Understand, Solve and Translate:
Bridging the Multilingual Mathematical Reasoning Gap

Hyunwoo Ko[1]*   Guijin Son[1]*   Dasol Choi[2]*

OneLineAI[1]   Yonsei University[2]
hyunwooko@onelineai.com   spthsrbwls123@yonsei.ac.kr

Abstract

Large language models (LLMs) demonstrate exceptional performance on complex reasoning tasks. However, despite their strong reasoning capabilities in high-resource languages (e.g., English and Chinese), a significant performance gap persists in other languages. To investigate this gap in Korean, we introduce HRM8K, a benchmark comprising 8,011 English-Korean parallel bilingual math problems. Through systematic analysis of model behaviors, we identify a key finding: these performance disparities stem primarily from difficulties in comprehending non-English inputs, rather than limitations in reasoning capabilities. Based on these findings, we propose UST (Understand, Solve, and Translate), a method that strategically uses English as an anchor for reasoning and solution generation. By fine-tuning the model on 130k synthetically generated data points, UST achieves a 10.91% improvement on the HRM8K benchmark and reduces the multilingual performance gap from 11.6% to 0.7%. Additionally, we show that improvements from UST generalize effectively to different Korean domains, demonstrating that capabilities acquired from machine-verifiable content can be generalized to other areas. We publicly release the benchmark, training dataset, and models[1].

1 Introduction

Large language models (LLMs) have made remarkable progress in reasoning tasks, often surpassing expert human performance (OpenAI, 2024; Anthropic, 2024). However, this exceptional reasoning capability is primarily observed in high-resource languages, with significant performance gaps in lower-resource languages (Huang et al., 2023; Li et al., 2024). This disparity likely stems from LLMs' difficulty in transferring their foundational

capabilities, including reasoning skills learned in high-resource languages like English or Chinese, to lower-resource languages (Chen et al., 2023; Dubey et al., 2024).

To investigate this gap in Korean mathematical reasoning, we introduce HRM8K, a bilingual benchmark comprising 8,011 questions in both Korean and English. The questions are carefully curated from existing benchmarks (Cobbe et al., 2021; Hendrycks et al., 2021; Gao et al., 2024) and Korean examinations to create a perfectly parallel evaluation structure. Through systematic evaluation on HRM8K, we reveal that the performance gap primarily stems from difficulties in comprehending non-English inputs, rather than limitations in reasoning capabilities. This finding challenges previous studies that suggest using English chain-of-thought (CoT) reasoning for multilingual questions (Shi et al., 2022), as we show that LLMs are heavily influenced by the input language itself.

Based on these insights, we propose UST (Understand, Solve, and Translate), a training method that strategically uses English as an anchor for reasoning and solution generation. Our approach builds on recent findings that LLMs effectively use English as a pivot language for processing multilingual inputs (Zhong et al., 2024). By training on 130k synthetically generated instances, UST achieves a 10.91% improvement on the HRM8K benchmark and reduces the multilingual performance gap from 11.6% to 0.7%. Furthermore, we demonstrate that these improvements generalize beyond mathematics to different Korean domains, suggesting broader applications.

In summary, the main contributions of this work are as follows:

- We identify through systematic analysis that multilingual performance gaps primarily stem from input comprehension difficulties rather than reasoning limitations.

---

*Equal Contribution
[1]https://huggingface.co/HAERAE-HUB

- We propose UST, a training method that effectively leverages English reasoning capabilities for non-English inputs, demonstrating significant performance improvements.

- We introduce HRM8K, the first Korean mathematics reasoning benchmark with 8,011 parallel questions, enabling systematic evaluation of multilingual reasoning capabilities.

2 Related Work

Mathematics Benchmarks Mathematical reasoning has emerged as a crucial capability for language models (Hurst et al., 2024; Alibaba, 2024; Zhao et al., 2024c), leading to the development of numerous benchmarks and datasets (Ling et al., 2017; Amini et al., 2019; Patel et al., 2021; Saxton et al., 2019). Traditional datasets such as GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021) primarily target grade-school to undergraduate-level problems, while more recent efforts introduce Olympiad-level challenges (Zheng et al., 2021; He et al., 2024; Huang et al., 2024; Fang et al., 2024; Gao et al., 2024).

Although these benchmarks prove valuable for evaluating English-language mathematical reasoning, fewer resources exist for non-English or bilingual math problems (Shi et al., 2022; Chen et al., 2023; Wu et al., 2024). In the Korean context, most benchmarks emphasize language understanding (Park, 2021; Son et al., 2023a), general knowledge (Son et al., 2023b; Kim et al., 2024), or commonsense reasoning (Son et al., 2024b,a), with mathematics being largely underrepresented. While the Open Ko-LLM Leaderboard (Park et al., 2024a) has begun translating some popular English benchmarks into Korean, those translated sets are not publicly accessible. Meanwhile, KMMLU (Son et al., 2024c) includes only about 100 math problems, insufficient for broader evaluation. To address this gap, we propose HRM8K, a large-scale bilingual Korean-English math benchmark comprising 8,011 problems, covering both competition-level Korean questions and parallel translations of existing English benchmarks.

Multilingual Reasoning and Language Models Recent LLMs have shown remarkable performance in English (OpenAI, 2024; Anthropic, 2024; Touvron et al., 2023), but many still underperform in multilingual scenarios (Lai et al., 2024; Dubey et al., 2024). Such performance discrepancies are attributed to limited exposure to lower-resource languages during pre-training. Consequently, much research has focused on enhancing multilingual reasoning skills, including methods that explicitly use English as a 'pivot' for cross-lingual tasks (Zhao et al., 2024b; Zhu et al., 2024). For example, PLUG (Zhang et al., 2023) aligns internal reasoning in different languages, enabling the model to leverage stronger English reasoning for other languages.

Despite these developments, few works have thoroughly examined how to best calibrate reasoning between high- and low-resource languages in *complex mathematical* contexts. Some studies investigate altering the fraction of multilingual data in training (Anonymous, 2024) or conduct smaller-scale experiments on bilingual math tasks (Shi et al., 2022), yet a clear, large-scale solution remains elusive. Against this backdrop, our work introduces UST, a multilingual reasoning method that intentionally routes math problems in lower-resource languages through English-based reasoning. We show that this strategy drastically narrows the performance gap and advances multilingual math capabilities.

3 HRM8K

In this section, we introduce the composition of the HRM8K benchmark and explain its construction process. We also conduct a contamination check to ensure data quality. Detailed information about each dataset and post-processing methods are provided in Appendix A.

3.1 Benchmark Formulation

The HRM8K benchmark is a bilingual math dataset that consists of two major subsets: **Korean School Math** (KSM) and **Prior Sets**. Each subset is available in both Korean and English (see Table 1 for details).

KSM This subset contains 1,428 challenging math problems sourced from Korean Olympiads and competition-level exams, irrespective of the target age group. As a result, even questions originally intended for younger students still require substantial reasoning ability to solve. To collect these questions, the authors manually captured screenshots and applied GPT-4o's OCR to convert the text, followed by a thorough validity check. (See Appendix G for the OCR prompt.)

