

KoSimpleQA: A Korean Factuality Benchmark with an Analysis of Reasoning LLMs

Donghyeon Ko^{*1}, Kyubyung Chae^{*2}, Yeguk Jin¹, Byungwook Lee¹, Chansong Jo¹,
Sookyo In¹, Jaehong Lee¹, Taesup Kim⁺², Donghyun Kwak⁺¹

¹Naver Cloud, ²Graduate School of Data Science, Seoul National University
{donghyeon.ko, donghyun.kwak}@navercorp.com, {kyubyung.chae, taesup.kim}@snu.ac.kr

Abstract

We present **Korean SimpleQA (KoSimpleQA)**, a benchmark for evaluating factuality in large language models (LLMs) with a focus on Korean cultural knowledge. KoSimpleQA is designed to be challenging yet easy to grade, consisting of 938 short, fact-seeking questions with unambiguous answers. We conduct a comprehensive evaluation across a diverse set of open-source LLMs of varying sizes that support Korean, and find that even the strongest model generates correct answer only 31.6% of the time, underscoring the challenging nature of KoSimpleQA. Notably, performance rankings on KoSimpleQA differ substantially from those on the English SimpleQA, highlighting the unique value of our dataset. Furthermore, we observe that reasoning helps mitigate the cross-lingual knowledge gap in LLMs, which refers to disparities in their ability to manifest knowledge across languages. KoSimpleQA can be found at <https://github.com/naver-ai/KoSimpleQA>.

1 Introduction

Large language models (LLMs) are increasingly deployed in applications that demand high factual reliability, ranging from information access to education and cultural content delivery. However, despite recent progress, LLMs still frequently generate factually incorrect outputs, a phenomenon widely known as *hallucination* (Banerjee et al., 2025). This has motivated the development of benchmarks that directly measure factual reliability, such as SimpleQA (Wei et al., 2024) and Chinese SimpleQA (He et al., 2024), which evaluate models on short, fact-seeking questions with unambiguous answers.

However, existing benchmarks primarily focus on English and Chinese, while no equivalent benchmark exists for Korean. Simply translating existing benchmarks such as SimpleQA into Korean is not sufficient, since many of their questions are rooted in Anglophone cultural contexts and do not meaningfully assess models trained primarily on Korean data. Evaluating LLMs in a particular language requires not only linguistic competence but also an understanding of the cultural knowledge associated with that language community.

Figure 1 illustrates this distinction: a model with only linguistic competence can respond fluently in Korean but still fails on a culturally grounded question, whereas a model equipped with Korean cultural knowledge produces the correct answer. The release of Chinese SimpleQA further demonstrates that benchmarks need to be adapted to each language community’s cultural context, rather than assuming a single benchmark can generalize across languages. In the same spirit, we propose **Korean SimpleQA (KoSimpleQA)**, which fills this gap for Korean by providing a benchmark centered on Korean cultural knowledge. Examples of KoSimpleQA are provided in Appendix A.

Following the design principles of SimpleQA-style benchmarks, KoSimpleQA is challenging yet easy to grade, consisting of short, fact-seeking questions with unambiguous answers. We performed multiple rounds of human validation to control the quality of the dataset. We then

* Equal contribution. † Co-corresponding authors.

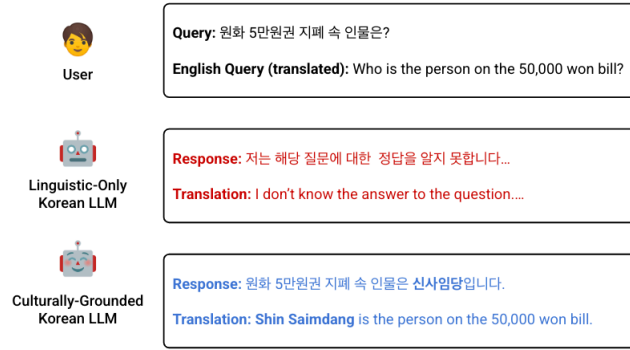


Figure 1: The importance of cultural knowledge for LLMs within a specific language community: an LLM with only linguistic competence responds fluently in Korean but fails, whereas one with Korean cultural knowledge correctly answers the question “Who is the person on the 50,000 won bill?”

systematically evaluated a broad range of open-source LLMs that support Korean, spanning different model families and parameter scales. Despite this diversity, no model generates correct answers more than 31.6% of the time, underscoring the difficulty of the benchmark and suggesting substantial room for further advancement in non-English, low-resource language settings. Moreover, performance rankings on KoSimpleQA diverge substantially from those on English SimpleQA, highlighting the distinct cultural dimension it captures and emphasizing the importance of localized, culturally grounded evaluations.

In addition, we analyze the behavior of recent reasoning LLMs on KoSimpleQA. While reasoning has been emphasized as a core capability of frontier models, its role in factual QA has not been systematically examined. Specifically, we find that leveraging a model’s reasoning ability can mitigate cross-lingual knowledge gap observed in non-reasoning responses, where the model fails to express knowledge it already possesses when the query language changes, and can help elicit its latent knowledge. In such cases, reasoning serves as a mechanism for bridging the gap between internal knowledge and its correct verbalization in factual question answering.

Our contributions are threefold:

- We release **Korean SimpleQA (KoSimpleQA)**, the first benchmark targeting factuality in Korean cultural contexts.
- We provide a comprehensive evaluation across diverse LLMs, establishing new baselines and showing differences from SimpleQA.
- We additionally analyze reasoning models, offering preliminary insights into the role of reasoning in factual QA.

2 Related Work

2.1 Korean Factuality Benchmarks

Several benchmarks have been proposed to evaluate factuality while incorporating Korean cultural knowledge, including KorNAT (Lee et al., 2024), CLick (Kim et al., 2024), HAE-RAE (Son et al., 2024), and KULTURE (Wang et al., 2024). However, these benchmarks share many of the limitations observed in traditional factuality benchmarks such as TruthfulQA (Lin et al., 2022) and CommonsenseQA (Talmor et al., 2019).

First, their relatively low difficulty has led to performance saturation, making it difficult to discriminate between top-tier models. This issue may stem from insufficient consideration of difficulty during data collection, as well as inherent limitations of the multiple-choice format. In multiple-choice settings, the answer candidate space is constrained, allowing models to guess correctly or rely on elimination strategies, which undermines the reliability of the evaluation (Tam et al., 2025; Singh et al., 2025; Li et al., 2024; Balepur et al., 2025).

Second, these benchmarks do not adequately account for the desirable behavior of models abstaining when uncertain (Madhusudhan et al., 2024). From a hallucination perspective, a model admitting uncertainty (“I don’t know”) is preferable to providing an incorrect answer.

However, existing factuality benchmarks often treat both abstentions and incorrect answers as equally wrong. Furthermore, the multiple-choice format inherently excludes abstention by forcing the model to select one of the provided options, thereby failing to capture this important aspect of model behavior.

2.2 SimpleQA and Cultural Adaptation

SimpleQA (Wei et al., 2024) addresses the limitations of traditional factuality benchmarks by presenting a significant challenge even to top-tier LLMs like GPT-4o (Hurst et al., 2024). This difficulty is ensured through a rigorous adversarial filtering process: human annotators were instructed to regenerate questions if a suite of four OpenAI models consistently answered them correctly. This ensures that the dataset consists only of non-trivial questions. To overcome the limitations of multiple-choice formats, SimpleQA adopts an open-ended (free-form) question format while carefully designing questions to elicit concise and uniquely identifiable answers, while still enabling reliable evaluation. Furthermore, SimpleQA introduces a new evaluation framework that prioritizes reliability, specifically measuring a model’s ability to abstain from answering (“know what it knows”) rather than hallucinating when uncertain.

Despite its effectiveness, the original SimpleQA primarily targets English-centric knowledge, leaving a gap in our understanding of LLM capabilities within diverse cultural and linguistic landscapes. Recent efforts have sought to extend this paradigm into multilingual or localized contexts. For instance, KoLasSimpleQA (Jiang et al., 2025) extends SimpleQA to nine languages, marking a valuable step toward multilingual factuality evaluation. However, it faces several limitations: the dataset remains publicly unavailable, constraining reproducibility; the sample size per language is relatively small; and its difficulty filtering is less rigorous than that of the original SimpleQA. Conversely, Chinese SimpleQA (He et al., 2024) focuses rigorously on Chinese-specific knowledge, revealing substantial model ranking discrepancies compared to the original benchmark, with localized community models often outperforming frontier global models.

Motivated by this shift toward localized factuality and the need for a rigorous, open-source benchmark for Korean, we introduce KoSimpleQA, the first factuality benchmark adopting the SimpleQA paradigm for Korean cultural knowledge. Notably, while Chinese SimpleQA was constructed using machine-generated data, KoSimpleQA aligns with the original SimpleQA in prioritizing human-curated QA pairs to ensure high factual reliability and to capture cultural context.

3 KoSimpleQA

KoSimpleQA is constructed through human annotation guided by carefully designed instructions (Section 3.1), followed by multiple rounds of verification (Section 3.2). This methodology ensures a non-trivial benchmark consisting of questions with a single, well-defined answer, enabling reliable evaluation of the factual accuracy of LLMs within the Korean cultural context. The evaluation metrics used in this benchmark are described in Section 3.3.

