responses reported by (Kang and Song, 2021). Likewise, for *sayngkakha-* 'think', a predicate typically associated with high certainty, the model shifted from an uncertain stance in the context-free task to confidently selecting 90% when a supportive context was provided. These shifts suggest that contextual embedding plays a crucial role in enabling the model to simulate human-like gradience in epistemic judgment.

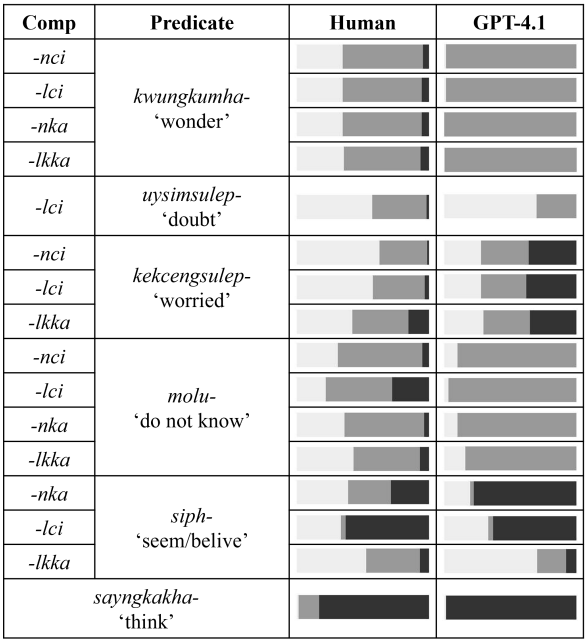

Figure 1: Overall result of Fill-in-the-Blank Task. The white, gray, and black bars represent the proportions of "yes" responses for 10%, 50%, and 90% commitment levels, respectively.

In sum, this section demonstrates that while the language model fails to exhibit gradient reasoning in isolation, the addition of contextual information enables it to move toward more human-aligned responses. The clearest evidence of this is the redistribution of responses across the 10%–50%–90% scale once context is available, with predicate-specific sensitivity emerging most notably for epistemically polarized verbs. This provides the first key finding of the study: that contextual grounding facilitates a more gradient and human-like pattern of certainty estimation in LLM.

4.2 GPT & Human Comparison

However, when compared to human responses, several critical discrepancies remain. Figure **??** presents a comparative visualization between

Comp	Predicate	Context-Free Task			Context-Based Task		
		10%	**50%**	**90%**	**10%**	**50%**	**90%**
-nci	*kwungkumha-* 'wonder'	0	100	0	1	100	0
-lci		0	100	0	1	100	0
-nka		0	100	0	0	99	0
-lkka		0	100	0	0	100	0
-lci	*uysimsulep-* 'doubt'	2	89	9	99	0	0
-nci	*kekcengsulep-* 'worry'	0	99	1	10	13	13
-lci		0	99	1	16	20	22
-lkka		0	99	1	17	20	20
-nci	*molu-* 'do not know'	0	100	0	11	100	0
-lci		0	100	0	3	100	0
-nka		0	100	0	11	100	0
-lkka		0	100	0	19	100	0
-nka	*siph-* 'seem/believe'	0	100	0	7	1	28
-lci		0	100	0	19	2	36
-lkka		0	100	0	76	24	8
-	*sayngkakha-* 'think'	0	88	12	1	0	90

Table 2: Result from Context-Free Task and Context-Based Task. In the context-free task, the language model was asked to choose which of the three certainty levels (10%, 50%, or 90%) best matched the given sentence; thus, the portions across the three levels sum to 100. In contrast, the context-based task required independent judgment of whether each sentence accurately reflected the context corresponding to 10%, 50%, or 90% certainty, resulting in separate counts for each condition.

model outputs and human acceptability-based judgments under context-provided conditions. Here, the light gray bars indicate the proportion of "yes" responses at the 10% certainty level, the medium gray bars for 50%, and black bars for 90%. At a glance, GPT-4.1's pattern appears superficially aligned with human ratings, particularly for certain predicates. However, closer inspection reveals notable differences in granularity and sensitivity.