Category	Subset	# of Instances	Short Description
KSM	KMO	730	Mathematics competition for high school students in South Korea, top-performers are selected as representatives for the IMO. (KMO)
1.4K Total	KJMO	62	KMO for junior students, up to age 13 (KJMO)
	CSAT	210	Questions from the Korean national university entrance exam and official mock exams, we only filter questions that have an error rate exceeding 70%. (CSAT)
	KMS	82	Math olympiad for university students, organized by the Korean Mathematical Society (KMS)
	TQ	344	Question from the national exam for math teacher certification (TQ)
Prior Sets	GSM8K	1,319	Grade school math word problems created by human problem writers (Cobbe et al., 2021)
6.5K Total	MATH	2,885	Competition-level mathematics problems. We only include questions with numeric answers (Hendrycks et al., 2021)
	Omni-MATH	1,909	Olympiad-level problems collected from international and Chinese math competitions. We only include questions with numeric answers (Gao et al., 2024)
	MMMLU	470	The MMLU (Hendrycks et al., 2020) dataset translated by professional human translators (OpenAI, 2024)

Table 1: Summary of dataset sources used in HRM8K

Prior Sets This subset comprises 6,583 problems drawn from established English math benchmarks, including GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021), Omni-MATH (Gao et al., 2024), and MMMLU (OpenAI, 2024). To streamline translation and evaluation, we include only instances with numeric answers, excluding problems that require text-based, proof-oriented, or equation-based final answers. In particular, proof-type questions would necessitate a more complex LLM-as-a-Judge approach (Zheng et al., 2023; Shi et al., 2024; Park et al., 2024b) instead of simpler machine verification. Lastly, from the MMMLU dataset, we select only three math-related subsets: `abstract_algebra`, `college_mathematics`, and `high_school_mathematics`. Questions for the MMMLU dataset is multiple-choice question answering format.

3.2 Benchmark Construction

We translate all instances in both subsets into English and Korean using GPT-4o (Hurst et al., 2024), and then conduct human review to remove any inaccurate translations. This fully parallel design enables a more direct analysis of multilingual performance gaps. Furthermore, because KSM is translated from Korean to English and the other benchmarks are translated from English to Korean, we avoid depending solely on a single translation direction. This bidirectional approach also helps detect translation artifacts: if a particular pattern appears

only in one direction, it may be due to translation-related issues rather than the inherent difficulty of the questions. For further details on the dataset construction, please refer to Appendix A.

3.3 Contamination Check

Benchmark contamination, where evaluation questions appear in the model's pretraining data, is increasingly recognized as a key concern in LLM evaluations (Deng et al., 2023; Roberts et al., 2023). Large-scale internet-crawled corpora (Gao et al., 2020; Weber et al., 2024) raise the likelihood of model memorization, potentially leading to inflated performance on evaluation benchmarks (Zhang et al., 2024; Zhao et al., 2024a). Ensuring that a newly proposed benchmark is entirely uncontaminated is nearly impossible, as many companies do not disclose the specifics of their pretraining mixtures (Aryabumi et al., 2024; Mishra et al., 2024), and logit-based detection methods are not yet well-established (Xu et al., 2024). In this work, we make our best effort to verify whether the dataset is included in publicly available large-scale Korean corpora. The contamination check is focused on the KSM subset, the only subset that was crawled in this work.

To ensure that the KSM subset is not present in common pretraining corpora, we perform a contamination check against FineWeb-2 (Penedo et al., 2024), the biggest Korean corpora available. This dataset contains 58 million Korean documents, totaling 95 GB, collected by the CommonCrawl foun-

dation (2013–2024). We first identify 149 documents that match the external sources used to compile HRM8K. Then, we search these documents for exact string matches from KSM's questions; no matches were found. We hypothesize that this absence arises because the authors manually downloaded PDF or HWP files and selectively extracted questions, making them unlikely to appear in standard web crawls. Consequently, we conclude that the KSM problems are highly unlikely to have been seen during the LLMs' pretraining phase.

4 Multilingual Performance Gaps

A recurring observation in large language models (LLMs) is that performance can vary significantly depending on the language of the prompt, even if the underlying task remains the same. We confirm this phenomenon on the HRM8K benchmark: as shown in Table 2, simply changing both the input and reasoning language from Korean to English yields an 11% performance boost, suggesting a notable gap in multilingual reasoning.

This section further investigates the causes of this gap. We first describe our experimental design (Section 4.1), then analyze the results (Section 4.2), and finally explore how multi-step prompting might mitigate these issues (Section 4.3).

4.1 Experimental Design

Let a model's final performance P be determined by two factors: the language of the *input* (L_{input}) and the language used for *reasoning* (L_{reason}). Formally,

$$P = f(L_{input}, L_{reason}).$$

In the context of solving Korean math problems, there are two key requirements:

Comprehension: The model must understand the question, which is provided in Korean:

$$P \propto \text{Comp.}(L_{input} = \text{Korean}).$$

Reasoning: It must also perform the reasoning steps in Korean:

$$P \propto \text{Reasoning}(L_{reason} = \text{Korean}).$$

To examine which factor is more critical, we evaluate three cross-lingual setups: (1) Korean-to-Korean (K2K), (2) Korean-to-English (K2E), and (3) English-to-English (E2E). We exclude the English-to-Korean (E2K) scenario because models

typically fail to maintain a Korean chain-of-thought when the input is given in English. Further details, including the prompts used, can be found in Appendices B and G.

4.2 Evaluation Results

Prompting Type		K2K	K2E	E2E
Language	L_{input}	Ko	Ko	En
	L_{reason}	Ko	En	En
Qwen2.5	1.5B	16	21 (+5)	37 (+21)
	7B	40	41 (+1)	51 (+11)
	72B	58	60 (+2)	63 (+5)
Llama3.1/2	1B	7	7 (0)	22 (+15)
	8B	28	26 (-2)	39 (+11)
	70B	45	45 (0)	55 (+12)
Average Delta			+1	+11

Table 2: Performance of Qwen2.5 and Llama3.1/2 models on the HRM8K benchmark depending on the input and reasoning language. Number in bracket denote its gain compared to the K2K prompt.

Table 2 summarizes the outcomes for each configuration. Overall, performance tends to increase with model size, and larger models show smaller gaps across languages. We highlight two findings:

Effect of Input Language Switching from Korean input (K2E) to an entirely English setup (E2E) yields an average improvement of 11%. In particular, Qwen2.5-7B and Llama-3.1-8B drop by 10% and 13%, respectively, when forced to process Korean input. This underscores the significance of L_{input} in model performance.

Effect of Reasoning Language In contrast, comparing K2K to K2E shows an average difference of only 1%, suggesting that the language of the *reasoning process* has a relatively small impact once the model has already ingested Korean input. Simply allowing the model to produce its chain-of-thought in English does not fully recover performance lost from reading a Korean prompt.

In short, enabling English-based reasoning alone is insufficient to close the multilingual gap. Instead, the limiting factor appears to be how well the model can *comprehend* Korean inputs.

4.3 Multi-Step Inference

A straightforward approach to alleviating these comprehension issues is to split the task into multiple inference steps, where the model explicitly translates the Korean question into English first.