3.1 Data Collection

We construct a dataset comprising 938 original, human-authored knowledge-seeking questions. We recruit annotators through two Korean crowdsourcing platforms, Selectstar¹ and Flitto². All human annotators are required to pass assessments designed by our data partners to verify their capability in producing high-quality data. Qualified annotators are

¹<https://selectstar.ai/>

²<https://www.flitto.com/>

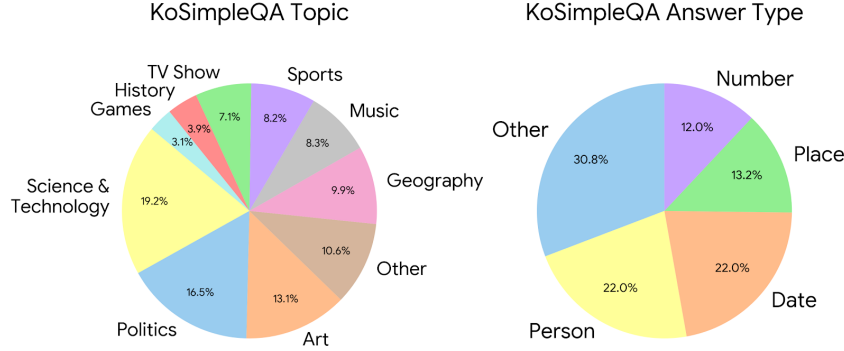


Figure 2: Distribution of topics and answer types in KoSimpleQA. The left chart shows the breakdown by topic category, and the right chart shows the distribution by answer type.

instructed to craft fact-based questions grounded in Korean cultural contexts. A detailed description of the annotator instructions and additional annotation statistics are provided in Appendix D.

Cultural Grounding. While many cross-lingual evaluations rely on translated versions of English-centric datasets (Artetxe et al., 2020; Bandarkar et al., 2024), a core objective of KoSimpleQA is to capture the nuance of Korean context. We define “culture” broadly, extending beyond traditional heritage (e.g., history, literature) to encompass modern socio-cultural trends and contemporary interests unique to native speakers. This scope ensures that the benchmark covers diverse topics of local significance, ranging from traditional holidays to details of popular modern variety shows (e.g., *I Live Alone*). Consequently, while keeping the question format simple, the content necessitates specific parametric knowledge of the Korean cultural sphere.

Difficulty requirement. Each question have to be sufficiently challenging such that at least one of the four strong closed-source LLMs we selected (HCX-005 (Yoo et al., 2024), gpt-4o-2024-05-13 (Hurst et al., 2024), gemini-2.0-flash³, and claude-3-5-sonnet-20240620⁴) failed to produce the correct answer. This criterion guaranteed that KoSimpleQA poses meaningful challenges even for current high-performing systems.

Unambiguous answers. Every question is required to admit a unique and definitive short answer. For example, a question regarding the symbolic meanings of the white background in the South Korean flag has three possible answers (i.e., brightness, purity, and national spirit). In this case, we specified that *all three* should be provided, thereby removing ambiguity. This unambiguity facilitates the verification of the answer.

Temporal constraint. To mitigate evaluation noise caused by the dynamic nature of real-world information, annotators are instructed to create items that could be answered with knowledge available up to December 31, 2023. This constraint is particularly vital for recurring events where the *latest* status changes periodically. For instance, a question inquiring about the “real name of the Final MVP of the League of Legends World Championship 2023” is permissible and unambiguous because the tournament is concluded and the outcome is fully settled before our cutoff date. By establishing this explicit deadline, we ensure that every item remains a stable benchmark, insulated from the influence of subsequent real-world developments.

Topics and Diversity. We adopt the topic categories defined in SimpleQA. We instruct human annotators to create data in a way that would elicit a diverse range of answer types. Figure 2 illustrates the resulting distribution of topic categories and answer types. The dataset covers a comprehensive range of subjects without skewing towards a single domain.

³<https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/>

⁴<https://www.anthropic.com/news/claude-3-5-sonnet>

Source verification. Annotators are instructed to record up to four reliable source links supporting each answer at the time of question creation. These references are subsequently utilized during the quality control phase to verify factual accuracy and ensure reproducibility. By explicitly linking each answer to verifiable evidence, we minimize the likelihood of annotation errors and enhance transparency in dataset construction.

3.2 Quality Control

To guarantee the factual reliability and clarity of KoSimpleQA, we implement a two-stage verification process incorporating an iterative feedback loop. First, we employ a cross-platform validation strategy: instances generated by one platform are cross-validated by annotators from the other to identify initial discrepancies. Second, a manual expert review is conducted by NLP researchers to enforce the rigorous standards defined in Section 3.1.

During this stage, we address various source of errors, including typographical errors, ambiguity, and time-sensitiveness. In particular, many cases related to clarity and time sensitivity are non-trivial and often overlooked without domain expertise in NLP, leading to substantial refinements at this stage. For example, some queries asking for dates are revised to explicitly request an appropriate level of temporal granularity (e.g., year, month, and day as needed), and questions whose answers can vary over time are reformulated to specify a particular temporal reference point.

3.3 Evaluation Metrics

Standard Metrics. We adopt the standard metric definitions from SimpleQA (Wei et al., 2024), which utilize evaluation results from an LLM judge. Following the initial release of the official implementation (simple-evals⁵), we employ GPT-4o as our judge model. However, we refine the NOT_ATTEMPTED criteria to prevent models from receiving credit for invalid refusals. These refinements are reflected in our evaluation prompts to ensure alignment between the metric logic and the automated grader. The metrics are defined as follows:

- **Correct (CO):** The proportion of predictions that fully cover the reference answer without contradicting it.
- **Not Attempted (NA):** The proportion of responses that explicitly refuse to answer or state ignorance (e.g., “I don’t know”). Crucially, unlike the original relaxed evaluation, we do not classify vague responses or false refusals (e.g., claiming that no record exists, even though such information is actually documented somewhere on the web) as NA; these are penalized as Incorrect. Detailed discussion on this classification and the corresponding evaluation prompts are provided in Appendix E.
- **Incorrect (IN):** The proportion of predictions that are not correct.
- **Correct Given Attempted (CGA):** The proportion of accurately answered questions among those that were attempted.
- **F-score:** The harmonic mean of CO and CGA, providing a balanced measure of precision and completeness.

Calibration Metrics. Beyond assessing the model’s propensity to refuse unknown questions (NA), a robust evaluation must account for its metacognitive reliability when it chooses to respond. This is critical because a model that provides *confidently incorrect* answers is fundamentally less reliable than one that acknowledges its uncertainty. We thus evaluate calibration, which we measure by eliciting the model’s verbalized confidence. Following Wei et al. (2024) and He et al. (2024), we instruct models to provide a confidence level from 0 to 100 alongside each response (see Appendix G for the prompt).

Prior work (Wei et al., 2024; He et al., 2024) primarily analyzed calibration plots, without detailed quantitative evaluations. In contrast, we adopt **Expected Calibration Error (ECE)** (Naeini et al., 2015) that is often employed (Xiong et al., 2024; Geng et al., 2024; Tian et al., 2023; Yoon et al., 2025) to enable quantitative analysis of calibration. ECE quantifies the average discrepancy between accuracy and predicted confidence across binned intervals, weighted by sample size. Importantly, these metrics are calculated exclusively on attempted questions to decouple the reliability of the model’s uncertainty estimation from its refusal behavior.

