BENCHHUB: A Unified Benchmark Suite for Holistic and Customizable LLM Evaluation

Eunsu Kim[1,*], **Haneul Yoo**[1,*], **Guijin Son**[2,3], **Hitesh Patel**[4], **Amit Agarwal**[4], **Alice Oh**[1]
[1]KAIST, [2]Yonsei University, [3]OnelineAI, [4]Oracle
kes0317@kaist.ac.kr, haneul.yoo@kaist.ac.kr, alice.oh@kaist.edu

Abstract

As large language models (LLMs) with advanced reasoning abilities continue to evolve, their capabilities are increasingly tested across heterogeneous contexts. To evaluate them effectively, benchmarks must move beyond fragmented datasets and narrow rankings, addressing the growing need to capture abilities that integrate multiple skills (e.g., reasoning and knowledge) across diverse domains (e.g., mathematics and culture). This complexity calls for a new paradigm of evaluation—flexible, domain-aware, and continuously updated. In this paper, we introduce BENCHHUB, a dynamic benchmark repository that empowers researchers and developers to evaluate LLMs effectively, with a focus on Korean and English. BENCHHUB aggregates and automatically classifies benchmark datasets from diverse domains, integrating 839k questions across 54 benchmarks. It is designed to support continuous updates and scalable data management, enabling flexible and customizable evaluation tailored to various domains or use cases. Through extensive experiments with various LLM families, we demonstrate that model performance varies significantly across domain-specific subsets, emphasizing the importance of domain-aware benchmarking. Furthermore, we extend BENCHHUB into 10 languages spanning resource levels. We believe BenchHub can encourage better dataset reuse, more transparent model comparisons, and easier identification of underrepresented areas in existing benchmarks, offering a critical infrastructure for advancing LLM evaluation research.

1 Introduction

Large language models (LLMs) have made remarkable strides, powering applications across diverse tasks, including research [6], industry [8], and everyday life [9]. As their roles expand—from open-ended reasoning to culturally sensitive decision-making [30]—LLMs are increasingly required to integrate multiple skills (e.g., reasoning and knowledge) across diverse domains (e.g., mathematics and culture). This complexity underscores the need for a new evaluation paradigm that goes beyond formulaic rankings, toward rigorous and comprehensive assessments of whether model behavior aligns with the nuanced objectives of specific users and applications.

In response, a wide range of evaluation efforts has emerged. On the one hand, holistic evaluation benchmarks [43, 52] and leaderboards based on user preference [11] or aggregated benchmarks [1] serve as popular community standards. While useful for broad comparisons, their aggregated scores obscure fine-grained strengths and weaknesses, often misaligning with the needs of specific applications [65]. On the other hand, specialized benchmarks target narrow aspects, such as law [42], medical advice [2], and finance [75], as well as specific tasks, including knowledge retrieval [21], reasoning [14, 96], and value alignment [55, 28]. While these datasets capture critical capabilities, their vast, fragmented, and overlapping nature creates a chaotic landscape. For instance, in the mathematics

[*]Equal contribution.

Submitted to NeurIPS 2025 Workshop on Efficient Reasoning. Do not distribute.

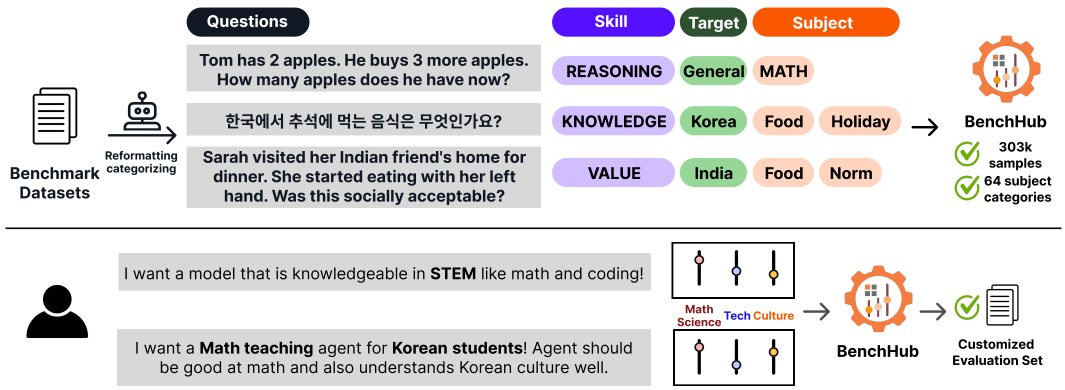


Figure 1: The concept of BENCHHUB. BENCHHUB automatically classifies and merges questions from existing benchmark datasets on a sample-wise basis. Through BENCHHUB, users can select test sets that align with their objectives and efficiently evaluate the models.

domain, numerous benchmarks exist, such as MATH [22] and GSM8k [14], which in turn partially overlap with broader collections (*e.g.*, MMLU [21]). This leaves researchers and practitioners with a dilemma: which benchmarks truly reflect their objective, and how can they compose a principled, customized evaluation suite tailored for diverse needs?

In this paper, we introduce BENCHHUB[2], a unified and customizable benchmark suite for holistic yet domain-aware LLM evaluation. BENCHHUB aggregates 839k questions from 54 benchmarks across 64 domains and 10 languages, mainly in English and Korean. We systematically categorize existing benchmarks by six dimensions: 1) tasks (*e.g.*, mathematical reasoning), 2) answer formats (*e.g.*, multiple-choice QA), 3) tool usage (*i.e.*, language-only or requirements to external tools), 4) skills (*i.e.*, knowledge, reasoning, or value/alignment), 5) coarse- and fine-grained subjects (*e.g.*, STEM–mathematics), and 6) targets (*i.e.*, culturally specific or agnostic). This design facilitates users to dynamically construct their own evaluation sets tailored to their needs, moving beyond rigid, predefined test sets (Figure 1). To ensure long-term, dynamic scalability, we further train and release a categorization model that seamlessly integrates new, unseen benchmarks into BENCHHUB.

Using BENCHHUB, we evaluate 14 open LLMs and uncover a crucial insight: model rankings fluctuate substantially depending on benchmark compositions and domain focus. This finding highlights the central issue of benchmark composition bias, which can significantly distort interpretations of model performance. We further validate BENCHHUB through 5 real-world use cases—such as legal, educational, and cultural applications—showing how domain-aware evaluation alters conclusions about model superiority. We hope BENCHHUB provides a foundation for the community to move beyond monolithic leaderboards toward domain-aware, trustworthy, and customizable evaluation.

2 Existing LLM Evaluation Benchmarks are Skewed

What aspects do the commonly used multi-domain datasets evaluate, and how is the distribution of domains represented across these datasets? To answer this question, we classify three representative holistic benchmarks (*i.e.*, Chatbot Arena [11], MixEval [52], and MMLU [21]) as multilabels using our fine-tuned classifiers (§ 3) in terms of coarse-grained subjects (Figure 2a) and tasks (Figure 2b).

Among them, Chatbot Arena includes only 25.5% of Humanities and Social Sciencce (HASS) questions, while both MixEval and MMLU comprise more than

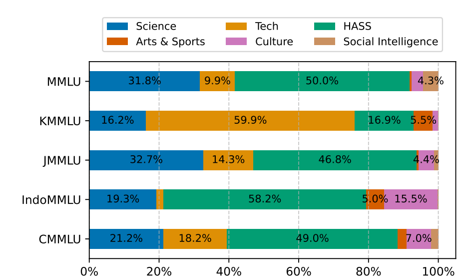
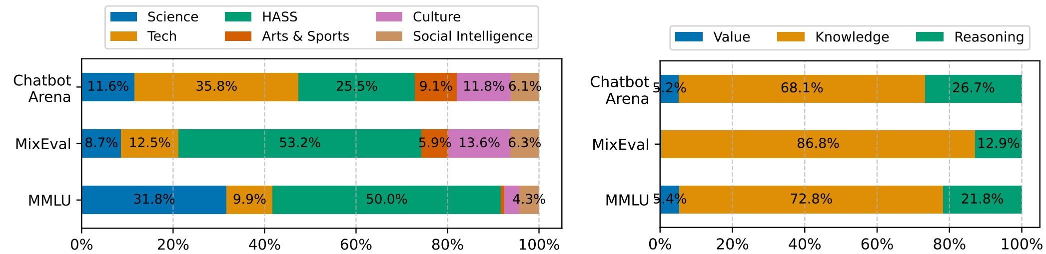


Figure 3: Distribution of MMLU in English, Korean, Japanese, Indonesian, and Chinese.

[2]We release our datasets and interactive platform at https://huggingface.co/BenchHub and our code at https://github.com/rladmstn1714/BenchHub.



(a) Subject (b) Task

Figure 2: Data distribution of existing evaluation benchmarks.

half of HASS questions. In addition, MixEval includes fewer than 0.30% of value alignment tasks and mostly focuses on measuring knowledge. Such disparities may lead to biased findings, where models that excel in certain domains may appear to perform better overall, potentially skewing the evaluation results.

Moreover, these biases are not limited to cross-benchmark comparisons but can also manifest within multilingual contexts. Figure 3 and Figure 10 illustrate data distributions of MMLU series datasets in 5 languages classified by the model (§ 3) in terms of coarse-grained subjects. For instance, MMLU in English emphasizes HASS, whereas Korean MMLU (KMMLU) [77] comprises 76.1% of STEM (Science, Technology, Engineering, and Mathematics) questions. This variation complicates the interpretation of performance differences, as it is challenging to discern whether the performance degradations in non-English are due to language proficiency or domain-specific knowledge.

Hence, instead of recklessly adopting existing holistic benchmarks, we recommend carefully selecting the benchmark suites for a reliable evaluation.

3 BenchHub

Consider a user who wants to determine "*Which model excels at both mathematics and understanding culture?*" As discussed in § 2, it remains unclear how to answer such specific, goal-oriented questions and how to construct their evaluation suite, as existing evaluation benchmarks [21, 43, 52] mainly provide general-purpose scores. To this end, we introduce BenchHub, a unified collection of LLM evaluation benchmarks across diverse domains. BenchHub integrates 54 benchmarks comprising 839k samples in 10 languages, with a primary focus on English and Korean as BenchHub-En and BenchHub-Ko, respectively. We design BenchHub around two core principles: 1) a fine-grained, multi-dimensional taxonomy to deconstruct model capabilities and 2) a fully automated pipeline to dynamically update and expand it with new datasets. In this section, we detail the taxonomy design (§ 3.1), the data curation (§ 3.2), the automated pipeline (§ 3.3), as well as interactive tools and utilities as a web-based platform(§ 3.4). Finally, we illustrate the multilingual extension of BenchHub —from English and Korean to eight additional languages—in § 3.5.

3.1 Taxonomy

We annotate each dataset with six orthogonal dimensions: three dataset-level attributes—**task, answer format, and tool usage**— and three sample-level attributes—**skill, subject, and target**. The full scheme is illustrated in Appendix D.

Dataset-level attributes:

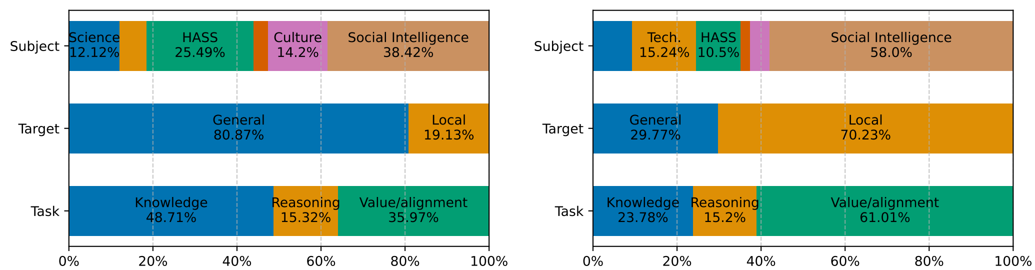
1. **Task** refers to the high-level family defined by the dataset authors (*e.g.*, mathematical reasoning, code generation, cultural understanding). This provides a general understanding of a dataset's purpose. We assign it automatically from the dataset's abstract or description using LLM inference.
2. **Answer format** specifies the expected response format: binary, multiple-choice QA (MCQA), short-form, free-form, open-ended (*e.g.*, story generation), and comparison (*e.g.*, determining which response is better between A and B). This is crucial for selecting appropriate evaluation prompts and formats.

3. **Tool Usage** indicates whether a task requires language capabilities only (*language-only*) or interaction with external tools such as *e.g.*, code interpreters, web browsers, calculators (*requires externals tools*). This dimension supports agentic evaluation, where models must decide when and how to invoke external resources.

