Benchmarks Are Not Atomic: Composition-Aware LLM Evaluation using BENCHHUB

Eunsu Kim [*1] **Haneul Yoo** [*1] **Guijin Son** [2 3] **Hitesh Patel** [4] **Amit Agarwal** [4] **Alice Oh** [1]

Abstract

LLM benchmarks are often treated as coherent measurement units, yet they are heterogeneous collections of instances spanning diverse domains, skills, formats, and contexts. As a result, aggregate benchmark scores can conflate model capability with benchmark composition, obscuring what existing benchmarks actually cover. We introduce BENCHHUB, a composition-aware evaluation framework that represents benchmarks as distributions over instance-level attributes. BENCHHUB integrates 54 benchmarks comprising 839K samples across 10 languages. It enables researchers to inspect benchmark contents, uncover reusable coverage hidden behind benchmark names, transparently compare benchmarks, and construct controllable evaluation sets for application-aligned model selection. Using BENCHHUB, we show that similarly motivated benchmarks can differ substantially in internal composition, existing benchmarks often contain reusable coverage beyond their stated purposes, models exhibit fine-grained category-level performance variation hidden by aggregate scores, and model rankings can shift under different reweighting and resampling configurations. Our results motivate evaluation practices that make benchmark composition explicit, inspectable, and controllable.

1. Introduction

The rapid development and deployment of large language models (LLMs) have led to the emergence of benchmarks at an unprecedented scale (Hendrycks et al., 2021a; Srivastava et al., 2023; Liang et al., 2023). Benchmark datasets serve as a primary instrument for assessing what LLMs can and cannot do and signaling what capabilities the community expects LLMs to demonstrate (Wang et al., 2024c; Sokol et al., 2025a). Accordingly, leaderboards and technical reports increasingly rely on aggregate scores across benchmark suites as compact indicators of model capability.

However, benchmark-level aggregate scores provide an incomplete, and sometimes misleading, view of evaluation (Ribeiro et al., 2020; Reuel et al., 2024). A benchmark is rarely a monolithic capability test; it is a heterogeneous mixture of instances that may differ substantially in what they require, what they cover, and how they are presented (Kim et al., 2025a;c). Collapsing this mixture into a single score obscures what is actually being assessed and which regions of the evaluation space remain underexplored. It also limits robust model analysis: two models with identical aggregate scores may exhibit entirely disjoint failure modes, and comparative conclusions can flip when an evaluation set is subtly reweighted, resampled, or swapped for a functionally similar benchmark (Alzahrani et al., 2024; Singh et al., 2025; Frick et al., 2025a). Thus, benchmarks should not be treated as atomic units; their internal composition is inherently part of the evaluation. What the community needs now is not merely more datasets, but better methodologies to understand and use the ones we already have.

To make current benchmarks inspectable, reusable, and controllable, we introduce BENCHHUB a dynamic infrastructure for transforming existing benchmarks into composition-aware evaluation resources. BENCHHUB integrates 54 benchmark datasets comprising 839K instances across 10 languages. Each instance is annotated with fine-grained attributes, including subject domain, required skill, input–output format, cultural specificity, and tool-use requirements, using an automated expansion pipeline backed by carefully designed taxonomies. Unlike prior work on instance-level benchmark annotation (Ye et al., 2024; Zhou et al., 2025), which typically operates over fixed evaluation suites, BENCHHUB provides a unified instance-level representation that continuously integrates and reorganizes existing benchmark collections into searchable, analyzable, and reconfigurable evaluation data. This representation enables researchers to inspect evaluation composition, analyze attribute coverage, interpret model performance at a finer

[1]KAIST [2]Seoul National University [3]OnelineAI [4]Oracle. Correspondence to: Eunsu Kim <kes0317@kaist.ac.kr>, Haneul Yoo <haneul.yoo@kaist.ac.kr>.

Accepted to the 1st Workshop on Combining Theory and Benchmarks, CTB@ICML 2026, Seoul, South Korea. Copyright 2026 by the author(s).

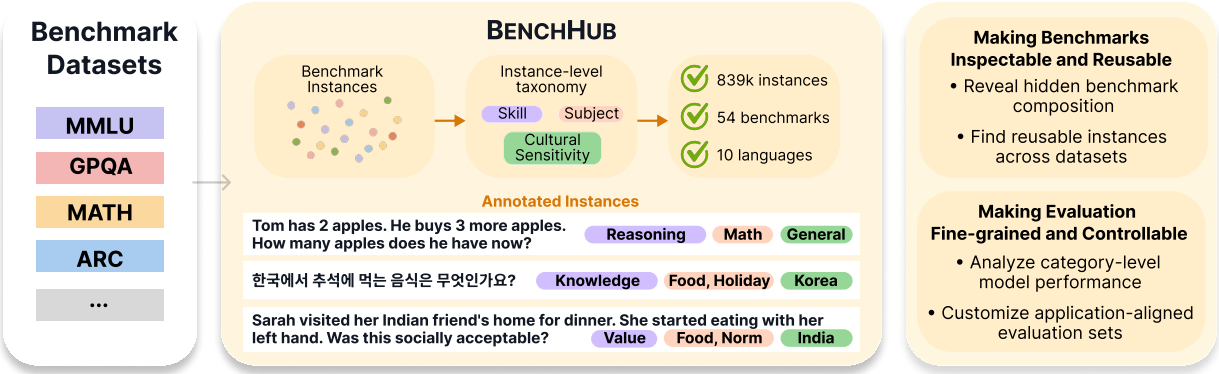


Figure 1. Conceptual overview of BENCHHUB. BENCHHUB automatically merges existing benchmark datasets and annotates questions at the instance level with a unified taxonomy, making the contents of existing benchmarks explicit and reusable beyond their dataset names. This enables users to construct objective-aligned test sets and conduct fine-grained, controllable model evaluation.

level, and construct or select custom evaluation sets aligned with specific operational goals.

Using BENCHHUB, we demonstrate the critical value of composition-aware evaluation in two areas.

- **Making benchmark datasets inspectable and reusable (§3):** We systematically reveal that existing benchmarks are internally heterogeneous and rarely compositionally interchangeable, even when designed for similar purposes. BENCHHUB exposes uneven coverage across attributes, and identifies reusable evaluation instances that remain hidden when practitioners select benchmarks by name alone.
- **Making model evaluation fine-grained and controllable (§4):** We empirically demonstrate that models exhibit structured performance disparities across composition dimensions that are entirely obscured by aggregate scores. BENCHHUB enables fine-grained, category-level analysis and shows how model rankings and deployment decisions can shift under plausible reweighting and resampling configurations.

Rather than diminishing the value of existing aggregate scores, which remain vital for high-level progress tracking, our findings complement them by offering a higher-resolution lens. We argue that evaluation practice should more explicitly account for composition: which instances are included, how they are distributed, and how sensitive conclusions are to these distributions. BENCHHUB bridges this gap, enabling rigorous benchmark auditing, revealing underrepresented evaluation space, and paving the way for targeted, robust model assessment.

2. BENCHHUB: Unified & Composition-aware LLM Evaluation Framework

BENCHHUB is a composition-aware evaluation framework that organizes existing LLM benchmarks into a shared instance-level representation. It integrates 54 benchmarks comprising 839k samples across 10 languages, with a primary focus on English and Korean through BENCHHUB-En and BENCHHUB-Ko, respectively. Rather than treating benchmarks as indivisible evaluation units, BENCHHUB represents them as distributions over compositional attributes, making benchmark composition explicit, inspectable, and controllable.

We design BENCHHUB around two core principles. First, it provides a fine-grained, multi-dimensional taxonomy for describing the composition of evaluation instances. Second, it uses an automated expansion pipeline to reformat, annotate, and integrate newly released datasets into the unified benchmark space. In this section, we describe the taxonomy design (§2.1), the contained dataset (§2.2), and the automated pipeline (§2.3).

To improve usability, BENCHHUB also provides a web-based platform with interactive tools and code utilities for data selection, customized evaluation-set construction, multilingual extensions, and automated model evaluation. Implementation details are provided in Appendix C.

2.1. Taxonomy

We annotate benchmarks using six complementary dimensions designed to capture the major factors that shape benchmark composition and evaluation behavior. These dimensions characterize *what* capability is evaluated, *how* responses are produced and assessed, and *which* knowledge or cultural context is required. Specifically, we define three dataset-level attributes—**task, answer format, and tool usage**—and three sample-level attributes—**skill, subject, and cultural-specificity**. Together, these dimensions define the composition space used throughout BENCHHUB to inspect benchmark contents, compare evaluation mixtures,

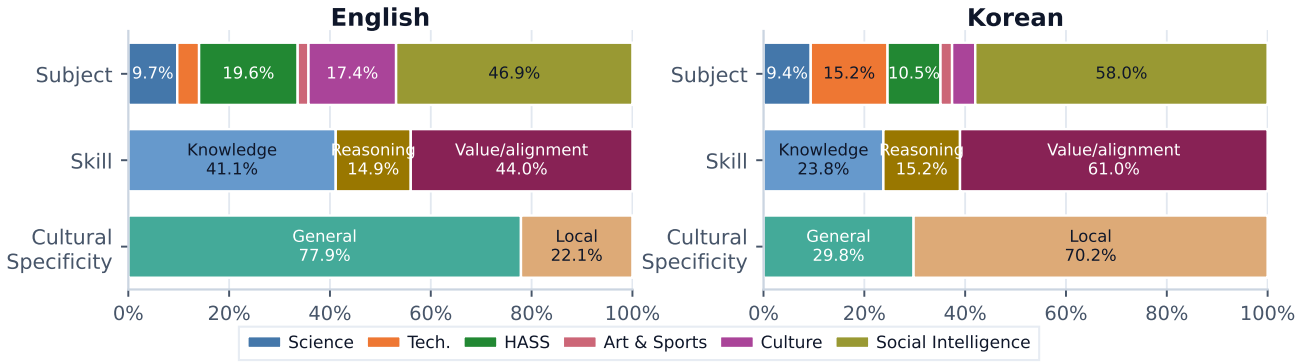


Figure 2. Data distribution of all datasets used in BENCHHUB by coarse-grained subjects, targets, and tasks. The English and Korean data include 183k and 137k questions, respectively.

and construct targeted evaluation sets. The full taxonomy is illustrated in Appendix F.

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, 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 code interpreters, web browsers, and calculators (*requires external tools*). This dimension supports agentic evaluation, where models must decide when and how to invoke external resources.