⁵<https://github.com/openai/simple-evals>

Models	CO $\uparrow$	NA	IN $\downarrow$	CGA $\uparrow$	F-score $\uparrow$	ECE $\downarrow$
<i>Korean Community LLMs</i>						
EXAONE 4.0-1.2B	1.6 $\pm$ 0.6	4.7 $\pm$ 1.2	93.7 $\pm$ 1.6	1.7 $\pm$ 0.6	1.6 $\pm$ 0.6	80.5 $\pm$ 0.8
EXAONE 4.0-32B	17.2 $\pm$ 0.4	1.2 $\pm$ 0.1	81.6 $\pm$ 0.5	17.4 $\pm$ 0.4	17.3 $\pm$ 0.4	68.4 $\pm$ 0.3
HCX SEED-0.5B	3.6 $\pm$ 1.1	2.3 $\pm$ 0.7	94.1 $\pm$ 1.8	3.6 $\pm$ 1.1	3.6 $\pm$ 1.1	75.3 $\pm$ 2.5
HCX SEED-1.5B	3.3 $\pm$ 0.4	2.6 $\pm$ 0.6	94.2 $\pm$ 0.9	3.4 $\pm$ 0.4	3.3 $\pm$ 0.4	90.9 $\pm$ 0.9
HCX SEED-3B	8.7 $\pm$ 0.6	5.4 $\pm$ 1.0	85.8 $\pm$ 1.5	9.2 $\pm$ 0.7	9.0 $\pm$ 0.6	81.2 $\pm$ 0.8
HCX SEED-14B	31.6 $\pm$ 0.9	6.1 $\pm$ 0.2	62.3 $\pm$ 1.1	33.6 $\pm$ 1.0	32.6 $\pm$ 0.9	61.5 $\pm$ 0.7
Kanana 1.5-2.1B	14.6 $\pm$ 0.9	9.3 $\pm$ 0.6	76.1 $\pm$ 0.9	16.1 $\pm$ 1.0	15.3 $\pm$ 0.9	78.8 $\pm$ 0.5
Kanana 1.5-8B	27.7 $\pm$ 0.5	1.2 $\pm$ 0.3	71.0 $\pm$ 0.3	28.1 $\pm$ 0.4	27.9 $\pm$ 0.5	64.1 $\pm$ 0.6
VARCO-8B	5.3 $\pm$ 0.4	4.8 $\pm$ 0.3	89.9 $\pm$ 0.3	5.5 $\pm$ 0.4	5.4 $\pm$ 0.4	86.7 $\pm$ 1.1
<i>Multilingual LLMs Supporting Korean</i>						
Gemma 3-1B	2.2 $\pm$ 0.1	0.9 $\pm$ 0.1	97.0 $\pm$ 0.1	2.2 $\pm$ 0.1	2.2 $\pm$ 0.1	83.6 $\pm$ 0.3
Gemma 3-4B	6.6 $\pm$ 0.2	0.4 $\pm$ 0.2	93.0 $\pm$ 0.1	6.6 $\pm$ 0.2	6.6 $\pm$ 0.2	92.2 $\pm$ 0.1
Gemma 3-12B	17.2 $\pm$ 0.3	0.1 $\pm$ 0.1	82.7 $\pm$ 0.3	17.2 $\pm$ 0.3	17.2 $\pm$ 0.3	80.8 $\pm$ 0.3
Gemma 3-27B	28.3 $\pm$ 0.1	0.0 $\pm$ 0.0	71.7 $\pm$ 0.1	28.3 $\pm$ 0.1	28.3 $\pm$ 0.1	70.0 $\pm$ 0.1
Llama 3.1-8B	5.3 $\pm$ 0.7	2.1 $\pm$ 1.2	92.6 $\pm$ 1.9	5.4 $\pm$ 0.8	5.3 $\pm$ 0.7	83.1 $\pm$ 2.8
Llama 3.1-70B	15.6 $\pm$ 0.3	0.2 $\pm$ 0.1	84.1 $\pm$ 0.4	15.7 $\pm$ 0.3	15.7 $\pm$ 0.3	71.4 $\pm$ 4.2
Qwen 3-0.6B	0.8 $\pm$ 0.3	0.7 $\pm$ 0.2	98.4 $\pm$ 0.4	0.8 $\pm$ 0.3	0.8 $\pm$ 0.3	85.0 $\pm$ 1.8
Qwen 3-1.7B	1.6 $\pm$ 0.4	0.7 $\pm$ 0.2	97.7 $\pm$ 0.6	1.6 $\pm$ 0.4	1.6 $\pm$ 0.4	91.1 $\pm$ 0.5
Qwen 3-4B	3.2 $\pm$ 0.3	1.0 $\pm$ 0.3	95.9 $\pm$ 0.5	3.2 $\pm$ 0.3	3.2 $\pm$ 0.3	88.9 $\pm$ 0.2
Qwen 3-8B	5.3 $\pm$ 0.4	0.7 $\pm$ 0.4	94.0 $\pm$ 0.3	5.3 $\pm$ 0.4	5.3 $\pm$ 0.4	89.9 $\pm$ 0.1
Qwen 3-14B	9.0 $\pm$ 0.9	1.1 $\pm$ 0.3	89.9 $\pm$ 0.9	9.1 $\pm$ 0.9	9.0 $\pm$ 0.9	84.7 $\pm$ 0.5
Qwen 3-32B	10.5 $\pm$ 0.2	1.0 $\pm$ 0.2	88.5 $\pm$ 0.1	10.6 $\pm$ 0.2	10.6 $\pm$ 0.2	81.9 $\pm$ 0.6

Table 1: Results of different models on KoSimpleQA. All metrics are reported on a scale of 0 to 100 for better readability. Values represent the mean of three seeds, and $\pm$ denotes the standard deviation. Models are grouped by family and listed in alphabetical order. $\downarrow$ indicates lower-is-better, and $\uparrow$ indicates higher-is-better. Not Attempted (NA) is reported without an arrow as it is not inherently directional.

4 Experimental Setup

To assess model behavior on KoSimpleQA, we evaluate a diverse set of baseline models. We describe the experimental setup below.

4.1 Baselines

Our evaluation encompasses a diverse suite of open-source models with fewer than 70 billion parameters, specifically selected for their proficiency in the Korean language. We categorize these models into two distinct groups: *Korean Community LLMs* and *Multilingual LLMs*. The Korean Community models include EXAONE 4.0 (LG AI Research et al., 2025), HyperCLOVA X (HCX) SEED (Team, 2025), Kanana 1.5 (Bak et al., 2025), and VARCO.⁶ The Multilingual models supporting Korean comprise Gemma 3 (Gemma Team et al., 2025), Llama 3.1 (Dubey et al., 2024), and Qwen3 (Yang et al., 2025).

4.2 Implementation Details

We adhered to the sampling hyperparameters specified in the SimpleQA framework⁷, setting the temperature to 1.0 and top_p to 1.0, while extending the maximum generation length to 8192 tokens to support reasoning models. All reported metrics represent the average across three runs with different random seeds. For ECE calculation, we used a bin size of 10.

5 Main Results

Table 1 presents the overall performance of baseline models on KoSimpleQA. Our analysis highlights three key observations regarding scaling behavior, role of pre-training data distribution and calibration.

⁶<https://huggingface.co/NCSOFT/Llama-VARCO-8B-Instruct>

⁷<https://github.com/openai/simple-evals>

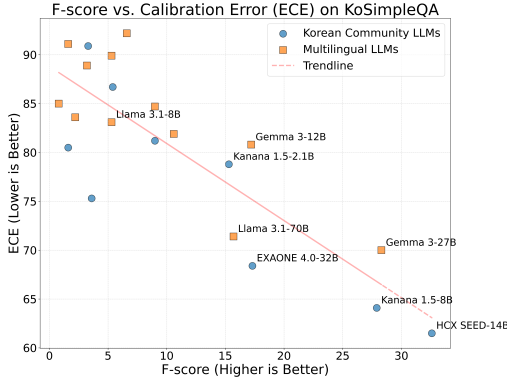


Figure 3: The relationship between performance (F-score) and calibration (ECE) across models on KoSimpleQA.

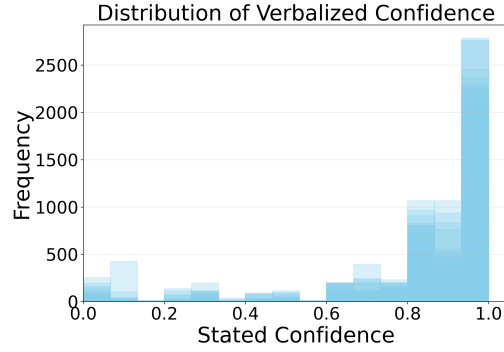


Figure 4: Distribution of stated confidence scores across all evaluated models. The confidence scores are uniformly binned into 15 intervals.

5.1 Scaling Behavior

Consistent with established scaling laws (Kaplan et al., 2020), performance monotonically improves with model size within each model family. For instance, scaling the HCX SEED family from 0.5B to 14B results in a dramatic F-score increase from 3.6 to 32.6. This trend is pervasive across both Korean and multilingual families, confirming that parametric capacity remains a foundational factor for encoding and accessing factual knowledge.

5.2 Role of Pre-training Data Distribution

While performance generally scales with model size, our results suggest that the composition of pre-training data may play a more decisive role than parameter count. Korean-native models consistently outperform significantly larger multilingual counterparts. Notably, HCX SEED-14B achieves the highest F-score (32.6), doubling the performance of Llama 3.1-70B (15.7) despite being five times smaller. Similarly, the compact Kanana 1.5-2.1B (15.3) exceeds the massive Qwen 3-32B (10.6). Even Gemma 3-27B, the strongest multilingual baseline, lags behind smaller Korean models.

Although exact training data distributions are rarely disclosed by model developers, Korean community LLMs are generally reported to focus primarily on Korean and English during development. Therefore, a plausible hypothesis for their performance advantage is that their pre-training corpora contain a higher proportion of Korean data than general multilingual LLMs. Following prior observations such as those discussed by Chang et al. (2024), differences in pretraining corpus composition may influence the factual knowledge that models can effectively express. These findings suggest that high-quality, region-specific pre-training is a critical determinant of factual performance, often outweighing the benefits of generic scaling. This points to targeted data curation and localized LLM development as a promising direction for better capturing culturally grounded knowledge in non-English, low-resource language settings.

5.3 Persistence of Overconfidence

As shown in Table 1, higher factual performance generally corresponds to improved *ECE* values on KoSimpleQA. This relationship is further evidenced by Figure 3, which illustrates a strong correlation between a model’s F-score and its calibration. Notably, the calibration trends align with the performance patterns previously observed regarding model scaling and cultural grounding. For example, culturally specialized models, such as HCX SEED-14B (*ECE* 61.5) and Kanana 1.5-8B (*ECE* 64.1), exhibit significantly lower calibration error than larger models like Llama 3.1-70B (*ECE* 71.4). To investigate the specific nature of these calibration failures, we examine the distribution of verbalized confidence in Figure 4. The results reveal a systemic tendency toward overconfidence across the dataset; even when generating incorrect responses, models frequently assign a confidence level near 100%.

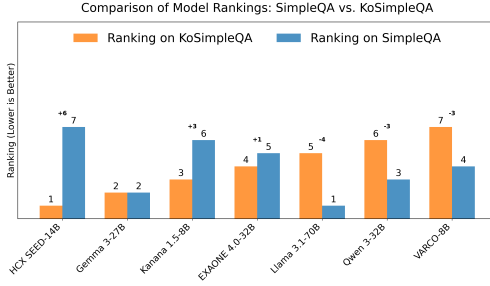


Figure 5: Rankings of the largest models in each LLM family on SimpleQA and KoSimpleQA.

