ScholarBench: A Bilingual Benchmark for Abstraction, Comprehension, and Reasoning Evaluation in Academic Contexts

Anonymous ACL submission

Abstract

Prior benchmarks for evaluating the domain-specific knowledge of large language models (LLMs) lack the scalability to handle complex academic tasks. To address this, we introduce `ScholarBench`, a benchmark centered on deep expert knowledge and complex academic problem-solving, which evaluates the academic reasoning ability of LLMs and is constructed through a three-step process. `ScholarBench` targets more specialized and logically complex contexts derived from academic literature, encompassing five distinct problem types. Unlike prior benchmarks, `ScholarBench` evaluates the abstraction, comprehension, and reasoning capabilities of LLMs across eight distinct research domains. To ensure high-quality evaluation data, we define category-specific example attributes and design questions that are aligned with the characteristic research methodologies and discourse structures of each domain. Additionally, this benchmark operates as an English-Korean bilingual dataset, facilitating simultaneous evaluation for linguistic capabilities of LLMs in both languages. The benchmark comprises 5,031 examples in Korean and 5,309 in English, with even state-of-the-art models like o3-mini achieving an average evaluation score of only 0.543, demonstrating the challenging nature of this benchmark.

1 Introduction

The emergence and application of large language models (LLMs) (OpenAI, 2024; Touvron et al., 2023; Team, 2024a) has significantly advanced performance across diverse natural language processing tasks and effectively addressed both conventional and complex challenges in the field. LLMs are trained on multilingual (Tang et al., 2024; Wang et al., 2025), general-purpose (Zhang et al., 2024b), and web-based data, enabling them to generalize across languages (Wu et al., 2025), handle interactions and code-switching (Huzaifah et al., 2024), and flexibly respond to queries across a wide range of domains (Wan et al., 2024).

Benchmarking initiatives are underway to evaluate LLM capabilities in language comprehension, generation, and reasoning, categorized by task types, domains, and languages (Clark et al., 2018; Wang et al., 2019; et al., 2021; Zheng et al., 2023; Hendrycks et al., 2021). These benchmarks critically enable objective comparisons among LLMs and identify areas for improvement. While existing benchmarks predominantly target general-purpose domains, such as MT-Bench (Zheng et al., 2023), MMLU (Hendrycks et al., 2021), C-EVAL (Huang et al., 2024), and Xiezhi (Gu et al., 2024), specialized domains necessitate distinct problem-solving approaches and domain-specific knowledge, underscoring the increasing demand for benchmarks tailored to these fields.

General-domain benchmarks predominantly utilize standardized examination questions, which typically focus on STEM disciplines. Consequently, they offer limited insight into specialized knowledge domains and inadequately capture LLMs' domain-specific problem-solving capabilities. To address these gaps, recent benchmark studies have focused on detailed evaluations of LLM performance in specialized tasks requiring deep expert knowledge and practical application.

For example, MultiMedQA (Singhal et al., 2023) and FinBen (Xie et al., 2024) specifically assess LLM within medical and financial domains, demonstrating their practical applicability in tasks such as information extraction and risk management. Similarly, ChemLLMBench (Guo et al., 2023) and DataSciBench (Zhang et al., 2024a) provide comprehensive assessments within chemistry and data science domains, respectively, elucidating both strengths and limitations of current LLMs. Nonetheless, existing domain-specific benchmarks remain inherently constrained in their generalizability across disciplines, providing insufficient support

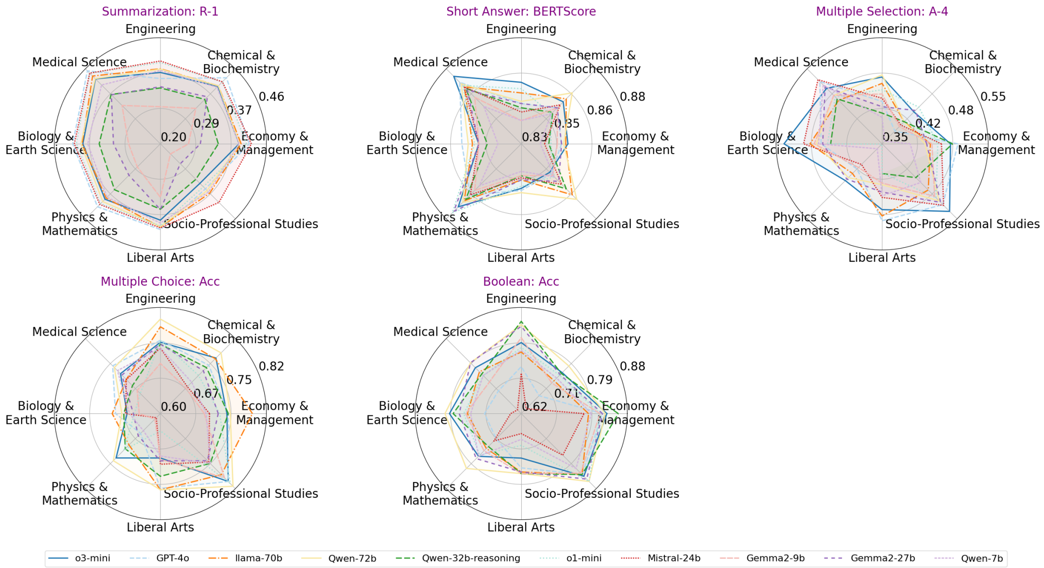

Figure 1: Model performance across categories for leading open- and closed-source LLMs on ScholarBench. Each column represents a task-specific evaluation metric. Main task-level results are reported in Table 2, while detailed performance analysis by category is provided in Appendix E.2.

for evaluating interdisciplinary and complex academic tasks.

In this paper, we introduce a new benchmark called ScholarBench, designed to evaluate the problem-solving capabilities of LLMs in academic domains, with a focus on their parametric knowledge, and analyze their performance in advanced reasoning tasks within scholarly environments. Academic-domain LLM benchmarks help enhance the practical applicability of LLMs in academic research, education, and specialized fields. ScholarBench offers three key features:

- **Domain and Attribute.** To systematically evaluate performance across interdisciplinary academic domains, we define four primary domains *Natural Sciences*, *Applied Sciences*, *Social Sciences*, and *Humanities* and further delineate a total of eight categories. In addition, we propose fine-grained attribute categories to capture the diversity of question types and enable nuanced assessments of generalization performance within and across academic domains.
- **Task and Evaluation.** ScholarBench leverages a diverse set of question types to concurrently assess multiple competencies of LLMs. Moving beyond simple item-level evaluation, the benchmark incorporates multidimensional assessment criteria including abstraction, reasoning, and comprehension to rigorously evaluate the practical problem-solving abilities and real-world applicability of LLMs. This approach offers a robust framework for holistic assessment of academic intelligence in LLMs.
- **Bilingual Ability.** We construct a bilingual (English–Korean) benchmark to enable precise evaluation of LLMs' cross-lingual knowledge transfer and multilingual understanding. By designing questions and examples that facilitate both direct and indirect comparisons across linguistic and cultural contexts, we enable in-depth analysis of model generalization and performance disparities in the presence of linguistic diversity.

Overall, ScholarBench aims to assess LLMs' performance in academic domains, analyze question type-specific results from multiple perspectives to identify model strengths and weaknesses, and provide insights for improving LLMs in the academic domains. The contributions of this paper are as follows:

- Performance comparison of LLMs across eight academic categories, revealing domain-specific strengths and weaknesses
- Fine-grained evaluation of LLM capabilities across 63 English and 65 Korean academic attributes
- Construction of parallel and non-parallel bilingual datasets to assess linguistic and terminological understanding in academic texts
- A systematic benchmark construction method grounded in complex academic data

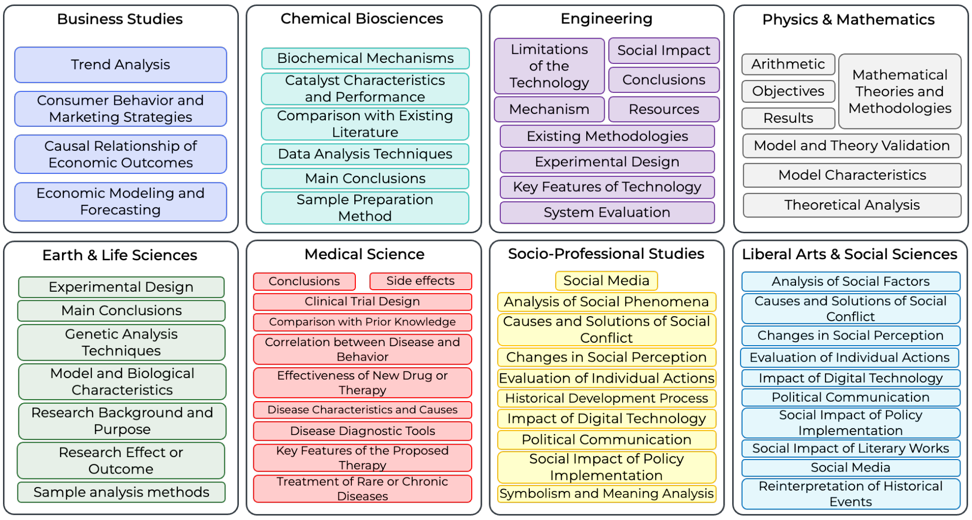

Figure 2: Taxonomy of academic categories and question attributes for English dataset.

2 Open Scholar Benchmark

We propose `ScholarBench`, a benchmark that enables performance evaluation and analysis of LLMs in academic domains, and introduce its construction methodology. To define the multidimensional evaluation of LLMs, we present three key concepts: **Abstraction (C1)** assesses the ability of LLMs to identify key information and summarize academic literature while maintaining the context. **Reasoning ability (C2)** evaluates the LLM's capacity for logical reasoning, demonstrated by its ability to infer answers based on parametric knowledge. The reasoning questions consist of short-answer, multiple-choice, multiple-selection, and true/false types, and are solved by the LLMs in a closed-book setting. **Comprehension (C3)** is evaluated in an open-book setting, where the LLM must identify and extract key information from a given academic paragraph to determine its ability to solve academic problems accurately.

2.1 Categories and Question Attributes

As shown in Figure 2, we categorize eight academic categories and 63 English question attributes (along with 65 Korean attributes) into four academic domains: *natural sciences*, *applied sciences*, *social sciences*, and *humanities*. English papers are selected from journals with the highest H5-index according to Google Scholar [1], while Korean papers are selected from journals with the highest five-year citation index based on the KCI (Korea Citation Index) [2]. By integrating these sources, a total of eight categories are derived. The attribute descriptions for each academic category are as follows:

Business Studies Focuses on key economic trends, consumer behaviors, and marketing strategies, leveraging economic modeling to forecast market dynamics.

Chemical Biosciences Covers chemical reaction mechanisms and biochemical processes through rigorous analysis and interpretation of data.

Engineering Centers on engineering methodologies and technological innovations by assessing performance, efficiency, and societal impacts.

Physics & Mathematics Focuses on theoretical modeling, mathematical reasoning, and rigorous validation of physical systems through analysis.

Earth & Life Sciences Combines theoretical biological modeling with analyses to elucidate quantitative relationships in biological phenomena.

Medical Science Addresses healthcare challenges by integrating clinical trial design, diagnostic evaluation, and drug efficacy analysis with evidence-based medical reasoning.

Socio-Professional Studies Covers topics at the intersection of society and professional practice, such as the effects of arts and physical education, psychological factors, and interpretation of cultural artifacts.

Liberal Arts & Social Sciences Explores causes and transformations of social phenomena, policy impact, and the role of digital media.

[1] https://scholar.google.com
[2] https://www.kci.go.kr/kciportal/main.kci

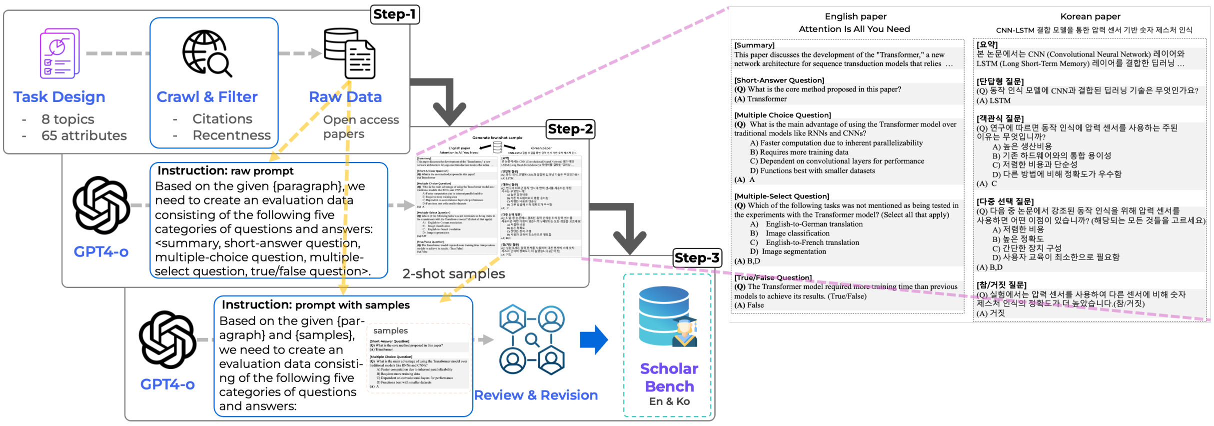

Figure 3: Data construction pipeline. For a step-by-step example of data construction, see Appendix F.3.

2.2 Question Design

Academic-domain benchmarks aim to evaluate whether LLMs can apply their learned knowledge to reason about and solve domain-specific problems. The proposed benchmark includes five types of questions: summarization, short answer, multiple choice, multiple selection, and true/false. These question types are designed not only to assess LLMs' abstraction, reasoning, and comprehension abilities, but also to ensure diversity, generality, and ease of evaluation.