Sample-level attributes:

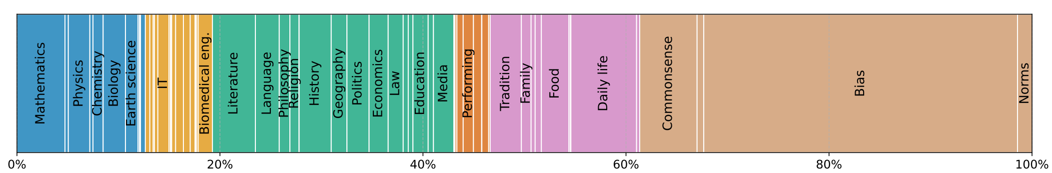
4. **Skill** captures the required ability to answer the question (*i.e.*, reasoning, knowledge, and value/alignment).
5. **Subject** denotes the knowledge domain. We define six coarse-grained categories—*Science*, *Technology*, *Humanities and Social Science (HASS)*, *Arts & Sports*, *Culture*, and *Social Intelligence*—along with 64 sub-categories, by integrating various knowledge classification systems. Each sample may have multiple subject labels.
6. **Target** represents the cultural or geographical focus. Culturally agnostic items are labeled as *General*; otherwise, we assign a *Local* tag. This supports evaluation under culturally-aware evaluation [73].

3.2 Datasets

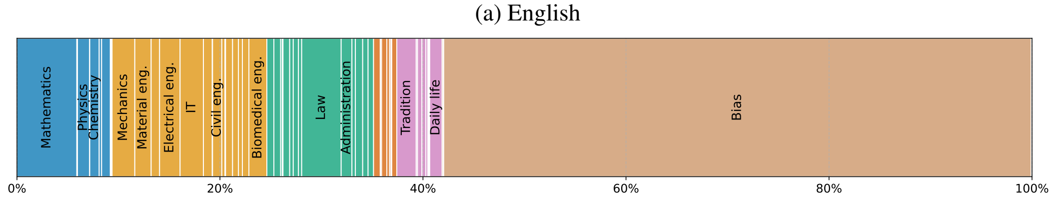


(a) English (b) Korean

Figure 4: Data distribution of all datasets used in this paper by coarse-grained subjects, targets, and tasks. The English and Korean data include 250,940 and 144,331 questions each.



(a) English



(b) Korean

Figure 5: Fine-grained data distribution of all datasets used in this paper in terms of subjects

We apply this taxonomy to 54 benchmarks in 10 languages, mainly covering English and Korean and totaling over 839k instances. Figures 2 and 5 show the overall statistics of English and Korean datasets included in our benchmark. For English and Korean, we include 31 English and 12 Korean language benchmarks with a total of 41 datasets. [3] We curate 1) general-purpose (*i.e.*, culturally agnostic) datasets commonly used by holistic evaluation benchmarks [94, 52] and 2) culture-specific

[3] We count the multilingual datasets—BLEnD [50] and CaLMQA [3]—in both.

datasets. We select English datasets spanning multiple cultures drawn from a recent survey [56], curating over 300 papers and datasets regarding LLM cultural awareness. For Korean, where public resources are fewer than in English, we include most datasets released after 2022. Table 2 in the Appendix provides a complete list of the datasets.

3.3 Automated and Dynamic Expansion

With benchmark datasets emerging at a rapid pace, it is crucial to flexibly manage them for holistic evaluation. To dynamically adapt to newly emerging datasets, we automate the entire dataset merging process using an LLM agent, which includes reformatting the datasets into our benchmark format and classifying each sample into categories. The processing pipeline for a newly introduced dataset is outlined as follows:

1. **Reformatting:** We first automatically parse, reformat, and map a new dataset to the standardized BENCHHUB scheme using an LLM-guided rule-based approach. If the dataset does not adhere to our predefined schema, an LLM agent (*e.g.*, GPT-4o or Gemini) is employed to map keys to the correct format.
2. **Metadata assignment:** The LLM agent extracts the meta-task description and infers the task, answer format, and tool usage from the dataset documentation (*e.g.*, abstract).
3. **Sample-level Categorization:** We then assign sample-level attributes (*i.e.*, skill, subject, and target type) using a fine-tuned Qwen-2.5-7B model (`BenchHub-Cat-7B`). [4]
4. **Merging:** The processed and annotated dataset is seamlessly merged into the main collections, thereby producing the next BENCHHUB release.

This automated pipeline allows BENCHHUB to continuously expand and provide more comprehensive evaluations as new datasets emerge. While we acknowledge the incompleteness of LLM-based expansion, we provide an empirical discussion of the reliability and robustness of this automated process in Appendix E.2.

3.4 Interactive Platform and Utilities

To proliferate our structured data into actionable insights for researchers and practitioners, we release an interactive web-based platform (Figure 8) and code utilities. The web demo allows users to filter out datasets by any category combinations, inspect statistics, download their customized subsets, and propose new datasets via pull requests. The code utilities offer two main features:

1. **Dataset Loader:** It filters the dataset to include only the categories selected by the user. It also allows the user to choose between returning the entire selected dataset or a filtered version with overlapping entries (including near-duplicates) removed, which is useful since multiple aggregated datasets may contain overlapping samples.
2. **Citation Report Generator:** For the customized dataset returned to the user, it produces a laTEX table of datasets with their sources and licenses, includes dataset statistics such as the number of instances, and provides a comprehensive citation list (e.g., BibTEX entries) to ensure proper credit to dataset authors.

For better reproducibility, we adopt HRET [39] [5], enabling direct evaluations on BENCHHUB. Design and implementation details of the platform and code utilities appear in Appendix B.

3.5 Multilingual Extension of BENCHHUB

While we focus on two languages (*i.e.*, Korean and English), we highlight that BENCHHUB is a language-agnostic, flexible framework that can be easily extended to other languages. To empirically guide this extension, we present `BenchHub-Multi-Cat-7B`[6], a multilingual categorizer supporting

[4] The model link will be added after the anonymous review period. Details on the training and validation of `BenchHub-Cat-7B` are provided in Appendix E.1.

[5] HRET is an evaluation toolkit supporting multiple datasets, including BENCHHUB.

[6] The model link will be added after the anonymous review period.

10 languages—English (En); 3 high-resource (Arabic (Ar), German (De), Dutch (Nl)); 3 mid-resource (Indonesian (Id), Korean (Ko), Ukrainian (Uk)); 3 low-resource (Swahili (Sw), Nepali (Ne), Kyrgyz (Ky)). Our multilingual categorizer achieves an average accuracy of 77.5% on fine-grained subject categorizations for unseen, out-of-domain data. Furthermore, we introduce BENCHHUB-multilingual, which extends our benchmark suite to a total of 10 languages consisting of 13 datasets and 444,402 samples. We hope BENCHHUB-multilingual to serve as a foundational step for reliable LLM evaluations in non-English languages. The details of the training procedure and the datasets for each language are provided in the Appendix F.

4 Evaluation Results using BENCHHUB

4.1 Evaluation of LLMs across diverse subjects

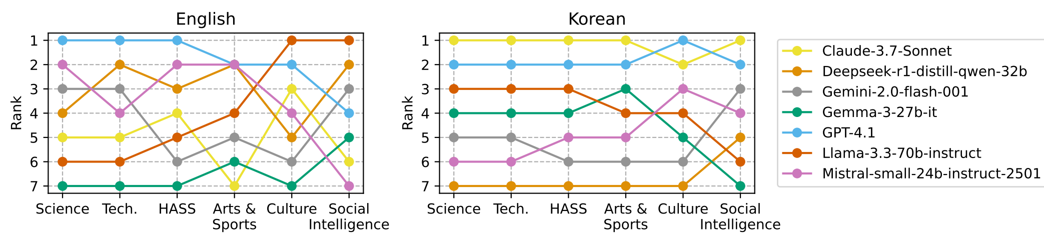


Figure 6: LLM evaluation ranking under BENCHHUB in terms of coarse-grained subjects

In this section, we evaluate seven LLMs across diverse subjects using BENCHHUB. We select 6,644 and 6,485 examples for English and Korean, respectively. To manage the large number of fine-grained categories, we sample up to 150 examples per category, fully including categories with 100–150 samples and merging categories with fewer than 80 samples into a miscellaneous group within the same coarse-grained classification. For evaluation, we extract the model's intended answer from MCQA questions by applying a set of regular expressions [49], while using an LLM as a parser extractor for short-form questions [7], similar to the approach in previous work [52].

We include one model from each commonly used LLM family. For proprietary models, we use GPT-4.1, Gemini-2.0-flash, and Claude 3.7 Sonnet [8]. Open models include Qwen-3-32b [89], DeepSeek-R1-Distill-Qwen-32B [15], Llama-3.3-70B [18], Mistral-Small-24B-Instruct, and gemma-2-27b-it [82].

Figure 6 presents model rankings by subject category. Our results show that frequent fluctuations in model rankings depend on the category. For example, Llama-3.3-70b ranks 6th in Science and Tech, but ranks as the top-performing model among seven models in Culture and Social Intelligence. This highlights the importance of domain-specific evaluation aligned with the evaluation context and objectives. The full results for each subject and model are in Table 12- 13 in the Appendix H.

4.2 Impact of Category Distribution on Model Ranking

In this section, we empirically validate the influence of category distributions within evaluation benchmarks on model rankings. Since this requires experiments on large datasets for statistical validation, we include 14 open models ranging from 1B to 72B parameters. We test on 27 English and 13 Korean datasets, comprising 16,898 and 18,977 MCQA samples, respectively. The number of answer choices per MCQA sample varies between 3 and 18. We extract the model's intended answer by applying a set of regular expressions [49]. The evaluated LLMs include:

- Qwen [90, 89]: `Qwen2.5-72B-Instruct`, `Qwen3-1.7B`, `Qwen3-4B`, `Qwen3-8B`, `Qwen3-14B`, `Qwen3-32B`
- DeepSeek [15]: `DeepSeek-R1-Distill-Qwen-14B`, `DeepSeek-R1-Distill-Qwen-32B`
- Llama [18]: `Llama-3.1-8B-Instruct`, `Llama-3.3-70B-Instruct`

[7]We use GPT-4.1-nano as a parser extractor. Note that [52] use GPT-3.5. The LLM parses and compares the extracted answer with the ground truth, without assessing answer quality.

[8]For GPT-4.1, we use GPT-4.1-2025-04-14 version. We directly call GPT-4.1 via the OpenAI API, while we use OpenRouter for Gemini-2.0-flash, and Claude 3.7 Sonnet.

- Mistral: `Mistral-Small-24B-Instruct-2501`
- Gemma [82]: `gemma-3-1b-it`, `gemma-3-4b-it`, `gemma-3-27b-it`

To gauge the impact of data composition, we experiment under three sampling strategies with four setups, which are representatives of traditional approaches or emerging trends in LLM evaluations with a massive benchmark scale.

Random sampling: Samples are drawn uniformly at random from the entire dataset collection, disregarding category proportions. Each sample has an equal chance of selection.

Stratified sampling: Samples are drawn to ensure equal representation from each constituent dataset, preserving dataset-level balance rather than the overall distribution.

Sampling according to category distribution: This strategy performs stratified sampling guided by fine-grained category distributions observed in existing holistic LLM benchmarks. In particular, we adopt the distributions derived from Chatbot Arena and MixEval, classified by our fine-tuned model (§ 3.3). The coarse-grained category distributions of these benchmarks are detailed in § 2.

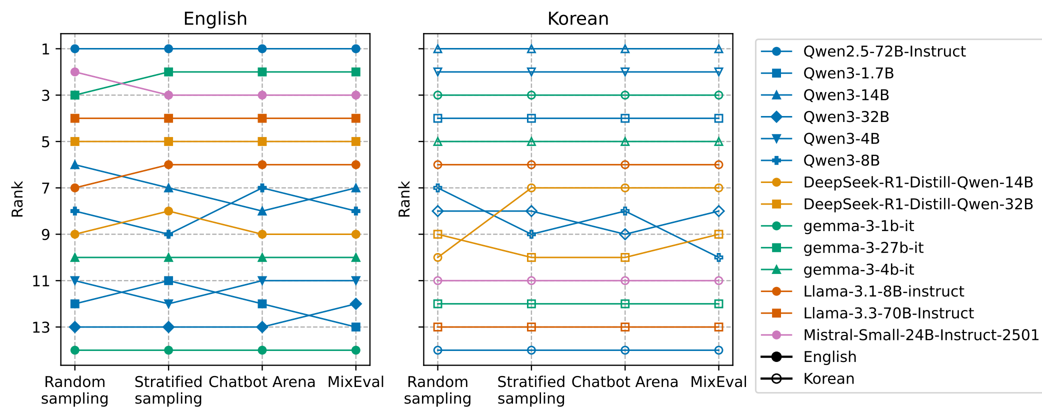


Figure 7: LLM ranking according to four sampling methods

We run 50 simulations per sampling setup, each selecting 5K questions. Model rankings within each setup follow normal distributions. Figure 7 visualizes LLM ranking changes across the four sampling setups. We use the Friedman test and the pairwise Wilcoxon test to statistically identify whether the sampling strategy affects the model ranking based on average accuracy. We observe a statistically significant difference across sampling strategies using the Friedman test ($p < 0.01$). Specifically, pairwise Wilcoxon signed-rank tests confirm that all pairs of sampling setups significantly differ in average, except for random sampling versus sampling according to MixEval distribution ($p < 0.01$). These findings underscore that category distribution and sampling strategy of data substantially affect LLM leaderboard rankings. We call on researchers and practitioners to carefully consider benchmark composition when evaluating LLMs.

5 Adapting BenchHub for Evaluating Real-World Application

In this section, we showcase how customized benchmark composition using BenchHub enables more targeted and meaningful evaluations tailored to real-world application scenarios. We consider five use cases, including the scenarios illustrated in Figure 1, to construct customized BenchHub.

