Cross-Lingual Factual Consistency of GPT-4o-mini: An Empirical Study of English and Korean

Seoha Lee
University of Tübingen
seoha.lee@student.uni-tuebingen.de

Abstract

Large language models are deployed globally, yet their factual behavior across languages remains relatively underexplored. This paper presents an empirical study of GPT-4o-mini's cross-lingual consistency on factual question answering, comparing English and Korean using 104 questions drawn from the MKQA benchmark (Longpre et al., 2021). During data preparation, we found that 46 of the 150 initially sampled Korean questions (31%) were malformed in the original crowd-sourced translations and had to be excluded, leaving a final sample of 104. We evaluate responses using strict string matching and multilingual semantic similarity, and show that both metrics underestimate true consistency for cross-script language pairs, with string matching doing so far more severely. Under multilingual semantic similarity, GPT-4o-mini produces consistent answers across the two languages for 15.4% of questions; manual error analysis of the inconsistent cases suggests the true consistency rate is approximately 46.2% once transliteration artifacts are accounted for. Korean accuracy is approximately three times lower than English. These results suggest that GPT-4o-mini does not treat languages equally for factual question answering.

1 Introduction

Large language models (LLMs) such as GPT-4o-mini are increasingly used across different languages. A reasonable expectation is that a model should give the same factual answer regardless of the language: if Paris is the capital of France, this should be true whether the question is asked in English or Korean. But LLMs are trained on text that is heavily dominated by English, while languages like Korean are much less represented (Blevins and Zettlemoyer, 2022). This raises a simple question: does the model's factual knowledge get worse or more inconsistent when queries move to less-represented languages?

This matters in practice. If a Korean-speaking user gets a different answer than an English speaker asking the same question, that is a reliability gap, one that would be invisible if the model is only ever tested in English. Despite this, there is not much empirical work specifically measuring cross-lingual *consistency* (as opposed to accuracy or translation quality).

This paper makes two contributions. First, we test GPT-4o-mini's factual consistency between English and Korean using 104 questions from the MKQA benchmark (Longpre et al., 2021). Second, we compare strict string matching with multilingual semantic similarity as evaluation metrics, and show that both metrics underestimate true consistency for language pairs that use different writing systems, with string matching doing so far more severely.

Our main findings: English accuracy is 42.3% and Korean accuracy is only 14.4%. Cross-lingual consistency is near zero under strict string matching (1.9%) but rises to 15.4% under semantic similarity. Manual error analysis suggests the true consistency rate is approximately 46.2%, meaning the 84.6% semantic inconsistency figure is best treated as an upper bound.

2 Related Work

Factual Knowledge in Language Models. Petroni et al. (2019) showed that language models store factual knowledge in their parameters and can retrieve it through prompts. Our study is similar in spirit but focuses on a generative LLM and measures consistency between languages rather than accuracy alone.

Training Data Composition. Blevins and

Zettlemoyer (2022) argue that the cross-lingual capabilities of English-pretrained models stem from non-English data present in their training corpora, and that this effect is weaker for languages like Korean that are less represented in English text. Our 2×2 accuracy table makes this concrete by separating cases where the model simply does not know the answer from cases where it knows in English but not Korean.

Dataset. The MKQA benchmark (Longpre et al., 2021), derived from Natural Questions and aligned across 26 languages, is the closest available resource for this kind of study. Prior work using MKQA has focused on retrieval-augmented systems; we use it to test a closed-book generative model, where any inconsistency must come from the model's own knowledge.

3 Methods

3.1 Dataset

We use MKQA (Longpre et al., 2021), a multilingual QA dataset with 10,000 questions derived from Natural Questions, each provided in 26 languages with gold answers. We focus on English and Korean because Korean uses a completely different writing system (Hangul), which makes it a good test case for cross-script evaluation.

We filtered questions to keep only those where both English and Korean fields were non-null and where the gold answer was at most 60 characters (to focus on short factual answers). This left 6,753 valid pairs, from which we randomly sampled 150 (seed=42).

