

Taylor & Francis
Taylor & Francis Group

Applied Artificial Intelligence
An International Journal

ISSN: 0883-9514 (Print) 1087-6545 (Online) Journal homepage: www.tandfonline.com/journals/uaai20

Evaluating Large Language Models on the 2026 Korean CSAT Mathematics Exam: Measuring Mathematical Ability in a Zero-Data-Leakage Setting

Goun Pyeon, Inbum Heo , Jeesu Jung, Taewook Hwang, Hyuk Namgoong, Hyein Seo, Yerim Han , Eunbin Kim, Hyeonseok Kang & Sangkeun Jung

To cite this article: Goun Pyeon, Inbum Heo , Jeesu Jung, Taewook Hwang, Hyuk Namgoong, Hyein Seo, Yerim Han , Eunbin Kim, Hyeonseok Kang & Sangkeun Jung (2026) Evaluating Large Language Models on the 2026 Korean CSAT Mathematics Exam: Measuring Mathematical Ability in a Zero-Data-Leakage Setting, Applied Artificial Intelligence, 40:1, 2684121, DOI: 10.1080/08839514.2026.2684121

To link to this article: https://doi.org/10.1080/08839514.2026.2684121

© 2026 The Author(s). Published with license by Taylor & Francis Group, LLC.

View supplementary material

Published online: 23 Jun 2026.

Submit your article to this journal

View related articles

View Crossmark data

This article has been awarded the Centre for Open Science 'Open Data' badge.

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=uaai20

APPLIED ARTIFICIAL INTELLIGENCE
2026, VOL. 40, NO. 1, e2684121 (26 pages)
https://doi.org/10.1080/08839514.2026.2684121

Taylor & Francis
by informa

OPEN ACCESS Check for updates

Evaluating Large Language Models on the 2026 Korean CSAT Mathematics Exam: Measuring Mathematical Ability in a Zero-Data-Leakage Setting

Goun Pyeon[a*], Inbum Heo[a*], Jeesu Jung[a*#], Taewook Hwang[b*], Hyuk Namgoong[a], Hyein Seo[c], Yerim Han[a], Eunbin Kim[a], Hyeonseok Kang[a], and Sangkeun Jung[a]

[a]Chungnam National University, Daejeon, South Korea; [b]AI Research Center SI Analytics Co. Ltd., Daejeon, Republic of Korea; [c]Korea Electric Power Research Institute, Naju, Republic of Korea

ABSTRACT

This study systematically evaluates the mathematical reasoning capabilities of Large Language Models (LLMs) using the 2026 Korean CSAT Mathematics section in a strictly contamination-free setting. To mitigate data leakage in existing benchmarks, we digitized all 46 questions (22 common and 24 elective) within two hours of the exam's public release, minimizing the chance that any item could be incorporated into model training data. We evaluated 24 state-of-the-art LLMs across input modalities (Text-only, Image-only, Text + Figure) and prompt languages (Korean, English). GPT-5 family models achieved perfect scores (100 points) under a limited set of language – modality configurations, while Grok 4, Qwen 3 235B, and Gemini 2.5 Pro also scored above 97 points. Notably, gpt-oss-20B achieved 95.7 points despite its relatively small size, demonstrating strong cost-effectiveness. Problem-level analysis identified Calculus as the weakest domain, with significant degradation on 4-point high-difficulty questions. Text input consistently outperformed image input, while the effect of prompt language varied by model. Reasoning enhancement experiments with GPT-5 show that increasing reasoning intensity improves performance (82.6→100) but roughly quadruples token usage and reduces efficiency, suggesting minimal reasoning can be more practical. Detailed results and model comparisons are available at https://isoft.cnu.ac.kr/csat2026/.

ARTICLE HISTORY
Received 9 February 2026
Revised 20 May 2026
Accepted 31 May 2026

Introduction

The mathematical reasoning abilities of Large Language Models (LLMs) have become a central indicator for assessing a model's logical thinking and problem-solving skills. Because mathematical problems demand clear answers and structured reasoning, they are widely used as benchmarks for evaluating a model's true reasoning capabilities (Hendrycks et al. 2021a). However, the rapid performance gains of recent models, traditional evaluation approaches have increasingly struggled to accurately distinguish an LLM's genuine mathematical competence.

In particular, given that LLMs are trained on vast web-scale corpora, it is difficult to verify whether a model has been exposed – directly or indirectly – to publicly available benchmark datasets or closely related materials (Dodge et al. 2021; Min et al. 2022). Even when a benchmark itself is not included in training data, paraphrased, synthetic, or structurally transformed versions of benchmark problems may still have been used in training or inference.

Such potential contamination to benchmark data can distort measured performance, making it difficult to determine whether a model solves problems using generalized mathematical reasoning or merely reproduces patterns memorized during training (Samuel, Zhou, and Zou 2025; Yao, Cheng, and Tan 2025). In other words, an LLM may perform well on a public benchmark simply because it has been implicitly optimized for those questions. Recent studies suggest that benchmark contamination may have occurred in the training data of major LLMs such as GPT-4 (OpenAI 2023) and Claude (Bai et al. 2022), raising the possibility that high scores may reflect *memorization or pattern replication* (McCoy, Pavlick, and Linzen 2019) rather than true reasoning ability

CONTACT Sangkeun Jung hugmanskj@gmail.com Department of Computer Science & Engineering, Chungnam National University, Daejeon 34134, South Korea

*These authors contributed equally.
#Jeesu Jung is currently affiliated with Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/08839514.2026.2684121

© 2026 The Author(s). Published with license by Taylor & Francis Group, LLC.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

(Balloccu et al. 2024; Deng et al. 2024). Thus, an evaluation dataset that is completely free from any risk of data leakage is essential for validating an LLM's genuine mathematical understanding and problem-solving ability.

In this context, the Mathematics section of the Korean College Scholastic Ability Test (CSAT) provides a uniquely leakage-free evaluation environment. The CSAT is a national-level exam created annually by the government, because test items remain undisclosed until the exam ends, rapid digitization and evaluation immediately after release ensures that no LLM could have incorporated similar problems into its training data. Moreover, with 46 questions covering core domains such as *Probability and Statistics*, *Calculus*, and *Geometry*, the CSAT Mathematics exam enables comprehensive measurement of an LLM's practical mathematical proficiency across the entire secondary curriculum.

This study establishes a new evaluation framework that systematically assesses the mathematical reasoning capabilities of state-of-the-art LLMs under a fully contamination-free setting, using the Mathematics section of the 2026 CSAT. By digitizing and preprocessing all questions and initiating model inference within two hours of their public release, we constructed a *Contamination-Free CSAT benchmark* that could not appear in the training data of any model. Beyond simple accuracy comparison, our framework varies input modality, prompting language, reasoning configuration, and solving time, enabling a multidimensional analysis of how LLMs interpret and solve real exam problems.

This work focuses on the following four research questions (RQs):

RQ1. What are the mathematical problem-solving abilities of LLMs on the 2026 CSAT?

RQ2. Which types of CSAT Mathematics problems are most challenging for LLMs?

RQ3. How does LLM mathematical performance vary across different input conditions?

RQ4. What impact does varying the strength of reasoning have on model performance?

Our results show that although several GPT-5 models achieved perfect scores under specific language – modality combinations, the GPT-5 family was the only group to reach such performance levels. In contrast, many other models exhibited notably poor performance on Calculus questions and on high-difficulty 4-point items, particularly short-answer problems.

Text-only input consistently outperformed image-based input, and increasing the reasoning strength of GPT-5 improved accuracy from 82.6 to 100 points, but with a 4–5× increase in token usage, sharply reducing efficiency. Notably, the small-scale gpt-oss-20B achieved 95.7 points at the lowest cost ($0.01), suggesting that model size is not a determining factor in mathematical performance.

The structure of this paper is as follows. 'Related Work' reviews prior work on math benchmarks and LLM evaluation. 'Korean CSAT-Math Evaluation Framework for LLMs' analyzes the CSAT problem structure and introduces our CSAT-Math evaluation framework, including the data construction pipeline and prompt design. 'CSAT-Math LLM Experiment Design' describes the evaluated models and experimental setup. 'LLM Experiment Results' presents experiments and results corresponding to the four research questions (RQ1–RQ4). Finally, 'Key Findings' summarizes the key findings, and 'Conclusion' concludes with limitations and future research directions.

A preliminary version of this work is available as a preprint (Pyeon et al. 2025).

Related work

Math benchmarks for LLMs

Math benchmarks for LLMs fall into two categories: (1) LLM-specific benchmarks and (2) real exam-based benchmarks.

LLM-specific benchmarks (GSM8K (Cobbe et al. 2021), CHAMP (Mao, Kim, and Zhou 2024), GSM-Plus (Li et al. 2024)) evaluate numerical reasoning, hint-dependency, and robustness. Multimodal benchmarks (MathVista (Lu et al. 2024), MATH-Vision (K. Wang et al. 2024)) incorporate visual information for integrated reasoning.

Real exam-based benchmarks use standardized tests (SAT, GRE, AIME, IMO). MATH (Hendrycks et al. 2021b), MathBench (Liu et al. 2024), AIME (Artificial Analysis 2025; Patel et al. 2024), and others assess conceptual understanding but remain predominantly English-language, Western-focused.

In contrast, our work systematically evaluates LLM math performance in a non-English educational setting using the Korean CSAT Mathematics exam. The CSAT is a nationwide high-stakes university entrance exam taken annually by over 500,000 students and is regarded as a national standard exam with high validity, reliability, and carefully balanced difficulty. By using all 46 questions—22 from the common curriculum and 24 from elective subjects – we conduct a comprehensive evaluation covering geometry, calculus, and probability and statistics.

Data leakage in LLM evaluation

Data leakage poses a critical threat to the reliability of LLM performance evaluation and has been a central topic of discussion in math and language benchmark research.

Recent studies (Balunovic et al. 2025; Hidayat et al. 2025; Xu et al. 2024) show contamination in major benchmarks (GSM8K, MATH, MMLU), with performance drops on private datasets, suggesting models may exploit memorized patterns rather than genuine reasoning.

To eliminate this risk, our study collects and processes the 2026 CSAT Mathematics exam immediately after public release and conducts model evaluation right away, ensuring that no LLM could have been exposed to the data beforehand. This procedure guarantees a fully contamination-free setting, blocking any possibility of prior contamination and enabling faithful measurement of mathematical reasoning using real, previously unseen exam problems.

Prior work on input modality and prompting language

LLM performance is influenced not only by model architecture but also by input modality and prompting language. Early LLMs were optimized for Text-only input, but recent advances have produced multimodal LLMs capable of jointly processing images, diagrams, and mathematical notation. This has driven interest in evaluating models on visually grounded mathematics tasks (Lu et al. 2024; K. Wang et al. 2024).

Prompting language is another key factor. Research on multilingual prompting shows that the chain-of-thought structure, reasoning trajectory, and solution strategy can vary depending on the language in which the problem is presented. Several studies report that English prompts often provide greater stability and accuracy (Park et al. 2025; Zhou et al. 2025).

Building on this literature, we analyze how input modality (text, image, text + visual information) and prompting language (Korean, English) affect model performance in a realistic exam setting. By crossing these factors, we evaluate and compare LLM mathematical reasoning from multiple perspectives.

Prior work on LLM reasoning and reasoning evaluation

Research on reasoning in LLMs focuses on how to elicit, regulate, and evaluate the reasoning process a model performs before generating a final answer. Chain-of-Thought (CoT) (X. Wang et al. 2023) proposed generating multiple reasoning paths in parallel and selecting the final answer using majority voting, improving stability and consistency. Zero-shot CoT (Kojima et al. 2022) demonstrated that simple prompts can induce multi-step reasoning without fine-tuning.

Recent studies examine how to scale test-time reasoning. Prior work has shown that increasing reasoning depth and inference-time compute – known as test-time scaling – can yield stronger performance (Snell et al. 2025). Zhou et al. (2025) present a visualization of diverse reasoning types and cognitive trajectories, enabling qualitative assessment of reasoning quality. These studies collectively highlight that the relationship between reasoning length and accuracy is not linear: excessively long reasoning may accumulate errors and degrade performance.

Our work quantitatively evaluates these insights in a real exam setting. By systematically varying the reasoning mode (e.g., reasoning_effort) on the same CSAT problems, we analyze the *trade-off* between reasoning strength, performance, and computational cost.