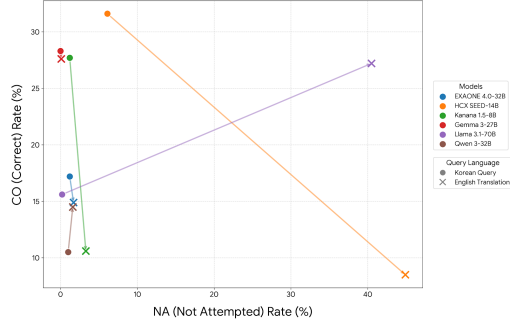


Figure 6: Impact of query language on model behavior. Arrows indicate the behavioral shift from Korean queries (circles) to English translations (crosses).

These findings highlight the need for overconfidence mitigation in culturally specialized contexts. Although our primary contribution lies in establishing reproducible benchmark analyses rather than introducing a new calibration algorithm, contextualizing these calibration failures paves the way for applying existing mitigation techniques (Wen et al., 2024) and serves as a foundational benchmark for future research in model calibration.

5.4 Comparison Between SimpleQA and KoSimpleQA

In Figure 5, we compare model rankings on English-centric SimpleQA versus our KoSimpleQA to assess cross-lingual factual alignment. We observed a notable *rank reversal*: Llama 3.1-70B drops from #1 on SimpleQA to #6 on KoSimpleQA, while HCX SEED-14B jumps from #8 to #1. This dramatic shift indicates that high performance on benchmarks reflecting knowledge mainly from one culture may not generalize to another culture. KoSimpleQA thus serves as a critical complement to existing evaluations, exposing the blind spots of high-resource benchmarks and properly validating models tailored for specific linguistic communities. Full results on SimpleQA are provided in Appendix F.

5.5 Same Knowledge, Different Behavior

While comparing the results on SimpleQA and KoSimpleQA, we observed substantial changes in metrics such as NA for certain models. To investigate whether these differences stem from language variation, we conduct an additional experiment. Specifically, we automatically translate the questions and answers of KoSimpleQA into English and produce a translated KoSimpleQA through a single round of human verification. As shown in Figure 6, even though only the language of the questions was changed, the models exhibit distinct behavioral shifts.

HCX SEED-14B shows a substantial decrease in CO accompanied by a large increase in NA, whereas Llama 3.1-70B exhibits pronounced increases in both CO and NA. The marked increase in NA is consistent with the observation reported by Ul Islam et al. (2025) that the degree of hallucination can vary across languages.

Notably, CO also changes significantly, confirming that a model’s ability to express internal knowledge can vary depending on the query language (Aggarwal et al., 2025). In this work, we use the term **cross-lingual knowledge gap** to refer strictly to these observed behavioral disparities in knowledge manifestation across languages, rather than to claim a single underlying cause. Such disparities may stem from multiple factors, including language-specific instruction tuning, refusal alignment, decoding behavior, or pre-training distributions.

6 Analysis of Reasoning LLMs

Recent advancements in LLMs have shown that explicit reasoning can significantly enhance performance in complex tasks. In this section, we analyze how such reasoning capabilities

Models	CO $\uparrow$	NA	IN $\downarrow$	CGA $\uparrow$	F-score $\uparrow$	ECE $\downarrow$
<i>Korean Community LLMs</i>						
EXAONE 4.0-1.2B-Think	1.4 (-0.2)	1.2 (-3.5)	97.4 (+3.7)	1.4 (-0.3)	1.4 (-0.2)	81.6 (+1.1)
EXAONE 4.0-32B-Think	16.1 (-1.1)	0.9 (-0.3)	83.0 (+1.4)	16.3 (-1.1)	16.2 (-1.1)	77.4 (+9.0)
HCX SEED-14B-Think	34.0 (+2.4)	1.6 (-4.5)	64.4 (+2.1)	34.5 (+0.9)	34.3 (+1.7)	55.8 (-5.7)
<i>Multilingual LLMs Supporting Korean</i>						
Qwen 3-0.6B-Think	0.8 (0.0)	1.4 (+0.7)	97.8 (-0.6)	0.8 (0.0)	0.8 (0.0)	81.1 (-3.9)
Qwen 3-1.7B-Think	2.4 (+0.8)	1.7 (+1.0)	95.9 (-1.8)	2.5 (+0.9)	2.4 (+0.8)	84.1 (-7.0)
Qwen 3-4B-Think	4.9 (+1.7)	1.6 (+0.6)	93.5 (-2.4)	5.0 (+1.8)	4.9 (+1.7)	82.0 (-6.9)
Qwen 3-8B-Think	7.8 (+2.5)	2.9 (+2.2)	89.2 (-4.8)	8.1 (+2.8)	7.9 (+2.6)	76.6 (-13.3)
Qwen 3-14B-Think	12.0 (+3.0)	2.4 (+1.3)	85.6 (-4.3)	12.3 (+3.2)	12.1 (+3.1)	73.2 (-11.5)
Qwen 3-32B-Think	12.5 (+2.0)	1.0 (0.0)	86.5 (-2.0)	12.6 (+2.0)	12.5 (+1.9)	74.6 (-7.3)

Table 2: Results of thinking models on KoSimpleQA compared to their base versions. Values in parentheses denote the delta (Δ) from the base model performance (blue: increase, red: decrease). Best scores are in **bold**.

interact with a model’s factual knowledge in the context of KoSimpleQA. We find that while reasoning cannot substitute for missing parametric knowledge, it serves as a powerful amplifier that helps models access and verify latent information.

6.1 Reasoning LLMs in Factual Knowledge

Table 2 presents the performance of models utilizing explicit reasoning (denoted as -Think). We observe an improvement in factual accuracy across most model families when reasoning is enabled, contrary to the intuition that reasoning would not be particularly helpful for factual question answering tasks. We further analyze this phenomenon in Section 6.2.

For the HCX SEED-14B model, the F-score increases from 32.6 (Base) to 34.3 (Think), achieving the highest accuracy in our benchmark. Beyond raw accuracy, a notable finding is the improvement in calibration. With reasoning enabled, HCX SEED-14B-Think reaches an *ECE* of 55.8. This suggests that the internal "thinking" process allows the model to better self-evaluate the correctness of its retrieved facts. Nevertheless, for models with lower parametric capacity, such as the Qwen 3-0.6B and 1.7B variants, reasoning provides negligible gains or even slightly inflates *ECE*.

6.2 Eliciting Suppressed Knowledge

Section 5.4 demonstrates that a model’s knowledge may not be fully expressed depending on the language of the query. Table 3 further illustrates how reasoning interacts with this phenomenon. Specifically, we observe that reasoning yields substantially larger gains in a model’s weaker language, while providing only marginal improvements in its dominant language. For instance, HCX SEED-14B, which is stronger in Korean, shows a modest gain of 2.4 in Korean but a significantly larger gain of 6.8 in English. Conversely, Qwen 3-32B, which is stronger in English, exhibits a gain of 0.4 in English but 2.0 in Korean.

These results suggest that reasoning does not primarily enhance factual question answering performance, but rather serves to mitigate cross-lingual knowledge gaps. This observation aligns with our intuition that, for tasks such as KoSimpleQA that emphasize knowledge memorization rather than reasoning, explicit reasoning provides limited additional benefit when the underlying knowledge is already accessible. Qualitative examples illustrating how reasoning can elicit latent knowledge are provided in Appendix B (Table 6).

Interestingly, Exaone does not exhibit substantial improvements with reasoning. We hypothesize that this discrepancy arises from differences in training methodologies for reasoning capabilities. This suggests that while reasoning has the potential to alleviate cross-lingual knowledge gaps, its effectiveness is contingent on how such capabilities are learned and integrated during training.

Another interesting finding is that reasoning substantially reduces NA rates, particularly in settings where models exhibit excessive refusal behavior. For example, HCX SEED-14B

Model	Metric	Korean Prompt		Translated (English)	
		Base	Think	Base	Think
EXAONE 4.0-32B	CO	17.2	16.1	14.9	15.3
	NA	1.2	0.9	1.7	1.6
HCX SEED-14B	CO	31.6	34.0	8.5	15.3
	NA	6.1	1.6	44.9	0.6
Qwen 3-32B	CO	10.5	12.5	14.5	14.9
	NA	1.0	1.0	1.6	2.1

Table 3: Effect of Reasoning in the Translation Experiment: Comparison of CO and NA.

shows a reduction of 44.3 in NA on English data. This observation is consistent with prior work (Yin et al., 2025), which reports that enabling reasoning can lead to decreased refusal behavior.

7 Conclusion

In this work, we introduce KoSimpleQA, a benchmark designed to evaluate short-form factuality within Korean cultural contexts. Our evaluation across diverse LLMs reveals that current models still struggle with factual accuracy in Korean, exhibiting performance rankings that diverge significantly from English-centric benchmarks. These findings highlight both the substantial room for improvement in non-English, low-resource language settings and the importance of localized, culturally grounded evaluation. Furthermore, our analysis of reasoning models indicates its potential to mitigate cross-lingual knowledge gap. We expect KoSimpleQA to serve as a vital foundation for developing more culturally attuned and factually reliable language models.