Instance-level attributes:

4. **Subject** denotes the knowledge domain. We define six coarse-grained categories—*Science*, *Technology*, *Humanities and Social Science (HASS)*, *Art & Sports*, *Culture*, and *Social Intelligence*—along with 64 subcategories, by integrating various knowledge classification systems (Appendix F). Each sample may have multiple subject labels.
5. **Skill** captures the required ability to answer the question (*i.e.*, knowledge, reasoning, and value/alignment). These categories reflect widely used capability abstractions in LLM evaluation settings.
6. **Cultural-specificity** represents the cultural or geographical focus. Culturally agnostic items are labeled as *General*; otherwise, we assign a *Local* tag. This dimension is important because model responses and their appropriateness may vary substantially across cultural contexts (Singh et al., 2024).

2.2. Datasets

We apply this taxonomy to 54 benchmarks in 10 languages, mainly covering English and Korean and totaling over 839k instances. Figures 2 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. [1] We curate (1) general-purpose (*i.e.*, culturally agnostic) datasets commonly used by holistic evaluation benchmarks (Ye et al., 2024; Ni et al., 2024) and (2) culture-specific datasets. We select English datasets spanning multiple cultures drawn from a recent survey (Pawar et al., 2024), 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. Tables 5–6 in Appendix provide a complete list of the datasets.

2.3. Automated and Dynamic Expansion

Since benchmark datasets are emerging at a rapid pace, BENCHHUB is designed to support continuous expansion. To integrate newly introduced datasets, we automate the dataset integration process using an LLM-assisted pipeline that reformats datasets into the BENCHHUB schema and assigns taxonomy labels to each instance. The pipeline consists of four steps:

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

[1] We count the multilingual datasets—BLEnD (Myung et al., 2024) and CaLMQA (Arora et al., 2024)—in both.

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 cultural-specificity) using a fine-tuned Qwen-2.5-7B-instruct model (`BenchHub-Cat-7B`).[2] Table 3 in Appendix reports the accuracy of the categorizer model for sample-level categorization.
4. **Merging:** The processed and annotated dataset is seamlessly merged into the main collections, thereby producing the next BENCHHUB release.

This automated pipeline enables BENCHHUB to continuously expand as new datasets emerge, supporting more comprehensive evaluations and analyses over time. Although LLM-based expansion is inherently imperfect, we empirically validate the categorization process: on human/metadata-aligned validation sets, `BenchHub-Cat-7B` achieves 0.871 accuracy for subject, 0.967 for skill, and 0.986 for cultural-specificity. We further examine the reliability, scalability, and robustness of the expansion process in Appendix D.

Responsible reuse. Because BENCHHUB aggregates existing benchmark datasets, it may inherit limitations or constraints from the original source datasets, such as contamination and dataset-specific reuse constraints, such as licensing and citation requirements. To support responsible reuse, BENCHHUB preserves source-dataset attribution for each instance and provides citation and license-reporting utilities that summarize source datasets, licenses, instance counts, citation guidelines, and BibTeX entries for customized evaluation sets.[3] We further examine the possible effect of contaminated instances through a controlled simulation in Appendix H.4.

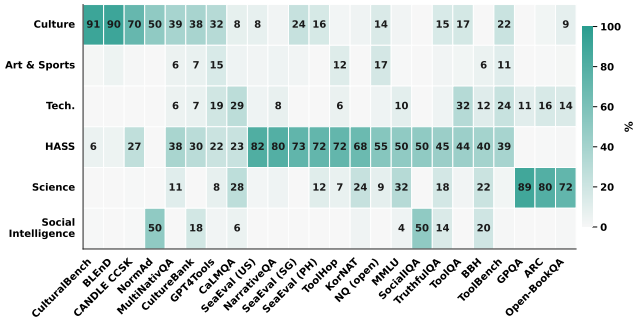
3. Auditing Benchmark Composition and Reusability

In this section, we use BENCHHUB to audit existing benchmarks as compositional resources rather than indivisible named datasets. Specifically, we examine the evaluation coverage hidden behind benchmark names and the extent to which it can be reused beyond their original purposes (Obs. 1). Second, we validate whether benchmarks designed for similar high-level goals actually instantiate comparable mixtures of domains, skills, and cultural contexts (Obs. 2).

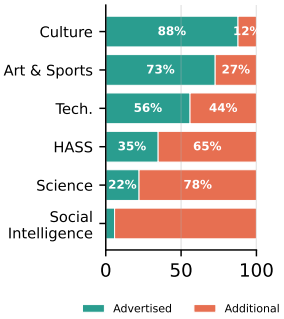
[2]Details on the training and validation of `BenchHub-Cat-7B` are provided in Appendix D.1.

[3]We will release the dataset and web demo upon acceptance.

Observation 1 *Benchmark Names Underdescribe Reusable Evaluation Coverage.*



(a) Subject distribution by benchmarks (p(subject|benchmark))



(b) Advertised vs. Additional

Figure 3. Subject coverage distribution and alignment in English benchmarks.

Figure 3a shows the subject distribution of each benchmark under the BENCHHUB taxonomy. While most benchmarks have a dominant advertised focus, many contain substantial secondary coverage that is not apparent from their names or high-level descriptions. For example, `NQ-open` includes culture-related questions about Indian currency and Chinese spice powder, while `ToolBench` includes `Art & Sports`-tagged instances requiring historical NBA roster knowledge. Such examples show that existing benchmarks often contain reusable evaluation material that would be missed by name-based dataset selection.

We quantify this mismatch by comparing each benchmark's advertised scope with its instance-level BENCHHUB label. The advertised scope is derived from the benchmark name, metadata, and, when available, author-provided categories. A sample is counted as *advertised* coverage if its BENCHHUB label matches this scope; otherwise, it is counted as *additional* coverage exposed by BENCHHUB. In Figure 3b, 65.1% of English samples fall outside their benchmarks' advertised subject categories. This indicates that existing benchmarks contain substantial reusable coverage beyond their name. For underrepresented domains, BENCHHUB provides a practical way to retrieve relevant instances that have already been embedded in existing benchmarks.

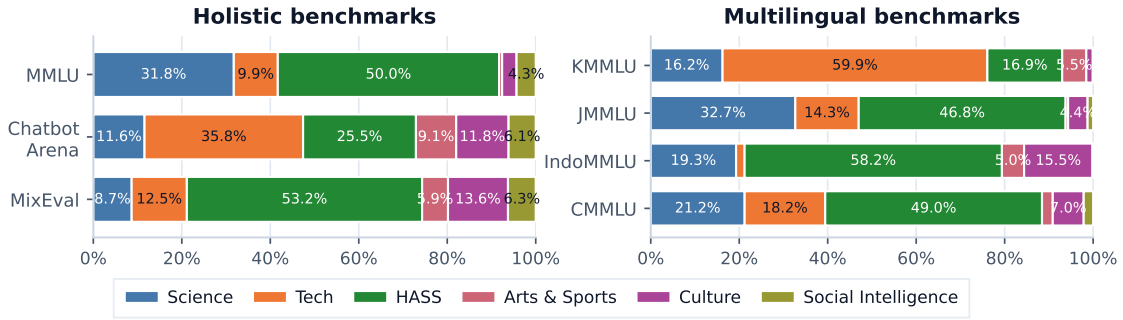


Figure 4. Subject distributions of existing evaluation benchmarks.

Takeaway. Benchmark names or their advertised scopes are incomplete interfaces to evaluation data. BENCHHUB exposes hidden instance-level, fine-grained coverage, making existing benchmarks more visible, auditable, and reusable for domain-specific evaluation.

Observation 2 *Similarly Motivated Benchmarks Encode Different Composition.*

In contrast to Obs. 1, benchmarks with similar high-level motivations may define different evaluation targets. Here, we compare two benchmarks: (1) holistic benchmarks in English for general-purpose LLMs and (2) MMLU-style benchmarks in multiple languages.

Holistic benchmarks operationalize general capability differently. Figure 4 compares Chatbot Arena (Chiang et al., 2024), MixEval (Ni et al., 2024), and MMLU (Hendrycks et al., 2021a) using our fine-tuned classifiers (§2.3). While they are all used to summarize general model capability, their subject distributions differ substantially. For example, Chatbot Arena contains only 25.5% of Humanities and Social Science (HASS) questions, whereas both MixEval and MMLU contain more than half HASS questions. Their skill distributions also diverge: MixEval includes fewer than 0.30% value/alignment tasks and is heavily skewed toward knowledge evaluation (Figure 12). These disparities indicate that holistic benchmarks do not instantiate distinct evaluation mixtures despite similar high-level positioning.

Multilingual benchmarks introduce composition shifts across languages. Composition differences are further amplified in multilingual settings. Figures 4 and 11 show that MMLU-style benchmarks vary substantially across languages. For example, English MMLU emphasizes HASS, whereas Korean MMLU (KMMLU) (Son et al., 2025b) comprises 76.1% STEM questions. Thus, cross-lingual performance differences on such benchmarks may reflect shifts in subject distributions and mislead decision-making. Composition-aware reporting is therefore essential to reliably interpret multilingual evaluation results.

Takeaway. Similarly motivated benchmarks are not necessarily compositionally interchangeable. Benchmark scores should be interpreted as performance on a specific mixture of domains, skills, formats, and languages, rather than as direct measurements of a single well-defined capability.

4. Composition-Aware Model Evaluation and Selection

In this section, we examine how the hidden and uneven mixtures of existing benchmarks shape model evaluation. Using BENCHHUB, we move from benchmark-level evaluation to composition-aware evaluation. First, we construct category-level evaluation sets across English and Korean benchmarks and show that model rankings vary substantially across different regions of the BENCHHUB taxonomy (Obs. 3). Second, we demonstrate that rankings are sensitive to the distribution of evaluation instances, revealing benchmark composition as a hidden variable in model comparison (Obs. 4). Finally, we show that making composition explicit enables use case-specific evaluation sets, leading to model-selection decisions that differ from those produced by general-purpose benchmark aggregation (Obs. 5).

4.1. Evaluation Setup

Dataset. We select 32,954 English and 21,541 instances for English and Korean, respectively. To manage the large number of fine-grained categories, we sample up to 150 examples per category, per problem format, fully including categories with 100–150 samples and merging categories with fewer than 80 samples into a miscellaneous group within the same coarse-grained category.

Models. We use 15 open-weight models: Qwen3-{4B,8B,14B,32B}, Qwen3-30B-A3B-Instruct-2507, Qwen3-Next-80B-A3B-Instruct); gemma-3-{4b,12b}-it; Meta-Llama-3-70B-Instruct; Ministral-8B-Instruct-2410, Mistral-Small-3.2-24B-Instruct-2506, Mixtral-8x7B-Instruct-v0.; and Olmo 3 (Olmo-3-1025-7B, Olmo-3-1125-32B, Olmo-3.1-32B-Instruct).

Evaluation details, including inference prompts and evaluation methods, are provided in Appendix G.