Korean CSAT-Math evaluation framework for LLMs

Overview of the CSAT

The Korean College Scholastic Ability Test (Korean CSAT) is a nationally administered, high-stakes standardized examination produced by the Korea Institute for Curriculum and Evaluation (KICE). It serves as the most authoritative assessment used for university admissions in Korea and evaluates whether students possess the academic abilities required for higher education.[1]

The CSAT is administered once per year and taken by roughly half a million students nationwide (554,174 examinees in 2026).

The CSAT is constructed under strict security protocols with isolated exam writers and reviewers, making it a near-impossible-to-leak examination.[2]

The CSAT consists of five subject areas: Korean, Mathematics, English, Korean History, and elective Social/Science/Vocational studies (Ministry of Education 2024). This work focuses exclusively on the Mathematics section.

Since 2022, the CSAT Mathematics section has consisted of 22 common-subject questions (Math I and Math II) and 8 elective questions selected from Probability and Statistics, Calculus, or Geometry, for a total of 30 questions to be solved in 100 minutes.

In this study, to evaluate overall mathematical competence and cross-subject reasoning, we include all 46 questions—22 common-subject problems and all 24 elective-subject problems from the three elective domains – regardless of the single-subject choice that human test-takers make.

Table 1 summarizes the structure, problem types, and scoring scheme.

- **Common Subject (22 problems)**: Math I (11 problems) and Math II (11 problems), taken by all test-takers.
- **Elective Subject (8 problems)**: Students select one among Probability and Statistics, Calculus, and Geometry.
- **Question Types**: Multiple-choice and short-answer.
  - **Short-answer**: Students provide an integer between 0 and 999.
  - **Multiple-choice**: Five answer options (1–5) are provided.
- **Scoring**: Each problem is worth 2, 3, or 4 points depending on required reasoning complexity.

  According to the official 2026 CSAT Mathematics curriculum guide,[3] each subject covers the following areas:
- **Common Subjects**
  - **Math I**: Exponential and logarithmic functions, trigonometric functions, sequences
  - **Math II**: Limits and continuity, differentiation, integration
- **Elective Subjects**
  - **Probability and Statistics**: Counting, probability, statistical inference
  - **Calculus**: Limits of sequences, differential calculus, integral calculus
  - **Geometry**: Conic sections, plane vectors, three-dimensional geometry and coordinates

Because the CSAT Mathematics exam has a systematically designed structure – including well-defined scope, problem types, and difficulty – it is particularly suited for evaluating LLM mathematical reasoning. The exam follows a consistent progression from common to elective subjects, as visualized in Figure 1.

Table 1. Structure and question types of the CSAT Mathematics Section.

		Common	Elective			
Subject		Math I + Math II	Probability	Calculus	Geometry	Total
Num. of Questions		11 + 11	8	8	8	46
Type	Multiple-choice	15	6	6	6	33
	Short-answer	7	2	2	2	13
Score	2-point problems	2	1	1	1	5
	3-point problems	6	4	4	4	18
	4-point problems	14	3	3	3	23
Total Score		74	26	26	26	152

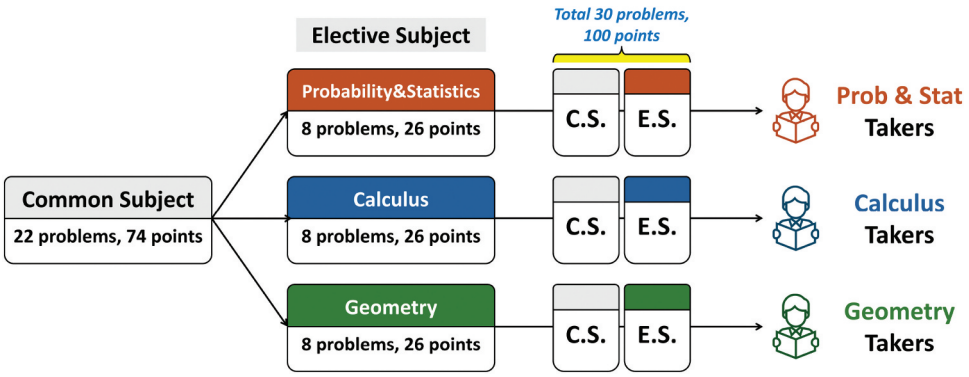


Figure 1. Overall structure of the CSAT Mathematics Section (common (C.S.) vs. Elective subjects (E.S.)).

Data construction pipeline

Our evaluation framework is designed to completely eliminate the possibility of prior contamination during model training. To achieve this, we constructed the evaluation dataset – including transcription, quality review, and aggregation – and began model evaluation within the first two hours after the official release of the exam. A total of seven researchers (six annotators and one reviewer) contributed to the manual dataset construction, and eight participated in the LLM evaluation and monitoring process. Figure 2 illustrates the full pipeline executed immediately after exam release.

The dataset construction process proceeds as follows:

(1) Annotators receive the Digitalized CSAT-Math template and their assigned question IDs.
(2) Immediately after the official exam release,[4] each annotator extracts all required information for their assigned questions.
(3) A reviewer aggregates all entries and checks for errors and consistency.
(4) A unified Digitalized CSAT-Math dataset is finalized.

Structure of the 2026 CSAT-Math question paper

The question booklet distributed to students contains a subject designation (common and elective subjects: Probability, Calculus, Geometry), a question-type header at the start of each section, the problem statement with its problem number and score (2, 3, or 4 points), visual materials (diagrams, graphs, or figures) accompanying the problem, occasional tables (e.g., z-score tables in Probability/Statistics), and answer choices (1–5) for multiple-choice items.

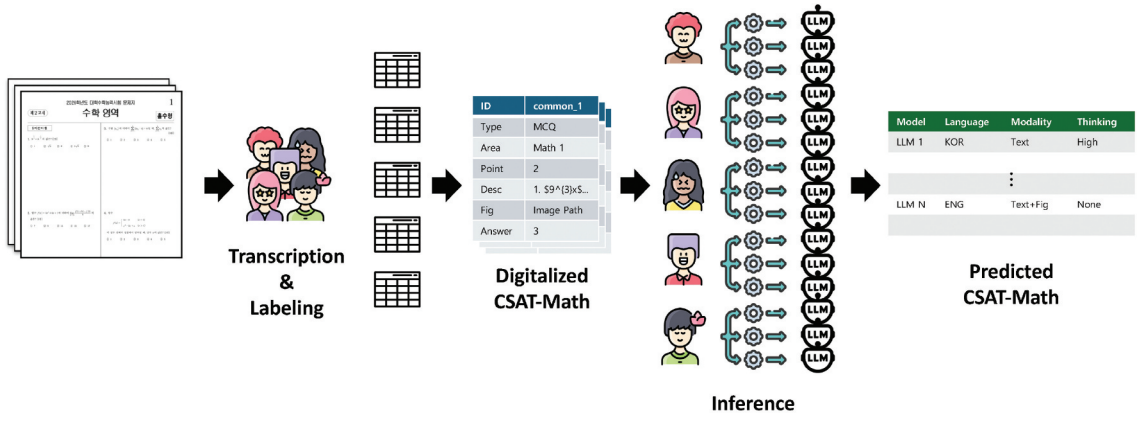


Figure 2. Manual dataset construction (transcription and labeling) and LLM evaluation pipeline.

Table 2. Schema of the Digitalized CSAT-Math dataset.

Field	Description
prob_id	Unique problem ID, e.g., 2026_odd_common_11 or 2026_odd_probability_30.
prob_type	Question type: multiple-choice or short-answer.
prob_area	Subject area: Math I, Math II, Probability, Calculus, or Geometry.
prob_point	Score value: 2, 3, or 4 points.
prob_desc	Problem text converted into Markdown + LaTeX.
prob_img_path	Cropped image of the full problem.
prob_fig_img_path	Cropped image of figures or diagrams only.
answer	Correct answer (1–5 for MCQ; integer 0–999 for short-answer).

Digitalized CSAT-Math data format

Annotators input structured information into the Digitalized CSAT-Math dataset according to the schema in Table 2. Of these, only variable fields (prob_area, prob_desc, image paths, answer) are filled in by annotators; structural fields remain fixed across years.

Digitalized CSAT-Math dataset construction

The goal of the dataset construction process is twofold: (1) to preserve the original structure and mathematical expressions of the exam problems as faithfully as possible, and (2) to standardize the data into a consistent format suitable for direct LLM input.

A total of seven annotators participated in building the Digitalized CSAT-Math dataset. Among the 46 questions, four annotators each handled 9 questions and one annotator handled 10 questions, producing the fields prob_desc, answer, and prob_area. Additionally, one annotator exclusively extracted all problem and figure images.

Text construction for each problem

To enable LLM-based evaluation using PDF CSAT problems, all questions were fully transcribed into LLM-friendly text. Each problem's stem, choices, score, tables, and mathematical expressions were encoded into the prob_desc field of the Digitalized CSAT-Math format. The prob_desc field represents a textual reconstruction of the "human-readable problem sheet."

Extraction of Structured content using markdown. CSAT problems contain structured elements such as boxed passages (Figure 3(f,i)), multiple-choice options (Figure 3(b,d)), and tables (Figure 3(k)) that lose readability in plain text. Markdown preserves these structures: general paragraphs as plain text (Figure 3(a,c,h)), boxed "보기" sections as block quotes using > (Figure 3(f,i)), enabling faithful reconstruction of the original layout.

Mathematical expressions Using LaTeX. Mathematical expressions were converted to LaTeX with inline formulas ($ ... $) (Figure 3(f,g,j)) and block expressions ($$ ... $$) (Figure 3(e)). Symbols and operators were faithfully reproduced using standard LaTeX commands to match the original formatting.

Representation of problem components (number, score, answer choices). Problem numbers, scoring labels (e.g., "[4점]"), and answer choices were transcribed exactly as printed. Multiple-choice options were preserved using original circled numerals (172–176), e.g., "172 2 173 3 174 4 175 5 176 6" (Figure 3(b,d)). This ensured that the Digitalized CSAT-Math text closely mirrored the actual problem layout.

Metadata construction for each problem

We define metadata as information that is required for evaluation and analysis but is not explicitly visible to LLMs in the original exam sheet. This includes the true *subject area* of each problem and its *correct answer*.

Actual subject area per problem. As described in Overview of the CSAT, CSAT Mathematics consists of Math I, Math II, and one of three elective subjects. However, within the 1–22 common-subject questions,

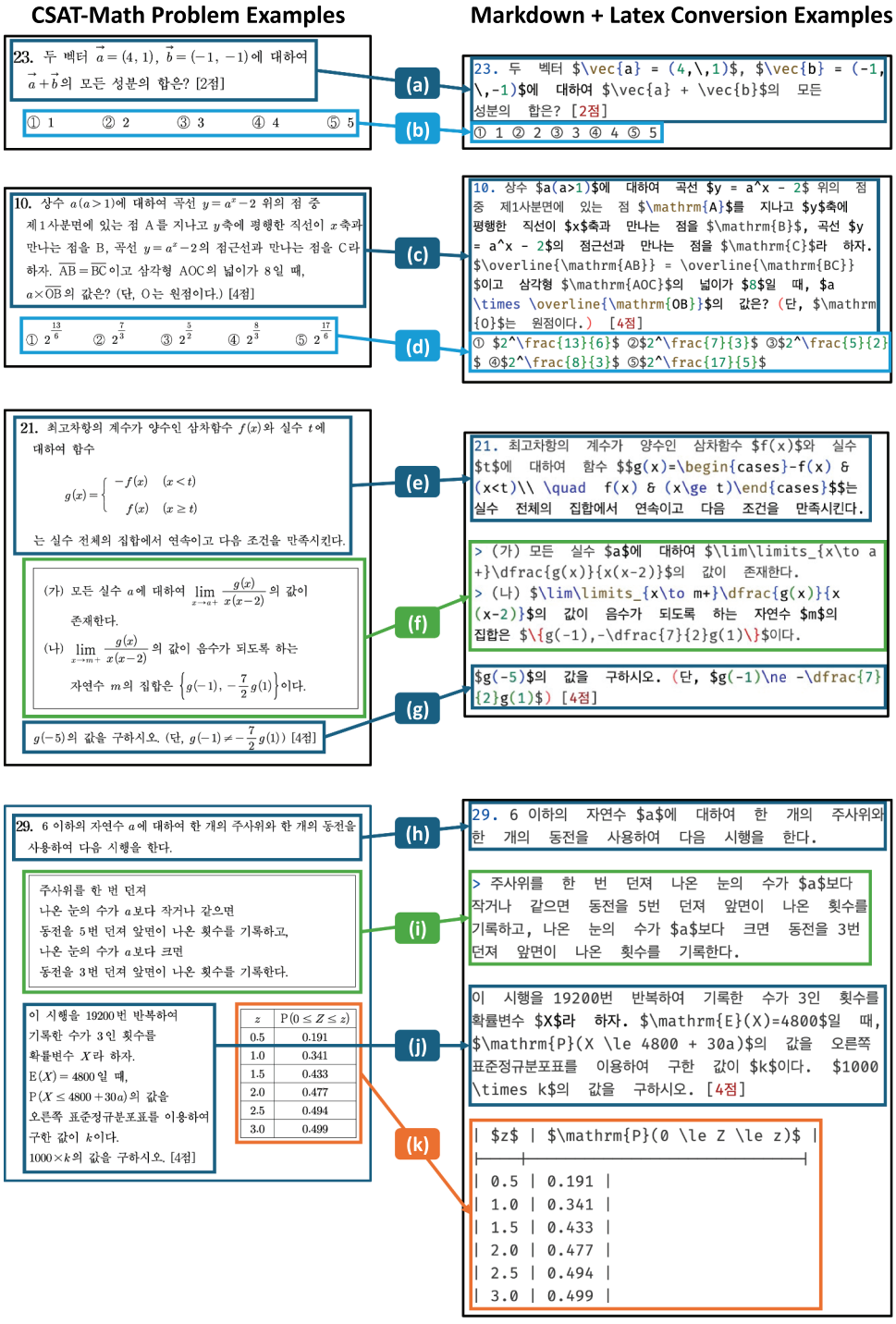


Figure 3. Example of a CSAT problem and its Digitalized CSAT-Math textual representation.

the subject source (Math I vs. Math II) is not indicated in the problem sheet and must be manually identified.