References

- Tushar Aggarwal, Kumar Tanmay, Ayush Agrawal, Kumar Ayush, Hamid Palangi, and Paul Pu Liang. Language models’ factuality depends on the language of inquiry, 2025. URL <https://arxiv.org/abs/2502.17955>.
- Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4623–4637, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.421. URL <https://aclanthology.org/2020.acl-main.421/>.
- Yunju Bak, Hojin Lee, Minho Ryu, Jiyeon Ham, Seungjae Jung, Daniel Wontae Nam, Taeyeon Eo, Donghun Lee, Doohae Jung, Boseop Kim, et al. Kanana: Compute-efficient bilingual language models. *arXiv preprint arXiv:2502.18934*, 2025.
- Nishant Balepur, Rachel Rudinger, and Jordan Lee Boyd-Graber. Which of these best describes multiple choice evaluation with LLMs? a) forced B) flawed C) fixable D) all of the above. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3394–3418, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.169. URL <https://aclanthology.org/2025.acl-long.169/>.
- Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 749–775, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.44. URL <https://aclanthology.org/2024.acl-long.44/>.

Sourav Banerjee, Ayushi Agarwal, and Saloni Singla. Llms will always hallucinate, and we need to live with this. In *Intelligent Systems Conference*, pp. 624–648. Springer, 2025.

Hoyeon Chang, Jinho Park, Seonghyeon Ye, Sohee Yang, Youngkyung Seo, Du-Seong Chang, and Minjoon Seo. How do large language models acquire factual knowledge during pretraining? *Advances in neural information processing systems*, 37:60626–60668, 2024.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Pettrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Pöder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL <https://arxiv.org/abs/2503.19786>.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koepl, Preslav Nakov, and Iryna Gurevych. A survey of confidence estimation and calibration in large language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6577–6595, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.366. URL <https://aclanthology.org/2024.naacl-long.366/>.

- Lukas Haas, Gal Yona, Giovanni D’Antonio, Sasha Goldshtein, and Dipanjan Das. Simpleqa verified: A reliable factuality benchmark to measure parametric knowledge, 2025. URL <https://arxiv.org/abs/2509.07968>.
- Yancheng He, Shilong Li, Jiaheng Liu, Yingshui Tan, Weixun Wang, Hui Huang, Xingyuan Bu, Hangyu Guo, Chengwei Hu, Boren Zheng, et al. Chinese simpleqa: A chinese factuality evaluation for large language models. *arXiv preprint arXiv:2411.07140*, 2024.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- Bowen Jiang, Runchuan Zhu, Jiang Wu, Zinco Jiang, Yifan He, Junyuan Gao, Jia Yu, Rui Min, Yinfan Wang, Haote Yang, et al. Evaluating large language model with knowledge oriented language specific simple question answering. *arXiv preprint arXiv:2505.16591*, 2025.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 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)*, pp. 3335–3346, 2024.
- Jiyoung Lee, Minwoo Kim, Seungho Kim, Junghwan Kim, Seunghyun Won, Hwaran Lee, and Edward Choi. KorNAT: LLM alignment benchmark for Korean social values and common knowledge. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 11177–11213, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.666. URL <https://aclanthology.org/2024.findings-acl.666/>.
- LG AI Research, Kyunghoon Bae, Eunbi Choi, Kibong Choi, Stanley Jungkyu Choi, Yemuk Choi, Seokhee Hong, Junwon Hwang, Hyojin Jeon, Kijeong Jeon, Gerrard Jeongwon Jo, Hyunjik Jo, Jiyeon Jung, Hyosang Kim, Joonkee Kim, Seonghwan Kim, Soyeon Kim, Sunkyoung Kim, Yireun Kim, Yongil Kim, Youchul Kim, Edward Hwayoung Lee, Haeju Lee, Honglak Lee, Jinsik Lee, Kyungmin Lee, Sangha Park, Yongmin Park, Sihoon Yang, Heuiyeon Yeen, Sihyuk Yi, and Hyeongu Yun. Exaone deep: Reasoning enhanced language models, 2025. URL <https://arxiv.org/abs/2503.12524>.
- Wangyue Li, Liangzhi Li, Tong Xiang, Xiao Liu, Wei Deng, and Noa Garcia. Can multiple-choice questions really be useful in detecting the abilities of LLMs? In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 2819–2834, Torino, Italia, May 2024. ELRA and ICCL. URL <https://aclanthology.org/2024.lrec-main.251/>.
- Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL <https://aclanthology.org/2022.acl-long.229/>.
- Nishanth Madhusudhan, Sathwik Tejaswi Madhusudhan, Vikas Yadav, and Masoud Hashemi. Do llms know when to not answer? investigating abstention abilities of large language models, 2024. URL <https://arxiv.org/abs/2407.16221>.
- Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAI’15*, pp. 2901–2907. AAAI Press, 2015. ISBN 0262511290.

- Shrutika Singh, Anton Alyakin, Daniel Alexander Alber, Jaden Stryker, Ai Phuong S Tong, Karl Sangwon, Nicolas Goff, Mathew de la Paz, Miguel Hernandez-Rovira, Ki Yun Park, Eric Claude Leuthardt, and Eric Karl Oermann. It is too many options: Pitfalls of multiple-choice questions in generative ai and medical education, 2025. URL <https://arxiv.org/abs/2503.13508>.
- Guijin Son, Hanwool Lee, Suwan Kim, Huiseo Kim, Jae cheol Lee, Je Won Yeom, Jihyu Jung, Jung woo Kim, and Songseong Kim. 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)*, pp. 7993–8007, 2024.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *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)*, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL <https://aclanthology.org/N19-1421/>.
- Zhi Rui Tam, Cheng-Kuang Wu, Chieh-Yen Lin, and Yun-Nung Chen. None of the above, less of the right parallel patterns in human and LLM performance on multi-choice questions answering. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20112–20134, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1031. URL <https://aclanthology.org/2025.findings-acl.1031/>.
- NAVER Cloud HyperCLOVA X Team. Hyperclova x think technical report. *arXiv preprint arXiv:2506.22403*, 2025.
- Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.330. URL <https://aclanthology.org/2023.emnlp-main.330/>.
- Saad Obaid Ul Islam, Anne Lauscher, and Goran Glavaš. How much do LLMs hallucinate across languages? on realistic multilingual estimation of LLM hallucination. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 29077–29098, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1481. URL <https://aclanthology.org/2025.emnlp-main.1481/>.
- Xiaonan Wang, Jinyoung Yeo, Joon-Ho Lim, and Hansaem Kim. KULTURE bench: A benchmark for assessing language model in Korean cultural context. In Nathaniel Oco, Shirley N. Dita, Ariane Macalinga Borlongan, and Jong-Bok Kim (eds.), *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pp. 914–927, Tokyo, Japan, December 2024. Tokyo University of Foreign Studies. URL <https://aclanthology.org/2024.paclic-1.88/>.
- Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*, 2024.
- Bingbing Wen, Chenjun Xu, Robert Wolfe, Lucy Lu Wang, Bill Howe, et al. Mitigating overconfidence in large language models: A behavioral lens on confidence estimation and calibration. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*, 2024.

- Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=gjeQKFxFpZ>.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.
- Qingyu Yin, Chak Tou Leong, Linyi Yang, Wenxuan Huang, Wenjie Li, Xiting Wang, Jaehong Yoon, YunXing, XingYu, and Jinjin Gu. Refusal falls off a cliff: How safety alignment fails in reasoning?, 2025. URL <https://arxiv.org/abs/2510.06036>.
- Kang Min Yoo, Jaegeun Han, Sookyo In, Heewon Jeon, Jisu Jeong, Jaewook Kang, Hyunwook Kim, Kyung-Min Kim, Munhyong Kim, Sungju Kim, et al. Hyperclova x technical report. *arXiv preprint arXiv:2404.01954*, 2024.
- Dongkeun Yoon, Seungone Kim, Sohee Yang, Sunkyoung Kim, Soyeon Kim, Yongil Kim, Eunbi Choi, Yireun Kim, and Minjoon Seo. Reasoning models better express their confidence. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL <https://openreview.net/forum?id=rbBtoVnduo>.

A Statistics and examples of KoSimpleQA

We present the detailed statistics of the KoSimpleQA dataset in Table 4. The benchmark comprises 938 high-quality question-answer pairs, manually curated to evaluate the factual accuracy of the LLMs. As shown in the table, the questions are designed to be concise, yet highly specific, with an average length of 29.22 characters. The corresponding ground-truth answers are brief and average only 5.24 characters. This focus on brevity ensures that the evaluation effectively decouples factual correctness from stylistic verbosity, aligning with our core objective of measuring pure factuality.

Statistics	Value	Statistics	Value
Question Length		Answer Length	
- Maximum	71	- Maximum	25
- Minimum	9	- Minimum	1
- Average	29.22	- Average	5.24

Table 4: Dataset statistics of KoSimpleQA.

Table 5 provides representative examples in topics and answer types. KoSimpleQA spans a wide range of topics, including Art, Geography, and Science & Technology, with answers categorized into distinct types such as Person, Number, and Place. As demonstrated by these examples, each query requires the retrieval of specific, unambiguous facts. For clarity in this report, we provide English translations alongside the original Korean text, though the primary evaluation is conducted using the native Korean prompts.