4.2. Evaluation Results of Open-weight Models using BENCHHUB

Observation 3 *Aggregate Scores Collapse Distinct Model Profiles.*

Rankings vary across composition dimensions. We first examine whether model rankings are stable across different categories of the BENCHHUB taxonomy. As illustrated in Figure 5, model rankings substantially fluctuate across subject, skill, and cultural-specificity categories. In English, the median model-level rank range, defined as the difference between a model's worst and best category-wise rank within a dimension, is 5 rank positions for Subject, 3 for Skill, and 2 for Cultural-specificity. Furthermore, pairwise model preference frequently flip: 29.5% of model pairs reverse their ordering between Science and Art & Sports. We observe a similar trend across Korean benchmarks(Figure 15).

Similar aggregate scores hide different strengths and weaknesses. To further demonstrate the masking effect of aggregation, we analyze model pairs whose overall accuracies differ by at most 5 percentage points. Figure 6 illustrates how BENCHHUB provides a higher-resolution view of model performance beyond aggregate scores. For two models and a taxonomy dimension, we define their *profile gap* as the mean absolute accuracy difference across categories in that dimension. While near-score pairs in English exhibit a median overall gap of only 1.24 pp, their category-level profile gaps remain substantially larger: a median of 4.70 pp for Subject, 5.02 pp for Skill, and 4.49 pp for Cultural-specificity. In extreme cases, the maximum profile gap reaches up to 8.46 pp for Subject and 7.86 pp for Skill. This suggests that models appearing functionally equivalent under aggregate scores can exhibit entirely disjoint strengths and weaknesses. By decomposing performance along fine-grained taxonomy dimensions, BENCHHUB helps uncover these hidden differences and improves the interpretability of model evaluation.

Takeaway. Model capability is not a single scalar. Aggregate scores inherently collapse these profiles, destroying the nuance required for rigorous model selection. When viewed through the BENCHHUB taxonomy, models reveal structured, interpretable performance profile.

Observation 4 *Benchmark Composition Shapes Model Rankings.*

We investigate whether evaluation conclusions, such as model rankings and model selection decisions, are robust to changes in benchmark composition. While benchmark scores are often treated as stable indicators of model capability, this implicitly assumes that evaluation outcomes are invariant to the distribution of evaluation instances.

To quantify how evaluation composition affects model comparison, we repeat 100 simulations to construct 1k-instance evaluation sets under four sampling strategies: (1) random sampling, (2) stratified sampling using the subject categories in BENCHHUB, and sampling following the distributions of (3) MixEval and (4) Chatbot Arena. These strategies induce substantially different subject mixtures.

These composition shifts lead to systematic changes in model scores and rankings. Across sampling strategies, the average model-level accuracy range is 3.39 pp in English and 4.95 pp in Korean, with maximum shifts of 4.95 and 7.25 pp, respectively. Rank distributions also shift broadly: after Benjamini–Hochberg correction, 53.3% model-by-method-pair comparisons in English and 54.4% in Korean show significant rank-distribution differences tested by Mann–Whitney U test ($p < 0.01$).

These results demonstrate that benchmark composition acts as a hidden intervention on model evaluation: even when global leaderboard structure is preserved, the evaluation distribution can shift scores, alter model-specific ranks, reverse pairwise preference, and change selected models.

Takeaway. Model rankings are not invariant properties of models alone. They also depend on how the evaluation set is constructed. Benchmark composition therefore acts as a hidden variable in model comparison: it shapes which capabilities are emphasized, which models appear strongest, and how leaderboard rankings should be interpreted.

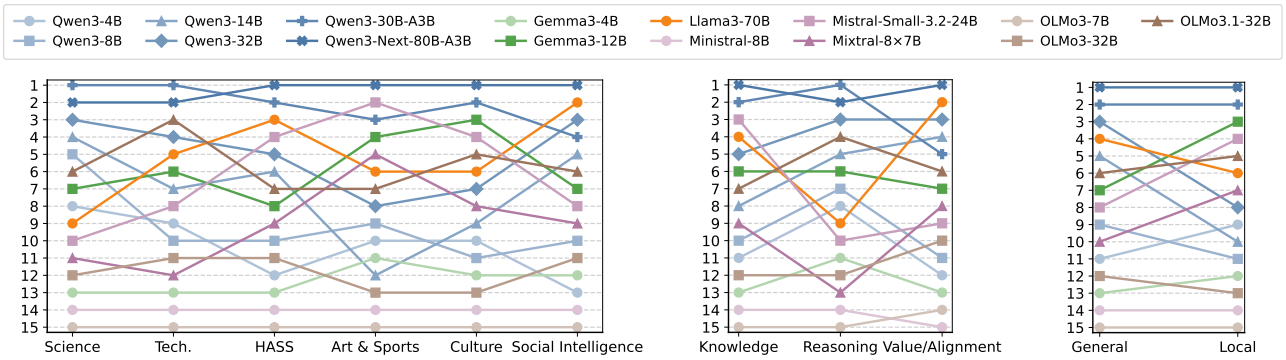
Observation 5 *Application-aligned Evaluation Can Change Model Selection.*

Since model rankings depend on benchmark composition, evaluation sets should be strongly aligned with the intended deployment context rather than inherited from a generic benchmark distribution. BENCHHUB supports this by allowing users to specify an explicit target composition over language, subject, skill, answer format, and cultural-specificity dimensions. The resulting score is computed over instances matching this declared profile, making the evaluation target transparent and reproducible.

We initiate this workflow with five application scenarios, including an English legal chatbot and a Korean docent agent. Full profile definitions with additional scenarios are provided in Appendix H.2.

(a) Legal chatbot with advanced reasoning: we construct an English test set combining legal domain knowledge, along with reasoning and free-form generation capabilities.
(b) Docent agent for Korean traditional arts: we construct a Korean test set focusing on expert knowledge in Korean traditional arts.

Compared with a generic stratified baseline, customized



(a) Subject (coarse-grained) *(b)* Skill *(c)* Cultural-specificity

Figure 5. LLM evaluation rankings using BENCHHUB (English).

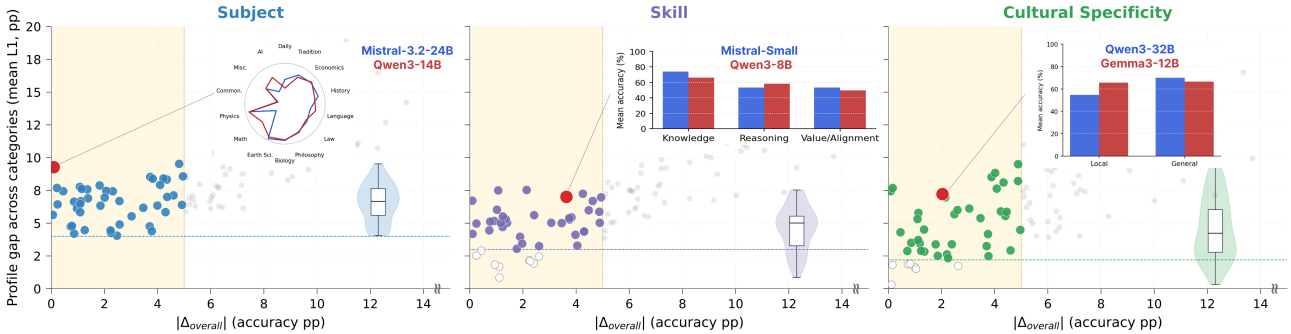


Figure 6. Model pairs with similar aggregate accuracy can show distinct category-level profiles. Each point is an unordered model pair. The shading region marks pairs within 5 %p differences in overall accuracy; red highlighted points indicate examples shown in the profile plots.

Table 1. Top-5 LLMs evaluated by BENCHHUB in real-world application scenarios.

Rank	(a) Legal chatbot with advanced reasoning (EN)		(b) Docent agent for Korean traditional art (KO)	
	Customized	Stratified	Customized	Stratified
1	Qwen3-30B-A3B-Instruct-2507	Qwen3-Next-80B-A3B-Instruct	Qwen3-Next-80B-A3B-Instruct	Qwen3-Next-80B-A3B-Instruct
2	Qwen3-Next-80B-A3B-Instruct	Qwen3-30B-A3B-Instruct-2507	Qwen3-32B	Qwen3-32B
3	Olmo-3.1-32B-Instruct	Qwen3-32B	Mistral-Small-3.2-24B-Instruct-2506	Qwen3-30B-A3B-Instruct-2507
4	gemma-3-12b-it	Olmo-3.1-32B-Instruct	Meta-Llama-3-70B-Instruct	Meta-Llama-3-70B-Instruct
5	Meta-Llama-3-70B-Instruct	Meta-Llama-3-70B-Instruct	gemma-3-12b-it	Qwen3-14B

evaluation changes the top-5 candidate set in every case. For the English legal chatbot, the top model changes from Qwen3-Next-80B-A3B-Instruct under the generic baseline to Qwen3-30B-A3B-Instruct-2507 under the customized profile. Even when the top model remains stable, the candidate set can shift: in the Korean docent scenario, Mistral-Small-3.2-24B-Instruct rises into the top-3 and Gemma3-12B enters the top-5, while Qwen3-30B-A3B drops out.

These results show that application-aligned evaluation is not merely a narrower version of generic benchmarking. Changing the target composition changes which capabilities are emphasized, which in turn can change the model selected for deployment. BENCHHUB therefore makes model selection a transparent composition-design problem rather than an implicit consequence of benchmark choice.

Takeaway. Generic benchmark rankings do not necessarily identify the best model for a specific use case. By making the target distribution explicit, BENCHHUB shows when application requirements change the selected model.

5. Related Work

As LLMs become integral to real-world generative AI systems, the focus on benchmarks and leaderboards has matured into evaluation *science* (Weidinger et al., 2025; Saxon et al., 2024). Our work contributes to this evolution by examining benchmark composition as both a property of the benchmark ecosystem and a factor that shapes model rank-

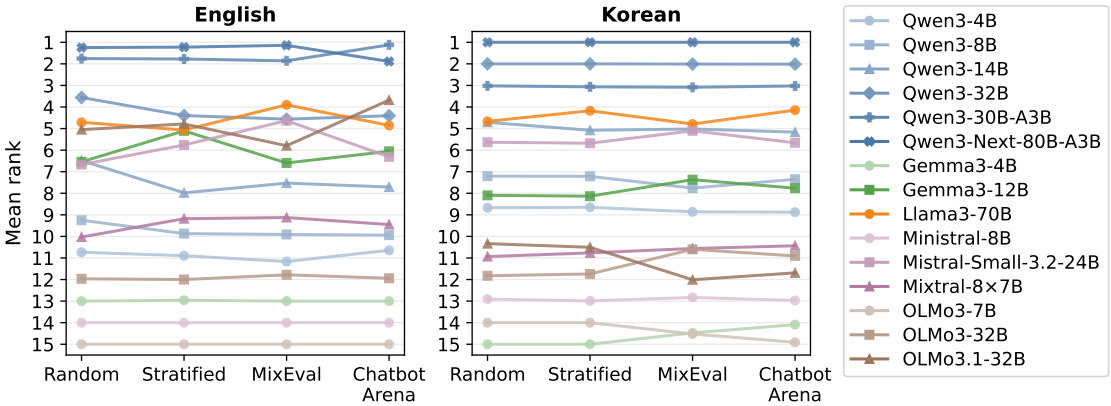


Figure 7. LLM ranking according to four sampling methods.

ings and selection decisions.

