NOVA-63: Native Omni-lingual Versatile Assessments of 63 Disciplines

Jinyang Zhang[1,2], **Kexin Yang**[3], **Yu Wan**[3], **Muyang Ye**[4], **Yue Fang**[1,2], **Baosong Yang**[3], **Fei Huang**[3], **Junyang Lin**[3], **Dayiheng Liu**[3*]

[1]School of Computer Science, Peking University, Beijing, China

[2]Key Lab of HCST (PKU), MOE; SCS, Peking University, China

[3]Alibaba Group

[4]Zhejiang University College of Computer Science and Technology

{jinyangzhang25}@stu.pku.edu.cn, {yangkexin.ykx,wanyu.wy}@alibaba-inc.com

Abstract

The multilingual capabilities of large language models (LLMs) have attracted considerable attention over the past decade. Assessing the accuracy with which LLMs provide answers in multilingual contexts is essential for determining their level of multilingual proficiency. Nevertheless, existing multilingual benchmarks generally reveal severe drawbacks, such as overly translated content (translationese), the absence of difficulty control, constrained diversity, and disciplinary imbalance, making the benchmarking process unreliable and showing low convincingness. To alleviate those shortcomings, we introduce NOVA-63 (Native Omni-lingual Versatile Assessments of 63 Disciplines), a comprehensive, difficult multilingual benchmark featuring 89,107 questions sourced from native speakers across 14 languages and 63 academic disciplines. Leveraging a robust pipeline that integrates LLM-assisted formatting, expert quality verification, and multi-level difficulty screening, NOVA-63 is balanced on disciplines with consistent difficulty standards while maintaining authentic linguistic elements. Extensive experimentation with current LLMs has shown significant insights into cross-lingual consistency among language families, and exposed notable disparities in models' capabilities across various disciplines. This work provides valuable benchmarking data for the future development of multilingual models. Furthermore, our findings underscore the importance of moving beyond overall scores and instead conducting fine-grained analyses of model performance. [1].

1 Introduction

The rapid advancement of Large Language Models (LLMs) has demonstrated remarkable capabilities across a wide array of natural language understanding and generation tasks. As most models are English-centric, evaluations default to English, such as GLUE (Wang et al., 2018), SuperGLUE (Wang et al., 2019), MMLU (Hendrycks et al., 2021a), BigBench (Srivastava et al., 2023), MMLU-Pro (Wang et al., 2024), SuperGPQA (Du et al., 2025), etc. However, these benchmarks that only focus on English overlook the importance of joint assessment and the benefits that multilingualism may bring to low-resource languages (Pfeiffer et al., 2022; Üstün et al., 2024; Aryabumi et al., 2024). Therefore, more benchmarks begin to assess LLMs' multi-language performance, e.g., MMMLU (Hendrycks et al., 2021a), MHellaswag (Dac Lai et al., 2023), MMLU-ProX (Xuan et al., 2025). However, these multilingual benchmarks are built using translation technology, which may introduce **"translationese"** artifacts (Bizzoni et al., 2020). Although some native-content benchmarks avoid these issues, they still have significant limitations. Specifically, they often exhibit **restricted difficulty** (Romanou et al., 2024) and either show **constrained diversity** (Hasan et al., 2021; Team, 2024) or **imbalance** in distribution across disciplines, due to inaccessibility.

To alleviate those problems above, we introduce **NOVA-63** (**N**ative **O**mni-lingual **V**ersatile **A**ssessments of 63 Disciplines), a general[2] multiple-choice benchmark containing 89,107 native questions across 14 languages and 63 academic secondary disciplines, covering 8 common language families and approximately 69% of the global population. [3] Specifically, we design a rigorous four-stage pipeline: (1) Data collection from native speakers to **avoid translationese**, (2)

*Corresponding authors. Work done when Jinyang Zhang is intern at Alibaba Group.

[1]Our dataset has been open-sourced at https://huggingface.co/datasets/zjy1298/NOVA-63

[2]In this paper, "general knowledge" refers to comprehensive coverage across a wide range of graduate-level academic disciplines, as opposed to domain-specific.

[3]Sourced from https://www.ethnologue.com and Wikipedia. See Appendix C for more statistics.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 7147–7189
November 4-9, 2025

Group	Benchmark	Native Content	Difficulty Control	Discipline Balancing	Lang. (#)	Effective Questions
English	MMLU (Hendrycks et al., 2021a)	✓	✗	✓	1	15,908
	MMLU-Pro (Wang et al., 2024)	✓	✓	✓	1	12,032
	SuperGPQA (Du et al., 2025)	✗	✓	✓	1	26,529
Multilingual	MMMLU (Hendrycks et al., 2021a)	✗	✗	✓	14	15,908
	MMLU-ProX (Xuan et al., 2025)	✗	✓	✓	13	11,829
	INCLUDE (Romanou et al., 2024)	✓	✗	✗	44	22,637
	NOVA-63 (this work)	✓	✓	✓	14	89,107

Table 1: Comparison of different benchmarks. Native Content indicates whether the benchmark includes translated questions. Difficulty Control shows if it implements systematic difficulty assessment and filtering. Discipline Balancing represents whether the question number is similar across disciplines. Lang. (#) shows the number of supported languages. Effective Questions shows the total question count, with translations counted once. And our effective questions are limited to the number of the open-source version of datasets, where the data of INCLUDE was obtained from CohereLabs/include-base-44, the most comprehensive open-source version available.

Meta-information annotation to capture problem attributes, (3) Multi-level difficulty screening and filtering with multiple LLMs to **guarantee difficulty**, and (4) Question Supplementation and final selection to **ensure diversity and balance** across disciplines. In particular, the diversified classification in stage (3) helps in model optimisation, while the difficulty annotation using multiple LLMs substantially improves both the complexity and robustness of the questions. Consequently, these advancements make the NOVA-63 more challenging.

We conducted extensive experiments on NOVA-63, collecting evaluation results with 62 LLMs (both open source and closed source, ranging from basic to chat/reasoning). These experiments verify the consistency of model capabilities within the language family and discover imbalances in model capabilities across disciplines. The main contributions are:

1. **We propose NOVA-63, a general large-scale discipline-balanced, native multilingual benchmark** with 89,107 questions in 14 languages across 63 academic disciplines, to comprehensively evaluate the multilingual capabilities of LLMs using native content.

2. **We introduce a comprehensive and generalizable data curation pipeline** that emphasizes native sourcing, rigorous quality, and difficulty control, multi-faceted classification guided by human experts to ensure robustness.

3. **We conduct thorough experimental evaluation and analysis of various LLMs on NOVA-63**, presenting a broad comparative study of their multilingual and multidisciplinary capabilities. This study provides insights into linguistic consistency within language families and highlights performance imbalances across disciplines.

2 Related Work

Recent work focuses on benchmarking the capability of LLMs on knowledge coverage. English benchmarks for LLMs vary in focus. Task-specific benchmarks such as GLUE (Wang et al., 2018), SuperGLUE (Wang et al., 2019), Hellaswag (Zellers et al., 2019), TruthfulQA (Lin et al., 2022), MATH (Hendrycks et al., 2021b), GSM8K (Cobbe et al., 2021), and GPQA (Rein et al., 2023) assess performance on particular tasks or domains. General-purpose benchmarks like MMLU (Hendrycks et al., 2021a), BigBench (Srivastava et al., 2023), MMLU-Pro (Wang et al., 2024), and SuperGPQA (Du et al., 2025) evaluate a model's overall language proficiency across diverse scenarios and disciplines.

To evaluate the LLMs' capability across languages, researchers developed various multilingual benchmarks. Many rely on English translation, including XNLI (Conneau et al., 2018), MMMLU (Hendrycks et al., 2021a), MHellaswag (Dac Lai et al., 2023), MGSM (Shi et al., 2022), MLogiQA (Liu et al., 2020), HumanEval-XL (Peng et al., 2024), MIFEval (Zhou et al., 2023), MMLU-ProX (Xuan et al., 2025), and P-MMEval (Zhang et al., 2024), which suffer from **translationese** (Bizzoni et al., 2020) and cultural context loss. Native multilingual benchmarks like XLSUM (Hasan et al., 2021) and FLORES-200 (Team, 2024) have **limited diversity** in task types. Although INCLUDE (Romanou et al., 2024) provides culturally aware native content, it **lacks**

discipline balance and difficulty control. While it covers 44 languages, the number of questions per language is imbalanced. Moreover, for some languages, it's still hard to evaluate the model's overall capabilities very well.

3 NOVA-63

Given the current lack of a native multilingual discipline balanced benchmark, we propose NOVA-63, a native multilingual general understanding benchmark that includes 14 languages, covering 8 common language families and approximately 69% of the global population. Questions for each language are divided into 13 primary, 63 secondary disciplines based on academic specialties, with a total of 89,107 questions. Our discipline setup references the settings in SuperGPQA (Du et al., 2025) for human graduate-level disciplines with changes in the multilingual contexts. Figure 1 shows detailed language and discipline information. To ensure statistical significance, we maintain a minimum of 50 questions per language in each discipline. Meanwhile, to facilitate evaluation, we set an upper limit of 150 questions. [4]

3.1 Data Selection Pipeline

Our data collection pipeline consists of four components, with the overview shown in Figure 2. For any cases requiring manual validation, we apply overall requirements for the human annotation process, annotator qualifications, and quality assurance, which are provided in Appendix A.

3.1.1 Data Collection

Initial Data Gathering To establish NOVA-63, we engage native speakers of each language to collect questions from local educational websites, academic publications, and exams. To ensure quality and difficulty, we prioritize questions in textbooks, educational platforms, and assessment materials from secondary education to the postgraduate level. Native speaker verification is implemented to guarantee the authenticity of native content. In this way, we are able to collect questions with local cultural characteristics in each discipline. [5]

Data formatting Due to the substantial corpus of questions collected, maintaining standardized formatting across diverse regional sources poses significant challenges for contributors. Thus, we develop a systematic approach utilizing LLMs to extract essential question components, including **question**, **options**, and **answer** via in-context learning (Brown et al., 2020). The questions are classified into multiple-choice questions (MCQ) and question-and-answer (QA) at the same time. To ensure high quality, we hire some people to check the information extracted from the model. [6]

3.1.2 Data Annotation

Quality Annotation and Filtering To ensure the questions' quality, we implement a rigorous annotation process focusing on three key dimensions:

- **Readability**: ensuring linguistic fluency and coherence, no grammatical errors, and the elimination of redundant expressions.
- **Completeness**: ensuring no multimedia dependencies, maintaining option integrity, and preserving contextual information.
- **Clarity**: Confirm question unambiguity and preserve essential technical elements (e.g., code snippets and math expressions).

After annotation, we will discard questions that lack Readability, Completeness, or Clarity to ensure the quality of questions. [7]

Classifying Questions We categorize the questions along two dimensions, which are preserved as metadata for each example[8]:

- **Academic Disciplines**: We categorise the questions according to human graduate specialisms to 13 primary disciplines, 63 secondary disciplines, and 262 tertiary disciplines. We adopt a hierarchical classification approach to determine the discipline of each question progressively. Our disciplines setup refers to superGPQA (Du et al., 2025) with changes in the multilingual context.
- **Cognitive Requirements**: We categorise questions according to their cognitive requirements, distinguishing between *Recitation-based* and *Reasoning-based* questions. The former emphasizes memorized knowledge,

[4] For a more detailed statistical analysis of language distribution across disciplines, please refer to appendix E.

[5] Details can be found in the Appendix A.1.1.

[6] The details of our extraction prompt and verification procedures are written in the Appendix A.1.2.

[7] For detailed annotation process, please refer to Appendix A.2.1.

[8] Please refer to Appendix A.2.2 for classification details.

Language Family	Language	Questions
Indo-European	English (en)	4,371
	French (fr)	6,000
	German (de)	5,917
	Italian (it)	5,353
	Portuguese (pt)	5,342
	Russian (ru)	6,285
	Spanish (es)	5,598
Afro-Asiatic	Arabic (ar)	6,446
Sino-Tibetan	Chinese (zh)	8,840
Austronesian	Indonesian (id)	5,740
Japonic	Japanese (ja)	6,541
Koreanic	Korean (ko)	7,492
Kra-Dai	Thai (th)	7,632
Austroasiatic	Vietnamese (vi)	7,550

Figure 1: A general overview of NOVA-63 in languages and disciplines. The left figure shows the distribution of questions by discipline. The right figure shows two levels of discipline classification and the statistics of the number of questions in each primary discipline. From the inside to the outside, showing primary disciplines, secondary disciplines (secondary disciplines are omitted if more than 5 in a primary discipline), and the number of questions in the corresponding primary discipline. A complete list of disciplines can be found at Appendix D.

the latter requires the comprehensive application and inference of understood concepts.[9]

3.1.3 Multi-level Difficulty Screening

Model Annotation We evaluate questions using various chat/reasoning models including Qwen2.5 series (Yang et al., 2024), QwQ (Yang et al., 2024), DeepSeek-r1 (Guo et al., 2025), DeepSeek-v3 (Liu et al., 2024), Llama3 series (Grattafiori et al., 2024), Gemma-3 series (Team et al., 2025), and Phi-4 (Abdin et al., 2024).[10]

Manual Validation After the models complete the questions, we conduct multi-level screening. First, to ensure difficulty, we select questions where Large LLMs had accuracy rates below 50%. Second, we use statistical methods to the fullest extent possible in resolving the issue of incorrect questions. Specifically, we identified two types of suspicious questions:

- **When incorrect answers concentrate in an option.** Given that correct solutions are generally reproducible and can be verified through various approaches, this clustered option, though initially marked as "incorrect," may actually be the right answer.
- **When smaller models achieve significantly higher accuracy than larger LLMs.** As scaling laws (Kaplan et al., 2020) have garnered widespread support, we speculate that multilingual capabilities are also improved with model size, making these unexpected results particularly noteworthy.

These two points are **not used as direct filtering criteria**. Rather, questions that fall into these two categories should undergo our manual verification before proceeding to the next round. Through manual inspection, erroneous questions are discarded, while validated questions are retained. [11]

3.1.4 Supplementation & Final Selection

Upon completing the above procedures, we conduct an initial statistical analysis of questions. However, the challenges in collecting native-language questions prevented us from obtaining statistically significant samples across all tertiary disciplines. As a result, we decide to organize and coordinate our final selection at the secondary discipline level.

Supplementation As some secondary disciplines have insufficient MCQs after the filtering

[9]The ratio of reasoning questions per language and discipline is in Appendix F.

[10]Details can be found in A.3.1.

[11]For detailed information regarding manual validation, please refer to Appendix A.3.2.

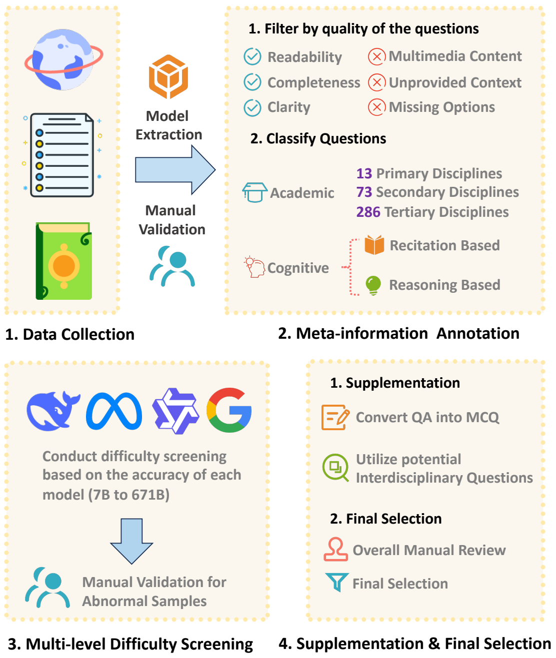

Figure 2: An overview of NOVA-63's data processing pipeline.

step described above, we employ two complementary approaches to enrich the question pool[12]:

- **Converting QAs to MCQs**: We use LLMs to evaluate the model-generated answers obtained during the difficulty annotation process described above, and select high-quality and challenging QA items with less than 50% accuracy. Then we generate interference options using incorrect solutions proposed by LLMs. Given that only 2425 MCQs are transformed, all of them undergo the aforementioned annotation and filtering procedures to ensure consistent quality and difficulty levels. Additionally, to ensure multilingual originality and linguistic correctness, these questions should all be manually reviewed, as generated interference options may introduce non-native or unnatural content.

- **Utilizing interdisciplinary questions**: According to our observation, certain questions appear to span multiple disciplines (e.g., Mechanics intersecting with Physics). Therefore, we identify potentially overlapping disciplines and use LLMs to determine whether a question belongs to the intersection of two secondary disciplines. After that, we assign the question to the discipline with the smaller question pool, aiming to balance distribution across disciplines.

Final Selection Finally, we conduct a manual review of all candidate questions, focusing on four key criteria:

- **Relevance**: Based on the question description, assess the relevance between the assigned discipline labels and the content of both the questions and their corresponding options.

- **Language Fluency and Originality**: Based on the question stem and its options, assess the overall text quality and identify any issues related to fluency or potential machine translation artifacts.

- **Question Completeness**: Determine whether the question can function as an independent and testable item and if it contains obvious errors. Given the high difficulty of the topic and the broad span of disciplines, we can only review the provided answers and explanations to identify any obvious logical inconsistencies.

After removing questions that failed manual review, we set a cap of 150 questions per discipline per language to maintain disciplinary balance. For disciplines exceeding this threshold, questions are randomly sampled to meet the limit.[13]

4 Experiment

4.1 Experiment Setting

Model Selection We conduct experiments on the NOVA-63 benchmark using both base models and chat/reasoning models. Specifically, for chat/reasoning models, we evaluate a diverse series of models, including Qwen2.5 series (Yang et al., 2024), QwQ (Team, 2025), Qwen3 series (Yang et al., 2025), Deepseek-v3 (Liu et al., 2024), Deepseek-r1 (Guo et al., 2025), Gemma3 series (Team et al., 2025), Phi-4 (Abdin et al., 2024), Llama3 series (Grattafiori et al., 2024), Llama4 series (Meta AI, 2025), Mistral series (Jiang et al., 2023), GPT-4 series (OpenAI, 2023), GPT-5 (OpenAI, 2025), Claude-3.7 sonnet (Anthropic, 2025a), Claude-4 sonnet (Anthropic, 2025b), and Grok-3 (xAI, 2025), Gemini-2.5 (Comanici et al., 2025).[14] As for base models,

[12]Please refer to Appendix A.4.1 for detailed supplementation procedures and manual reviews.

[13]Please refer to Appendix A.4.2 for details.

[14]If not otherwise specified in the name, we use the thinking mode by default for the Qwen3 series.

we include the following models in our evaluation: Qwen2.5 (Yang et al., 2024) series, Qwen3 (Yang et al., 2025) series, Gemma3 (Team et al., 2025) series, Llama3 (Grattafiori et al., 2024) series.[15]

Evaluation Metrics Accuracy is used as the primary evaluation metric for NOVA-63, measured as the proportion of correctly answered questions across all disciplines and languages. Considering the multi-level disciplinary hierarchy, we define three aggregation strategies for computing overall scores: question-level averaging, secondary discipline averaging, and primary discipline averaging. The main results adopt **primary discipline averaging** to avoid overrepresentation of disciplines with more subcategories. [16]

4.2 Main Results

Table 2 presents an overview of the evaluation results. Overall, open-source models underperform compared to closed-source models in the chat/reasoning model category. In the base model category, the Qwen series stands out as the top-performing family. Interestingly, base models tend to surpass their Instruct counterparts on NOVA-63, with the Qwen series showing the most pronounced advantage. We speculate that this may be attributed to differences in evaluation methodologies. Notably, we conduct additional experiments and analyses in the Appendix B.2.3 to verify the robustness of the experiment about the base models, especially the Qwen2.5 series, which demonstrate exceptionally strong performance in the Chinese language test.

Comparison between Chat Models and Reasoning Models The performance gap between reasoning and chat variants is notably illustrated through our analysis. QwQ-32B and Deepseek-r1, both optimized for reasoning, consistently outperform their chat counterparts: QwQ-32B achieves 48.5% average score versus Qwen2.5-32B-Instruct's 42.3%, while Deepseek-r1 scores 51.9% compared to Deepseek-v3's 49.6%. This advantage is particularly evident in Chinese language performance, where these reasoning variants show significant leads (QwQ: 57.5% vs. 45.1%; Deepseek-r1: 68.3% vs. 57.6%).

We suspect this pattern emerges from two factors: the prevalence of reasoning questions in NOVA-63 and the potentially better cross-lingual generalization of reasoning capabilities. However, in contrast, Claude-4-Sonnet performs slightly better (52.9% on average) than its 'thinking' variant (52.7% on average), likely due to its better multi-language optimization, since the Claude-4-sonnet-thinking variant shows stronger performance in higher-resource Western languages.

Scaling Laws in Model Families Qwen3 family demonstrates clear scaling benefits across model sizes. Performance improves consistently from Qwen3-0.6B to Qwen3-4B, with average accuracy rising from 37.4% to 46.5%. This trend continues with larger variants: Qwen3-8B (47.6%), Qwen3-14B (49.7%), and Qwen3-30B-A3B (50.1%), indicating a logarithmic relationship between model size and multilingual capability. Similar scaling patterns are observed in other model series. However, the scaling law is not absolute. For instance, Qwen3-32B (49.2%) shows a slight regression compared to Qwen3-14B (49.7%). These finding suggest that simply increasing model size may not guarantee better multilingual performance.

Model Evolution Analysis Comparing Qwen3 with its predecessor Qwen2.5 reveals substantial improvements in multilingual capabilities. For instance, at the 32B scale, Qwen3 outperforms Qwen2.5 by a notable margin (49.2% vs. 42.3%). The improvements are particularly pronounced in European languages: Qwen3-32B achieves improvements of 11.3% in English (55.2% vs. 43.9%) and 5.7% in French (45.5% vs. 39.8%). These enhancements reflect better training strategies and data quality. Moreover, the consistent improvement across model scales suggests that the advances stem from fundamental improvements in architecture and training methodology, rather than just scaling up parameters.