Question	Answer	Topic	Answer Type
Which number is Director Bong Joon-ho’s Parasite among South Korea’s films that have surpassed 10 million admissions? <i>봉준호 감독의 ‘기생충’은 대한민국 몇 번째 천만영화인가요?</i>	26th <i>26번째</i>	Art	Number
Which university in South Korea was the first to connect to the Internet? <i>대한민국 최초로 인터넷이 연결된 대학교는?</i>	Seoul Nat’l Univ. <i>서울대학교</i>	Science & Technology	Place
What is the depth of the deepest point in South Korean waters in meters? <i>대한민국 해역 중 가장 깊은 곳의 수심은 몇 m인가요?</i>	2,985m	Geography	Number
Which show won Program of the Year at the 2022 MBC Entertainment Awards? <i>2022 MBC 방송 연예대상에서 올해의 예능 프로그램상을 수상한 프로그램은?</i>	I Live Alone <i>나 혼자 산다</i>	TV Show	Other
Who was the first non-Italian to receive the International Puccini Award? <i>이탈리아인이 아닌 사람으로서는 처음으로 국제 푸치니상을 수상한 사람은?</i>	Sumi Jo <i>조수미</i>	Music	Person
In what year was the soju brand ‘Sunyang’ released? <i>제품명이 ‘선양’인 소주는 몇 년도에 출시됐어?</i>	2023 <i>2023년</i>	Other	Date

Table 5: Examples from KoSimpleQA. English translations are provided with the original Korean text displayed in *italics*.

B Qualitative Analysis of Reasoning mode

To understand *when* reasoning succeeds or fails, we examine representative traces. Table 6 presents a case study on “1st-Gen New Towns,” illustrating a central qualitative finding: reasoning is not a substitute for parametric knowledge—it is an amplifier.

Case A: Reasoning with cultural knowledge (HCX). The base HCX SEED-14B response exhibits *attribute hallucination*, conflating properties of two different cities (Sanbon and Ilsan). When reasoning is enabled, the model explicitly enumerates the canonical set of “1st-Gen New Towns” (e.g., Bundang, Ilsan) and uses this structure to filter candidates, ultimately identifying “Ilsan.” Here, reasoning successfully *elicits* latent knowledge and constrains the output.

Case B: Reasoning without cultural knowledge (Qwen). In contrast, Qwen 3-32B appears to lack the relevant cultural/policy ontology. Its reasoning trace attempts a logical derivation but operates on incorrect premises, eventually fabricating a non-existent “Hannam

Model	Base (Direct Generation)	Think (Chain-of-Thought)
<i>Q: Which 1st-Gen New Town has low density and many garden houses? (Target: Ilsan)</i>		
HCX SEED-14B	[Base Answer] "Among 1st-gen new towns... **Sanbon** fits this description... and many garden houses are located there." Output: Sanbon (✗) Error: <i>Attribute Hallucination.</i> Incorrectly attributes the target features (low density, garden houses) to Sanbon, conflating it with Ilsan.	[Thinking Process] "Thought: Among 1st-gen new towns (Bundang, Ilsan, Pyeongchon, Sanbon, Jungdong)... **Ilsan** has the lowest density and garden houses in Jeongbalsan..." Output: Ilsan (✓) Key: <i>Domain Knowledge Retrieval.</i> Accurately recalled the policy-defined list.
Qwen 3-32B	[Base Answer] "Among the 1st-gen new towns centered in the **Seoul outskirts and Gyeonggi area (Uiwang, Siheung, Yongin, Bucheon, Seongnam, etc.)**, **Yongin Phase 1** is famous..." Output: Yongin Phase 1 (✗) Error: <i>Ontological Collapse.</i> Redefines '1st-Gen New Town' as generic suburbs, failing to distinguish the specific 5 policy towns.	[Thinking Process] "I need to recall 1st-generation new towns... Examples include Seohae, Jeongnung, Jangan-gu... Hannam New Town is the most likely." Output: Hannam New Town (✗) Error: <i>Reasoning on False Premises.</i> The reasoning module inherits the same ontological gap, hallucinating lists instead of retrieving facts.

Table 6: Qualitative Analysis of Reasoning Strategies. We present the English translation first for readability, followed by the original Korean text in gray. **Row 1 (HCX)** demonstrates how Chain-of-Thought reasoning corrects the base model’s attribute hallucination by retrieving the precise list of towns. **Row 2 (Qwen)** illustrates an ontological collapse where the model fails to define the category ‘1st-Gen New Town’ in both modes.

New Town.” This reflects an **ontological collapse**: without the requisite facts in the model parameters, reasoning can simply produce more plausible-sounding hallucinations.

C Results of Translated KoSimpleQA

Table 7 provides the comprehensive evaluation results for the English-translated version of KoSimpleQA across all evaluated models.

Models	CO $\uparrow$	NA	IN $\downarrow$	CGA $\uparrow$	F-score $\uparrow$	ECE $\downarrow$
<i>Korean Community LLMs</i>						
VARCO-8B	9.0 $\pm$ 6.4	6.3 $\pm$ 3.2	84.7 $\pm$ 9.6	9.3 $\pm$ 1.0	9.3 $\pm$ 1.1	78.1 $\pm$ 0.0
Kanana 1.5-2.1B	5.5 $\pm$ 0.6	9.4 $\pm$ 1.9	85.1 $\pm$ 1.3	6.1 $\pm$ 0.6	5.8 $\pm$ 0.6	83.0 $\pm$ 1.3
Kanana 1.5-8B	10.6 $\pm$ 0.7	3.3 $\pm$ 1.5	86.1 $\pm$ 2.1	10.9 $\pm$ 0.9	10.7 $\pm$ 0.8	61.5 $\pm$ 1.2
EXAONE 4.0-1.2B	1.5 $\pm$ 0.2	2.9 $\pm$ 2.1	95.6 $\pm$ 2.3	1.6 $\pm$ 0.2	1.6 $\pm$ 0.2	90.1 $\pm$ 0.2
EXAONE 4.0-32B	14.9 $\pm$ 0.3	1.7 $\pm$ 0.5	83.5 $\pm$ 0.2	15.1 $\pm$ 0.3	15.0 $\pm$ 0.3	66.5 $\pm$ 0.3
EXAONE 4.0-32B-Think	15.3 $\pm$ 1.7	1.6 $\pm$ 0.8	83.2 $\pm$ 1.8	15.5 $\pm$ 1.7	15.4 $\pm$ 1.7	72.1 $\pm$ 1.8
HXC SEED-0.5B	1.2 $\pm$ 0.3	2.7 $\pm$ 0.2	96.1 $\pm$ 0.3	1.3 $\pm$ 0.3	1.3 $\pm$ 0.3	76.1 $\pm$ 6.1
HXC SEED-1.5B	2.1 $\pm$ 0.4	18.6 $\pm$ 7.0	79.4 $\pm$ 6.6	2.5 $\pm$ 0.3	2.3 $\pm$ 0.4	79.8 $\pm$ 2.9
HXC SEED-3B	4.8 $\pm$ 0.6	9.7 $\pm$ 1.7	85.6 $\pm$ 1.6	5.3 $\pm$ 0.6	5.0 $\pm$ 0.6	71.5 $\pm$ 2.8
HXC SEED-14B	8.5 $\pm$ 0.8	44.9 $\pm$ 4.9	46.7 $\pm$ 4.5	15.4 $\pm$ 1.4	10.9 $\pm$ 0.9	65.7 $\pm$ 2.1
HXC SEED-14B-Think	15.3 $\pm$ 0.7	0.6 $\pm$ 0.4	84.1 $\pm$ 1.1	15.4 $\pm$ 0.8	15.3 $\pm$ 0.7	58.3 $\pm$ 0.2
<i>Multilingual LLMs Supporting Korean</i>						
Llama 3.1-8B	6.7 $\pm$ 1.1	69.7 $\pm$ 6.2	23.6 $\pm$ 5.2	22.1 $\pm$ 2.3	10.2 $\pm$ 1.2	58.1 $\pm$ 2.6
Llama 3.1-70B	27.2 $\pm$ 2.3	40.5 $\pm$ 9.8	32.4 $\pm$ 7.6	46.0 $\pm$ 3.8	34.0 $\pm$ 1.0	29.8 $\pm$ 3.6
Gemma 3-1B	3.7 $\pm$ 0.2	1.2 $\pm$ 1.2	95.1 $\pm$ 1.1	3.7 $\pm$ 0.2	3.7 $\pm$ 0.2	91.1 $\pm$ 0.2
Gemma 3-4B	8.7 $\pm$ 0.1	0.2 $\pm$ 0.2	91.1 $\pm$ 0.3	8.8 $\pm$ 0.1	8.8 $\pm$ 0.1	86.9 $\pm$ 0.1
Gemma 3-12B	18.4 $\pm$ 0.8	0.1 $\pm$ 0.2	81.4 $\pm$ 0.9	18.5 $\pm$ 0.8	18.5 $\pm$ 0.8	73.9 $\pm$ 0.7
Gemma 3-27B	27.6 $\pm$ 0.8	0.1 $\pm$ 0.2	72.3 $\pm$ 1.0	27.6 $\pm$ 0.8	27.6 $\pm$ 0.8	62.3 $\pm$ 0.6
Qwen 3-0.6B	1.6 $\pm$ 0.2	3.6 $\pm$ 1.0	94.8 $\pm$ 1.2	1.7 $\pm$ 0.2	1.7 $\pm$ 0.2	91.0 $\pm$ 2.7
Qwen 3-0.6B-Think	1.8 $\pm$ 0.3	4.8 $\pm$ 1.3	93.5 $\pm$ 1.1	1.9 $\pm$ 0.2	1.8 $\pm$ 0.2	85.7 $\pm$ 0.8
Qwen 3-1.7B	2.8 $\pm$ 0.4	2.6 $\pm$ 1.0	94.6 $\pm$ 0.8	2.9 $\pm$ 0.4	2.9 $\pm$ 0.4	80.3 $\pm$ 1.0
Qwen 3-1.7B-Think	3.3 $\pm$ 0.3	3.3 $\pm$ 1.0	93.4 $\pm$ 0.9	3.4 $\pm$ 0.3	3.4 $\pm$ 0.3	77.4 $\pm$ 0.1
Qwen 3-4B	6.7 $\pm$ 0.4	2.1 $\pm$ 0.8	91.2 $\pm$ 1.3	6.8 $\pm$ 0.5	6.7 $\pm$ 0.5	80.6 $\pm$ 0.7
Qwen 3-4B-Think	7.7 $\pm$ 0.4	3.1 $\pm$ 1.2	89.2 $\pm$ 0.8	8.0 $\pm$ 0.4	7.8 $\pm$ 0.4	69.6 $\pm$ 0.2
Qwen 3-8B	9.3 $\pm$ 0.1	2.1 $\pm$ 0.6	88.6 $\pm$ 0.5	9.5 $\pm$ 0.1	9.4 $\pm$ 0.1	79.6 $\pm$ 0.3
Qwen 3-8B-Think	10.3 $\pm$ 0.4	3.7 $\pm$ 0.7	86.0 $\pm$ 1.1	10.7 $\pm$ 0.5	10.5 $\pm$ 0.5	63.8 $\pm$ 0.6
Qwen 3-14B	12.8 $\pm$ 0.2	3.2 $\pm$ 0.3	83.9 $\pm$ 0.4	13.3 $\pm$ 0.3	13.0 $\pm$ 0.2	70.5 $\pm$ 0.3
Qwen 3-14B-Think	13.3 $\pm$ 0.3	3.6 $\pm$ 1.0	83.1 $\pm$ 0.9	13.8 $\pm$ 0.3	13.6 $\pm$ 0.3	57.7 $\pm$ 0.5
Qwen 3-32B	14.5 $\pm$ 0.2	1.6 $\pm$ 0.3	83.9 $\pm$ 0.5	14.8 $\pm$ 0.3	14.7 $\pm$ 0.2	70.1 $\pm$ 0.5
Qwen 3-32B-Think	14.9 $\pm$ 1.1	2.1 $\pm$ 0.3	83.0 $\pm$ 1.2	15.2 $\pm$ 1.1	15.1 $\pm$ 1.1	62.2 $\pm$ 1.0