To evaluate whether LLMs can adequately perform academic-domain tasks using only their pre-trained knowledge, we conduct experiments under both closed-book and open-book settings. In the closed-book setting, no external information is provided except for summarization tasks requiring the model to rely solely on its internal knowledge. To ensure evaluation diversity and analyze the impact of information access on performance, we additionally include the open-book setting.

Among the five question types, summarization tasks require the model to condense a given paragraph into its key content. Short-answer questions involve understanding the query and generating a specific, correct response (Rajpurkar et al., 2018). Multiple-choice questions assess the model's decision-making based on provided options. Multiple-selection questions demand higher discriminative ability, as the model must identify all correct answers. Finally, true/false questions, commonly used in benchmarks such as BoolQ (Clark et al., 2019) and TruthfulQA (Lin et al., 2022), evaluate factual reasoning by requiring binary judgments based on a given question.

These question types are comparable to those used in general-purpose benchmarks, allowing us to assess LLMs' problem-solving capabilities in specialized domains and to estimate the difficulty of the benchmark.

2.3 Data Construction Process

To ensure high-quality question-answer pairs aligned with each academic domain, we conduct a three-step data construction pipeline. Figure 3 illustrates the overall pipeline, which consists of the following steps. In step 1, tasks are designed based on eight academic categories and 63 English and 65 Korean attributes, and source materials are collected by crawling and filtering publicly available academic papers according to recency and citation count. In step 2, five types of evaluation questions, summarization, short-answer, multiple-choice, multiple-selection, and true/false, are generated from the source paragraphs using GPT-4o. In step 3, to ensure varying question types and difficulties across turns, two questions from the initial outputs are randomly sampled to iteratively refine the prompting strategy. Finally, the generated questions undergo expert review and revision to improve quality before being finalized in `ScholarBench`. Motivated by our observation that question diversity and difficulty vary depending on the source documents, we performed a comprehensive evaluation integrating passages, generated questions, and corresponding answers. Detailed prompts, data samples, and diversity-related statistics used for question generation are available in Appendix F.

Academic Papers Collection. The criteria for selecting academic papers are as follows. For English papers, we collect a total of 200 articles from 80 Open Access journals that fall under our academic categories and are ranked in the top 10 of the

Language	Evaluation A	Evaluation B	Assessment A	Assessment B	Kappa Coefficient
English	4.50	4.13	4.11	4.08	0.614
Korean	4.19	3.81	3.97	3.91	0.706

Table 1: Human evaluation and data difficulty assessment using a 1–5 scale and Kappa coefficient.

H5-index according to Google Scholar. For Korean papers, we collect 1,916 articles from 138 journals selected using the same criteria, based on the KCI.

Review and Revision. The data verification process is conducted by 15 reviewers, 11 for English and 4 for Korean. All reviewers are university students, and they follow the guidelines. The review process consists of three stages: **Paragraph evaluation**, where reviewers assess whether the selected paragraph appropriately reflects the core concept of the paper; **Answer verification**, where the correctness of the provided answer for each item is checked; **Human evaluation**, which is conducted only if the answer is correct, to assess the overall quality of the example. The main focus points during the review process are as follows:

- All five question types are derived from a single paragraph, and the evidence required to answer each question must be explicitly present in the paragraph.
- Questions should be designed to avoid the use of referential noun phrases (e.g., demonstratives or determiners) that do not directly point to specific content in the paragraph. Due to certain attributes such as research objectives and conclusions (see Table 2), expressions like "this study" may be automatically generated. During the review process, we ensure such referential phrases are excluded from the question prompt.
- In `ScholarBench`, short-answer questions are designed to have a single correct answer. Ambiguous cases that could allow multiple valid answers are excluded, enabling evaluation of whether the LLM can generate a precise response. This design is important for assessing reasoning ability with parametric knowledge.

To measure inter-annotator agreement during the benchmark validation process, we report the average scores given by two annotators across two evaluation dimensions—*Evaluation* and *Assessment*—as shown in Table 1. For this, we randomly sample 1% of the examples in each language and have the annotators evaluate them independently. The resulting Kappa coefficients (McHugh, 2012) indicate a high level of agreement: 0.614 for English and 0.706 for Korean, demonstrating that `ScholarBench` provides consistent and reliable annotations.

3 Experiments

In this chapter, we evaluate the performance of publicly available English-Korean bilingual models to determine the validity of the `ScholarBench`.

3.1 Target Models

- **Three API-based models** include o3-mini, o1-mini, GPT-4o (OpenAI et al., 2024)
- **Seven open-source model families** include Llama3.3 (Grattafiori et al., 2024), Mistral (Jiang et al., 2024), Qwen2.5 (Team, 2024b), Gemma2 (Team, 2024a), Bllossom, Exaone (Research et al., 2024), Trillion (Han et al., 2025).

A detailed description of each model can be found in the Appendix C.1.

3.2 Evaluation Metrics

We use ROUGE (Lin, 2004) to evaluate summarization tasks, and accuracy to assess performance on multiple-choice, multiple-selection, and boolean questions. For short-answer questions, we use BERTScore (Zhang et al., 2020) to assess semantic similarity. Accuracy is used to evaluate multiple-selection, multiple-choice, and boolean questions. In multiple selection evaluation, performance is assessed by varying the target number of correct answers (2-4). For a target of 2, an accuracy point is given upon selecting any two correct options, irrespective of the total options presented (e.g., 2 out of 4 total). The evaluation metrics used in this paper are shown in Appendix D.

3.3 Overall Performance

Table 2 presents the evaluation results for API-based and open-source models on `ScholarBench`. The evaluation comprises five tasks: summarization, short-answer, multiple-selection, multiple-choice question (MCQ), and boolean question, showing representative metrics for each task. This setup allows for identifying performance variations across tasks and specialization tendencies of models. The table includes the average across all metric results to provide an overview of the models' overall performance.

Model	Summarization			Short Answer	Multiple Selection			MCQ	Boolean	Avg
	R-1	R-2	R-L	BERTScore	A-2	A-3	A-4			
o3-mini	0.392	0.130	0.320	**0.860**	**0.666**	0.518	**0.482**	0.728	0.786	**0.543**
o1-mini	0.409	0.151	0.348	0.857	0.567	0.469	0.436	0.702	0.771	0.523
GPT-4o	**0.417**	0.151	0.356	0.857	0.586	0.497	0.465	0.736	0.743	0.534
Qwen-72b	0.396	0.151	0.345	0.859	0.522	0.468	0.452	**0.755**	**0.811**	0.529
Llama-70b	0.397	0.146	0.341	0.858	0.578	0.485	0.455	0.746	0.769	0.530
Bllossom-70b	0.349	0.129	0.299	0.848	0.650	0.463	0.440	0.683	0.724	0.509
Qwen-32b-reasoning	0.350	0.116	0.303	0.853	0.539	**0.640**	0.423	0.721	0.793	0.527
Exaone-32b	0.321	0.094	0.267	0.852	0.590	0.466	0.431	0.713	0.751	0.498
Exaone-32b-reasoning	0.316	0.092	0.267	0.847	0.492	0.397	0.359	0.686	0.667	0.458
Gemma2-27b	0.329	0.117	0.283	0.856	0.577	0.479	0.453	0.707	0.796	0.511
Mistral-24b	0.414	**0.159**	**0.359**	0.853	0.584	0.488	0.458	0.694	0.696	0.523
Gemma2-9b	0.294	0.096	0.248	0.851	0.556	0.520	0.444	0.684	0.774	0.496
Exaone-8b	0.317	0.092	0.265	0.846	0.577	0.417	0.386	0.692	0.756	0.483
Mistral-8b	0.402	0.151	0.350	0.842	0.504	0.374	0.355	0.656	0.582	0.468
Llama-8b	0.381	0.136	0.327	0.845	0.501	0.419	0.395	0.658	0.556	0.469
Bllossom-8b	0.346	0.129	0.301	0.844	0.537	0.419	0.383	0.633	0.581	0.464
Qwen-7b	0.388	0.144	0.338	0.847	0.559	0.452	0.423	0.699	0.756	0.512
Trilion-7b	0.264	0.103	0.232	0.850	0.556	0.384	0.348	0.647	0.654	0.449

Table 2: Overall evaluation results of `ScholarBench` on all curated prompts under closed-book settings. For performance metrics, *R* denotes ROUGE scores (ROUGE-1, ROUGE-2, and ROUGE-L), and *A* indicates multiple selection settings, where the appended number signifies the count of correct answers. Only the summarization task is evaluated using paragraph-level input; all subsequent tasks are evaluated without paragraph-level input. The *Avg* column reports the average over all listed metrics. Bold and underline indicate the first and second ranks per metric.

Summarization. In summarization tasks, GPT-4o and Mistral-24b demonstrate strong overall performance. Specifically, GPT-4o achieves the top score in ROUGE-1, while Mistral-24b leads in both ROUGE-2 and ROUGE-L. We hypothesize that while GPT-4o is effective in capturing simple information, Mistral-24b excels in comprehensively processing longer context. Furthermore, among the small LLMs, Mistral-8b exhibits performance comparable to o1-mini, suggesting that even a small model can achieve competitive results depending on the architecture design and training strategy.

Short Answer. The short-answer task reveals only marginal performance differences among top models, and most of them show a high BERTScore of 0.84 or higher. This appears to be because semantic similarity-based evaluation is more sensitive to the naturalness and consistency of expressions than to complex inferences. In addition, small LLMs show relatively score in this task, suggesting that the models can be adjusted to trade-off balance between model efficiency and accuracy.

Multiple Selection. This task exhibits a tendency for model performance to vary as the number of options increases. Specifically, o3-mini achieves the best performance in the A-2 setting, while Qwen-32b-reasoning excels in A-3, indicating that the optimal model differs depending on the number of options. A general performance degradation is observed in the A-4 setting, implying that as the number of correct answers increases, the difficulty in achieving partial correctness or the challenge of avoiding distractors increases. The strong performance of Qwen-32b-reasoning in A-3 suggests that reasoning-specific tuning can be effective for addressing specific types of complex multiple selection questions.

MCQ. In the MCQ, Qwen-72b, Llama-70b, GPT-4o achieve top scores. These models increase accuracy in answer selection, due to their ability to effectively discriminate between candidate options based on semantic similarity and identify distractor. On the other hand, models under 8b parameters exhibit decreased performance on this task. These results reflect a correlation between model size and problem-solving ability in the MCQ task, which requires logical comparitive judgement, reasoning, long context dependecy.

Boolean. On this task, Qwen-72b demonstrates top performance, achieving a score of 0.811, which highlights its strong capability for binary classification. Qwen-32b-reasoning and Gemma2-27b also achieve top-tier accuracy, there is a significant contribution of explicit reasoning-oriented fine-tuning to performance on this task. The significantly decreased accuracy of small LLMs suggests that model parameter size considerably impacts performance. However, among small LLMs,

Model	R-1	R-2	R-L	BERTScore
Paragraph-Only Prompting				
o1-mini	0.409	0.151	0.348	0.885
Qwen-72b	0.396	0.151	0.345	0.879
Mistral-8b	0.402	0.151	0.350	0.883
Prompting with Paragraph and Category				
o1-mini	0.413	0.152	0.349	0.886
Qwen-72b	0.412	0.160	0.357	0.882
Mistral-8b	0.408	0.155	0.355	0.884

Table 3: Evaluating abstraction ability using summarization under different prompting settings.

Gemma2-9b, Exaone-8b, and Qwen-7b maintain scores above 0.75, indicating that problem-solving ability can be improved through learning strategy even with a small parameter size.

Average Performance. The average performance across all metrics (the *Avg* column) reveals well-balanced capabilities encompassing generation, understanding, classification, and reasoning. As shown in Table 2, o3-mini achieves the highest average score (0.543), followed by models such as GPT-4o, o1-mini, Qwen-72b, Llama-70b, Qwen-32b-reasoning, Mistral-24b, and Gemma2-9b, which also demonstrate top-tier performance with scores above 0.52. LLMs exhibit stable and consistent performance, even on academic domain-specific tasks. Meanwhile, some medium and small-sized models also show significant average scores. Accordingly, this highlights that `ScholarBench` was designed not for a biased comparison centered on a single task, but to evaluate various problem types and difficulty levels.

4 Analysis

In this section, we evaluate and analyze the abstraction, comprehension, reasoning, and bilingual abilities. For more detailed experimental results, please refer to Appendix G.

4.1 LLM Capability Analysis

Abstraction For abstraction (summarization) tasks, we compare the performance variations in selected models depending on whether a category was added to the prompt. In Table 3, the models demonstrate good summarization performance even on academic domain passages. Furthermore, when prompting with a domain-specific category, a slight but noticeable performance gain is observed.

Comprehensibility `ScholarBench` includes paragraphs supporting each question for the evaluation of comprehension ability. This feature is aimed at measuring the ability to extract key information from the provided passage and generate correct responses. Table 5 presents the performance results when prompting with both the question and the paragraph. Overall performance improves for all models, indicating that the inclusion of paragraphs contributes to improving the models' problem-solving capabilities, and suggesting that the models are able to understand the paragraph well and infer the correct answer. For LLMs, not only simple reasoning ability but also the ability to interpret and utilize the given context is crucial. Thus, a benchmark designed to evaluate comprehension ability is valuable.

Reasoning Table 4 presents an example of a reasoning question, answer, and the supporting paragraph. Based solely on simple information explicitly listed in the paragraph, deriving the correct answer to the question *What factors contribute to the educational challenges faced by Dominican girls of Haitian descent?* is difficult. Accurate selection for this question requires contextual inference and integration of the impact of factors such as economic difficulties and gender roles and cultural norms on education. This question type evaluates models' ability to relationally connect detailed information within the paragraph and their understanding of the background and context. Additionally, *What impact can the lack of documentation have on youth education?* requires causal inference between documentation issues and education. Although this information is not explicitly written in the paragraph, it necessitates contextual reasoning such as *lack of documentation* $\rightarrow$ *inability to take exams* $\rightarrow$ *exclusion from education*. The boolean question presents the core argument of the paragraph in reverse, and measures judgment for logical negation.