Benchmark Auditing. Recent work has begun to audit benchmarks themselves, asking whether benchmark can provide reliable measurement signals. This includes studies of data contamination, benchmark validity, discriminative power, item quality, and benchmark documentation (Kim et al., 2025b; Sokol et al., 2025b; Bean et al., 2025; Zhou et al., 2026). However, they do not characterize the broader benchmark ecosystem in terms of covered, missing, or redundantly reused content across datasets. In contrast, BENCHHUB provides an ecosystem-level view by mapping benchmark instances into a shared taxonomy, revealing composition skew, hidden coverage beyond dataset names, and cross-benchmark redundancy for more composition-aware analysis and reuse.

Benchmark aggregation and instance-level annotation. Prior work has expanded LLM evaluation by aggregating diverse benchmarks into broad evaluation suites. HELM evaluates models across multiple scenarios and metrics to capture trade-offs beyond accuracy (Liang et al., 2023), while MixEval constructs benchmark mixtures aligned with real-world user queries (Ni et al., 2024). Although aggregation improves coverage, it can obscure the internal composition of evaluation sets. A complementary line of work addresses this through instance-level annotations. FLASK provides human annotations of fine-grained skills, domains, and difficulty levels, improving interpretability but limiting scalability (Ye et al., 2024). ADeLe instead uses automated demand-level annotations to improve the explanatory and predictive power of evaluation (Zhou et al., 2025). BENCHHUB extends this direction with a comprehensive, automated, and dynamically extensible framework for composition-aware evaluation, making benchmark contents labelable, inspectable, and reusable at scale.

Controllable model evaluation. Recent work shows that model rankings and selection decisions are sensitive to benchmark choice and evaluation-set composition (Siska et al., 2024; Garbacea et al., 2026; Mishra & Arunkumar, 2021). Motivated by this, prior studies have proposed controllable evaluation frameworks, including user-adjustable leaderboards, instance reweighting, and prompt-specific ranking methods (Mishra & Arunkumar, 2021; Jung et al., 2026; Frick et al., 2025b). BENCHHUB advances this line of work by making benchmark composition explicit and controllable at the instance level. Rather than relying only on benchmark- or prompt-level customization, BENCHHUB maps instances into a shared taxonomy, enabling users to construct and rebalance evaluation sets for specific domains, skills, cultural contexts, and deployment needs.

6. Conclusion

We introduce BENCHHUB, a composition-aware evaluation framework that organizes existing LLM benchmarks into a shared instance-level representation. BENCHHUB integrates 54 benchmarks comprising 839K samples across 10 languages, with a primary focus on English and Korean. By making implicit benchmark composition explicit, BENCHHUB shows that benchmark-level usage can obscure the reusable structure of evaluation data. This reveals that existing benchmarks may support broader evaluation purposes than their names or high-level descriptions suggest. BENCHHUB also enables transparent comparisons among benchmarks designed for similar purposes by exposing differences in their internal composition. Beyond benchmark analysis, BENCHHUB provides a higher-resolution view of model performance through fine-grained instance-level categories, while also supporting the construction of controllable evaluation sets tailored to specific application needs. Together, our findings suggest that treating benchmarks as compositional resources, rather than fixed atomic units, can make LLM evaluation more interpretable, reusable, and application-aligned.

References

Alzahrani, N., Alyahya, H., Alnumay, Y., AlRashed, S., Alsubaie, S., Almushayqih, Y., Mirza, F., Alotaibi, N., Al-Twairesh, N., Alowisheq, A., Bari, M. S., and Khan, H. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13787–13805, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.744. URL https://aclanthology.org/2024.acl-long.744/.

Arora, S., Karpinska, M., Chen, H.-T., Bhattacharjee, I., Iyyer, M., and Choi, E. CaLMQA: Exploring culturally specific long-form question answering across 23 languages. *arXiv preprint arXiv:2406.17761*, 2024. URL https://arxiv.org/abs/2406.17761.

Austin, J., Odena, A., Nye, M., Bosma, M., Michalewski, H., Dohan, D., Jiang, E., Cai, C., Terry, M., Le, Q., et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021. URL https://arxiv.org/abs/2108.07732.

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

Bean, A. M., Kearns, R. O., Romanou, A., Hafner, F. S., Mayne, H., Batzner, J., Foroutan, N., Schmitz, C., Korgul, K., Batra, H., Deb, O., Beharry, E., Emde, C., Foster, T., Gausen, A., Grandury, M., Han, S., Hofmann, V., Ibrahim, L., Kim, H., Kirk, H. R., Lin, F., Liu, G. K.-M., Luettgau, L., Magomere, J., Rystrøm, J., Sotnikova, A., Yang, Y., Zhao, Y., Bibi, A., Bosselut, A., Clark, R., Cohan, A., Foerster, J., Gal, Y., Hale, S. A., Raji, I. D., Summerfield, C., Torr, P. H. S., Ududec, C., Rocher, L., and Mahdi, A. Measuring what matters: Construct validity in large language model benchmarks, 2025. URL https://arxiv.org/abs/2511.04703.

Bisk, Y., Zellers, R., Le bras, R., Gao, J., and Choi, Y. PIQA: Reasoning about physical commonsense in natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7432–7439, Apr. 2020. doi: 10.1609/aaai.v34i05.6239. URL https://ojs.aaai.org/index.php/AAAI/article/view/6239.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. D. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. URL https://arxiv.org/abs/2107.03374.

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhu, B., Zhang, H., Jordan, M., Gonzalez, J. E., and Stoica, I. Chatbot arena: An open platform for evaluating LLMs by human preference. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8359–8388. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/chiang24b.html.

Chiu, Y. Y., Jiang, L., Lin, B. Y., Park, C. Y., Li, S. S., Ravi, S., Bhatia, M., Antoniak, M., Tsvetkov, Y., Shwartz, V., et al. CulturalBench: a robust, diverse and challenging benchmark on measuring the (lack of) cultural knowledge of LLMs. *arXiv preprint arXiv:2410.02677*, 2024. URL https://arxiv.org/abs/2410.02677.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. URL https://arxiv.org/abs/1803.05457.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Frick, E., Chen, C., Tennyson, J., Li, T., Chiang, W.-L., Angelopoulos, A. N., and Stoica, I. Prompt-to-leaderboard: Prompt-adaptive LLM evaluations. In Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., and Zhu, J. (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 17672–17689. PMLR, 13–19 Jul 2025a. URL https://proceedings.mlr.press/v267/frick25a.html.

Frick, E., Chen, C., Tennyson, J., Li, T., Chiang, W.-L., Angelopoulos, A. N., and Stoica, I. Prompt-to-leaderboard, 2025b. URL https://arxiv.org/abs/2502.14855.

Garbacea, C., Wang, H., and Tan, C. Personalized benchmarking: Evaluating llms by individual preferences, 2026. URL https://arxiv.org/abs/2604.18943.

Guo, Z., Cheng, S., Wang, H., Liang, S., Qin, Y., Li, P., Liu, Z., Sun, M., and Liu, Y. StableToolBench: Towards stable large-scale benchmarking on tool learning of large language models. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational*

Linguistics: ACL 2024, pp. 11143–11156, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.664. URL https://aclanthology.org/2024.findings-acl.664/.

Hamotskyi, S., Levbarg, A.-I., and Hänig, C. Eval-UA-tion 1.0: Benchmark for evaluating Ukrainian (large) language models. In Romanyshyn, M., Romanyshyn, N., Hlybovets, A., and Ignatenko, O. (eds.), *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*, pp. 109–119, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.unlp-1.13/.

Hasan, M. A., Hasanain, M., Ahmad, F., Laskar, S. R., Upadhyay, S., Sukhadia, V. N., Kutlu, M., Chowdhury, S. A., and Alam, F. NativQA: Multilingual culturally-aligned natural query for LLMs. *arXiv preprint arXiv:2407.09823*, 2024. URL https://arxiv.org/abs/2407.09823.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the math dataset. In Vanschoren, J. and Yeung, S. (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021b. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/be83ab3ecd0db773eb2dc1b0a17836a1-Paper-round2.pdf.

Hijazi, F., AlHarbi, S., AlHussein, A., Shairah, H. A., AlZahrani, R., AlShamlan, H., Knio, O., and Turkiyyah, G. Arablegaleval: A multitask benchmark for assessing arabic legal knowledge in large language models. In *The Second Arabic Natural Language Processing Conference*, 2024. URL https://openreview.net/forum?id=3EHYXqKKLA.

Hsu, P.-L., Dai, Y., Kothapalli, V., Song, Q., Tang, S., Zhu, S., Shimizu, S., Sahni, S., Ning, H., and Chen, Y. Liger kernel: Efficient triton kernels for llm training. *arXiv preprint arXiv:2410.10989*, 2024.

Isbarov, J., Akhundjanova, A., Hajili, M., Huseynova, K., Gaynullin, D., Rzayev, A., Tursun, O., Saetov, I., Kharisov, R., Belginova, S., Kenbayeva, A., Alisheva, A., Turdubaeva, A., Köksal, A., Rustamov, S., and Ataman, D. TUMLU: A Unified and Native Language Understanding Benchmark for Turkic Languages, 2025. URL https://arxiv.org/abs/2502.11020.

Jin, J., Kim, J., Lee, N., Yoo, H., Oh, A., and Lee, H. KoBBQ: Korean bias benchmark for question answering. *Transactions of the Association for Computational Linguistics*, 12:507–524, 2024. doi: 10.1162/tacl_a_00661. URL https://aclanthology.org/2024.tacl-1.28/.

Jung, M., Lee, M., Kim, Y., Choi, S., and Kahng, M. Who defines "best"? towards interactive, user-defined evaluation of llm leaderboards, 2026. URL https://arxiv.org/abs/2604.21769.

Kim, D., Shim, G., Chun, Y., Kim, M., Park, C., and Lim, H. Benchmark profiling: Mechanistic diagnosis of LLM benchmarks. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V. (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 15635–15650, Suzhou, China, November 2025a. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.789. URL https://aclanthology.org/2025.emnlp-main.789/.