Figure 1: Example of UST process. When presented with a problem in Korean, the model generates Korean answers through the following processes highlighted in yellow: *Understanding the Question*, *Solving the Question*, and *Translating the Solution into Korean*.
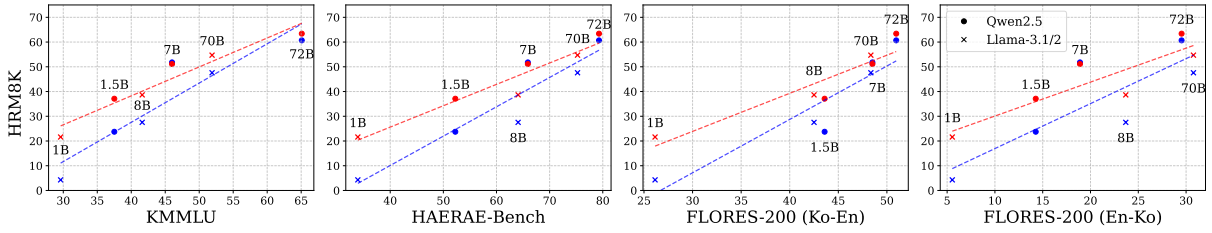

Figure 2: Comparison of HRM8K performance (vertical axis) and three additional benchmarks (KMMLU, HAERAE-Bench, FLORES-200) for Qwen2.5 and Llama-3.1/2 across different model sizes. TE2E (blue) translates Korean input to English before solving; E2E (red) uses an English prompt from the start.

We denote this as *Translated-English* (TE) in the first step, followed by an English-to-English (E2E) reasoning step in the second inference. This overall pipeline, called TE2E, uses a five-shot prompt for the translation stage.

In Figure 2, we compare TE2E (blue) and E2E (red) on three additional benchmarks: KMMLU (Son et al., 2024c), HAERAE-Bench (Son et al., 2023b), and FLORES (NLLB Team et al., 2024). For larger models with stronger Korean proficiency, TE2E and E2E become more similar, implying that multi-step inference can indeed help if the model already translates Korean accurately. However, for smaller models with weaker Korean skills, the performance gap remains pronounced. (Further details on the TE prompt are provided in Appendix G.)

5 Understand, Solve and Translate

Based on the experiments in Section 4, we conjecture that for smaller models improving their ability to understand Korean questions and allowing it to reason in English can address or bypass the constraints limiting LLMs in solving Korean questions. To validate this hypothesis, we fine-tune LLMs on a custom dataset designed to guide models through three stages: understanding Korean questions, solving them in English, and translating solutions back to Korean.

In this section, we explain the details of our training dataset (Section 5.1), report performance gains (Section 5.3), and conduct ablations on the effectiveness of the training (Section 5.4).

Models	GSM8K	MATH	OMNI_MATH	MMMLU	KSM	Avg.
Proprietary or Large Models						
GPT-4o	91.21	74.45	30.75	68.72	22.83	57.59
GPT-4o-Mini	87.57	70.68	26.45	63.40	19.40	53.50
Qwen2.5-72B-Instruct	90.07	72.06	30.96	66.60	23.46	56.63
Llama-3.1-70B-Instruct	79.08	56.05	19.85	60.00	13.10	45.61
Qwen2.5-7B-Instruct						
K2K-Prompting	66.41	50.36	18.96	50.00	11.83	39.52
K2E-Prompting	65.20	54.59	20.22	49.79	16.67	41.29
E2E-Prompting	81.35	68.87	27.29	57.02	21.08	51.12
Ours	80.06	68.53	27.19	57.23	19.12	50.43

Table 3: Evaluation results on HRM8K, comparing large-scale models (top) with different prompting strategies for Qwen2.5-7B-Instruct (bottom). Our UST-trained model achieves comparable performance to E2E-Prompting.

5.1 Training Dataset Construction

We create the UST (Understand, Solve, and Translate) dataset designed for training multilingual mathematical reasoning. Using GPT-4o-Mini, we generate cross-lingual Chain-of-Thought (CoT) examples that consist of three stages: (1) English Understanding Stage: Breaking down Korean questions and explaining their context and objectives in English; (2) English Solution Stage: Solving the mathematical problem in English; (3) Korean Solution Stage: Translating the English solution back to Korean. The dataset construction process follows these specific steps:

Step 1: Seed data collection   We source our initial data from two datasets: OpenMathInstruct-2 (Toshniwal et al., 2024) and NuminaMath-CoT (LI et al., 2024). From OpenMathInstruct-2's 14 million instruction samples, we randomly select 5 million instances and translate them to Korean using GPT-4o-Mini. For quality control, we use Qwen2.5-Math-RM-72B (Yang et al., 2024b), a reward model specialized for evaluating the quality of math question and output pairs, to score these instances and retain the top 50,000.

However, since OpenMathInstruct-2 primarily contains MATH and GSM8K augmentations, it lacks sufficient olympiad-level problems. To address this limitation, we supplement our dataset with samples from NuminaMath-CoT, specifically selecting problems from Aops Forum, AMC, Synthetic AMC, AIME, and Olympiads, and apply the same processing steps.

Step 2: Generating the Understanding Stage
Our previous experiments show that models per-

form better on Korean questions when translation and reasoning are separated into distinct steps. This suggests that traditional training methods, which emphasize immediate reasoning, may limit models' ability to process non-English inputs effectively. Such observation aligns with Zhu et al. (2024)'s finding that translation-specific training enhances multilingual reasoning capabilities. Based on these insights and recent advances in longer CoT generation, we introduce an **Understanding Stage** to our training pipeline. In this stage, we use GPT-4o-Mini to create structured breakdowns of Korean questions in English, providing both the original question and its solution to ensure alignment between understanding and problem-solving.

Step 3: Generating the Korean Solution Stage
Our seed datasets originally include an English solution for each sample. In this step, we translate each solution into Korean. At every generation stage, we include prompts to discard samples with incorrect translations or solutions. The final version of the dataset contains approximately 130k samples. Further details of the prompt used are available in Appendix G.

	HRM8K	ELO	Token Consum.
K2K	39.52	807	2,202
UST	50.43	1145	7,854
M.S.I	51.78	978	11,764

Table 4: Comparison of K2K prompting, UST model (ours), and Multi-Step Inference (M.S.I). For more details on M.S.I see Appendix D.

Model Configuration	Stage Language (Understand / Solve)	Accuracy (%)
Baseline (K2K)	Korean / Korean	15.95
Cross-lingual (K2E)	Korean / English	21.15
English-only (E2E)	English / English	37.10
Ablation Studies		
No Understanding	- / Korean	36.49
Korean Understanding	Korean / Korean	34.82
No Understanding	- / English	43.30
Full UST	English / English	**44.43**

Table 5: Ablation studies on different configurations of UST. The top section shows baseline prompting results, while the bottom section examines the impact of different language settings for each stage.

5.2 Fine-Tuning with UST

We fine-tune Qwen2.5-7B using a standard autoregressive objective to generate all three stages (understanding, solving, and translating) in a single inference. Special tokens are inserted between stages to enable selective generation during inference. The model architecture remains unchanged, without parameter freezing or additional parameters. The training process runs for 3 epochs (approximately 11 hours) on four H100 80GB HBM3 GPUs using DeepSpeed ZeRO-1 parallelism (Rajbhandari et al., 2020). Detailed training configurations and hyperparameters are provided in Appendix C.

5.3 Performance Analysis

Effects of Targeted Training Table 3 shows the performance of our Qwen2.5-7B model trained on the UST dataset. Our model achieves higher accuracy (50.43%) than both baseline approaches: K2K (39.52%) and K2E (41.29%) prompting, highlighting the effectiveness of targeted training for English reasoning. Furthermore, this performance is comparable to E2E prompting (51.12%), suggesting that our model successfully recovers the capabilities observed under ideal conditions where both questions and reasoning are in English.