During data preparation, we noticed that many Korean questions looked broken: some were just a single consonant, others were cut off mid-sentence. We ran two inspection passes and found that 46 of the 150 questions (31%) had malformed Korean translations. A few examples:

- ㅈ: a single Hangul consonant with no meaning
- "미국 대법": means "US Supreme" but has no question
- "윟": a corrupted Hangul character
- "second world countries같은게": mixed English and Korean

These were problems in the original MKQA data, not something we introduced. We removed all 46 across two inspection passes: the first pass identified the most clearly malformed questions (reducing the sample to n=126), and the second pass caught the remaining cases (reducing to the final n=104). We confirmed that the main results were stable across all three versions of the dataset. The final sample is 104 questions.

3.2 Model

We queried GPT-4o-mini via the OpenAI API. Each question was sent independently with no conversation history, so the model had to rely entirely on what it already knows. We used temperature=0 to make results reproducible and set max tokens to 60. The system prompt told the model to answer in one short phrase and in the language of the question.

3.3 Evaluation

We define cross-lingual consistency as the degree to which the model's factual responses remain stable across languages, measured independently of factual correctness; it captures specifically whether a knowledge difference exists between languages, not whether the knowledge itself is accurate.

We measure cross-lingual consistency with two metrics:

Strict matching. Both strings are lowercased and punctuation is removed. For accuracy evaluation, we check whether the gold answer appears as a substring of the response or vice versa. When applied to consistency (the comparison row in Table 3), we check whether the two normalized response strings are identical, since there is no short/long asymmetry between two free-form responses. This metric is simple but sensitive to any surface-level difference.

Semantic similarity. We encode both strings using `paraphrase-multilingual-MiniLM-L12-v2` (Reimers and Gurevych, 2020), a sentence embedding model that maps text from over 50 languages into the same vector space. We then compute cosine similarity between the two embeddings. Because the model was trained on multilingual data, semantically equivalent phrases in different

scripts (like “Paris” and “파리”) get similar embeddings, which makes it much better suited for cross-script comparison than string matching.

It is worth noting explicitly that the two metrics are not applied symmetrically across all tasks. Accuracy (Tables 1 and 2) is evaluated under strict matching, following standard practice in factual QA evaluation where the goal is verifying that the model’s response contains the correct answer string. Consistency (Table 3), on the other hand, is evaluated under semantic similarity. This asymmetry is deliberate: because the two responses being compared in the consistency task come from different writing systems, character-level matching would flag every cross-script pair as inconsistent regardless of meaning, making it unsuitable as a consistency measure for English–Korean.

We set the similarity threshold at 0.75. We started with 0.85 based on the model documentation, but then manually checked 20 borderline response pairs (similarity scores between 0.70 and 0.90) and judged whether each pair was saying the same thing. All pairs above 0.75 were genuine matches, and all four pairs that were not genuine matches fell below 0.75. So we used 0.75 as the final threshold. That said, 20 pairs is a fairly small sample for a threshold calibration, so this value should be treated as approximate rather than precisely determined (see also Section 5.3).

Figure 1 shows precision, recall, and F1 across thresholds from 0.70 to 0.95, computed on the manually labelled pairs. Precision is 1.0 at all thresholds (no false positives in the calibration sample), while recall decreases as the threshold rises. F1 peaks at τ=0.75 (F1 = 0.968), corroborating the manual inspection result.

Wilson score confidence intervals (95%) are reported throughout given the small sample size (n=104).

Why string matching is not enough for cross-script pairs. When the model answers in Korean, the response contains Hangul characters. The English response contains Latin characters. These share no characters at all, so string matching will always call them inconsistent - even when they mean exactly the same thing. Semantic similarity solves this because it works at the meaning level, not the character level. Our results show that semantic matching finds a consistency rate eight times higher than strict matching (15.4% vs. 1.9%), which shows that a lot of apparently inconsistent cases are actually the same answer written in different scripts.