Kim, D., Shim, G., Chun, Y., Kim, M., Park, C., and Lim, H. Benchmark profiling: Mechanistic diagnosis of llm benchmarks, 2025b. URL https://arxiv.org/abs/2510.01232.

Kim, E., Suk, J., Oh, P., Yoo, H., Thorne, J., and Oh, A. CLIcK: A benchmark dataset of cultural and linguistic intelligence in Korean. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N. (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 3335–3346, Torino, Italia, May 2024a. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.296/.

Kim, S., Suk, J., Cho, J. Y., Longpre, S., Kim, C., Yoon, D., Son, G., Cho, Y., Shafayat, S., Baek, J., Park, S. H., Hwang, H., Jo, J., Cho, H., Shin, H., Lee, S., Oh, H., Lee, N., Ho, N., Joo, S. J., Ko, M., Lee, Y., Chae, H., Shin, J., Jang, J., Ye, S., Lin, B. Y., Welleck, S., Neubig, G., Lee, M., Lee, K., and Seo, M. The BiGGen bench: A principled benchmark for fine-grained evaluation of language models with language models. In Chiruzzo, L., Ritter, A., and Wang, L. (eds.), *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)*, pp. 5877–5919, Albuquerque, New Mexico, April 2025c. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.303.

URL https://aclanthology.org/2025.naacl-long.303/.

Kim, Y., Choi, Y., Choi, E., Choi, J., Park, H. J., and Hwang, W. Developing a pragmatic benchmark for assessing Korean legal language understanding in large language models. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5573–5595, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.319. URL https://aclanthology.org/2024.findings-emnlp.319/.

Ko, H., Son, G., and Choi, D. Understand, solve and translate: Bridging the multilingual mathematical reasoning gap. *arXiv preprint arXiv:2501.02448*, 2025. URL https://arxiv.org/abs/2501.02448.

Kočiský, T., Schwarz, J., Blunsom, P., Dyer, C., Hermann, K. M., Melis, G., and Grefenstette, E. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018. doi: 10.1162/tacl_a_00023. URL https://aclanthology.org/Q18-1023/.

Koto, F., Li, H., Shatnawi, S., Doughman, J., Sadallah, A., Alraeesi, A., Almubarak, K., Alyafeai, Z., Sengupta, N., Shehata, S., Habash, N., Nakov, P., and Baldwin, T. ArabicMMLU: Assessing massive multitask language understanding in Arabic. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5622–5640, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.334. URL https://aclanthology.org/2024.findings-acl.334/.

Kweon, S., Choi, B., Chu, G., Song, J., Hyeon, D., Gan, S., Kim, J., Kim, M., Park, R. W., and Choi, E. KorMedMCQA: multi-choice question answering benchmark for korean healthcare professional licensing examinations. *arXiv preprint arXiv:2403.01469*, 2024. URL https://arxiv.org/abs/2403.01469.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q., and Petrov, S. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Lee, H., Hong, S., Park, J., Kim, T., Kim, G., and Ha, J.-w. KoSBI: A dataset for mitigating social bias risks towards safer large language model applications. In Sitaram, S., Beigman Klebanov, B., and Williams, J. D. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pp. 208–224, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-industry.21. URL https://aclanthology.org/2023.acl-industry.21/.

Lee, H., Choi, D., Kim, S., Jung, I., Baek, S., Son, G., Hwang, I., Lee, N., and Hong, S. Redefining evaluation standards: A unified framework for evaluating the korean capabilities of language models, 2025. URL https://arxiv.org/abs/2503.22968.

Lee, J., Kim, M., Kim, S., Kim, J., Won, S., Lee, H., and Choi, E. KorNAT: LLM alignment benchmark for Korean social values and common knowledge. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 11177–11213, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.666. URL https://aclanthology.org/2024.findings-acl.666/.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Re, C., Acosta-Navas, D., Hudson, D. A., Zelikman, E., Durmus, E., Ladhak, F., Rong, F., Ren, H., Yao, H., WANG, J., Santhanam, K., Orr, L., Zheng, L., Yuksekgonul, M., Suzgun, M., Kim, N., Guha, N., Chatterji, N. S., Khattab, O., Henderson, P., Huang, Q., Chi, R. A., Xie, S. M., Santurkar, S., Ganguli, S., Hashimoto, T., Icard, T., Zhang, T., Chaudhary, V., Wang, W., Li, X., Mai, Y., Zhang, Y., and Koreeda, Y. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=iO4LZibEqW. Featured Certification, Expert Certification, Outstanding Certification.

Lin, S., Hilton, J., and Evans, O. TruthfulQA: Measuring how models mimic human falsehoods. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/.

Lovenia, H., Mahendra, R., Akbar, S. M., Miranda, L. J. V., Santoso, J., Aco, E., Fadhilah, A., Mansurov, J., Imperial, J. M., Kampman, O. P., Moniz, J. R. A., Habibi, M.

R. S., Hudi, F., Montalan, R., Ignatius, R., Lopo, J. A., Nixon, W., Karlsson, B. F., Jaya, J., Diandaru, R., Gao, Y., Amadeus, P., Wang, B., Cruz, J. C. B., Whitehouse, C., Parmonangan, I. H., Khelli, M., Zhang, W., Susanto, L., Ryanda, R. A., Hermawan, S. L., Velasco, D. J., Kautsar, M. D. A., Hendria, W. F., Moslem, Y., Flynn, N., Adilazuarda, M. F., Li, H., Lee, J., Damanhuri, R., Sun, S., Qorib, M. R., Djanibekov, A., Leong, W. Q., Do, Q. V., Muennighoff, N., Pansuwan, T., Putra, I. F., Xu, Y., Chia, T. N., Purwarianti, A., Ruder, S., Tjhi, W., Limkonchotiwat, P., Aji, A. F., Keh, S., Winata, G. I., Zhang, R., Koto, F., Yong, Z.-X., and Cahyawijaya, S. SEACrowd: A multilingual multimodal data hub and benchmark suite for Southeast Asian languages. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5155–5203, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.296. URL https://aclanthology.org/2024.emnlp-main.296/.

Mihaylov, T., Clark, P., Khot, T., and Sabharwal, A. Can a suit of armor conduct electricity? a new dataset for open book question answering. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J. (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2381–2391, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1260. URL https://aclanthology.org/D18-1260/.

Mishra, S. and Arunkumar, A. How robust are model rankings: A leaderboard customization approach for equitable evaluation, 2021. URL https://arxiv.org/abs/2106.05532.

Muennighoff, N., Soldaini, L., Groeneveld, D., Lo, K., Morrison, J., Min, S., Shi, W., Walsh, P., Tafjord, O., Lambert, N., Gu, Y., Arora, S., Bhagia, A., Schwenk, D., Wadden, D., Wettig, A., Hui, B., Dettmers, T., Kiela, D., Farhadi, A., Smith, N. A., Koh, P. W., Singh, A., and Hajishirzi, H. Olmoe: Open mixture-of-experts language models, 2025. URL https://arxiv.org/abs/2409.02060.

Myung, J., Lee, N., Zhou, Y., Jin, J., Putri, R. A., Antypas, D., Borkakoty, H., Kim, E., Perez-Almendros, C., Ayele, A. A., Gutiérrez-Basulto, V., Ibáñez García, Y., Lee, H., Muhammad, S. H., Park, K., Rzayev, A. S., White, N., Yimam, S. M., Pilehvar, M. T., Ousidhoum, N., Camacho-Collados, J., and Oh, A. BLEnD: A benchmark for llms on everyday knowledge in diverse cultures and languages. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 78104–78146. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/8eb88844dafefa92a26aaec9f3acad93-Paper-Datasets_and_Benchmarks_Track.pdf.

Nguyen, T.-P., Razniewski, S., Varde, A., and Weikum, G. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM Web Conference 2023*, WWW '23, pp. 1907–1917, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394161. doi: 10.1145/3543507.3583535. URL https://doi.org/10.1145/3543507.3583535.

Ni, J., Xue, F., Yue, X., Deng, Y., Shah, M., Jain, K., Neubig, G., and You, Y. MixEval: Deriving wisdom of the crowd from LLM benchmark mixtures. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 98180–98212. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/b1f34d7b4a03a3d80be8e72eb430dd81-Paper-Conference.pdf.

Nyachhyon, J., Sharma, M., Thapa, P., and Bal, B. K. Consolidating and developing benchmarking datasets for the nepali natural language understanding tasks, 2025. URL https://arxiv.org/abs/2411.19244.

OpenAI, :, Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., Mądry, A., Baker-Whitcomb, A., Beutel, A., Borzunov, A., Carney, A., Chow, A., Kirillov, A., Nichol, A., Paino, A., Renzin, A., Passos, A. T., Kirillov, A., Christakis, A., Conneau, A., Kamali, A., Jabri, A., Moyer, A., Tam, A., Crookes, A., Tootoochian, A., Tootoonchian, A., Kumar, A., Vallone, A., Karpathy, A., Braunstein, A., Cann, A., Codispoti, A., Galu, A., Kondrich, A., Tulloch, A., Mishchenko, A., Baek, A., Jiang, A., Pelisse, A., Woodford, A., Gosalia, A., Dhar, A., Pantuliano, A., Nayak, A., Oliver, A., Zoph, B., Ghorbani, B., Leimberger, B., Rossen, B., Sokolowsky, B., Wang, B., Zweig, B., Hoover, B., Samic, B., McGrew, B., Spero, B., Giertler, B., Cheng, B., Lightcap, B., Walkin, B., Quinn, B., Guarraci, B., Hsu, B., Kellogg, B., Eastman, B., Lugaresi, C., Wainwright, C., Bassin, C., Hudson, C., Chu, C., Nelson, C., Li, C., Shern, C. J., Conger, C., Barette, C., Voss, C., Ding, C., Lu, C., Zhang, C., Beaumont, C., Hallacy, C., Koch, C., Gibson, C., Kim, C., Choi, C., McLeavey, C., Hesse, C., Fischer, C., Winter, C., Czarnecki, C., Jarvis, C., Wei, C., Koumouzelis, C., Sherburn, D., Kappler, D., Levin, D., Levy, D., Carr, D., Farhi, D., Mely, D., Robinson, D., Sasaki, D., Jin, D., Valladares, D., Tsipras, D., Li, D., Nguyen, D. P., Findlay, D., Oiwoh, E., Wong, E., Asdar, E., Proehl, E., Yang, E., Antonow, E., Kramer, E., Peterson, E., Sigler, E., Wallace, E., Brevdo, E., Mays, E., Khorasani, F., Such, F. P., Raso, F., Zhang,