Effects of Single-Pass Translation Instances in the UST dataset integrate translation with understanding and solving in a single inference. To examine this design choice, we first verify that the translation stage does not compromise performance. Out of the 8,011 questions in HRM8K, we observe only 15 cases (0.18%) where translation fails, all due to context length limitations. Notably, the translation stage serves purely as a user-friendly feature without affecting the model's

problem-solving capabilities.

For generating Korean solutions, we compare three approaches: (1) direct generation in Korean (K2K), (2) our single-pass UST, and (3) multi-step inference (MSI). MSI is a direct re-implementation of UST through prompting that separately performs three steps: translating the Korean question to English, solving in English, and re-translating the solution back to Korean. In Table 4, we evaluate these methods across three metrics: accuracy on HRM8K, response quality via ELO ratings[2], and computational efficiency through token consumption. While our model achieves accuracy comparable to MSI, it demonstrates superior response quality - preferred in 87.32% of direct comparisons (excluding ties). Furthermore, our approach consumes only 66% of the tokens required by MSI, making it computationally more efficient. Detailed evaluation methodology and prompts are provided in Appendices E, G.

5.4 Ablation Analysis

The UST dataset consists of three stages: Understand, Solve, and Translate. Having confirmed that the translation stage does not impact performance, we now examine the roles of the Understanding and Solving stages. We conduct experiments with four different configurations by varying both the presence and language of these stages. In our experiments, '-' indicates the omission of a stage, while 'Korean' or 'English' specifies the language used. For computational efficiency, we use a smaller model (Qwen2.5-1.5B-Instruct) and randomly sample 50k instances from the original dataset.

_____

[2]ELO rating is a widely adopted metric for comparing relative quality between language models through pairwise comparisons.

Models	HUMSS	STEM	Applied Science	Other	Avg.
Qwen2.5-7B	37.3	45.0	42.4	36.5	40.3
(Ours)	**39.3**	**49.5**	**45.3**	**40.3**	**43.6**

Table 6: Evaluation results on KMMLU (Son et al., 2024c).

Models	BN	DE	EN	ES	FR	JA	RU	SW	TE	TH	ZH	Av.
Qwen2.5-7B	60.8	79.6	**90.8**	80.0	77.2	70.0	**84.0**	**18.4**	31.2	**76.4**	**82.8**	68.3
(Ours)	**65.6**	**80.4**	90.0	**83.6**	**80.0**	**75.2**	83.2	14.8	**48.8**	71.6	81.6	**70.4**

Table 7: Evaluation results on MGSM (Shi et al., 2022).

Our experiments reveal that the original configuration - both stages in English - achieves the highest performance (44.43%) among all variants. A notable finding is that adding a Korean understanding stage actually decreases performance (36.49% → 34.82%). We attribute this counterintuitive result to two factors. First, when solving in Korean, an explicit understanding stage may be redundant as it essentially serves as another form of translation. Second, and more importantly, this suggests that chain-of-thought reasoning is most effective when conducted in the model's preferred language (English). This aligns with our observation that models show weaker reasoning capabilities in non-English languages, likely due to limited exposure during pre-training.

6 Tracing the Performance Gains

In this work, we demonstrate that routing through English understanding and reasoning steps enhances model performance on HRM8K. To understand the source of these improvements, we evaluate our UST-trained model on two additional benchmarks: KMMLU (Son et al., 2024c) and MGSM (Shi et al., 2022).

Our model shows consistent improvements across all KMMLU categories, with the largest gains in STEM (+4.5) and the smallest in HUMSS (+2.0)[3]. However, when tested on MGSM, the model shows performance drops in Swahili (-3.6) and Thai (-4.8). These contrasting results suggest that our gains stem not from general improvements in mathematical reasoning, but rather from enhanced Korean-specific capabilities and better Korean-to-English reasoning alignment.

Our findings align with our initial goal: addressing the performance gap between English and Korean reasoning on identical questions (Section 5.3). The effectiveness of our approach is demonstrated in two ways. First, it recovers most of the performance achieved with E2E prompting (Table 9). Second, it shows successful transfer to new domains (Table 6), suggesting that reasoning capabilities learned from machine-verifiable mathematics can generalize effectively. Most importantly, our method provides a simple path for non-English language users to benefit from the advanced reasoning capabilities typically available only in English.

7 Conclusion

In this paper, we propose UST, a training method that leverages English as an anchor language to enhance reasoning capabilities in Korean, and introduce HRM8K, a benchmark of 8,011 English-Korean parallel mathematics problems. Our analysis reveals that the performance gap in multilingual reasoning primarily stems from difficulties in processing non-English inputs. Through extensive experiments, we demonstrate that UST effectively bridges this gap and shows promising generalization to various Korean domains beyond mathematics. Our approach offers a simple yet effective solution for non-English language users to benefit from advanced reasoning capabilities typically available only in English, suggesting a practical direction for improving multilingual reasoning capabilities in language models.

---

[3]This aligns with the nature of CoT, which primarily enhances reasoning capabilities (Sprague et al., 2024) rather than factual knowledge required for HUMSS questions.

References

Alibaba. 2024. Qwq: Reflect deeply on the boundaries of the unknown.

Aida Amini, Saadia Gabriel, Peter Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. *arXiv preprint arXiv:1905.13319*.

Anonymous. 2024. Enhancing multilingual reasoning in LLMs: Insights from cross-linguistic correlations and optimal data proportions. In *Submitted to The Thirteenth International Conference on Learning Representations*. Under review.

Anthropic. 2024. Claude 3.5 sonnet.

Viraat Aryabumi, John Dang, Dwarak Taluporu, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, et al. 2024. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*.

Meriem Boubdir, Edward Kim, Beyza Ermis, Sara Hooker, and Marzieh Fadaee. 2023. Elo uncovered: Robustness and best practices in language model evaluation. *arXiv preprint arXiv:2311.17295*.

Nuo Chen, Zinan Zheng, Ning Wu, Linjun Shou, Ming Gong, Yangqiu Song, Dongmei Zhang, and Jia Li. 2023. Breaking language barriers in multilingual mathematical reasoning: Insights and observations. *arXiv preprint arXiv:2310.20246*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

CSAT. Korea institute for curriculum and evaluation. https://www.suneung.re.kr/boardCnts/list.do?boardID=1500234&m=0403&s=suneung.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2023. Investigating data contamination in modern benchmarks for large language models. *arXiv preprint arXiv:2311.09783*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Meng Fang, Xiangpeng Wan, Fei Lu, Fei Xing, and Kai Zou. 2024. Mathodyssey: Benchmarking mathematical problem-solving skills in large language models using odyssey math data. *arXiv preprint arXiv:2406.18321*.

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, et al. 2024. Omni-math: A universal olympiad level mathematic benchmark for large language models. *arXiv preprint arXiv:2410.07985*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. *arXiv preprint arXiv:2305.07004*.

Zhen Huang, Zengzhi Wang, Shijie Xia, Xuefeng Li, Haoyang Zou, Ruijie Xu, Run-Ze Fan, Lyumanshan Ye, Ethan Chern, Yixin Ye, et al. 2024. Olympicarena: Benchmarking multi-discipline cognitive reasoning for superintelligent ai. *arXiv preprint arXiv:2406.12753*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024. Click: A benchmark dataset of cultural and linguistic intelligence in korean. *arXiv preprint arXiv:2403.06412*.