Table 7: **Results of different models on Translated KoSimpleQA.** All metrics are reported on a scale of 0 to 100 for better readability. Values represent the average of three seeds $\pm$ standard deviation. Models are ordered by family. $\downarrow$ indicates lower-is-better, and $\uparrow$ indicates higher-is-better.

D Annotation Guidelines and Statistics

D.1 Annotation guidelines for KoSimpleQA

To ensure high-quality data generation, all annotators were provided with the following standardized criteria (refer to Table 8). The primary objective was to construct a fact-centric dataset that minimizes hallucination and maximizes objective verifiability.

[Task Requirements]

- Ground all inquiries in South Korean contexts, covering topics that require
 - localized knowledge.
- Verify all facts as of Dec 31, 2023. Avoid "drifting" facts (e.g., "Who is the
 - current champion?") unless a specific timeframe is explicitly provided to
 - ensure the answer remains invariant.
- Target "middle-to-high" difficulty questions that require specific expertise or
 - niche knowledge, avoiding trivial "Google-able" facts.

[Response and Reference Standards]

- Provide a minimum of two and a maximum of four reputable source URLs for each
 - question to ensure factual reliability. Strictly avoid the use of unverified
 - community-driven wikis (e.g., NamuWiki).
- Ensure responses contain only the core entity or fact (e.g., "Sejong the Great"
 - instead of "The answer is Sejong the Great"). Strip away all conversational
 - fillers or redundant context to maintain short-form conciseness.
- List all valid answers separated by commas if a question has multiple correct
 - responses, provided that the set of answers is exhaustive and remains invariant
 - over time.

[Topic Categorization]

Please generate questions for the following topics:

- Science & Tech: Natural sciences, engineering, and digital innovations.
- Politics: Domestic political systems, legislation, and administrative history.
- Art: Visual arts, architecture, and fine arts.
- Geography: Topography, administrative divisions, and local landmarks.
- Sports: Professional leagues, historical records, and athletic events.
- Music: From traditional Korean music (Gugak) to contemporary K-pop.
- TV Shows: Dramas, variety shows, and broadcasting history.
- History: Ancient history through to modern Korean historical events.
- Games: PC, mobile, and console gaming culture and industry facts.
- Other: Miscellaneous topics not covered by the above categories.

Table 8: **Annotation guidelines for KoSimpleQA.** The instructions were provided to all human workers.

D.2 Annotation Statistics

Subject to the vendor’s corporate policies and non-disclosure agreements (NDAs), we report the available annotation statistics below.

- **Annotator counts:** Each crowdsourcing platform hired two to four annotators.
- **Inter-annotator agreement:** Among the initial 1,501 collected samples, 338 cases were filtered out through cross-platform validation. This resulted in an inter-annotator agreement rate of 77.48%.
- **Rejection rate during experts verification process:** Among the remaining 1,164 samples after the cross-platform validation, 226 samples are additionally filtered during experts verification process, which results in a rejection rate of 19.42%. A detailed types of samples filtered during the experts verification process is provided in Section 3.2.

E Autorater Modification

Haas et al. (2025) highlight the limitations of the original SimpleQA prompt (Wei et al., 2024) specifically in grading numeric ranges and hedged responses. Beyond these issues, we identified an additional vulnerability: the autorater remained overly permissive in assigning the NOT_ATTEMPTED (NA) label to outputs that should be classified as **INCORRECT**. This occurs primarily in cases of **false refusals**, **irrelevant answers**, and **text degeneration**.

To address this, we further refine the evaluation prompt by incorporating targeted *counter-examples* that clarify the boundaries of the NA category. Specifically, we augment the prompt with three distinct failure cases, as illustrated in Table 9:

1. **False refusal:** Responses that fabricate justifications to claim ignorance (e.g., inventing phrases such as “since there are no clear records”) are labeled as **INCORRECT** rather than NA.
2. **Irrelevant answer:** Outputs that identify the wrong entity or fail to provide the specific attribute requested (e.g., providing a name when a date was asked) are classified as **INCORRECT**.
3. **Text degeneration:** Incoherent, corrupted, or mixed-language gibberish is strictly labeled as **INCORRECT**.

These refinements ensure that the NA metric reflects genuine instances where the model recognizes its lack of knowledge, rather than obscuring actual hallucinations or generation failures behind the NA label. The full revised prompt is provided in Appendix H. To verify the effectiveness of these refinements, we manually reviewed the outputs from the HCX SEED 14B think model and confirmed that while the initial prompt yielded a 75.27% agreement rate with human experts on NA vs. IN classification, our refined prompt boosted this agreement rate to 97.58%.

Error Type	Question	Model Response & Error Analysis	Change
False Refusal (Fabricates reasons to claim ignorance)	<i>Who was the first non-Italian to receive the International Puccini Award?</i> (Ref: Sumi Jo)	“It is generally referred to as ‘Persona Non Grata’, but since there are no clear records ..., it is difficult to reveal the exact name.” → <i>Falsely claims historical records are unavailable to justify the failure.</i>	NOT_ATTEMPTED ↓ INCORRECT
Irrelevant Answer (Speculation / Wrong Entity)	<i>What is the courtesy name (Ja) of the Joseon painter who entered government service with the help of Kim Chang-jip?</i> (Ref: Jeong Seon’s Ja ‘Wonbaek’)	“The painter helped by Kim Chang-jip is Kim Hong-do . He is a representative painter ..., and his artistic talents shone...” → <i>Identifies the wrong person (Kim Hong-do) and fails to provide the requested ‘courtesy name’.</i>	NOT_ATTEMPTED ↓ INCORRECT
Text Degeneration (Gibberish / Broken Text)	<i>Tell me the winning piece of Seong-jin Cho at the International Chopin Piano Competition.</i> (Ref: Chopin Piano Concerto No. 1)	“Player Seong-jin Cho is a basic pianist famous for ... waukee primaries program... his coodagonium is... [Broken Text]” → <i>Response collapses into incoherent gibberish and mixed languages.</i>	NOT_ATTEMPTED ↓ INCORRECT

Table 9: **Reclassification of Model Responses under Refined Evaluation Criteria.** We addressed the ambiguity where valid errors (e.g., false refusal, irrelevant answer, and text degeneration) were previously misclassified as NOT_ATTEMPTED. By enforcing stricter guidelines, these responses are now correctly identified as INCORRECT.

F Results on Original SimpleQA

Table 10 reports the comprehensive evaluation results on the original English SimpleQA dataset (Wei et al., 2024).