Bilingual Ability `ScholarBench` is partially constructed as parallel data in English and Korean for 18.7% of all questions. Figure 4 shows the results of cross-lingual evaluation based on this parallel data. This experiment analyzes the cross-lingual generalization capabilities of multilingual LLMs, reflecting complex factors such as semantic consistency across languages, reasoning coherence, and potential information loss during translation. Most models exhibit a Ko > En trend, while some global models show the opposite pattern (En > Ko).

Passage (excerpt)	Question	Answer	Reasoning
Dominican youth of Haitian descent face significant barriers to education due to lack of documentation, societal prejudice, and economic hardship. Many children receive Dominican birth certificates, only to have them cancelled arbitrarily. High dropout rates are observed, especially for females, who may encounter additional challenges from traditional gender roles. Despite these adversities, some students continue their education.	What specific challenge related to documentation impacts school participation for Dominican females of Haitian descent?	Lack of documentation	**Yes**
	What factors contribute to the educational challenges faced by Dominican girls of Haitian descent? (Select all that apply)	a) Cultural attitudes like machismo, b) Economic hardship	**Yes**
	What impact can the lack of documentation have on youth education?	a) Denial of access to national exams	**Yes**
	The absence of documentation does not affect the educational success of Dominican females of Haitian descent. (True/False)	False	**Yes**

Table 4: Example of a reasoning-based question set. The passage is excerpted to retain core information about cultural capital, systemic oppression, and education resilience, allowing evaluation of LLMs' contextual reasoning.

Model	Summarization			Short Answer	Multiple Selection			MCQ	Boolean	Avg
	R-1	R-2	R-L	BERTScore	A-2	A-3	A-4			
o3-mini	0.392	0.130	0.320	0.898	0.783	0.611	0.591	0.886	0.875	0.609
o1-mini	0.409	0.151	0.348	0.898	**0.795**	0.656	0.632	0.883	0.872	0.627
GPT-4o	**0.417**	**0.151**	0.356	0.897	0.793	**0.667**	**0.643**	0.886	0.854	**0.629**
Qwen-72b	0.396	0.151	0.345	0.902	0.746	0.650	0.628	**0.900**	**0.896**	0.624
Llama-70b	0.397	0.146	0.341	**0.903**	0.751	0.620	0.598	0.875	0.873	0.612
Mistral-24b	0.414	0.159	**0.359**	0.901	0.756	0.580	0.560	0.865	0.827	0.602

Table 5: Evaluation of comprehension ability in `ScholarBench` under open-book settings: effect of paragraph prompting on task performance. Summarization results correspond to ROUGE-1 scores as reported in Table 2.

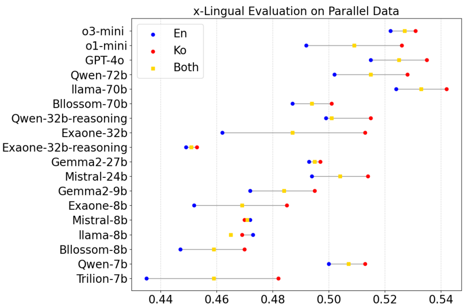

Figure 4: Model-wise performance on parallel data across En, Ko, and Both language settings.

5 Conclusion

We propose `ScholarBench`, a new benchmark dataset for evaluating the capabilities of LLMs in academic domains, and analyze the experimental results. While previous benchmark's studies focus on specific tasks or general domains, `ScholarBench` is designed to include various abilities and question types required in academic fields, such as abstraction, comprehension, reasoning, and bilingual ability. By defining eight academic domain categories and 63 corresponding attributes, we enhance the diversity of problems and provide a guide for more detailed and systematic evaluation of problem-solving abilities.

As analysis results utilizing `ScholarBench`, we confirmed the significant variations that LLMs show depending on the task types, academic fields, and model characteristics. For each task, we analyze model performance based on the problem characteristics such as reliance on structured information, context dependency, expression diversity, causal reasoning, relational understanding, and logical negation. This enables the evaluation of both *topic generalization ability* and *topic-specific reasoning ability* to handle the complex knowledge and reasoning within specific academic fields.

For future work, we envision further expanding `ScholarBench` to include more academic domains and new task types, such as RAG and multimodal. Additionally, we plan to specifically analyze model error types to identify the root causes of performance degradation and utilize `ScholarBench` to develop LLM learning and tuning strategies specialized for academic domains. Finally, we expect that `ScholarBench` will contribute to the development of more capable and reliable AI systems for the academic community.

Limitations

While ScholarBench provides a comprehensive evaluation of academic problem-solving across various fields, its scope is largely confined to English-language academic settings and structured tasks. This means it may not fully capture the reasoning challenges unique to multilingual or less formal educational contexts, thus limiting its generalizability beyond standardized academic domains.

Our constructed benchmark dataset, `ScholarBench`, was designed to evaluate academic problem-solving abilities based on scholarly articles. However, a primary limitation is that the data sources were confined solely to paragraph text from these articles. Real scholarly papers, beyond plain text, contain information across diverse modalities such as figures, tables, algorithms, and diagrams, which are often essential for understanding the research content. Evaluating models on academic paper data typically requires the ability to integrate and reason over information from various modalities. Since the current dataset does not explicitly incorporate or utilize these multimodal elements, this poses a direct limitation in accurately measuring the academic problem-solving capabilities of models that leverage multimodal information. Future work is needed to overcome this limitation by constructing an extended dataset that includes and utilizes diverse modalities.

To effectively evaluate the model's comprehensibility of specific text passages, `ScholarBench` is structured such that paragraphs related to each question are provided as prompts alongside the question. This setup is valuable for assessing how accurately a model interprets and responds based on given contextual information. However, a recent major trend in NLP, such as retrieval-augmented generation (RAG) (Lewis et al., 2020; Gao et al., 2024; Li et al., 2024), emphasizes the overall pipeline performance, which includes finding relevant external knowledge and generating responses based on it. The current dataset structure presents a limitation in that it cannot directly evaluate the performance of the Retrieval phase necessary before problem-solving in a RAG setup. As future work, we plan to leverage the currently developed questions but extend the evaluation process to require models to first retrieve and extract the necessary information from the original full papers or a collection of related documents.

Ethical Considerations

References

T. Tony Cai and Rong Ma. 2022. Theoretical foundations of t-sne for visualizing high-dimensional clustered data. *Preprint*, arXiv:2105.07536.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *NAACL*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *Preprint*, arXiv:1803.05457.

Sungjoon Park et al. 2021. KLUE: Korean language understanding evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Zhouhong Gu, Xiaoxuan Zhu, Haoning Ye, Lin Zhang, Jianchen Wang, Yixin Zhu, Sihang Jiang, Zhuozhi Xiong, Zihan Li, Weijie Wu, Qianyu He, Rui Xu, Wenhao Huang, Jingping Liu, Zili Wang, Shusen Wang, Weiguo Zheng, Hongwei Feng, and Yanghua Xiao. 2024. Xiezhi: An ever-updating benchmark for holistic domain knowledge evaluation. In *AAAI*, pages 18099–18107.

Taicheng Guo, Kehan Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2023. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Sungjun Han, Juyoung Suk, Suyeong An, Hyungguk Kim, Kyuseok Kim, Wonsuk Yang, Seungtaek Choi, and Jamin Shin. 2025. Trillion 7b technical report. *Preprint*, arXiv:2504.15431.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2024. C-eval: a multi-level multi-discipline chinese evaluation suite for foundation models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Muhammad Huzaifah, Weihua Zheng, Nattapol Chanpaisit, and Kui Wu. 2024. Evaluating code-switching translation with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6381–6394, Torino, Italia. ELRA and ICCL.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

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. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics*, pages 3335–3346. ELRA and ICCL.

Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2024. Mt-eval: A multi-turn capabilities evaluation benchmark for large language models. *Preprint*, arXiv:2401.16745.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024. Retrieval augmented generation or long-context LLMs? a comprehensive study and hybrid approach. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 881–893, Miami, Florida, US. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering. *Preprint*, arXiv:2203.14371.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

LG AI Research, Soyoung An, Kyunghoon Bae, Eunbi Choi, Kibong Choi, Stanley Jungkyu Choi, Seokhee Hong, Junwon Hwang, Hyojin Jeon, Gerrard Jeongwon Jo, Hyunjik Jo, Jiyeon Jung, Yountae Jung, Hyosang Kim, Joonkee Kim, Seonghwan Kim, Soyeon Kim, Sunkyoung Kim, Yireun Kim, and 14 others. 2024. Exaone 3.5: Series of large language models for real-world use cases. *Preprint*, arXiv:2412.04862.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, and 57 others. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *Preprint*, arXiv:2406.05967.

Mobashir Sadat and Cornelia Caragea. 2022. Scinli: A corpus for natural language inference on scientific text. *Preprint*, arXiv:2203.06728.

Mobashir Sadat and Cornelia Caragea. 2024. Mscinli: A diverse benchmark for scientific natural language inference. *Preprint*, arXiv:2404.08066.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Pavel Sirotkin. 2013. On search engine evaluation metrics. *Preprint*, arXiv:1302.2318.

Yu Song, Santiago Miret, and Bang Liu. 2023. Matscinlp: Evaluating scientific language models on materials science language tasks using text-to-schema modeling. *Preprint*, arXiv:2305.08264.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.

Gemma Team. 2024a. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Qwen Team. 2024b. Qwen2.5: A party of foundation models.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kuckreja, Mykola Maslych, Wafa Al Ghallabi, Mihail Mihaylov, Chao Qin, Abdelrahman M Shaker, Mike Zhang, Mahardika Krisna Ihsani, Amiel Esplana, Monil Gokani, and 50 others. 2024. All languages matter: Evaluating lmms on culturally diverse 100 languages. *Preprint*, arXiv:2411.16508.

Zhen Wan, Yating Zhang, Yexiang Wang, Fei Cheng, and Sadao Kurohashi. 2024. Reformulating domain adaptation of large language models as adapt-retrieve-revise: A case study on Chinese legal domain. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5030–5041, Bangkok, Thailand. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. *Preprint*, arXiv:1804.07461.

Jiayi Wang, Yao Lu, Maurice Weber, Max Ryabinin, David Adelani, Yihong Chen, Raphael Tang, and Pontus Stenetorp. 2025. Multilingual language model pretraining using machine-translated data. *Preprint*, arXiv:2502.13252.

Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. 2025. The semantic hub hypothesis: Language models share semantic representations across languages and modalities. In *The Thirteenth International Conference on Learning Representations*.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, Yijing Xu, Haoqiang Kang, Ziyan Kuang, Chenhan Yuan, Kailai Yang, Zheheng Luo, Tianlin Zhang, Zhiwei Liu, GUOJUN XIONG, and 15 others. 2024. Finben: An holistic financial benchmark for large language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Dan Zhang, Sining Zhoubian, Min Cai, Fengzu Li, Lekang Yang, Wei Wang, Tianjiao Dong, Ziniu Hu, Jie Tang, and Yisong Yue. 2024a. Datascibench: An LLM agent benchmark for data science.

Liang Zhang, Qin Jin, Haoyang Huang, Dongdong Zhang, and Furu Wei. 2024b. Respond in my language: Mitigating language inconsistency in response generation based on large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4177–4192, Bangkok, Thailand. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *Preprint*, arXiv:2304.06364.

A Related Work

Benchmark studies have focused on evaluating specific capabilities or domains, including domain-specific benchmarks, multilingual and cultural benchmarks, and multi-domain knowledge benchmarks.

Domain-Specific Benchmarks Domain-specific benchmarks aim to evaluate whether large language models (LLMs) can go beyond general language understanding to comprehend and apply specialized knowledge and terminology within specific fields. In the scientific domain, benchmarks such as SciNLI (Sadat and Caragea, 2022), MSciNLI (Sadat and Caragea, 2024), and Matsci-NLP (Song et al., 2023) have been introduced. In the medical domain, MedMCQA (Pal et al., 2022) has been widely adopted, and FinBen (Xie et al., 2024) is proposed for the financial domain. Unlike general-domain NLI datasets, SciNLI evaluates scientific language understanding based on research papers. Compared to other NLI datasets, a characteristic feature is the low vocabulary overlap between the premise and hypothesis, which necessitates a deeper understanding rather than reliance on surface-level lexical cues. MSciNLI extends SciNLI by covering domains such as hardware, networks, software engineering, and security, thus increasing data diversity. Matsci-NLP focuses on evaluating LLM performance in the materials science domain. The dataset consists of seven tasks covering topics such as superconductors, fuel cells, and glass. MedMCQA is a large-scale multiple-choice QA benchmark constructed from actual medical entrance exams. MedMCQA covers 2,400 medical topics and 21 medical subjects, and contains a total of more than 194,000 entrance exam questions. Each data point includes a question, correct and incorrect answer options, and evaluates 10 or more reasoning skills to assess language comprehension skills in medical subjects and topics. FinBen is designed to assess the capabilities of LLMs in financial tasks. This dataset contains 36 different data sets spanning 24 financial tasks, including information extraction, text analysis, forecasting, and risk management.

Multilingual and Culturally-Aware Benchmarks With the growing adoption of LLMs by diverse user groups, understanding cultural context and supporting low-resource languages have become increasingly important. ALM-Bench (Vayani et al., 2024) evaluates LLMs' ability to understand and reason over text and culturally grounded images across 100 languages. This benchmark is characterized by its ability to comprehensively evaluate cultural characteristics, including culturally diverse features ranging from low-resource languages to specific regional dialects. CVQA (Romero et al., 2024) presents a culturally diverse visual question answering benchmark consisting of questions and images collected from 30 countries across four continents. Each question is written in both English and the local language, allowing assessment of both multilingual and English-only models.