KJMO. Korean junior mathematical olympiad. https://www.kms.or.kr/kjmo/.

KMO. Korean mathematical olympiad. https://www.kmo.or.kr/kmo/sub07.html.

KMS. Korean university students mathematics competition. https://www.kms.or.kr/conference/sub10.html.

Wen Lai, Mohsen Mesgar, and Alexander Fraser. 2024. Llms beyond english: Scaling the multilingual capability of llms with cross-lingual feedback. *arXiv preprint arXiv:2406.01771*.

Jia LI, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. 2024. Numinamath. [https://huggingface.co/AI-MO/NuminaMath-CoT](https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf).

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

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. *arXiv preprint arXiv:1705.04146*.

LMSYS. 2023. Chatbot arena: Benchmarking llms in the wild with elo ratings.

Meta AI. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models.

Mayank Mishra, Matt Stallone, Gaoyuan Zhang, Yikang Shen, Aditya Prasad, Adriana Meza Soria, Michele Merler, Parameswaran Selvam, Saptha Surendran, Shivdeep Singh, et al. 2024. Granite code models: A family of open foundation models for code intelligence. *arXiv preprint arXiv:2405.04324*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

OpenAI. 2024. Learning to reason with llms.

OpenAI. 2024. Multilingual massive multitask language understanding (mmlu). https://huggingface.co/datasets/openai/MMMLU.

Chanjun Park, Hyeonwoo Kim, Dahyun Kim, Seonghwan Cho, Sanghoon Kim, Sukyung Lee, Yungi Kim, and Hwalsuk Lee. 2024a. Open ko-llm leaderboard: Evaluating large language models in korean with ko-h5 benchmark. *arXiv preprint arXiv:2405.20574*.

Junsoo Park, Seungyeon Jwa, Meiying Ren, Daeyoung Kim, and Sanghyuk Choi. 2024b. Offsetbias: Leveraging debiased data for tuning evaluators. *arXiv preprint arXiv:2407.06551*.

Sungjoon Park. 2021. Klue: Korean language understanding evaluation. *arXiv preprint arXiv:2105.09680*.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are nlp models really able to solve simple math word problems? *arXiv preprint arXiv:2103.07191*.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Martin Jaggi, Leandro von Werra, and Thomas Wolf. 2024. Fineweb2: A sparkling update with 1000s of languages.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

Manley Roberts, Himanshu Thakur, Christine Herlihy, Colin White, and Samuel Dooley. 2023. To the cutoff... and beyond? a longitudinal perspective on llm data contamination. In *The Twelfth International Conference on Learning Representations*.

David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. 2019. Analysing mathematical reasoning abilities of neural models. *arXiv preprint arXiv:1904.01557*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Lin Shi, Chiyu Ma, Wenhua Liang, Weicheng Ma, and Soroush Vosoughi. 2024. Judging the judges: A systematic investigation of position bias in pairwise comparative assessments by llms. *arXiv preprint arXiv:2406.07791*.

Guijin Son, Hyunjun Jeon, Chami Hwang, and Hanearl Jung. 2024a. Krx bench: Automating financial benchmark creation via large language models. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and the 4th Workshop on Economics and Natural Language Processing@ LREC-COLING 2024*, pages 10–20.

Guijin Son, Hyunwoo Ko, Hoyoung Lee, Yewon Kim, and Seunghyeok Hong. 2024b. Llm-as-a-judge & reward model: What they can and cannot do. *arXiv preprint arXiv:2409.11239*.

Guijin Son, Hanwool Lee, Nahyeon Kang, and Moonjeong Hahm. 2023a. Removing non-stationary knowledge from pre-trained language models for entity-level sentiment classification in finance. *arXiv preprint arXiv:2301.03136*.

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

Guijin Son, Hanwool Lee, Suwan Kim, Huiseo Kim, Jaecheol Lee, Je Won Yeom, Jihyu Jung, Jung Woo Kim, and Songseong Kim. 2023b. Hae-rae bench: Evaluation of korean knowledge in language models. *arXiv preprint arXiv:2309.02706*.

Guijin Son, Dongkeun Yoon, Juyoung Suk, Javier Aula-Blasco, Mano Aslan, Vu Trong Kim, Shayekh Bin Islam, Jaume Prats-Cristià, Lucía Tormo-Bañuelos, and Seungone Kim. 2024d. Mm-eval: A multilingual meta-evaluation benchmark for llm-as-a-judge and reward models. *arXiv preprint arXiv:2410.17578*.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2024. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. *arXiv preprint arXiv:2409.12183*.

Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisacanin, Alexan Ayrapetyan, and Igor Gitman. 2024. Openmathinstruct-2: Accelerating ai for math with massive open-source instruction data. *arXiv preprint arXiv:2410.01560*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

TQ. Korean teacher qualification examination. https://blog.naver.com/headracer.

Maurice Weber, Daniel Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, et al. 2024. Redpajama: an open dataset for training large language models. *arXiv preprint arXiv:2411.12372*.

Yanan Wu, Jie Liu, Xingyuan Bu, Jiaheng Liu, Zhanhui Zhou, Yuanxing Zhang, Chenchen Zhang, Zhiqi Bai, Haibin Chen, Tiezheng Ge, et al. 2024. Conceptmath: A bilingual concept-wise benchmark for measuring mathematical reasoning of large language models. *arXiv preprint arXiv:2402.14660*.

Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. 2024. Benchmarking benchmark leakage in large language models. *arXiv preprint arXiv:2404.18824*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024b. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Dylan Slack, Qin Lyu, et al. 2024. A careful examination of large language model performance on grade school arithmetic. *arXiv preprint arXiv:2405.00332*.

Zhihan Zhang, Dong-Ho Lee, Yuwei Fang, Wenhao Yu, Mengzhao Jia, Meng Jiang, and Francesco Barbieri. 2023. Plug: Leveraging pivot language in cross-lingual instruction tuning. *arXiv preprint arXiv:2311.08711*.

Qihao Zhao, Yangyu Huang, Tengchao Lv, Lei Cui, Qinzheng Sun, Shaoguang Mao, Xin Zhang, Ying Xin, Qiufeng Yin, Scarlett Li, et al. 2024a. Mmlu-cf: A contamination-free multi-task language understanding benchmark. *arXiv preprint arXiv:2412.15194*.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024b. How do large language models handle multilingualism? *arXiv preprint arXiv:2402.18815*.

Yu Zhao, Huifeng Yin, Bo Zeng, Hao Wang, Tianqi Shi, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2024c. Marco-o1: Towards open reasoning models for open-ended solutions. *arXiv preprint arXiv:2411.14405*.

Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. 2021. Minif2f: a cross-system benchmark for formal olympiad-level mathematics. *arXiv preprint arXiv:2109.00110*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. 2024. Beyond english-centric llms: What language do multilingual language models think in? *arXiv preprint arXiv:2408.10811*.

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. 2024. Question translation training for better multilingual reasoning. *arXiv preprint arXiv:2401.07817*.

A Dataset Details

This section provides additional details regarding the construction and composition of the HRM8K benchmark.

A.1 Dataset Sources

KSM This subset consists of problems from Korean mathematics examinations and competitions:

- **College Scholastic Ability Test (CSAT)**[4]: The Korean counterpart to the SAT, which serves as the primary university entrance examination. We include only math problems with historical error rates exceeding 70%.