F., von Lohmann, F., Sulit, F., Goh, G., Oden, G., Salmon, G., Starace, G., Brockman, G., Salman, H., Bao, H., Hu, H., Wong, H., Wang, H., Schmidt, H., Whitney, H., Jun, H., Kirchner, H., de Oliveira Pinto, H. P., Ren, H., Chang, H., Chung, H. W., Kivlichan, I., O'Connell, I., O'Connell, I., Osband, I., Silber, I., Sohl, I., Okuyucu, I., Lan, I., Kostrikov, I., Sutskever, I., Kanitscheider, I., Gulrajani, I., Coxon, J., Menick, J., Pachocki, J., Aung, J., Betker, J., Crooks, J., Lennon, J., Kiros, J., Leike, J., Park, J., Kwon, J., Phang, J., Teplitz, J., Wei, J., Wolfe, J., Chen, J., Harris, J., Varavva, J., Lee, J. G., Shieh, J., Lin, J., Yu, J., Weng, J., Tang, J., Yu, J., Jang, J., Candela, J. Q., Beutler, J., Landers, J., Parish, J., Heidecke, J., Schulman, J., Lachman, J., McKay, J., Uesato, J., Ward, J., Kim, J. W., Huizinga, J., Sitkin, J., Kraaijeveld, J., Gross, J., Kaplan, J., Snyder, J., Achiam, J., Jiao, J., Lee, J., Zhuang, J., Harriman, J., Fricke, K., Hayashi, K., Singhal, K., Shi, K., Karthik, K., Wood, K., Rimbach, K., Hsu, K., Nguyen, K., Gu-Lemberg, K., Button, K., Liu, K., Howe, K., Muthukumar, K., Luther, K., Ahmad, L., Kai, L., Itow, L., Workman, L., Pathak, L., Chen, L., Jing, L., Guy, L., Fedus, L., Zhou, L., Mamitsuka, L., Weng, L., McCallum, L., Held, L., Ouyang, L., Feuvrier, L., Zhang, L., Kondraciuk, L., Kaiser, L., Hewitt, L., Metz, L., Doshi, L., Aflak, M., Simens, M., Boyd, M., Thompson, M., Dukhan, M., Chen, M., Gray, M., Hudnall, M., Zhang, M., Aljubeh, M., Litwin, M., Zeng, M., Johnson, M., Shetty, M., Gupta, M., Shah, M., Yatbaz, M., Yang, M. J., Zhong, M., Glaese, M., Chen, M., Janner, M., Lampe, M., Petrov, M., Wu, M., Wang, M., Fradin, M., Pokrass, M., Castro, M., de Castro, M. O. T., Pavlov, M., Brundage, M., Wang, M., Khan, M., Murati, M., Bavarian, M., Lin, M., Yesildal, M., Soto, N., Gimelshein, N., Cone, N., Staudacher, N., Summers, N., LaFontaine, N., Chowdhury, N., Ryder, N., Stathas, N., Turley, N., Tezak, N., Felix, N., Kudige, N., Keskar, N., Deutsch, N., Bundick, N., Puckett, N., Nachum, O., Okelola, O., Boiko, O., Murk, O., Jaffe, O., Watkins, O., Godement, O., Campbell-Moore, O., Chao, P., McMillan, P., Belov, P., Su, P., Bak, P., Bakkum, P., Deng, P., Dolan, P., Hoeschele, P., Welinder, P., Tillet, P., Pronin, P., Tillet, P., Dhariwal, P., Yuan, Q., Dias, R., Lim, R., Arora, R., Troll, R., Lin, R., Lopes, R. G., Puri, R., Miyara, R., Leike, R., Gaubert, R., Zamani, R., Wang, R., Donnelly, R., Honsby, R., Smith, R., Sahai, R., Ramchandani, R., Huet, R., Carmichael, R., Zellers, R., Chen, R., Chen, R., Nigmatullin, R., Cheu, R., Jain, S., Altman, S., Schoenholz, S., Toizer, S., Miserendino, S., Agarwal, S., Culver, S., Ethersmith, S., Gray, S., Grove, S., Metzger, S., Hermani, S., Jain, S., Zhao, S., Wu, S., Jomoto, S., Wu, S., Shuaiqi, Xia, Phene, S., Papay, S., Narayanan, S., Coffey, S., Lee, S., Hall, S., Balaji, S., Broda, T., Stramer, T., Xu, T., Gogineni, T., Christianson, T., Sanders, T., Patwardhan, T., Cunninghman, T., Degry, T., Dimson, T., Raoux, T., Shadwell, T., Zheng, T., Underwood, T., Markov, T., Sherbakov, T., Rubin, T., Stasi, T., Kaftan, T., Heywood, T., Peterson, T., Walters, T., Eloundou, T., Qi, V., Moeller, V., Monaco, V., Kuo, V., Fomenko, V., Chang, W., Zheng, W., Zhou, W., Manassra, W., Sheu, W., Zaremba, W., Patil, Y., Qian, Y., Kim, Y., Cheng, Y., Zhang, Y., He, Y., Zhang, Y., Jin, Y., Dai, Y., and Malkov, Y. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P. M., and Bowman, S. BBQ: A hand-built bias benchmark for question answering. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2086–2105, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.165. URL https://aclanthology.org/2022.findings-acl.165/.

Pawar, S., Ramrakhiyani, N., Sinha, A., Apte, M., and Palshikar, G. Why generate when you can discriminate? a novel technique for text classification using language models. In Graham, Y. and Purver, M. (eds.), *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 1099–1114, St. Julian's, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-eacl.74. URL https://aclanthology.org/2024.findings-eacl.74/.

Pfister, J. and Hotho, A. SuperGLEBer: German language understanding evaluation benchmark. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7904–7923, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.438. URL https://aclanthology.org/2024.naacl-long.438/.

Ponti, E. M., Glavaš, G., Majewska, O., Liu, Q., Vulić, I., and Korhonen, A. XCOPA: A multilingual dataset for causal commonsense reasoning. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2362–2376, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.185. URL https://aclanthology.org/2020.emnlp-main.185/.

Putri, R. A. and Oh, A. IDK-MRC: Unanswerable questions for Indonesian machine reading comprehension. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Proceedings*

of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 6918–6933, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.465. URL https://aclanthology.org/2022.emnlp-main.465/.

Qin, Y., Liang, S., Ye, Y., Zhu, K., Yan, L., Lu, Y., Lin, Y., Cong, X., Tang, X., Qian, B., Zhao, S., Tian, R., Xie, R., Zhou, J., Gerstein, M., Li, D., Liu, Z., and Sun, M. Toolllm: Facilitating large language models to master 16,000+ real-world apis. *arXiv preprint arXiv:2307.16789*, 2023. URL https://arXiv.org/abs/2307.16789.

Rajbhandari, S., Rasley, J., Ruwase, O., and He, Y. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–16. IEEE, 2020.

Rao, A., Yerukola, A., Shah, V., Reinecke, K., and Sap, M. NormAd: A framework for measuring the cultural adaptability of large language models. In Chiruzzo, L., Ritter, A., and Wang, L. (eds.), *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)*, pp. 2373–2403, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.120. URL https://aclanthology.org/2025.naacl-long.120/.

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., and Bowman, S. R. GPQA: A graduate-level google-proof Q&A benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Reuel, A., Hardy, A., Smith, C., Lamparth, M., Hardy, M., and Kochenderfer, M. J. BetterBench: Assessing AI benchmarks, uncovering issues, and establishing best practices. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 21763–21813. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/26889e8359e7ef8a7f5d77457364ca55-Paper-Datasets_and_Benchmarks_Track.pdf.

Ribeiro, M. T., Wu, T., Guestrin, C., and Singh, S. Beyond accuracy: Behavioral testing of NLP models with CheckList. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Rizqullah, M., Purwarianti, A., and Aji, A. F. QASiNa: Religious domain question answering using sirah nabawiyah. In *preprint / arXiv*, 2023. URL https://arxiv.org/abs/2310.08102.

Sakaguchi, K., Bras, R. L., Bhagavatula, C., and Choi, Y. WinoGrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106, August 2021. ISSN 0001-0782. doi: 10.1145/3474381. URL https://doi.org/10.1145/3474381.

Sap, M., Rashkin, H., Chen, D., Le Bras, R., and Choi, Y. Social IQa: Commonsense reasoning about social interactions. In Inui, K., Jiang, J., Ng, V., and Wan, X. (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4463–4473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1454. URL https://aclanthology.org/D19-1454/.

Saxon, M., Holtzman, A., West, P., Wang, W. Y., and Saphra, N. Benchmarks as microscopes: A call for model metrology. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=bttKwCZDkm.

Shi, W., Li, R., Zhang, Y., Ziems, C., Yu, S., Horesh, R., Paula, R. A. D., and Yang, D. CultureBank: An online community-driven knowledge base towards culturally aware language technologies. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4996–5025, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.288. URL https://aclanthology.org/2024.findings-emnlp.288/.

Singh, S., Romanou, A., Fourrier, C., Adelani, D. I., Ngui, J. G., Vila-Suero, D., Limkonchotiwat, P., Marchisio, K., Leong, W. Q., Susanto, Y., Ng, R., Longpre, S., Ko, W.-Y., Ruder, S., Smith, M., Bosselut, A., Oh, A., Martins, A. F. T., Choshen, L., Ippolito, D., Ferrante, E., Fadaee, M., Ermis, B., and Hooker, S. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *arXiv preprint arXiv:2412.03304*, 2024. URL https://arxiv.org/abs/2412.03304.

Singh, S., Nan, Y., Wang, A., Dsouza, D., Kapoor, S., Üstün, A., Koyejo, S., Deng, Y., Longpre, S., Smith, N., Ermis,

B., Fadaee, M., and Hooker, S. The leaderboard illusion. In Belgrave, D., Zhang, C., Lin, H., Pascanu, R., Koniusz, P., Ghassemi, M., and Chen, N. (eds.), *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/70a93f260a51123b3c0e33ecd1b4de97-Paper-Datasets_and_Benchmarks_Track.pdf.

Siska, C., Marazopoulou, K., Ailem, M., and Bono, J. Examining the robustness of LLM evaluation to the distributional assumptions of benchmarks. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10406–10421, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.560. URL https://aclanthology.org/2024.acl-long.560/.

Sokol, A., Daly, E., Hind, M., Piorkowski, D., Zhang, X., Moniz, N., and Chawla, N. BenchmarkCards: Standardized documentation for large language model benchmarks. In Belgrave, D., Zhang, C., Lin, H., Pascanu, R., Koniusz, P., Ghassemi, M., and Chen, N. (eds.), *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025a. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/76175f4355e2f67cf91be468c8860070-Paper-Datasets_and_Benchmarks_Track.pdf.

Sokol, A., Daly, E., Hind, M., Piorkowski, D., Zhang, X., Moniz, N., and Chawla, N. Benchmarkcards: Standardized documentation for large language model benchmarks, 2025b. URL https://arxiv.org/abs/2410.12974.