Models	CO	NA	IN	CGA	F-score
<i>Korean Community LLMs</i>					
VARCO-8B	4.9	11.7	83.4	5.5	5.2
Kanana 1.5-2.1B	1.7	17.7	80.6	2.0	1.8
Kanana 1.5-8B	3.6	8.1	88.3	3.9	3.7
EXAONE 4.0-1.2B	0.9	7.5	91.6	1.0	1.0
EXAONE 4.0-32B	4.1	5.5	90.4	4.3	4.2
HCX SEED-0.5B	0.6	10.9	88.5	0.7	0.6
HCX SEED-1.5B	0.4	56.2	43.3	1.0	0.6
HCX SEED-3B	1.6	22.9	75.5	2.1	1.8
HCX SEED-14B	2.4	61.7	35.8	6.3	3.5
<i>Multilingual LLMs Supporting Korean</i>					
Gemma 3-1B	1.9	2.1	96.0	2.0	1.9
Gemma 3-4B	4.2	0.3	95.5	4.2	4.2
Gemma 3-12B	5.5	0.3	94.2	5.6	5.6
Gemma 3-27B	9.5	0.3	90.2	9.5	9.5
Llama 3.1-8B	1.7	89.0	9.4	15.1	3.0
Llama 3.1-70B	12.9	65.3	21.8	37.1	19.1
Qwen 3-0.6B	1.0	5.8	93.2	1.1	1.0
Qwen 3-1.7B	2.1	3.8	94.2	2.1	2.1
Qwen 3-4B	3.1	3.7	93.1	3.2	3.2
Qwen 3-8B	4.0	3.9	92.1	4.2	4.1
Qwen 3-14B	5.2	4.5	90.4	5.4	5.3
Qwen 3-32B	5.7	1.6	92.7	5.8	5.8

Table 10: Performance comparison on the original English SimpleQA (Wei et al., 2024). Evaluation was conducted using the original prompt without modification.

G Calibration Prompt

Table 11 outlines the prompt template used to elicit both the model’s answer and its associated confidence score. We explicitly instruct the model to quantify its confidence on a scale of 0 to 100 and request the output in JSON format to ensure structured parsing. The table provides the original Korean prompt utilized in our experiments alongside its English translation for reference.

Instruction (Korean)
질문은 다음과 같습니다: <pre>{question}</pre> 이 질문에 대한 가장 적절한 답변과 0점에서 100점까지의 척도로 자신감을 표현해 주세요. 다음 JSON 형식으로 답변해 주세요: <pre>{ "answer": "당신의 답변", "confidence_score": 당신의 자신감 점수 }</pre>
Translation (English)
<i>The question is as follows:</i> <pre>{question}</pre> <i>Please provide the most appropriate answer and a confidence score on a scale of 0 to 100 for this question. Please answer in the following JSON format:</i> <pre>{ "answer": "Your Answer", "confidence_score": Your Confidence Score }</pre>

Table 11: Calibration Prompt template

H Korean SimpleQA Evaluation Prompt

The following is the updated evaluation prompt for Korean SimpleQA. Modifications introduced by SimpleQA Verified (Haas et al., 2025) relative to the original SimpleQA (Wei et al., 2024) are highlighted in red, while our additional changes are highlighted in blue.

Your job is to look at a question, a gold target, and a predicted answer, and then assign a grade of either ["CORRECT", "INCORRECT", "NOT_ATTEMPTED"]. First, I will give examples of each grade, and then you will grade a new example.

The following are examples of CORRECT predicted answers.

Question: What are the names of Barack Obama's children?
 Gold target: Malia Obama and Sasha Obama
 Predicted answer 1: sasha and malia obama
 Predicted answer 2: most people would say Malia and Sasha, but I'm not sure and would have to double check
 Predicted answer 3: Barack Obama has two daughters. Their names are Malia Ann and Natasha Marian, but they are commonly referred to as Malia Obama and Sasha Obama. Malia was born on July 4, 1998, and Sasha was born on June 10, 2001.
 Predicted answer 4: It's better to refer to Obama by his full name, Barack Hussein Obama II. He has two daughters: Malia and Sasha.

These predicted answers are all CORRECT because:

- They fully contain the important information in the gold target.
- They do not contain any information that contradicts the gold target.
- Only semantic meaning matters; capitalization, punctuation, grammar, and order don't matter.
- Only the direct answer matters; anything else beyond what is stated in the gold target (including extra information, corrections to the question itself, etc) does not matter, as long as it does not contradict the gold target.
- Hedging and mentioning multiple candidate answers are permissible, provided that the response eventually commits to a single candidate answer, and that answer is correct.

The following are examples of INCORRECT predicted answers.

Question: What are the names of Barack Obama's children?
 Gold target: Malia and Sasha
 Predicted answer 1: Malia.
 Predicted answer 2: Malia, Sasha, and Susan.
 Predicted answer 3: Barack Obama does not have any children.
 Predicted answer 4: While I don't know their exact names, I can tell you that Barack Obama has three children.
 Predicted answer 5: It's possible you may mean Betsy and Olivia. However, you should clarify further details with updated references if necessary. Is that the correct answer?
 Predicted answer 6: It may be the case that Obama's child is named James. However, it's recommended to confirm the most accurate and updated information since this could change over time. This model may not always reflect the most current information.
 Predicted answer 7: I don't know their names, and such records are usually hidden.

These predicted answers are all INCORRECT because:

- A factual statement in the answer contradicts the gold target.
- A hedged statement (e.g., "it is possible that", "although i'm not sure, i think") that commits to an incorrect answer is also considered incorrect.
- Predictions that claim not to know the answer but also provide unsupported or invented reasons must be graded as INCORRECT, not NOT_ATTEMPTED.

The following are examples of NOT_ATTEMPTED predicted answers.

Question: What are the names of Barack Obama's children?
 Gold target: Malia and Sasha
 Predicted answer 1: I don't know.

Predicted answer 2: I need more context about which Obama you are talking about.
 Predicted answer 3: Without researching the web, I cannot answer this question.
 However, I can tell you that Barack Obama has two children.
 Predicted answer 4: Barack Obama has two children. I know that one of them is Malia, but I'm not sure about the other one.
 Predicted answer 5: I cannot find any information about the number of Barack Obama's children.
 Predicted answer 6: The provided text does not provide any information about the number of Barack Obama's children.
 Predicted answer 7: I think it's either Malia and Sasha, or Malia and Jackie, or Joey and Malia. I'm not sure which one.
 ...

These predicted answers are all NOT_ATTEMPTED because:

- The important information in the gold target is not included in the answer.
- No statements in the answer contradict the gold target.
- Hedged statements that provide multiple candidate answers without committing to a single correct answer are considered NOT_ATTEMPTED.

Also note the following things:

- For grading questions where the answer is a number, the gold target will also specify the allowed range, and any predicted answer that falls in that range should be considered correct. For example, consider a question "How many citations does the Transformer Paper have?" with gold target "120k (acceptable range: anything between 118k and 122k)".
 - Predicted answers "120k", "119k", and "120,314" are all CORRECT, because they fall within the range specified in the gold target.
 - Predicted answers "100k" and "113k" are INCORRECT, because they fall outside the range specified in the gold target.
 - Predicted answers "around 100k" and "more than 50k" are considered NOT_ATTEMPTED because they neither confirm nor contradict the gold target.
- The gold target may contain more information than the question. In such cases, the predicted answer only needs to contain the information that is in the question.
 - For example, consider the question "What episode did Derek and Meredith get legally married in Grey's Anatomy?" with gold target "Season 7, Episode 20: White Wedding". Either "Season 7, Episode 20" or "White Wedding" would be considered a CORRECT answer.
- Do not punish predicted answers if they omit information that would be clearly inferred from the question.
 - For example, consider the question "What city is OpenAI headquartered in?" and the gold target "San Francisco, California". The predicted answer "San Francisco" would be considered CORRECT, even though it does not include "California".
 - Consider the question "What award did A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity win at NAACL '24?", the gold target is "Outstanding Paper Award". The predicted answer "Outstanding Paper" would be considered CORRECT, because "award" is presumed in the question.
 - For the question "What is the height of Jason Wei in meters?", the gold target is "1.73 m (acceptable range: anything between 1.72 m and 1.74 m)". The predicted answer "1.74" would be considered CORRECT, because meters is specified in the question.
 - For the question "What is the name of Barack Obama's wife?", the gold target is "Michelle Obama". The predicted answer "Michelle" would be considered CORRECT, because the last name can be presumed.
- Do not punish for typos in people's name if it's clearly the same name.
 - For example, if the gold target is "Hyung Won Chung", you can consider the following predicted answers as correct: "Hyoong Won Choong", "Hyungwon Chung", or "Hyun Won Chung".
- Predicted answers should only be graded as NOT_ATTEMPTED when the model clearly and explicitly indicates that it does not know the answer, cannot answer, or lacks sufficient information, and does not provide any alternative answer, explanation, or guess.
 - Responses that express uncertainty but still attempt to answer (e.g., by speculating, describing, or giving partial explanations) must be graded as INCORRECT rather than NOT_ATTEMPTED.
- NOT_ATTEMPTED should only be used when the model explicitly states that it cannot answer or does not know, and does not offer any alternative answer, explanation, or speculation.

- If the response contains broken code/text, looks like a model failure (gibberish, repeated phrases), or is completely irrelevant to the question, it must be graded as INCORRECT, not NOT_ATTEMPTED.

Here is a new example. Simply reply with either CORRECT, INCORRECT, NOT ATTEMPTED. Don't apologize or correct yourself if there was a mistake; we are just trying to grade the answer.

Question: {question}

Gold target: {target}

Predicted answer: {predicted_answer}

Grade the predicted answer of this new question as one of:

A: CORRECT

B: INCORRECT

C: NOT_ATTEMPTED

Just return the letters "A", "B", or "C", with no text around it.