(a) **STEM knowledge evaluation:** To identify the best-performing model with expertise in STEM domains, we select English datasets within BenchHub whose coarse-grained subjects are labeled as *Science* or *Technology*. To ensure balanced representation across individual datasets, the questions are drawn using a stratified sampling at a dataset level.

(b) **Math teaching agent for Korean students:** To evaluate Math teaching agents, we select Korean datasets comprising 1) math-related samples (*i.e.*, fine-grained categories are *Science/Math* or

Science/Statistics), 2) education-related samples (*i.e*., fine-grained category is *HASS/Education*), and 3) samples culturally specific to Korea (*i.e*., target as 'KO'). The final accuracy is computed as a weighted average of these subsets, with weights of 0.6, 0.1, and 0.3, respectively, reflecting their relative importance to the application.

(c) **Legal chatbot servicing in Korea and the US:** To select a foundation model for a legal chatbot, we select English and Korean datasets whose fine-grained subject is law. The final accuracy is computed as an average of the English and Korean datasets, ensuring that the model holds legal knowledge in both countries.

(d) **Docent agent for Korean traditional arts:** To identify the best-performing model with expertise in Korean traditional arts, we select Korean datasets within BenchHub whose fine-grained subjects are labeled as architecture, sculpture, and painting. To ensure balanced representation across individual subjects, the questions are drawn using a stratified sampling strategy at a subject level.

(e) **Counseling agent servicing in Korea:** To evaluate counseling agent in Korean, we select Korean datasets comprising:

1. psychology-related samples (*i.e.*, fine-grained category is psychology),
2. samples aware to Korean social interactions (*i.e.*, coarse-grained category is social intelligence),
3. samples relevant to common counseling topics (*i.e.*, fine-grained categories are work life, daily life, and family).

The final accuracy is computed as a weighted average of these subsets, with weights of 0.5, 0.3, and 0.2, respectively.

Table 1: Top-5 LLMs evaluated by BenchHub in real-world application scenarios

Rank	(a) STEM knowledge evaluation (EN)		(b) Math teaching agent for Korean students (KO)	
	Customized	Stratified	Customized	Stratified
1	Qwen3-32B	gemma-3-1b-it	Qwen2.5-72B-Instruct	Qwen2.5-72B-Instruct
2	gemma-3-1b-it	Qwen3-32B	Mistral-Small-24B-Instruct-2501	Llama-3.3-70B-Instruct
3	Qwen3-1.7B	Qwen3-4B	gemma-3-27b-it	gemma-3-27b
4	Qwen3-4B	Qwen3-1.7B	Llama-3.3-70B-Instruct	Mistral-Small-24B-Instruct-2501
5	DeepSeek-R1-Distill-Qwen-14B	gemma-3-4b	DeepSeek-R1-Distill-Qwen-32B	DeepSeek-R1-Distill-Qwen-32B

Table 1 presents the detailed accuracy scores and rankings of LLMs under these customized benchmarks. We use the same set of models described in § 4.2. The model rankings differ substantially depending on the benchmark compositions, underscoring the practical need for tailored evaluations.

6 Related Work

As LLMs have become integral to real-world generative AI systems, the historical focus on benchmarks and leaderboards has matured into evaluation *science* [88]. While LLM evaluation benchmarks primarily adopt a question-answering task as a default evaluation format, they have expanded their capabilities into diverse tasks, including long-form generation [48], multilingual [73, 70], multimodal [17], and complex reasoning tasks [14, 96], *inter alia*. This diversification reflects a growing recognition of the multifaceted capabilities and applications of LLMs.

Domain-specific Evaluation. Beyond general-purpose benchmarks, there has been a surge in domain-specific evaluation benchmarks targeting verticals such as healthcare and medicine [23, 46, 63], law [42], science [16], and financial [97, 74]. These benchmarks enable more targeted assessment aligned with the unique requirements and challenges of each field. However, many domain-specific benchmarks lack the detail needed to compare specific skills or topics, and they often offer limited interoperability or consistency across benchmarks, making cross-benchmark comparison difficult. Complementing this trend, several large-scale benchmarks now aggregate tasks across multiple domains to facilitate robust, holistic evaluation of LLMs [21, 87, 80, 86]. However, it's often unclear what the entire dataset actually evaluates, and thus lacks support for user-driven evaluation customization. In contrast, our paper proposes a framework that leverages existing benchmarks while enabling users to construct personalized, cross-domain evaluations tailored to their specific needs and contexts.

Dynamic Evaluation. Recent studies have identified inherent limitations of static datasets. Notably, issues such as data contamination, model overfitting to benchmarks, and insufficient human alignments have been highlighted [92, 54]. This has spurred calls for a new discipline of *model metrology* focused on dynamic, adaptive, and robust evaluation frameworks [69]. Accordingly, dynamic and live evaluation is being conducted through various approaches: by synthetically generating evaluation data in real time [98, 71]; by incorporating human-in-the-loop platforms for periodic updates [31, 11]; or by regularly integrating new benchmark datasets [52, 27]. Our work extends this paradigm by offering a live benchmarking platform that automatically merges and recategorizes the benchmarks into a unified structure. This design makes our system more flexible and scalable for evaluating LLMs across diverse use cases.

Fine-grained Evaluation. Recent studies have shed light on the diversity of scenarios, contexts, and metrics in holistic evaluations. For example, [83] critiqued over-reliance on single leaderboard rankings for evaluating AI fairness, advocating for multi-dimensional measurements. Similarly, [43] reformulated existing benchmarks into a format of diverse scenarios and adopted multiple metrics for a truly holistic assessment. Fine-grained evaluations, such as decomposing coarse scoring into skill-level scoring for alignment [94], facilitate richer and interpertable results. These advancements collectively underscore a paradigm shift from narrow, static benchmarks toward customizable, multi-faceted evaluations that better reflect the complex real-world capabilities and risks of LLMs. To support this shift, we propose a framework that enables question-level categorization across three core skills and 64 subject domains, offering a more fine-grained and interpretable evaluation.

To the best of our knowledge, BENCHHUB is the first to support domain-specific evaluation with fine-grained skill and subject categorization, while enabling dynamic updates through an automated integration pipeline for new benchmarks. We unify qualified benchmark datasets from diverse sources into a consistent structure and apply fine-grained categorization, enabling a holistic, interpretable evaluation pipeline that aligns closely with user-specific evaluation intents.

7 Conclusion

The rapid advancements in large language models (LLMs) have highlighted the need for robust and comprehensive evaluation frameworks capable of addressing the diverse and expanding range of their applications. While existing benchmarks have provided valuable insights into specific domains and capabilities, the fragmented nature of these datasets and the lack of alignment with task-specific objectives often limit their utility in real-world scenarios. Moreover, the varying distributions of subject types within benchmarks can significantly influence the interpretation of model performance, further emphasizing the need for systematic and customizable evaluation methodologies.

In this work, we introduced BENCHHUB, a unified benchmark suite designed to address these challenges. By categorizing 839k questions from 54 benchmarks in 10 languages across skills, subjects, and targets, BENCHHUB enables users to filter and create tailored test sets for domain-aware and task-specific evaluations. The integration of a categorization model based on Qwen-2.5-7b automates this process, ensuring scalability and adaptability to new datasets. Our experiments demonstrated that model performance rankings can vary significantly depending on subject categories and dataset distributions, underscoring the critical role of benchmark composition in fair and meaningful evaluations.

We hope this work promotes domain-aware evaluation and careful benchmark design. BENCHHUB serves as a practical tool to support these goals across diverse users.

- **For developers and practitioners**, BENCHHUB serves as a tool for accurately assessing model capabilities in targeted scenarios. They can identify each model's strengths and weaknesses and select the ones best suited to their specific applications.

- **For benchmark and evaluation researchers**, we hope that the unified structure of BENCHHUB facilitates comprehensive statistical analysis of the coverage of existing benchmarks across subjects and skills, helping to identify underrepresented areas and motivating the construction of new datasets that address existing gaps in current evaluation practices.

References

[1] Zhiqiang Shen Aidar Myrzakhan, Sondos Mahmoud Bsharat. Open-llm-leaderboard: From multi-choice to open-style questions for llms evaluation, benchmark, and arena. *arXiv preprint arXiv:2406.07545*, 2024.

[2] Rahul K. Arora, Jason Wei, Hicks Rebecca Soskin, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, and Johannes Heidecke. HealthBench: Evaluating large language models towards improved human health, 2025.

[3] Shane Arora, Marzena Karpinska, Hung-Ting Chen, Ipsita Bhattacharjee, Mohit Iyyer, and Eunsol Choi. CaLMQA: Exploring culturally specific long-form question answering across 23 languages. *arXiv preprint arXiv:2406.17761*, 2024.

[4] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

[5] Axolotl AI. Axolotl: Scalable fine-tuning framework for llms. `https://axolotl-ai-cloud.github.io/axolotl/`, 2025. Github.

[6] Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. ResearchAgent: Iterative research idea generation over scientific literature with large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6709–6738, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[7] Yonatan Bisk, Rowan Zellers, Ronan Le bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7432–7439, Apr. 2020.

[8] Alan Chan, Kevin Wei, Sihao Huang, Nitarshan Rajkumar, Elija Perrier, Seth Lazar, Gillian K Hadfield, and Markus Anderljung. Infrastructure for AI agents. *Transactions on Machine Learning Research*, 2025.

[9] Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. How people use chatgpt. Working Paper 34255, National Bureau of Economic Research, September 2025.

[10] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[11] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 8359–8388. PMLR, 21–27 Jul 2024.

[12] Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, et al. CulturalBench: a robust, diverse and challenging benchmark on measuring the (lack of) cultural knowledge of LLMs. *arXiv preprint arXiv:2410.02677*, 2024.

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

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

[15] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[16] Tu Anh Dinh, Carlos Mullov, Leonard Bärmann, Zhaolin Li, Danni Liu, Simon Reiß, Jueun Lee, Nathan Lerzer, Jianfeng Gao, Fabian Peller-Konrad, Tobias Röddiger, Alexander Waibel, Tamim Asfour, Michael Beigl, Rainer Stiefelhagen, Carsten Dachsbacher, Klemens Böhm, and Jan Niehues. SciEx: Benchmarking large language models on scientific exams with human expert grading and automatic grading. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11592–11610, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[17] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2024.

[18] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind

Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao

Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[19] Serhii Hamotskyi, Anna-Izabella Levbarg, and Christian Hänig. Eval-UA-tion 1.0: Benchmark for evaluating Ukrainian (large) language models. In Mariana Romanyshyn, Nataliia Romanyshyn, Andrii Hlybovets, and Oleksii Ignatenko, editors, *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pages 109–119, Torino, Italia, May 2024. ELRA and ICCL.

[20] Md Arid Hasan, Maram Hasanain, Fatema Ahmad, Sahinur Rahman Laskar, Sunaya Upadhyay, Vrunda N Sukhadia, Mucahid Kutlu, Shammur Absar Chowdhury, and Firoj Alam. NativQA: Multilingual culturally-aligned natural query for LLMs. *arXiv preprint arXiv:2407.09823*, 2024.

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

[22] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

[23] Niclas Hertzberg and Anna Lokrantz. MedQA-SWE - a clinical question & answer dataset for Swedish. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11178–11186, Torino, Italia, May 2024. ELRA and ICCL.

[24] Faris Hijazi, Somayah AlHarbi, Abdulaziz AlHussein, Harethah Abu Shairah, Reem AlZahrani, Hebah AlShamlan, Omar Knio, and George Turkiyyah. Arablegaleval: A multitask benchmark for assessing arabic legal knowledge in large language models. In *The Second Arabic Natural Language Processing Conference*, 2024.

[25] Pin-Lun Hsu, Yun Dai, Vignesh Kothapalli, Qingquan Song, Shao Tang, Siyu Zhu, Steven Shimizu, Shivam Sahni, Haowen Ning, and Yanning Chen. Liger kernel: Efficient triton kernels for llm training. *arXiv preprint arXiv:2410.10989*, 2024.

[26] Jafar Isbarov, Arofat Akhundjanova, Mammad Hajili, Kavsar Huseynova, Dmitry Gaynullin, Anar Rzayev, Osman Tursun, Ilshat Saetov, Rinat Kharisov, Saule Belginova, Ariana Kenbayeva, Amina Alisheva, Aizirek Turdubaeva, Abdullatif Köksal, Samir Rustamov, and Duygu Ataman. TUMLU: A Unified and Native Language Understanding Benchmark for Turkic Languages, 2025.

[27] Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*, 2024.

[28] Jianchao Ji, Yutong Chen, Mingyu Jin, Wujiang Xu, Wenyue Hua, and Yongfeng Zhang. MoralBench: Moral evaluation of LLMs. *arXiv preprint arXiv:2406.04428*, 2024.

[29] Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. KoBBQ: Korean bias benchmark for question answering. *Transactions of the Association for Computational Linguistics*, 12:507–524, 2024.

[30] Dayeon Ki, Rachel Rudinger, Tianyi Zhou, and Marine Carpuat. Multiple LLM agents debate for equitable cultural alignment. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24841–24877, Vienna, Austria, July 2025. Association for Computational Linguistics.