Son, G., Lee, H., Kim, S., Kim, H., Lee, J. c., Yeom, J. W., Jung, J., Kim, J. w., and Kim, S. HAE-RAE bench: Evaluation of Korean knowledge in language models. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N. (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 7993–8007, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.704/.

Son, G., Ko, H., and Choi, D. Multi-step reasoning in Korean and the emergent mirage. In Prabhakaran, V., Dev, S., Benotti, L., Hershcovich, D., Cao, Y., Zhou, L., Cabello, L., and Adebara, I. (eds.), *Proceedings of the 3rd Workshop on Cross-Cultural Considerations in NLP (C3NLP 2025)*, pp. 10–21, Albuquerque, New Mexico, May 2025a. Association for Computational Linguistics. ISBN 979-8-89176-237-4. doi: 10.18653/v1/2025.c3nlp-1.2. URL https://aclanthology.org/2025.c3nlp-1.2/.

Son, G., Lee, H., Kim, S., Kim, S., Muennighoff, N., Choi, T., Park, C., Yoo, K. M., and Biderman, S. KMMLU: Measuring massive multitask language understanding in Korean. In Chiruzzo, L., Ritter, A., and Wang, L. (eds.), *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)*, pp. 4076–4104, Albuquerque, New Mexico, April 2025b. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.206. URL https://aclanthology.org/2025.naacl-long.206/.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., Xiang, A., Parrish, A., Nie, A., Hussain, A., Askell, A., Dsouza, A., Slone, A., Rahane, A., Iyer, A. S., Andreassen, A. J., Madotto, A., Santilli, A., Stuhlmüller, A., Dai, A. M., La, A., Lampinen, A. K., Zou, A., Jiang, A., Chen, A., Vuong, A., Gupta, A., Gottardi, A., Norelli, A., Venkatesh, A., Gholamidavoodi, A., Tabassum, A., Menezes, A., Kirubarajan, A., Mullokandov, A., Sabharwal, A., Herrick, A., Efrat, A., Erdem, A., Karakaş, A., Roberts, B. R., Loe, B. S., Zoph, B., Bojanowski, B., Özyurt, B., Hedayatnia, B., Neyshabur, B., Inden, B., Stein, B., Ekmekci, B., Lin, B. Y., Howald, B., Orinion, B., Diao, C., Dour, C., Stinson, C., Argueta, C., Ferri, C., Singh, C., Rathkopf, C., Meng, C., Baral, C., Wu, C., Callison-Burch, C., Waites, C., Voigt, C., Manning, C. D., Potts, C., Ramirez, C., Rivera, C. E., Siro, C., Raffel, C., Ashcraft, C., Garbacea, C., Sileo, D., Garrette, D., Hendrycks, D., Kilman, D., Roth, D., Freeman, C. D., Khashabi, D., Levy, D., González, D. M., Perszyk, D., Hernandez, D., Chen, D., Ippolito, D., Gilboa, D., Dohan, D., Drakard, D., Jurgens, D., Datta, D., Ganguli, D., Emelin, D., Kleyko, D., Yuret, D., Chen, D., Tam, D., Hupkes, D., Misra, D., Buzan, D., Mollo, D. C., Yang, D., Lee, D.-H., Schrader, D., Shutova, E., Cubuk, E. D., Segal, E., Hagerman, E., Barnes, E., Donoway, E., Pavlick, E., Rodolà, E., Lam, E., Chu, E., Tang, E., Erdem, E., Chang, E., Chi, E. A., Dyer, E., Jerzak, E., Kim, E., Manyasi, E. E., Zheltonozhskii, E., Xia, F., Siar, F., Martínez-Plumed, F., Happé, F., Chollet, F., Rong, F., Mishra, G., Winata, G. I., de Melo, G., Kruszewski, G., Parascandolo, G., Mariani, G., Wang, G. X., Jaimovitch-Lopez, G., Betz, G., Gur-Ari, G., Galijasevic, H., Kim, H., Rashkin, H., Hajishirzi, H., Mehta, H., Bogar, H., Shevlin, H. F. A., Schuetze, H., Yakura, H., Zhang, H., Wong, H. M., Ng, I., Noble, I., Jumelet, J., Geissinger, J.,

Kernion, J., Hilton, J., Lee, J., Fisac, J. F., Simon, J. B., Koppel, J., Zheng, J., Zou, J., Kocon, J., Thompson, J., Wingfield, J., Kaplan, J., Radom, J., Sohl-Dickstein, J., Phang, J., Wei, J., Yosinski, J., Novikova, J., Bosscher, J., Marsh, J., Kim, J., Taal, J., Engel, J., Alabi, J., Xu, J., Song, J., Tang, J., Waweru, J., Burden, J., Miller, J., Balis, J. U., Batchelder, J., Berant, J., Frohberg, J., Rozen, J., Hernandez-Orallo, J., Boudeman, J., Guerr, J., Jones, J., Tenenbaum, J. B., Rule, J. S., Chua, J., Kanclerz, K., Livescu, K., Krauth, K., Gopalakrishnan, K., Ignatyeva, K., Markert, K., Dhole, K., Gimpel, K., Omondi, K., Mathewson, K. W., Chiafullo, K., Shkaruta, K., Shridhar, K., McDonell, K., Richardson, K., Reynolds, L., Gao, L., Zhang, L., Dugan, L., Qin, L., Contreras-Ochando, L., Morency, L.-P., Moschella, L., Lam, L., Noble, L., Schmidt, L., He, L., Oliveros-Colón, L., Metz, L., Senel, L. K., Bosma, M., Sap, M., Hoeve, M. T., Farooqi, M., Faruqui, M., Mazeika, M., Baturan, M., Marelli, M., Maru, M., Ramirez-Quintana, M. J., Tolkiehn, M., Giulianelli, M., Lewis, M., Potthast, M., Leavitt, M. L., Hagen, M., Schubert, M., Baitemirova, M. O., Arnaud, M., McElrath, M., Yee, M. A., Cohen, M., Gu, M., Ivanitskiy, M., Starritt, M., Strube, M., Swędrowski, M., Bevilacqua, M., Yasunaga, M., Kale, M., Cain, M., Xu, M., Suzgun, M., Walker, M., Tiwari, M., Bansal, M., Aminnaseri, M., Geva, M., Gheini, M., T, M. V., Peng, N., Chi, N. A., Lee, N., Krakover, N. G.-A., Cameron, N., Roberts, N., Doiron, N., Martinez, N., Nangia, N., Deckers, N., Muennighoff, N., Keskar, N. S., Iyer, N. S., Constant, N., Fiedel, N., Wen, N., Zhang, O., Agha, O., Elbaghdadi, O., Levy, O., Evans, O., Casares, P. A. M., Doshi, P., Fung, P., Liang, P. P., Vicol, P., Alipoormolabashi, P., Liao, P., Liang, P., Chang, P. W., Eckersley, P., Htut, P. M., Hwang, P., Miłkowski, P., Patil, P., Pezeshkpour, P., Oli, P., Mei, Q., Lyu, Q., Chen, Q., Banjade, R., Rudolph, R. E., Gabriel, R., Habacker, R., Risco, R., Millière, R., Garg, R., Barnes, R., Saurous, R. A., Arakawa, R., Raymaekers, R., Frank, R., Sikand, R., Novak, R., Sitelew, R., Bras, R. L., Liu, R., Jacobs, R., Zhang, R., Salakhutdinov, R., Chi, R. A., Lee, S. R., Stovall, R., Teehan, R., Yang, R., Singh, S., Mohammad, S. M., Anand, S., Dillavou, S., Shleifer, S., Wiseman, S., Gruetter, S., Bowman, S. R., Schoenholz, S. S., Han, S., Kwatra, S., Rous, S. A., Ghazarian, S., Ghosh, S., Casey, S., Bischoff, S., Gehrmann, S., Schuster, S., Sadeghi, S., Hamdan, S., Zhou, S., Srivastava, S., Shi, S., Singh, S., Asaadi, S., Gu, S. S., Pachchigar, S., Toshniwal, S., Upadhyay, S., Debnath, S. S., Shakeri, S., Thormeyer, S., Melzi, S., Reddy, S., Makini, S. P., Lee, S.-H., Torene, S., Hatwar, S., Dehaene, S., Divic, S., Ermon, S., Biderman, S., Lin, S., Prasad, S., Piantadosi, S., Shieber, S., Misherghi, S., Kiritchenko, S., Mishra, S., Linzen, T., Schuster, T., Li, T., Yu, T., Ali, T., Hashimoto, T., Wu, T.-L., Desbordes, T., Rothschild, T., Phan, T., Wang, T., Nkinyili, T., Schick, T., Kornev, T., Tunduny, T., Gerstenberg, T., Chang, T., Neeraj, T., Khot, T., Shultz, T., Shaham, U., Misra, V., Demberg, V., Nyamai, V., Raunak, V., Ramasesh, V. V., vinay uday prabhu, Padmakumar, V., Srikumar, V., Fedus, W., Saunders, W., Zhang, W., Vossen, W., Ren, X., Tong, X., Zhao, X., Wu, X., Shen, X., Yaghoobzadeh, Y., Lakretz, Y., Song, Y., Bahri, Y., Choi, Y., Yang, Y., Hao, S., Chen, Y., Belinkov, Y., Hou, Y., Hou, Y., Bai, Y., Seid, Z., Zhao, Z., Wang, Z., Wang, Z. J., Wang, Z., and Wu, Z. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=uyTL5Bvosj. Featured Certification.

Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q., Chi, E., Zhou, D., and Wei, J. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.824. URL https://aclanthology.org/2023.findings-acl.824/.

Talmor, A., Herzig, J., Lourie, N., and Berant, J. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Burstein, J., Doran, C., and Solorio, T. (eds.), *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)*, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421/.

Wang, B., Liu, Z., Huang, X., Jiao, F., Ding, Y., Aw, A., and Chen, N. SeaEval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. In Duh, K., Gomez, H., and Bethard, S. (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 370–390, Mexico City, Mexico, June 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.22. URL https://aclanthology.org/2024.naacl-long.22/.

Wang, X., Yeo, J., Lim, J.-H., and Kim, H. KULTURE Bench: A benchmark for assessing language model in Korean cultural context. *arXiv preprint arXiv:2412.07251*, 2024b. URL https://arxiv.org/abs/2412.07251.

Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo,

S., Ren, W., Arulraj, A., He, X., Jiang, Z., Li, T., Ku, M., Wang, K., Zhuang, A., Fan, R., Yue, X., and Chen, W. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 95266–95290. Curran Associates, Inc., 2024c. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ad236edc564f3e3156e1b2feafb99a24-Paper-Datasets_and_Benchmarks_Track.pdf.