The Advantage of Pre-training The Qwen series demonstrates outstanding fundamental capabilities, significantly outperforming other models of similar scale across overall and multilingual distributions. This underscores the decisive contribution of large-scale multilingual pre-training, and also explains the strong performance of Qwen chat/reasoning models. Additionally, Qwen and

[15] The names of the models in our paper may differ from the nomenclature of the Hugging Face or the official website, and we have published the full list of models used, the source information for each model and evaluate settings in Appendix B.1.

[16] Detailed descriptions of each calculation method, along with additional performance metrics, are provided in the Appendix B.1 and B.2.

Group	Model	En	Fr	De	It	Pt	Ru	Es	Ar	Zh	Id	Ja	Ko	Th	Vi	Avg
	Chat & Reasoning Models															
	Qwen3-0.6B	29.9	38.7	39.6	37.1	37.9	38.3	37.5	36.8	23.8	40.2	41.6	40.9	39.9	41.8	37.4
	Gemma-3-1B-it	25.0	36.1	36.1	37.4	35.9	38.0	36.0	37.7	23.2	34.5	38.6	39.6	39.2	39.0	35.4
<7B	Qwen3-1.7B	38.3	39.9	44.5	40.9	40.6	42.1	41.6	42.3	30.6	43.1	40.4	42.9	42.1	42.7	40.9
	Gemma-3-4B-it	33.1	41.9	45.0	42.9	41.5	42.7	44.3	40.6	24.4	41.2	44.3	44.8	43.3	44.9	41.1
	Qwen3-4B	**45.0**	**46.2**	**48.4**	**46.9**	**44.3**	**47.7**	**47.8**	**46.2**	**40.3**	**48.0**	**45.7**	**47.6**	**48.6**	**48.6**	**46.5**
	Qwen2.5-7B-Instruct	35.4	38.4	38.4	38.0	37.7	38.2	38.1	38.2	33.7	39.3	38.9	37.4	38.1	37.5	37.7
	Aya-8B	29.6	41.7	39.1	42.4	36.7	41.2	36.9	37.0	24.1	38.5	44.1	41.7	35.3	42.4	37.9
	Llama3-8B-Instruct	32.9	42.3	38.1	39.1	38.1	37.8	38.5	41.8	23.6	41.6	42.8	39.4	40.9	40.0	38.3
	Qwen3-8B	46.7	44.5	48.2	50.2	48.2	48.1	47.4	47.1	43.1	48.9	48.1	50.3	47.2	48.8	47.6
7-14B	Gemma-3-12B-it	38.5	45.4	46.6	47.9	44.6	46.9	45.5	49.4	24.9	48.5	46.6	47.5	45.6	46.1	44.6
	Mistral-Nemo-Instruct	33.5	38.1	39.4	39.1	36.9	38.6	37.0	35.8	27.4	39.5	37.3	35.9	35.3	36.6	36.4
	Phi-4	51.2	**47.7**	**50.5**	**52.7**	46.7	49.9	**49.9**	44.9	26.8	48.3	48.9	49.3	47.1	48.1	47.3
	Qwen2.5-14B-Instruct	41.0	41.7	42.1	45.3	41.3	45.7	42.3	43.4	39.0	45.0	44.6	45.0	43.8	44.2	43.2
	Qwen3-14B	**52.3**	46.4	49.4	49.6	**50.2**	**50.4**	47.2	**49.7**	**47.8**	**51.8**	**49.2**	**51.0**	**48.5**	**52.1**	**49.7**
	GPT-oss-20B	50.5	48.5	50.6	50.9	48.3	48.3	49.5	49.1	32.8	50.9	48.6	50.0	48.8	47.8	48.2
	Mistral-Small-Instruct	38.4	45.5	44.2	48.5	41.9	44.6	42.8	38.1	25.3	42.5	45.0	44.0	39.2	42.8	41.6
	Magistral-Small-2507	43.8	43.3	44.4	45.2	42.8	43.0	42.2	41.3	33.1	45.7	40.3	39.5	39.2	39.8	41.7
	Gemma-3-27B-it	45.3	**51.6**	51.4	52.5	**49.6**	50.0	50.0	51.3	28.7	51.4	47.8	48.7	49.1	50.4	48.4
	Qwen3-30B-A3B-Instruct-2507	**55.7**	50.7	**51.9**	**52.8**	48.7	**52.2**	**52.7**	**51.9**	54.7	**56.2**	**53.6**	**54.1**	**53.1**	**54.4**	**53.0**
14-32B	Qwen3-30B-A3B-Thinking-2507	55.5	49.0	50.9	52.8	49.6	50.9	51.2	50.4	51.6	52.0	49.5	51.9	50.8	51.5	51.2
	Qwen3-30B-A3B	53.1	48.9	51.4	50.0	48.9	49.7	49.9	50.2	51.2	51.8	47.1	50.3	48.6	50.5	50.1
	Aya-32B	36.2	46.8	45.0	48.5	45.3	45.4	44.5	42.1	26.6	42.5	46.3	45.7	39.8	47.2	43.0
	QwQ-32B	53.0	47.9	49.2	49.5	47.1	48.1	47.3	44.8	**57.5**	47.4	45.3	48.3	46.4	46.8	48.5
	Qwen2.5-32B-Instruct	43.9	39.8	43.7	44.1	41.6	43.7	40.4	39.9	45.1	44.5	39.7	42.5	40.4	42.8	42.3
	Qwen3-32B	55.2	45.5	49.4	50.8	47.7	48.4	47.7	47.9	56.9	49.7	48.0	47.5	47.2	47.6	49.2
	Llama3-70B-Instruct	48.5	**56.6**	53.6	55.7	52.7	51.6	52.3	50.2	30.6	53.1	52.0	52.8	52.5	52.1	51.0
	Llama4-scout	51.5	55.6	53.3	55.0	53.1	52.6	52.9	52.2	40.0	54.4	49.7	50.8	51.3	52.6	51.8
	GPT-oss-120B	55.9	49.3	50.8	52.4	47.7	50.3	50.9	48.6	38.4	52.8	49.6	49.4	47.4	48.5	49.4
	Mistral-Large-Instruct	43.1	43.8	48.2	48.9	45.3	45.9	46.2	45.4	34.0	47.5	44.6	44.7	42.0	42.4	44.4
>32B	Qwen3-235B-A22B-Instruct-2507	**64.5**	54.8	54.9	57.3	52.8	**57.2**	**54.6**	**55.8**	**70.9**	**58.0**	**54.8**	**57.6**	**56.8**	**57.2**	**57.7**
	Qwen3-235B-A22B-Thinking-2507	59.8	50.6	54.4	55.1	50.2	51.1	52.8	52.5	60.2	53.2	50.8	55.3	53.5	54.1	53.8
	Qwen3-235B-A22B	57.5	51.0	51.0	52.0	49.5	53.4	50.0	51.8	61.5	50.4	49.9	50.8	50.9	50.7	52.2
	Llama4-maverick	59.5	54.2	**57.8**	**58.4**	**54.5**	55.7	53.4	53.4	49.0	55.7	54.3	54.3	52.6	55.6	54.9
	Deepseek-r1	60.6	49.8	51.5	52.7	49.1	49.8	48.9	49.9	68.3	51.7	48.0	48.8	48.0	49.5	51.9
	Deepseek-v3	56.8	49.0	49.4	51.1	47.1	48.3	47.5	49.7	57.6	50.8	47.4	46.5	45.7	47.6	49.6
	GPT-5	**67.3**	51.2	53.0	54.4	50.8	52.3	53.5	51.6	61.3	56.4	49.3	49.5	49.5	51.4	53.7
	Gemini-2.5-flash	62.8	52.2	53.3	55.5	49.6	49.6	53.4	53.1	53.6	54.3	51.4	50.6	51.8	49.8	52.9
	Gemini-2.5-pro	66.3	52.8	55.2	56.6	53.3	53.1	55.2	55.8	64.4	56.6	51.5	54.3	52.2	52.9	55.7
	Qwen3-max-preview	66.3	**60.3**	**58.7**	**61.5**	**57.6**	**60.1**	**59.1**	**60.1**	**72.9**	**62.4**	**56.5**	**60.4**	**60.7**	**59.3**	**61.1**
	ChatGPT-4o-latest	60.4	54.4	55.3	55.4	51.4	55.0	54.1	52.6	42.0	52.6	54.5	52.7	50.6	51.8	53.1
Close-sourced	Claude3.7-sonnet-thinking	61.2	51.6	50.9	53.3	49.8	51.5	50.2	53.6	46.7	54.4	47.3	48.3	49.2	49.7	51.3
	Claude3.7-sonnet	60.4	51.2	50.1	53.5	51.2	50.2	51.9	55.5	44.2	55.4	49.0	49.3	50.6	49.9	51.6
	Claude4-Sonnet-thinking	63.1	51.9	53.1	56.1	49.7	51.6	54.0	52.6	55.8	53.5	48.5	49.9	48.4	50.0	52.7
	Claude4-Sonnet	63.0	52.9	52.5	54.0	51.3	50.9	53.1	51.4	55.7	53.4	50.1	52.6	50.0	50.0	52.9
	GPT-4.1	62.0	52.6	54.4	54.6	50.5	53.4	51.8	52.7	44.0	54.0	51.9	53.2	48.6	50.9	52.5
	Grok-3	61.4	53.0	51.7	54.9	50.0	51.9	53.1	52.9	45.3	57.8	52.2	52.6	51.6	51.3	52.8
	Base models															
	Qwen3-0.6B-Base	34.2	48.0	44.4	44.4	46.6	42.8	45.6	48.4	29.6	44.0	44.9	46.4	44.6	46.9	43.6
	Gemma-3-1B-pt	24.3	21.6	20.8	17.9	20.0	20.4	27.5	20.9	26.4	23.5	20.6	23.7	22.1	19.3	22.1
<7B	Qwen3-1.7B-Base	39.1	49.6	47.4	48.6	47.4	47.9	48.3	51.4	36.3	49.6	47.0	50.2	50.6	49.0	47.3
	Gemma-3-4B-pt	33.0	42.7	41.6	39.9	42.6	40.5	43.4	43.3	26.0	46.2	42.0	41.0	42.0	44.0	40.6
	Qwen3-4B-Base	**43.2**	**54.2**	**56.1**	**56.9**	**56.5**	**56.8**	**55.1**	**58.4**	**46.4**	**57.6**	**54.8**	**57.9**	**56.6**	**58.4**	**54.9**
	Qwen2.5-7B	42.3	50.5	46.7	48.9	48.4	45.7	48.6	48.9	56.3	48.6	45.8	47.5	42.8	46.4	47.7
	Meta-Llama-3-8B	32.7	38.7	39.6	38.6	37.2	36.5	39.7	39.3	25.0	38.7	39.3	39.8	39.2	39.0	37.4
	Qwen3-8B-Base	48.6	56.6	58.0	57.9	58.8	57.4	57.4	57.7	52.0	57.8	55.0	58.5	57.6	60.8	56.7
7-14B	Gemma-3-12B-pt	41.8	52.2	49.8	49.8	49.5	48.2	48.4	51.3	29.1	52.6	48.6	51.2	49.6	48.2	47.9
	Qwen2.5-14B	44.8	49.4	50.5	49.6	50.2	48.1	48.7	53.2	**62.5**	50.3	45.7	49.4	48.3	49.9	50.0
	Qwen3-14B-Base	**53.8**	**61.2**	**61.9**	**60.9**	**62.8**	**60.4**	**62.2**	**61.6**	62.0	**62.6**	**58.8**	**63.0**	**61.2**	**65.3**	**61.3**
	Gemma-3-27B-pt	48.1	55.0	53.4	53.4	53.6	50.9	54.8	54.1	30.6	54.4	49.2	53.8	52.7	52.6	51.2
14-32B	Qwen3-30B-A3B-Base	47.4	**59.9**	**60.0**	**59.8**	**60.1**	**57.3**	**59.6**	**59.1**	56.1	**60.7**	**54.5**	**60.5**	**57.1**	**59.9**	**58.0**
	Qwen2.5-32B	**50.6**	53.0	53.3	54.9	54.0	53.4	55.0	52.9	**73.5**	56.1	51.4	52.1	50.6	52.9	54.5
>32B	Meta-Llama-3-70B	44.6	51.8	52.0	**53.8**	51.3	50.4	52.8	50.1	30.6	53.9	48.7	48.9	46.4	48.5	48.8
	Qwen2.5-72B	**53.7**	**53.9**	**52.7**	53.0	**53.2**	**52.2**	**53.3**	**55.0**	**76.8**	**56.3**	**51.0**	**54.4**	**52.3**	**54.0**	**55.1**

Table 2: Comparison across 14 languages on performance averaged by primary disciplines. The best model in a column within each size interval we mark in **bold**.

DeepSeek achieve outstanding results on Chinese tasks, surpassing other models. This underscores the significant impact of both the quantity and quality of Chinese corpora in the pre-training process for downstream effectiveness.

5 Analysis

5.1 Consistency of Linguistic Competence within the Language Family

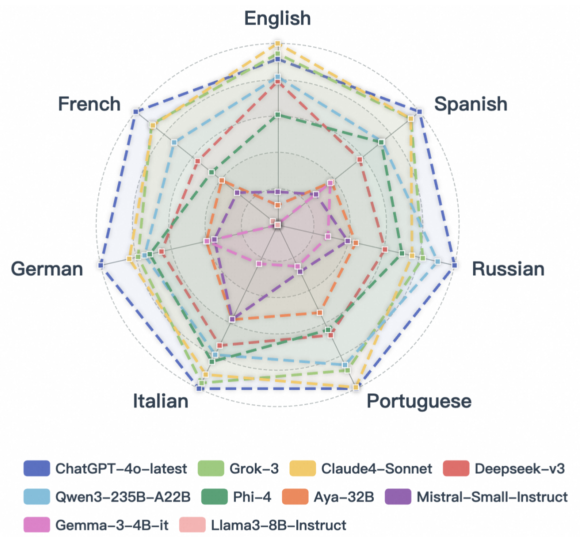

Figure 3: LLMs' performance across Indo-European languages. Models from various families and sizes are sampled to ensure generalisability. Scores per language are normalised between 0 (minimum) and 1 (maximum).

To explore the transfer of LLMs' ability within cognate languages, we compare performances on Indo-European languages in Figure 3. Our analysis reveals a strong cross-lingual consistency within this language family, as evidenced by the consistent hexagonal patterns and graphical nested relationship in Figure 3. Such strong consistency might be attributed not only to shared linguistic features but also to the cultural proximity among these language communities. Since our questions are sourced from native speakers, they tend to naturally incorporate regional cultural contexts.[17]

5.2 Imbalanced model performance in disciplines

The evaluation results from NOVA-63 reveal significant performance gaps across academic disciplines among current LLMs. As shown in figure 4, no single model excels in every discipline — there is no

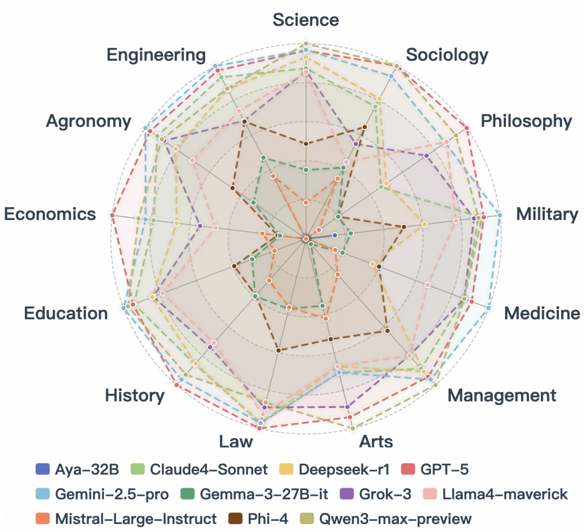

Figure 4: LLMs' performance in **English** across different disciplines. Models with the highest average scores from various LLM families (see Table 2) are selected. Scores per discipline are normalised to 0 (minimum) and 1 (maximum).

"one-size-fits-all" solution among current LLMs.[18] For instance, GPT-5 achieves the highest overall performance in English among the LLMs shown in the figure, yet it underperforms in disciplines such as Management and Medicine.

Moreover, model performance varies significantly across academic disciplines, with some domains showing greater consistency than others. For example, models in the Science and Engineering domains exhibit greater performance consistency, as evidenced by the fewer and less complex interconnections among them in Figure 4. In contrast, models across other disciplines display higher variability in their performance. For example, Qwen3-235B-A22B outperforms other models in History but demonstrates only moderate performance in Sociology, despite the commonly assumed conceptual overlap between the two disciplines. These findings underscore the importance of moving beyond overall scores and instead conducting fine-grained analyses of model performance across individual disciplines. Relying solely on aggregated metrics can obscure critical differences in model capabilities. NOVA-63 benchmark exactly emphasizes subject-specific evaluation, delivers a deeper, more informative view of LLM performance, offering critical insights into their practical applicability across varied disciplines.

[17] A comparative analysis of LLMs' performance across other language families is provided in the Appendix B.2.4.

[18] Please see the Appendix B.2.5 for analyses of performance across different disciplines in other languages

6 Conclusion

In this paper, we present NOVA-63, a native multilingual and discipline-balanced challenging benchmark, constructed through a rigorous four-stage data curation pipeline that integrates automated processing with expert supervision. By conducting extensive experiments, we uncover critical insights into the consistency of linguistic capabilities within language families and identify significant disparities in model performance across different disciplines. These findings contribute to a deeper understanding of the multilingual proficiency of LLMs and offer actionable guidance for future model development and optimization.

In the future, we plan to expand NOVA-63 to incorporate additional languages and disciplines, as well as investigate more effective mechanisms for difficulty control, in order to keep pace with the rapid advancement of LLM capabilities.

7 Data Availability and Usage

Our dataset is freely available for research purposes and can be accessed at `https://huggingface.co/datasets/zjy1298/NOVA-63`.

We released this dataset under the MIT License. This means that anyone is free to use, copy, modify, distribute, and reuse our data, provided that the original copyright notice and license information are retained.

To ensure the validity and fairness of the benchmark evaluation, we explicitly require all users not to use this dataset for model training or training data augmentation, and prohibit any inclusion of this data in training datasets. We will clearly state the above usage restrictions in the license file and user agreement when releasing the dataset. We also encourage researchers to conduct self-assessments in their work to avoid any potential risk of data leakage, thus ensuring the fairness and scientific integrity of benchmark evaluations.

Limitation

The limitations of our work are as follows:

- Because our native collection of multilingual questions requires the help of native speakers, and we need to filter and balance disciplines and difficulty, we only provide problems in 14 languages. Since in other languages it is difficult to ensure that we have more than a certain statistical number of questions in most of the secondary-level disciplines, we will collect more problems in other languages and do the same filtering and balancing in our future work.
- For the convenience of the assessment, we use a multiple-choice format for the assessment. Because the questions themselves are sufficiently difficult, we do not expand on the question options or generate distractors. The difficulty of the questions will be further enhanced in our future work.

Ethics Statement

This work requires manual annotation and validation across multiple languages (details in Appendix A). We compensate our annotators (native speakers) at rates above their local minimum hourly wages. All annotators are clearly informed about the purpose of the data collection and their rights in the annotation process. We have ensured that our annotation guidelines explicitly address the need to avoid cultural biases, offensive content, personal privacy and inappropriate stereotypes across different languages and cultures.

We believe this work will contribute to the healthy development of truly multilingual AI systems through responsible evaluation and assessment. Our goal is to promote the development of language models that can serve diverse linguistic communities effectively and ethically.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Anthropic. 2025a. Claude 3.7 sonnet and claude code. `https://www.anthropic.com/news/claude-3-7-sonnet`. Accessed: 2024-05-15.

Anthropic. 2025b. Introducing Claude 4. `https://www.anthropic.com/news/claude-4`. Accessed: 2025-05-22.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, and 2 others. 2024. Aya 23: Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.

Yuri Bizzoni, Tom S. Juzek, Cristina España-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich. 2020. How human is machine translationese? comparing human and machine translations of text and speech. In *Proceedings of the 17th International Conference on Spoken Language Translation, IWSLT 2020, Online, July 9 - 10, 2020*, pages 280–290. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit S. Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 81 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2475–2485. Association for Computational Linguistics.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *arXiv e-prints*, pages arXiv–2307.

Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, and 1 others. 2025. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *arXiv preprint arXiv:2502.14739*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Samin Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 4693–4703. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

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), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3214–3252. Association for Computational Linguistics.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*.

Meta AI. 2025. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Accessed: 2024.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

OpenAI. 2025. Introducing GPT-5. OpenAI's most capable model, featuring advanced routing and unified system architecture.

Qiwei Peng, Yekun Chai, and Xuhong Li. 2024. Humaneval-xl: A multilingual code generation benchmark for cross-lingual natural language generalization. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 8383–8394. ELRA and ICCL.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States. Association for Computational Linguistics.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. GPQA: A graduate-level google-proof q&a benchmark. *CoRR*, abs/2311.12022.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A Haggag, Alfonso Amayuelas, and 1 others. 2024. Include: Evaluating multilingual language understanding with regional knowledge. *arXiv preprint arXiv:2411.19799*.

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

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 431 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Trans. Mach. Learn. Res.*, 2023.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nat.*, 630(8018):841–846.

Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2018, Brussels, Belgium, November 1, 2018*, pages 353–355. Association for Computational Linguistics.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

xAI. 2025. Grok 3 beta: The age of reasoning agents. https://x.ai/news/grok-3. Accessed: 2024-05-15.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Yun Xing, Junjue Wang, Huitao Li, Xin Li, Kunyu Yu, Nan Liu, Qingyu Chen, Douglas Teodoro, Edison Marrese-Taylor, Shijian Lu, Yusuke Iwasawa, Yutaka Matsuo, and Irene Li. 2025. Mmlu-prox: A multilingual benchmark for advanced large language model evaluation. *CoRR*, abs/2503.10497.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4791–4800. Association for Computational Linguistics.