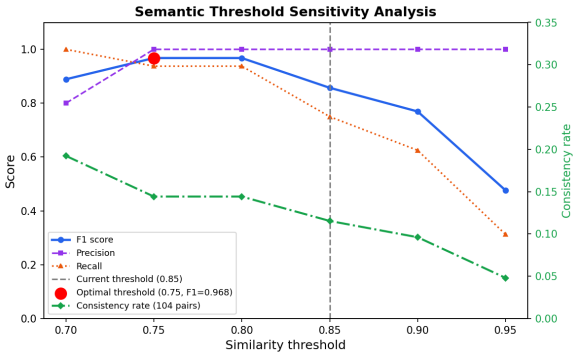

Figure 1: Semantic threshold sensitivity analysis. F1 peaks at τ=0.75 (F1 = 0.968); precision is 1.0 throughout and recall decreases monotonically as the threshold rises.

4 Results

4.1 Accuracy by Language

Accuracy is measured not to define consistency, but to decompose the sources of inconsistency. It distinguishes cases where the model lacks knowledge entirely from cases where knowledge exists in English but is inaccessible in Korean. Table 1 shows accuracy for English and Korean under strict matching. Figure 2 shows the same results as a bar chart.

Language	Strict Accuracy [95% CI]
English	42.3% [33.3–51.9%]
Korean	14.4% [8.9–22.4%]

Table 1: Factual accuracy by language under strict matching (n=104). CIs are Wilson score 95% intervals.

English accuracy is 42.3% and Korean accuracy is 14.4%, about three times lower. It is worth noting that even the English accuracy is not high, which matters for interpreting the results (see Section 5).

Table 2 breaks this down further by showing all four combinations of whether each language got the question right or wrong.

	KR Correct	KR Wrong
EN Correct	14 (13.5%)	30 (28.8%)
EN Wrong	1 (1.0%)	59 (56.7%)

Table 2: Accuracy cross-tabulation by language (n=104). Row totals: EN Correct 44 (42.3%), EN Wrong 60 (57.7%). Column totals: KR Correct 15 (14.4%), KR Wrong 89 (85.6%).

The biggest cell is "both wrong" at 56.7% (59 questions). For more than half the questions, the model got it wrong in both languages, meaning the problem is not really about Korean specifically, it is that the model just does not know these facts. The second biggest cell is "EN correct, KR wrong" at 28.8% (30 questions), which is the case where the model knows the answer in English but not Korean. Only 1 question had the opposite pattern (KR correct, EN wrong).

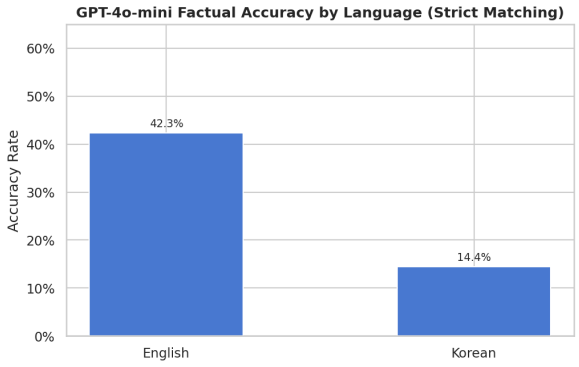

Figure 2: Factual accuracy by language.

4.2 Cross-Lingual Consistency

Table 3 shows how often the model gave the same answer in both languages, under strict and semantic matching. Figure 3 shows the same results visually.

Metric	n	Rate [95% CI]
Consistent (semantic, τ=0.75)	16 / 104	15.4% [9.7–23.5%]
Inconsistent (semantic)	88 / 104	84.6% [76.5–90.3%]
Strict matching (shown for comparison):		
Consistent (strict)	2 / 104	1.9% [0.5–6.7%]

Table 3: Cross-lingual consistency (EN↔KR). Primary measure is semantic similarity: 16 consistent + 88 inconsistent = 104. Strict matching is shown for comparison. CIs are Wilson score 95% intervals.