Because this information is not officially provided, we manually annotated each problem in the 1–22 range as belonging to Math I or Math II, constructing the prob_area field accordingly.

Ground-truth answers. The answer field holds the correct answer for each problem. Multiple-choice answers were stored as integers corresponding to their options (e.g., 172 → 1, 176 → 5). Short-answer

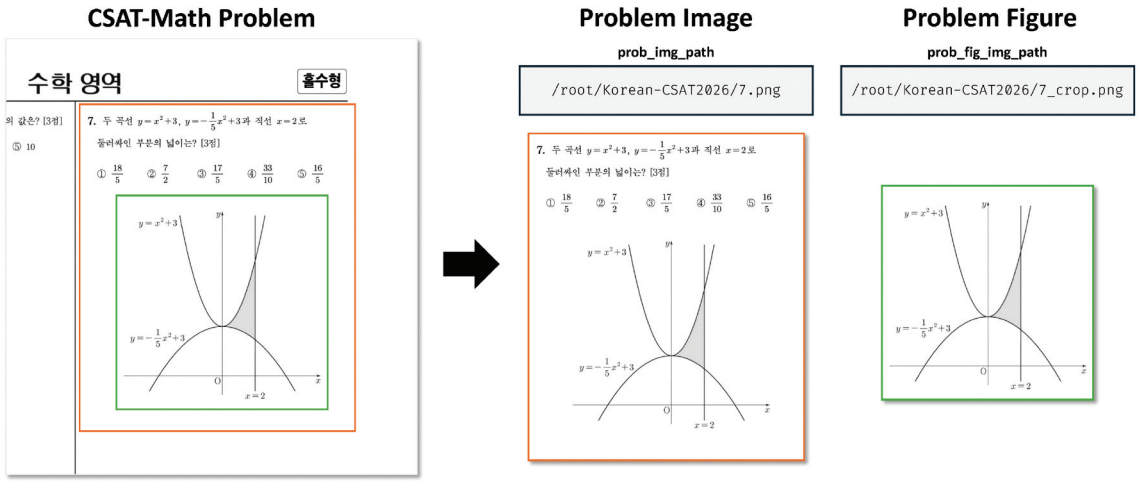


Figure 4. Examples of extracted CSAT problem images.

problems were recorded as integers within the 0–999 range. Expressing all answers in a unified numerical form enables consistent scoring across both multiple-choice and short-answer problems.

Image data construction

Image extraction. Many CSAT-Math questions include diagrams or visual elements. In the 2026 exam, 8 of 46 questions contained such components. To study how visual information affects LLM mathematical reasoning, we extracted two image types from the original PDF:

- **Problem Image**: Full problem region (prob_img_path)
- **Problem Figure**: Diagrams/Graphs/Visual elements (prob_fig_img_path)

Figure 4 shows examples of these extracted image components.

Quality assurance

Annotators and reviewers rendered each problem using Markdown editors (Obsidian[5] or StackEdit[6]) to verify visual alignment with the original PDF problem sheet. They checked Markdown + LaTeX rendering accuracy – including formula formatting, line breaks, symbols, and layout. A final pass by the supervisor ensured consistency and correctness across all entries.

To ensure transcription correctness in a systematic and reproducible way, we adopted three complementary measures: (i) a formal Labelling Guide that fixes all transcription conventions (TSV field rules, Markdown + LaTeX notations, handling of figures/tables/boxes, and answer encoding); (ii) pilot training on the 2022 CSAT exam, used both to develop the Guide and to train all 2026 annotators before annotation; and (iii) a multi-pass review of the aggregated dataset by an experienced reviewer with cross-year CSAT labelling experience, followed by a supervisor pass. Full details – including the contents of the Labelling Guide and the review workflow – are provided in Appendix G.

Prompt template for problem solving

LLM prompts consist of two components: (1) the problem, containing the CSAT-Math content, and (2) the instruction, which guides the model on how to solve the problem in a CSAT-appropriate manner.

To evaluate multimodal and multilingual capabilities, we combined three problem input modalities with two instruction-language settings, yielding six prompt conditions. Figure 5 illustrates these combinations.

Input Modalities. Because CSAT questions may include both textual and visual elements, we evaluate LLMs under three input formats:

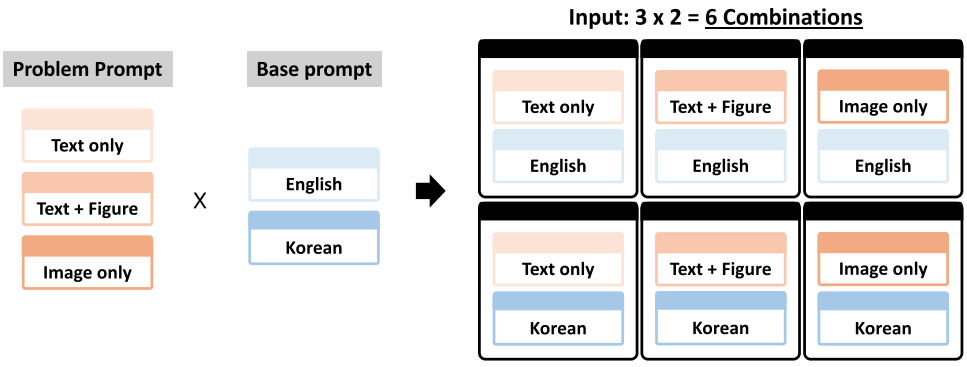


Figure 5. Example of prompt composition.

- **Text only**: The problem is provided solely as text (prob_desc). All diagrams or figures are omitted. This modality is supported by all models.
- **Image only**: The model receives the full problem as an image (Problem Image) along with the instruction. This setting mirrors actual test-taking conditions and evaluates the model's OCR and visual reasoning capabilities. Text-only LLMs are excluded from this condition.
- **Text + Figure**: The textual portion is provided as prob_desc, while diagrams or graphs are separately provided as images (Problem Figure). This hybrid setting requires both text understanding and visual interpretation. Text-only models are excluded.

Prompting Language. We use two instruction-language settings:

- **Korean**: Both the problem and the instruction are presented in Korean.
- **English**: Korean problem text with an English instruction.

We deliberately keep the original Korean problem text in both conditions and vary only the instruction language. This preserves the construct validity of evaluating LLMs on a Korean national exam (pre-translation would distort mathematical notation and idiomatic CSAT-specific phrasing), maintains comparable information content across text and image modalities (where the latter captures the Korean exam sheet as-is), and avoids introducing an external translator as a confounder – our English-instruction prompts instead include "Translate the problem to English" as the first solving step (Figure 6), so translation is performed internally by the same model under test.

Combining three modalities and two languages yields six conditions (Figure 5). Text-only models are evaluated under the two text-based conditions, while multimodal models are evaluated under all six. Examples of full prompt configurations are shown in Figure 6.

CSAT-Math LLM experiment design

In this section, we describe the overall experimental design. Specifically, we (1) list the evaluated LLMs and their execution environment, and (2) explain the goals and procedures aligned with the four core research questions (RQs).

Model list

We selected models satisfying four criteria: (i) strong baseline mathematical performance (the 24 highlighted models from the AIME 2025 benchmark[7] as of October 5, 2025); (ii) accessibility via OpenRouter on the same date, for reproducibility through a unified API interface; (iii) a release time preceding the 2026 CSAT date (November 13, 2025) to preserve a zero-leakage environment; and (iv) diversity, balancing proprietary families (GPT, Claude, Gemini) and open-weight families (Qwen, Llama, DeepSeek) across parameter scales

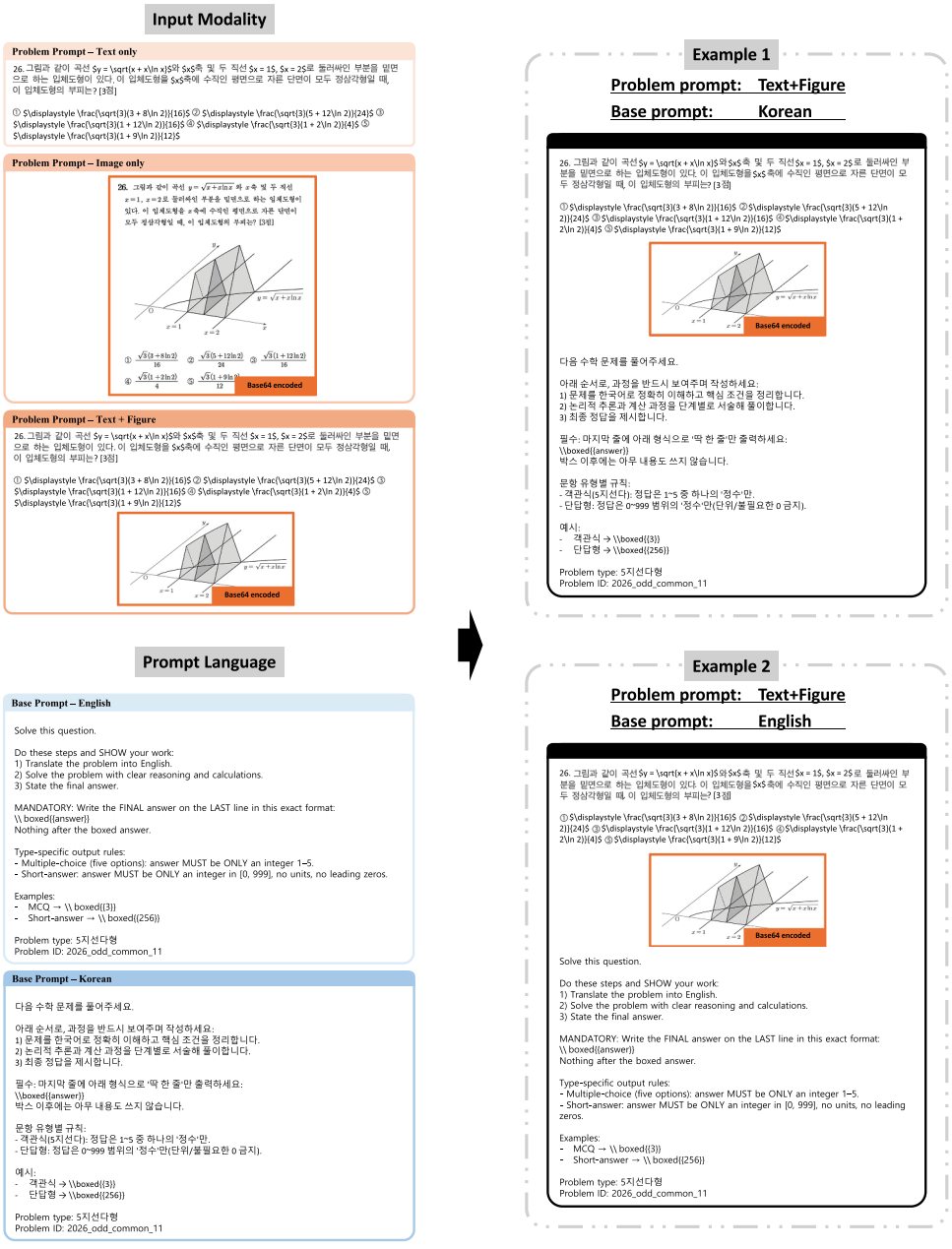


Figure 6. Example of a full prompt.

and architectures. Table 3 summarizes the final set, including provider, API pricing, parameter scale, and whether a built-in reasoning/thinking mode is supported.