Yidan Zhang, Boyi Deng, Yu Wan, Baosong Yang, Haoran Wei, Fei Huang, Bowen Yu, Junyang Lin, and Jingren Zhou. 2024. P-mmeval: A parallel multilingual multitask benchmark for consistent evaluation of llms. *CoRR*, abs/2411.09116.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *CoRR*, abs/2311.07911.

A Data Selection Pipeline Details

For all manual processes, we maintain high standards for our annotators. We require multilingual annotators to be **native speakers** with **postgraduate qualification**.

Additionally, we require our annotators to hold **demonstrate strong English communication skills**. We compensate our annotators at rates above their local minimum hourly wages.

All the annotators are recruited through a crowdsourcing platform and our annotator population is evenly distributed, with 3 annotators assigned for each language. And all manual annotations were cross-validated by two annotators. In the case of disagreement, a third annotator acted as an arbiter through a voting mechanism. This process is supported by the crowdsourcing platform and does not require manual assignment of arbiters.

All the prompts and requests have been given **some typographical treatment** to be **displayed beautifully in LaTeX**. **All non-English examples are accompanied by English translation.**

A.1 Data Collection

A.1.1 Selected credible data sources

When collecting data, we have certain requirements for the authenticity, difficulty, and reliability of the collected data. Our requirements for native speakers are as follows (Requirement 1):

Requirement 1

When collecting questions, please ensure the following:

- **Authenticity**: Ensure that questions are collected from local sources, such as practice websites, books, etc., rather than in other countries.
- **Difficulty**: Ensure the questions are at least high school difficulty.
- **Reliability**: Ensure that the sources of the questions are trusted by local people.

Eventually, please return a JSONL file. You can give me data in two possible formats, depending on the formatting of the question:

1. “question_text”:“”, “answer_text”:“”, which means that the question is unformatted HTML text, etc. The content of “question_text” should contain the question, options, and the content of “answer_text” should contain the complete answer.
2. “question”:“”, “options”: [] (IF HAVE), “answer”:“”, means the question is well formatted, question and answer are strings representing the question and answer, and options should be a list of strings representing the options.

By meeting these requirements, we believe native speakers can find questions that meet our criteria.

After our review, in fact, in every language, native speakers are very responsible and most of the questions collected are well formatted. Comparatively speaking, there are more “text” formed collected questions in Vietnamese, Chinese and Indonesian. However, we found that even in the well-formatted questions collected, there are irregular line breaks, HTML tags, etc. in the options, questions and answers, which may be due to copying the questions directly from the browser. So we need to do some further cleaning.

A.1.2 Data Preprocessing and Question Format Normalization

We use LLMs to clean and format the questions' content in the first step. Our purpose is (prompt shown as Prompt 1):

- Remove irrelevant line breaks, HTML tags, potential advertisements, and other redundant information from the question.
- Filter out text containing multimedia content such as images and sound recordings.
- Format the question into a structured form with **question**, **options**, **answer**, and **question_type**.

Prompt 1

Please process a question and its detailed answer scraped from the internet according to the following requirements:
1. Clean the content of the question and answer:

- Remove HTML tags and unnecessary consecutive line breaks
- Retain tags such as <sup>, <sub> that are crucial for solving the problem

2. Question Classification:

- Classify into: multiple-choice question (choice) or non-multiple-choice question (non-choice)
- Special case: If the question includes `<img>` tags which are necessary for answering, set `question_type` to "error"
- Do not embellish content; focus only on extraction and format refinement

3. Structured Information Extraction:

- Split question part into "question" (problem description) and "options" (list of options)
- For non-multiple-choice questions, the options field should be empty
- Extract key answer field from provided answer
- For choice questions: indicate correct options using letters (e.g., ABCD)
- For non-choice questions: place the cleaned answer in the answer field

Note: Some questions have options split across multiple lines; ensure they are not separated during extraction.
[Question]
{question}
[Answer]
{answer}
Please output in the following JSON format without additional explanation:

```
{
  "question_type": "",
  "question": "",
  "options": ["A. Option A", ...],
  "answer": ""
}
```

Here {question} represents the original poorly formatted content of the question and {answer} represents the original poorly formatted content of the answer. We will extract the JSON content to get the well-formatted question.

After LLM extraction, we performed manual verification through sampling to evaluate the extraction quality. The requirement is shown below (Requirement 2):

Requirement 2

Given the original collected text and model-extracted results, we aimed to verify:
- Whether the model successfully filtered out multimedia content
- Whether the model preserved the original question information without accidentally removing critical content
- Whether the extracted content contains any garbled characters and whether all components of the extracted questions are complete

Ultimately, please return the following indicator:

1. "Have Multimodal content?": "YES or NO"
2. "Missing Part": "Specify which part is missing, or null if there is none"

3. “Whether Well-formatted”: “YES or NO”

After examining a sample of questions, we find that:

1. About 0.2% of the questions still contained implicit multimodal parts that are not explicitly present in an explicit manner, such as <img>.

2. 0.7% of the questions reported missing options, and answers, and we observed that the options of the original questions are just incomplete, independent of the extraction of the model.

3. With the exception of 2, no format issues are reprted, and we have reason to believe in the capability of LLMs in this type of simple extraction task.

We will discard any partially incomplete questions in this stage.

A.2 Data Annotation

A.2.1 Quality Annotation and Filter

We use LLMs to evaluate the quality of questions in this step. Our purpose is:

- Assess the readability of questions, including grammar, logic coherence, and fluency.
- Verify the completeness of question content, including necessary context and components, and reconfirm that no multimedia context is included.
- Evaluate the clarity and consistency of expression and formatting.

Given the complexity in measuring completeness and the diversity of practical scenarios, we incorporated numerous empirically observed cases into our evaluation protocol, resulting in a particularly sophisticated prompt (Prompt 2) for assessing this criterion.

Prompt 2

Please evaluate the given question (including its options) based on the following criteria. Rate each dimension and provide brief justification:
1. Readability (3-level scale):

- 1 point: Contains grammatical errors, weak logic, poor fluency, or high repetition
- 2 points: Few grammatical errors, good logic, and fluency
- 3 points: No errors, strong logic coherence, and fluent expression

2. Completeness (2-level scale):

- 0 points: Missing essential information, necessary graphs, options, context, or parts
- 1 point: Contains all required information and components

Note: Domain-specific terms and minor punctuation issues do not affect completeness
3. Clarity and Consistency (3-level scale):

- 1 point: Ambiguous expressions, poor formatting, misuse of symbols
- 2 points: Generally clear with minor issues in formatting or symbol usage
- 3 points: Clear expression, well-formatted, accurate use of symbols

[Question]
{text}
Please output in the following JSON format:

```
{
  "Readability": {
    "score": _,
    "reason": ""
  },
  "Completeness": {
    "score": _,
    "reason": ""
  },
  "Clarity and Consistency": {
    "score": _,
    "reason": ""
  }
}
```

Here {question} represents the concatenations of the question and options. We will extract the json content to get the metrics.

Adhering to the principle of quality over quantity, we selected only questions that achieved full scores across all criteria.

Here, we give a particular question of what kinds of topics are incomplete:

Example 1

Original Question (Chinese):
文中三次提到「那盏旧煤油灯」，分别出现在不同的情节关键点。下列哪一项最准确地概括了它在全文中的象征意义变化?
English Translation:
The "old oil lamp" is mentioned three times in the text, appearing at different plot points. Which of the following most accurately summarizes the changes in its symbolic meaning throughout the text?
Origin Options (Chinese):

A) 从「家庭的温暖」到「战争的残酷」，最后变为「记忆的虚无」

B) 始终象征「主人公对童年的怀念」，没有明显变化

C) 从「希望的指引」到「孤独的陪伴」，最后成为「未完成的遗憾」

D) 仅作为背景道具出现，无特殊象征意义

English Translation of Options:

A) From "family warmth" to "war cruelty", finally becoming "emptiness of memory"

B) Consistently symbolizes "protagonist's nostalgia for childhood" without significant change

C) From "hope's guidance" to "lonely companionship", finally becoming "unfinished regret"

D) Appears only as a background prop with no special symbolic meaning

Note: Obviously there should be an article preceding this, but it is lost during the collection and extraction process.

Similarly, the presence of proper nouns without explanation does not affect the completeness of the question, as in the following question:

Example 2

Original Question (Spanish):
¿Qué técnica se utiliza para modelar la incertidumbre en el rendimiento de las arquitecturas en NAS?
English Translation:
What technique is used to model uncertainty in the performance of architectures in NAS?
Origin Options (Spanish)

A) Redes bayesianas

B) Métodos de Monte Carlo

C) Regresión lineal

D) Árboles de decisión

English Translation of Options:

A) Bayesian networks

B) Monte Carlo methods

C) Linear regression

D) Decision trees

Note: NAS as a proper noun does not need explanation.

A.2.2 Classify Questions

We implement a three-level hierarchical classification system for academic disciplines (see complete discipline list in Appendix D). The model is required to classify each question into exactly one category. When the model indicates that a question does not belong to any subcategory at a given level, we revert to the previous level and repeat the subsequent classification process. If the same situation persists, we maintain the classification at the current level. The prompt used is as follows (Prompt 3):

Prompt 3

Based on the provided discipline list and question, determine which discipline in the discipline list the question belongs to.

discipline list: {discipline list}

Question: {question}

Instructions:

- Select **exactly one discipline** from the discipline list that best matches the question.
- You **must choose the closest match** from the provided discipline list. Do not create or infer new disciplines outside of the list. For example, you should classify Geography under Science and output "Science"; you should classify Political Science under Law and output "Law". Please only make the selection.
- If absolutely no match is possible after careful evaluation, output "[None]". This should only be used in rare cases where none of the disciplines is remotely relevant.

Output format:
You must output **strictly and exclusively** in the following JSON format:

```
{
    "discipline": "discipline name"
}
```

We use the same prompt three times for a question. Here, {discipline list} represents the list of disciplines (the first classification is the list of primary disciplines, the second classification will be the secondary disciplines belonging to the primary discipline, and the tertiary classification is similar) and {question} represents the content of the question to be classified.

For cognitive ability classification of questions, we adopt a few-shot approach to help the model better distinguish between two types of questions. The classification is primarily based on whether the question requires reasoning and analysis beyond basic knowledge. Our prompt is as follows (Prompt 4):

Prompt 4

You will act as a question classification assistant, and your task is to help identify whether the following questions belong to the "recitation-based" or "reasoning-based" category. Please make your judgment based on the following criteria:

1. Recitation-based questions: These questions primarily test the student's ability to remember specific knowledge points, facts, or information. Typically, the answers can be found directly in textbooks or other study materials without the need for additional analysis or reasoning.

- Example: "What is the capital of France?" (direct memory)
- Example: "List three famous works of Shakespeare." (direct recall)

2. Reasoning-based questions: These questions not only require students to have a certain knowledge base but also to be able to use that knowledge to analyze problems, solve issues, or perform logical reasoning. These questions often do not have ready-made answers and require students to think and draw conclusions on their own.

- Example: "According to Newton's second law 'F=ma', if the mass of an object remains unchanged but the acceleration doubles, what happens to the force acting on the object?" (Apply a formula to calculate)
- Example: "Why is 'To Kill a Mockingbird' considered a significant work in American literature? Please explain in terms of its themes and social impact." (comprehensive analysis)

[Question Start]
{text}
[Question End]

Please read the question carefully and categorize it as either "recitation-based" or "reasoning-based".
Finally, please provide the type ("recitation-based" or "reasoning-based") and a brief reason in JSON format:

```
{
    "reason": "",
    "type": "",
}
```

Here {text} represents the concatenation of the question and options. We will extract the JSON content to get the cognitive ability classification of questions. The "reason" we would not systematically check. It rather acts as a COT to help make results more credible.

Typically, reasoning questions are usually confined to STEM disciplines, but the fact is that questions in the humanities and social sciences can and do require integrated analyses and reasoning, such as the following one (Example 3). The ratio of reasoning questions per language and discipline is in Appendix F.

Example 3

Original Question (German):
Wie steht Nietzsche zur Rolle des Subjekts in der Kunst?
English Translation:
What is Nietzsche's position on the role of the discipline in art?
Original Option (German):

A) Das Subjekt ist irrelevant

B) Das Subjekt ist der zentrale Punkt der Kunst

C) Das Subjekt sollte die Kunst kontrollieren

D) Das Subjekt ist eine Illusion

English Translation of Options:

A) The subject is irrelevant

B) The subject is the central point of art

C) The subject should control art

D) The subject is an illusion

LLM's reason: This question requires reasoning rather than mere memorization because it asks for understanding Nietzsche's philosophical stance on subjectivity in art. To answer correctly, one needs to:
1. Understand Nietzsche's broader philosophical framework about the nature of the subject
2. Connect his general philosophy with his specific views on art
3. Analyze how these perspectives intersect Rather than simply recalling a stated position, the answer requires synthesizing Nietzsche's various philosophical ideas about subjectivity, consciousness, and artistic creation.

A.3 Multi-level Difficulty Screening

A.3.1 Model Annotation

In our model annotation process, we use contemporary, commonly used chat/reasoning models to determine the difficulty of the questions. According to the size of the models, we divide them into **small models** ($\leq$ 14B) and **large models** (>14B).

- **Small Models**: Gemma-3-4B-it, Qwen2.5-7B-Instruct, Llama3-8B-Instruct, Gemma-3-12B-it, Qwen2.5-14B-Instruct, Phi-4.
- **Large Models**: Gemma-3-27B-it, Qwen2.5-32B-Instruct, QwQ-32B, Llama3-70B-Instruct, Qwen2.5-72B-Instruct, Deepseek-v3, Deepseek-r1.

During our model annotation in difficulty screening, the prompts used for difficulty assessment differ slightly from the final evaluation. To facilitate answer extraction and further verification of potentially unprocessed multimedia content, we require models to output in JSON format.

For the model annotation, we employ zero-shot settings with **max_tokens=2048** (4096 for reasoning models), **seed=42**, and **temperature=0**. The prompt used is as follows (Prompt 5):

Prompt 5

Carefully read the given question, think about it thoroughly, and provide an answer. If the question requires the use of graphs, videos (such as drawing questions), or cannot be answered without them, please directly return {"error": "This question requires the use of graphs to be answered"}.
[Question]
{question}
[Output Format]
The final result should be presented in JSON format

```
{
    "answer": "your answer"
}
```

Here {question} represents the question to be annotated.

For robust answer extraction, we implement the following hierarchical procedure:

- If multimedia content is detected in the question, we'll directly drop the question.

- If the question is a non-multiple choice question, we will temporarily keep the question. Since we cannot directly use the rule-based method to judge the correctness of the model's answer, we should use the model to judge the correctness of the answer later.

- If the output follows valid JSON format, we directly extract the answer from it.

- If the lowercase output contains the string "answer", we use the first uppercase letter following the last occurrence of "answer".

- If all above methods fail, we default to using the last uppercase letter in the output as the answer.

We observe a small number of cases where model answers could not be extracted, typically due to the model **either failing to reach a conclusion or entering repetitive generation patterns** within the max token limit. In this case, we default to using the last capital letter in the output as the answer.

A.3.2 Manual Validation

Following the model difficulty assessment, we conduct manual verification for two categories of questions (we have already stated the reason in the main text):

- Question less than 30 percent model correct with more than 50 percent (half) probability of choosing an incorrect answer.

- Cases where smaller LLMs ($\leq$ 14B) significantly outperformed larger LLMs in accuracy, requiring verification of question correctness by native speakers with university or higher education from the respective regions.

The requirements for the verification are as follows (Requirement 3). Here, we require that native speakers have a **postgraduate degree or higher in the relevant discipline** when reviewing the correctness of a question.

Requirement 3

Please complete the following tasks:

- Evaluate the question for linguistic accuracy, grammatical correctness, and natural expression.

- Identify any ambiguity, multiple valid interpretations, or misleading phrasing in the question.

- Check if possible visual elements (e.g., images, graphs, diagrams) and context are present, and check whether they are essential to answering the question.

- Confirm the correctness of the provided answer.

Return your judgment using two columns:

1. correctness: "YES/NO"

2. reason: "If the problem is wrong, please tell us the specific reason."

After manual inspection of the samples, we find that:

1. During our manual inspection of abnormal samples, we also identified some correct questions where smaller models outperform larger ones. We hypothesize that this phenomenon may be attributed to the specialized training of current models focusing on higher-order reasoning abilities, potentially at the expense of knowledge recall capabilities (such as Example 4). When the accuracy of the small model is less than 1.2 times the accuracy of the large model, the question correctness rate reaches 97%. Therefore, following the principle that quality outweighs quantity, we have set a threshold of 1.2. All questions where the small model's accuracy exceeds 1.2 times the large model's accuracy will be discarded.

2. For cases where less than 30% of the models answer correctly, yet a particular incorrect option is selected by more than 50% of the models, manual inspection revealed an error rate as high as 90%. As a precaution, we discard all such questions.

Example 4

Question:
Which chemical analysis can be used to determine the origin of wood in musical instruments?
Options

A) Isotope analysis

B) Microbiological analysis

C) Spectral analysis

D) Image analysis

Answer: A

Note: In this case, Qwen-2.5-7B-Instruct and Qwen-2.5-14B-Instruct correctly identified the answer, while larger models like Qwen-2.5-72B-Instruct and DeepSeek-R1 provided incorrect responses.

In this step, although our strategy as such does not guarantee that all such questions are correct, we**balance between the correctness of the questions and the number of questions to the maximum extent**.

Apart from these, during manual inspection, we still identified 0.2% of samples containing images, where the images are implicitly embedded in the question context. Here is an example (Example 5):

Example 5

The S River is the primary source of hydropower and irrigation for the region. If a chemical plant leaks pollutants into its waters, which nearby settlement would face the MOST severe consequences?
A. Greenhill Village
B. Riverside Chemical Plant
C. Oakwood School & Hospital
D. Northford Factory Complex

Note: This question implicitly requires a map or diagram showing the relative positions of these locations along the S River, making it impossible to answer without visual information.

A.4 Supplemantation & Final Selection

A.4.1 Supplementation

Given the large number of unused non-multiple-choice questions and the insufficient number of questions on certain disciplines in some languages, we attempt to convert them into the multiple-choice format. Through our Multi-level Difficulty Screening process, we obtain numerous LLM answers and solution processes. We first use the following prompt to evaluate model-generated answers (Prompt 6):

Prompt 6

We have a question, the content of the question is as follows:
{question}

The official standard answer for this question is:
[Standard Answer]
{reference_answer}

Now, we have a student's response as follow:
{response}

Please score strictly according to the standard answer. Our scoring ignores any process and only considers whether the final answer is correct or not. Your response must only include "Correct" and "Incorrect" options.
The final output format should be in the following JSON format:

```
{
   "judge": "Correct or Incorrect",
    "reason": ""
}
```

Here, {question} represents the concatenation of question and options, {reference_answer} represents the correct answer, and {response} represents the answer of LLM to be evaluated.

Subsequently, we select questions with error rates exceeding 50% for conversion, with the following requirements:

- Maintain the original question topic and correct answer.
- Generate distractors based on model-generated answers and common calculation errors.

The prompt used for conversion is as follows (Prompt 7):

Prompt 7

I will provide a question, the standard answer, and students' responses. Please:
First, analyze the differences between the students' answers and the standard answer:

- Assess the correctness of each student's answer

- Identify common patterns and misconceptions in incorrect answers
- Analyze the possible erroneous thought processes of the students

Then, adapt the original question into a deceptive multiple-choice question designed to mislead students as much as possible, but **ensure that your question has a definite basis in the original answer and that your answer must be correct**.
Since your question will be treated as a stand-alone question, if the original question has the context you need, copy it into your "Question" to **make sure there is no missing information**.
1. Retain the core knowledge points of the original question 2. Design 4 options (A-D), including:

- The correct answer
- Distractors based on identified common errors
- Options that seem plausible but contain subtle errors

Option design principles:

- Use discovered error patterns to design distractors
- Include typical thinking misconceptions
- Contain partially correct but incomplete answers
- For calculation problems, use results from common calculation errors

Attention: You must use the language from the Origin Question to generate new questions (if the Origin Question is in German, you must generate the question and options in German accordingly, and so on). You can learn from the expressions in the question and answer to ensure the language is authentic.
[Original Question]
{question}
[Reference Answer]
{answer}
[Students' Solutions]
{solutions}

Please think step by step about how to design your question. Your multiple-choice question should be as misleading as possible for students, but your question must have a definite basis in the original answer, and your answer must be correct.
Output format:

```
{
    "Question": "",
    "Options": [],
    "Answer": "(A/B/C/D)",
    "Reason": "",
}
```

Here, {question} represents the original question, {answer} represents the correct answer, and {solutions} represents the entire output of the wrong model, including the answer and process. Finally, we will extract the json part of the model output as the new MCQs. To ensure the difficulty and quality of the questions, we will still use the same standards for all the previous scoring and screening processes for this batch of questions. Although we are mainly asking the model to do the extraction task, the question may still suffer from "translationese". Since we only converted 2425 questions as needed, they will **all go through the final manual review** in Appendix A.4.2.

Besides, in order to utilize cross-disciplinary questions, we manually identified potential cross-disciplinary relationships among secondary disciplines, as shown in Table 5. To address the imbalance in question distribution across disciplines, we implemented a cross-disciplinary question identification approach for disciplines with overlapping domains. The following prompt is used to identify questions with equal relevance to multiple disciplines (Prompt 8):

Prompt 8

Please evaluate whether the following question from {original discipline} demonstrates equal relevance to "{discipline}".
A question is considered interdisciplinary if its concepts, theories, applications, or examples are equally applicable to both domains.
Provide your output in JSON format.

Question for evaluation: {text}

Required Output Format:

```
{
    "Answer": "Yes/No"
}
```

Here **{text} represents the concatenation of question content and options**, {original discipline} and {discipline} represent the question's original discipline and possible intersecting disciplines. Here we show an example found in actual operation. This Indonesian question is indeed interdisciplinary (Example 6).