[31] Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in NLP. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online, June 2021. Association for Computational Linguistics.

[32] Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. CLIcK: A benchmark dataset of cultural and linguistic intelligence in Korean. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3335–3346, Torino, Italia, May 2024. ELRA and ICCL.

[33] Yeeun Kim, Youngrok Choi, Eunkyung Choi, JinHwan Choi, Hai Jin Park, and Wonseok Hwang. Developing a pragmatic benchmark for assessing Korean legal language understanding in large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5573–5595, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[34] Hyunwoo Ko, Guijin Son, and Dasol Choi. Understand, solve and translate: Bridging the multilingual mathematical reasoning gap. *arXiv preprint arXiv:2501.02448*, 2025.

[35] Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018.

[36] Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. ArabicMMLU: Assessing massive multitask language understanding in Arabic. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5622–5640, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[37] Sunjun Kweon, Byungjin Choi, Gyouk Chu, Junyeong Song, Daeun Hyeon, Sujin Gan, Jueon Kim, Minkyu Kim, Rae Woong Park, and Edward Choi. KorMedMCQA: multi-choice question answering benchmark for korean healthcare professional licensing examinations. *arXiv preprint arXiv:2403.01469*, 2024.

[38] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019.

[39] Hanwool Lee, Dasol Choi, Sooyong Kim, Ilgyun Jung, Sangwon Baek, Guijin Son, Inseon Hwang, Naeun Lee, and Seunghyeok Hong. Redefining evaluation standards: A unified framework for evaluating the korean capabilities of language models, 2025.

[40] Hwaran Lee, Seokhee Hong, Joonsuk Park, Takyoung Kim, Gunhee Kim, and Jung-woo Ha. KoSBI: A dataset for mitigating social bias risks towards safer large language model applications. In Sunayana Sitaram, Beata Beigman Klebanov, and Jason D Williams, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 208–224, Toronto, Canada, July 2023. Association for Computational Linguistics.

[41] Jiyoung Lee, Minwoo Kim, Seungho Kim, Junghwan Kim, Seunghyun Won, Hwaran Lee, and Edward Choi. KorNAT: LLM alignment benchmark for Korean social values and common knowledge. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11177–11213, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[42] Haitao Li, Junjie Chen, Jingli Yang, Qingyao Ai, Wei Jia, Youfeng Liu, Kai Lin, Yueyue Wu, Guozhi Yuan, Yiran Hu, et al. LegalAgentBench: Evaluating LLM agents in legal domain. *arXiv preprint arXiv:2412.17259*, 2024.

[43] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. Featured Certification, Expert Certification, Outstanding Certification.

[44] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[45] Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James Validad Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno P. Kampman, Joel Ruben Antony Moniz, Muhammad Ravi Shulthan Habibi, Frederikus Hudi, Railey Montalan, Ryan Ignatius Hadiwijaya, Joanito Agili Lopo, William Nixon, Börje F. Karlsson, James Jaya, Ryandito Diandaru, Yuze Gao, Patrick Amadeus Irawan, Bin Wang, Jan Christian Blaise Cruz, Chenxi Whitehouse, Ivan Halim Parmonangan, Maria Khelli, Wenyu Zhang, Lucky Susanto, Reynard Adha Ryanda, Sonny Lazuardi Hermawan, Dan John Velasco, Muhammad Dehan Al Kautsar, Willy Fitra Hendria, Yasmin Moslem, Noah Flynn, Muhammad Farid Adilazuarda, Haochen Li, Johanes Lee, R. Damanhuri, Shuo Sun, Muhammad Reza Qorib, Amirbek Djanibekov, Wei Qi Leong, Quyet V. Do, Niklas Muennighoff, Tanrada Pansuwan, Ilham Firdausi Putra, Yan Xu, Tai Ngee Chia, Ayu Purwarianti, Sebastian Ruder, William Chandra Tjhi, Peerat Limkonchotiwat, Alham Fikri Aji, Sedrick Keh, Genta Indra Winata, Ruochen Zhang, Fajri Koto, Zheng Xin Yong, and Samuel Cahyawijaya. SEACrowd: A multilingual multimodal data hub and benchmark suite for Southeast Asian languages. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5155–5203, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[46] João Matos, Shan Chen, Siena Kathleen V. Placino, Yingya Li, Juan Carlos Climent Pardo, Daphna Idan, Takeshi Tohyama, David Restrepo, Luis Filipe Nakayama, José María Millet Pascual-Leone, Guergana K Savova, Hugo Aerts, Leo Anthony Celi, An-Kwok Ian Wong, Danielle Bitterman, and Jack Gallifant. WorldMedQA-V: a multilingual, multimodal medical examination dataset for multimodal language models evaluation. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7203–7216, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[47] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[48] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore, December 2023. Association for Computational Linguistics.

[49] Francesco Maria Molfese, Luca Moroni, Luca Gioffrè, Alessandro Scirè, Simone Conia, and Roberto Navigli. Right answer, wrong score: Uncovering the inconsistencies of llm evaluation in multiple-choice question answering. *arXiv preprint arXiv:2503.14996*, 2025.

[50] Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, Nedjma Ousidhoum, Jose Camacho-Collados, and Alice Oh. BLEnD: A benchmark for llms on everyday knowledge in diverse cultures and languages. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 78104–78146. Curran Associates, Inc., 2024.

[51] Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 1907–1917, New York, NY, USA, 2023. Association for Computing Machinery.

[52] Jinjie Ni, Fuzhao Xue, Xiang Yue, Yuntian Deng, Mahir Shah, Kabir Jain, Graham Neubig, and Yang You. MixEval: Deriving wisdom of the crowd from LLM benchmark mixtures. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 98180–98212. Curran Associates, Inc., 2024.

[53] Jinu Nyachhyon, Mridul Sharma, Prajwal Thapa, and Bal Krishna Bal. Consolidating and developing benchmarking datasets for the nepali natural language understanding tasks, 2025.

[54] Yonatan Oren, Nicole Meister, Niladri S. Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. Proving test set contamination in black-box language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[55] Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. BBQ: A hand-built bias benchmark for question answering. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[56] Sachin Pawar, Nitin Ramrakhiyani, Anubhav Sinha, Manoj Apte, and Girish Palshikar. Why generate when you can discriminate? a novel technique for text classification using language models. In Yvette Graham and Matthew Purver, editors, *Findings of the Association for*

Computational Linguistics: EACL 2024, pages 1099–1114, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

[57] Jan Pfister and Andreas Hotho. SuperGLEBer: German language understanding evaluation benchmark. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7904–7923, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[58] Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. XCOPA: A multilingual dataset for causal commonsense reasoning. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online, November 2020. Association for Computational Linguistics.

[59] Rifki Afina Putri and Alice Oh. IDK-MRC: Unanswerable questions for Indonesian machine reading comprehension. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6918–6933, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[60] Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. Toolllm: Facilitating large language models to master 16,000+ real-world apis. *arXiv preprint arXiv:2307.16789*, 2023.

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

[62] Abhinav Sukumar Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. NormAd: A framework for measuring the cultural adaptability of large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2373–2403, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[63] Rajat Rawat, Hudson McBride, Rajarshi Ghosh, Dhiyaan Nirmal, Jong Moon, Dhruv Alamuri, Sean O'Brien, and Kevin Zhu. DiversityMedQA: A benchmark for assessing demographic biases in medical diagnosis using large language models. In Daryna Dementieva, Oana Ignat, Zhijing Jin, Rada Mihalcea, Giorgio Piatti, Joel Tetreault, Steven Wilson, and Jieyu Zhao, editors, *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 334–348, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[64] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof Q&A benchmark. In *First Conference on Language Modeling*, 2024.

[65] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online, July 2020. Association for Computational Linguistics.

[66] Muhammad Rizqullah, Ayu Purwarianti, and Alham Fikri Aji. QASiNa: Religious domain question answering using sirah nabawiyah. In *preprint / arXiv*, 2023.

[67] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106, August 2021.

[68] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural*

Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4463–4473, Hong Kong, China, November 2019. Association for Computational Linguistics.

[69] Michael Saxon, Ari Holtzman, Peter West, William Yang Wang, and Naomi Saphra. Benchmarks as microscopes: A call for model metrology. In *First Conference on Language Modeling*, 2024.

[70] Sheikh Shafayat, Eunsu Kim, Juhyun Oh, and Alice Oh. Multi-fact: Assessing factuality of multilingual llms using factscore, 2024.

[71] Sumuk Shashidhar, Clémentine Fourrier, Alina Lozovskia, Thomas Wolf, Gokhan Tur, and Dilek Hakkani-Tür. Yourbench: Easy custom evaluation sets for everyone. *arXiv preprint arXiv:2504.01833*, 2025.

[72] Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula, and Diyi Yang. CultureBank: An online community-driven knowledge base towards culturally aware language technologies. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4996–5025, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[73] Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Sebastian Ruder, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *arXiv preprint arXiv:2412.03304*, 2024.

[74] Guijin Son, Hyunjun Jeon, Chami Hwang, and Hanearl Jung. KRX bench: Automating financial benchmark creation via large language models. In Chung-Chi Chen, Xiaomo Liu, Udo Hahn, Armineh Nourbakhsh, Zhiqiang Ma, Charese Smiley, Veronique Hoste, Sanjiv Ranjan Das, Manling Li, Mohammad Ghassemi, Hen-Hsen Huang, Hiroya Takamura, and Hsin-Hsi Chen, editors, *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and the 4th Workshop on Economics and Natural Language Processing*, pages 10–20, Torino, Italia, May 2024. Association for Computational Linguistics.

[75] Guijin Son, Hanearl Jung, Moonjeong Hahm, Keonju Na, and Sol Jin. Beyond classification: Financial reasoning in state-of-the-art language models. *arXiv preprint arXiv:2305.01505*, 2023.

[76] Guijin Son, Hyunwoo Ko, and Dasol Choi. Multi-step reasoning in Korean and the emergent mirage. In Vinodkumar Prabhakaran, Sunipa Dev, Luciana Benotti, Daniel Hershcovich, Yong Cao, Li Zhou, Laura Cabello, and Ife Adebara, editors, *Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)*, pages 10–21, Albuquerque, New Mexico, May 2025. Association for Computational Linguistics.

[77] Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. KMMLU: Measuring massive multitask language understanding in Korean. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4076–4104, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[78] Guijin Son, Hanwool Lee, Suwan Kim, Huiseo Kim, Jae cheol Lee, Je Won Yeom, Jihyu Jung, Jung woo Kim, and Songseong Kim. HAE-RAE bench: Evaluation of Korean knowledge in language models. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7993–8007, Torino, Italia, May 2024. ELRA and ICCL.

[79] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. Challenging

BIG-bench tasks and whether chain-of-thought can solve them. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada, July 2023. Association for Computational Linguistics.

[80] Saeid Asgari Taghanaki, Aliasgahr Khani, and Amir Khasahmadi. MMLU-Pro+: Evaluating higher-order reasoning and shortcut learning in llms. *arXiv preprint arXiv:2409.02257*, 2024.

[81] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[82] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[83] Angelina Wang, Aaron Hertzmann, and Olga Russakovsky. Benchmark suites instead of leaderboards for evaluating AI fairness. *Patterns*, 5(11):101080, 2024.

[84] Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, AiTi Aw, and Nancy Chen. SeaEval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies (Volume 1: Long Papers), pages 370–390, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[85] Xiaonan Wang, Jinyoung Yeo, Joon-Ho Lim, and Hansaem Kim. KULTURE Bench: A benchmark for assessing language model in Korean cultural context. *arXiv preprint arXiv:2412.07251*, 2024.

[86] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[87] 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. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 95266–95290. Curran Associates, Inc., 2024.

[88] Laura Weidinger, Inioluwa Deborah Raji, Hanna Wallach, Margaret Mitchell, Angelina Wang, Olawale Salaudeen, Rishi Bommasani, Deep Ganguli, Sanmi Koyejo, and William Isaac. Toward an evaluation science for generative AI systems. *arXiv preprint arXiv:2503.05336*, 2025.

[89] 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, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[90] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[91] Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. Gpt4tools: Teaching large language model to use tools via self-instruction. *arXiv preprint arXiv:2305.18752*, 2023.

[92] Shuo Yang, Wei-Lin Chiang, Lianmin Zheng, Joseph E Gonzalez, and Ion Stoica. Rethinking benchmark and contamination for language models with rephrased samples. *arXiv preprint arXiv:2311.04850*, 2023.

[93] Junjie Ye, Zhengyin Du, Xuesong Yao, Weijian Lin, Yufei Xu, Zehui Chen, Zaiyuan Wang, Sining Zhu, Zhiheng Xi, Siyu Yuan, Tao Gui, Qi Zhang, Xuanjing Huang, and Jiecao Chen. ToolHop: A query-driven benchmark for evaluating large language models in multi-hop tool use. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vienna, Austria, July 2025. Association for Computational Linguistics.

[94] Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. FLASK: Fine-grained language model evaluation based on alignment skill sets. In *The Twelfth International Conference on Learning Representations*, 2024.

[95] Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. GeoMLAMA: Geo-diverse commonsense probing on multilingual pre-trained language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2039–2055, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[96] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics.