Configuring model reasoning

Modern LLMs increasingly move beyond simple pattern matching and incorporate built-in reasoning mechanisms that analyze problem structure and generate intermediate thoughts. In this work, we define reasoning as "the internal thinking or multi-step inference process a model performs before producing a final answer." Typically, such reasoning processes consume separate reasoning tokens, distinct from the visible output tokens.

While both proprietary and open-weight models have begun to expose some reasoning functionality, these are often available only through dedicated "reasoning models" or with limited, coarse-grained options. Fine-grained control over reasoning strength is rarely supported. Therefore, for systematic comparison, we

Table 3. List of LLMs, providers, prices, model scales, and reasoning/thinking support. "1 M input/output" indicates the OpenRouter-listed price per 1 M input/output tokens. Model scale is the number of active parameters (estimated as ∼* when unavailable). Reasoning support is marked if available and "–" otherwise.

Type		Model	Model Provider	1 M Input	1 M Output	Scale	Reasoning Support
Closed	GPT	GPT-5 Codex	openai	$1.25	$10.00	∼ 635B *	–
		GPT-5	openai	$1.25	$10.00	∼ 330B *	
		GPT-5-mini	openai	$0.25	$2.00	∼ 27B *	
		GPT-5-nano	openai	$0.05	$0.40	∼ 15B *	
	Grok	Grok 4	xai	$6.00	$30.00	∼ 213B *	
		Grok 4 Fast	xai	$0.40	$1.00	∼ 40B *	–
	Claude	Claude 4.5 Sonnet	anthropic	$6.00	$22.50	∼ 100B *	
		Claude 4.5 Haiku	anthropic	$1.00	$5.00	∼ 30B *	
		Claude 4.1 Opus	anthropic	$15.00	$75.00	∼ 400B *	
	Gemini	Gemini 2.5 Pro	google-vertex/global	$2.50	$15.00	∼ 128B *	
		Gemini 2.5 Flash	google-vertex/global	$0.30	2.50	∼ 27B *	–
Open	GPT	gpt-oss 20B	hyperbolic	$0.04	$0.04	3.6B	
	Qwen	Qwen3 235B A22B	siliconflow/fp8	$0.30	$1.50	22B	–
		Qwen3 235B A22B Thinking 2507	siliconflow/fp8	$0.45	$3.50	22B	
	Deepseek	Deepseek R1 0528	novita/fp8	$0.27	$0.41	∼ 37B *	
		Deepseek V3.2 Exp	novita/fp8	$0.70	$2.50	∼ 37B *	–
	GLM	GLM 4.6	z-ai	$0.60	$2.20	32B	–
	Magistral	Magistral Medium 2506	mistral	$2.00	$5.00	45B	–
		Magistral Medium 2506 Thinking	mistral	$2.00	$5.00	45B	
	MiniMax	MiniMax M2	minimax	$0.30	$1.20	∼ 10B *	–
	Llama	Llama 4 Maverick	friendli	$0.20	$0.60	17B	–
		Llama 3.3 Nemotron Super 49B V1.5	deepinfra/fp8	$0.10	$0.40	49B	–
	Kimi	Kimi K2 0905	moonshotai	$0.60	$2.50	32B	–
		Kimi K2 0905 Thinking	moonshotai	$0.60	$2.50	32B	

conduct a dedicated reasoning study only on GPT-5, which allows us to control both reasoning strength and explanation length.

GPT-5 exposes two configurations: (1) Reasoning_Effort: internal reasoning strength (amount of reasoning tokens), and (2) Text_Verbosity: length of the final explanation. These two parameters are independent: for example, one can use high reasoning effort with a concise explanation (high effort + low verbosity), or minimal reasoning with a long explanation (minimal effort + high verbosity).

Reasoning_Effort. GPT-5 provides four discrete levels for internal reasoning:

- **minimal**: minimizes reasoning tokens; fastest responses
- **low**: basic reasoning; sufficient for simpler problems
- **medium (default)**: balanced reasoning; stable for typical math problems
- **high**: deep reasoning; suitable for difficult, complex problems

Text_Verbosity. Text_Verbosity controls the length of the final response, independently of Reasoning_Effort:

- **low**: concise answers; may omit detailed reasoning steps
- **medium (default)**: standard explanatory length
- **high**: detailed, extended explanations

Reasoning_effort × text_verbosity combinations

Combining the four effort levels with the three verbosity levels yields 12 distinct configurations. Table 4 summarizes their qualitative characteristics. We treat each configuration as a separate experimental condition and compare their behavior on the CSAT.

API and runtime environment

All experiments were conducted via the OpenRouter API[8] under fixed hyperparameters chosen for deterministic, reproducible evaluation: Temperature $= 0.0$ and Top-$\mathbf{p}$ $= 0.0$ (to eliminate sampling

Table 4. Qualitative characteristics of GPT-5 reasoning effort × text verbosity combinations.

Effort	Verbosity	Characteristics
minimal	low	Minimal reasoning × short output
	medium	Minimal reasoning × medium output
	high	Minimal reasoning × long output
low	low	Low reasoning × short output
	medium	Low reasoning × medium output
	high	Low reasoning × long output
medium	low	Medium reasoning × short output
	medium	Medium reasoning × medium output (default)
	high	Medium reasoning × long output
high	low	High reasoning × short output
	medium	High reasoning × medium output
	high	High reasoning × long output

variance, since our analysis isolates the effects of modality, language, and reasoning configuration), Max tokens unconstrained (capped only at the model's maximum), an exponential-backoff retry policy (up to 5 attempts with $5s \rightarrow 10s \rightarrow 20s \rightarrow 40s \ldots$ backoff), and no timeout.

Experiments were run by eight researchers on personal PCs and servers (Ubuntu 22.04 LTS or Windows 10, Python 3.10, openai==2.8.0 for API calls, pandas==2.3.3 for result aggregation). Up to 16 processes were executed in parallel across distinct (model, input-condition) pairs; within each process, all 46 questions are solved strictly sequentially (Q_{i+1} is sent only after Q_i's response is received), so the per-run total solving time equals the wall-clock time of a sequential solver.

LLM response logging and predicted CSAT-Math format

For every API call, the problem-solving script logs the full API request, the full model response, the extracted final answer, the per-problem response time (KST), API token usage (input/output/total), and any error type and number of retries. From these logs, we derive the Predicted CSAT-Math format, whose fields are summarized in Table 5.

Using this logs, we analyze correlations between performance and latency, modality-specific behavior, and stability across model families and sizes.

Goals of the CSAT-Math experiments

To answer the four research questions defined in 'Introduction' (RQ1–RQ4), we use the LLM response logs and the Predicted CSAT-Math data described in 'LLM Response Logging and Predicted CSAT-Math Format'. By systematically varying input conditions, problem characteristics, and reasoning configurations, we examine LLM mathematical problem-solving ability from multiple angles.

RQ1: Compare overall performance of recent LLMs and derive a ranking based on total score and accuracy over all 46 problems.

RQ2: Analyze which concepts and item types exhibit concentrated errors, considering subject area (Math I/II, Calculus, Geometry, Probability and Statistics), score weight (2/3/4 points), problem format (multiple-choice vs. short-answer), and structural/difficulty factors.

RQ3: Compare how input modality and prompting language affect accuracy, error patterns, and latency.

RQ4: Quantitatively evaluate how changes in reasoning strength affect performance, cost, and latency.

This evaluation framework cross-combines multiple axes—(i) model family and size, (ii) prompting language, (iii) input modality, and (iv) reasoning configuration – to measure LLM math capabilities in a fine-grained manner and to isolate the experimental conditions relevant to each RQ.

Table 5. Fields of the Predicted CSAT-Math format derived from per-call logs.

Field	Description
prob_id	Unique problem ID.
prob_type	Problem type: multiple-choice or short-answer.
prob_area	Subject area: Math I, Math II, Probability and Statistics, Calculus, or Geometry.
answer	Ground-truth answer (172–176 for MCQ, integer 0–999 for short-answer).
model_index	Internal model index.
model_id	OpenRouter model identifier (e.g., openai/gpt-5-codex).
provider	Serving provider (e.g., novita/fp8); affects quantization and pricing.
content_mode	Input modality: text_only, image_only, or text_fig.
language	Prompt language: ko or en.
variant_tag	Tag used for naming log files.
start_time_kst, end_time_kst	Per-problem solve start/end times (KST).
model_answer	Parsed answer; extracted from the content of the last \ boxed{} required by the instruction.
elapsed_sec	Per-problem solve time, computed from start/end_time_kst (seconds, millisecond precision).
prompt_tokens, completion_tokens, total_tokens	Input, output, and total token counts.
attempt_count	Number of attempts (1 = succeeded on first try; 0 = unrecovered failure).
error	Logged error type, if any.
request_body	Additional options passed in the model call (e.g., {"Reasoning_effort": "high," "text_verbosity": "low"}).

Overall experimental configuration

Table 6 summarizes our experimental design. We evaluate 24 models under all combinations of prompt language (2) × input modality (3), and additionally run a reasoning ablation for the GPT-5 family over all Reasoning_Effort and Text_Verbosity combinations ($4 \times 3 = 12$). This multi-faceted setup enables a precise analysis of math performance, modality sensitivity, language effects, and the trade-off between reasoning strength, cost, and latency.

Evaluation metrics

We compute evaluation metrics over the full Korean-CSAT-Math benchmark under a unified scoring scheme. Although human test-takers answer only one elective subject – solving 30 problems in total (22 common + 8 elective) – we require LLMs to answer all three elective sets, yielding 46 problems overall.

Table 7 summarizes the seven metrics we report per (model, condition) pair. Even open-weight models (e.g., gpt-oss-20B) incur API costs when accessed via OpenRouter; a consistent OpenRouter-based formula is applied to all models. As a worked example for the Score/Normalized Score, a model that correctly answers five 2-point, fourteen 3-point, and seventeen 4-point problems obtains $(2 \times 5) + (3 \times 14) + (4 \times 17) = 120$, normalized to $120 \times 100/152 = 78.9$.

LLM experiment results

LLM score leaderboard (Text-Only / Korean)

The overall performance of each model under the proposed Korean-CSAT LLM evaluation framework is summarized in Table A.8 in Appendix A. More detailed results and the latest leaderboard are

Table 6. Summary of experimental settings. We evaluate 24 models across language and input-modality conditions, and conduct additional reasoning configuration experiments (Reasoning_effort × Text_Verbosity) for GPT-5.

Axis	# Levels	Levels
Model	24	LLM and multimodal models (Table 3)
Language	2	Korean, English
Input modality	3	Text, Image, Text + Figure
GPT-5 Reasoning_Effort	4	Minimal, Low, Medium, High
GPT-5 Text_Verbosity	3	Low, Medium, High
Base combinations	$24 \times 2 \times 3$	Conditions shared by all models
GPT-5 extra combos	4×3	Effort–Verbosity combinations

Table 7. Evaluation metrics computed per (model, condition) pair over the 46 problems.

Metric	Definition
Score	Sum of the point values of correctly answered problems.
Normalized Score	Score scaled to a 100-point base, where 152 is the maximum: score × 100/152 (cf. Table 1).
Latency	Time from API invocation to the full response, recorded as elapsed_sec in seconds (millisecond precision).
Input Token Usage	Number of tokens in the prompt.
Output Token Usage	Number of tokens generated in the completion.
Total Token Usage	Sum of input and output tokens (the total_tokens field).
Cost ($)	Per-problem inference cost, computed from the OpenRouter pricing policy for each model (Table 3) using input and output token counts.

available on the official website.[9] Below, we present key findings observed under this specific configuration.

We focus on Text-only/Korean conditions, as some models lack image support and the benchmark is Korean-based.