Example 6

Original Question (Indonesian):
Apa perbedaan antara proses pembakaran sempurna dan tidak sempurna dalam kembang api?
English Translation:
What's the difference between complete and incomplete combustion processes in fireworks?
Origin Options (Indonesian):

A) Pembakaran sempurna menghasilkan CO_2 dan H_2O sementara pembakaran tidak sempurna hanya menghasilkan CO.

B) Pembakaran sempurna menghasilkan panas dan gas yang cukup untuk kembang api meluncur tinggi dan meledak, sementara pembakaran tidak sempurna menghasilkan asap berbau dan mengurangi efektivitas ledakan.

C) Pembakaran sempurna menghasilkan CO_2, H_2O tanpa bahan bakar tersisa, sementara pembakaran tidak sempurna menghasilkan CO, karbon (jelaga) dan senyawa organik yang tidak terbakar.

D) Pembakaran sempurna menghasilkan suara keras, cahaya terang dan warna yang cerah, sementara pembakaran tidak sempurna menghasilkan lebih banyak asap, warna yang buruk dan suara yang lebih lemah.

English Translation of Options:

A) Complete combustion produces CO_2 and H_2O while incomplete combustion only produces CO.

B) Complete combustion produces enough heat and gas for the firework to launch high and explode, while incomplete combustion produces smelly smoke and reduces explosion effectiveness.

C) Complete combustion produces CO_2, H_2O with no remaining fuel, while incomplete combustion produces CO, carbon (soot) and unburned organic compounds.

D) Complete combustion produces loud sound, bright light and vibrant colors, while incomplete combustion produces more smoke, poor colors and weaker sound.

Note: This question demonstrates the intersection between Weapon Science and Technology (pyrotechnic engineering) and Chemistry. While the context of fireworks falls under pyrotechnic engineering, understanding the distinction between complete and incomplete combustion processes requires fundamental chemical knowledge.

A.4.2 Final Selection

In the end, we review a sample of the questions pool, and we manually review all the questions converted from QA. Our manual verification requirements are as follows (Requirement 4):

Requirement 4

The annotation task focuses on three main aspects:

- **Discipline Relevance Assessment**: Evaluate the correlation between assigned discipline labels and question content.
- **Text Quality Evaluation**: Consider fluency, accuracy, and ambiguity.
- **Machine translation traces**: Availability of machine translation of texts from other languages.

- **Question Completeness**: Whether the question unclear or missing information that requires external context or relevance to other issues. Whether there are any obvious logic errors in the answers to the questions.

Finally, could you please output the following:

- "Relevance": Rate as: "High", "Medium", or "Low", If rated "Low", suggest a more appropriate discipline from the provided list.
- "Overall Quality": Rate as: "High Quality", "Medium Quality", or "Low Quality". If rated "Low Quality", please give your reason.
- "Machine Translation Artifacts": Rate as:"Severe (affects comprehension)", "Minor (1-2 instances)", "None apparent".
- "Completeness": Rated as "Complete", "Incomplete (ambiguous or missing information or requires external context or linked to other questions or obvious error in answer)".

During manual review, we indeed discover some questions with poor language fluency and errors, mostly due to incorrect usage of professional terminology such as Example 7.

Example 7

Original Question (German):
Welche wichtige Rolle spielt die Zellmembran bei der Aufrechterhaltung des Gleichgewichts in der Zelle?
English Translation:
What important role does the cell membrane play in maintaining cell equilibrium?
Origin Options (German):

A) Sie führt die selektive Durchlässigkeit durch

B) Sie macht die Energie-Umwandlung

C) Sie speichert die Nährstoffe ab

D) Sie produziert die Proteine

English Translation of Options:

A) It performs selective permeability

B) It makes the energy conversion

C) It stores the nutrients

D) It produces the proteins

Note: The machine translation shows typical issues such as literal word-by-word translation, redundant particles, oversimplified verb choices, excessive use of definite articles, and overly simple sentence structures that deviate from standard German academic language conventions.

There are even some common words used incorrectly, such as Example 8.

Example 8

Original Question (Japanese):
画像処理で一般的に使用される「フィルタ」の種類にはどれがありますか？
Translated Question:
Which of the following types of "filters" are commonly used in image processing?
Original Options (Japanese):

A) 平均フィルタ

B) メディアンフィルタ

C) ガウスフィルタ

D) すべての上記

Translated Options:

A) Mean filter

B) Median filter

C) Gaussian filter

D) All the above

Note: The question marker "にはどれがありますか" is misused in this context.

Besides, we also find some translation artifacts in NOVA-63. Here is a typical example, such as Example 9.

Example 9

Original Question((Japanese)):
夏目漱石の「こころ」における主人公の孤独感は、どのように家族のダイナミクスと関連していますか？
Translated Question:
How does the protagonist's sense of loneliness in Natsume Soseki's "Kokoro" relate to family dynamics?
Original Options((Japanese)):
A) 家族からの距離感が孤独を強調する
B) 家族は常に主人公を支えている
C) 家族は物語において重要ではない
D) 孤独感は社会的な要因によるものである
Translated Options:
A) The distance from family emphasizes loneliness
B) Family constantly supports the protagonist
C) Family is not important in the story
D) The sense of loneliness is due to social factors

Note: The term "家族のダイナミクス" (family dynamics) appears to be a direct translation from English, which is inappropriate in this context, as it refers to a therapeutic term in Japanese. This suggests the question is likely translated from another language and doesn't fit the natural Japanese expression.

Besides all the above case studies, we put the detailed statistics in Table 3, and put the detailed statistics of MCQs converted from QA in Table 4.

Based on our manual review results in Table 3, the overall quality of our collected questions is highly reliable. Low-quality questions' ratio are no more than 3% in all our languages, while high-quality questions' ratio reach more than 95%.

The case of Overall Quality "Medium" is usually because there are minor traces of translation in the question. In fact, it is not uncommon to see imported words (katakana) being used in languages due to cultural exchanges between countries, such as Spanish and Japanese. What's more, there're words that cannot be expressed in the local language but can only be expressed in a foreign language. (Just to be on the safe side, we will use the ***langdetect*** to do a final language check after manual review to filter out questions using other languages.)

The percentage of questions that are complete is even higher, at an average of 98.8, with the lowest language at 97.2 per cent complete. Some of the questions labeled incomplete lack proper nouns. But in reality, in the context of the discipline, we don't need to explain these terms, such as the following one (Example 10), so the percentage of titles that are complete should be higher than what we currently have in the Table 3.

Example 10

Original Question (Indonesian):
Apa efek dari perubahan iklim terhadap DPL di Indonesia?
Translated Question:
What are the effects of climate change on deep sea in Indonesia?
Original Options (Indonesian):
A) Meningkatkan suhu air laut
B) Mengurangi jumlah spesies invasif
C) Meningkatkan populasi ikan
D) Menurunkan kadar oksigen di laut
Translated Options:
A) Increase in sea water temperature
B) Decrease in invasive species
C) Increase in fish population
D) Decrease in ocean oxygen levels

Note: The annotators feel the need to give background knowledge about the Indonesian deep sea, which is not needed. In fact, this kind of cultural background knowledge is unique to our kind of native topics.

As for Relevance, we find that in many cases, because the annotator does not understand the context of the subject, they can only take it literally, e.g., 'mechanics' is considered by many to be similar to the mechanics of physics, but in fact it is not! But even then, we still have a maximum of 3.6 per cent of questions that are considered low relevant.

For the questions converted from QA, the results of our full-volume check in Table 4 are not as good as the overall sampling, which demonstrates the necessity for our full-volume review of these topics. Relatively speaking, the proportion of low quality is particularly high for Thai and Arabic. After the full-volume review, we will simply delete the questions with low quality or low relevance or severe translation traces, or incompleteness. Whether the questions rated as "Medium" could be added to NOVA-63, refer to the annotators' suggestion.

After filtering out questions that failed to meet these criteria, we applied the following selection rules:

- For discipline-language combinations with fewer than 150 questions: retain all questions.
- For discipline-language combinations with 150 or more questions: randomly sample 150 questions from the set.

B Evaluation

B.1 Evaluation Setting

For chat/reasoning models, we test:

- Aya series: Aya-8B, Aya-32B
- Qwen2.5 series: Qwen2.5-7B-Instruct, Qwen2.5-14B-Instruct, Qwen2.5-32B-Instruct, Qwen2.5-72B-Instruct
- QwQ-32B
- Qwen3 series: Qwen3-0.6B, Qwen3-1.7B, Qwen3-4B, Qwen3-8B, Qwen3-14B, Qwen3-32B, Qwen3-30B-A3B, Qwen3-30B-A3B-Instruct-2507, Qwen3-30B-A3B-Thinking-2507, Qwen3-235B-A22B, Qwen3-235B-A22B-Instruct-2507, Qwen3-235B-A22B-Thinking-2507
- Deepseek-v3
- Deepseek-r1
- Gemma3 series: Gemma3-1B-it, Gemma3-4B-it, Gemma3-12B-it, Gemma3-27B-it
- Phi-4
- Llama3 series: Llama3-8B-Instruct, Llama3-70B-Instruct
- Llama4 series: Llama4-scout, Llama4-maverick
- Mistral series: Mistral-Nemo-Instruct(12B), Mistral-Small-Instruct(22B), Magistral-Small(24B), Mistral-Large(123B).
- GPT-4 series: GPT-oss-20B(open-source), GPT-oss-120B(open-source), GPT-5, ChatGPT-4o-latest, GPT-4.1[19].
- Cluade3.7 sonnet: Claude3.7-sonnet[20], Claude3.7-sonnet-thinking[21]
- Cluade4 sonnet: Claude4-sonnet[22], Claude4-sonnet-thinking[23]
- Grok-3
- Gemini-2.5 series: Gemini-2.5-flash, Gemini-2.5-pro

For base models, we test:

- Qwen2.5 series: Qwen2.5-7B, Qwen2.5-14B, Qwen2.5-32B, Qwen2.5-72B
- Qwen3 series: Qwen3-0.6B-Base, Qwen3-1.7B-Base, Qwen3-4B-Base, Qwen3-8B-Base, Qwen3-14B-Base, Qwen3-30B-A3B-Base
- Gemma3 series: Gemma3-1B-pt, Gemma3-4B-pt, Gemma3-12B-pt, Gemma3-27B-pt
- Llama3 series: Llama3-8B, Llama3-70B

By default, for open-source models, we use the inference parameters (e.g., temperature, top_p, top_k) recommended in their HuggingFace demos. For a small number of models without provided examples, we set temperature=0.7 and top_p=0.95 for reasoning models, and use greedy decoding for chat models. Closed-source models are evaluated with their respective default parameters. For all base models, we apply greedy decoding.

The models mentioned in this article and the corresponding access addresses are shown in the Table 6.

For base models, we employ a five-shot approach with default parameters except for the following specifications while the zero-shot approach is used for char/reasoning models. For all models, we just used greedy generation with $temperature = 0, seed = 42$. All other parameters remain at their default values. The specific prompts used for both zero-shot and five-shot evaluations are provided below. For chat/reasoning models with instruction-following capabilities, we design our prompts to facilitate answer extraction by requiring models to end their responses with "The answer is". Our prompt design draws inspiration from the Chain-of-Thought (CoT) prompts used in MMLU, allowing

[19] GPT-4.1-2025-04-14

[20] Claude3.7-sonnet-20250514

[21] Claude3.7-sonnet-thinking-20250514

[22] Claude4-sonnet-20250514

[23] Claude4-sonnet-thinking-20250514

Language	Relevance			Overall Quality			Translation Traces			Completeness	
	Low	Med	High	Low	Med	High	Severe	Minor	None	No	Yes
English	0.1	0.6	99.3	0.3	2.4	97.3	0.0	0.0	100.0	0.7	99.3
French	0.7	1.0	98.3	1.0	0.6	98.4	0.6	1.3	98.1	0.9	99.1
German	0.3	0.8	98.9	0.0	1.1	98.9	0.0	0.8	99.2	0.1	99.9
Italian	0.1	1.2	98.7	0.2	0.0	99.8	0.2	8.6	91.2	0.4	99.6
Portuguese	0.4	4.3	95.3	0.0	0.0	100.0	0.0	0.0	100.0	0.0	100.0
Russian	0.6	0.9	97.5	0.0	0.7	99.3	0.1	0.7	99.2	0.6	99.4
Spanish	0.1	6.1	93.8	0.1	2.9	97.1	0.1	8.4	91.5	1.8	98.2
Arabic	3.6	2.2	94.2	2.9	1.4	95.7	2.8	1.8	95.4	2.4	97.6
Chinese	0.3	4.3	95.4	2.7	2.2	95.1	0.0	0.9	99.1	2.8	97.2
Indonesian	0.1	0.0	99.9	0.1	2.8	98.9	0.0	1.0	99.0	1.9	98.1
Japanese	1.7	2.7	95.6	1.3	1.3	97.4	1.4	4.7	93.9	0.6	99.4
Korean	0.8	5.6	93.6	0.1	3.9	96.0	0.9	3.4	95.7	1.8	98.2
Thai	0.4	0.6	99.0	0.6	0.0	99.4	0.8	5.7	93.5	2.0	98.0
Vietnamese	0.1	1.4	98.5	0.7	0.4	98.9	0.7	0.4	98.9	0.7	99.3

Table 3: Results of manual review on a sample of questions. (“Yes” represent “complete” and “No” represent “Incomplete”)

Language	Relevance			Overall Quality			Translation Traces			Completeness	
	Low	Med	High	Low	Med	High	Severe	Minor	None	No	Yes
English	1.5	2.5	96.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	100.0
French	1.6	3.6	94.8	0.0	0.9	99.1	0.0	2.6	97.4	1.8	98.2
German	0.0	5.5	94.5	0.0	0.0	100.0	0.0	0.5	99.5	0.0	100.0
Italian	1.0	3.5	95.5	0.0	12.0	88.0	0.0	0.0	100.0	0.0	100.0
Portuguese	0.5	3.5	96.0	0.0	0.0	100.0	1.0	0.0	99.0	0.5	99.5
Russian	3.0	1.0	96.0	2.5	15.0	83.5	2.0	10.5	87.5	0.5	99.5
Spanish	0.0	5.0	95.0	0.0	15.0	85.0	0.0	0.0	100.0	0.0	100.0
Arabic	1.5	4.5	94.0	0.0	43.0	57.0	0.0	43.5	56.5	2.0	98.0
Indonesian	0.0	0.0	100.0	0.0	9.5	90.5	0	0.7	99.3	0.0	100.0
Japanese	2.5	3.5	94.0	0.5	3.0	96.5	0.5	3.0	96.5	0.0	100.0
Korean	2.0	2.5	95.5	0.5	10.0	89.5	0.5	10.5	89.5	14.0	86.0
Thai	1.5	2.5	96.0	32.5	33.5	34.0	32.5	33.5	34.0	1.5	98.5
Vietnamese	0.0	0.0	100.0	1.8	0.0	98.2	1.8	0.0	98.2	1.8	98.2

Table 4: Results of manual review on questions converted from QAs (No Chinese questions are converted, “Yes” represent “complete” and “No” represent “Incomplete”).

Original Discipline	Potentially Related Disciplines *("Others" indicates cases where questions could not be classified into any listed secondary disciplines)*
Surveying and Mapping Science and Geography Technology	Environmental Science and Engineering, Geological Resources and Geological Engineering, Computer Science and Technology, Others
Physical Oceanography	Oceanography, Geophysics, Environmental Science and Engineering, Atmospheric Science, Others
Food Science and Engineering	Agricultural Engineering, Chemical Engineering and Technology, Environmental Science and Engineering, Biology, Others
Weapon Science and Technology	Military Studies, Mechanics, Electronic Science and Technology, Control Science and Engineering, Others
Art Studies	Musicology, Language and Literature, Journalism and Communication, History, Others
Veterinary Medicine	Animal Husbandry, Biology, Agricultural Engineering, Public Health and Preventive Medicine, Others
Hydraulic Engineering	Civil Engineering, Environmental Science and Engineering, Geological Resources and Geological Engineering, Atmospheric Science, Others
Journalism and Communication	Art Studies, Language and Literature, Public Administration, Library, Information and Archival Management, Others
Public Administration	Business Administration, Library, Information and Archival Management, Political Science, Sociology, Others
Mechanical Engineering	Materials Science and Engineering, Electrical Engineering, Manufacturing Automation, Control Science and Engineering, Others
Musicology	Art Studies, Language and Literature, History, Psychology, Others
Physics	Chemistry, Mathematics, Astronomy, Engineering Mechanics, Others
Traditional Medicine	Pharmacy, Public Health and Preventive Medicine, Biology, Clinical Medicine, Others
Stomatology	Clinical Medicine, Biology, Traditional Medicine, Public Health and Preventive Medicine, Others
Textile Science and Engineering	Materials Science and Engineering, Chemical Engineering and Technology, Mechanical Engineering, Industrial Engineering, Others
Architecture	Civil Engineering, Transportation Engineering, Naval Architecture and Ocean Engineering, Environmental Science and Engineering, Others
Mechanics	Civil Engineering, Physics, Engineering Thermophysics, Structural Engineering, Others
Animal Husbandry	Veterinary Medicine, Agricultural Science, Biology, Crop Science, Others
Naval Architecture and Ocean Engineering	Mechanical Engineering, Civil Engineering, Hydraulic Engineering, Transportation Engineering, Others
Geography	Geology, Environmental Science and Engineering, Surveying and Mapping Science and Technology, Urban Planning, Others
Language and Literature	Art Studies, History, Journalism and Communication, Psychology, Others
Atmospheric Science	Environmental Science and Engineering, Geophysics, Oceanography, Meteorology, Others
Metallurgical Engineering	Materials Science and Engineering, Chemical Engineering and Technology, Mining Engineering, Engineering Thermophysics, Others
Petroleum and Natural Gas Engineering	Environmental Science and Engineering, Chemical Engineering and Technology, Geological Resources and Geological Engineering, Civil Engineering, Others
Transportation Engineering	Civil Engineering, Environmental Science and Engineering, Architectural Engineering, Urban Planning, Others
Military Studies	Weapon Science and Technology, Political Science, Engineering Mechanics, Logistics and Equipment Management, Others

Table 5: Potential Cross-disciplinary Relationships

Model Name in Our Paper	Hugging Face Link/API Website
Chat/Reasoning Models	
Aya-8B	https://huggingface.co/CohereLabs/aya-expanse-8b
Aya-32B	https://huggingface.co/CohereLabs/aya-expanse-32b
Qwen2.5-7B-Instruct	https://huggingface.co/Qwen/Qwen2.5-7B-Instruct
Qwen2.5-14B-Instruct	https://huggingface.co/Qwen/Qwen2.5-14B-Instruct
Qwen2.5-32B-Instruct	https://huggingface.co/Qwen/Qwen2.5-32B-Instruct
Qwen2.5-72B-Instruct	https://huggingface.co/Qwen/Qwen2.5-72B-Instruct
QwQ-32B	https://huggingface.co/Qwen/QwQ-32B
Qwen3-0.6B	https://huggingface.co/Qwen/Qwen3-0.6B
Qwen3-1.7B	https://huggingface.co/Qwen/Qwen3-1.7B
Qwen3-4B	https://huggingface.co/Qwen/Qwen3-4B
Qwen3-8B	https://huggingface.co/Qwen/Qwen3-8B
Qwen3-14B	https://huggingface.co/Qwen/Qwen3-14B
Qwen3-32B	https://huggingface.co/Qwen/Qwen3-32B
Qwen3-30B-A3B	https://huggingface.co/Qwen/Qwen3-30B-A3B
Qwen3-30B-A3B-Instruct-2507	https://huggingface.co/Qwen/Qwen3-30B-A3B-Instruct-2507
Qwen3-30B-A3B-Thinking-2507	https://huggingface.co/Qwen/Qwen3-30B-A3B-Thinking-2507
Qwen3-235B-A22B	https://huggingface.co/Qwen/Qwen3-235B-A22B
Qwen3-235B-A22B-Instruct-2507	https://huggingface.co/Qwen/Qwen3-235B-A22B-Instruct-2507
Qwen3-235B-A22B-Thinking-2507	https://huggingface.co/Qwen/Qwen3-235B-A22B-Thinking-2507
Qwen3-max-preview	https://help.aliyun.com/en/model-studio/models
DeepSeek-v3	https://huggingface.co/deepseek-ai/deepseek-v3
DeepSeek-r1	https://huggingface.co/deepseek-ai/deepseek-r1
Gemma3-1B-it	https://huggingface.co/google/gemma-3-1b-it
Gemma3-4B-it	https://huggingface.co/google/gemma-3-4b-it
Gemma3-12B-it	https://huggingface.co/google/gemma-3-12b-it
Gemma3-27B-it	https://huggingface.co/google/gemma-3-27b-it
Phi-4	https://huggingface.co/microsoft/phi-4
Llama3-8B-Instruct	https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
Llama3-70B-Instruct	https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct
Llama3.1-405B-Instruct	https://huggingface.co/meta-llama/Meta-Llama-3.1-405B-Instruct
Llama4-scout	https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E-Instruct
Llama4-maverick	https://huggingface.co/meta-llama/Llama-4-Maverick-17B-128E-Instruct
Mistral-Nemo-Instruct	https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407
Mistral-Small-Instruct	https://huggingface.co/mistralai/Mistral-Small-Instruct-2409
Magistral-Small	https://huggingface.co/mistralai/Magistral-Small-2507
Mistral-Large-Instruct	https://huggingface.co/mistralai/Mistral-Large-Instruct-2411
GPT-oss-20B	https://huggingface.co/openai/gpt-oss-20b
GPT-oss-120B	https://huggingface.co/openai/gpt-oss-120b
GPT-5	https://openai.com/api
ChatGPT-4o-latest	https://openai.com/api
GPT-4.1	https://openai.com/api
Claude3.7-sonnet	https://claude.ai
Claude3.7-sonnet-thinking	https://claude.ai
Claude4-sonnet	https://claude.ai
Claude4-sonnet-thinking	https://claude.ai
Grok-3	https://x.ai
Gemini-2.5-flash	https://gemini.google.com
Gemini-2.5-pro	https://gemini.google.com
Base Models	
Qwen2.5-7B	https://huggingface.co/Qwen/Qwen2.5-7B
Qwen2.5-14B	https://huggingface.co/Qwen/Qwen2.5-14B
Qwen2.5-32B	https://huggingface.co/Qwen/Qwen2.5-32B
Qwen2.5-72B	https://huggingface.co/Qwen/Qwen2.5-72B
Qwen3-0.6B	https://huggingface.co/Qwen/Qwen3-0.6B-Base
Qwen3-1.7B	https://huggingface.co/Qwen/Qwen3-1.7B-Base
Qwen3-4B-Base	https://huggingface.co/Qwen/Qwen3-4B-Base
Qwen3-8B-Base	https://huggingface.co/Qwen/Qwen3-8B-Base
Qwen3-14B-Base	https://huggingface.co/Qwen/Qwen3-14B-Base
Qwen3-30B-A3B-Base	https://huggingface.co/Qwen/Qwen3-30B-A3B-Base
Gemma3-1B-pt	https://huggingface.co/google/gemma-3-1b-pt
Gemma3-4B-pt	https://huggingface.co/google/gemma-3-4b-pt
Gemma3-12B-pt	https://huggingface.co/google/gemma-3-12b-pt
Gemma3-27B-pt	https://huggingface.co/google/gemma-3-27b-pt
Llama3-8B	https://huggingface.co/meta-llama/Meta-Llama-3-8B
Llama3-70B	https://huggingface.co/meta-llama/Meta-Llama-3-70B

Table 6: Model Names and Links. For the open source model, we provide the Hugging Face link here, and for the closed source model, we provide the API call URL.

models to show their reasoning process. The specific prompt format is shown below (Prompt 9):

Prompt 9

Q: {question}
Please conclude your answer with 'answer is {A/B/C/D}'.
A: Let's think step by step.