First, the model lacks the gradient distribution across certainty levels that characterizes human responses. For example, in predicates like *kwungkumha-* 'wonder' and *molu-* 'do not know', human participants displayed a probabilistic spread, with secondary selections at 10% or 90% even when 50% was most frequent. In contrast, the model's responses clustered almost exclusively around 50%, failing to exhibit the nuanced variation observed in human reasoning. This one-sided concentration suggests a categorical bias in model predictions, in which a single certainty level dominates across instances of a given predicate. Second, although humans varied their judgments depending on the complementizer, even after controlling for the predicate (e.g., *kwungkumha-* 'wonder' vs. *kekcengsulep* 'worried'), the model did not show this sensitivity. While human participants adjusted their certainty based on fine-grained morphosyntactic cues, GPT-4.1 treated different complementizers more uniformly, leading to flatter variation. This indicates that, unlike human judgments, which are jointly shaped by the predicate and complementizer, model responses are less influenced by the compositional semantics of these constructions. Third, the predicate *kekcengsulep-* 'worried' showed the most distinct divergence. Whereas human judgments predominantly clustered around 10%, the model provided an ambiguous distribution, often split across categories without a strong preference. This discrepancy becomes clearer when examining the model's justifications: GPT-4.1 frequently classified *kekcengsulep-* 'worried' as an emotion-descriptive predicate rather than an epistemic one, asserting that it is not appropriate for evaluating certainty. This semantic misclassification suggests that the model does not recognize the epistemic implications embedded in emotion predicates like worry. Finally, one notable exception is observed with the predicate *siph-* 'seem' in conjunction with the complementizer *-lkka*, forming the construction

-lkka siph-. Here, the model's response distribution aligned closely with human data, accurately reflecting a lower degree of certainty. This suggests that, in certain cases where complementizer–predicate collocations are highly conventionalized, the model can replicate human-like interpretations. However, such alignment remains the exception rather than the rule.

5 Conclusion

This study examined whether large language models can distinguish different levels of certainty expressed through Korean periphrastic constructions. The experiment was conducted in two ways: a context-free task and a context-based task. In the context-free setting, the model consistently preferred a neutral judgment of 50%, regardless of the sentence form. When context was provided, the model began to show more variation in its responses and partially aligned with human judgments for certain combinations. However, the model did not capture the gradual variation that human speakers exhibited, suggesting a limited understanding of fine-grained certainty comprehension in Korean.

Notably, a closer comparison between human and model responses revealed several persistent divergences. First, while human responses exhibited gradient distributions across all certainty levels, the model's predictions were frequently concentrated on a single value, especially 50%. Second, human responses showed variation based on both the predicate and the complementizer, reflecting sensitivity to their interaction, whereas the model's outputs appeared largely invariant across complementizers for a given predicate. Third, for certain epistemic predicates such as *kekcengsulep-* 'worried', the model not only failed to align with human judgments but also rationalized its misclassification by labeling the predicate as non-epistemic, thereby revealing limitations in semantic categorization. These discrepancies underscore that even when context is provided, the model lacks the interpretive mechanisms necessary to represent the probabilistic and compositional aspects of Korean epistemic modality. However, the experiment also revealed only a few cases of convergence. For instance, in the expression *-lkka siph-*, the model produced outputs that closely resembled human judgments. This suggests that when complementizer–predicate pairs co-occur frequently or are structurally salient, language models may succeed in mimicking human-like certainty evaluations. Nevertheless, such instances remain isolated exceptions rather than generalizable patterns.

While these findings provide valuable insight, several limitations remain. Certainty is inherently a gradient rather than a binary concept, and further work is needed to capture this gradience more effectively. In this study, the focus was on complementizers and predicates; future analyses should also examine how epistemic constructions behave across syntactic positions, such as in embedded versus main clauses. Additionally, unlike other predicates, *siph-* 'seem' exhibited a notable tendency for the model to select *-lkka* in 10% certainty contexts, indicating a different complementizer selection pattern from predicates such as *kwungkumha-* 'wonder' or *molu-* 'do not know'. Given the unique semantic behavior of *siph-* 'seem' in Korean, further investigation is warranted. Finally, although Korean has distinctive features in expressing epistemic modality, it employs such markers less frequently than languages with richer epistemic systems, such as Italian. To determine whether language models encode the concept of certainty in a cross-linguistic sense, experiments should be extended to other languages.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Arum Kang and Sanghoun Song. 2021. A study on subjunctive mood in korean: Using corpus and experimental linguistic data. In *23rd Seoul International Conference on Generative Grammar*. The Korean Generative Grammar Circle. Written in Korean.