Model-wise performance analysis

As shown in Figure 7, GPT-5 Codex is the only model that achieved a perfect normalized score of 100 in the main analysis. Other GPT-5 family models (GPT-5: 95.7, GPT-5-mini: 93.5), Grok 4 models (Grok 4: 97.8, Grok 4 Fast: 95.7), and Gemini 2.5 Pro (91.3) also surpassed 90 points, demonstrating strong performance among recent proprietary models. Notably, gpt-oss-20B achieves performance comparable to the top tier despite being a relatively lightweight open-weight model.

Most of the remaining models scored in the 70–90 range. For example, Qwen3 235B A22B Thinking 2507 and GLM 4.6 achieved scores close to 90, while DeepSeek V3.2 Exp and Kimi K2 Thinking performed in the 80-point range. In contrast, the Claude family and Gemini 2.5 Flash underperformed relative to other proprietary models, and Llama 4 Maverick ranked last with a score of 21.7. These results indicate substantial variance in mathematical problem-solving capability across LLMs.

Model-wise latency analysis

Human test-takers are given 100 minutes to solve 30 problems, whereas our setup requires LLMs to solve all 46 problems. To account for the increased number of questions, we scale the time budget to 152 minutes and evaluate models under both unlimited-time and time-limited settings. We emphasize that each (model, condition) run solves the 46 problems strictly sequentially within a single process (Q_{i+1} is sent only after Q

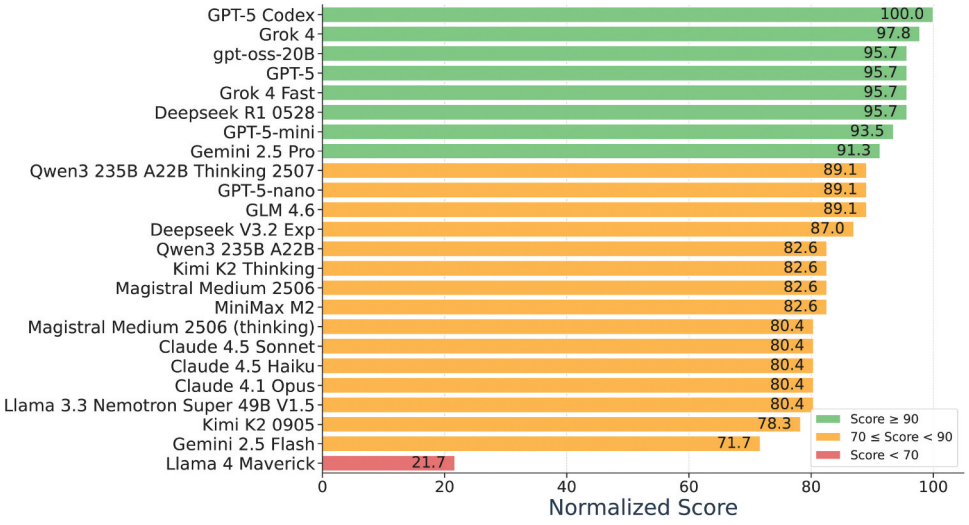


Figure 7. Model-wise performance comparison (input modality: Text-only, prompt language: Korean).

$_i$'s response is received), so the reported total solving time equals the wall-clock time of a sequential solver and is directly comparable to the 152-minute budget; the across-condition parallelism noted in Secion "CSAT-Math LLM Experiment Design" only accelerates aggregate throughput and does not affect per-run timing.

Unlimited-time performance

Figure 8 plots model latency and performance under the Text-only, Korean prompt condition. A notable observation is that most models solved all 46 problems within 100 minutes, i.e., within the human time budget for only 30 problems. Even GPT-5 Codex, the top-performing model, completed all questions in 72 minutes.

Figure 8 shows the latency–score distribution under unlimited reasoning, Text-only input, and Korean prompts.

Time-limited performance

According to Figure 8, four models – Qwen3 235B A22B Thinking 2507, GLM 4.6, DeepSeek R1 0528, and Kimi K2 Thinking – failed to complete all 46 problems within the 152-minute limit. Figure 9 compares their scores under both conditions.

Under the 152-minute limit, all four models show reduced scores. Kimi K2 Thinking is most affected (82.6→12.5 points due to 565-minute total solving time), while Qwen3 235B loses 24 points and GLM 4.6 loses 3 points.

Model-wise cost analysis

Figure 10 shows the total API cost for each model to solve all 46 problems. Proprietary models are relatively expensive: Grok 4 costs $11.48, Gemini 2.5 Pro costs $4.82, and Claude 4.1 Opus costs $4.01. By contrast, open-weight models are much cheaper: gpt-oss-20B costs $0.01, while DeepSeek V3.2 Exp and Llama 4 Maverick each cost $0.04. Remarkably, gpt-oss-20B achieves a score of 95.7 (3rd) at the lowest cost ($0.01). Overall, some models show disproportionate cost increases for modest gains, while others achieve near – state-of-the-art scores at a fraction of the cost.

Model-wise efficiency analysis

We define three efficiency scores to analyze performance against time and cost jointly:

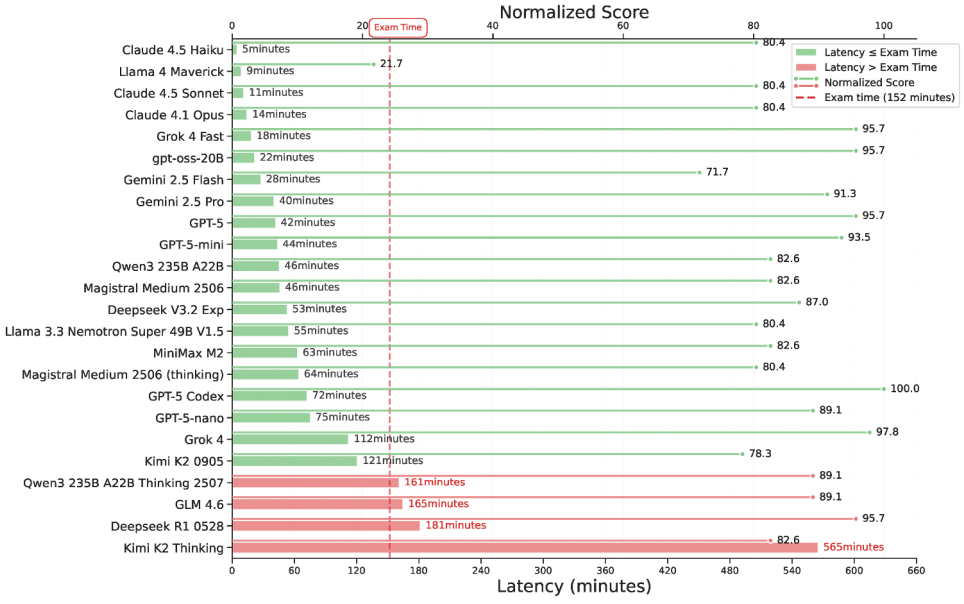


Figure 8. Latency vs. score by model (Text-only, Korean).

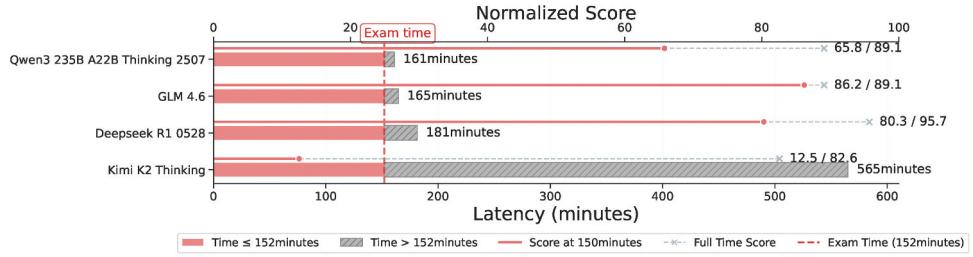


Figure 9. Time-limited performance of models exceeding the 152-minute budget.

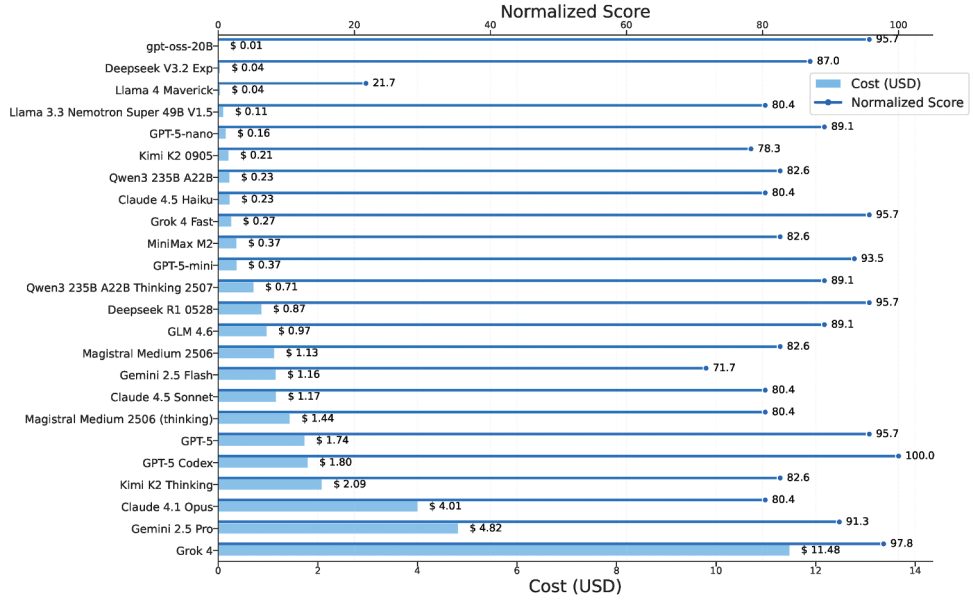


Figure 10. Cost vs. score by model (Text-only, Korean).

$$(1)\quad \mathrm{Eff}_t = \frac{\mathrm{score}}{\mathrm{latency(min)}} \qquad (2)\quad \mathrm{Eff}_c = \frac{\mathrm{score}}{\mathrm{cost(\$)}} \qquad (3)\quad \mathrm{Eff}_{t,c} = \frac{\mathrm{score}}{\frac{\mathrm{latency}}{152\ \mathrm{min}} + \frac{\mathrm{cost}}{1\ \$}}$$

Eff$t$ in (1) captures *time efficiency*: higher values indicate a higher score in less time. Eff$c$ in (2) captures *cost efficiency*: for the same performance, lower-cost models have higher Eff$c$. Finally, Efft, c in (3) measures *joint time – cost efficiency*: higher values indicate high performance with relatively low time and monetary cost.

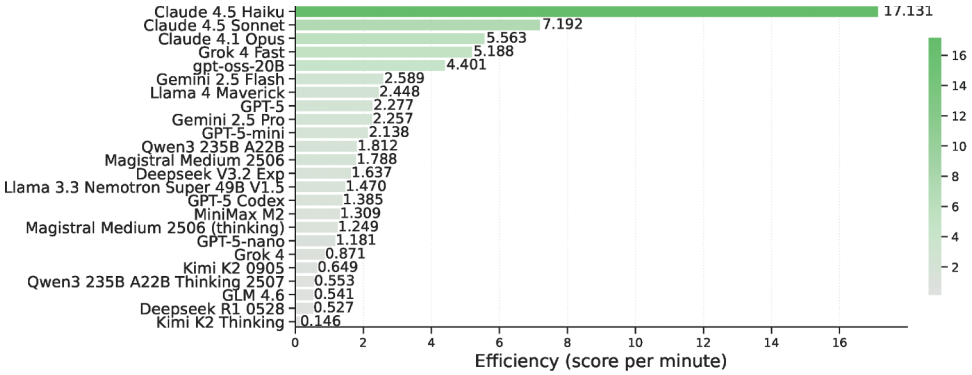
Figure 11a visualizes time-normalized performance using Eff_t. Claude 4.5 Haiku achieves the highest Eff_t, solving all 46 problems in 5 minutes with a score of 80. Other Claude variants rank second and third, followed by Grok 4 Fast and gpt-oss-20B, indicating strong time efficiency.

Figure 11b shows cost-normalized performance via Eff_c. gpt-oss-20B achieves a normalized score of 95.7 (3rd) at a cost of $0.01 (1st), resulting in by far the highest Eff_c. DeepSeek V3.2 Exp ranks second, while Grok 4, the most expensive model ($11.48), ranks last on this metric.