For base models, we adopt a probability-based evaluation approach. We concatenate the prompt with each option and calculate their output probabilities. The probability of each option is computed as:

$$P(o) = \frac{\exp(\text{logit}_o)}{\sum_{o \in \{A,B,C,D\}} \exp(\text{logit}_o)}, \quad (1)$$

$$\text{prediction} = argmax_{o \in \{A,B,C,D\}} P(o), \quad (2)$$

where logit_o represents the model's output logit for option o, and the final prediction is the option with the highest probability.

We borrowed from MMLU and used the following prompt for both the few-shot sample and the question to compose the final query. Since we do not divide our benchmark between test and training sets, we defaulted to **using the first five questions from the same discipline as the few-shots, and for the first five questions themselves we use the last five questions as the few-shots**. The prompt structure is shown below (Prompt 10).

Prompt 10

{question}
A. {option_A}
B. {option_B}
C. {option_C}
D. {option_D}
Answer:

To evaluate model performance comprehensively, we consider three averaging methods:

- Question-level averaging: Each question contributes equally to the final score $\text{Score}_{\text{question}} = \frac{1}{N}\sum_{i=1}^{N} \text{correct}_i$
- Secondary discipline averaging: Each secondary discipline has equal weight. $\text{Score}_{\text{secondary}} = \frac{1}{M}\sum_{j=1}^{M} \frac{\sum_{i \in D_j} \text{correct}_i}{|D_j|}$
- Primary discipline averaging: Each primary discipline has equal weight, with secondary disciplines equally weighted within their primary discipline. $\text{Score}_{\text{primary}} = \frac{1}{K}\sum_{k=1}^{K} \frac{1}{|S_k|}\sum_{j \in S_k} \frac{\sum_{i \in D_j} \text{correct}_i}{|D_j|}$

where N is the total number of questions, M is the number of secondary disciplines, K is the number of primary disciplines, $correct\_i$ is 1 if the answer is correct else 0, D_j represents questions in secondary discipline j, and S_k represents secondary disciplines in primary discipline k.

B.2 Detailed Results

B.2.1 Performance of models with question-level averaging on 14 languages

Please see Table 7.

B.2.2 Performance of models with secondary discipline averaging on 14 languages

Please see Table 8.

B.2.3 Base model experimental results robustness

In response to the extraordinarily good performance of the Qwen Base series of models on Chinese (since the Qwen Instruct series of models didn't achieve such a good performance), we conduct some experiments to discuss whether there's data leakage and whether the models benefit from the presence of some kind of option label preference in the Chinese question options. To evaluate the robustness of Qwen Base models and investigate potential option-order bias, we experimented by randomly shuffling the order of options (including option labels) in our benchmark. The results in Figure 5 reveal two important findings:

First, the performance generally decreases after shuffling (indicated by red bars in the figure), suggesting that the models may have developed certain preferences for option ordering during training. This implies potential data leakage or position bias in the learning process. However, the degradation is relatively modest, with most models showing less than 5% performance drop across all evaluation metrics.

Second, and notably, even with shuffled options, the larger Qwen Base models maintain strong performance, with the best model (Qwen2.5-72B) still achieving around 68% accuracy. This demonstrates both the robust capabilities of Qwen Base models

Group	Model	En	Fr	De	It	Pt	Ru	Es	Ar	Zh	Id	Ja	Ko	Th	Vi	Avg
								Base models								
	Qwen3-0.6B	31.0	38.7	39.2	37.4	38.9	38.7	38.4	37.2	24.8	40.1	41.0	40.3	38.6	39.4	37.4
	Gemma-3-1B-it	24.2	36.3	37.4	36.9	37.1	38.5	36.2	36.2	24.5	36.9	38.7	39.5	38.5	37.2	35.6
<7B	Qwen3-1.7B	39.8	41.6	43.9	42.0	42.0	43.3	41.3	42.5	29.1	44.0	41.9	43.4	41.5	41.9	41.3
	Gemma-3-4B-it	33.4	42.5	44.3	43.8	42.6	44.0	44.4	42.4	25.4	42.0	44.2	44.5	43.5	43.2	41.4
	Qwen3-4B	**47.7**	**46.5**	**49.7**	**47.9**	**46.2**	**48.8**	**48.9**	**46.0**	**36.8**	**49.3**	**46.2**	**48.3**	**47.1**	**48.3**	**47.0**
	Qwen2.5-7B-Instruct	38.0	39.1	39.4	39.2	38.2	39.1	39.3	38.6	31.8	39.6	38.5	37.7	38.4	38.5	38.2
	Aya-8B	30.5	40.4	39.2	40.9	37.0	40.5	39.0	39.2	25.0	39.6	42.1	41.1	34.4	41.1	37.9
	Llama3-8B-Instruct	32.9	39.9	38.6	38.4	38.3	37.2	38.7	41.3	24.8	40.9	41.2	38.9	39.3	38.6	37.8
	Qwen3-8B	48.5	46.1	49.7	49.6	48.0	49.0	47.9	46.8	40.6	49.0	48.1	49.7	47.6	48.9	47.8
7-14B	Gemma-3-12B-it	39.0	46.6	47.7	47.9	45.2	47.4	47.1	48.1	26.0	47.7	46.7	47.9	45.5	46.1	44.9
	Mistral-Nemo-Instruct	34.6	37.6	40.0	39.0	38.2	38.6	38.7	35.9	26.6	40.5	37.9	35.5	35.7	37.4	36.9
	Phi-4	50.6	**48.6**	**51.9**	**51.9**	47.9	51.7	**51.7**	45.5	28.1	48.7	48.5	50.0	47.1	48.7	47.9
	Qwen2.5-14B-Instruct	43.5	43.3	45.1	45.7	42.2	45.8	43.8	43.5	35.5	43.8	44.4	44.2	43.6	44.4	43.5
	Qwen3-14B	**53.4**	47.3	51.5	50.6	**49.9**	**51.8**	49.0	**49.9**	**45.4**	**51.3**	**49.3**	**50.9**	**47.8**	**52.0**	**50.0**
	GPT-oss-20B	53.6	50.0	50.9	51.8	49.4	50.5	51.2	49.3	34.9	50.5	49.1	50.1	48.1	47.6	49.1
	Mistral-Small-Instruct	38.4	45.3	45.8	46.8	42.5	43.6	43.9	39.6	25.7	43.5	44.5	42.4	38.2	41.5	41.6
	Magistral-Small-2507	46.0	43.9	45.6	45.0	45.1	43.5	43.1	41.9	31.3	44.1	41.3	40.8	39.6	41.2	42.3
	Gemma-3-27B-it	46.5	**52.4**	53.3	52.8	50.8	51.9	52.4	**51.5**	28.8	52.6	49.3	50.4	49.7	51.0	49.5
	Qwen3-30B-A3B-Instruct-2507	**57.0**	51.3	**53.7**	**53.9**	50.1	**53.2**	**53.4**	51.4	52.2	**55.3**	**52.9**	**53.4**	**52.1**	**53.9**	**53.1**
14-32B	Qwen3-30B-A3B-Thinking-2507	55.3	49.6	53.4	53.2	**51.0**	51.7	52.4	50.4	49.0	52.7	49.6	52.1	50.3	51.7	51.6
	Qwen3-30B-A3B	53.9	48.9	52.0	50.9	49.3	51.0	50.0	49.6	47.7	51.5	46.8	50.6	47.9	50.5	50.0
	Aya-32B	36.6	46.7	45.9	46.9	46.7	45.4	45.6	43.1	25.5	44.7	45.7	45.6	39.1	45.8	43.1
	QwQ-32B	52.7	48.9	50.8	50.8	47.9	48.9	49.0	46.8	**53.4**	47.5	46.0	47.3	46.3	47.5	48.8
	Qwen2.5-32B-Instruct	46.9	41.2	45.9	45.2	41.7	44.4	43.1	41.8	41.3	44.1	41.0	42.2	41.2	43.2	43.1
	Qwen3-32B	56.5	47.3	51.7	51.7	48.9	50.5	49.5	47.4	53.3	50.4	48.2	48.6	46.1	48.6	49.9
	Llama3-70B-Instruct	48.1	54.7	53.6	54.4	51.4	53.3	52.4	51.1	30.2	54.0	52.2	52.0	51.8	51.8	50.8
	Llama4-scout	53.9	**54.8**	54.7	55.3	54.6	54.2	55.4	54.1	39.4	54.6	51.4	51.3	51.0	52.4	52.7
	GPT-oss-120B	57.8	50.4	51.9	52.9	50.6	51.9	52.7	49.3	40.3	52.4	50.1	49.4	47.9	49.6	50.5
	Mistral-Large-Instruct	43.0	43.3	48.8	48.7	46.3	46.1	46.3	45.2	34.5	47.9	45.3	45.4	43.0	42.7	44.8
>32B	Qwen3-235B-A22B-Instruct-2507	**63.2**	54.6	56.4	57.0	54.3	**57.9**	**55.8**	**56.0**	**66.9**	**58.3**	54.4	**57.2**	**55.6**	**57.1**	**57.5**
	Qwen3-235B-A22B-Thinking-2507	56.6	51.8	54.6	54.5	51.3	52.1	53.1	52.0	56.1	54.4	50.9	54.0	52.2	53.5	53.4
	Qwen3-235B-A22B	56.3	50.4	53.2	52.7	50.4	53.2	51.9	51.8	57.5	52.0	49.9	52.4	49.6	51.0	52.3
	Llama4-maverick	60.4	54.4	**57.4**	**58.5**	**55.3**	57.4	55.7	54.1	49.0	57.0	**54.7**	55.4	53.3	54.8	55.5
	Deepseek-r1	62.9	49.2	52.6	53.3	49.8	51.8	52.8	49.0	64.9	51.0	47.9	49.1	48.5	50.7	52.4
	Deepseek-v3	59.1	49.0	51.3	51.6	49.4	50.3	50.9	49.3	56.1	51.1	47.6	47.7	47.2	49.4	50.7
	GPT-5	**66.2**	50.7	53.9	53.3	52.4	53.4	53.4	50.0	60.4	54.2	48.7	50.0	49.2	51.5	53.4
	Gemini-2.5-flash	62.8	51.9	54.2	55.1	51.8	51.9	55.1	53.1	53.5	54.5	51.5	49.9	51.2	50.9	53.4
	Gemini-2.5-pro	65.9	53.1	56.7	56.8	55.4	55.6	56.7	55.2	62.7	55.3	52.0	54.6	52.2	53.5	56.1
	Qwen3-max-preview	65.6	**58.9**	**60.0**	**60.6**	**58.3**	**60.8**	**59.2**	**59.4**	**70.4**	**61.9**	**56.4**	**60.1**	**58.8**	**59.2**	**60.7**
	ChatGPT-4o-latest	59.2	53.4	56.3	55.4	53.7	55.5	55.7	52.8	42.7	52.5	54.2	53.5	50.6	52.1	53.4
Close-sourced	Claude3.7-sonnet-thinking	61.1	50.8	51.5	53.2	50.6	52.2	51.9	52.7	48.2	53.4	47.2	48.7	49.3	50.6	51.5
	Claude3.7-sonnet	59.1	50.9	52.1	52.8	50.8	51.9	52.8	54.3	46.0	54.6	48.0	49.3	49.9	50.8	51.7
	Claude4-Sonnet-thinking	63.7	52.7	54.1	54.7	52.5	53.3	54.8	52.3	53.1	52.7	49.5	51.1	48.4	51.0	53.1
	Claude4-Sonnet	63.3	52.9	53.5	55.0	53.5	52.4	54.3	50.9	52.8	53.2	50.4	53.3	49.7	51.3	53.3
	GPT-4.1	60.9	52.6	55.4	54.6	51.8	54.1	53.5	52.5	44.0	52.6	51.9	53.1	49.2	51.3	52.7
	Grok-3	61.2	52.4	53.7	54.7	51.9	53.4	54.4	51.8	46.3	55.6	51.9	52.2	51.1	51.1	53.0
								Base models								
	Qwen3-0.6B-Base	36.9	46.5	44.5	43.8	47.5	44.1	44.2	47.6	29.0	44.2	45.3	46.5	43.8	46.1	43.6
	Gemma-3-1B-pt	25.3	21.2	21.9	19.9	20.2	19.8	28.1	20.6	25.3	23.3	20.5	24.8	22.6	20.4	22.4
<7B	Qwen3-1.7B-Base	39.4	48.4	46.9	47.3	48.5	47.6	46.9	52.2	32.7	49.2	48.4	50.5	49.1	49.6	46.9
	Gemma-3-4B-pt	36.3	42.0	40.7	41.2	42.3	39.3	41.4	43.3	26.3	44.7	40.9	39.0	41.2	41.8	40.0
	Qwen3-4B-Base	**46.1**	**54.9**	**56.1**	**56.3**	**56.7**	**58.0**	**55.6**	**58.3**	**41.9**	**57.7**	**55.5**	**57.9**	**55.8**	**57.6**	**54.9**
	Qwen2.5-7B	43.5	47.9	45.6	47.1	48.0	45.0	46.8	47.6	52.2	47.7	45.3	45.1	42.4	45.8	46.4
	Meta-Llama-3-8B	36.0	37.8	38.2	37.2	37.5	35.6	38.2	39.5	25.4	39.5	38.3	38.8	38.5	39.1	37.1
7-14B	Qwen3-8B-Base	49.4	57.3	57.9	57.5	58.8	58.3	56.7	58.2	48.7	58.3	56.4	58.2	57.4	59.6	56.6
	Gemma-3-12B-pt	42.6	50.7	48.7	49.5	49.6	47.5	48.5	50.6	28.7	51.9	48.6	49.3	48.1	47.2	47.3
	Qwen2.5-14B	46.7	48.8	49.0	48.8	49.2	47.6	48.5	50.7	**58.1**	49.3	45.8	47.6	47.5	48.5	49.0
	Qwen3-14B-Base	**52.5**	**61.6**	**61.5**	**60.8**	**63.1**	**60.7**	**60.9**	**61.5**	57.5	**63.1**	**60.3**	**62.9**	**60.4**	**63.9**	**60.8**
	Gemma-3-27B-pt	47.0	53.6	52.7	52.6	53.2	50.2	53.7	53.6	31.1	54.0	49.9	52.9	51.1	52.0	50.5
14-32B	Qwen3-30B-A3B-Base	47.1	**57.8**	**58.9**	**59.1**	**58.8**	**58.0**	**58.5**	**58.7**	54.6	**61.0**	**56.3**	**59.6**	**56.7**	**58.4**	**57.4**
	Qwen2.5-32B	**51.4**	52.4	53.6	53.5	53.7	53.1	54.0	52.5	**69.2**	55.3	51.4	51.5	51.0	52.1	53.9
>32B	Meta-Llama-3-70B	44.7	**51.9**	51.7	51.9	52.0	50.7	**52.3**	49.8	30.2	53.9	49.2	48.8	47.9	49.7	48.9
	Qwen2.5-72B	**52.7**	**51.9**	**53.3**	**52.3**	**53.1**	**51.8**	**52.3**	**54.1**	**74.3**	**55.5**	**51.2**	**53.0**	**51.2**	**53.0**	**54.3**

Table 7: Performance averaged by **questions** comparison across 14 languages for different chat/reasoning/base model sizes. The best model in a column within each size interval we mark in **bold**.