- **Korean Mathematical Olympiad (KMO)**[5]: The Korean equivalent to the International Mathematical Olympiad (IMO), primarily designed for middle- and high-school students. These problems require advanced mathematical knowledge and critical thinking.

- **Korean Junior Mathematical Olympiad (KJMO)**[6]: An elementary-school version of KMO aimed at identifying mathematical talent at an early stage.

- **Korean University Mathematical Olympiad (KMS)**[7]: A university-level competition featuring advanced topics in calculus, linear algebra, number theory, geometry, and discrete mathematics.

- **Korean National Teacher Qualification Test (TQ)**[8]: A standardized examination for teacher certification that focuses on mathematical pedagogy and content expertise.

Prior Sets This subset incorporates problems drawn from established English mathematics benchmarks, filtered to include only those with numeric answers:

- **GSM8K** (Cobbe et al., 2021): A collection of 8.5K grade-school math word problems that emphasize multi-step reasoning and elementary arithmetic.

- **MATH** (Hendrycks et al., 2020): A comprehensive benchmark of 12.5K high-school competition-level problems spanning seven mathematical domains, each accompanied by detailed step-by-step solutions.

- **Omni-MATH** (Gao et al., 2024): An advanced dataset containing 4.4K Olympiad-level problems across 33 sub-domains and 10 difficulty tiers, designed to push the limits of current LLM capabilities.

- **MMMLU** (OpenAI, 2024): A multilingual extension of the MMLU benchmark (Hendrycks et al., 2020), covering various STEM fields (e.g., abstract algebra, college mathematics, and high-school mathematics). It is available in 14 languages produced by professional translators.

A.2 Post-processing

For the **KSM** subset, we performed a manual verification and editing procedure to ensure high-quality OCR results. Specifically, we developed a review application in Streamlit[9], illustrated in Figure 3, which compares the original problem text against the OCR output. Two main factors were verified:

- *Content Completeness*: Confirming that all parts of the problem statement are accurately captured and that no text is omitted.

- *LATEX Integrity*: Ensuring that mathematical symbols and equations are correctly transcribed in LATEX format.

Based on these checks, we corrected errors and added any missing content. For instance, monetary symbols ($) enclosing LATEX symbols were removed to enhance clarity. Erroneous OCR outputs were manually fixed, and missing text was supplemented as needed. Figure 3 illustrates our interactive review interface.

B Additional Details in Experiments

We experiment with six multilingual language models reported to have been pretrained on Korean data: three Qwen2.5 Instruct models (1.5B, 7B, and 72B parameters) (Yang et al., 2024a) and three Llama-3.1/2 Instruct models (1B, 8B, and 70B parameters) (Dubey et al., 2024; Meta AI, 2024). For simplicity, we omit the word "Instruct" in references to these models, although all are instruction-tuned.

Unless otherwise noted, we set the sampling temperature to 0.7 and top_p to 0.95, with a minimum of 8 tokens and maximum of 2,048 tokens for the output. While lower temperatures are often used in

---

[4] https://www.suneung.re.kr/boardCnts/list.do?boardID=1500234&m=0403&s=suneung&searchStr=
[5] https://www.kmo.or.kr/kmo/sub07.html
[6] https://www.kms.or.kr/board/list.html?code=junior2
[7] https://www.kms.or.kr/board/list.html?code=conf12
[8] https://blog.naver.com/headracer

[9] https://streamlit.io/

KSM Question image & MetaData Review

RULES

1. 문제 이미지에 따라서 알맞게 문제가 OCR 되었는지 판단해주세요.
2. LaTeX 수식의 형태는 '$$'가 포함되지 않도록 교정해주세요.

Question Image

1. 함수 $f(x) = e^x\left(ax^3 + bx^2\right)$과 양의 실수 t에 대하여 닫힌

구간 $\left[-t, \ t\right]$에서 함수 $f(x)$의 최댓값을 $M\left(t\right)$, 최솟값을 $m(t)$

라 할 때, 두 함수 $M\left(t\right), m(t)$는 다음 조건을 만족시킨다.

(가) 모든 양의 실수 t에 대하여 $M\left(t\right) = f(t)$이다.

(나) 양수 k에 대하여 닫힌 구간 $\left[k, \ k+2\right]$에 있는 임의의 실수 t

에 대해서만 $m(t) = f(-t)$가 성립한다.

Category
수능/모의고사

Source
nan

Difficulty
미적분

Question

함수 \(f(x) = e^x (ax^{3} + bx^{2}) \)과 양의 실수 \(t \)에 대하여 닫힌 구간 \([-t, \ t] \)에서 함수 \(f(x) \)의 최댓값을 \(M(t) \), 최솟값을 \(m(t) \)라 할 때, 두 함수 \(M(t), m(t) \)는 다음 조건을 만족시킨다.

\[
\begin{aligned}
&\(\text{(가)} \) \text{모든 양의 실수} \ t \text{에 대하여} \ M(t) = f(t) \ \text{이다.} \\
&\(\text{(나)} \) \text{양수} \ k \ \text{에 대하여 닫힌 구간} \ \left[k, \ k + 2 \right] \ \text{에 있는 임의의 실수} \ t \\
& \quad \text{에 대하여서만} \ m(t) = f(-t) \ \text{가 성립한다.} \\
&\(\text{(다)} \) \int_{1}^{5} \left\(e^{t} \times m(t) \right\) \ dt = \frac{7}{3} - 8e
\end{aligned}
\]

\(f(k+1) = \frac{q}{p} e^{k+1} \)일 때, \(p + q \)의 값을 구하시오. (단, \(a \)와 \(b \)는 0이 아닌 상수, \(p \)와 \(q \)는 서로소인 자연수이고, \(\lim_{x \to \infty} \frac{x^{3}}{e^{x}} = 0 \)이다.)

Answer

Figure 3: Screenshot of our Streamlit-based OCR validation tool, used to compare source documents with OCR outputs and correct any errors.

pass@1 settings, we observed that extremely low temperatures sometimes cause models to revert to their preferred language (often English or Chinese). Hence, to maintain the specified response language, we employ a slightly higher temperature with moderate top_p.

C Fine-tuning Details

We fine-tune our models on H100 80GB GPUs using DeepSpeed ZeRO. Specifically, we train Qwen2.5-7B-Instruct on the UST approach and conduct ablation analyses with Qwen2.5-1.5B-Instruct under various settings. To maximize GPU utilization, we use a batch size of 96 per GPU across four GPUs for Qwen2.5-7B-Instruct, and a batch size of 128 per GPU across two GPUs for Qwen2.5-1.5B-Instruct. Table 8 summarizes the relevant hyperparameters.

D Multi-Step Inference

Multi-Step Inference (M.S.I) is a direct re-implementation of UST solely via prompting (i.e., without additional training). It is carried out in three steps:

1. **Translated-English (TE)**: Translate the original Korean prompt into English. We use the template shown in Figure 12.

2. **Translated-English-to-English (TE2E)**: Solve the translated problem in English using the template in Figure 9.

3. **Translated-English-to-English-to-Korean (TE2E2K)**: Translate the English solution back into Korean, following the template in Figure 13.

This multi-step approach echoes the three-stage UST pipeline (Understand, Solve, Translate) but relies on separate inferences with task-specific prompts.