Multi-Domain Knowledge Benchmarks Multi-domain knowledge benchmarks evaluate the generalization capabilities of LLMs by testing them across a wide range of subjects. These benchmarks aim to assess how well LLMs perform in diverse academic and professional areas. Notable datasets include MMLU (Hendrycks et al., 2021), Xiezhi (Gu et al., 2024), and AGIEval (Zhong et al., 2023). MMLU includes problems on diverse topics across 57 domains and is composed of multiple-choice questions of varying difficulty. It evaluates the comprehensive understanding and problem-solving abilities of large language models by assessing how well they understand diverse knowledge. Xiezhi includes 13 subjects and 516 diverse fields, and consists of a total of 249,587 multiple-choice questions. To evaluate multiple-choice questions, it uses Mean Reciprocal Rank (MRR) (Sirotkin, 2013), which calculates a ranking score, as the evaluation metric. A notable feature is the continuous updating of the benchmark by automatically generating data from open academic resources and labeling it with trained models. AGIEval is a benchmark dataset that evaluates the performance of large language models based on human-centric standardized tests such as college entrance exams, law school admission tests, mathematics competitions, and other professional qualification exams. It excludes subjective questions and includes objective questions like multiple-choice and fill-in-the-blank.

Evaluating Conversational Capabilities Existing datasets for evaluating the conversational capabilities of large language models include MT-Bench (Zheng et al., 2023), MT-Eval (Kwan et al., 2024), and CLIcK (Kim et al., 2024). MT-Bench consists of 80 high-quality multi-turn questions designed to evaluate conversational flow and instruction following. It covers 8 task categories: writing, role-playing, extraction, reasoning, math, coding, knowledge I (STEM), and knowledge II (humanities/social science). Each task category comprises 10 multi-turn problems. Model outputs are obtained for these tasks, and the model responses are then evaluated based on criteria including context understanding, accuracy, consistency of reasoning steps, and whether responses meet user expectations. MT-Eval is an extended benchmark of MT-Bench that refines the evaluation of multi-turn conversational capabilities by evaluating abilities such as remembering and utilizing previously mentioned information, answering various questions within the same topic, following progressively complex instructions, and responding to questions based on previous responses.

B Data Statistics

Topic Categories	Ko	En
Business Studies	124	174
Chemical Biosciences	125	124
Engineering	125	139
Medical Science	124	111
Earth & Life Sciences	125	130
Physics & Mathematics	118	149
Socio-Professional Studies	124	146
Liberal Arts & Social Sciences	150	150
Total	1,015	1,123

Table 6: Data statistics for topic categories

Problem Type	Ko	En
Summarization	1,004	1,108
Multiple choice	1,010	1,048
Multiple selection	1,003	1,056
Short answer	1,006	1,027
Boolean	1,008	1,070
Total	5,031	5,309

Table 7: Dataset statistics for question types

Table 6 presents the number of questions for each academic category, and Table 7 provides the distribution of question types. The benchmark was designed to maintain an even distribution of the five question formats, which include summarization, short answer, multiple choice, multiple selection, and true or false, across all academic domains.

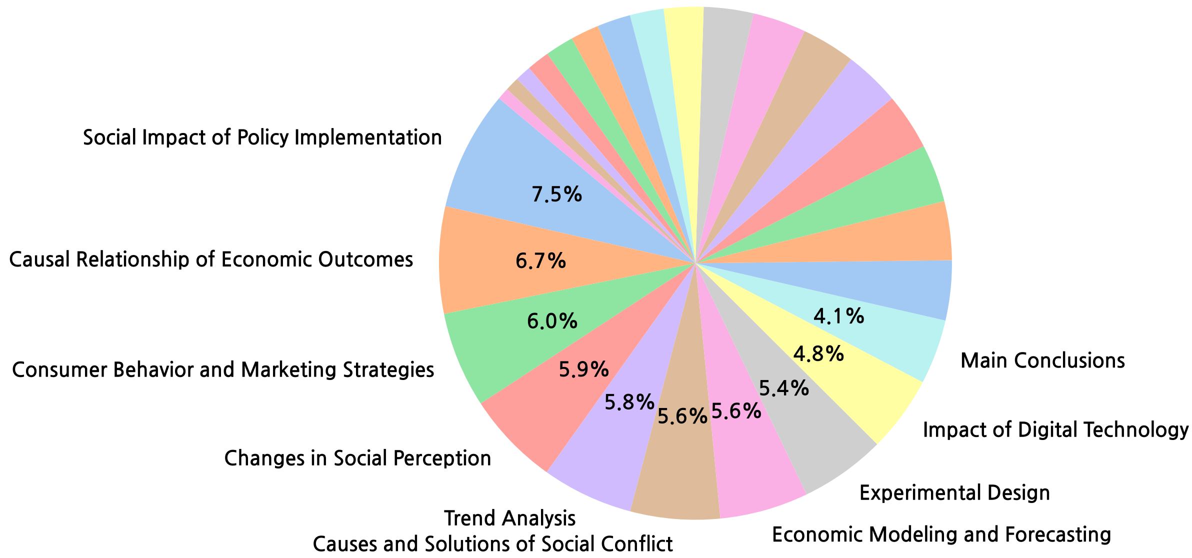

Figure 5: Frequency distribution of 27 attributes commonly shared across all question types. The balanced distribution without overconcentration on specific attributes suggests that the benchmark enables fair model evaluation across a diverse range of attributes.

B.1 Domain Diversity

Figure 5 illustrates the distribution of question attributes that are commonly shared across all task types, as defined in Section 2.1. A statistical analysis of their frequency reveals a mean of 177.7, a standard deviation of 89.3, and a coefficient of variation of 0.50. These results suggest that the attributes are evenly distributed, without excessive concentration in specific task types. Such a balanced distribution indicates that `ScholarBench` enables fair and comprehensive model evaluation across a diverse range of attributes.

B.2 Query Length Distribution

We show query word length distributions for each attribute of eight domains in Figure 6 and 7. Queries contain 10-20 words, though some instances exceed 30 words depending on the academic domain. Even when domains share the same attribute, query distributions are similar. However, variations in query length are also observed across different attributes. Specific attributes such as *Consumer Behavior and Marketing Strategies*, *Key Features of* *, *Existing Methodologies*, and *Disease Diagnostic Tools* tend to have shorter queries. In contrast, attributes such as *Research related*, *Results of* *, and *Problem of* * show a wider range of lengths. This is considered a reflection of the unique characteristics and research topics pertinent to each academic domain, as captured by the query attributes.

The Korean dataset in Figure 7 also shows a similar pattern.

B.3 Paragraph and Summary Length Distribution

We show word length distributions for paragraphs and summaries for each domain in Figure 8. Overall, paragraph lengths vary across academic domains and exhibit a wide distribution range. For instance, paragraphs within the *Business Studies* and *Earth & Life Sciences* domains show a relatively longer median and a wider interquartile range (IQR, representing the box length), indicating significant variance in paragraph length. In contrast, fields such as *Chemical Biosciences* and *Medical Science* tend to have comparatively shorter paragraphs with a more condensed distribution. It can also be observed that each domain includes very long paragraphs, resulting in extended whiskers in the plots.

In contrast to paragraph lengths, summary lengths demonstrate a consistent and uniform distribution across all academic domains. Most summaries contain between 50 and 80 words, with the median concentrating around 60-65 words in ev-

Model	Model Name
llama-70b	meta-llamaLlama-3.3-70B-Instruct
llama-8b	meta-llamaLlama-3.1-8B-Instruct
Mistral-24b	mistralaiMistral-Small-24B-Instruct-2501
Mistral-8b	mistralaiMinistral-8B-Instruct-2410
Qwen-72b	QwenQwen2.5-72B-Instruct
Qwen-32b-reasoning	QwenQwQ-32B
Qwen-7b	QwenQwen2.5-7B-Instruct
Trilion-7b	trillionlabsTrillion-7B-preview
Gemma2-27b	googlegemma-2-27b-it
Gemma2-9b	googlegemma-2-9b-it
Bllossom-70b	Bllossomllama-3-Korean-Bllossom-70B
Bllossom-8b	MLP-KTLimllama-3-Korean-Bllossom-8B
Koni-8b	KISTI-KONIKONI-Llama3.1-8B-R-Preview-20250320
Exaone-32b-reasoning	LGAI-EXAONEEXAONE-Deep-32B
Exaone-32b	LGAI-EXAONEEXAONE-3.5-32B-Instruct
Exaone-8b	LGAI-EXAONEEXAONE-3.5-7.8B-Instruct

Table 8: Model cards

ery domain. The interquartile range for summary lengths is also remarkably narrower compared to paragraph lengths, indicating effective length control during summary generation. This suggests that the summarization task encourages the generation of consistent-length summaries despite varying input lengths.

C Applied Hyperparameters

C.1 Model Cards

We show model cards used in experiments of this paper in Table 8.

C.2 Inference Hyperparameters

Parameter	Value
TensorParallelSize	4
DType	bfloat16
GpuMemoryUtilization	0.95
Seed	42
Temperature	0
MaxTokens	32k
TopK	1
DoSample	False
BatchSize	1

Table 9: vLLM Inference Hyperparameters

We conducted inference using vLLM to support various experiments and evaluations. For a fair comparison across models, all hyperparameters were uniformly set to commonly used default values. To ensure reproducibility and consistency in model outputs, we set the temperature to 0, enabling greedy decoding. Furthermore, to prevent response loss and performance degradation often observed with large batch processing in the vLLM environment, and to maintain consistent output quality, we fixed the batch size to 1. The detailed hyperparameter settings are summarized in Table 9.

D Evaluation Metrics

- **BERTScore** (Zhang et al., 2020) computes semantic similarity between a generated sequence and a reference sequence by taking the maximum similarity between each token in one sequence and the tokens in the other. We use BERTScore to assess semantic alignment between model-generated answers and reference texts:

$$\text{BERTScore} = \frac{1}{N}\sum_{i=1}^{N}\max_{j}\cos(\mathbf{h}_i^{\text{Ref}}, \mathbf{h}_j^{\text{Gen}})$$

- **BLEURT** (Sellam et al., 2020) evaluates the semantic similarity between the generated and reference texts using a pre-trained language model. Unlike BLEURT, BLEU (Papineni et al., 2002) measures n-gram precision between the generated and reference texts, focusing on how much of the generated content is contained within the reference.
- **ROUGE** (Lin, 2004) evaluates n-gram recall, assessing how much of the reference content is covered in the generated text. Variants such as ROUGE-1, ROUGE-2, and ROUGE-L measure different aspects of textual overlap. ROUGE is particularly well-suited for evaluating summarization, as it reflects how well the generated summary captures key information from the source text.

E Additional Analysis

E.1 Query Embedding Analysis

To validate the semantic diversity of the proposed `ScholarBench`, we conducted a t-SNE-based visualization analysis of the query embeddings (Cai and Ma, 2022). Since `ScholarBench` is a bilingual dataset comprising both Korean and English, it is essential that semantically similar sentences are mapped to a shared embedding space regardless of language. To this end, we employed LaBSE (Feng et al., 2022) as the sentence encoder, which is trained to map semantically equivalent sentences across multiple languages into the same vector space.

Figure 9 presents the result of applying t-SNE to the query embeddings from `ScholarBench`.

Notably, the embeddings are widely dispersed without forming clusters based on question type. This suggests that the queries exhibit rich semantic variation and are not restricted to specific task formats, thereby enabling fairer and more comprehensive evaluation of language models.

E.2 Category Analysis

To analyze the characteristics that LLMs have within each domain, Figure 1 visualizes the performance of the top-10 models from the main experiments for the eight academic domains. First, in the summarization (R-1) task, most models demonstrate high scores in the Engineering and Medical Science domain, whereas they exhibit relatively lower scores in Economy & Management and Socio-Professional Studies. Summarization is relatively easier in structured technical texts, whereas economics and social science texts are more challenging due to contextual dependency and linguistic variability. Despite mild overall variation, model performance is consistently lower in the Economy & Management and Liberal Arts domains in the short-answer task. The Multiple Selection (A-4) task shows large variance in model performance. Models achieve high accuracy in Biology & Earth Science, but perform poorly on categories that require abstract or complex reasoning, such as Liberal Arts and Physics & Mathematics. Model reasoning ability in multiple-selection questions varies by category. For both MCQ and boolean formats, performance is higher in Engineering and Chemical & Biochemistry, but lower in Biology, Liberal Arts, and Physics. The Qwen-32b-reasoning model performs particularly well in reasoning-focused domains, likely due to its pretraining objectives. In contrast, o3-mini and Mistral-24b show stable performance across domains, reflecting stronger generalization.

According to these results, the average performance of models is insufficient to explain domain-specific performance characteristics. Thus, both domain generalization ability and domain-specific reasoning ability should be considered in LLM evaluation. Fine-grained benchmarks like `ScholarBench` are effective in quantitatively revealing these imbalances in performance distribution.

E.3 Short Answer Bilingual Results

Table 10 presents the short answer performance of 10 selected models for both Korean and English, with model selection informed by the results detailed in Table 2. Overall, the BERTScore across the two languages is largely similar, exhibiting only minor differences that are likely attributable to inherent linguistic characteristics. Beyond BERTScore, which quantifies semantic similarity, BLEURT was also utilized for evaluation. Notably, the linguistic discrepancy between the two languages appears more pronounced when assessed using BLEURT.

Model	English BERTScore	English BLEURT	Model	Korean BERTScore	Korean BLEURT
o3-mini	0.852	0.328	o3-mini	0.868	0.403
GPT-4o	0.851	0.342	GPT-4o	0.864	0.367
o1-mini	0.851	0.334	o1-mini	0.863	0.362
llama-70b	0.850	0.311	llama-70b	0.866	0.359
Qwen-72b	0.852	0.337	Qwen-72b	0.867	0.375
Qwen-32b-r	0.846	0.305	Qwen-32b-r	0.861	0.354
Mistral-24b	0.847	0.313	Mistral-24b	0.859	0.335
Gemma2-27b	0.847	0.299	Gemma2-27b	0.864	0.347
Gemma2-9b	0.847	0.307	Gemma2-9b	0.855	0.318
Qwen-7b	0.846	0.286	Qwen-7b	0.848	0.279

Table 10: Short answer evaluation results on English and Korean. Qwen-32b-r is Qwen-32b-reasoning.