Group	Model	En	Fr	De	It	Pt	Ru	Es	Ar	Zh	Id	Ja	Ko	Th	Vi	Avg
	Chat & Reasoning Models															
	Qwen3-0.6B	31.8	39.1	39.8	38.0	39.4	38.3	38.8	37.0	24.7	39.8	40.7	40.8	38.5	39.9	37.6
	Gemma-3-1B-it	24.9	36.4	38.1	37.2	37.9	38.9	35.9	36.6	24.3	37.1	39.0	39.6	38.3	37.6	35.8
<7B	Qwen3-1.7B	40.1	42.2	43.8	42.2	43.7	43.6	42.3	42.5	29.3	44.0	42.6	43.8	41.7	41.5	41.7
	Gemma-3-4B-it	33.4	43.2	44.6	44.7	43.9	44.2	45.0	42.2	25.6	41.5	44.1	44.4	43.6	43.6	41.7
	Qwen3-4B	**47.5**	**47.6**	**49.5**	**48.4**	**47.2**	**49.1**	**49.9**	**46.4**	**36.7**	**48.8**	**46.5**	**48.3**	**47.1**	**48.5**	**47.2**
	Qwen2.5-7B-Instruct	37.9	39.8	40.2	40.0	38.7	39.2	39.3	39.1	31.9	40.6	38.5	37.7	38.5	38.8	38.6
	Aya-8B	31.0	41.1	39.9	40.7	37.8	40.9	40.0	39.3	24.9	39.1	42.1	41.1	34.5	40.9	38.1
	Llama3-8B-Instruct	32.9	40.1	38.8	38.8	39.3	38.1	40.0	41.0	24.9	41.5	40.7	38.8	39.6	39.4	38.1
	Qwen3-8B	49.5	47.7	50.4	50.1	48.1	49.1	48.6	46.5	40.9	49.0	47.9	49.2	47.7	48.7	48.1
7-14B	Gemma-3-12B-it	39.4	47.6	47.8	49.9	45.9	47.6	47.2	48.4	25.8	48.4	46.4	47.8	45.5	46.6	45.3
	Mistral-Nemo-Instruct	34.8	39.0	40.5	39.0	38.9	39.4	38.7	36.6	26.6	41.4	38.1	35.2	36.0	37.6	37.3
	Phi-4	50.4	**49.6**	**52.0**	**52.6**	48.0	**52.3**	**51.0**	45.4	28.4	49.4	48.1	50.0	47.4	48.9	48.1
	Qwen2.5-14B-Instruct	42.7	44.3	45.4	45.8	43.2	46.3	43.7	43.1	35.8	43.0	44.5	44.1	43.8	44.4	43.6
	Qwen3-14B	**54.0**	48.8	51.6	52.2	**50.1**	51.4	49.1	**49.9**	**45.5**	**51.2**	**49.2**	**50.6**	**47.8**	**51.9**	**50.2**
	GPT-oss-20B	53.1	51.1	51.4	52.0	49.7	50.5	52.1	50.0	35.0	51.0	49.1	50.0	48.4	47.6	49.4
	Mistral-Small-Instruct	39.5	46.5	46.1	47.5	43.1	44.6	45.2	39.8	25.7	44.0	44.3	41.9	38.4	41.7	42.0
	Magistral-Small-2507	46.1	45.1	46.1	45.5	45.4	44.2	43.5	42.3	31.7	44.1	41.5	41.1	40.3	41.4	42.7
	Gemma-3-27B-it	46.3	**53.6**	53.7	54.0	**51.5**	52.3	**53.1**	**51.8**	28.8	53.0	49.0	49.9	49.9	51.4	49.9
	Qwen3-30B-A3B-Instruct-2507	**57.5**	52.2	53.7	**54.3**	50.4	**53.7**	52.6	51.3	52.3	**56.5**	**52.9**	**53.2**	**52.1**	**53.9**	**53.3**
14-32B	Qwen3-30B-A3B-Thinking-2507	55.7	50.5	**54.0**	53.5	51.4	52.2	52.2	50.1	49.1	52.8	49.5	51.7	50.3	51.6	51.8
	Qwen3-30B-A3B	54.3	49.8	52.5	50.9	48.4	51.2	50.2	49.7	47.7	52.2	47.2	50.6	47.9	50.4	50.2
	Aya-32B	37.0	47.8	46.1	47.0	47.8	46.4	45.3	44.1	25.5	44.5	45.4	45.4	39.1	45.9	43.4
	QwQ-32B	53.3	50.1	51.3	52.0	48.3	50.0	49.1	46.8	**53.7**	47.2	45.6	47.2	46.6	47.4	49.2
	Qwen2.5-32B-Instruct	47.1	42.1	46.4	45.9	42.3	45.5	43.3	42.5	41.8	44.2	40.5	41.9	41.7	43.3	43.5
	Qwen3-32B	57.0	48.3	52.4	52.0	50.1	51.0	49.1	47.4	53.2	50.8	48.1	48.2	46.3	48.1	50.2
	Llama3-70B-Instruct	48.4	54.9	53.8	55.0	50.8	54.0	52.9	51.5	30.2	54.8	52.1	51.8	51.8	52.2	51.0
	Llama4-scout	54.3	55.4	55.2	56.3	54.4	54.9	55.3	53.8	39.8	55.0	51.7	51.2	50.9	52.2	52.9
	GPT-oss-120B	57.7	51.9	51.8	52.7	50.3	52.9	52.6	49.2	40.4	53.3	50.1	49.0	48.0	49.7	50.7
	Mistral-Large-Instruct	43.3	44.2	48.6	48.2	46.1	47.0	46.7	45.7	33.7	46.4	44.4	44.0	42.5	44.3	44.7
	Qwen3-235B-A22B-Instruct-2507	**63.1**	**55.8**	56.6	57.9	53.9	**58.5**	55.5	**56.6**	**67.0**	**59.0**	54.0	**57.1**	**55.6**	**56.6**	**57.7**
>32B	Qwen3-235B-A22B-Thinking-2507	57.4	52.6	54.7	55.1	51.9	52.4	53.0	52.7	56.2	54.3	50.6	53.8	52.4	52.8	53.6
	Qwen3-235B-A22B	55.9	51.2	53.6	53.4	50.5	54.2	51.7	52.6	57.7	52.1	50.5	52.3	49.7	50.9	52.6
	Llama4-maverick	60.6	55.0	**57.9**	**59.2**	**55.6**	57.8	**55.6**	54.5	49.0	57.4	**54.2**	54.9	53.3	55.0	55.7
	Deepseek-r1	62.9	50.7	52.6	53.9	50.3	52.7	53.6	49.0	65.0	51.4	47.6	49.0	48.6	50.5	52.7
	Deepseek-v3	58.6	50.4	51.3	52.4	49.2	51.1	51.0	49.6	56.3	50.9	47.7	47.4	47.1	49.3	50.9
	GPT-5	**65.6**	51.8	53.8	53.6	52.6	53.5	53.4	50.6	60.3	55.5	48.5	49.7	49.6	51.9	53.6
	Gemini-2.5-flash	62.7	53.0	54.7	55.8	52.4	52.5	55.0	53.6	53.3	55.3	51.2	49.7	51.5	50.9	53.7
	Gemini-2.5-pro	65.6	53.9	57.6	57.1	55.3	55.9	56.8	55.7	62.6	55.6	52.3	54.4	52.0	53.7	56.3
	Qwen3-max-preview	65.3	**60.3**	**60.1**	**60.4**	**58.1**	**61.2**	**59.2**	**59.5**	**70.5**	**62.2**	**56.1**	**59.7**	**58.9**	**59.1**	**60.7**
	ChatGPT-4o-latest	59.0	54.4	56.2	55.6	53.8	55.6	55.5	53.2	42.9	52.8	53.9	53.2	50.6	52.0	53.5
Close-sourced	Claude3.7-sonnet-thinking	61.4	52.0	52.0	53.4	50.4	53.1	52.4	53.3	48.1	53.7	47.1	48.2	49.2	50.5	51.8
	Claude3.7-sonnet	58.8	52.2	52.0	53.4	50.5	52.6	53.3	54.5	46.2	54.9	47.5	48.7	49.9	50.8	51.8
	Claude4-Sonnet-thinking	63.5	53.7	54.5	55.0	52.5	53.4	55.5	52.8	53.1	53.0	49.6	50.6	48.7	50.9	53.4
	Claude4-Sonnet	63.0	54.0	53.2	56.0	53.4	52.9	54.9	51.3	53.2	53.0	50.3	53.0	49.8	51.1	53.5
	GPT-4.1	60.7	53.3	54.9	54.2	51.9	53.7	53.4	53.1	44.1	53.1	51.7	52.6	49.1	51.4	52.7
	Grok-3	60.9	53.4	53.5	55.1	52.8	54.2	54.9	51.8	46.4	56.5	51.9	51.9	51.3	51.1	53.3
	Base models															
	Qwen3-0.6B-Base	36.8	47.0	44.7	44.7	47.8	44.4	45.3	47.5	29.0	43.9	45.6	46.9	43.9	46.3	43.8
	Gemma-3-1B-pt	25.9	21.2	21.6	19.1	19.5	19.6	27.3	20.4	25.8	23.7	20.4	24.7	22.3	20.6	22.3
<7B	Qwen3-1.7B-Base	40.6	49.2	47.6	48.2	48.7	48.0	47.9	51.5	33.1	49.0	48.4	50.3	49.1	49.7	47.2
	Gemma-3-4B-pt	36.6	43.1	41.7	41.7	43.6	39.8	41.9	43.6	26.5	45.3	41.4	39.3	41.6	41.8	40.6
	Qwen3-4B-Base	**46.6**	**56.0**	**57.0**	**57.2**	**57.5**	**57.6**	**56.0**	**57.8**	**42.4**	**57.6**	**55.3**	**57.7**	**56.2**	**57.3**	**55.2**
	Qwen2.5-7B	44.1	48.6	46.8	47.0	49.3	45.3	48.7	48.5	52.3	47.5	45.4	45.0	42.7	46.0	46.9
	Meta-Llama-3-8B	35.7	38.0	39.1	37.9	38.8	36.2	39.0	39.4	25.8	40.1	38.4	38.8	38.6	39.0	37.5
7-14B	Qwen3-8B-Base	49.7	58.8	58.7	58.1	59.4	58.5	57.3	58.0	49.3	57.9	56.2	58.1	57.3	59.5	56.9
	Gemma-3-12B-pt	43.4	51.7	50.2	50.5	50.7	47.4	49.6	50.6	28.9	52.0	48.5	48.8	48.4	47.3	47.7
	Qwen2.5-14B	47.4	49.1	49.9	49.6	50.6	48.0	49.1	50.9	**58.4**	49.1	46.2	47.7	47.7	48.4	49.4
	Qwen3-14B-Base	**53.1**	**62.2**	**61.4**	**61.8**	**63.4**	**60.1**	**61.1**	**60.8**	57.7	**63.1**	**60.3**	**62.7**	**60.4**	**63.4**	**60.8**
	Gemma-3-27B-pt	47.4	54.8	54.3	53.9	54.4	50.4	54.6	54.1	31.2	53.6	50.2	52.9	51.5	51.5	51.1
14-32B	Qwen3-30B-A3B-Base	47.4	**58.2**	**59.2**	**59.5**	**58.7**	**57.8**	**58.8**	**58.4**	54.2	**60.9**	**56.1**	**59.2**	**56.3**	**58.0**	**57.3**
	Qwen2.5-32B	**52.3**	52.9	54.5	54.4	55.3	53.6	54.7	52.6	**69.3**	55.6	51.5	51.4	51.4	51.7	54.4
>32B	Meta-Llama-3-70B	45.4	**53.1**	52.0	**53.4**	52.4	51.3	52.8	50.1	30.6	54.8	49.3	48.7	47.8	49.7	49.4
	Qwen2.5-72B	**53.2**	52.0	**54.3**	52.5	**53.7**	**51.5**	**53.1**	**54.5**	**74.0**	**55.7**	**51.8**	**52.9**	**51.4**	**52.7**	**54.5**

Table 8: Performance averaged by **secondary** disciplines comparison across 14 languages for different chat/reasoning/base model sizes. The best model in a column within each size interval we mark in **bold**.

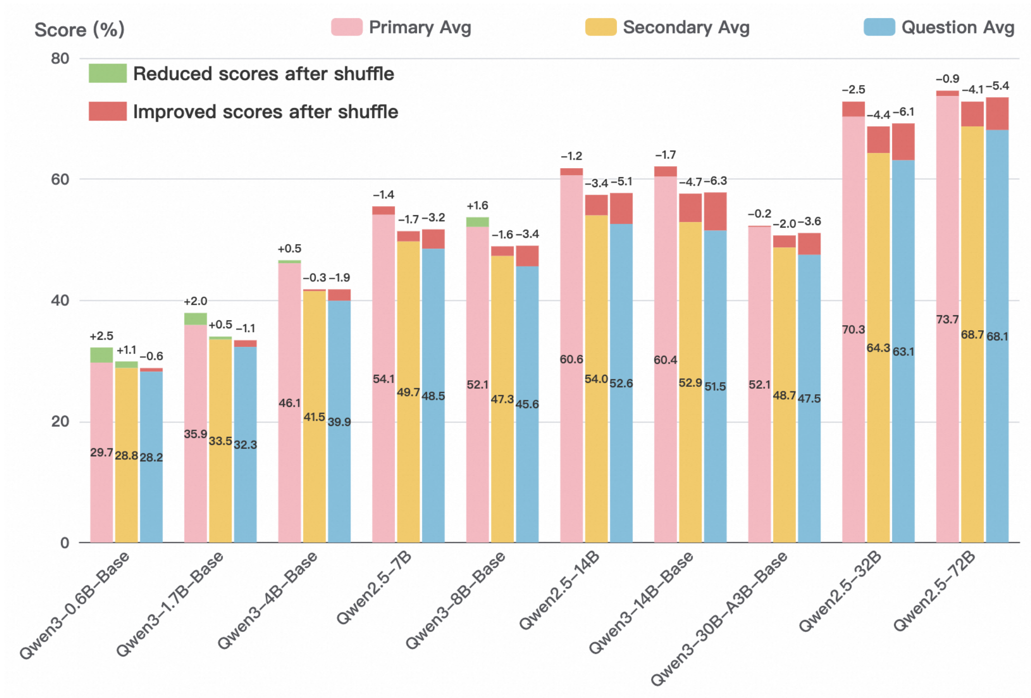

Figure 5: Performance comparison before and after option shuffling across different LLMs. The bars show the average scores by primary disciplines (pink), secondary disciplines (yellow), and questions (blue). Green bars indicate score reduction after shuffling, while red bars show score improvement after shuffling. The numbers inside the bars represent the lower score between pre- and post-shuffle results, while the numbers above the bars indicate the change in performance.

in Chinese language understanding and the reliability of our experimental results. The minimal impact of option shuffling on overall performance validates the fundamental strength of these models in handling Chinese multiple-choice questions.

B.2.4 Inconsistency of Linguistic Competence across Different Language Families

While our analysis reveals strong cross-lingual consistency among Indo-European languages, we observe significant performance disparities across different language families, as shown in Figure 6. Unlike the consistent hexagonal patterns seen within Indo-European languages, the performance of languages across different language families (such as Arabic from Afro-Asiatic, Thai from Kra-Dai, and Korean from Koreanic) shows irregular patterns and notable performance drops in Figure 6. While Claude4-Sonnet excels in English, it shows notably weaker performance in Chinese, Indonesian, and Arabic. While Qwen3-235B-A22B performs well in Chinese, its performance is mediocre in all other languages in the figure. This inconsistency suggests that the advantages derived from shared linguistic features and cultural proximity in Indo-European languages do not extend to other

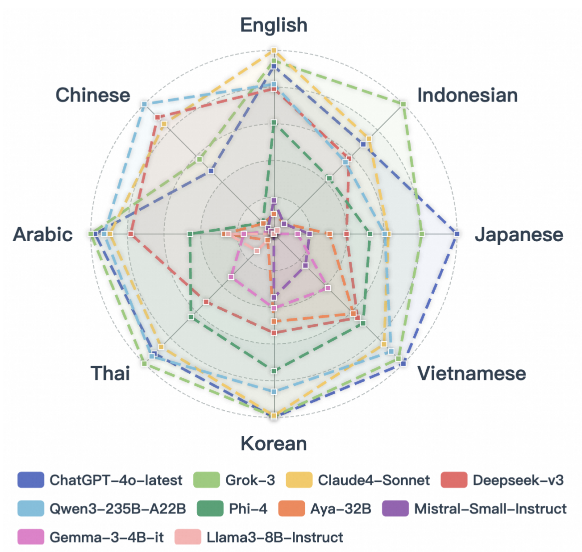

Figure 6: LLMs' performance across different language families. Models from various families and sizes are sampled to ensure generalisability. Scores per language are normalised between 0 (minimum) and 1 (maximum).

language families.

B.2.5 Imbalance of model performance in disciplines

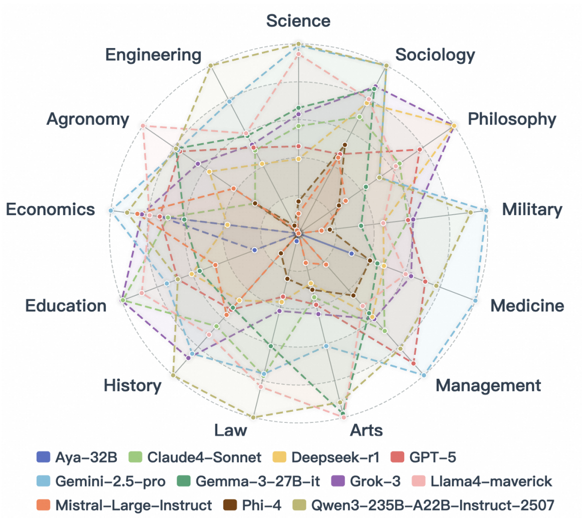

Figure 7: LLMs' performance in **Arabic** across different disciplines. Models with the highest scores from various LLM families (see Table 2) are selected. Scores per discipline are normalised to 0 (minimum) and 1 (maximum).

For Arabic (Figure 7), Indonesian (Figure 8), and Spanish (Figure 9), Portuguese(Figure 10), Russian(Figure 11), German(Figure 12), French(Figure 13), Italian(Figure 14),Japanese(Figure 15), .we can find that the pattern of their radargrams is similar to the performance on English. For example, the results in Arabic demonstrate similar performance imbalances across disciplines as observed in English. Figure 7 reveals that while Grok-3 achieves strong performance in Philosophy and Sociology, it shows notable weaknesses in Engineering-related fields. Similarly, Qwen3-235B-A22B-Instruct-2507 exhibits excellence in Science and Engineering but underperforms in Medicine and Management. This uneven performance across academic domains, observed consistently across different language models, reinforces our finding that current LLMs face fundamental challenges in achieving truly balanced capabilities across diverse fields of knowledge in these languages.

On the other hand, the pattern of radargrams on Chinese (Figure 16), Korean (Figure 17), Thai (Figure 18), and Vietnamese (Figure 19) is quite different. The most distinct ones are Thai and Chinese.

The Vietnamese results present an intriguing pat-

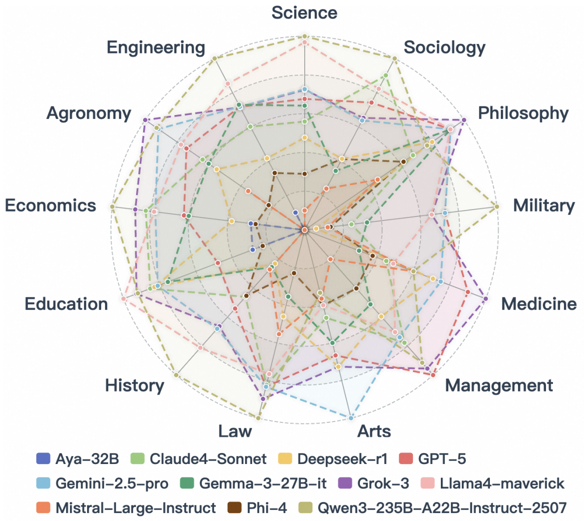

Figure 8: LLMs' performance in **Indonesian** across different disciplines. Models with the highest scores from various LLM families (see Table 2) are selected. Scores per discipline are normalised to 0 (minimum) and 1 (maximum).

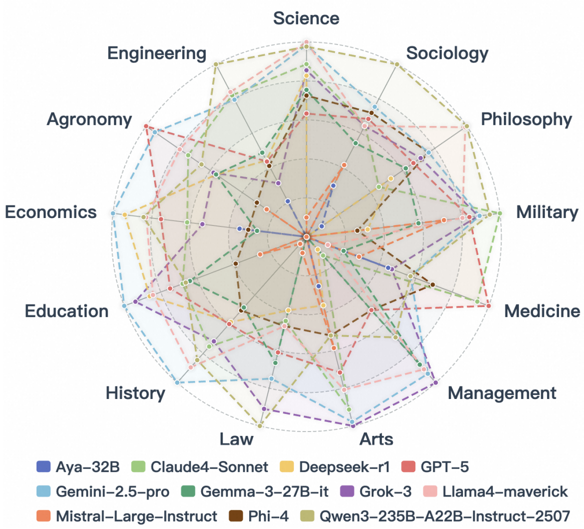

Figure 9: LLMs' performance in **Spanish** across different disciplines. Models with the highest scores from various LLM families (see Table 2) are selected. Scores per discipline are normalised to 0 (minimum) and 1 (maximum).

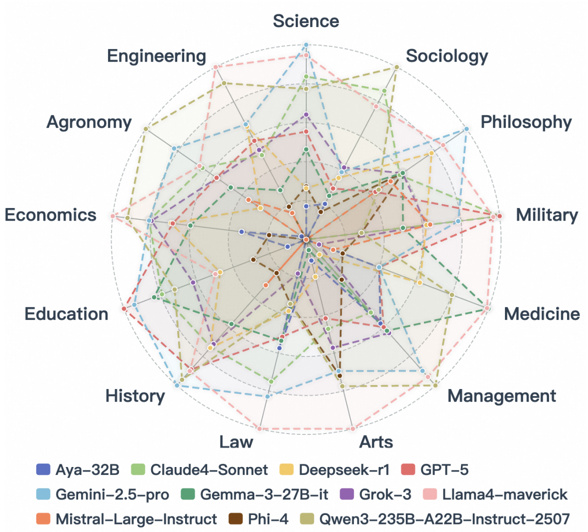

Figure 10: LLMs' performance in **Portuguese** across different disciplines. Models with the highest scores from various LLM families (see Table 2) are selected. Scores per discipline are normalised to 0 (minimum) and 1 (maximum).

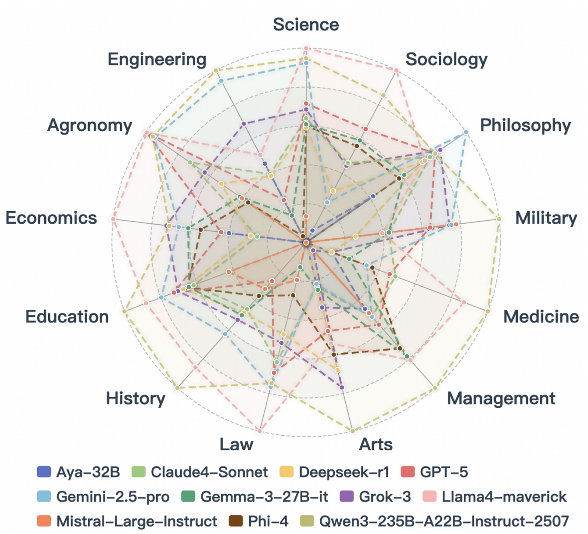

Figure 11: LLMs' performance in **Russian** across different disciplines. Models with the highest scores from various LLM families (see Table 2) are selected. Scores per discipline are normalised to 0 (minimum) and 1 (maximum).

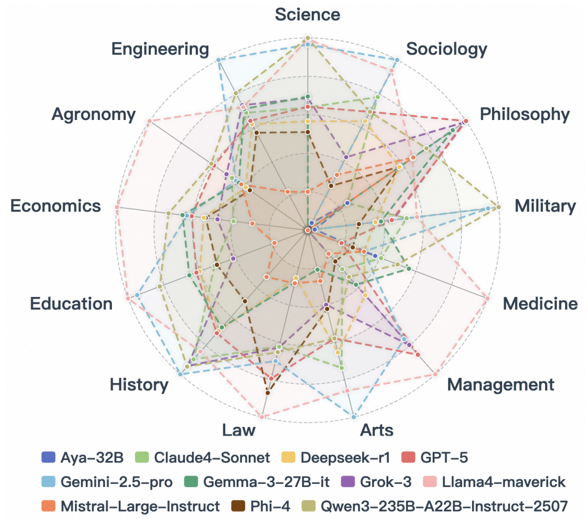

Figure 12: LLMs' performance in **German** across different disciplines. Models with the highest scores from various LLM families (see Table 2) are selected. Scores per discipline are normalised to 0 (minimum) and 1 (maximum).

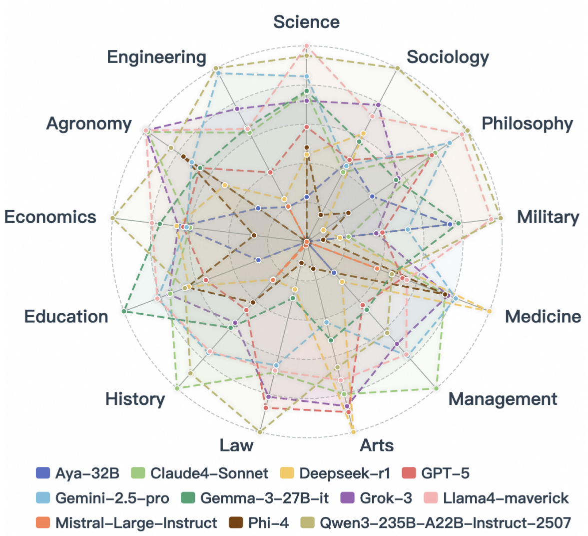

Figure 13: LLMs' scores on **French** in different disciplines. We sampled relatively strong models from different LLM families to ensure generalisability. For each discipline, we normalised the scores with a minimum score of 0 and a maximum score of 1.

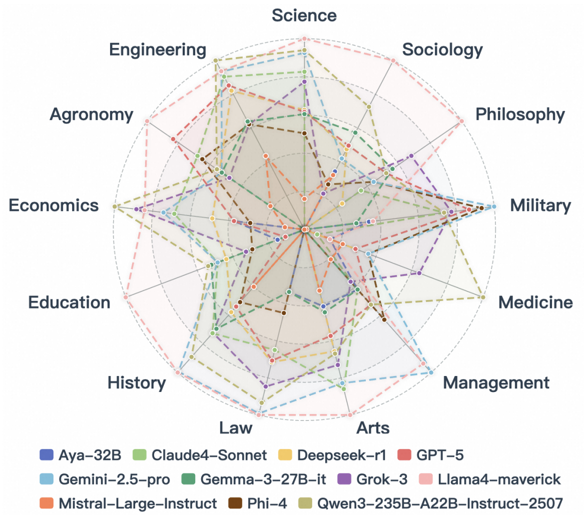

Figure 14: LLMs' performance in **Indonesian** across different disciplines. Models with the highest scores from various LLM families (see Table 2) are selected. Scores per discipline are normalised to 0 (minimum) and 1 (maximum).