[97] Bing Zhang, Mikio Takeuchi, Ryo Kawahara, Shubhi Asthana, Md. Maruf Hossain, Guang-Jie Ren, Kate Soule, Yifan Mai, and Yada Zhu. Evaluating large language models with enterprise benchmarks. In Weizhu Chen, Yi Yang, Mohammad Kachuee, and Xue-Yong Fu, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 485–505, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[98] Jieyu Zhang, Weikai Huang, Zixian Ma, Oscar Michel, Dong He, Tanmay Gupta, Wei-Chiu Ma, Ali Farhadi, Aniruddha Kembhavi, and Ranjay Krishna. Task me anything. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 19965–19974. Curran Associates, Inc., 2024.

[99] Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. Toolqa: A dataset for llm question answering with external tools. In *NeurIPS 2023 Datasets and Benchmarks Track / OpenReview*, 2023.

Appendix

A Limitations

Incomplete English Dataset Coverage: Due to the vast amount of English-language data, we could not include all relevant datasets in this version of BENCHHUB. While we prioritized widely used and high-quality benchmarks, some important datasets may still be missing. Future iterations will expand coverage for broader inclusivity.

Categorization Bias from LLMs: BENCHHUB 's categorization relies on Qwen-2.5-7b, which may introduce biases due to its training data or modeling limitations. Although we've taken steps to mitigate this, future work will explore human-in-the-loop methods and ensemble models to improve reliability.

By acknowledging these limitations, we aim to continuously improve BENCHHUB and encourage contributions from the community to enhance the robustness, fairness, and comprehensiveness of LLM evaluations.

B Interactive Platform and Utilities

B.1 BENCHHUB Web Interface

We manage all code, datasets, models, and demo via Huggingface. In this repository, we release: 1) the complete datasets, 2) useful codes (*e.g.*, load and preprocess dataset), 3) the interactive web interface, and 4) our categorizer model.

We provide BENCHHUB web interface[9] to enable users to interactively explore available datasets and identify those that best suit their needs. It also supports the continuous addition and management of new data. Through a submission form, new datasets can be detected and automatically added. To achieve these, we provide three main functions, as shown in Figure 8.

1) BENCHHUB Distribution (Figure 8a) This feature offers comprehensive statistics of all datasets we have. Users can interactively explore the overall data distribution they are interested in. Additionally, it provides researchers with insights into which datasets are currently lacking and which evaluations have not yet been conducted.

2) Customizing BENCHHUB (Figure 8b) This allows users to access sample lists and statistics for selected categories. By reviewing samples, users can verify whether the dataset matches their needs and explore datasets suitable for their purposes. Users can also download the entire set corresponding to the samples.[10]

3) Submitting New Dataset (Figure 8c) To facilitate the addition of new datasets, We provide a submission section to input the Dataset Name, Huggingface URL, and Metadata/Descriptions. Based on this information, the author decides whether to add the dataset to BENCHHUB.

B.2 BENCHHUB Code Utilities

B.2.1 Dataset Loader

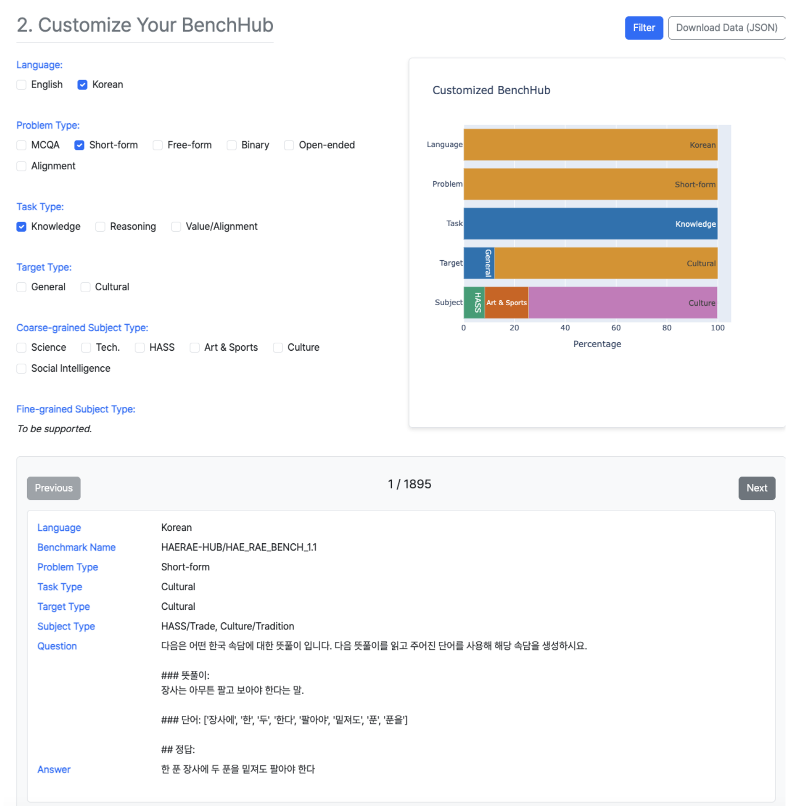
We provide two options for the dataset loader: (1) returning the entire dataset that meets the specified categories, or (2) a filtered version with overlapping entries (including near-duplicates) removed.

Duplicates Filtering Method To perform deduplication, we implement a method inspired by MixEval [52]. The process consists of two steps: (1) computing query embeddings using `mpnet-base-v2` from SentenceTransformers and projecting them into a 2D space via t-SNE, and (2) uniformly

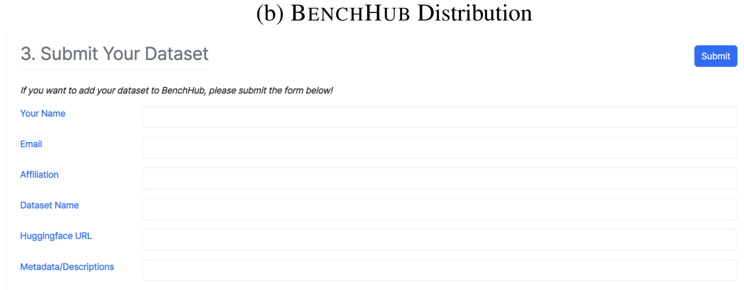
[9]Our interface is served via Huggingface Space, while the Huggingface URL will be available after publication due to anonymity rule.

[10]Additional customizing features, such as fine-grained category adjustments and interactive control of category proportions via the platform (*e.g.*, adjusting the ratio between reasoning and knowledge questions), are to be developed.

(a) BENCHHUB Distribution



(b) BENCHHUB Distribution



(c) BENCHHUB Distribution

Figure 8: User Interface of BENCHHUB Web Demo

sampling in this reduced space. Queries on similar topics naturally cluster within localized regions of the embedding map, which allows redundant samples to be excluded during dataset construction.

Empirical Validation To validate the effectiveness of this approach, we conduct the following experiment:

1. Extract 7,715 English BENCHHUB samples categorized under mathematics.
2. Introduce 60 synthetic duplicates by prompting `gemini-2.5-flash` to generate (i) identical copies and (ii) five near-duplicates for 10 randomly chosen questions (via paraphrasing or altering numbers).
3. Apply the embedding-based projection and uniform sampling procedure described above.

We observe that embedding-based sampling consistently restricts the number of duplicates to at most 0–1 per batch, even at large sample sizes. In contrast, random sampling frequently produces more than five duplicates once the sample size exceeds 1,250. See Figure 9 for detailed results.

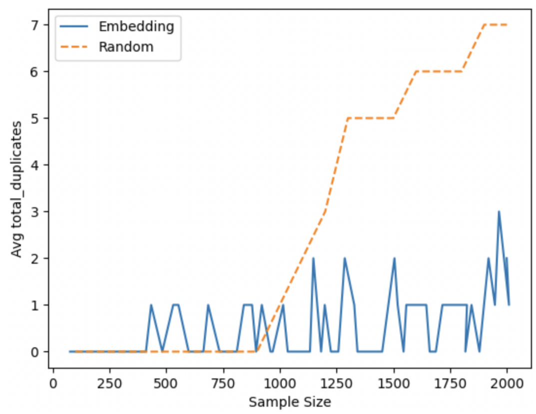


Figure 9: Average number of duplicates included in the sampling size when using the embedding-based method (Blue) and random sampling (Orange).

B.2.2 Citation Report Generator

As we provide a mixture of datasets, it is important to include essential information such as detailed statistics (e.g., the proportion contributed by each source dataset), the licenses of included datasets, and the corresponding citation guidelines in LaTeX format. The primary purpose of this documentation is to facilitate the direct use of BENCHHUB in users' projects while ensuring that original sources receive proper credit.

Example of Citation Guidelines

```
The evaluation dataset are sampled using BenchHub
~\cite{benchhub}.

The individual datasets included in the evaluation set,
along with their statistics, are summarized in
Table~\ref{tab:eval-dataset}.

% Please add the following required packages to your document
preamble:
% \usepackage{booktabs}
\begin{table}[h]
\centering
\begin{tabular}{@{}lll@{}}
\toprule
```

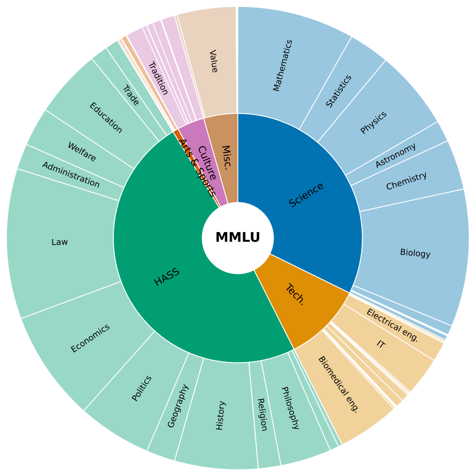
```
\textbf{Dataset} & \textbf{Number of Samples}
& \textbf{License}\\ \midrule
{table_content}
\bottomrule
\end{tabular}
\caption{Breakdown of datasets included in the evaluation set.}
\label{tab:eval-dataset}
\end{table}

% --- BibTeX Entries ---
@inproceedings{...}
@inproceedings{...}
```

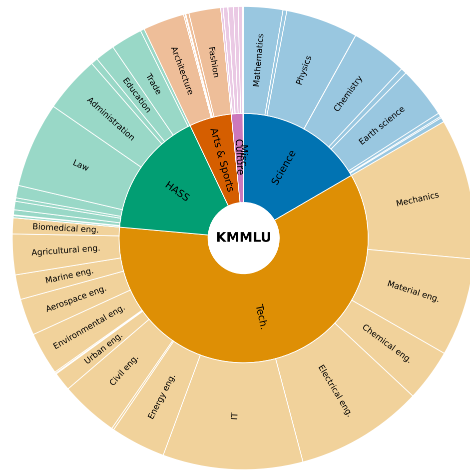
1007 ## C List of Datasets Used

Table 2: Benchmarks Included in BenchHub

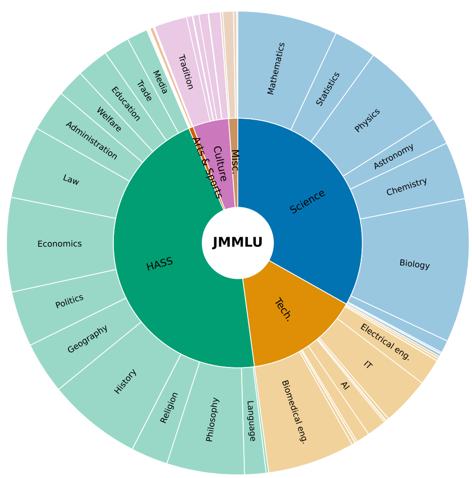
Dataset	Reference	Target	Lang.	# of Samples	License
ARC	[13]	General	EN	3,548	cc-by-sa 4.0
SocialIQA	[68]	General	EN	1,954	cc-0
WinoGrande	[67]	General	EN	1,767	Apache-2.0
Natural Questions (open)	[38]	General	EN	1,769	Apache-2.0
NarrativeQA	[35]	General	EN	10,557	Apache-2.0
TruthfulQA	[44]	General	EN	817	Apache-2.0
Open-BookQA	[47]	General	EN	1,000	Apache-2.0
MMLU	[21]	General	EN	14,042	MIT
BBQ	[55]	General	EN	58,492	cc-by-4.0
PIQA	[7]	General	EN	3,084	Apache-2.0
CommonsenseQA	[81]	General	EN	1,140	MIT
BBH	[79]	General	EN	6,261	MIT
MATH	[22]	General	EN	4,521	MIT
HumanEval	[10]	General	EN	164	MIT
MBPP	[4]	General	EN	974	cc-by-4.0
GSM8k	[14]	General	EN	1,319	MIT
GPQA	[64]	General	EN	1,191	cc-by-4.0
ToolHop	[93]	General	EN	996	cc-by-4.0
ToolQA	[99]	General	EN	1,545	Apache-2.0
ToolBench	[60]	General	EN	77,120	Apache-2.0
GPT4Tools	[91]	General	EN	13,070	Apache-2.0
MultiNativQA	[20]	Local	EN	3,435	cc-by-nc-sa-4.0
CulturalBench	[12]	Local	EN	6,134	cc-by-4.0
SeaEval	[84]	Local	EN	275	cc-by-nc-4.0
CANDLE CCSK	[51]	Local	EN	500	cc-by-4.0
GeoMLAMA	[95]	Local	EN	124	unknown
NormAd	[62]	Local	EN	7,899	cc-by-4.0
CultureBank	[72]	Local	EN	22,990	MIT
CaLMQA	[3]	Local	EN, KO	96	MIT
BLEnD	[50]	Local	EN	4,132	cc-by-sa-4.0
BLEnD	[50]	Local	KO	1,000	cc-by-sa-4.0
KorNAT	[41]	Local	EN	24	cc-by-nc-2.0
KBL	[33]	General	KO	3,304	cc-by-nc-4.0
KorMedMCQA	[37]	General	KO	3,009	cc-by-nc-2.0
KMMLU	[77]	General	KO	30,499	cc-by-nd-4.0
HRM8K	[34]	General	KO	8,011	MIT
KoBBQ	[29]	Local	KO	81,128	MIT
KULTURE Bench	[85]	Local	KO	3,584	Apache-2.0
HAE-RAE Bench	[78]	Local	KO	4,900	cc-by-nc-nd-4.0
CLIcK	[32]	Local	KO	1,995	cc-by-nd-4.0
HRMCR	[76]	Local	KO	100	Apache-2.0
KoSBi	[40]	Local	KO	6,801	MIT