E.4 Analysis of Summary Length and Performance

Figures 10 and 11 visualize the relationship between summary length and evaluation scores for English and Korean summaries, respectively. In each figure, the upper plots show the distribution of summary lengths across various performance score ranges, while the lower plots illustrate the regression relationship between summary scores and their corresponding lengths. Across both languages, a low correlation is observed between evaluation metrics and summary length, suggesting that summary length does not significantly influence performance. Furthermore, this indicates that the constructed dataset is not optimized for a particular length. Instead, it maintains consistent quality across a wide range of lengths according to the evaluation metrics, thus reflecting its balanced nature.

E.5 Human Evaluation

Figure 12 compares human evaluation results with the performance of various LLMs on a randomly sampled 1% subset of the entire dataset. The examples used in this comparison are identical to those in Table 1, and all evaluations are conducted with access to the corresponding paragraph.

Similar to the high inter-annotator agreement observed in the human evaluation (Cohen's Kappa:

0.614 for English and 0.706 for Korean), the distribution of model performance on the sampled data shows clear differentiation. The performance gap between large-scale and smaller models is evident, and the consistency between model predictions and human judgments further supports the *reliability* of the `ScholarBench` as a quantitative evaluation standard.

In the graph, the average score from human evaluation (i.e., the Human Upper Bound) is presented as a reference line, above which most models do not reach. This demonstrates that `ScholarBench` is *sensitive* enough to capture fine-grained differences in model performance.

F Prompt Template

F.1 Synthetic Data Generation

The method for generating paragraphs is as follows. We extract paragraphs from full research papers using a sliding window approach. Specifically, we set the window size to 33% of the full document and slide it with a 20% overlap to generate each paragraph.

The example below illustrates the prompt used for generating synthetic data. This prompt includes attribute definitions and standards for each item, defined as criteria, alongside few-shot examples sampled from the 1st synthetic data. The prompt for generating the 1st synthetic data is applied without including few-shot examples.

English data generation prompt

```
To create an evaluation set,
we need to generate five types
of questions and answers as
follows:
{summary, short_answer,
multiple_choice,
multiple_select, true_false}.
⟨Order⟩
1. Generate questions for all
types except 'summary' based
on the given ⟨Topic⟩. If the
⟨Topic⟩ does not match the
document content, feel free
to create an appropriate topic.
2. For all question types
except 'summary', create
questions using multi-hop
reasoning. Please provide the
reasoning process explaining
why it is multi-hop.
3. Verify that the generated
questions match the provided
document.
4. Evaluate whether the
generated answers correctly
respond to the questions.
5. Confirm that the generated
answers are consistent with the
provided document.
6. Identify the question that
is most similar to the ⟨Topic⟩
⟨Topic⟩:
⟨Format⟩:
[Summary]
Write a summary.
[short_answer question]
(Q) Write a question.
(A) Write an answer.
[multiple_choice question]
(Q) Write a question.
A) Write choice.
B) Write choice.
C) Write choice.
D) Write choice.
(A) Write the correct answer.
[multiple_select question]
(Q) Write a question.
A) Write choice.
B) Write choice.
C) Write choice.
D) Write choice.
(A) Write the correct answers.
[true_false question]
(Q) Write a question. +
(True/False)
(A) Write the correct answer.
```

Korean data generation prompt

```
평가셋을 만들기 위해,우리는 다음과
같은 다섯 가지 유형의 질문과 답변을
만들어야 합니다:
{요약, 단답형 질문, 객관식 질문, 다중
선택 질문, 참/거짓 질문}.
⟨순서⟩
1. '요약'을 제외한 유형에서 ⟨주제⟩에
맞게 질문을 생성하세요. 만약 ⟨주제⟩와
```

문서 내용이 일치하지 않는다면, 질문 주제를 임의로 생성하세요.
2. 요약을 제외한 문제 유형 multi-hop으로 생성하세요.왜 multi-hop인지 사고 과정을 제공해 주세요.
3. 생성한 질문이 주어진 문서와 일치하는지 평가하세요.
4. 생성한 답변이 질문에 대하여 올바른지 평가하세요.
5. 생성한 답변이 주어진 문서와 일치하는지 평가하세요.
6. ⟨주제⟩와 가장 유사한 질문을 알려주세요.
⟨주제⟩:
⟨형식⟩:
[요약]
요약을 작성하시오.
[단답형 질문]
(Q) 질문을 작성하시오.
(A) 답변을 작성하시오.
[객관식 질문]
(Q) 질문을 작성하시오.
A) 선택지를 작성하시오.
B) 선택지를 작성하시오.
C) 선택지를 작성하시오.
D) 선택지를 작성하시오.
(A) 정답을 작성하시오.
[다중선택 질문]
(Q) 질문을 작성하시오.
A) 선택지를 작성하시오.
B) 선택지를 작성하시오.
C) 선택지를 작성하시오.
D) 선택지를 작성하시오.
(A) 정답을 작성하시오.
[참/거짓 질문]
(Q) 질문을 작성하시오. + (참/거짓)
(A) 정답을 작성하시오.

F.2 Evaluation Prompt

Multiple-choice question prompt

A multiple-choice question with a single correct answer is provided. The Question contains the given question text. The Choices include four answer options for the question, and you must select the most appropriate one. fewshot1 and fewshot2 provide examples of selecting the most appropriate answer from the Choices for the given Question.
The question is presented in the following format:
Question: {question}
Choices: {choices}
fewshot1: {fewshot1}
fewshot2: {fewshot2}
Refer to the few shot examples, read the Question, and output only the letter corresponding to the correct answer from the Choices. Do not provide any additional explanations, reasons, or detailed content. Only output the letter of the correct answer.

Multiple selection question prompt

A multiple selection question with one or more correct answers is provided. The Question contains the given question text. The Choices include four answer options for the question, and you must select the most appropriate one or more answers. fewshot1 and fewshot2 provide examples of selecting the most appropriate one or more answers from the Choices for the given Question.
The question is presented in the following format:
Question: {question}
Choices: {choices}
fewshot1: {fewshot1}
fewshot2: {fewshot2}
Refer to the fewshot examples, read the Question, and output the letter(s) corresponding to the correct answer(s) from the Choices in Python list format. Do not provide any additional explanations, reasons, or detailed content. Only output the list of correct answer

```
letters.
```

Short answer prompt

```
A short answer question is
provided. The Question contains
the given question text.
fewshot1 and fewshot2 provide
examples of short-answer
responses to the Question. The
question is presented in the
following format:
Question: {question}
fewshot1: {fewshot1}
fewshot2: {fewshot2}
Refer to the fewshot examples,
read the Question, and provide
a short-answer response. Answer
only with keywords or short
phrases. Do not use complete
sentences or provide additional
details or explanations. Only
output the correct answer.
```

Boolean

```
A True or False question is
provided, where the correct
answer is either 0 or 1. The
Question contains the given
question text. fewshot1 and
fewshot2 provide examples
of determining whether the
Question is true or false. The
question is presented in the
following format:
Question: {question}
fewshot1: {fewshot1}
fewshot2: {fewshot2}
Refer to the fewshot examples,
read the Question, and
determine whether it is true
or false. Output 1 if true
and 0 if false. Do not provide
any additional explanations,
reasons, or details. Only
output the corresponding
number.
```

Summarization

```
A paragraph is provided. The
Paragraph is the text to
be summarized. fewshot1 and
fewshot2 provide examples of
creating a simple and clear
summary of the given paragraph.
Read the following Paragraph
and provide a brief and clear
summary. Output only the
summary.
Paragraph: {paragraph}
fewshot1: {fewshot1}
fewshot2: {fewshot2}
```

The evaluation prompts used in this paper are as follows: multiple choice question F.2, multiple selection question F.2, short answer F.2, boolean F.2, and summarization F.2.

F.3 Example of Benchmark Construction

Tables 11 and 12 show examples of benchmark data at the 1st synthetic, 2nd synthetic, and human annotation stages, as constructed via the pipeline shown in Figure 3. The examples present representative questions: short-answer, multiple-choice, and boolean questions in order.

F.3.1 Short-Answer

We design a multi-stage process for enhancing question quality. For the initial stage in Table 11, short-answer questions generated using GPT often suffer from broad phrasing and ambiguous answer candidates. For example, a question like "Which industries have experienced significant advancements due to GPT applications?" is open to multiple interpretations, making it difficult to derive a single correct answer. To address this, in the second stage, an automatic refinement process is applied to narrow the scope of the question and include specific clues, enhancing answer steerability. In this process, by inserting meaning-based clues such as *personalized learning* and *automation of academic tasks*, we enhance the reasoning ability required for the model to understand the context and derive the correct answer. Finally, through human annotation, we enhance the naturalness and clarity of the question phrasing and further refine the questions by adding higher-level meaning-based clues such as *knowledge transfer*, *intelligent tutoring systems*, and *content generation*. Through

Step-1: 1st synthetic data
question: Which industries have experienced significant advancements due to GPT applications? answer: Education and healthcare?
question: What are some of the main technologies used in GPT models? (Select all that apply) choices: [a) Deep learning models, b) Rule-based algorithms, c) Transformer architecture, d) Manual data labeling] answer: [a, c]
question: What technology does GPT mainly utilize? choices: [a) Rule-based systems, b) Deep learning models, c) Supervised learning, d) Semantic analysis] answer: b
question: GPT only follows hand-coded rules to generate text. answer: False
Step-2: 2nd synthetic data
question: Which industry has been notably transformed by GPT through personalized learning and automation of academic tasks? answer: Education
question: Based on the principles underlying GPT, which technologies contribute to its ability to generalize language patterns beyond memorization? (Select all that apply) choices: [a) Pretraining on large-scale textual data, b) Rule-based decision trees, c) Attention-based neural networks, d) Explicit grammar rules] answer: [a, c]
question: Based on GPT's ability to generate human-like text through large-scale training data and neural architectures, which underlying technology enables this capability? choices: [a) Rule-based systems, b) Deep learning models, c) Supervised learning, d) Semantic analysis] answer: b
question: GPT operates based on predefined rules rather than learning from data patterns. answer: False
Step-3: Human annotation
question: Which industry, often associated with knowledge transfer and learning, has seen significant transformation through GPT-driven innovations such as intelligent tutoring systems and automated content generation? answer: Education
question: Considering GPT's architecture and learning process, which of the following elements enable it to generate semantically coherent and contextually relevant responses by leveraging hierarchical representations of language? choices: [a) Transformer-based deep neural networks, b) Unsupervised pretraining on diverse corpora, c) Rule encoding for language syntax, d) Self-attention mechanisms enabling contextual word representation] answer: [a, b, d]
question: Considering that GPT generates coherent text by learning statistical patterns from large datasets using multi-layered neural networks, which of the following best describes the core technology it is built upon? choices: [a) Rule-based systems, b) Deep learning models, c) Supervised learning, d) Semantic analysis] answer: b
question: GPT produces human-like language by identifying statistical patterns in large-scale data rather than depending on predefined rule sets. answer: False

Table 11: Example of step-by-step data generation process for Enlgish.

Step-1: 1st synthetic data
question: 군나르 뮈르달이 구축한 발전경제학 모델의 두 가지 주요 특징은 무엇인가? answer: 사회민주주의적, 사회공학적
question: 뮈르달 발전경제학의 한계로 지적된 요소를 모두 고르시오. choices: [a) 지나치게 구체적인 정책 중심의 접근, b) 추상적이고 일반적인 수준에 머문 이론, c) 사회공학적 전제의 지속, d) 발전 자체에 대한 근본적 반성의 결여] answer: [b, c, d]
question: 뮈르달의 발전경제학 접근법이 갖는 주요 한계점은 무엇인가? choices: [a) 미국적 자본집약형 모델을 과도하게 강조했다, b) 구체적 맥락의 복잡성에 적용하기에 너무 일반적이고 추상적이었다, c) 발전을 단순한 거시적 지표의 성장으로만 보았다, d) 문화상대주의적 접근을 완전히 배제했다] answer: b
question: 뮈르달의 발전경제학은 사회공학적 전제를 완전히 제거했다. (참/거짓) answer: 거짓
Step-2: 2nd synthetic data
question: 군나르 뮈르달의 주요 경제 모델은 무엇인가? answer: 사회민주주의
question: 군나르 뮈르달의 발전경제학이 접근하는 방식의 주요 특징은 무엇인가요? (모두 선택) choices: [a) 거시적 지표의 성장, b) 전 사회적인 변화와의 결합, c) 사회공학적 접근, d) 정책 도입 전후의 사회적 영향 분석] answer: [b, c]
question: 군나르 뮈르달의 발전경제학 접근법에서 강력한 중앙의 권위에 근거한 발전을 상정할 수 밖에 없는 이유는 무엇인가? choices: [a) 사회적 변화의 복잡성, b) 사회공학적 접근의 고수, c) 근대화 이론에 대한 반대, d) 미국 주도의 자본주의 질서] answer: a
question: 군나르 뮈르달의 발전경제학은 단순한 경제 지표의 성장에 초점을 맞추고 있다. (참/거짓) answer: 거짓
Step-3: Human annotation
question: 군나르 뮈르달의 주요 경제 모델 중 자본주의와 민주주의의 원칙을 조화시키면서 경제적 평등과 사회적 정의를 추구하는 정치 이념을 뜻하는 것은 무엇인가? answer: 사회민주주의
question: 군나르 뮈르달의 발전경제학에서 나타나는 주요 특징은 무엇인가? (모두 선택) choices: [a) 거시적 지표의 성장, b) 전 사회적인 변화와의 결합, c) 사회공학적 접근, d) 정책 도입 전후의 사회적 영향 분석] answer: [b, c]
question: 군나르 뮈르달의 발전경제학 접근법에서 강력한 중앙의 권위에 근거한 발전을 상정할 수 밖에 없는 이유는 무엇인가? choices: [a) 사회적 변화의 복잡성, b) 사회공학적 접근의 고수, c) 근대화 이론에 대한 반대, d) 미국 주도의 자본주의 질서] answer: b
question: 군나르 뮈르달의 발전경제학은 단순한 경제 지표의 성장에 초점을 맞추고 있다. (참/거짓) answer: 거짓

Table 12: Example of step-by-step data generation process for Korean.

these staged improvements, questions are refined to require semantic inference and contextual understanding rather than simple information retrieval, and are designed to be solvable using LLMs' parametric knowledge without an accompanying paragraph. This contributes to precisely evaluating models' complex language abilities.