E Evaluation Methods

E.1 ELO Rating

We use an Elo rating system to compare responses produced by different approaches in a pairwise manner (LMSYS, 2023). Elo ratings are computed in two parts: (1) *Expected Score*, which gauges each model's probability of winning based on current ratings; and (2) *Rating Update*, which adjusts the ratings after each match.

Expected Score. Given two models A and B with Elo ratings R_A and R_B, their expected scores E_A and E_B are:

$$
E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}},
$$
$$
E_B = \frac{1}{1 + 10^{(R_A - R_B)/400}}.
$$

Rating Update. After each pairwise comparison, the rating of model A is updated as follows:

$$
R'_A = R_A + K \times (S_A - E_A),
$$

Table 8: Hyperparameters for fine-tuning and ablation studies.

Base Model	Batch Size	Learning Rate	Scheduler	Optimizer	Max Length	# GPUs
Qwen2.5-7B	96	2e-5	Cosine	AdamW	8192	4
Qwen2.5-1.5B	128	2e-5	Cosine	AdamW	8192	2

where $S_A \in \{0, 1\}$ is the actual score (1 if A is preferred, and 0 otherwise). The constant K modulates the step size of the rating update; we set $K = 4$ for more stable ratings. We also randomly shuffle match order and apply bootstrapping over 1,000 iterations to mitigate dependence on match sequence (Boubdir et al., 2023).

E.2 Token Consumption

We measure token consumption via a simplified metric that accounts for both input and output tokens. For a dataset of N samples, let T_{input} and T_{output} be the number of input and output tokens, respectively, for each sample. The total token cost L_{model} for each model is:

$$L_{\text{model}} = \frac{\sum_{i=1}^{N} \left(T_{i,\text{input}} + 3 \times T_{i,\text{output}} \right)}{N},$$

where we weight output tokens by a factor of 3 to reflect their higher processing cost, following cost ratios from common LLM providers (e.g., OpenAI, Mistral AI, Alibaba Cloud, and Deepseek AI).
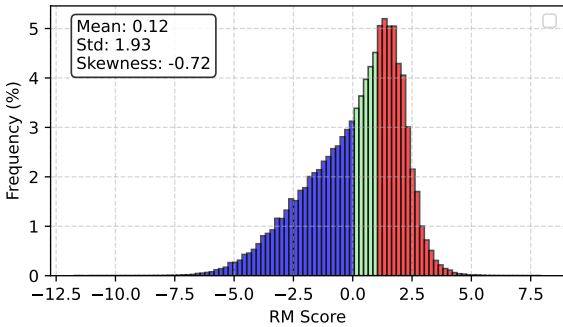
F Cross-Lingual Application of RMs

Figure 4: Reward model evaluation result on UST dataset. The samples were categorized into three groups based on the reward model score: high (RM Score > 1, red), low (RM Score < 0, blue), and medium ($0 \leq$ RM Score ≤ 1, green).

While creating the UST we leverage Qwen2.5-Math-RM-72B a reward model (RM) originally intended to be used in English or Chinese. We observe whether such RMs can be applied with further post-training for language transfer. In Figure 4, we

illustrate the score distribution on our initial dataset. The distribution shows to be right-skewed with a gradual tapering off towards the left. We create two datasets high and low. The high consists of samples with a score higher than 1 (colored in red) and low consists samples with a score lower than 0 (colored in blue). The high is used as our final dataset. For comparison we train a Qwen2.5-7B-Instruct model on the low dataset with identical number of instances.
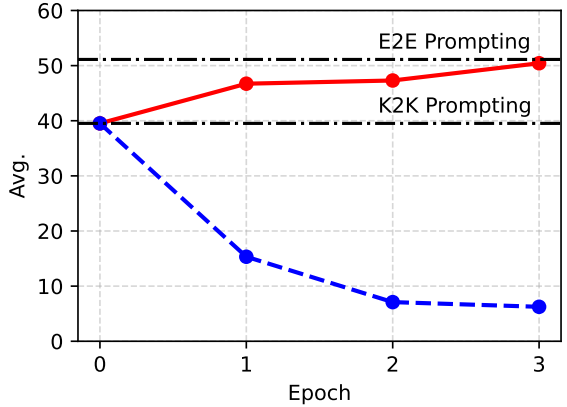

Figure 5: Qwen2.5-7B-Instruct model performance trends across epochs during training on high (red) and low (blue) datasets. The evaluation results of the original Qwen2.5-7B-Instruct model for K2K and E2E prompting were depicted with the dash-dotted lines.

As shown in Figure 5, training on the High subset progressively improves the model's performance. In contrast, using the Low subset degrades performance, with the model's score dropping to an average of 6.25 on HRM8K. This result suggests that RMs can be applied to new languages without additional training, aligning with previous studies (Son et al., 2024b,d). Accordingly, we choose to use the filtered subset.

G Prompt Templates

This section provides the complete text of all prompts used in our experiments, evaluations, and dataset construction. Each prompt is presented in a separate figure, preserving the original structure while enhancing clarity and consistency.

Table 9: Evaluation result for Qwen and Llama models on HRM8K. All models are instruction-tuned, but they are abbreviated for simplicity.

Model	GSM8K	MATH	Omni-MATH	MMMLU	KSM	Avg.
Korean-to-Korean (K2K)						
Qwen2.5-1.5B	28.13	20.69	8.64	18.51	3.78	15.95
Qwen2.5-7B	66.41	50.36	18.96	50.00	11.83	39.52
Qwen2.5-72B	89.46	74.73	30.07	69.79	25.35	57.88
Llama-3.2-1B	7.88	10.50	4.77	10.43	2.80	7.28
Llama-3.1-8B	57.47	31.20	11.73	32.98	5.25	27.73
Llama-3.1-70B	78.62	56.12	20.38	57.87	11.83	44.96
Korean-to-English (K2E)						
Qwen2.5-1.5B	31.92	26.86	10.32	30.85	5.81	21.15
Qwen2.5-7B	65.20	54.59	20.22	49.79	16.67	41.29
Qwen2.5-72B	89.23	77.68	32.74	70.43	27.73	59.56
Llama-3.2-1B	7.20	9.57	5.08	11.49	2.94	7.26
Llama-3.1-8B	55.04	31.13	11.16	27.66	5.18	26.03
Llama-3.1-70B	77.63	55.18	19.43	58.94	12.82	44.80
English-to-English (E2E)						
Qwen2.5-1.5B	65.50	49.25	16.87	45.53	8.33	37.10
Qwen2.5-7B	81.35	68.87	27.29	57.02	21.08	51.12
Qwen2.5-72B	94.31	83.33	37.72	70.00	31.65	63.40
Llama-3.2-1B	43.44	27.24	9.90	23.83	3.64	21.61
Llama-3.1-8B	79.45	48.11	16.08	42.34	7.21	38.64
Llama-3.1-70B	93.33	67.90	24.83	70.43	17.09	54.71

1. **OCR Prompt (Figure 6)**: Performs OCR on an image of a Korean math problem and extracts the text. Used to build the KSM subset (Section 3.1).

2. **Question Translation Prompts (Figures 7–8)**: Translate math questions into Korean or English, respectively. Used for creating bilingual pairs in HRM8K (Section 3.2).

3. **Solution Generation Prompts (Figures 9–10)**: Evaluate model performance on HRM8K under different reasoning-language conditions (Section 4.1).