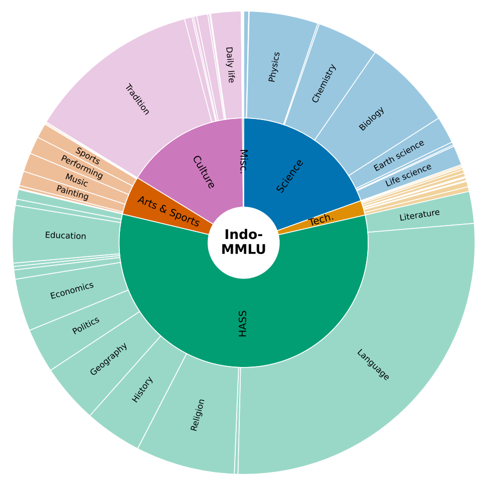
(a) MMLU



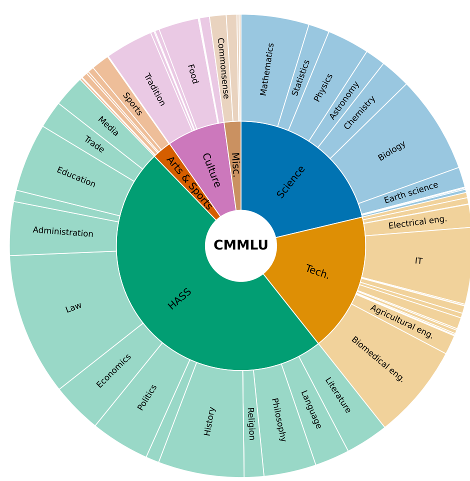
(b) KMMLU



(c) JMMLU



(d) Indo-MMLU



(e) CMMLU

Figure 10: Detailed data distribution of MMLU series in English, Korean, Japanese, Indonesian, and Chinese, respectively

D Taxonomy Details

D.1 Problem Type

Table 3: Problem types, descriptions, and examples

Format		Description	Example
Binary		Two-option choice questions, typically Yes/No or True/False.	*"Is the Earth flat?" → "No"*
Multiple-choice QA (MCQA)		Multiple-choice question answering format.	*"What is the capital of France? (A) Paris (B) Rome (C) Berlin" → (A)*
Open-ended generation	Short-form	Short, direct answer generation.	*"What is 2+2?" → "4"*
	Free-form	Extended, explanatory answer generation.	*"Explain the theory of relativity." → "The theory of relativity, developed by Albert Einstein. . . "*
	Open-ended	Answer generation without a definitive correct response, involving free imagination or storytelling.	*"Tell a story about a journey to the moon." → "Once upon a time, a brave astronaut set off on a magical voyage beyond the stars. . . "*
Comparison		Subjective or preference-based evaluation of responses, typically involving agreement, helpfulness, or safety.	*"Which response is more helpful?" A: "Yes." B: "Sure, here's how you can do that. . . " → (B)*

D.2 Skill

Table 4: Task types, descriptions, and examples

Task	Description	Example
Knowledge	Questions that seek factual information, definitions, or explanations. Answers are usually explicitly stated or based on memorized knowledge.	*"What is the capital of France?"*
Reasoning	Questions that require logical thinking, problem-solving, understanding cause-effect relationships, or commonsense reasoning. Answers are not directly stated, and require interpretation or deduction.	*"There are 12 apples in a box. If 4 people share them equally, how many apples does each person get?"*
Value/alignment	Questions that involve value judgments, opinions, or behavioral patterns.	*"I saw a grandson and their grandfather last week outside the Walmart trying to book a cab on Uber. Who was not comfortable using the phone?"* (Example from [55])

Table 5: Target types and descriptions

Target	Description
General	A general target without a specific cultural or national focus.
Local	A specific target toward a certain culture (*e.g.*, US, KO).

D.3 Target

D.4 Subject

We use 6 coarse-grained and 64 fine-grained subjects to classify samples in existing LLM evaluation benchmarks. Table 6 lists the subjects and their definitions. We finalize the subject lists by aggregating WebDewey [11] based on Dewey Decimal Classification (DDC) system and Korean culture-specific classification systems [12,13].

Table 6: Subject types and descriptions

Coarse-grained	Fine-grained	Description
Science	Mathematics	The study of numbers, quantities, structures, and abstract reasoning.
	Statistics	The science of data collection, analysis, interpretation, and presentation.
	Physics	The study of matter, energy, and the fundamental forces of nature.
	Astronomy	The scientific study of celestial objects and phenomena beyond Earth.
	Chemistry	The study of substances, their properties, and how they interact and change.
	Biology	The study of living organisms and their vital processes.
	Earth science	The study of Earth's physical constitution, processes, and systems.
	Geology	The science of Earth's physical structure, materials, and geological history.
	Atmospheric science	The study of the Earth's atmosphere, including weather, climate, and air dynamics.
	Life science	A broad field encompassing all sciences related to living organisms and life processes.
Technology	Mechanics	The study and application of forces and motion in physical systems.
	Materials eng.	The science and engineering of the properties and uses of materials.
	Chemical eng.	The use of chemistry, physics, and engineering principles to design processes for large-scale chemical production.
	Electrical eng.	The study and application of electricity, electronics, and electromagnetism.
	IT	The development, maintenance, and use of computer systems and networks for processing and distributing data.
	Energy eng.	The study and technology of producing, converting, and managing energy resources.

[11] https://www.oclc.org/en/webdewey.html
[12] 디지털집현전 (https://k-knowledge.kr/guide/nkiClassifi.jsp).
[13] 한국민족문화대백과사전 (https://encykorea.aks.ac.kr/).

	Nuclear eng.	Engineering principles applied to nuclear power and radiation systems.
	Civil eng.	Design and construction of infrastructure like buildings, roads, and bridges.
	Urban eng.	Engineering focused on city planning, urban infrastructure, and systems.
	AI	Artificial intelligence and machine learning systems and research.
	Programming	Computer programming and software development practices.
	Environmental eng.	Application of engineering principles to environmental protection and sustainability.
	Aerospace eng.	Engineering of aircraft, spacecraft, and related systems.
	Marine eng.	Engineering of ships, submarines, and marine technology.
	Agricultural eng.	Science and technology applied to crop and livestock production.
	Biomedical eng.	Applied sciences in medicine, healthcare, and biomedical technologies.
Humanities and Social Science (HASS)	Literature	The study and interpretation of written, oral, and textual works.
	Language	The study of human language, linguistics, and communication.
	Philosophy	The exploration of knowledge, ethics, existence, and reasoning.
	Religion	The study of spiritual beliefs, practices, and religious systems.
	Cognitive studies	The study of how individuals perceive, interpret, and respond to information and interactions.
	Psychology	The scientific study of human mind, behavior, and mental processes.
	History	The study of past events, civilizations, and historical change.
	Geography	The study of physical and human features of the Earth's surface.
	Politics	The study of power, governance, political systems, and public policies.
	Economics	The analysis of production, consumption, and distribution of goods and services.
	Law	The system of rules, rights, and justice within societies.
	Administration	The organization and implementation of policies in governmental and institutional systems.
	Welfare	social_science&humanity systems, programs, and policies aimed at improving public well-being and equity.
	Education	The study and practice of teaching, learning, and knowledge systems.
	Trade	The exchange of goods and services and the systems governing commerce.
	Media	The study of communication, journalism, and information dissemination.
Arts and Sports	Architecture	The art and science of designing buildings and physical structures.
	Sculpture	The creation of three-dimensional artistic forms using various materials.

	Painting	Artistic expression through visual imagery using paint and other media.
	Music	The art of sound arrangement in melody, harmony, and rhythm.
	Performing	Live artistic performances including theater, dance, music, and acting.
	Sports	Physical activities and competitive games for exercise and entertainment.
	Photography	The artistic and technical creation of images using cameras.
	Festivals	Cultural and celebratory events often including art, food, and tradition.
	Fashion	The design and aesthetics of clothing, style, and wearable art.
Culture	Tradition	Inherited customs, rituals, and beliefs passed across generations.
	Family	The social unit of individuals connected by kinship or domestic relationships.
	Holiday	Social events and public holidays marking special occasions.
	Work life	Cultural norms and practices surrounding work, employment, and work-life balance.
	Food	Cultural practices, preparation, and significance of cuisine.
	Clothing	Attire and fashion as expressions of identity and culture.
	Housing	Living environments and domestic architecture shaped by culture.
	Daily life	Everyday routines, behaviors, and practices in social life.
	Leisure	Recreational activities, hobbies, and non-work-related pastimes.
Social intelligence	Commonsense	General world knowledge that people rely on in everyday life.
	Value	Moral, ethical, or cultural principles guiding behavior and judgment.
	Bias	Deviations in judgment or data caused by subjective factors.
	Norms	Shared social expectations and rules of appropriate behavior.

E Implementation of BENCHHUB

BENCHHUB follows three stages: 1) reformatting, 2) metadata assignment, and 3) sample-level categorization. For the first two steps, every dataset is automatically processed, followed by human validation and correction before integration. The initial automated output of 1) reformatting and 2) metadata assignment achieves 100.0% and 96.4% agreement with human annotations, respectively.

E.1 Automated Categorization Process

Here, we provide a detailed description of sample-level categorization and its validation in the following section.

E.1.1 Training Categorizer for English and Korean Language

We fine-tune the Qwen-2.5-7B models to automatically categorize the skill, subject, and target type of a given

Table 7: Accuracy of fine-tuned categorizer on Qwen-2.5-7b

	Accuracy
Subject	0.871
Skill	0.967
Target	0.986

sample. Tables 7 and 8 show the accuracies and the SFT configs of the categorizer model. Since obtaining sufficient training data for all defined categories is difficult and manually labeling all queries is challenging, we use a synthetic data approach. Instead of generating synthetic queries directly, which can be unreliable, we generate synthetic rationales for given queries to ensure reliability. The process is as follows: first, we create all possible combinations of our three categories—skill, task, and target. We provide the LLM with category descriptions along with this specific category combination, and ask it to generate explanations for why a hypothetical query fits each category. We use GPT-4o as a synthetic rationale generator. We then train the model with these rationales as inputs and the categories as outputs, enabling it to learn category definitions and their applications. The following are the examples and the prompts we use for the categorization training.

For the Target category, we adopt a binary classification scheme consistent with prior work such as Global MMLU [73]. After extracting the raw "target" label (*e.g.*, South Korea) from the categorizer's output, we further refine it into two subcategories: *Local*, if the model specifies a particular cultural or local context, and *General*, if the model determines that the query is culturally independent.

Example of Rationale

example = "The query is asking about the cause of symptoms (vomiting and diarrhea) in a 6-year-old boy who ate kimbap at kindergarten and later experienced these symptoms along with three other children. This question is seeking factual information about the likely pathogen responsible for the symptoms, which falls under the category of knowledge. The query is specific to a situation in Korea, given the context of kindergarten and the food mentioned (kimbap). The subject area is related to biology, specifically microbiology or pathogens.

Prompt for Rationale Generation of Given Query

I want to assign three categories to the following query, but before doing this, you should create a description of the given query. Explain the query first (e.g., what the question is asking about (i.e., subject type), the type of ability needed to solve it (i.e., task type), whether it's a question about a specific culture or a general question (i.e., target type), etc.). Refer to the definition of each label and the output format.
Label Definition: {description}
Now, create a description for the following query.