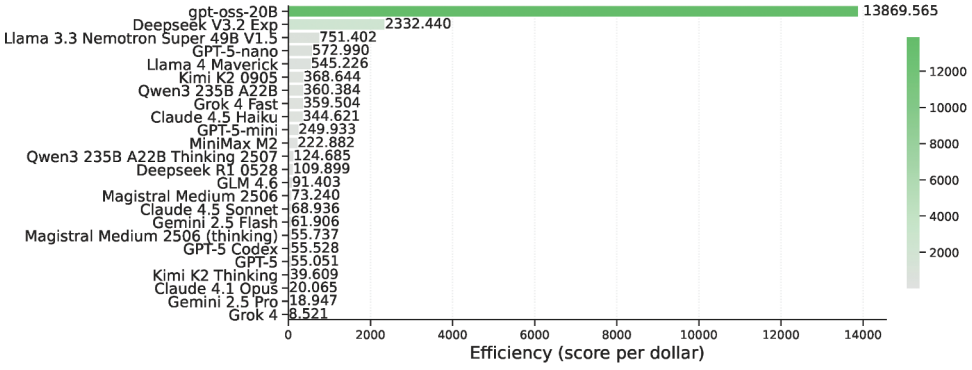
Figure 11c visualizes $\mathrm{Eff}_{t,c}$, capturing overall performance per combined time and monetary resource. Under this joint metric, gpt-oss-20B is the best model. Claude 4.5 Haiku ranks second: despite its relatively lower cost efficiency, its exceptional time efficiency boosts its overall $\mathrm{Eff}_{t,c}$.

Full results for all input combinations are reported in Appendix B.

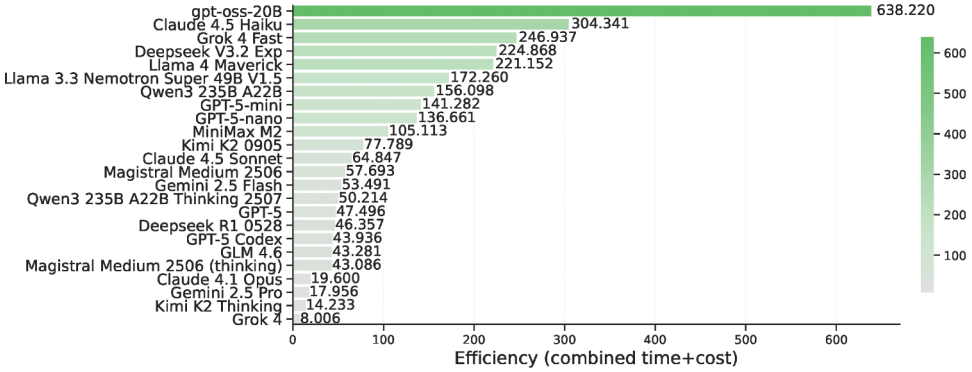
In summary, models that are optimal in terms of time efficiency differ from those that are optimal in terms of cost efficiency. If users prioritize fast responses, or if they need multiple repeated runs, Claude 4.5 Haiku and Claude 4.5 Sonnet are attractive choices; if they prioritize low cost, gpt-oss-20B and DeepSeek V3.2 Exp offer highly favorable performance – resource trade-offs.



(a) Time-based Efficiency Score (Eff_t) across models (Text-only, Korean)



(b) Cost-based Efficiency Score (Eff_c) across models (Text-only, Korean)



(c) Joint time–cost Efficiency Score ($\mathrm{Eff}_{t,c}$) across models (Text-only, Korean)

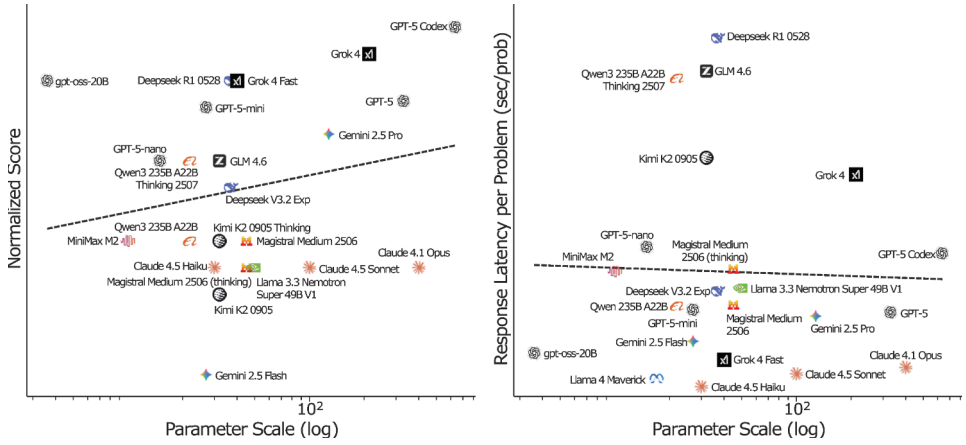
Figure 11. Efficiency comparison across performance, time, and cost (Text-only, Korean).

LLM performance under varying input and prompt conditions

Performance differences by model parameter size

Performance by model size. Figure 12 illustrates how model parameter size relates to performance and per-problem latency. As shown in Figure 12(a), larger models generally achieve higher scores, but this trend is far from strictly linear. For example, gpt-oss-20B contains roughly seven times fewer active parameters than Qwen3 235B A22B, yet it outperforms the larger model. This demonstrates that parameter count alone is not a definitive determinant of math problem-solving ability.

Relationship between model size and latency. Figure 12(b) shows that some large models maintain relatively short latencies despite their size, whereas several mid-sized models show extremely long per-problem processing times. This indicates that parameter size produces a complex trade-off structure involving accuracy, latency, and efficiency; no single dimension is sufficient for fully characterizing model performance.



(a) Performance by model size. Llama 4 Maverick is omitted for legibility.

(b) Latency by model size. Kimi K2 Thinking (737 sec/problem) is omitted for legibility.

Figure 12. Model performance and per-problem latency by parameter size (Text-only, Korean).

Full results for all modality – language combinations are presented in 10.

In summary, although larger models generally show stronger performance, model size is not an absolute predictor of accuracy, and certain smaller models (e.g., gpt-oss-20B) outperform larger models in both accuracy and efficiency.

Performance differences by input modality

Figure 13(a) visualizes how input modality affects proprietary models. Key observations are as follows:

- **GPT-5 family**: Achieves the highest scores under Text-only or Text + Figure. GPT-5 Codex reaches **100 points** with Text + Figure. Image-only input lowers performance.
- **Claude family**: Image-only input sharply degrades performance across all variants; Text and Text + Figure perform similarly.
- **Grok family**: Similar to Claude – Image-only performs worst; Text and Text + Figure perform better.
- **Gemini family**: Displays the opposite trend, with Image-only outperforming other modalities, indicating strong visual-processing capability.

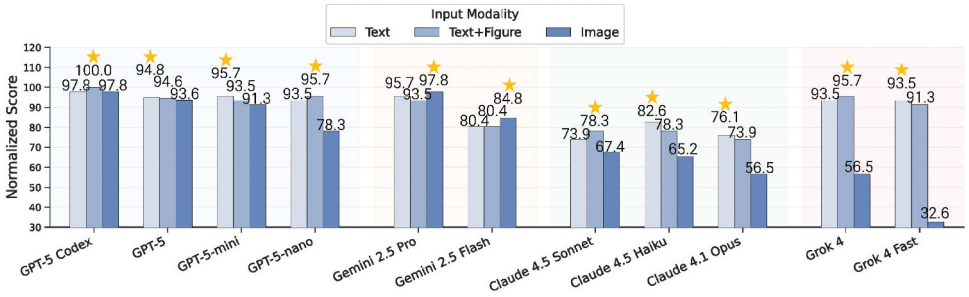
Figure 13(b) shows win rates by input modality, separated by prompt language and weight accessibility. As in Figure 13(a), proprietary models prefer Text and Text + Figure, with lower win rates for Image-only. Open-weight models differ slightly: with Korean prompts, Image input has a near-zero win rate, but this is based on only three image-capable open-weight models (Qwen3 235B A22B, Qwen3 235B A22B Thinking 2507, Llama 4 Maverick), limiting interpretability.

Performance differences by prompting language

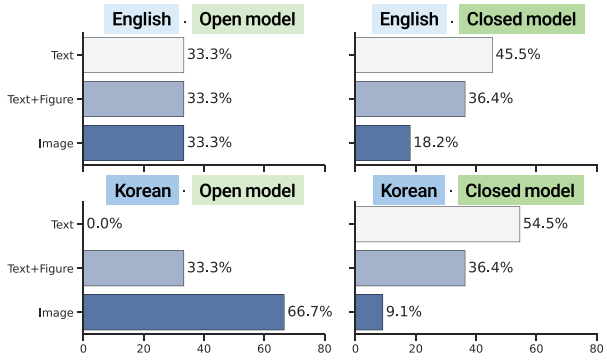
Figure 14(a) shows performance differences between Korean and English prompts across model families. GPT-5 Codex and gpt-oss-20B perform better with Korean prompts, while Claude and Grok models also show an advantage with Korean input. In contrast, Gemini models and major open-weight models (Qwen, Kimi, etc.) exhibit substantially higher performance with English prompts; Korean prompts lead to significant drops in accuracy.

Figure 14(b) presents win rates by prompting language grouped by input modality and weight accessibility. Across all conditions, English prompts show a clear advantage. Even among Text-only inputs, open-weight models benefit more from English prompts than proprietary models. Although results for image-based prompts are influenced by the small number of image-capable open-weight models, English prompts consistently help across modalities.

We hypothesize that two complementary factors plausibly explain these patterns: (i) an overall English-instruction advantage consistent with the well-documented English skew of pretraining and instruction-tuning corpora for current LLMs (Park et al. 2025; Zhou et al. 2025), and (ii) model-specific differences in Korean multilingual alignment that determine the *direction* of the gap. We emphasize that these are



(a) Model-family performance by input modality. Highest score per model is marked with a star.



(b) Win rate by input modality, prompt language, and weight accessibility

Figure 13. Performance and win rate comparison across input modalities.

hypotheses rather than conclusions of our experimental design: a more conclusive analysis would require comparing models for which the language composition of training data is disclosed (e.g., Korean-specialized LLMs) or running controlled fine-tuning ablations on Korean instruction-tuning ratios. Full discussion is provided in Appendix H; we leave such a mechanistic study to future work.

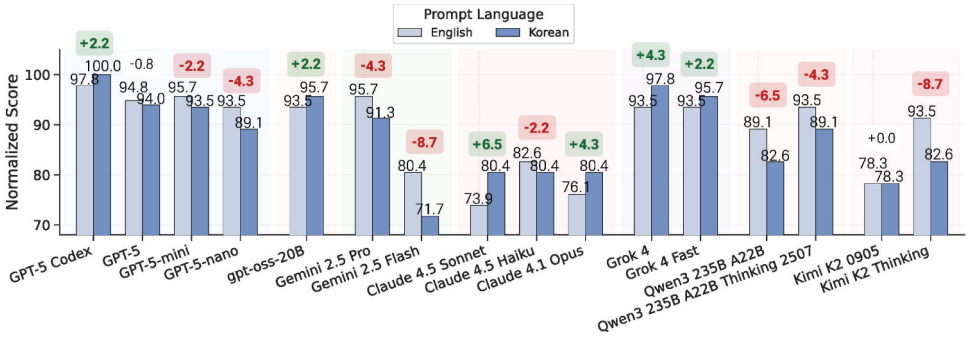
Qualitative error analysis: format compliance, not mathematical understanding. A per-problem analysis of disagreements between the two prompting languages shows that the gap is not universal – English prompts dominate 54.8% of disagreements while Korean prompts dominate 45.2%. More importantly, the EN-correct/KO-wrong cases are dominated by *final-answer formatting* failures (e.g., a correct value emitted directly instead of the required option index in 1–5, or a final answer that is missing or unparsable) rather than failures of mathematical reasoning, as illustrated by Llama 4 Maverick on common_2 (correct value 13 under both languages, but only the English prompt maps it to the required option 4). The full discordance counts, per-category error breakdown, and additional examples are provided in Appendix J. The higher English-prompt accuracy observed for several smaller and faster models on this Korean exam therefore stems primarily from more stable *answer-format compliance*—option-index mapping, boxed-answer formatting, and final-answer extractability – rather than from superior comprehension of Korean mathematics itself.