Similarly for Korean, as shown in Table 12, problems generated in the 1st synthetic data have multiple correct answers. In contrast, when generating the 2nd synthetic data by sampling from the 1st generated data, problems with a single correct answer are generated. This demonstrates that the automatic data generation pipeline model proposed in this paper assists in generating problems with unique answers. Subsequently, through a review process, we enhance the completeness of the problems by adding idioms (or phrases) that enable inference of a single correct answer.

F.3.2 Multiple Selection

A 3-step process is followed to enhance the quality of multiple-choice questions. Initial questions of step 1 in Table 11 are broad, such as "What are some of the main technologies used in GPT models?", and present general options, making them solvable based solely on superficial information. Consequently, this results in a limitation where models can achieve high accuracy by relying on simple keyword matching. In step 2, by including conceptual keywords in the question, such as *its ability to generalize* and *language patterns beyond memorization*, we improve them to require understanding and reasoning about GPT's working principles, rather than simple knowledge retrieval. Finally, in the human annotation stage, we incorporate higher-level concepts into the questions, such as *hierarchical representations of language*, *contextually relevant responses*, and *semantically coherent*, and also refine the options to subtly distinguish the roles of technical components, elevating the questions to a level where models must understand the function of each component and infer the correct answer. Through these staged improvements, questions are progressively improved from surface information extraction types to meaning-based inference types, enabling a more refined evaluation of GPT models' complex language understanding and reasoning abilities.

F.3.3 Multiple Choice

Consistent with (Joshi et al., 2017), questions demanding inference capabilities are known to exhibit a higher difficulty than those that do not. As illustrated in Table 12, the 1st synthetic data generation stage produces questions focusing on general patterns, such as inquiring about *what* and *major limitations*. In contrast, the 2nd synthetic data stage generates questions that explicitly require inferential reasoning by exploring the interrelations between theoretical approaches and their underlying premises. This demonstrates that providing initial sample data enables the generation of higher-difficulty questions. All generated data undergoes an additional review process to ensure that the answers to the questions are grounded in the original source data and constitute valid responses.

F.3.4 Boolean

For the boolean presented in Table 11, 12 the 1st synthetic data generation stage produced questions with clear distinctions, largely due to the use of absolute (all-or-nothing) expressions. These questions typically require only a straightforward factual verification to answer. Conversely, the questions generated in the 2nd synthetic data stage posed significant challenges in logical judgment, necessitating a thorough examination of the overall context for resolution.

Consequently, it can be observed that the iterative question generation method proposed in this paper demands greater knowledge and inference capabilities compared to questions generated via single prompting. During the review process, questions that do not require modification are retained as 2nd-stage data, ensuring the quality of the higher-difficulty set.

G Full evaluation results

G.1 Evaluation results for comprehensibility

Table 13 presents the evaluation results for comprehensibility across all models using paragraph-augmented prompting. This setting provides each model with an extended input that includes contextual paragraphs to assess its ability to understand and interpret academic content more effectively.

G.2 Evaluation results for comprehensibility with CoT

Table 14 shows the results of the comprehensibility evaluation when Chain-of-Thought (CoT) prompt-

Model	Summarization			Short Answer	Multiple Selection			MCQ	Boolean	Avg
	R-1	R-2	R-L	BERTScore	A-2	A-3	A-4			
o3-mini	0.392	0.130	0.320	0.898	0.783	0.611	0.591	0.886	0.875	0.609
o1-mini	0.409	0.151	0.348	0.898	0.795	0.656	0.632	0.883	0.872	0.627
GPT-4o	0.417	0.151	0.356	0.897	0.793	0.667	0.643	0.886	0.854	0.629
Qwen-72b	0.396	0.151	0.345	0.902	0.746	0.650	0.628	0.900	0.896	0.624
llama-70b	0.397	0.146	0.341	0.903	0.751	0.620	0.598	0.875	0.873	0.612
Bllossom-70b	0.349	0.129	0.299	0.880	0.676	0.528	0.507	0.767	0.840	0.553
Qwen-32b-reasoning	0.350	0.116	0.303	0.896	0.706	0.572	0.552	0.870	0.862	0.581
Exaone-32b	0.321	0.094	0.267	0.889	0.463	0.611	0.583	0.876	0.879	0.554
Exaone-32b-reasoning	0.316	0.092	0.267	0.886	0.697	0.558	0.542	0.848	0.840	0.561
Gemma2-27b	0.329	0.117	0.283	0.891	0.625	0.516	0.497	0.736	0.767	0.529
Mistral-24b	0.414	0.159	0.359	0.901	0.756	0.580	0.560	0.865	0.827	0.602
Gemma2-9b	0.294	0.096	0.248	0.883	0.577	0.471	0.449	0.695	0.761	0.497
Exaone-8b	0.317	0.092	0.265	0.883	0.746	0.567	0.537	0.855	0.883	0.572
Mistral-8b	0.402	0.151	0.350	0.891	0.708	0.548	0.524	0.848	0.861	0.587
llama-8b	0.381	0.136	0.327	0.895	0.701	0.558	0.536	0.832	0.827	0.577
Bllossom-8b	0.346	0.129	0.301	0.883	0.558	0.464	0.435	0.757	0.729	0.511
Qwen-7b	0.388	0.144	0.338	0.896	0.713	0.573	0.550	0.844	0.856	0.589
Trilion-7b	0.264	0.103	0.232	0.878	0.469	0.355	0.335	0.783	0.754	0.464

Table 13: Full table of comprehensibility evaluation results from paragraph-based prompting experiments (Table 5).

Model	Summarization			Short Answer	Multiple Selection			MCQ	Boolean	Avg
	R-1	R-2	R-L	BERTScore	A-2	A-3	A-4			
o3-mini	0.391	0.130	0.321	0.898	0.780	0.612	0.593	0.886	0.876	0.610
o1-mini	0.409	0.151	0.348	0.897	0.786	0.651	0.628	0.885	0.874	0.626
GPT-4o	0.414	0.149	0.352	0.898	0.796	0.669	0.646	0.880	0.847	0.628
Qwen-72b	0.317	0.116	0.278	0.896	0.741	0.634	0.613	0.885	0.886	0.596
llama-70b	0.383	0.141	0.336	0.902	0.731	0.619	0.596	0.881	0.875	0.607
Bllossom-70b	0.220	0.079	0.192	0.849	0.548	0.423	0.406	0.628	0.739	0.454
Exaone-32b	0.201	0.059	0.169	0.815	0.721	0.570	0.548	0.787	0.661	0.503
Gemma2-27b	0.137	0.038	0.119	0.833	0.481	0.404	0.391	0.523	0.617	0.394
Mistral-24b	0.358	0.142	0.315	0.895	0.714	0.580	0.561	0.843	0.834	0.582
Gemma2-9b	0.090	0.019	0.082	0.834	0.467	0.394	0.379	0.539	0.604	0.379
Exaone-8b	0.164	0.045	0.141	0.556	0.541	0.414	0.392	0.804	0.827	0.432
Mistral-8b	0.324	0.119	0.283	0.882	0.580	0.498	0.476	0.821	0.822	0.534
llama-8b	0.368	0.132	0.322	0.872	0.553	0.464	0.435	0.618	0.794	0.507
Bllossom-8b	0.181	0.064	0.159	0.873	0.367	0.313	0.300	0.611	0.651	0.391
Qwen-7b	0.252	0.082	0.215	0.864	0.640	0.501	0.480	0.793	0.836	0.518
Trilion-7b	0.180	0.063	0.160	0.816	0.145	0.116	0.111	0.556	0.577	0.303

Table 14: Overall evaluation results for paragraph w/ CoT from Table 13.

ing was applied. This experimental setting prompts models to generate intermediate reasoning steps before producing a final response, aiming to enhance interpretability and answer quality.

G.3 Overall evaluation results for English

Table 15 summarizes the English-only evaluation results extracted from the full paragraph-augmented prompting experiments. The results reflect the models' performance specifically on English inputs across all academic domains and question types.

G.4 Overall evaluation results for Korean

Table 16 reports the evaluation results for Korean-language inputs, also based on the full paragraph-augmented prompting experiments. This analysis focuses on assessing multilingual capability by isolating performance on Korean prompts.

Model	Summarization			Short Answer	Multiple Selection			MCQ	Boolean	Avg
	R-1	R-2	R-L	BERTScore	A-2	A-3	A-4			
o3-mini	0.392	0.130	0.320	0.898	0.783	0.611	0.591	0.886	0.875	0.609
o1-mini	0.409	0.151	0.348	0.898	0.795	0.656	0.632	0.883	0.872	0.627
GPT-4o	0.417	0.151	0.356	0.897	0.793	0.667	0.643	0.886	0.854	0.629
Qwen-72b	0.396	0.151	0.345	0.902	0.746	0.650	0.628	0.900	0.896	0.624
llama-70b	0.397	0.146	0.341	0.903	0.751	0.620	0.598	0.875	0.873	0.612
Bllossom-70b	0.349	0.129	0.299	0.880	0.676	0.528	0.507	0.767	0.840	0.553
Qwen-32b-reasoning	0.350	0.116	0.303	0.896	0.706	0.572	0.552	0.870	0.862	0.581
Exaone-32b	0.321	0.094	0.267	0.889	0.463	0.611	0.583	0.876	0.879	0.554
Exaone-32b-reasoning	0.316	0.092	0.267	0.886	0.697	0.558	0.542	0.848	0.840	0.561
Gemma2-27b	0.329	0.117	0.283	0.891	0.625	0.516	0.497	0.736	0.767	0.529
Mistral-24b	0.414	0.159	0.359	0.901	0.756	0.580	0.560	0.865	0.827	0.602
Gemma2-9b	0.294	0.096	0.248	0.883	0.577	0.471	0.449	0.695	0.761	0.497
Exaone-8b	0.317	0.092	0.265	0.883	0.746	0.567	0.537	0.855	0.883	0.572
Mistral-8b	0.402	0.151	0.350	0.891	0.708	0.548	0.524	0.848	0.861	0.587
llama-8b	0.381	0.136	0.327	0.895	0.701	0.558	0.536	0.832	0.827	0.577
Bllossom-8b	0.346	0.129	0.301	0.883	0.558	0.464	0.435	0.757	0.729	0.511
Qwen-7b	0.388	0.144	0.338	0.896	0.713	0.573	0.550	0.844	0.856	0.589
Trilion-7b	0.264	0.103	0.232	0.878	0.469	0.355	0.335	0.783	0.754	0.464

Table 15: Overall evaluation results for English based on Table 2.

Model	Summarization			Short Answer	Multiple Selection			MCQ	Boolean	Avg
	R-1	R-2	R-L	BERTScore	A-2	A-3	A-4			
o1-mini	0.482	0.206	0.406	0.863	0.521	0.434	0.402	0.683	0.747	0.527
o3-mini	0.466	0.180	0.370	0.868	0.612	0.497	0.464	0.720	0.756	0.548
GPT-4o	0.493	0.206	0.416	0.864	0.556	0.498	0.469	0.739	0.694	0.548
Qwen-72b	0.472	0.209	0.409	0.867	0.461	0.424	0.408	0.752	0.825	0.536
llama-70b	0.457	0.193	0.389	0.866	0.526	0.465	0.436	0.752	0.798	0.542
Bllossom-70b	0.465	0.197	0.389	0.847	0.695	0.417	0.406	0.642	0.787	0.538
Qwen-32b-reasoning	0.430	0.166	0.368	0.861	0.545	0.454	0.419	0.697	0.777	0.524
Exaone-32b	0.411	0.147	0.343	0.861	0.548	0.445	0.414	0.716	0.713	0.511
Exaone-32b-reasoning	0.368	0.128	0.313	0.848	0.477	0.384	0.356	0.680	0.575	0.459
Gemma2-27b	0.442	0.183	0.375	0.864	0.523	0.443	0.418	0.710	0.827	0.532
Mistral-24b	0.488	0.214	0.420	0.859	0.533	0.461	0.436	0.717	0.611	0.526
Gemma2-9b	0.412	0.165	0.342	0.855	0.493	0.426	0.401	0.670	0.796	0.507
llama-8b	0.440	0.181	0.372	0.845	0.421	0.368	0.346	0.657	0.485	0.457
Mistral-8b	0.472	0.204	0.407	0.844	0.426	0.356	0.343	0.628	0.610	0.477
Exaone-8b	0.393	0.138	0.330	0.852	0.554	0.426	0.402	0.687	0.735	0.502
Bllossom-8b	0.468	0.200	0.402	0.843	0.479	0.390	0.354	0.616	0.481	0.470
Qwen-7b	0.472	0.203	0.407	0.848	0.512	0.440	0.412	0.685	0.722	0.522
Trilion-7b	0.441	0.192	0.382	0.860	0.543	0.390	0.359	0.659	0.596	0.491

Table 16: Overall evaluation results for Korean based on Table 2.

H Qualitative Results

Table 17 shows qualitative results.