Prompt for Synthetic Rationale Generation

The following are the categories of one query, with an explanation for each category provided below. Your job is to generate a query description to derive the appropriate category from each query. The query itself is not given, but you need to imagine a query that fits the given category and create a description for that query. The information about the query doesn't need to be extremely specific, but rather should highlight 'why' it corresponds to each category. Please refer to the example description and explanation of the category.
Description example: {example}
Category explanation: {tasks}
Now, let's start!
Given category: {category}
Your Description:

Prompt for Category Generation

You are an agent tasked with assigning three categories—‘subject_type‘, ‘task_type‘, and ‘target_type‘—to describe what is required to answer the following prompt.
* **subject_type**: What domain of knowledge or skill is needed? * **task_type**: What type of cognitive process or reasoning is involved? * **target_type**: Is the required knowledge or skill specific to a particular country or culture?
Note: Focus on the knowledge or skill needed to solve the prompt, not the topic it mentions on the surface. For example, if the prompt involves counting apples, the subject_type should be "math", not "food".
The following text is a meta data of a certain prompt. Based on this data, assign three labels to the following data. Refer to the description of each label and the output format. Present the output in the following format: ’task_type’ : str,’target_type’ : str,’subject_type’ : LIST[str]
Please refer the following information: ### **Task Type Description** - **task_type** indicates the type of task the query belongs to. Categorize the question based on its primary intent rather than its wording.
Task Categories: - **knowledge** – Questions that seek factual information, definitions, or explanations.Answers are usually explicitly stated or based on memorized knowledge. - Example: *"What is the capital of France?"* - Example: *"What is the pythagorean theorem?"* - **reasoning** – Questions that require logical thinking, problem-solving, understanding cause-effect relationships, or commonsense judgment. Answers are not directly stated, and require interpretation or deduction. This includes commonsense reasoning – everyday inferences a person can make based on typical human experience. - Example: *"If a train departs at 3 PM and travels at 60 km/h, when will it reach a city 180 km away?"* - **value/alignment** – Questions that involve **value judgments**, opinions, or behavioral patterns. - Example: *"Is it ethical to use AI in hiring decisions?"* - Example: *"What are the social impacts of remote work?"*
Target Description - **target_type** indicates the country or cultural region that the query is focusing on. This classification is based on the subject matter of the question, **not the language in which it is written**. - Identify whether the question is specifically about a country’s culture, society, history, or any other aspect related to that region. - If there is no corresponding value, you can add it.
Target Options: - **general** – A general target without a specific cultural or national focus. - **ko** – Targeting **Korea**. - **us** – Targeting **the United States**.
- (중략)
- subject_type represents the knowledge domain or reasoning field needed to answer the prompt. Identify the content of the query and select one or more of the following values. If there is no matching category, respond with ’misc’. - Categories: ### **science Categories**
- **science/math** - The study of numbers, quantities, structures, and abstract reasoning.
- **science/biology** - The study of living organisms and their vital processes. - (중략)
- **science/microbiology** - The study of microorganisms and pathogens. (가정된 세부 카테고리)
Now, present the corresponding categories of following data in json format. Data: "query": "What causes vomiting and diarrhea in a child after eating kimbap?", "answer": "Likely bacterial infection such as Salmonella or E. coli.", "category": null

—

"subject_type": ["science/biology", "science/microbiology"], "task_type": "knowledge", "target_type": "ko"

E.2 Reliability of Automated Categorization

E.2.1 Influence of Categorization Accuracy on Model Evaluation

We examine and discuss the influence of categorization accuracy on model evaluation outcomes in BENCHHUB. To quantify and simulate the categorizing errors, we conduct an ablation study in which the categorization error rate is systematically varied and controlled. Following the experimental setups described in § 4.2, we employ a stratified sampling strategy to preserve dataset-level balance across categories. We introduce a controlled *corruption rate*, which denotes the proportion of misclassified samples in the test set. We increment the corruption rate from 0.0% to 10.0% in 0.5% steps. For each corruption level, we perform 50 independent simulation runs to ensure statistical robustness. We compare the model rankings obtained from the corrupted test sets to the baseline rankings derived from the original, uncorrupted set.

We demonstrate that categorization errors up to 1.5% yield negligible disruption to model rankings, confirmed by Spearman's rank correlation coefficient and Wilcoxon Signed-Rank test. This finding suggests a notable resilience of the evaluation framework to minor categorization inaccuracies. It is noteworthy that this robustness extends beyond simple misclassification scenarios to dynamic, real-world settings tailored for users. Introducing a small fraction of samples comprising undefined categories is less likely to cause significant shifts in model rankings. Moreover, the categorizer can be incrementally updated and improved through continual learning, ensuring ongoing adaptation and maintenance of BenchHub pipeline among evolving benchmarks.

E.2.2 Enhancing Categorization Robustness

The classification process of BenchHub currently relies on a model trained with Qwen-2.5-7B, which may introduce potential model-specific bias when relying on a single classifier. As a possible direction for improving categorization, we additionally train classifiers using Llama-3.1-8B and Mistral-7B-v0.1 with the same training data and procedure. We then construct a multi-agent classification system in which the predictions from all three models (Qwen, Llama, and Mistral) are aggregated via majority voting. This system achieves a 2.4%p increase in agreement with human labels compared to Qwen-2.5-7B alone. Among the individual classifiers, Qwen-2.5-7B achieves the best standalone performance, and we expect that leveraging larger foundation models will further amplify the benefits of majority voting.

While majority voting improves robustness, it also triples the computational cost for training and inference. As an alternative, we implement a confidence-based hybrid approach: majority voting is invoked only when the classifier's confidence (measured by average logit probability) falls below a threshold of -0.04. This method enhances agreement by 1.4%p while substantially reducing the additional cost, thereby offering a practical trade-off between robustness and efficiency.

E.3 Experimental Setups

We use Axolotl [5] for the SFT training in § 3.3. We train `Qwen2.5-7B-Instruct` with DeepSpeed-Zero3 [61] on 4 A6000 48GB GPUs for 5 hours per run. We follow the method of **(author?)** [25] for optimization.

E.4 License

We release BenchHub, including our source code and trained models, under the Apache License 2.0. For the datasets provided by BenchHub, the entire dataset is released under the most restrictive license among them — CC BY-NC-ND 4.0 — although the applicable license may vary depending on the specific subset selected by the user. The license for each dataset is listed in Table 2.

E.5 Instructions and System Prompts

> Please read the following passage and answer the question. Choose one answer from `{label set}`. ↵ Passage: `{passage}` ↵ Question: `{question}` ↵ Choices: `{choices}` ↵ Answer:

> 다음 지문을 참고하여 질문에 답하여라. 답은 보기 중 하나를 `{label set}` 중에서 고르시오. ↵ 지문: `{passage}` ↵ 질문: `{question}` ↵ 보기: `{choices}` ↵ 답:

> Answer the following question. Choose one answer from `{label set}`. ↵ Question: `{question}` ↵ Choices: `{choices}` ↵ Answer:

> 다음 질문에 답하여라. 답은 보기 중 하나를 `{label set}` 중에서 고르시오. ↵ 질문: `{question}` ↵ 보기: `{choices}` ↵ 답:

F Multilingual Expansion of BENCHHUB

F.1 Multilingual Categorizer

We fine-tune Qwen-2.5-7B on ten languages (English; three high-resource languages: Arabic, German, Dutch; three mid-resource languages: Indonesian, Korean, Ukrainian; and three low-resource languages: Swahili, Nepali, Kyrgyz). For the training dataset, we use 20,000 samples from Global MMLU [73], with 2,000 samples per language. Since Global MMLU provides human-validated fine-grained subject categories, we adopt these categories while mapping them to our taxonomy. The training method and configurations follow those used in the categorizer for Korean and English (Appendix E.1).

Table 9: Categorizer Accuracy in G-MMLU (in-domain) and M-MMLU (out-domain)

language	G-MMLU	M-MMLU
ar	0.765	0.767
de	0.789	0.833
id	0.800	0.808
ky	0.681	–
ne	0.709	–
nl	0.804	–
sw	0.614	0.653
uk	0.765	–

We validate the categorizer on 2,850 Global MMLU samples (285 samples per language) that were not used during fine-tuning (in-domain), and on 1,225 Multilingual MMLU samples (245 samples per language) from outside the training distribution (out-of-domain). Our model achieves 75.3% accuracy in-domain and 77.5% accuracy out-of-domain for fine-grained subject categorization. Table 9 reports detailed results for both evaluation settings. Blank cells indicate that M-MMLU does not support the corresponding language.

F.2 Multilingual Dataset

Table 10 indicates the benchmarks included in BENCHHUB-multilingaul. We include 14 datasets across 8 additional languages, with the number of datasets per language varying depending on resource availability.

G Customized BENCHHUB

We provide three additional examples of real-world use cases of BENCHHUB:

(c) **Legal chatbot servicing in Korea and the US:** To select a foundation model for a legal chatbot, we select English and Korean datasets whose fine-grained subject is law. The final accuracy is computed as an average of the English and Korean datasets, ensuring that the model holds legal knowledge in both countries.

(d) **Docent agent for Korean traditional arts:** To identify the best-performing model with expertise in Korean traditional arts, we select Korean datasets within BENCHHUB whose fine-grained subjects are labeled as architecture, sculpture, and painting. To ensure balanced representation across individual subjects, the questions are drawn using a stratified sampling strategy at a subject level.

(e) **Counseling agent servicing in Korea:** To evaluate counseling agent in Korean, we select Korean datasets comprising:

1. psychology-related samples (*i.e.*, fine-grained category is psychology),
2. samples aware to Korean social interactions (*i.e.*, coarse-grained category is social intelligence),
3. samples relevant to common counseling topics (*i.e.*, fine-grained categories are work life, daily life, and family).

The final accuracy is computed as a weighted average of these subsets, with weights of 0.5, 0.3, and 0.2, respectively.

Table 11 presents the top-5 model rankings across these scenarios. The fluctuations in model rankings among the three scenarios also underscore the practical need for tailored evaluations using BENCHHUB.

Table 10: Benchmarks Included in BENCHHUB-multilingual

Dataset	Reference	Target	# of Samples	License
Language: AR				
G-MMLU	[73]	General/Local	14,042	apache-2.0
ArabLegalEval	[24]	Local	15,311	-
ArabicMMLU	[36]	General/Local	14,455	cc-by-nc-sa-4.0
Language: DE				
G-MMLU	[73]	General/Local	14,042	apache-2.0
GermanQUAD	[57]	General	2,204	cc-by-4.0
MLQA	[57]	General	4,517	cc-by-sa3.0
Language: NL				
G-MMLU	[73]	General/Local	14,042	apache-2.0
Language: ID				
G-MMLU	[73]	General/Local	14,042	apache-2.0
Eli5-indo	nlp/eli5_id	General	245,274	-
facQA	[45]	General	1,564	cc-by-sa-4.0
idkmrc	[59]	Local	1,198	cc-by-sa4.0.
QASiNa	[66]	Local	133	MIT.
TyDi QA	[45]	General	4,276	Apache-2.0
xcopa	[58]	Local	4,001	cc-by-4.0
Language: UK				
G-MMLU	[73]	General/Local	14,042	apache-2.0
UA-CBT (Eval-UA-tion 1.0)	[19]	Local	2,129	cc-by-4.0
Language: Sw				
G-MMLU	[73]	General/Local	14,042	apache-2.0
Language: Ne				
G-MMLU	[73]	General/Local	14,042	apache-2.0
Winogrande-Nepali	[53]	General	8,135	MIT
Language: Ky				
G-MMLU	[73]	General/Local	14,042	apache-2.0
TUMLU	[26]	Local	785	-

Table 11: Top 5 LLMs evaluated by customized BENCHHUB across three scenarios

Rank	(c) Legal chatbot	(d) Docent for Korean art	(e) Counseling agent
1	Qwen3-32B	Qwen2.5-72B-Instruct	Qwen2.5-72B-Instruct
2	gemma-3-1b-it	gemma-3-27b-it	Qwen3-8B
3	Qwen3-8B	Llama-3.3-70B-Instruct	gemma-3-27b-it
4	Qwen3-1.7B	Qwen3-32B	DeepSeek-R1-Distill-Qwen-32B
5	Mistral-Small-24B-Instruct-2501	Mistral-Small-24B-Instruct-2501	Mistral-Small-24B-Instruct-2501

H Experimental Results

See Table 12-13 for the scores (accuracies) of the models across subject types.