Analysis of LLM performance by mathematical problem characteristics

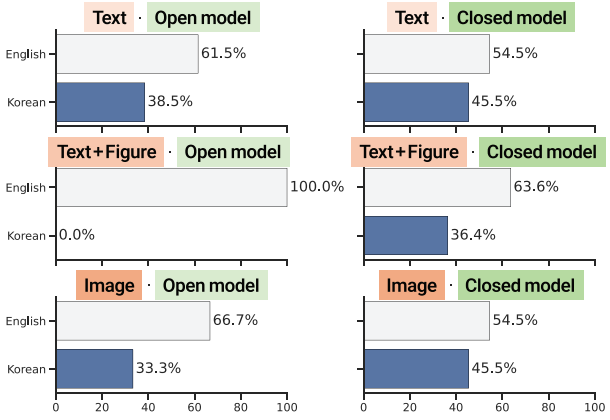
Performance differences by problem category

Figure 15 shows LLM performance by subject category. Labels indicate normalized scores and raw scores (score divided by the category maximum).

Models demonstrate their weakest performance in Calculus, with Geometry ranking as the second most challenging category. The difficulty of Calculus stems from its demand for complex symbolic manipulation and multi-step logical reasoning, whereas Geometry introduces additional variability across models due to its reliance on spatial interpretation and visual understanding.



(a) Model-family performance by prompting language. Green = Korean better, Red = English better, White = difference < 2 points.



(b) Win rate by prompting language across modalities and accessibility

Figure 14. Performance and win rate comparison across prompting languages.

Averaged across models, performance is lowest in Calculus (77.7 pts), followed by Geometry (79.2 pts), Common Math (83.9 pts), and Probability & Statistics (84.7pts). Full modality – language results are provided in Appendix D.

Performance differences by problem attributes

Differences by problem type (multiple-choice vs. short-answer). Figure 16(a) compares performance by problem type. LLMs perform much better on multiple-choice questions: the average Normalized Score is 91 for five-option problems, versus 68 for short-answer problems.

However, this gap is not solely due to format: in CSAT-Math, most 4-point (hardest) problems are short-answer, confounding type, and difficulty.

Differences by problem score (difficulty level). A consistent decline in accuracy is observed as problem difficulty increases. Normalized scores for 2- and 3-point questions range from 93 to 97, while 4-point questions drop to 71. This suggests that the CSAT-Math difficulty scale transfers naturally to LLMs, making the hardest questions difficult for both humans and models.

Figure 16(b) further breaks down performance by the intersection of type (multiple-choice/short-answer) and score (2/3/4). Performance is high for all 2–3-point combinations, but all 4-point combinations – multiple-choice or short-answer – show large accuracy drops and increased variance across models.

Overall, LLMs perform well on easier combinations (2–3 points, multiple-choice), but struggle significantly on harder combinations, especially 4-point short-answer questions. Complete results for all modality – language combinations are detailed in Appendix E.

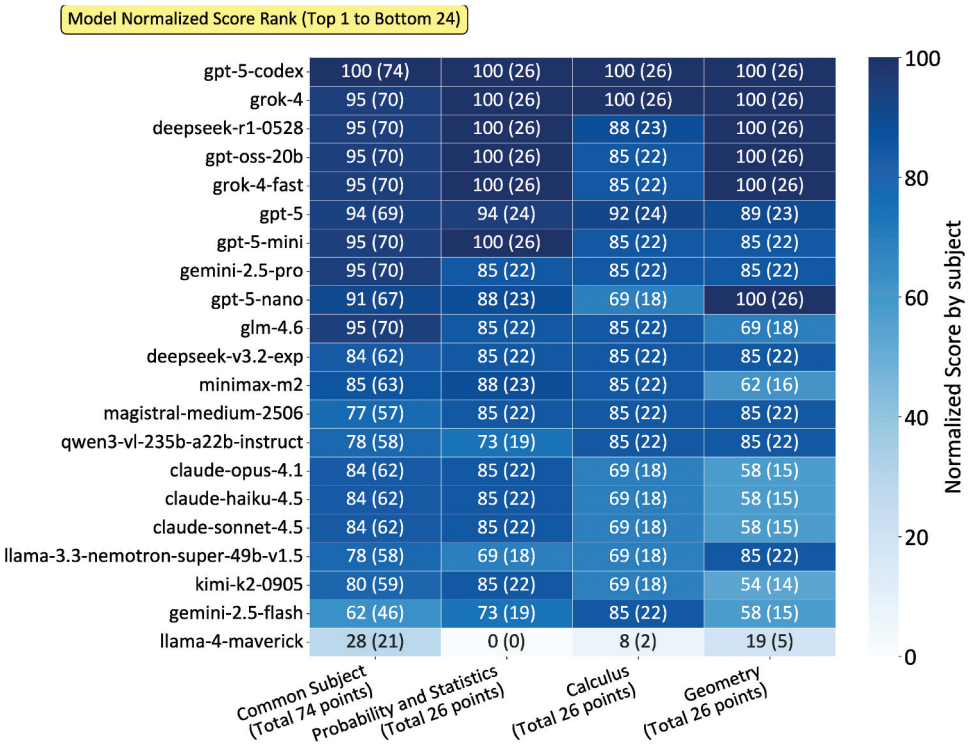


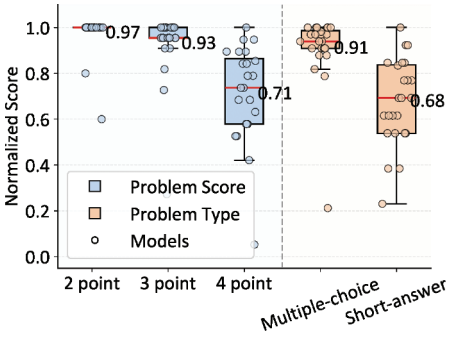
Figure 15. LLM performance by problem category. Normalized scores are computed by dividing each model's score by the category maximum. Models are sorted by overall normalized score.

Effectiveness of reasoning-oriented inference vs. baseline

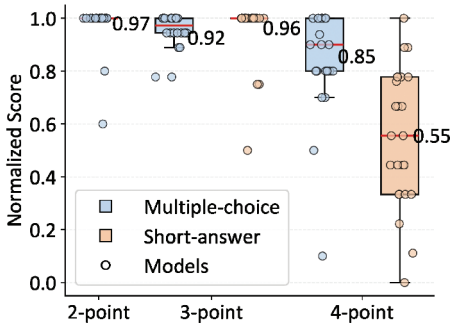
Impact of reasoning configuration on model performance

Contributions of Reasoning_Effort and Text_Verbosity. Figure 17(a) shows GPT-5 performance across Reasoning_Effort – Text_Verbosity settings: higher Reasoning_Effort consistently improves accuracy, while higher Text_Verbosity yields almost no change.

This aligns with the definitions provided in the GPT-5 documentation.[10] Text_Verbosity controls the number of output tokens – its effect is limited to explanation length and does not alter internal reasoning. In contrast, Reasoning_Effort explicitly controls the number of internal reasoning tokens generated prior to producing an answer, thus directly affecting the model's internal inference process. Therefore, the observed performance gains primarily reflect increased reasoning effort, not verbosity.



(a) Performance distribution by score (2/3/4 points) and problem type (multiple-choice/short-answer).



(b) Performance by score–type combinations. (No 2-point short-answer problems in 2026 CSAT.)

Figure 16. LLM performance by problem characteristics. (a) examines score and type independently; (b) analyzes joint score–type combinations.

Reasoning cost and the resulting efficiency trade-off

Reasoning configurations induce large changes in cost (Figure 17; full per-configuration efficiency trends in Appendix F). The medium-effort/minimal-verbosity setting attains the best efficiency scores: minimal verbosity reduces tokens while preserving accuracy, maximizing return per time and per dollar. In contrast, high-effort/high-verbosity is worst on all efficiency metrics – despite the highest accuracy (100.0), only 2–3 points above typical medium-effort settings, it requires 4–5× more tokens and yields efficiency scores below half of baseline. Higher reasoning effort thus delivers only modest performance gains at substantially higher cost; baseline or minimal settings provide the best practical trade-off, and high effort is useful only when score maximization is the sole objective regardless of cost.

An extension of this analysis to three open-weight non-thinking/thinking model pairs (Qwen3 235B A22B, Magistral Medium 2506, Kimi K2) – where the same qualitative trade-off recurs with strongly model-specific magnitudes, including a near-total collapse for Kimi K2 Thinking under the time-limited setting – is provided in Appendix K.

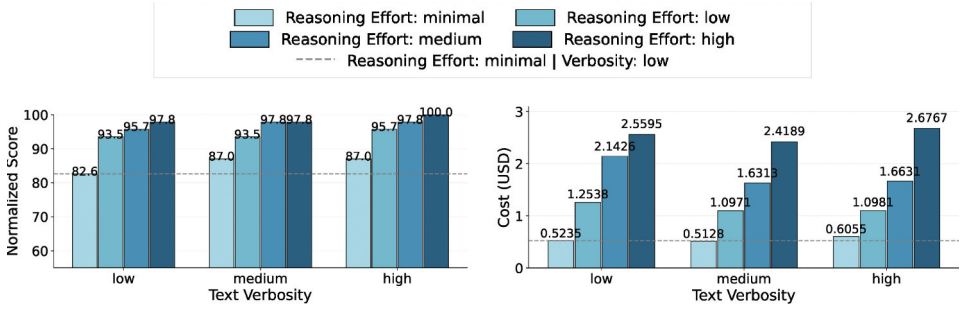
To clarify the contribution of CSAT-Math relative to established mathematical reasoning benchmarks (AIME, IMO, GSM8K, MATH, MathVista, MATH-Vision), Appendix I provides a detailed comparison. In summary, CSAT-Math is intended as a complement rather than a replacement: the top of our leaderboard largely overlaps with the strongest performers on AIME-style benchmarks (Artificial Analysis 2025; Patel et al. 2024), indicating that CSAT-Math captures general mathematical reasoning ability rather than dataset-specific artifacts, while the differences highlight dimensions – multimodal sensitivity, curriculum- and difficulty-specific weaknesses, efficiency-aware evaluation, and contamination-free evaluation – that public English-centric benchmarks underexplore.

Key findings

We summarize the main findings from our zero-leakage evaluation of 24 LLMs on the 2026 Korean CSAT Mathematics exam, organized by RQ.

RQ1. Overall Mathematical Reasoning Ability Under a strictly zero-leakage CSAT evaluation, perfect performance was achieved only by the GPT-5 family, indicating that CSAT-level mathematical reasoning remains limited to the most advanced models. However, gpt-oss-20B attained 95.7 points despite its smaller scale, demonstrating that model size alone does not determine reasoning ability and that strong cost – performance trade-offs are achievable.

RQ2. Effects of Input Conditions Text-only inputs consistently yielded the highest accuracy, with Text + Figure closely matching them, while Image-only inputs underperformed across nearly all models. Prompt language effects depended on model scale: large models benefited from Korean prompts, whereas smaller models showed notable gains with English, suggesting higher sensitivity to multilingual imbalance.



(a) GPT-5 performance across `Reasoning_Effort`-`Text_Verbosity` settings

(b) Cost variation across `Reasoning_Effort` and `Text_Verbosity` combinations

Figure 17. Comparison of Reasoning_Effort and Text_Verbosity in GPT-5 configuration.

RQ3. Effects of Problem Characteristics Performance dropped sharply on high-difficulty (4-point) problems, especially in Calculus and Geometry, highlighting persistent weaknesses in multi-step logical and spatial reasoning. The CSAT's human-designed difficulty structure induced clear performance stratification in LLMs, mirroring human examinee patterns.

RQ4. Effects of Enhanced Reasoning Increasing test-time reasoning substantially improved accuracy but caused a four- to five-fold increase in token usage and significant efficiency loss. Performance gains were driven primarily by reasoning effort rather than verbosity, indicating that excessive reasoning is often impractical and that adaptive reasoning strategies are necessary.

Conclusion

This study introduces a benchmark and evaluation framework designed to assess the mathematical reasoning capabilities of state-of-the-art LLMs under a fully contamination-free setting using the 2026 Korean CSAT Mathematics exam. All 46 problems were digitized within two hours of the exam's public release, and 24 models were evaluated across systematically varied input modalities, prompt languages, and reasoning intensities.