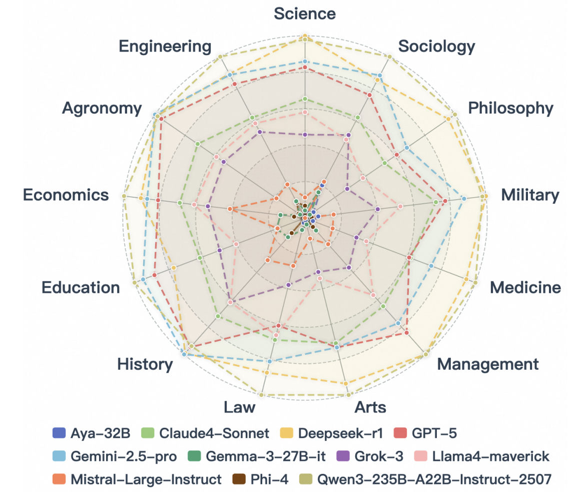

Figure 16: LLMs' performance in **Chinese** across different disciplines. Models with the highest scores from various LLM families (see Table 2) are selected. Scores per discipline are normalised to 0 (minimum) and 1 (maximum).

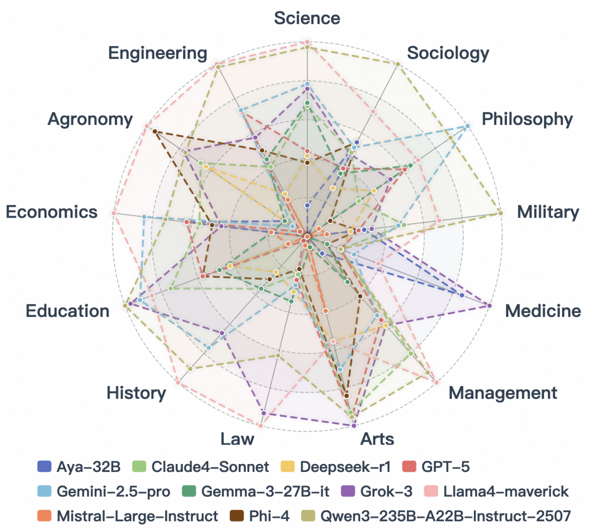

Figure 15: LLMs' performance in **Japanese** across different disciplines. Models with the highest scores from various LLM families (see Table 2) are selected. Scores per discipline are normalised to 0 (minimum) and 1 (maximum).

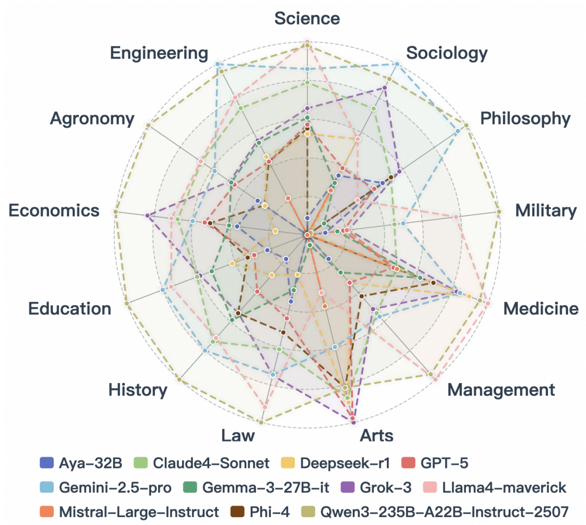

Figure 17: LLMs' performance in **Korean** across different disciplines. Models with the highest scores from various LLM families (see Table 2) are selected. Scores per discipline are normalised to 0 (minimum) and 1 (maximum).

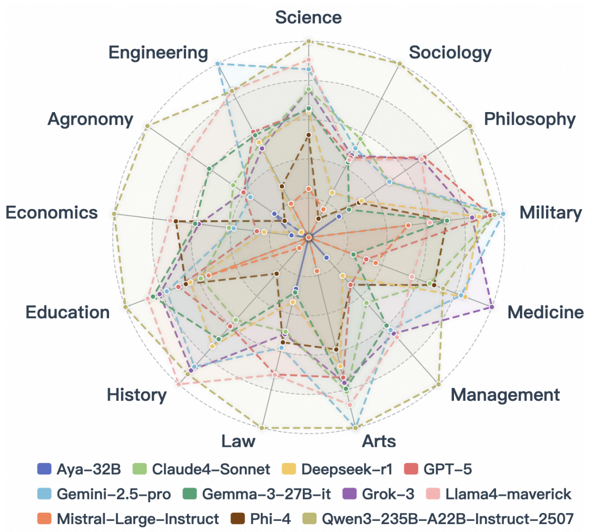

Figure 18: LLMs' performance in **Thai** across different disciplines. Models with the highest scores from various LLM families (see Table 2) are selected. Scores per discipline are normalised to 0 (minimum) and 1 (maximum).

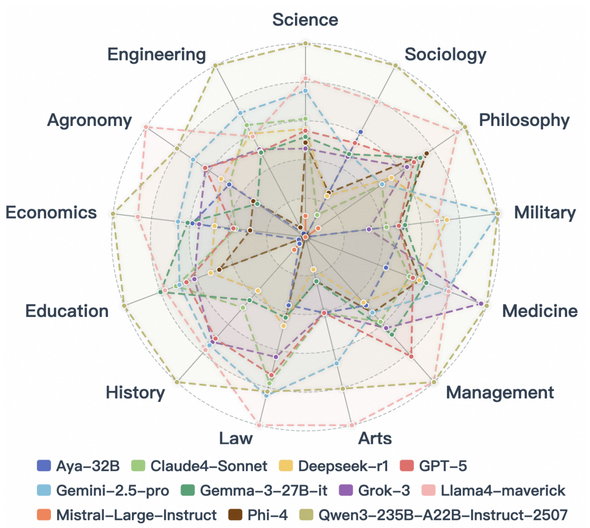

Figure 19: LLMs' performance in **Vietnamese** across different disciplines. Models with the highest scores from various LLM families (see Table 2) are selected. Scores per discipline are normalised to 0 (minimum) and 1 (maximum).

tern of performance consistency across disciplines. Figure 19 shows that Qwen3-235B-A22B-Instruct-2507, in particular, demonstrates remarkably balanced capabilities across diverse fields, from Science and Engineering to Sociology and Economics, suggesting effective training optimization for Vietnamese language understanding. While the characteristic discipline-specific variations are still present, they are less notable compared to other languages. Moreover, we find that Qwen3-235B-A22B-Instruct-2507's performance is also excellent in Chinese, Korean, Thai, etc. These results may reflect Qwen3's strong foundation and excellent targeted optimisation in multilingualism, though subtle performance variations in specialized disciplines indicate that achieving perfect cross-discipline uniformity remains challenging even with well-optimized language models.

The Chinese results reveal a distinctive pattern that sets them apart from other languages, particularly in the exceptional performance of models specifically optimized for Chinese language processing. Figure 16 demonstrates that Qwen3-235B-A22B-Instruct-2507 exhibits remarkable capabilities across multiple disciplines. Most strikingly, Qwen3-235B-A22B-Instruct-2507 achieves unprecedented balance across traditionally disparate fields, showing strong performance not only in STEM disciplines but also in disciplines like Economics and History. This pattern suggests effective domain adaptation in Chinese language models, likely benefiting from extensive pretraining on Chinese academic resources. The superior performance in Chinese of Chinese-optimized models in our benchmark underscores the importance of language-specific optimization in developing truly capable language models. It also highlights the potential for achieving more balanced cross-disciplinary performance through targeted model development.

C Detailed information about languages

Please see Table 9 for detailed information on languages.

D Full list of disciplines

A complete list of the three levels of disciplines we use is demonstrated in Table 10.

Code	Full Name	Language Family	Speakers (M)
en	English	Indo-European	1,500
zh	Chinese	Sino-Tibetan	1,400
es	Spanish	Indo-European	595
ar	Arabic	Afro-Asiatic	400
fr	French	Indo-European	300
pt	Portuguese	Indo-European	270
ru	Russian	Indo-European	260
id	Indonesian	Austronesian	200
de	German	Indo-European	135
ja	Japanese	Japonic	130
vi	Vietnamese	Austroasiatic	86
it	Italian	Indo-European	85
ko	Korean	Koreanic	80
th	Thai	Kra-Dai	80
	Total: 5,521 (∼69% of total world population)		

Table 9: Detailed information on the languages used in our articles and population statistics. The data are partially sourced from https://www.ethnologue.com and Wikipedia.

E Detailed question distribution in discipline and language

Please see the Table 11, 12, 13.

F Detailed distribution of Cognitive Requirements for questions in each language and discipline

Detailed distribution of Cognitive Requirements for questions in each language and discipline is shown in the Table 14, 15, 16. The distribution of reasoning questions across languages and disciplines reveals interesting patterns in our benchmark's cognitive requirements. The benchmark maintains a balanced combination of reasoning and recitation-based questions. Due to our strict difficulty control strategy, the questions are more difficult, so we can clearly see a higher proportion of overall reasoning questions. This balance is particularly evident in fundamental disciplines like Mathematics, Physics, and Economics, where deep conceptual understanding requires both analytical reasoning and factual knowledge. The proportion of reasoning questions derived from these basic disciplines is even higher in applied disciplines, such as Applied Economics, Chemical Engineering, and Technology, because usually in real-life applications, we not only need to master the discipline knowledge, but also to adapt to real-life scenarios.

Notable variations exist across different disciplines and languages. STEM fields generally show a higher proportion of reasoning questions across most languages, particularly in disciplines like Mathematics, Engineering, and Computer Science. However, some specialized fields like Chemistry and Clinical Medicine maintain a lower ratio of reasoning questions (<30% in many languages), reflecting the importance of factual medical and chemical knowledge. The humanities and social sciences display mixed patterns. Unlike our stereotypes, the percentage of reasoning questions in disciplines in the humanities and social sciences, such as Sociology and Philosophy, is also very high. Today's humanities and social sciences are more than just memorisation, many of them require comprehensive analysis, which people may overlook.

Language-wise variations also emerge, with Chinese and English versions generally maintaining higher proportions of reasoning questions compared to other languages. We speculate that there are two reasons for this. One is that it is relatively easier to collect questions in Chinese and English than in other languages, and the number of questions collected is also relatively large, so the puzzles we screened out are more biased in favour of reasoning questions. On the other hand, most of the models are optimised for English first, so English knowledge type questions are not considered difficult for the models, which also leads to a climb in the proportion of reasoning questions in English. We use more Qwen-family models for filtering difficulty, and Qwen-family models are better optimised for Chinese, which may also lead to a higher proportion of reasoning questions in Chinese, as the general Chinese questions are not defined as 'difficult' by us.

Although the cognitive requirement of the questions is carefully analysed in this paper. However, we do not review the reasoning and recitation questions in the benchmark separately in this article, because the number of reasoning or recitation questions on a given discipline in a given language may not reach statistical significance. This work will be left for future work.

Primary Disciplines	Secondary & Third Disciplines
Engineering	**Weapon Science and Technology**: Military Chemistry and Pyrotechnics; Weapon Systems Science and Engineering **Mechanics**: Fundamentals of Dynamics and Control; Rigid Body Mechanics; Solid Mechanics; Theoretical Fluid Mechanics; Theoretical Mechanics **Petroleum and Natural Gas Engineering**: Poromechanics and Reservoir Physics; Oil and Gas Field Development and Storage & Transportation Engineering **Civil Engineering**: Geotechnical Engineering; Urban Infrastructure Engineering; Structural Engineering; Bridge and Tunnel Engineering **Food Science and Engineering**: Food Biochemistry; Food Processing and Storage Engineering **Surveying and Mapping Science and Technology**: Geodesy and Surveying Engineering; Digital Surveying and Remote Sensing Applications; Cartography and Geographic Information Engineering **Metallurgical Engineering**: Non-ferrous Metallurgy; Physical Chemistry of Metallurgical Process; Principles of Metallurgy; Iron and Steel Metallurgy **Hydraulic Engineering**: Hydraulics and Hydrology; Water conservancy and Hydropower Engineering **Computer Science and Technology**: Computer Architecture; Computer Networks; Operating Systems; Pattern Recognition; Advanced Programming Languages; Databases; Formal Languages; Principles of Computer Organization; Computer Software and Theory; Data Structures **Optical Engineering**: Optoelectronic Technology; Laser Technology; Theoretical Optics; Applied Optics **Electrical Engineering**: Power Systems and Automation; Power Electronics and Electrical Drives; High Voltage and Insulation Technology; Electrical Theory and New Technologies **Electronic Science and Technology**: Microelectronics and Solid-State Electronics; Electromagnetic Field and Microwave Technology; Circuits and Systems **Information and Communication Engineering**: Optical Fiber Communication; Communication and Information Systems; Antenna and Radio Communication; Communication Principles; Signal and Information Processing **Transportation Engineering**: Traffic Information Engineering and Control; Vehicle Operation Engineering; Transportation Planning and Management; Road and Railway Engineering **Power Engineering and Engineering Thermophysics**: Power Machinery and Engineering; Refrigeration and Cryogenic Engineering; Fluid Machinery and Engineering; Engineering Thermophysics; Heat Transfer; Internal Combustion Engineering; Engineering Fluid Mechanics; Thermal Energy Engineering **Materials Science and Engineering**: Materials Processing Engineering; Materials Physics and Chemistry **Environmental Science and Engineering**: Environmental Engineering; Environmental Science; Environmental and Resource Protection **Chemical Engineering and Technology**: Mass Transport and Separation Process in Chemical Engineering; Fluid Flow and Heat Transfer in Chemical Engineering; Chemical Transport Engineering; Elements of Chemical Reaction Engineering **Mechanical Engineering**: Manufacturing Automation; Mechatronic Engineering **Architecture**: Architectural Design and Theory; Architectural History; Urban Planning and Design **Nuclear Science and Technology**: Radiation Protection and Nuclear Technology Applications; Nuclear Energy and Reactor Technology **Control Science and Engineering**: Guidance, Navigation and Control; Operations Research and Cybernetics; Detection Technology and Automatic Equipment; Control Theory and Control Engineering **Instrument Science and Technology**: Instrument Science and Technology **Geological Resources and Geological Engineering**: Geological Resources and Geological Engineering **Textile Science and Engineering**: Textile Engineering; Textile Chemistry and Dyeing Engineering; Textile Materials Science **Naval Architecture and Ocean Engineering**: Ship Mechanics and Design Principles; Marine Engineering **Aeronautical & Astronautical Science & Technology**: Aeronautical & Astronautical Science & Technology
Science	**Chemistry**: Radiochemistry; Inorganic Chemistry; Analytical Chemistry; Electrochemistry; Organic Chemistry; Polymer Chemistry and Physics; Physical Chemistry **Mathematics**: Functions of Complex Variables; Fundamental Mathematics; Discrete Mathematics; Numerical Analysis; Cryptography; Ordinary Differential Equations; Number Theory; Polynomials and Series Expansions; Functions of Real Variables; Fuzzy Mathematics; Computational Mathematics; Combinatorial Mathematics; Stochastic Processes; Advanced Algebra; Mathematical Analysis; Probability and Statistics; Group Theory; Geometry and Topology; Graph Theory; Special Number Theory **Physics**: Relativity; Thermodynamics; Quantum Mechanics; Solid State Physics; Particle and Nuclear Physics; Polymer Physics; Thermodynamics and Statistical Physics; Acoustics; Subatomic and Atomic Physics; Atomic and Molecular Physics; Statistical Mechanics; Semiconductor Physics; Electrodynamics; Fluid Physics **Atmospheric Science**: Meteorology; Atmospheric Physics and Atmospheric Environment; Weather Dynamics **Biology**: Microbiology; Genetics; Cell Biology; Biophysics; Ecology; Biochemistry and Molecular Biology; Physiology; Zoology; Botany **Geography**: Human Geography; Physical Geography **Astronomy**: Astronomical Observation and Technology; Astrophysics; Stellar and Interstellar Evolution; Cosmology; Solar System Science **Physical Oceanography**: Physical Oceanography
Law	**Legal Studies**: Contract Law; Civil and Commercial Law; Criminal Law; Procedural Law; International Law; Military Law; Law and Social Governance; Constitutional and Administrative Law; Legal Theory and Legal History **Political Science**: Political Science
Arts (Literature and Arts)	**Journalism and Communication**: Journalism and News Practice; Communication and Broadcasting; History and Theory of Journalism and Media Management **Language and Literature**: Japanese Language and Literature; Linguistics and Applied Linguistics; Philology and Bibliography; Literary Theory; French Language and Literature; Literary History **Art Studies**: Dance Studies; Broadcasting and Television Art; Design Arts; Film Studies; Fine Arts; Drama and Opera Studies **Musicology**: Harmony; Musical Forms and Analysis; Instrumentation and Performance; Composition; Music History, Education, and Technology; Pitch and Scales
Management	**Business Administration**: Tourism Management and Technological Economics Management; Business and Accounting Management **Public Administration**: Social Medicine and Health Management; Education Economics, Management and Social Security; Land Resource Management and Administrative Management **Management Science and Engineering**: Management Science and Engineering
Medicine	**Clinical Medicine**: Dermatology and Venereology; Pediatrics; Oncology; Emergency Medicine; Imaging and Nuclear Medicine; Nursing and Rehabilitation Medicine; Geriatric Medicine; Obstetrics and Gynecology; Psychiatry and Mental Health; Internal Medicine; Surgery; Clinical Laboratory Diagnostics; Neurology; Ophthalmology; Anesthesiology; Otorhinolaryngology **Public Health and Preventive Medicine**: Maternal, Child and Adolescent Health; Nutrition and Food Hygiene; Health Toxicology and Environmental Health; Epidemiology and Health Statistics **Pharmacy**: Pharmaceutical Analysis; Pharmaceutics; Medicinal Chemistry; Microbiology and Biochemical Pharmacy; Pharmacology **Basic Medicine**: Forensic Medicine; Pathogen Biology; Human Anatomy and Histology-Embryology; Radiation Medicine; Immunology; Pathology and Pathophysiology **Stomatology**: Basic Stomatology; Clinical Stomatology **Traditional Medicine**: Traditional Pharmacy; Traditional Health Preservation; Traditional Medicine Theory
Education	**Pedagogy**: Theory of Curriculum and Instruction; Preschool Education; Educational Technology and Principles; Special Education **Psychology**: Psychology
Military (Militray Science)	**Military Studies**: Military Management; Military Thought and History; Military Logistics and Equipment; Military Command and Information Systems
Philosophy	**Philosophy**: Religious Studies; Philosophical Aesthetics; Ethics; Logic; Philosophy of Science and Technology
Economics	**Applied Economics**: Quantitative Economics; Finance; International Trade; Labor Economics; Public Finance; Economic Statistics; National and Defense Economics; Industrial Economics **Theoretical Economics**: Political Economy; Economic History; Western Economics
Agronomy	**Crop Science**: Crop Science **Aquaculture**: Aquaculture **Forestry**: Landscape Plants and Ornamental Horticulture; Forest Cultivation and Genetic Breeding **Animal Husbandry**: Animal Nutrition and Feed Science; Animal Rearing and Breeding **Veterinary Medicine**: Veterinary Medicine
Sociology	**Sociology**: Social and Folklore Studies; Demography and Anthropology
History	**History**: World History; Archaeology and Museology; Historical Geography

Table 10: Full list of the three levels of disciplines we use.