The primary measure is semantic similarity: 16 of 104 questions were consistent (15.4%), and 88 were inconsistent (84.6%). For comparison, strict matching found only 2 consistent cases (1.9%). The 14 additional cases caught by semantic matching are pairs where the answers mean the same thing but look different due to script differences or Korean grammatical endings.

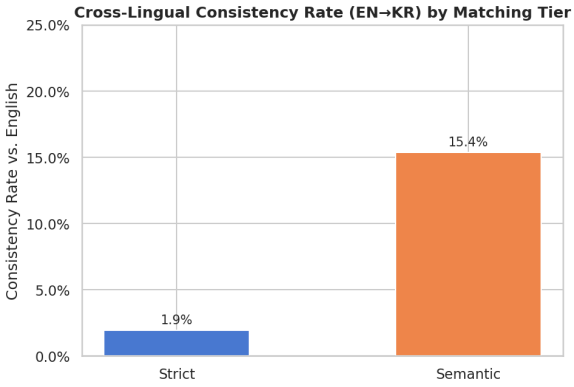

Figure 3: Cross-lingual consistency by metric (EN–KR).

4.3 Error Analysis

We manually looked at the 88 inconsistent cases and assigned each one to a category. Table 4 and Figure 4 show the results.

Category	Count	Percentage
knowledge_gap	52	59.1%
transliteration_error	32	36.4%
regional_variant	2	2.3%
language_bleed	2	2.3%
Total	**88**	**100%**

Table 4: Error categories for the 88 inconsistent cases (T=0.75).

The most common type was **knowledge_gap** (59.1%, 52 cases): the Korean response was simply wrong or irrelevant in a way that had nothing to do with script differences. For example, when asked who played Leatherface in Texas Chain Saw Massacre, the English response correctly named Gunnar Hansen while the Korean response named a completely different person. When asked when Sheetz started 24-hour service, English said 2004 and Korean said 1983.

The second most common type was **transliteration_error** (36.4%, 32 cases): both responses were actually saying the same thing,

but the embedding model scored them below the 0.75 threshold because of how Korean handles names and grammar. For example, "Roberta Flack" in English and "로버타 플랙이 불렀습니다" in Korean are the same answer, with the Korean adding a verb ending meaning "sang it."

The remaining 4 cases (4.6%) were split between **language_bleed** (model responded in English despite a Korean question) and **regional_variant** (both answers were plausible but referred to different regional versions of the same thing).

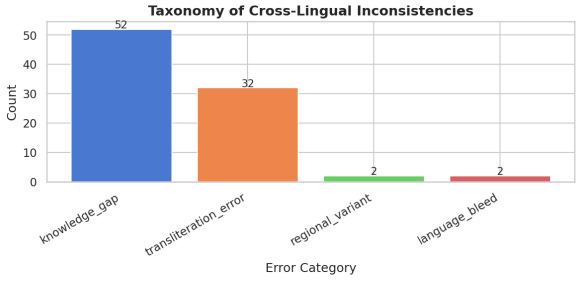

Figure 4: Error categories among the 88 inconsistent cases (T=0.75).

5 Discussion

5.1 What the Accuracy Gap Actually Means

Korean accuracy (14.4%) is about three times lower than English (42.3%). But the cross-tabulation in Table 2 shows this is not a simple story. The majority of failures (56.7%) are cases where the model got it wrong in both languages. This means the main problem is that the model simply does not know these facts; the language is almost beside the point for most questions.

That said, 28.8% of questions were ones where the model got it right in English but wrong in Korean. These are the genuinely cross-lingual failures, where the knowledge exists in the model but is not accessible in Korean. The fact that the reverse almost never happened (only 1 case of KR correct, EN wrong) confirms that Korean knowledge is essentially a subset of English knowledge in this model.