Table 12: Results of all models across fine-grained categories (English)

Subject	gpt-4.1	claude-3.7-sonnet	gemini-2.0	gemma-3-27b	DeepSeek-R1-32B	Llama-3.3-70B	Mistral-24B
Tech							
Urban eng.	0.882	0.765	0.824	0.625	0.765	0.588	0.882
Nuclear eng.	1.000	0.750	0.500	0.500	0.500	1.000	1.000
Marin eng.	1.000	0.667	1.000	0.500	1.000	1.000	1.000
Biomedical eng.	0.963	0.828	0.716	0.563	0.743	0.779	0.794
Mechanics	0.943	0.829	0.829	0.559	0.706	0.647	0.941
Materials eng.	0.987	0.920	0.760	0.595	0.811	0.784	0.932
IT	0.904	0.735	0.783	0.598	0.690	0.724	0.782
Environmental eng.	0.957	0.739	0.855	0.652	0.797	0.754	0.928
Energy eng.	0.953	0.802	0.791	0.628	0.826	0.767	0.872
Electrical eng.	0.877	0.816	0.825	0.609	0.722	0.704	0.800
Programming	1.000	0.913	0.826	0.667	0.611	0.556	0.722
Civil eng.	1.000	0.769	0.923	0.750	0.750	0.750	1.000
Chemical eng.	0.714	0.571	0.571	0.429	0.714	0.714	0.571
AI	0.931	0.984	0.817	0.474	0.420	0.355	0.330
Agricultural eng.	1.000	0.867	0.800	0.705	0.864	0.795	0.932
Aerospace eng.	1.000	0.833	1.000	1.000	0.833	0.833	1.000
Science							
Statistics	0.879	0.803	0.803	0.452	0.563	0.600	0.622
Physics	0.892	0.800	0.842	0.549	0.689	0.705	0.713
Mathematics	0.918	0.956	0.872	0.756	0.717	0.587	0.711
Life science	0.965	0.798	0.781	0.565	0.809	0.678	0.904
Geology	0.990	0.816	0.776	0.688	0.792	0.656	0.885
Earth science	0.979	0.798	0.840	0.692	0.788	0.779	0.942
Chemistry	0.863	0.814	0.762	0.510	0.650	0.697	0.720
Biology	0.959	0.730	0.818	0.533	0.767	0.769	0.835
Atmospheric science	0.990	0.753	0.753	0.739	0.783	0.641	0.935
Astronomy	0.965	0.843	0.843	0.704	0.835	0.809	0.852
HASS							
Welfare	0.896	0.722	0.729	0.576	0.654	0.737	0.797
Trade	0.944	0.807	0.800	0.494	0.811	0.767	0.856
Cognitive studies	0.620	0.524	0.481	0.500	0.580	0.662	0.629
Religion	0.912	0.877	0.895	0.724	0.914	0.860	0.948
Politics	0.909	0.759	0.693	0.635	0.767	0.767	0.872
Philosophy	0.875	0.664	0.632	0.455	0.711	0.623	0.651
Media	0.857	0.864	0.759	0.667	0.889	0.778	0.722
Literature	0.950	0.850	0.850	0.684	0.950	0.750	0.950
Law	0.750	0.596	0.610	0.294	0.540	0.518	0.679
Language	0.736	0.548	0.518	0.420	0.526	0.519	0.504
History	0.911	0.864	0.578	0.463	0.786	0.857	0.881
Geography	0.911	0.804	0.804	0.628	0.773	0.886	0.818
Education	0.957	0.793	0.793	0.580	0.795	0.652	0.848
Economics	0.893	0.809	0.695	0.574	0.597	0.713	0.752
Administration	0.899	0.797	0.732	0.551	0.819	0.819	0.841
Social Intelligence							
Value	0.699	0.890	0.788	0.653	0.599	0.857	0.619
Norms	0.816	0.658	0.605	0.516	0.613	0.581	0.710
Commonsense	0.837	0.765	0.749	0.871	0.877	0.856	0.837
Bias	0.000	1.000	0.333	0.349	0.333	0.324	0.288
Culture							
Work life	0.778	0.667	0.704	0.600	0.720	0.700	0.720
Tradition	0.833	0.881	0.950	0.618	0.806	0.800	0.784
Housing	1.000	1.000	0.750	1.000	1.000	0.750	0.750
Food	0.534	0.479	0.479	0.360	0.553	0.675	0.456
Family	0.913	0.739	0.609	0.591	0.659	0.705	0.818
Daily life	0.600	0.521	0.475	0.355	0.590	0.676	0.532
Clothing	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Holiday	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Arts % Sports							
Sports	0.781	0.578	0.453	0.714	0.929	0.786	0.857
Sculpture	1.000	1.000	1.000	0.500	1.000	0.500	1.000
Photography	1.000	0.600	0.800	0.400	0.400	0.800	0.800
Performing	0.846	0.846	0.769	0.673	0.654	0.808	0.846
Painting	1.000	0.600	0.900	0.600	0.900	0.700	1.000
Music	1.000	1.000	0.800	0.900	0.900	0.900	0.800
Festivals	0.500	1.000	1.000	1.000	0.500	1.000	0.500
Fashion	1.000	0.800	1.000	0.800	0.800	0.600	0.600
Architecture	1.000	0.857	0.714	0.429	1.000	0.571	1.000

Table 13: Results of all models across fine-grained categories (Korean)

Subject	gpt-4.1	claude-3.7-sonnet	gemini-2.0	gemma-3-27b	DeepSeek-R1-32B	Llama-3.3-70B	Mistral-24B
Tech							
Urban eng.	0.552	0.634	0.559	0.504	0.507	0.543	0.468
Nuclear eng.	0.676	0.647	0.618	0.676	0.559	0.588	0.588
Marine eng.	0.688	0.826	0.625	0.569	0.521	0.611	0.569
Biomedical eng.	0.838	0.805	0.409	0.727	0.507	0.767	0.713
Mechanics	0.661	0.709	0.563	0.537	0.495	0.487	0.420
Materials eng.	0.720	0.820	0.560	0.608	0.510	0.619	0.608
IT	0.854	0.877	0.667	0.727	0.756	0.803	0.742
Environmental eng.	0.591	0.649	0.480	0.456	0.427	0.462	0.368
Energy eng.	0.587	0.674	0.551	0.507	0.457	0.457	0.399
Electrical eng.	0.688	0.778	0.646	0.549	0.535	0.549	0.500
Programming	0.667	0.722	0.667	0.667	0.667	0.667	0.833
Civil eng.	0.517	0.669	0.530	0.503	0.391	0.497	0.430
Chemical eng.	0.711	0.809	0.641	0.596	0.539	0.574	0.560
AI	0.861	0.829	0.676	0.694	0.618	0.657	0.703
Agricultural eng.	0.605	0.605	0.539	0.464	0.386	0.506	0.428
Aerospace eng.	0.757	0.786	0.579	0.621	0.564	0.629	0.579
Science							
Statistics	0.813	0.813	0.571	0.571	0.582	0.549	0.615
Physics	0.826	0.870	0.644	0.626	0.595	0.603	0.542
Mathematics	0.842	0.889	0.848	0.385	0.487	0.359	0.359
Life science	0.783	0.783	0.635	0.635	0.609	0.739	0.635
Geology	0.755	0.765	0.627	0.608	0.422	0.618	0.510
Earth science	0.701	0.769	0.627	0.604	0.552	0.575	0.575
Chemistry	0.760	0.829	0.643	0.574	0.612	0.643	0.512
Biology	0.852	0.875	0.586	0.766	0.664	0.742	0.711
Atmospheric science	0.719	0.688	0.625	0.531	0.531	0.656	0.563
Astronomy	1.000	1.000	1.000	0.900	1.000	1.000	0.800
HASS							
Welfare	0.783	0.745	0.516	0.755	0.742	0.724	0.705
Trade	0.856	0.767	0.658	0.752	0.752	0.766	0.731
Religion	0.846	0.860	0.714	0.805	0.706	0.812	0.856
Psychology	1.000	1.000	1.000	1.000	0.000	1.000	0.000
Politics	0.806	0.858	0.714	0.717	0.634	0.667	0.703
Philosophy	0.843	0.897	0.715	0.791	0.718	0.757	0.757
Media	0.942	0.928	0.897	0.877	0.755	0.876	0.877
Literature	0.836	0.914	0.760	0.700	0.739	0.798	0.800
Law	0.604	0.555	0.463	0.510	0.416	0.544	0.530
Language	0.807	0.906	0.763	0.648	0.685	0.750	0.705
History	0.775	0.794	0.691	0.622	0.526	0.603	0.570
Geography	0.711	0.778	0.698	0.594	0.522	0.631	0.597
Education	0.732	0.816	0.586	0.701	0.603	0.755	0.660
Economics	0.814	0.820	0.606	0.704	0.701	0.692	0.656
Administration	0.731	0.766	0.598	0.691	0.635	0.711	0.675
Social Intelligence							
Value	0.848	0.879	0.697	0.818	0.818	0.788	0.758
Norms	0.884	0.881	0.881	0.881	0.810	0.721	0.762
Commonsense	0.835	0.873	0.822	0.718	0.757	0.748	0.767
Bias	0.993	0.966	0.951	1.000	1.000	0.846	1.000
Culture							
Work life	0.926	0.926	0.826	0.921	0.768	0.921	0.921
Tradition	0.962	0.960	0.858	0.917	0.819	0.900	0.911
Leisure	1.000	1.000	1.000	0.500	0.500	1.000	0.500
Housing	0.824	0.824	0.647	0.735	0.676	0.676	0.676
Food	0.850	0.923	0.769	0.744	0.684	0.789	0.821
Family	0.826	0.792	0.696	0.652	0.818	0.864	0.800
Daily life	0.837	0.837	0.823	0.751	0.682	0.738	0.764
Clothing	0.793	0.793	0.690	0.621	0.655	0.759	0.655
Holiday	0.643	0.602	0.602	0.620	0.616	0.674	0.654
Arts & Sports							
Sports	0.960	0.960	0.818	0.960	0.917	0.913	0.864
Sculpture	0.923	0.833	0.833	1.000	0.727	0.917	0.833
Photography	0.800	0.855	0.655	0.768	0.600	0.667	0.655
Performing	0.950	0.950	0.911	0.930	0.752	0.884	0.918
Painting	0.931	0.932	0.833	0.896	0.794	0.837	0.918
Music	0.912	0.971	0.758	0.909	0.667	0.879	0.909
Festivals	0.941	1.000	1.000	0.941	0.882	0.813	0.941
Fashion	0.626	0.626	0.524	0.565	0.490	0.571	0.456
Architecture	0.745	0.778	0.641	0.711	0.658	0.664	0.618

Tables 14 and 15 details the accuracy of 14 open LLMs across four different sampling strategies in English and Korean, respectively.

Table 16 details the accuracy of 14 open LLMs evaluated by the customized BENCHHUB in five different scenarios.

Table 14: Evaluation results of 14 open LLMs in English across four different sampling strategies

Model	Random	Stratified	Chatbot Arena	MixEval
Qwen2.5-72B-Instruct	0.688	0.694	0.680	0.661
Qwen3-1.7B	0.810	0.833	0.811	0.798
Qwen3-14B	0.729	0.763	0.737	0.723
Qwen3-32B	0.817	0.852	0.816	0.789
Qwen3-4B	0.784	0.845	0.788	0.779
Qwen3-8B	0.734	0.779	0.733	0.729
DeepSeek-R1-Distill-Qwen-14B	0.743	0.778	0.747	0.730
DeepSeek-R1-Distill-Qwen-32B	0.717	0.748	0.721	0.704
gemma-3-1b-it	0.874	0.962	0.888	0.870
gemma-3-27b-it	0.702	0.707	0.690	0.677
gemma-3-4b-it	0.746	0.799	0.755	0.743
Llama-3.1-8B-instruct	0.732	0.749	0.726	0.707
Llama-3.3-70B-Instruct	0.704	0.733	0.712	0.689
Mistral-Small-24B-Instruct-2501	0.696	0.713	0.696	0.686

Table 15: Evaluation results of 14 open LLMs in Korean across four different sampling strategies

Model	Random	Stratified	Chatbot Arena	MixEval
Qwen2.5-72B-Instruct	0.697	0.708	0.723	0.692
Qwen3-1.7B	0.453	0.492	0.478	0.486
Qwen3-14B	0.360	0.376	0.363	0.371
Qwen3-32B	0.605	0.613	0.618	0.609
Qwen3-4B	0.370	0.444	0.383	0.406
Qwen3-8B	0.597	0.623	0.617	0.625
DeepSeek-R1-Distill-Qwen-14B	0.613	0.613	0.615	0.606
DeepSeek-R1-Distill-Qwen-32B	0.612	0.635	0.638	0.623
gemma-3-1b-it	0.466	0.474	0.469	0.468
gemma-3-27b-it	0.661	0.666	0.665	0.649
gemma-3-4b-it	0.507	0.533	0.510	0.519
Llama-3.1-8B-instruct	0.531	0.562	0.547	0.541
Llama-3.3-70B-Instruct	0.671	0.674	0.683	0.666
Mistral-Small-24B-Instruct-2501	0.624	0.647	0.646	0.630

Table 16: Evaluation results of 14 open LLMs using customized BENCHHUB across five use cases

Model	(a)	(b)	(c)	(d)	(e)
Qwen2.5-72B-Instruct	0.604	0.657	0.658	0.595	0.670
Qwen3-1.7B	0.711	0.477	0.703	0.383	0.624
Qwen3-4B	0.667	0.420	0.556	0.300	0.599
Qwen3-8B	0.629	0.568	0.718	0.430	0.665
Qwen3-14B	0.642	0.429	0.531	0.316	0.499
Qwen3-32B	0.798	0.523	0.663	0.529	0.648
DeepSeek-R1-Distill-Qwen-14B	0.657	0.554	0.653	0.479	0.647
DeepSeek-R1-Distill-Qwen-32B	0.626	0.609	0.654	0.488	0.660
Llama-3.1-8B-Instruct	0.650	0.581	0.602	0.393	0.627
Llama-3.3-70B-Instruct	0.651	0.612	0.637	0.562	0.659
Mistral-Small-24B-Instruct-2501	0.619	0.632	0.661	0.523	0.660
gemma-3-1b-it	0.762	0.465	0.704	0.364	0.551
gemma-3-4b-it	0.632	0.529	0.641	0.391	0.632
gemma-3-27b-it	0.611	0.614	0.651	0.582	0.664