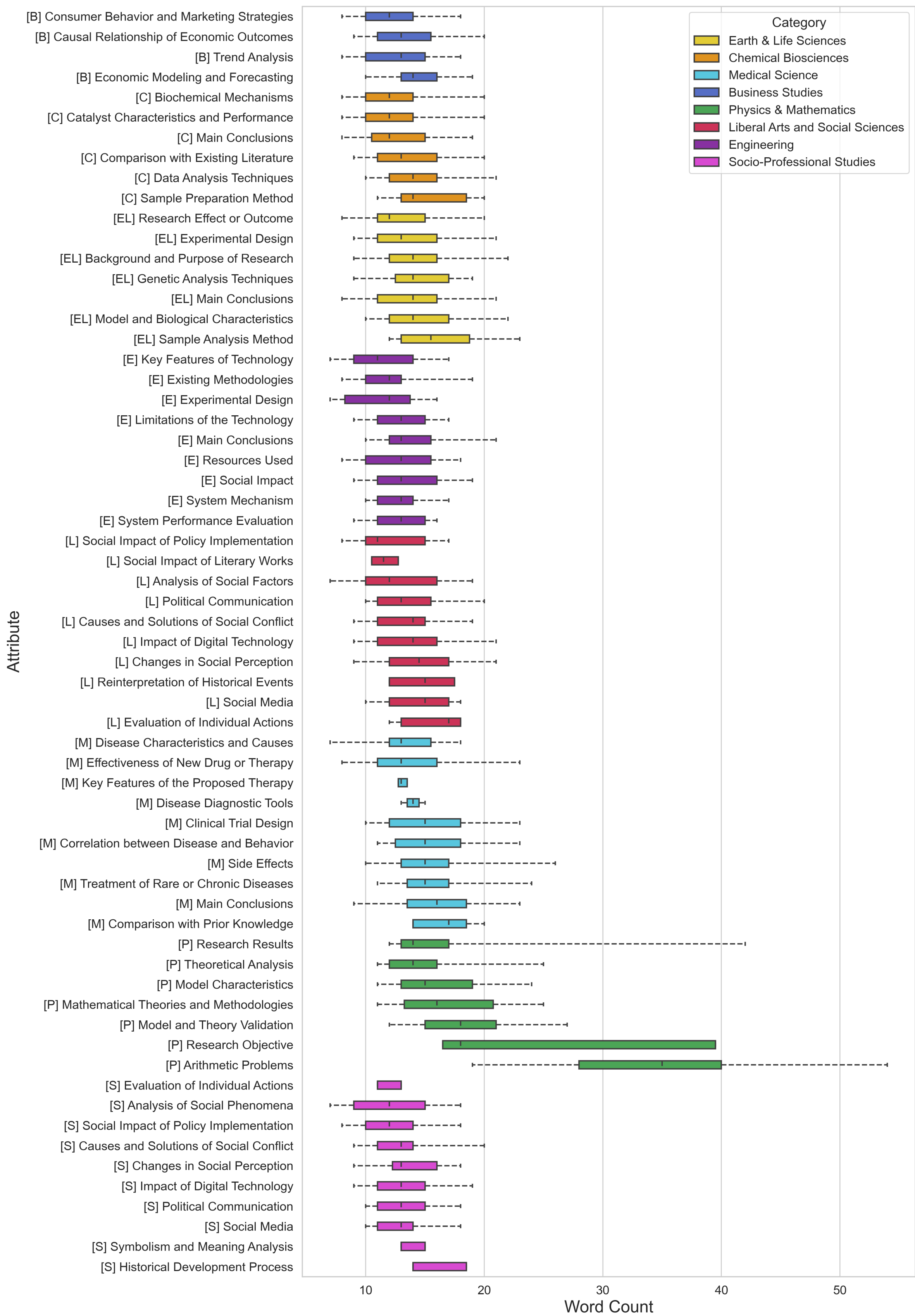

Figure 6: Distribution of query length for each attribute in English categories. The initials preceding each attribute represent abbreviations of the corresponding categories, as follows: [B] Business Studies, [C] Chemical Biosciences, [E] Engineering, [EL] Earth & Life Sciences, [L] Liberal Arts and Social Sciences, [M] Medical Science, [P] Physics & Mathematics, and [S] Socio-Professional Studies.

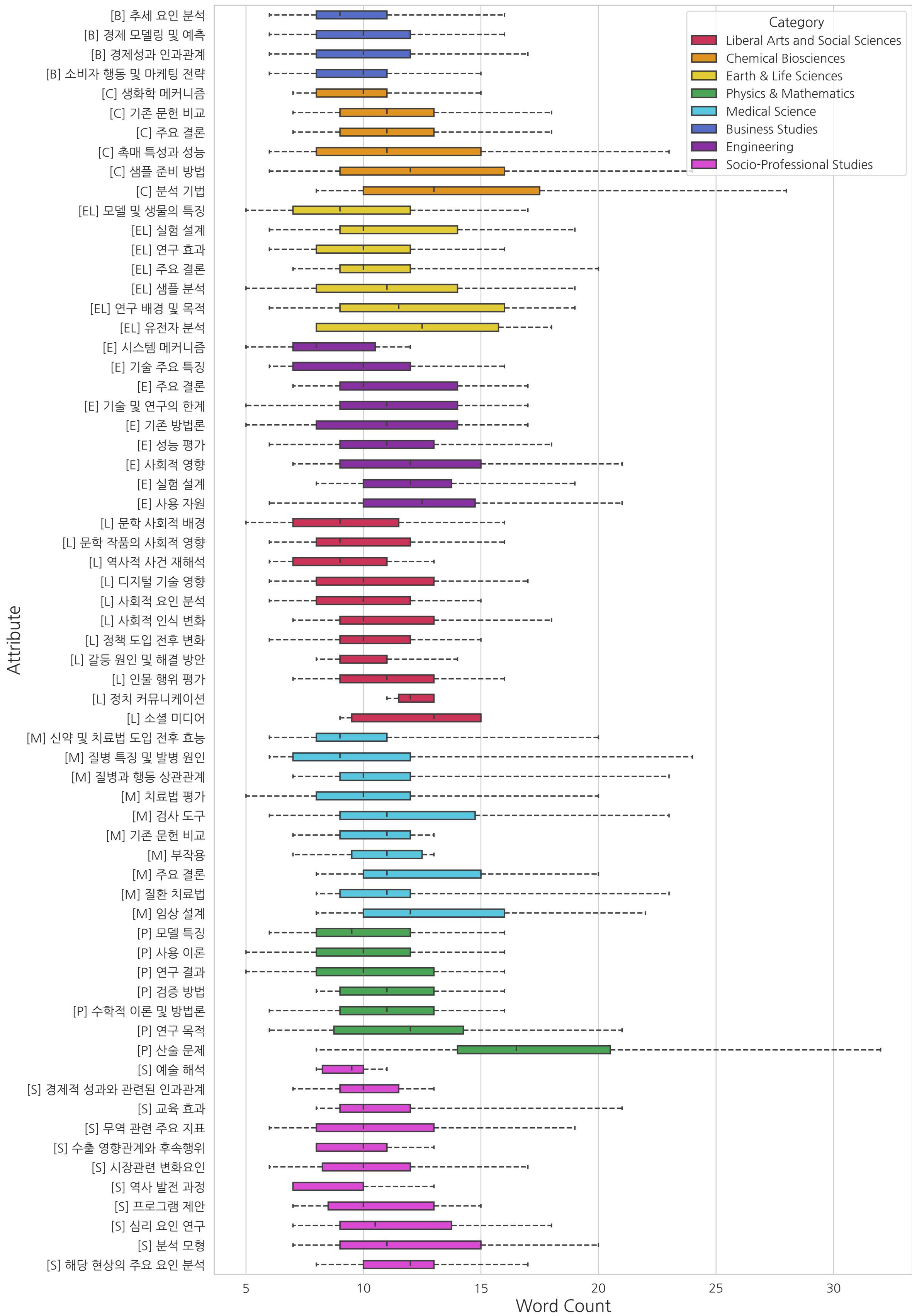

Figure 7: Distribution of query length for each attribute in Korean categories.

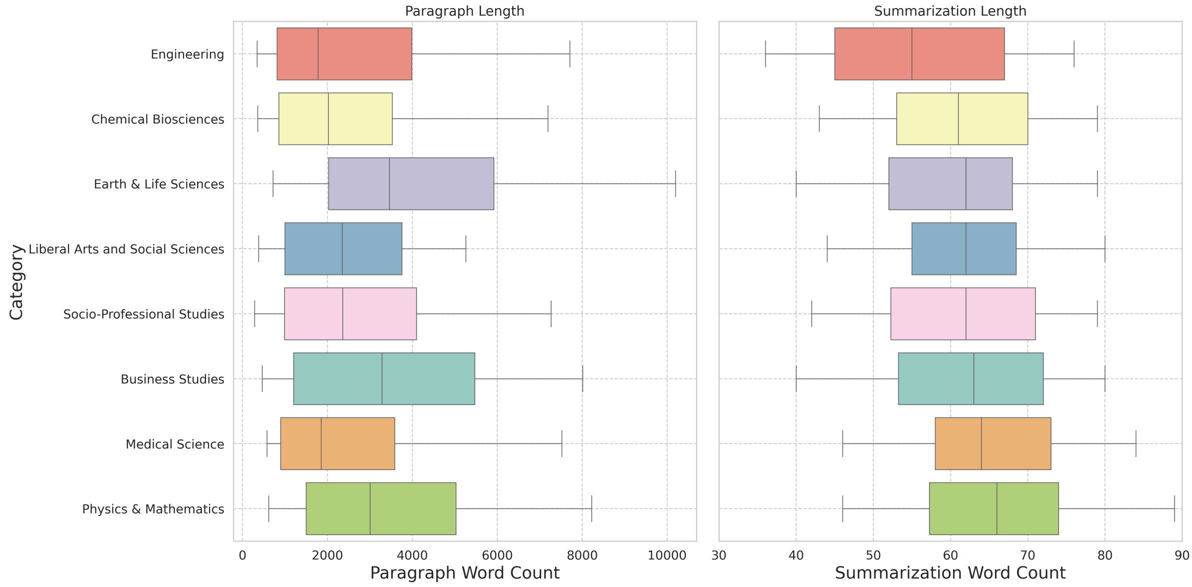

Figure 8: Length distribution of paragraph and summary for each domain.

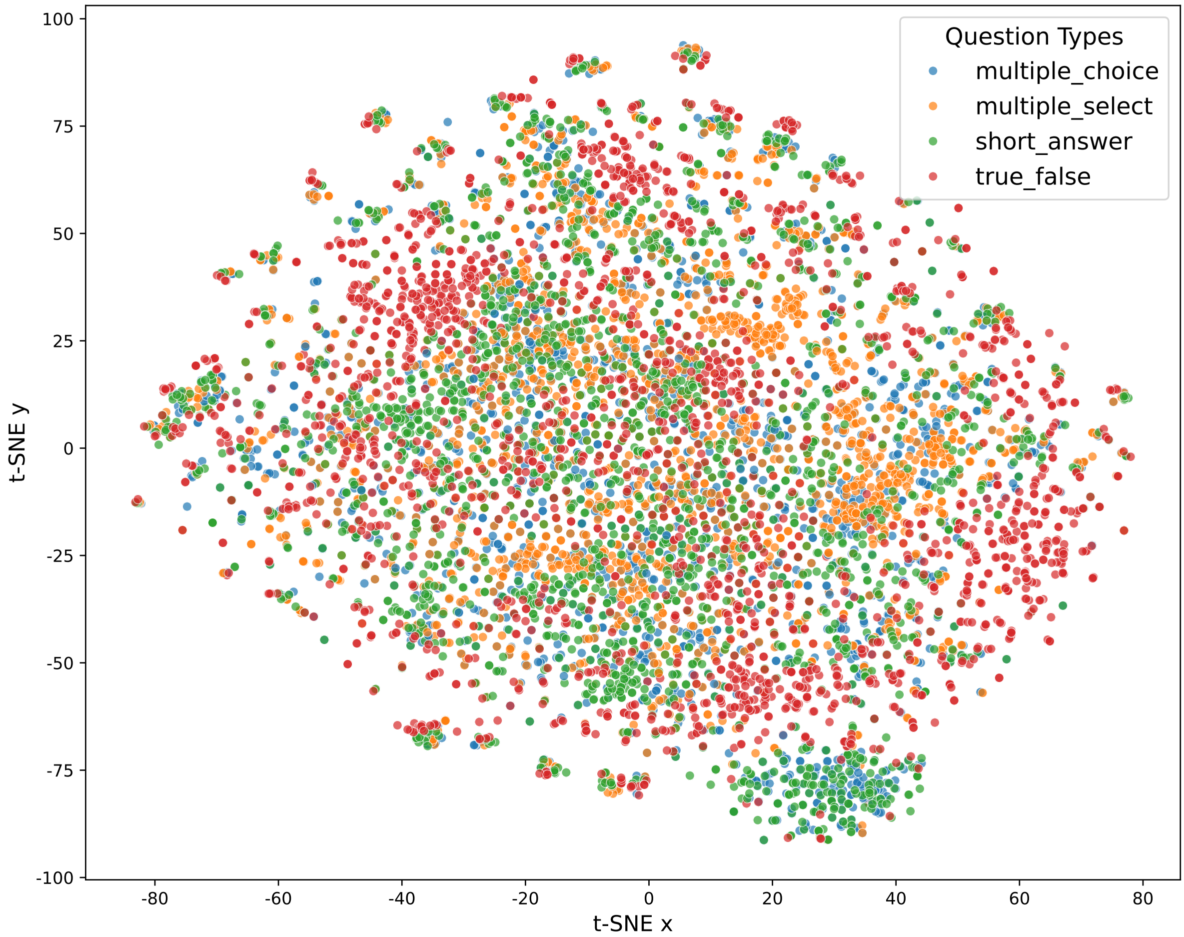

Figure 9: t-SNE visualization of query embeddings in Scholar Bench. The wide and overlapping distribution across question types suggests that the queries are semantically diverse and not bound to specific task types, enabling fairer evaluation.