Arum Kang and Suwon Yoon. 2019. The subjunctive complementizer in korean: The interaction between inquisitiveness and nonveridicality. In *Proceedings of the 12th Generative Linguistics in the Old World & the 21st Seoul International Conference on Generative Grammar*, pages 343–358.

Arum Kang and Suwon Yoon. 2020. From inquisitive disjunction to nonveridical equilibrium: Modalized questions in korean. *Linguistics*, 58(1):207–244.

Lea Krause, Wondimagegnhue Tufa, Selene Báez Santamaría, Angel Daza, Urja Khurana, and Piek Vossen. 2023. Confidently wrong: exploring the calibration and expression of (un) certainty of large language models in a multilingual setting. In *Proceedings*

of the workshop on multimodal, multilingual natural language generation and multilingual WebNLG Challenge (MM-NLG 2023), pages 1–9.

Meng Li, Michael Vrazitulis, and David Schlangen. 2025. Representations of fact, fiction and forecast in large language models: Epistemics and attitudes. *arXiv preprint arXiv:2506.01512*.

National Institute of the Korean Language. 2009. The 21st century sejong project corpus. https://korean.go.kr.

Frank Robert Palmer. 2001. *Mood and modality*. Cambridge university press.

Hyeok Son. 2016. A study on use of epistemic modality markers. *Language Facts and Perspectives*, 39:249–285. Written in Korean.

Mirac Suzgun, Tayfun Gur, Federico Bianchi, Daniel E Ho, Thomas Icard, Dan Jurafsky, and James Zou. 2024. Belief in the machine: Investigating epistemological blind spots of language models. *arXiv preprint arXiv:2410.21195*.

Zhiqiu Xia, Jinxuan Xu, Yuqian Zhang, and Hang Liu. 2025. A survey of uncertainty estimation methods on large language models. *arXiv preprint arXiv:2503.00172*.

Gal Yona, Roee Aharoni, and Mor Geva. 2024. Can large language models faithfully express their intrinsic uncertainty in words? *arXiv preprint arXiv:2405.16908*.