Table 11: Discipline-language distribution. We mark yellow for grids with less than 50, and red for those with less than 20. (Part 1)

Discipline	Arabic	Chinese	English	French	German
Aeronautical & Astronautical Science & Technology	90	97	37	80	78
Animal Husbandry	128	147	75	45	81
Applied Economics	151	148	131	150	150
Aquaculture	84	143	45	51	74
Architecture	44	148	44	43	53
Art Studies	51	146	62	150	126
Astronomy	100	148	81	92	101
Atmospheric Science	53	146	68	44	41
Basic Medicine	150	139	110	150	150
Biology	150	150	70	151	151
Business Administration	150	149	88	100	127
Chemical Engineering and Technology	150	150	41	155	152
Chemistry	116	150	43	94	103
Civil Engineering	150	150	85	103	101
Clinical Medicine	165	129	140	161	168
Computer Science and Technology	150	149	99	150	150
Control Science and Engineering	134	149	73	129	111
Crop Science	90	147	65	87	90
Electrical Engineering	150	148	62	150	150
Electronic Science and Technology	124	148	95	110	100
Environmental Science and Engineering	150	150	72	190	164
Food Science and Engineering	65	148	67	44	54
Forestry	128	150	54	76	91
Geography	50	136	50	56	75
Geological Resources and Geological Engineering	48	150	36	40	43
History	82	142	48	150	100
Hydraulic Engineering	110	149	75	88	78
Information and Communication Engineering	150	150	87	120	125
Instrument Science and Technology	123	149	39	87	74
Journalism and Communication	29	147	59	100	59
Language and Literature	30	137	90	150	73
Legal Studies	150	146	93	150	150
Management Science and Engineering	82	33	45	91	86
Materials Science and Engineering	150	148	54	159	100
Mathematics	150	146	42	160	150
Mechanical Engineering	70	150	79	45	71
Mechanics	150	147	4	100	104
Metallurgical Engineering	99	150	97	66	95
Military Studies	69	149	51	100	54
Musicology	50	19	98	73	56
Naval Architecture and Ocean Engineering	93	150	99	49	60
Nuclear Science and Technology	106	150	68	72	76
Optical Engineering	150	148	96	150	150
Pedagogy	101	148	49	100	88
Petroleum and Natural Gas Engineering	75	149	79	53	48
Pharmacy	150	126	96	100	100
Philosophy	150	145	45	150	123
Physical Oceanography	27	149	27	23	27
Physics	48	148	68	19	45
Political Science	68	98	51	100	76
Power Engineering and Engineering Thermophysics	150	150	52	150	150
Psychology	56	99	64	51	47
Public Administration	47	150	54	37	25
Public Health and Preventive Medicine	153	137	87	153	150
Sociology	157	143	48	168	154
Stomatology	31	127	58	26	25
Surveying and Mapping Science and Technology	86	148	148	63	76
Textile Science and Engineering	76	149	67	75	100
Theoretical Economics	128	150	73	67	100
Traditional Medicine	70	174	57	24	82
Transportation Engineering	106	147	150	76	95
Veterinary Medicine	47	129	31	31	26
Weapon Science and Technology	56	149	50	23	35

Table 12: Discipline-language distribution. We mark yellow for grids with less than 50, and red for those with less than 20. (Part 2)

Discipline	**Indonesian**	**Italian**	**Japanese**	**Korean**	**Portuguese**
Aeronautical & Astronautical Science & Technology	63	55	74	71	54
Animal Husbandry	100	57	81	150	71
Applied Economics	150	150	150	150	151
Aquaculture	91	61	79	75	61
Architecture	45	25	55	89	35
Art Studies	158	150	150	150	150
Astronomy	75	61	108	119	69
Atmospheric Science	21	30	40	54	21
Basic Medicine	150	151	151	150	151
Biology	151	150	151	150	153
Business Administration	121	120	128	150	95
Chemical Engineering and Technology	150	174	151	150	188
Chemistry	106	69	80	139	56
Civil Engineering	86	75	102	150	108
Clinical Medicine	157	156	156	156	162
Computer Science and Technology	150	160	150	150	150
Control Science and Engineering	100	76	149	150	101
Crop Science	86	73	86	101	74
Electrical Engineering	150	137	150	150	151
Electronic Science and Technology	100	101	120	125	63
Environmental Science and Engineering	151	149	141	150	128
Food Science and Engineering	61	49	81	98	29
Forestry	150	83	76	107	80
Geography	57	52	59	56	45
Geological Resources and Geological Engineering	35	28	60	52	41
History	100	150	150	150	150
Hydraulic Engineering	83	45	58	113	65
Information and Communication Engineering	150	100	150	150	100
Instrument Science and Technology	62	51	62	101	49
Journalism and Communication	57	78	103	150	61
Language and Literature	68	150	150	150	150
Legal Studies	100	151	150	150	150
Management Science and Engineering	58	82	83	77	84
Materials Science and Engineering	132	146	150	150	123
Mathematics	176	116	115	150	160
Mechanical Engineering	26	38	59	59	44
Mechanics	125	107	109	149	107
Metallurgical Engineering	103	69	100	117	29
Military Studies	34	75	111	134	47
Musicology	49	72	101	108	56
Naval Architecture and Ocean Engineering	31	17	86	99	36
Nuclear Science and Technology	84	100	96	103	95
Optical Engineering	150	150	150	150	130
Pedagogy	106	82	150	150	63
Petroleum and Natural Gas Engineering	47	25	66	85	37
Pharmacy	150	102	99	150	82
Philosophy	149	150	150	150	150
Physical Oceanography	23	23	31	42	12
Physics	13	30	40	70	23
Political Science	39	84	78	84	87
Power Engineering and Engineering Thermophysics	100	100	150	150	125
Psychology	45	33	79	60	40
Public Administration	65	38	50	101	36
Public Health and Preventive Medicine	153	136	150	150	151
Sociology	150	152	147	150	168
Stomatology	26	25	45	49	15
Surveying and Mapping Science and Technology	84	34	100	145	54
Textile Science and Engineering	54	62	108	116	44
Theoretical Economics	89	50	102	150	89
Traditional Medicine	31	38	85	116	34
Transportation Engineering	96	34	103	133	73
Veterinary Medicine	42	42	42	73	27
Weapon Science and Technology	26	24	55	66	9

Table 13: Discipline-language distribution. We mark yellow for grids with less than 50, and red for those with less than 20. (Part 3)

Discipline	**Russian**	**Spanish**	**Thai**	**Vietnamese**
Aeronautical & Astronautical Science & Technology	50	45	88	73
Animal Husbandry	71	73	150	135
Applied Economics	150	150	150	150
Aquaculture	52	82	75	85
Architecture	27	30	82	75
Art Studies	150	153	150	150
Astronomy	123	81	146	143
Atmospheric Science	54	41	81	55
Basic Medicine	150	152	150	150
Biology	150	151	150	151
Business Administration	151	100	105	105
Chemical Engineering and Technology	151	163	150	150
Chemistry	102	89	150	150
Civil Engineering	108	85	150	150
Clinical Medicine	165	163	150	163
Computer Science and Technology	150	151	150	150
Control Science and Engineering	145	97	150	150
Crop Science	93	75	76	76
Electrical Engineering	150	150	150	150
Electronic Science and Technology	100	102	135	146
Environmental Science and Engineering	161	140	150	150
Food Science and Engineering	47	69	125	128
Forestry	67	88	150	150
Geography	51	56	138	150
Geological Resources and Geological Engineering	42	44	65	57
History	127	132	150	150
Hydraulic Engineering	61	53	116	106
Information and Communication Engineering	150	101	150	150
Instrument Science and Technology	77	59	94	111
Journalism and Communication	99	62	102	97
Language and Literature	150	153	150	150
Legal Studies	150	150	150	150
Management Science and Engineering	87	80	82	75
Materials Science and Engineering	151	165	150	128
Mathematics	150	155	150	150
Mechanical Engineering	44	27	56	47
Mechanics	150	105	150	150
Metallurgical Engineering	89	100	150	150
Military Studies	103	46	100	150
Musicology	65	47	100	132
Naval Architecture and Ocean Engineering	52	26	62	42
Nuclear Science and Technology	100	106	126	110
Optical Engineering	150	150	150	150
Pedagogy	103	69	150	150
Petroleum and Natural Gas Engineering	71	36	98	91
Pharmacy	100	123	150	150
Philosophy	150	150	150	150
Physical Oceanography	30	27	54	55
Physics	50	27	68	53
Political Science	70	74	52	63
Power Engineering and Engineering Thermophysics	150	101	150	150
Psychology	56	28	96	93
Public Administration	41	48	118	116
Public Health and Preventive Medicine	152	150	150	150
Sociology	162	149	150	150
Stomatology	36	21	55	28
Surveying and Mapping Science and Technology	58	37	130	150
Textile Science and Engineering	78	66	114	118
Theoretical Economics	101	78	150	150
Traditional Medicine	93	48	107	124
Transportation Engineering	89	51	125	83
Veterinary Medicine	37	40	89	74
Weapon Science and Technology	43	28	72	62

Table 14: The distribution of questions on the cognitive requirement for each language and each major, in the form of {number of reasoning questions}/{number total questions}, where we mark yellow for grids with less than 40% of inference questions and red for those with less than 20%. (Part 1)

Discipline	Arabic	Chinese	English	French	German
Aeronautical & Astronautical Science & Technology	70/ 90	88/ 97	34/ 37	37/ 80	38/ 78
Animal Husbandry	117/128	93/147	56/ 75	22/ 45	43/ 81
Applied Economics	130/151	144/148	110/131	106/150	105/150
Aquaculture	74/ 84	89/143	27/ 45	27/ 51	45/ 74
Architecture	30/ 44	103/148	28/ 44	21/ 43	35/ 53
Art Studies	37/ 51	116/146	29/ 62	51/150	38/126
Astronomy	50/100	130/148	44/ 81	29/ 92	25/101
Atmospheric Science	39/ 53	121/146	27/ 68	13/ 44	19/ 41
Basic Medicine	109/150	119/139	75/110	52/150	42/150
Biology	100/150	146/150	39/ 70	53/151	58/151
Business Administration	132/150	131/149	66/ 88	74/100	83/127
Chemical Engineering and Technology	117/150	128/150	37/ 41	81/155	77/152
Chemistry	44/116	141/150	41/ 43	33/ 94	39/103
Civil Engineering	100/150	138/150	76/ 85	35/103	30/101
Clinical Medicine	121/165	123/129	95/140	57/161	64/168
Computer Science and Technology	108/150	140/149	72/ 99	48/150	69/150
Control Science and Engineering	96/134	125/149	63/ 73	45/129	38/111
Crop Science	81/ 90	146/147	44/ 65	55/ 87	48/ 90
Electrical Engineering	108/150	131/148	45/ 62	75/150	49/150
Electronic Science and Technology	75/124	121/148	82/ 95	53/110	62/100
Environmental Science and Engineering	95/150	112/150	45/ 72	78/190	83/164
Food Science and Engineering	35/ 65	93/148	58/ 67	12/ 44	27/ 54
Forestry	102/128	95/150	33/ 54	49/ 76	60/ 91
Geography	23/ 50	125/136	20/ 50	20/ 56	27/ 75
Geological Resources and Geological Engineering	30/ 48	104/150	16/ 36	20/ 40	23/ 43
History	38/ 82	135/142	17/ 48	46/150	30/100
Hydraulic Engineering	73/110	121/149	66/ 75	38/ 88	48/ 78
Information and Communication Engineering	98/150	122/150	61/ 87	47/120	36/125
Instrument Science and Technology	69/123	126/149	22/ 39	22/ 87	19/ 74
Journalism and Communication	26/ 29	120/147	32/ 59	59/100	22/ 59
Language and Literature	19/ 30	119/137	69/ 90	62/150	30/ 73
Legal Studies	108/150	126/146	52/ 93	64/150	42/150
Management Science and Engineering	70/ 82	31/ 33	40/ 45	50/ 91	63/ 86
Materials Science and Engineering	92/150	105/148	38/ 54	70/159	35/100
Mathematics	110/150	142/146	38/ 42	87/160	77/150
Mechanical Engineering	42/ 70	131/150	73/ 79	13/ 45	28/ 71
Mechanics	110/150	136/147	3/ 4	59/100	48/104
Metallurgical Engineering	55/ 99	108/150	75/ 97	22/ 66	38/ 95
Military Studies	43/ 69	76/149	16/ 51	36/100	29/ 54
Musicology	49/ 50	5/ 19	32/ 98	23/ 73	22/ 56
Naval Architecture and Ocean Engineering	39/ 93	112/150	92/ 99	11/ 49	17/ 60
Nuclear Science and Technology	67/106	81/150	58/ 68	34/ 72	32/ 76
Optical Engineering	109/150	87/148	90/ 96	75/150	56/150
Pedagogy	80/101	126/148	32/ 49	52/100	56/ 88
Petroleum and Natural Gas Engineering	51/ 75	131/149	66/ 79	16/ 53	22/ 48
Pharmacy	99/150	90/126	49/ 96	43/100	25/100
Philosophy	111/150	142/145	24/ 45	74/150	59/123
Physical Oceanography	17/ 27	145/149	15/ 27	11/ 23	19/ 27
Physics	25/ 48	135/148	62/ 68	7/ 19	31/ 45
Political Science	51/ 68	85/ 98	18/ 51	57/100	29/ 76
Power Engineering and Engineering Thermophysics	80/150	121/150	47/ 52	72/150	47/150
Psychology	50/ 56	99/ 99	51/ 64	36/ 51	28/ 47
Public Administration	31/ 47	105/150	28/ 54	20/ 37	12/ 25
Public Health and Preventive Medicine	137/153	107/137	51/ 87	95/153	88/150
Sociology	141/157	139/143	28/ 48	134/168	108/154
Stomatology	24/ 31	91/127	20/ 58	4/ 26	11/ 25
Surveying and Mapping Science and Technology	42/ 86	103/148	117/148	16/ 63	34/ 76
Textile Science and Engineering	33/ 76	98/149	45/ 67	14/ 75	38/100
Theoretical Economics	87/128	146/150	55/ 73	28/ 67	49/100
Traditional Medicine	51/ 70	154/174	5/ 57	11/ 24	47/ 82
Transportation Engineering	66/106	105/147	124/150	36/ 76	55/ 95
Veterinary Medicine	40/ 47	118/129	14/ 31	20/ 31	17/ 26
Weapon Science and Technology	34/ 56	78/149	50/ 50	9/ 23	20/ 35

Table 15: The distribution of questions on the cognitive requirement for each language and each major, in the form of {number of reasoning questions}/{number total questions}, where we mark yellow for grids with less than 40% of inference questions and red for those with less than 20%. (Part 2)

Discipline	Indonesian	Italian	Japanese	Korean	Portuguese
Aeronautical & Astronautical Science & Technology	44/ 63	40/ 55	42/ 74	38/ 71	36/ 54
Animal Husbandry	62/100	28/ 57	51/ 81	98/150	46/ 71
Applied Economics	114/150	97/150	87/150	113/150	123/151
Aquaculture	63/ 91	33/ 61	51/ 79	46/ 75	42/ 61
Architecture	27/ 45	13/ 25	25/ 55	52/ 89	24/ 35
Art Studies	76/158	16/150	97/150	95/150	44/150
Astronomy	38/ 75	16/ 61	48/108	55/119	41/ 69
Atmospheric Science	7 21	13/ 30	23/ 40	25/ 54	10/ 21
Basic Medicine	63/150	30/151	40/151	42/150	59/151
Biology	70/151	45/150	78/151	61/150	82/153
Business Administration	78/121	71/120	77/128	101/150	64/ 95
Chemical Engineering and Technology	73/150	74/174	76/151	70/150	107/188
Chemistry	20/106	18/ 69	27/ 80	21/139	18/ 56
Civil Engineering	43/ 86	27/ 75	21/102	72/150	36/108
Clinical Medicine	43/157	23/156	47/156	61/156	37/162
Computer Science and Technology	77/150	64/160	79/150	77/150	81/150
Control Science and Engineering	41/100	38/ 76	53/149	66/150	55/101
Crop Science	58/ 86	40/ 73	35/ 86	39/101	39/ 74
Electrical Engineering	72/150	52/137	66/150	56/150	88/151
Electronic Science and Technology	61/100	64/101	49/120	33/125	56/ 63
Environmental Science and Engineering	61/151	57/149	51/141	74/150	70/128
Food Science and Engineering	21/ 61	11/ 49	38/ 81	40/ 98	15/ 29
Forestry	108/150	50/ 83	39/ 76	62/107	52/ 80
Geography	24/ 57	11/ 52	32/ 59	33/ 56	28/ 45
Geological Resources and Geological Engineering	14/ 35	13/ 28	35/ 60	27/ 52	23/ 41
History	54/100	44/150	75/150	92/150	97/150
Hydraulic Engineering	42/ 83	18/ 45	30/ 58	44/113	42/ 65
Information and Communication Engineering	69/150	40/100	54/150	55/150	63/100
Instrument Science and Technology	21/ 62	11/ 51	20/ 62	28/101	19/ 49
Journalism and Communication	35/ 57	19/ 78	72/103	96/150	40/ 61
Language and Literature	22/ 68	43/150	92/150	88/150	95/150
Legal Studies	34/100	32/151	89/150	86/150	58/150
Management Science and Engineering	43/ 58	43/ 82	45/ 83	31/ 77	55/ 84
Materials Science and Engineering	52/132	59/146	59/150	59/150	66/123
Mathematics	108/176	43/116	18/115	55/150	77/160
Mechanical Engineering	9/ 26	15/ 38	25/ 59	29/ 59	20/ 44
Mechanics	58/125	62/107	48/109	71/149	74/107
Metallurgical Engineering	25/103	17/ 69	50/100	46/117	15/ 29
Military Studies	24/ 34	19/ 75	55/111	76/134	32/ 47
Musicology	14/ 49	21/ 72	61/101	49/108	22/ 56
Naval Architecture and Ocean Engineering	14/ 31	5/ 17	52/ 86	46/ 99	20/ 36
Nuclear Science and Technology	40/ 84	38/100	49/ 96	51/103	65/ 95
Optical Engineering	60/150	56/150	60/150	60/150	68/130
Pedagogy	74/106	38/ 82	102/150	109/150	35/ 63
Petroleum and Natural Gas Engineering	22/ 47	10/ 25	42/ 66	49/ 85	25/ 37
Pharmacy	44/150	10/102	30/ 99	47/150	39/ 82
Philosophy	69/149	43/150	92/150	102/150	78/150
Physical Oceanography	13/ 23	9/ 23	23/ 31	24/ 42	6/ 12
Physics	4/ 13	11/ 30	32/ 40	28/ 70	12/ 23
Political Science	34/ 39	33/ 84	42/ 78	56/ 84	50/ 87
Power Engineering and Engineering Thermophysics	48/100	37/100	45/150	46/150	58/125
Psychology	26/ 45	14/ 33	57/ 79	36/ 60	25/ 40
Public Administration	35/ 65	12/ 38	32/ 50	69/101	26/ 36
Public Health and Preventive Medicine	79/153	59/136	86/150	97/150	94/151
Sociology	114/150	102/152	107/147	104/150	130/168
Stomatology	4/ 26	5/ 25	30/ 45	22/ 49	5/ 15
Surveying and Mapping Science and Technology	33/ 84	13/ 34	56/100	50/145	25/ 54
Textile Science and Engineering	11/ 54	8/ 62	47/108	30/116	21/ 44
Theoretical Economics	34/ 89	13/ 50	48/102	83/150	52/ 89
Traditional Medicine	1/ 31	5/ 38	49/ 85	46/116	14/ 34
Transportation Engineering	47/ 96	16/ 34	56/103	67/133	39/ 73
Veterinary Medicine	21/ 42	21/ 42	29/ 42	43/ 73	20/ 27
Weapon Science and Technology	12/ 26	12/ 24	32/ 55	44/ 66	7/ 9

Table 16: The distribution of questions on the cognitive requirement for each language and each major, in the form of {number of reasoning questions}/{number total questions}, where we mark yellow for grids with less than 40% of inference questions and red for those with less than 20%. (Part 3)

Discipline	Russian	Spanish	Thai	Vietnamese
Aeronautical & Astronautical Science & Technology	26/ 50	26/ 45	62/ 88	56/ 73
Animal Husbandry	51/ 71	36/ 73	108/150	89/135
Applied Economics	113/150	107/150	128/150	105/150
Aquaculture	35/ 52	55/ 82	56/ 75	54/ 85
Architecture	14/ 27	22/ 30	55/ 82	45/ 75
Art Studies	68/150	48/153	91/150	56/150
Astronomy	12/123	36/ 81	39/146	50/143
Atmospheric Science	21/ 54	19/ 41	42/ 81	17/ 55
Basic Medicine	41/150	44/152	59/150	58/150
Biology	75/150	50/151	70/150	78/151
Business Administration	119/151	64/100	86/105	73/105
Chemical Engineering and Technology	63/151	92/163	74/150	83/150
Chemistry	22/102	18/ 89	44/150	55/150
Civil Engineering	39/108	34/ 85	85/150	87/150
Clinical Medicine	56/165	41/163	58/150	55/163
Computer Science and Technology	91/150	68/151	74/150	83/150
Control Science and Engineering	73/145	50/ 97	76/150	101/150
Crop Science	49/ 93	41/ 75	33/ 76	46/ 76
Electrical Engineering	49/150	63/150	78/150	84/150
Electronic Science and Technology	62/100	57/102	62/135	67/146
Environmental Science and Engineering	68/161	52/140	88/150	82/150
Food Science and Engineering	14/ 47	26/ 69	42/125	53/128
Forestry	54/ 67	60/ 88	108/150	104/150
Geography	25/ 51	27/ 56	92/138	90/150
Geological Resources and Geological Engineering	14/ 42	21/ 44	27/ 65	31/ 57
History	67/127	65/132	87/150	78/150
Hydraulic Engineering	19/ 61	34/ 53	74/116	59/106
Information and Communication Engineering	55/150	38/101	71/150	82/150
Instrument Science and Technology	22/ 77	14/ 59	47/ 94	33/111
Journalism and Communication	53/ 99	25/ 62	65/102	40/ 97
Language and Literature	76/150	67/153	112/150	42/150
Legal Studies	69/150	61/150	82/150	71/150
Management Science and Engineering	57/ 87	49/ 80	68/ 82	56/ 75
Materials Science and Engineering	58/151	59/165	48/150	59/128
Mathematics	98/150	89/155	90/150	89/150
Mechanical Engineering	24/ 44	16/ 27	32/ 56	31/ 47
Mechanics	78/150	77/105	108/150	89/150
Metallurgical Engineering	26/ 89	28/100	35/150	35/150
Military Studies	34/103	22/ 46	61/100	104/150
Musicology	22/ 65	14/ 47	59/100	34/132
Naval Architecture and Ocean Engineering	16/ 52	8/ 26	37/ 62	18/ 42
Nuclear Science and Technology	37/100	46/106	48/126	57/110
Optical Engineering	54/150	52/150	68/150	53/150
Pedagogy	69/103	48/ 69	120/150	104/150
Petroleum and Natural Gas Engineering	23/ 71	15/ 36	51/ 98	41/ 91
Pharmacy	29/100	21/123	58/150	65/150
Philosophy	63/150	58/150	87/150	64/150
Physical Oceanography	8/ 30	15/ 27	30/ 54	24/ 55
Physics	14/ 50	13/ 27	43/ 68	22/ 53
Political Science	41/ 70	41/ 74	28/ 52	41/ 63
Power Engineering and Engineering Thermophysics	47/150	33/101	69/150	78/150
Psychology	38/ 56	13/ 28	69/ 96	65/ 93
Public Administration	19/ 41	24/ 48	90/118	88/116
Public Health and Preventive Medicine	103/152	75/150	84/150	73/150
Sociology	138/162	120/149	106/150	106/150
Stomatology	11/ 36	5/ 21	25/ 55	9/ 28
Surveying and Mapping Science and Technology	17/ 58	15/ 37	72/130	74/150
Textile Science and Engineering	17/ 78	12/ 66	21/114	31/118
Theoretical Economics	52/101	32/ 78	89/150	78/150
Traditional Medicine	41/ 93	12/ 48	29/107	9/124
Transportation Engineering	48/ 89	29/ 51	87/125	53/ 83
Veterinary Medicine	18/ 37	21/ 40	71/ 89	55/ 74
Weapon Science and Technology	8/ 43	15/ 28	42/ 72	31/ 62