Experimental results show that while the strongest commercial models achieve scores above 90 on CSAT-level mathematics, their performance varies substantially depending on problem characteristics (e.g., Calculus, high-point problems), input conditions (language and modality), and the presence of time constraints. Although reasoning-enhanced modes improve absolute accuracy, they incur high computational and monetary costs, making baseline configurations more practical and cost-efficient for many use cases.

This work provides three main contributions. First, it establishes a fully contamination-free evaluation environment that fundamentally eliminates the data leakage concerns present in prior benchmarks. Second, it formalizes a standardized digitization pipeline that converts human-targeted exam materials into LLM-ready evaluation data. Third, it proposes a practical assessment perspective beyond accuracy to incorporate time, cost, and overall efficiency.

The study nevertheless has several limitations, including restrictions arising from the use of OpenRouter APIs, the absence of Korean-specialized models, the reasoning experiments being limited to the GPT-5 family, and the lack of direct comparison with human test-taker statistics. Future work includes longitudinal CSAT analyses, alignment with official scoring distributions, extensions to additional subject areas, and evaluation of Korean-optimized models.

Ultimately, this study highlights the need to shift LLM assessment from "how many problems the model gets correct" to "how efficiently and reliably the model problems under real-world constraints." We expect that the proposed evaluation framework will serve as a new reference point for assessing and deploying LLMs in high-stakes exam settings such as the CSAT.

Notes

1. https://suneung.re.kr/sub/info.do?m=0101s=suneung
2. https://www.news1.kr/society/education/5973136
3. https://www.suneung.re.kr/boardCnts/fileDown.do?fileSeq=ead4803d0df8cda516056b38fef4c7ae
4. Released at 14:10 KST via the KICE website: https://cdn2.kice.re.kr/suneung-26/index.html
5. https://obsidian.md/
6. https://stackedit.io/
7. https://artificialanalysis.ai/evaluations/aime-2025
8. https://openrouter.ai
9. https://isoft.cnu.ac.kr/csat2026/
10. https://platform.openai.com/docs/guides/latest-model

Author contributions

CRediT: **Inbum Heo:** Conceptualization, Investigation, Methodology, Software, Visualization, Writing – review & editing; **Jeesu Jung:** Conceptualization, Investigation, Validation, Writing – original draft; **Taewook Hwang:** Software, Validation, Writing – original draft; **Hyuk Namgoong:** Data curation, Investigation, Resources; **Hyein Seo:** Data curation, Investigation; **Yerim Han:** Data curation, Investigation; **Eunbin Kim:** Data curation, Investigation; **Hyeonseok Kang:** Data curation, Investigation; **Sangkeun Jung:** Funding acquisition, Project administration, Supervision, Validation, Writing – review & editing.

Data availability statement

The model inference outputs and evaluation results that support the findings of this study are openly available in Zenodo at https://doi.org/10.5281/zenodo.18530701, distributed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license. The deposited dataset contains 347 TSV files comprising per-model inference results across all experimental conditions (3 input modalities, 2 prompt languages, 2 reasoning modes) and 9 ablation summary tables reporting accuracy, weighted scores, latency, token usage, and cost for all evaluated models. These data are sufficient to reproduce all tables, figures, and statistical analyses reported in this paper.

The original exam materials – including problem text descriptions, scanned images (54 PNG files, 7.40 MB), and answer keys for 46 problems (33 multiple-choice and 13 short-answer items with point values of 2, 3, or 4) from the 2026 Korean CSAT Mathematics section (Odd Type) – are not included in the deposited dataset due to copyright restrictions imposed by the Korea Institute for Curriculum and Evaluation (KICE). These materials were used under license for this study and can be obtained directly from the official KICE website (https://www.suneung.re.kr). Detailed descriptive statistics of the nonpublic dataset, including distributions by subject area, problem type, point value, and image metadata, are provided in the project repository to support reproducibility.

Open Scholarship Statement



This article has earned the Center for Open Science badge for Open Data. The data are openly accessible at https://doi.org/10.5281/zenodo.18530701

Disclosure statement

No potential conflict of interest was reported by the author(s).

During the preparation of this work, the authors used ChatGPT and Claude in order to assist with translation and English editing. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Funding

This work was supported by Chungnam National University, the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) [RS-2025-00556217314820926401 01], and the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2019-II190004]; BK21 FOUR Program by Chungnam National University Research Grant, 2026.

ORCID

Goun Pyeon http://orcid.org/0009-0004-2420-5956
Jeesu Jung http://orcid.org/0000-0003-1684-0517
Taewook Hwang http://orcid.org/0000-0003-2440-4707
Hyuk Namgoong http://orcid.org/0000-0001-5166-3064
Hyein Seo http://orcid.org/0000-0003-3107-0880
Eunbin Kim http://orcid.org/0009-0008-2704-3752
Sangkeun Jung http://orcid.org/0000-0003-3531-0618

CRediT (Contributor roles taxonomy)

• Goun Pyeon: Conceptualization, Methodology, Formal analysis, Investigation, Writing – Review & Editing, Visualization
• Inbum Heo: Conceptualization, Methodology, Software, Investigation, Writing – Review & Editing, Visualization
• Jeesu Jung: Conceptualization, Validation, Investigation, Writing – Original Draft
• Taewook Hwang: Conceptualization, Validation, Software, Writing – Original Draft
• Hyuk Namgoong: Data Curation, Investigation, Resources
• Hyein Seo: Data Curation, Investigation
• Yerim Han: Data Curation, Investigation
• Eunbin Kim: Data Curation, Investigation
• Hyeonseok Kang: Data Curation, Investigation
• Sangkeun Jung: Validation, Supervision, Project Administration, Writing – Review & Editing, Funding Acquisition

References

Artificial Analysis. 2025. "Aime 2025 Benchmark Leaderboard."

Bai, Y., S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, J. Kerr, J. Mueller, J. Ladish, J. Landau, K. Ndousse, K. Lukošiūtė, L. Lovitt, M. Sellitto, N. Elhage, N. Schiefer, N. Mercado, N. Dassarma, R. Lasenby, R. Larson, S. Ringer, S. Johnston, S. Kravec, S. E. Showk, S. Fort, T. Lanham, T. Telleen-Lawton, T. Conerly, T. Henighan, T. Hume, S. Bowman, Z. Hatfield-Dodds, B. Mann, D. Amodei, N. Joseph, S. McCandlish, T. B. Brown, and J. Kaplan. 2022. "Constitutional Ai: Harmlessness from Ai Feedback." Arxiv.

Balloccu, S., P. Schmidtová, M. Lango, and O. Dusek. 2024. "Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs Mar."

Balunovic, M., J. Dekoninck, I. Petrov, N. Jovanović, and M. Vechev. 2025. "Matharena: Evaluating LLMs on Uncontaminated Math Competitions." The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track, San Diego, CA.

Cobbe, K., V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, et al. 2021. "Training Verifiers to Solve Math Word Problems." arXiv:2110.14168.

Deng, C., Y. Zhao, X. Tang, M. Gerstein, and A. Cohan. 2024. "Investigating Data Contamination in Modern Benchmarks for Large Language Models." Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Mexico City, Mexico. Association for Computational Linguistics.

Dodge, J., M. Sap, A. Marasović, W. Agnew, G. Ilharco, D. Groeneveld, M. Mitchell, and M. Gardner. 2021. "Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus." Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic.

Hendrycks, D., C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. 2021a. "Measuring Mathematical Problem Solving with the Math Dataset." *NeurIPS*.

Hendrycks, D., C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. 2021b. "Measuring Mathematical Problem Solving with the MATH Dataset." Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).

Hidayat, N. S., M. D. A. Kautsar, A. F. Wicaksono, and F. Koto. 2025. "Simulating Training Data Leakage in Multiple-Choice Benchmarks for Llm Evaluation." Proceedings of the 5th Workshop on Evaluation and Comparison of NLP Systems Proceedings of the 5th Workshop on Evaluation and Comparison of NLP Systems, 21–39. https://doi.org/10.18653/v1/2025.eval4nlp-1.3.

Kojima, T., S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. 2022. "Large Language Models are Zero-Shot Reasoners." Proceedings of the 36th International Conference on Neural Information Processing Systems, Red Hook, NY, USA. Curran Associates Inc.

Li, Q., L. Cui, X. Zhao, L. Kong, and W. Bi. 2024. "GSM-Plus: A Comprehensive Benchmark for Evaluating the Robustness of LLMs as Mathematical Problem Solvers." Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Bangkok, Thailand. Association for Computational Linguistics.

Liu, H., Z. Zheng, Y. Qiao, H. Duan, Z. Fei, F. Zhou, W. Zhang, S. Zhang, D. Lin, and K. Chen. 2024. "MathBench: Evaluating the Theory and Application Proficiency of LLMs with a Hierarchical Mathematics Benchmark." In *Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics*, edited by Ku Lun-Wei, Martins Andre and Srikumar Vivek, 6884--6915. Bangkok, Thailand: Association for Computational Linguistics.

Lu, P., H. Bansal, T. Xia, J. Liu, C. Li, H. Hajishirzi, H. Cheng, K.-W. Chang, M. Galley, and J. Gao. 2024. "Mathvista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts." The Twelfth International Conference on Learning Representations, Vienna, Austria.

Mao, Y., Y. Kim, and Y. Zhou. 2024. "CHAMP: A Competition-Level Dataset for Fine-Grained Analyses of LLMs' Mathematical Reasoning Capabilities." In *Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics*, edited by Ku Lun-Wei, Martins Andre and Srikumar Vivek, 13256–13274. Bangkok, Thailand.

McCoy, R. T., E. Pavlick, and T. Linzen. 2019. "Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy.

Min, S., X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer. 2022. "Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?" Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates.

Ministry of Education. 2024. Republic of Korea, Announcement on the 2028 College Scholastic Ability Test Structure and Scoring System, Accessed December 10, 2025, Press release, in Korean.

OpenAI. 2023. Gpt-4 Technical Report, arXiv preprint arXiv:2303.08774.

Park, C., J. Kim, J. Lee, S. Bae, J. Choo, and K. M. Yoo. 2025. "Cross-Lingual Collapse: How Language-Centric Foundation Models Shape Reasoning in Large Language Models." arXiv:2506.05850.

Patel, B., S. Chakraborty, W. A. Suttle, M. Wang, A. S. Bedi, and D. Manocha. 2024. "Aime: Ai System Optimization via Multiple Llm Evaluators." arXiv:2410.03131.

Pyeon, G., I. Heo, J. Jung, T. Hwang, H. Namgoong, H. Seo, Y. Han, E. Kim, H. Kang, and S. Jung. 2025. "Evaluating Large Language Models on the 2026 Korean Csat Mathematics Exam: Measuring Mathematical Ability in a Zero-Data-Leakage Setting." arXiv:2511.18649. https://arxiv.org/abs/2511.18649.

Samuel, V., Y. Zhou, and H. P. Zou. 2025. "Towards Data Contamination Detection for Modern Large Language Models: Limitations, Inconsistencies, and Oracle Challenges." Proceedings of the 31st International Conference on Computational Linguistics, 5058–5070. Abu Dhabi, United Arab Emirates.

Snell, C. V., J. Lee, K. Xu, and A. Kumar. 2025. "Scaling LLM Test-Time Compute Optimally Can Be More Effective Than Scaling Parameters for Reasoning." The Thirteenth International Conference on Learning Representations, Singapore EXPO.

Wang, K., J. Pan, W. Shi, Z. Lu, H. Ren, A. Zhou, M. Zhan, and H. Li. 2024. "Measuring Multimodal Mathematical Reasoning with Math-Vision Dataset." Advances in Neural Information Processing Systems, Curran Associates, Inc., 95095–95169. https://doi.org/10.52202/079017-3014.

Wang, X., J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou. 2023. "Self-Consistency Improves Chain of Thought Reasoning in Language Models." arXiv:2203.11171.

Xu, R., Z. Wang, R.-Z. Fan, and P. Liu. 2024. "Benchmarking Benchmark Leakage in Large Language Models." arXiv:2404.18824.

Yao, J., R. Cheng, and K. C. Tan. 2025. "Var-Math: Probing True Mathematical Reasoning in Llms via Symbolic Multi-Instance Benchmarks." *arXiv Preprint arXiv: 2507.12885*.

Zhou, Z., Z. Zhu, X. Li, M. Galkin, X. Feng, S. Koyejo, J. Tang, and B. Han. 2025. "Landscape of Thoughts: Visualizing the Reasoning Process of Large Language Models." arXiv:2503.22165.