5.2 Interpreting the Inconsistency Rate

The 84.6% semantic inconsistency figure is an upper bound rather than a precise measurement. Of the 88 cases labeled inconsistent, 36.4% (32 cases) are transliteration artifacts: pairs where a human judge found the answers equivalent, but the embedding model scored them below the 0.75 threshold due to script differences or grammatical inflection. Counting these as consistent, the implied true inconsistency rate drops to approximately 53.8% and the implied true consistency rate rises to approximately 46.2% (48 of 104 questions). The semantic metric therefore also underestimates true consistency, not only strict matching, though it does so far less severely. Even at this lower estimate, a bilingual user would receive conflicting answers for roughly half of all questions.

The breakdown across error categories: 59.1% of inconsistent cases are genuine knowledge gaps where the Korean answer is factually wrong (e.g., different person named, different year given), while 36.4% are *transliteration_error* cases where the content is equivalent but expressed differently across scripts (e.g., "Roberta Flack" vs. a Korean response meaning "Roberta Flack sang it"). The remaining 4.6% are split between *language_bleed* and *regional_variant*.

5.3 Limitations

We only tested two languages and one model, so the results may not generalize. The semantic threshold was chosen by manually reviewing only 20 borderline pairs, which is a small sample (as noted in Section 3.3). The error analysis categories were assigned by a single annotator without an inter-annotator agreement check; the transliteration_error count (32 cases) directly drives the ∼46.2% implied true consistency estimate, so that figure should be treated with corresponding caution. The MKQA benchmark had data quality issues in the Korean translations, and the sample size of n=104 is small enough that the confidence intervals are quite wide. Finally, filtering for short answers may have made the questions easier than average, which could mean the accuracy numbers are slightly higher than they would be on a more representative sample.

5.4 Future Work

The most obvious next step would be testing other models like Claude or Gemini to see if the pattern we found is specific to GPT-4o-mini or more general. A larger study with more ques-

tions would also give more reliable results and allow more detailed breakdowns by question type. There is also a real gap in available benchmarks: there is currently no high-quality multilingual dataset for open-domain factual QA with verified answers across many languages. Building one would benefit the whole research community.

6 Conclusion

We tested GPT-4o-mini's factual consistency between English and Korean using 104 questions from MKQA. English accuracy (42.3%) was about three times higher than Korean accuracy (14.4%). Cross-lingual consistency was near zero under strict string matching (1.9%) and 15.4% under semantic similarity. Manual error analysis of the 88 semantically inconsistent cases found that 59.1% were genuine knowledge gaps and 36.4% were transliteration artifacts (cases where the two answers expressed the same content in different scripts but fell below the embedding threshold). Counting those as consistent raises the implied true consistency rate to approximately 46.2% (48/104), so the 84.6% semantic inconsistency figure is best interpreted as an upper bound rather than a precise measurement.

The main takeaway is that GPT-4o-mini does not treat English and Korean equally for factual question answering. Some of this gap is because the model does not know certain facts in either language, but a meaningful portion (28.8% of questions) represents cases where the model knows the answer in English but cannot produce it in Korean. We also showed that both strict string matching and semantic similarity underestimate true cross-lingual consistency for language pairs with different writing systems, with string matching doing so far more severely. Semantic similarity is the more appropriate proxy metric, but human error analysis remains necessary to identify residual transliteration artifacts that even the embedding model misclassifies as inconsistent.

References

Terra Blevins and Luke Zettlemoyer. 2022. Language contamination helps explain the cross-lingual capabilities of English pretrained models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1557–1572. Association for Computational Linguistics.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 2463–2473. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4512–4525. Association for Computational Linguistics.

A Prompt Template

The system prompt used for all queries was:

```
Answer the following factual question
in one short phrase or sentence. Do
not explain your reasoning. Do not
repeat the question.
```

The question was passed directly as the user message. All queries used temperature=0 and max tokens=60, with no retrieval augmentation or conversation history.