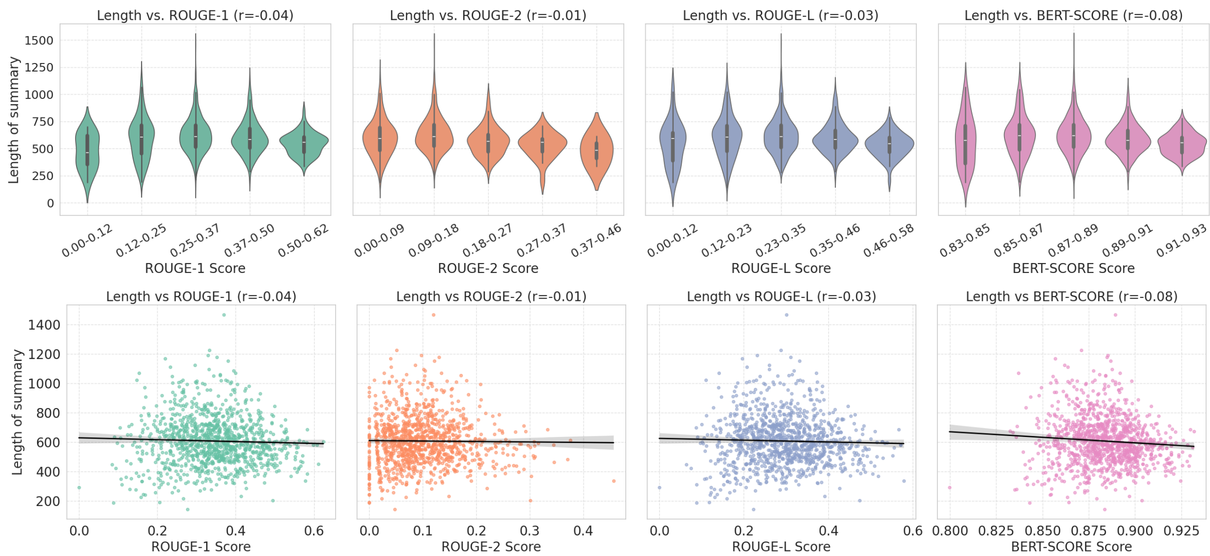

Figure 10: Length distribution of summaries and regression analysis for English.

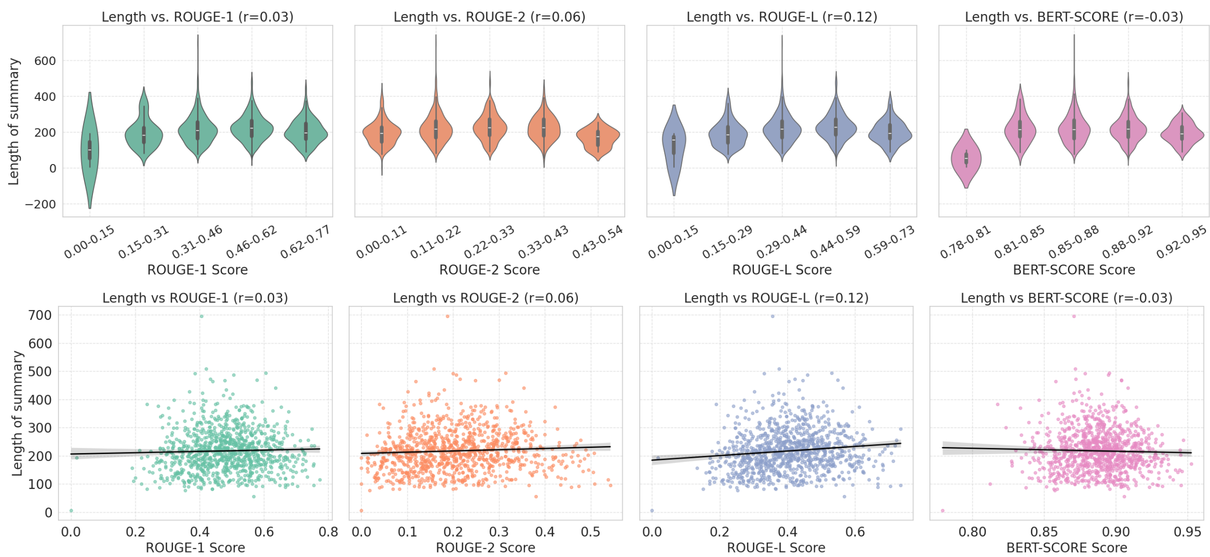

Figure 11: Length distribution of summaries and regression analysis for Korean.

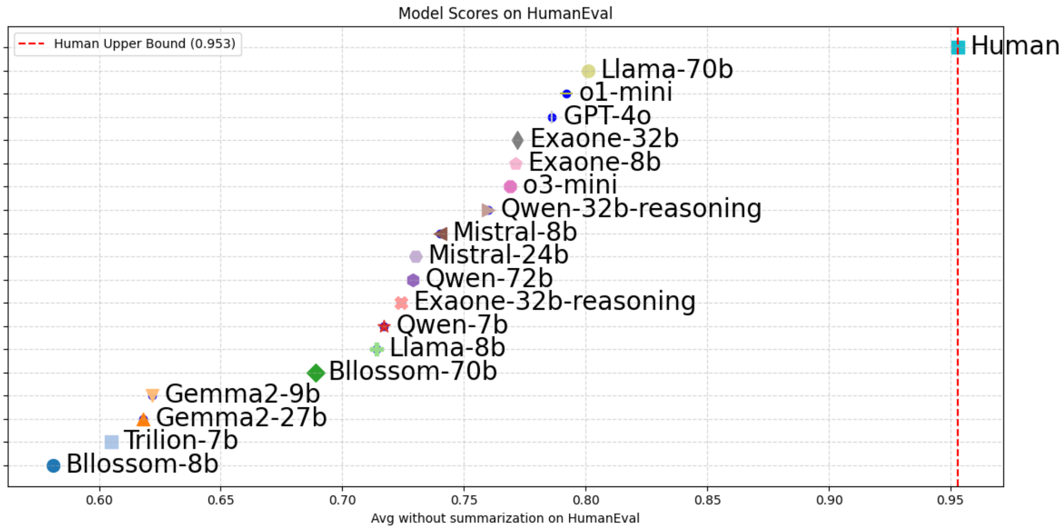

Figure 12: Comparison between human evaluation and model performance on 1% of the data, using the same examples as in Table 1. All evaluations are conducted with access to the corresponding paragraph context.

Category	Paragraph (excerpt)	Question	Answer
Economy & Management	In modern organizations, most forms of overt gender discrimination (i.e., blatant mistreatment or overtly sexistjokes) have become less socially acceptable and have beenreplaced with subtle and often unintentional slights, knownas microaggressions that denigrate women (Capodilupoet al., 2010; Cardador, 2017; Cortina et al., 2013; Yang & Carroll, 2018). To illustrate, Tracy Chou, an experienced software engineer, ... work strategies women use, and the buffers that influence their sensemaking process.	What is the term for comments that subtly and unintentionally denigrate women's competence in professional settings?	Subtle discriminatory comments
		Which of the following are types of gender microaggressions encountered by women in STEM? (Select all that apply)	a) Microassault, b) Microinsult, c) Microinvalidation
		What is one effect of microaggressions on women in STEM?	c) Negative psychological outcomes
		Microaggressions may negatively influence a woman's professional identity. (True/False)	True
Chemical & Biochemistry	The average size from Cryo-TEM was around 57% smaller than that from SN-FSHS-CICS, which might be partially attributedthe physical difference in the size characterized: Under CryoTEM, only the electron-dense region, presumably the core ensemble of lipids and RNA, is captured, ... Subsequently, thebetween each fluorescently tagged payload and LNP size was better visualized by projecting the 3D data onto the corresponding planes for the Cy3-siRNA payload (Figure 4d) ...	How does increasing PEG lipid content in LNPs affect the siRNA payload?	Increased PEG decreases siRNA payload
		Which techniques were used for data analysis of siRNA LNPs size and loading? (Select all that apply)	a) SN-FSHS-CICS, b) Cryo-TEM
		How does the average siRNA payload per LNP change with PEG concentration?	c) It decreases with PEG concentration.
		The increasing percentage of PEG in formulations leads to larger average sizes of LNPs. (True/False)	FALSE
Engineering	It's essential to recognize that GPTs might occasionally make mistakes or give poor answers, particularly when dealing with complicated or ambiguous queries. This highlights the necessity of continual model training, thorough testing, and modification to guarantee that they consistently meet consumer needs. To confirm the efficacy and dependability of using GPTs specifically in the e-commerce area, more research and testing are required. ...	What is the main aspect HCI addresses in terms of GPT usability?	User interaction efficiency
		Which ethical issues are related to GPT models? (Select all that apply)	a) Privacy concerns, b) Data bias
		What is a potential disadvantage of HCI in GPT models?	b) Potential for biases
		HCI techniques improve GPT usability but might introduce biases. (True/False)	True
Medical Science	our results showed that targeting all three subpopulations with 4-1BB activation and not only the stem-like T cells with OX40 activation endowed HBV-specific CD8[+] T cells with robust antiviral activity. The mechanism behind this observation remains uncertain and may be linked to lower TSL numbers, their potential distinct localization, or the differential biological effects downstream of these two co-stimulatory receptors.[37][,][38] The potential of 4-1BB agonism for initiating anti-tumor T cell responses is well recognized. ...	What therapeutic target is known to reinvigorate dysfunctional HBV-specific CD8[+] T cells?	4-1BB
		What factors affect the proliferation of CD8[+] T cells in the context of dysregulation? (Select all that apply)	a) Co-stimulation, b) Cytokine environment, c) Ag engagement
		Which molecule is expressed exclusively by the Dys-TSL population?	b) OX40
		The activation of OX40 leads to a significant increase in CD8[+] T cells' ability to produce IFN-g. (True/False)	False
Biology & Earth Science	... provided in Table S3. Upon analysing the dyeing performance depicted in Fig. 3A, it be comes clear that both the wool fibers with and without mordant exhibited comparable chlorophyll uptakes, with the unmordanted wool fibers even demonstrating higher chlorophyll uptake values. These results are particularly intriguing, as the fixation of natural dyes in textile fibers typically requires mordanting processes prior to the dyeing cycle in order to enhance the dye uptake (Guesmi et al., 2013; Zhao et al., 2020a). Moreover, achieving a natural dye uptake exceeding 70% without any type of optimization ...	What solvent is primarily used in the ABS process to extract chlorophyll?	Ethanol
		Which components contribute to the recovery process in ABS? (Select all that apply)	a) CuSO4, b) Chlorophyll derivatives, c) Sodium hydroxide, d) Ethanol
		What was observed regarding dye uptake in unmordanted fibers compared to mordanted fibers?	b) Unmordanted fibers showed equal or greater dye uptake.
		The ABS process in the study was shown to have potential health risks associated with pollution. (True/False)	TRUE
Physics & Mathematics	... We then show posterior distributions obtained, respectively, with runs adopting 3k, 6k, and 10k live points, demonstrating the gradual convergence to slightly lower values of radii and larger uncertainties... ... the main mode identified with the ST-U model and reported in panel (A) of the same Figure. The omitting component is always associated with the smaller, closer-to-the-equator, hot spot (labeled as primary in panel (A) of Figure 3). The location and size of the masking element can vary significantly within the identified mode... ... The omitting component is always associated with the smaller, closer-to-the-equator, hot spot (labeled as primary in panel (A) of Figure 3). The location and size of the masking element can vary significantly within the identified mode ...	The NICER instrument collected 1.936 Ms (megaseconds) of data from PSR J0030+0451 over a specific time period. Convert this time into days. (1 Ms = $10^{(}6)seconds)$	22.4 days
		What components are included in the ST+PST model? (Select all that apply)	a)Primary hot spot ,b)Secondary hot spot emitting, c)Secondary hot spot masking
		The data for PSR J0030+0451 included multiple inference runs with various live point (LP) settings. If one inference run used 10,000 live points and another used half that amount, how many live points did the second run use?	c) 5,000
		Multimodal structures in a posterior surface suggest the existence of multiple solutions or interpretations for a model's parameters. (True/False)	TRUE

Table 17: Qualitative results for each topic category I.

Category	Paragraph (excerpt)	Question	Answer
Social Science	When management seeks to incorporate all stakeholders and their demands, this may only be feasible in a sequential manner where some stakeholders lose with regard to some aspects in the short term and others gain, but then sequential negotiations can help creating packages that foster sustainable development at a societal and planetary level. ... Moreover, Tesla's success has produced a wave of start-ups across the world vying to make EVs at a lower cost than Tesla can. While EVs are not a perfect solution, and Tesla is not a perfect ...	How do firms engage in sustainable entrepreneurship by working with others?	Collaborative innovation
		What actions are essential to enhance sustainability? (Select all that apply)	a) Co-creating policies, c) Engaging in responsible lobbying
		Which action is essential for improving sustainability in corporate practices?	c) Engaging in stakeholder collaboration
		Firms can solve complex sustainability issues solely on their own without any external collaboration. (True/False)	FALSE
Humanities, Literature & Arts	Navigational capital refers to the ability to maneuver through institutions created to exclude groups or classes of people (i.e. the Dominican education system, which both symbolically and physically excludes people of Haitian descent). Social capital refers to people and relationships that provide emotional and instrumental support when navigating systems, like schools and government bureaucracies. Linguistic capital includes the cognitive flexibility and social skills that come with the ability to navigate multiple languages. Familial capital involves the history, memory and cultural intuition that one gains through an extended ...	What specific challenge related to documentation impacts school participation for Dominican females of Haitian descent?	Lack of documentation
		What factors contribute to the educational challenges faced by Dominican girls of Haitian descent? (Select all that apply)	a) Cultural attitudes like machismo, b) Economic hardship
		What impact can the lack of documentation have on youth education?	a) Denial of access to national exams
		The absence of documentation does not affect the educational success of Dominican females of Haitian descent. (True/False)	FALSE

Table 18: Qualitative results for each topic category II.