A Example Single Sentence Set For 16 Predicate–Complementizer Combinations

Comp	Predicate	Sentences
-nci	*kwungkumha-* ‘wonder’	Minci-ka nayil hakkyo-ey o-nunci kwungkumha-ta Minci-NOM tomorrow school-LOC come-PRES.whether wonder-DECL *‘I wonder if Minci will come to school tomorrow.’*
-lci		Minci-ka nayil hakkyo-ey o-lci kwungkumha-ta Minci-NOM tomorrow school-LOC come-FUT.whether wonder-DECL *‘I wonder if Minci will come to school tomorrow.’*
-nka		Minci-ka nayil hakkyo-ey o-nunka kwungkumha-ta Minci-NOM tomorrow school-LOC come-PRES.whether wonder-DECL *‘I wonder if Minci will come to school tomorrow.’*
-lkka		Minci-ka nayil hakkyo-ey o-lkka kwungkumha-ta Minci-NOM tomorrow school-LOC come-FUT.whether wonder-DECL *‘I wonder if Minci will come to school tomorrow.’*
-lci	*uysimsulep-* ‘doubt’	Minci-ka nayil hakkyo-ey o-lci uysimsulep-ta Minci-NOM tomorrow school-LOC come-FUT.whether be.doubtful-DECL *‘I doubt Minci will come to school tomorrow.’*
-nci	*kekcengsulep-* ‘worry’	Minci-ka nayil hakkyo-ey o-nunci kekcengsulep-ta Minci-NOM tomorrow school-LOC come-PRES.whether be.worry-DECL *‘I’m worried whether Minci will come to school tomorrow.’*
-lci		Minci-ka nayil hakkyo-ey o-lci kekcengsulwe-ta Minci-NOM tomorrow school-LOC come-FUT.whether be.worry-DECL *‘I’m worried whether Minci will come to school tomorrow.’*
-lkka		Minci-ka nayil hakkyo-ey o-nunka kekcengsulep-ta Minci-NOM tomorrow school-LOC come-PRES.whether be.worry-DECL *‘I’m worried whether Minci will come to school tomorrow.’*
-nci	*molu-* ‘do not know’	Minci-ka nayil hakkyo-ey o-nunci molu-keyss-ta Minci-NOM tomorrow school-LOC come-PRES.whether not.know-MODAL-DECL *‘I do not know if Minci will come to school tomorrow.’*
-lci		Minci-ka nayil hakkyo-ey o-lci molu-keyss-ta Minci-NOM tomorrow school-LOC come-FUT.whether not.know-MODAL-DECL *‘I do not know if Minci will come to school tomorrow.’*
-nka		Minci-ka nayil hakkyo-ey o-nunka molu-keyss-ta Minci-NOM tomorrow school-LOC come-PRES.whether wonder-DECL *‘I do not know if Minci will come to school tomorrow.’*
-lkka		Minci-ka nayil hakkyo-ey o-lkka molu-keyss-ta Minci-NOM tomorrow school-LOC come-FUT.whether not.know-MODAL-DECL *‘I do not know if Minci will come to school tomorrow.’*
-lci	*siph-* ‘seem/belive’	Minci-ka nayil hakkyo-ey o-lci siph-ta Minci-NOM tomorrow school-LOC come-FUT.whether seem-DECL *‘It seems as if Minci will come to school tomorrow.’*
-nka		Minci-ka nayil hakkyo-ey o-nunka siph-ta Minci-NOM tomorrow school-LOC come-PRES.whether seem-DECL *‘It seems as if Minci will come to school tomorrow.’*
-lkka		Minci-ka nayil hakkyo-ey o-lkka siph-ta Minci-NOM tomorrow school-LOC come-FUT.whether seem-DECL *‘It seems as if Minci will come to school tomorrow.’*
-	*sayngakha-* ‘think’	Minci-ka nayil hakkyo-ey o-l-gerago sayngkakha-n-ta Minci-NOM tomorrow school-LOC come-FUT-will think-PRES-DECL *‘I think Minci will come to school tomorrow.’*

B Prompt Used in Each Experiment

Task	English Translation	Korean (Original)
Context-Free Task	Respond with one of the following options, indicating the degree of certainty of the speaker in the given sentence: 10% (very unlikely), 50% (uncertain), or 90% (very likely). Be sure to respond with only one of the three values, and explain the reason in a single sentence along with your answer. **Sentence:** I wonder if Minci will come to school tomorrow.	주어진 문장의 발화자가 확신하는 정도를 10% (가능성이 거의 없음), 50% (전혀 예상할 수 없음), 90% (매우 가능성이 큼) 중 하나로 응답하세요. **문장:** 민지가 내일 학교에 오는지 궁금하다.
Context-Based Task	Does the following sentence fit the given context? Start your answer with either 'yes' or 'no', and explain your reason in a single sentence. **Context(90%):** Sun-i thinks it is very likely that Min-ci will come to school tomorrow. **Context(50%):** Sun-i thinks it is very likely that Min-ci will come to school tomorrow. **Context(10%):** Sun-i thinks it is very likely that Min-ci will come to school tomorrow. **Question:** Does the following sentence appropriately reflect Sun-i's thoughts in response to In-ho's question, "Will Minci come to school tomorrow?" **Sentence:** I wonder if Minci will come to school tomorrow.	다음 문장은 맥락(context)에 잘 부합합니까? '예.' 또는 '아니오.'로만 시작하고, 이유는 대답과 함께 하나의 문장으로 설명하세요. **맥락(90%):** 순이는 민지가 내일 학교에 올 가능성이 매우 크다고 생각한다. **맥락(50%):** 순이는 민지가 내일 학교에 올지 안 올지 전혀 예상할 수 없다. **맥락(10%):** 순이는 민지가 내일 학교에 올 가능성이 거의 없다고 생각한다. **질문:** 인호의 "민지가 내일 학교에 올까?"라는 질문에 다음 문장은 순이의 생각을 잘 반영하는가? **문장:** 민지가 내일 학교에 오는지 궁금하다.

For the Context-Based Task, only single context was provided at a time, and the model was asked to respond with Yes or No.