4. **Understanding Generation Prompt (Figure 11)**: Produces a structured breakdown of the problem for the *Understanding* stage of UST (Section 5.1).

5. **Model Translation Prompts (Figures 12–13)**: Used in multi-step inference (M.S.I) to translate the question or solution between Korean and English (Section 4.3).

6. **LLM-as-a-Judge Prompt (Figure 14)**: Conducts pairwise response comparisons and produces a verdict, enabling ELO-based evaluation (Section 5.3).

OCR Prompt

You will be given an image containing a Korea math question. Your task is to conduct an OCR to retrieve the question in text format.

Follow the following roles:

1. return the question only, nothing else.
2. If, the image contains the answer ignore it. Do not return it with the question.
3. Put extra care on notations and equations make sure they are identical.

Figure 6: Prompt to perform OCR and extract the mathematical question from the given image. Both the prompt and the screenshot of the problem were provided to the model for OCR processing.

Table 10: Translated-English-to-English (TE2E) prompting evaluation result on HRM8K.

Model	GSM8K	MATH	Omni-MATH	MMMLU	KSM	Avg.
Translated-English-to-English (TE2E)						
Qwen2.5-1.5B	36.24	34.52	12.05	29.36	6.51	23.74
Qwen2.5-7B	79.53	70.78	28.86	59.36	20.38	51.78
Qwen2.5-72B	89.16	78.13	34.05	70.43	31.65	60.68
Llama-3.2-1B	2.96	5.41	1.94	9.36	1.75	4.28
Llama-3.1-8B	53.90	36.43	13.04	29.57	4.76	27.54
Llama-3.1-70B	78.17	60.83	22.26	60.43	16.25	47.59

Question Translation Prompt: En→Ko

You are a professional English-to-Korean translator specializing in academic content. Your task is to translate math problems provided in English into clear, natural, and precise Korean referring to given examples. Follow the instructions below:

INSTRUCTIONS:
1. You SHOULD NOT solve the problem and translate only the given question — do not include any additional commentary.
2. Preserve all mathematical symbols, notations, formatting, and existing choices exactly as presented.
3. Use fluent, natural Korean that aligns with academic standards for math problems.
4. Ensure the translation conveys the meaning and context accurately.

INPUT:
{question}

Figure 7: Translation prompt to translate English math questions into Korean. This prompt is utilized to translate the *Prior Sets* data, sourced from English math benchmarks, into Korean. The bracketed part is a placeholder to fill in the question.

HRM8K Solution Prompt: English Reasoning

Solve the given question.
After solving the problem, state your final answer in the following format: $\\boxed{N}$.

{question} Respond in English.

Figure 9: Solution generation prompt to evaluate the models in an English reasoning setup, such as K2E and E2E. The bracketed part is a placeholder to fill in the question.

Question Translation Prompt: Ko→En

You are a professional Korean-to-English translator specializing in academic content. Your task is to translate math problems provided in Korean into clear, natural, and precise English referring to given examples. Follow the instructions below:

INSTRUCTIONS:
1. You SHOULD NOT solve the problem and translate only the given question — do not include any additional commentary.
2. Preserve all mathematical symbols, notations, formatting, and existing choices exactly as presented.
3. Use fluent, natural English that aligns with academic standards for math problems.
4. Ensure the translation conveys the meaning and context accurately.

Figure 8: Translation prompt to translate Korean math questions into English. This prompt is utilized to translate the *KSM* data, sourced from Korean math examinations and competitions, into English. The bracketed part is a placeholder to fill in the question.

HRM8K Solution Prompt: English Reasoning

Solve the given question.
After solving the problem, state your final answer in the following format: $\\boxed{N}$.

{question} Respond in Korean.

Figure 10: Solution generation prompt to evaluate the models in a Korean reasoning setup, such as K2K. The bracketed part is a placeholder to fill in the question.

Understanding Generation Prompt

[User]
Solve the following problem:
{question}

[Assistant]
{solution}

[User]
I'm planning to generate a step-by-step guide for the solution. The step-by-step solution will be provided to students to guide their solution. Accordingly, it should be clear and straightforward, guiding the student through the problem-solving process. However, it must not reveal the answer as it will disturb the students' solution. Generate the breakdown. It should assist with understanding the question and planning how to solve it. The generation will be directly provided to the student, accordingly do not include notes like 'not reveal the answer', or a evaluation of your own breakdown. Write in first person view: e.g I will , I can .

Figure 11: Generation prompt to generate *Understanding* stage of UST. Given a problem and its corresponding answer, a prefix conversation history is constructed where the user asks the problem and the assistant provides the ground-truth answer. Subsequently, the user instructs the assistant to generate an understanding of the problem. The bracketed parts are a placeholder to fill in the question and its ground-truth solution.

Model Translation Prompt: Translated-English (TE)

You are a professional Korean-to-English translator specializing in academic content. Your task is to translate math problems provided in Korean into clear, natural, and precise English referring to given examples. Follow the instructions below:

INSTRUCTIONS:
1. You SHOULD NOT solve the problem and translate only the given question — do not include any additional commentary.
2. Preserve all mathematical symbols, notations, formatting, and existing choices exactly as presented.
3. Use fluent, natural English that aligns with academic standards for math problems.
4. Ensure the translation conveys the meaning and context accurately.

INPUT:
[1st Korean Question Example]

OUTPUT:
[1st English Translation Result]

. . .

INPUT:
[5th Korean Question Example]

OUTPUT:
[5th English Translation Result]

INPUT:
{question}

OUTPUT:

Figure 12: Translation prompt to generate Translated-English (TE) problem by translating the given Korean problem into English. The bracketed part is a placeholder to fill in the question.

Model Translation Prompt: Translated-English-to-English-to-Korean (TE2E2K)

You are a professional Korean-to-English translator specializing in academic content. Your task is to translate math problems provided in Korean into clear, natural, and precise English referring to given examples. Follow the instructions below:

INSTRUCTIONS:
1. You SHOULD NOT solve the problem and translate only the given question — do not include any additional commentary.
2. Preserve all mathematical symbols, notations, formatting, and existing choices exactly as presented.
3. Use fluent, natural English that aligns with academic standards for math problems.
4. Ensure the translation conveys the meaning and context accurately.

INPUT:
[1st English Solution Example]

OUTPUT:
[1st Korean Translation Result]

. . .

INPUT:
[5th English Solution Example]

OUTPUT:
[5th Korean Translation Result]

INPUT:
{question}

OUTPUT:

Figure 13: Translation prompt to generate Translated-English-to-English-to-Korean solution by translating the given English solution into Korean. The bracketed part is a placeholder to fill in the question.

LLM-as-a-judge Prompt

[System]
Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user question displayed below. You should choose the assistant that follows the user's instructions and answers the user's question better. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of their responses. Begin your evaluation by comparing the two responses and provide a short explanation. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible. After providing your explanation, output your final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]" if assistant B is better. There is no option for a tie, you should choose "[[A]]" or "[[B]]".

[User Question]
{question}

[The Start of Assistant A's Answer]
{model_a_answer}
[The End of Assistant A's Answer]

[The Start of Assistant B's Answer]
{model_b_answer}
[The End of Assistant B's Answer]

Figure 14: Evaluation prompt to conduct llm-as-a-judge pairwise evaluation between model A's response and model B's response to a given question. The bracketed parts are the placeholders to fill in the question, model A's answer, and model B's answer.