Weidinger, L., Raji, I. D., Wallach, H., Mitchell, M., Wang, A., Salaudeen, O., Bommasani, R., Ganguli, D., Koyejo, S., and Isaac, W. Toward an evaluation science for generative AI systems. *arXiv preprint arXiv:2503.05336*, 2025. URL https://arxiv.org/abs/2503.05336.

Yang, R., Song, L., Li, Y., Zhao, S., Ge, Y., Li, X., and Shan, Y. Gpt4tools: Teaching large language model to use tools via self-instruction. *arXiv preprint arXiv:2305.18752*, 2023. URL https://arXiv.org/abs/2305.18752.

Ye, J., Du, Z., Yao, X., Lin, W., Xu, Y., Chen, Z., Wang, Z., Zhu, S., Xi, Z., Yuan, S., Gui, T., Zhang, Q., Huang, X., and Chen, J. ToolHop: A query-driven benchmark for evaluating large language models in multi-hop tool use. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2995–3021, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.150. URL https://aclanthology.org/2025.acl-long.150/.

Ye, S., Kim, D., Kim, S., Hwang, H., Kim, S., Jo, Y., Thorne, J., Kim, J., and Seo, M. FLASK: Fine-grained language model evaluation based on alignment skill sets. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=CYmF38ysDa.

Yin, D., Bansal, H., Monajatipoor, M., Li, L. H., and Chang, K.-W. GeoMLAMA: Geo-diverse commonsense probing on multilingual pre-trained language models. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2039–2055, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.132. URL https://aclanthology.org/2022.emnlp-main.132/.

Zhou, H., Huang, H., Zhao, Z., Han, L., Wang, H., Chen, K., Yang, M., Bao, W., Dong, J., Xu, B., Zhu, C., Cao, H., and Zhao, T. Lost in benchmarks? rethinking large language model benchmarking with item response theory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Zhou, L., Pacchiardi, L., Martínez-Plumed, F., Collins, K. M., Moros-Daval, Y., Zhang, S., Zhao, Q., Huang, Y., Sun, L., Prunty, J. E., Li, Z., Sánchez-García, P., Chen, K. J., Casares, P. A. M., Zu, J., Burden, J., Mehrbakhsh, B., Stillwell, D., Cebrian, M., Wang, J., Henderson, P., Wu, S. T., Kyllonen, P. C., Cheke, L., Xie, X., and Hernández-Orallo, J. General scales unlock ai evaluation with explanatory and predictive power, 2025. URL https://arxiv.org/abs/2503.06378.

Zhuang, Y., Yu, Y., Wang, K., Sun, H., and Zhang, C. Toolqa: A dataset for llm question answering with external tools. In *NeurIPS 2023 Datasets and Benchmarks Track / OpenReview*, 2023. URL https://openreview.net/forum?id=pV1xV2RK6I.

A. Limitations

By acknowledging these limitations, we aim to continuously improve BENCHHUB and encourage contributions from the community.

Incomplete English Dataset Coverage. Due to the vast amount of English evaluation 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.

Categorization Bias from LLMs. BENCHHUB's automated categorization relies on Qwen-2.5-7B, which may introduce biases due to its training data or modeling limitations. Although we validate robustness to categorization errors up to 1.5% (Appendix D.2), future work will explore human-in-the-loop methods and ensemble approaches.

Potential Data Contamination. BENCHHUB aggregates existing benchmarks and thus inherits any contamination present in its sources. While BENCHHUB itself does not introduce new contamination risk, results may be influenced if evaluated models were exposed to included benchmark samples. We provide a controlled contamination simulation in Appendix H.4.

B. Broader Impact Statement

BENCHHUB aims to make LLM evaluation more transparent, reusable, and application-aligned by exposing the instance-level composition of existing benchmarks. Its potential positive impacts include helping researchers and practitioners better understand what benchmarks actually contain, identify underrepresented subjects or skills, and construct evaluation sets that better match their intended applications. Its positive impacts can further extend to multilingual evaluation by enabling comparisons of benchmark coverage across languages and helping identify underrepresented languages and cultures in existing evaluation resources. By moving beyond benchmark names and aggregate scores, BENCHHUB can also support more interpretable model comparison and reduce the risk of selecting models based on incomplete or misleading benchmark-level summaries.

At the same time, composition-aware evaluation can introduce potential risks if used uncritically. First, users may over-rely on automatically assigned instance-level labels, even though such labels can be incomplete or noisy. This could lead to inappropriate conclusions about model capabilities or benchmark coverage. Second, the ability to construct customized evaluation sets may be misused to selectively report favorable model rankings or to overfit evaluation choices to a desired conclusion. Third, because BENCHHUB reveals reusable contents within existing benchmarks, it could encourage benchmark reuse in contexts for which the original data were not designed.

To mitigate these risks, we provide validation analyses of the automated categorization process and encourage users to inspect sampled instances, account for label uncertainty, and clearly document the criteria used to construct customized evaluation sets. To support transparent and reproducible evaluation, we provide a code utility that reports both aggregate and fine-grained results, along with the composition of the selected evaluation data.

C. Interactive Platform and Utilities of BENCHHUB

We provide a web interface (Appendix C.1), additional code utilities for data loading and evaluation (Appendix C.2).

- Our interface will be available on Hugging Face Space upon acceptance.
- Our code utilities are available at `https://github.com/rladmstn1714/BenchHub`.

C.1. BENCHHUB Web Interface

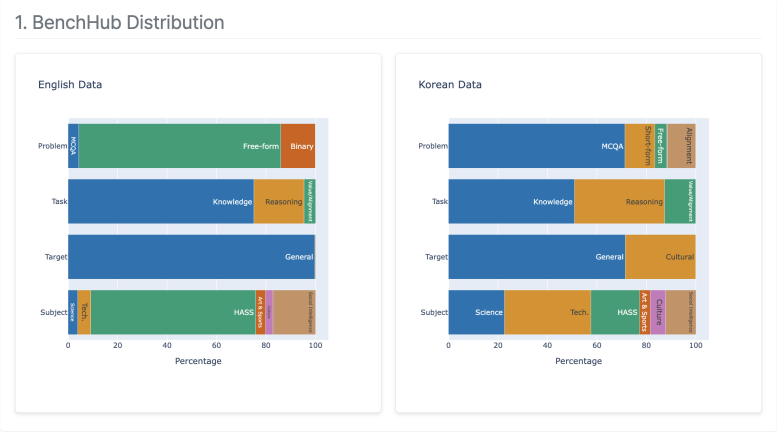
We provide the BENCHHUB web interface (Figure 8) to enable users to interactively explore available datasets and identify those that best suit their needs. It also supports the continuous addition and management of new data. Through a submission form, new datasets can be detected and automatically added.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

(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.

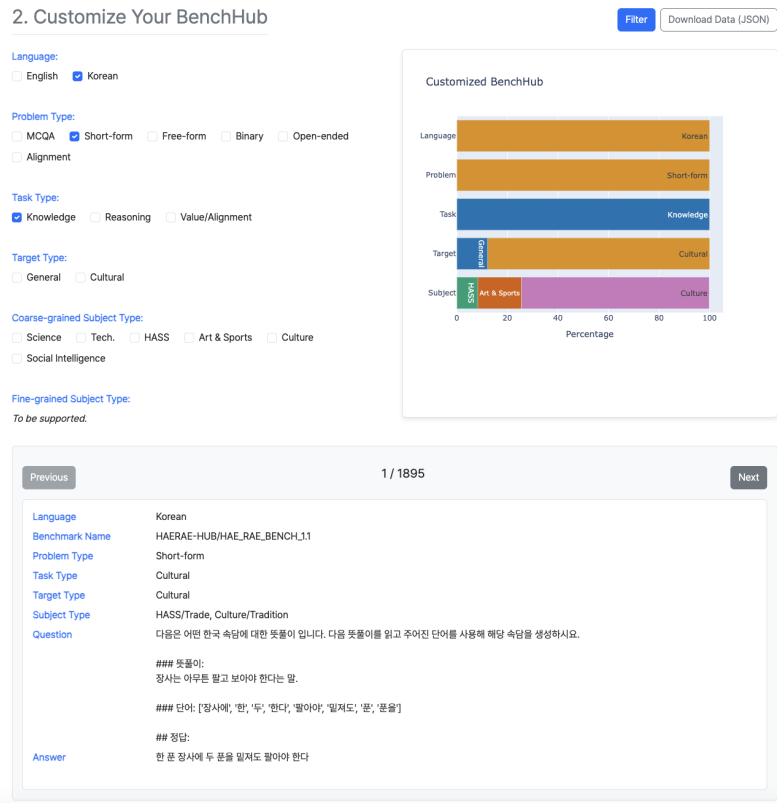
(2) Customizing BENCHHUB (Figure 8b) This allows users to access sample lists and statistics for selected categories. By reviewing samples, users can verify whether the dataset matches their needs and explore datasets suitable for their purposes. Users can also download the entire set corresponding to the samples.[4]

(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.

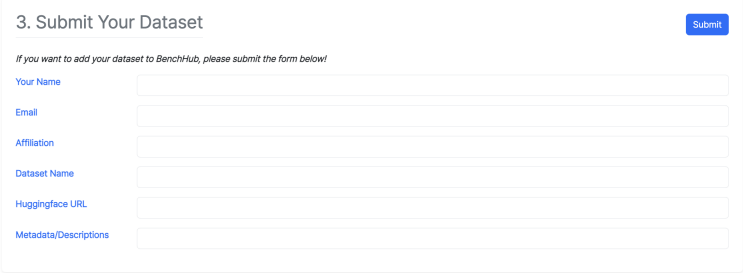
[4]Additional customizing features, such as fine-grained category adjustments and interactive control of category proportions via the platform (*e.g.*, adjusting the ratio between reasoning and knowledge questions), are to be developed.



(a) BENCHHUB Distribution



(b) Dataset Selection



(c) New Dataset Submission

Figure 8. User Interface of BENCHHUB Web Demo.

C.2. BENCHHUB Code Utilities

To proliferate our structured data into actionable insights for researchers and practitioners, we release 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:

C.2.1. DATASET LOADER

The dataset loader filters the aggregated benchmark pool according to the categories selected by the user. It provides two loading options: (1) returning all instances that match the specified categories, or (2) returning a deduplicated version in which overlapping entries, including near-duplicates, are removed. The latter is useful because multiple aggregated benchmarks may contain overlapping or semantically similar samples.

Duplicate filtering method. To perform deduplication, we implement an embedding-based sampling method inspired by MixEval (Ni et al., 2024). The process consists of two steps: (1) computing query embeddings using `mpnet-base-v2` from SentenceTransformers and projecting them into a two-dimensional space via t-SNE, and (2) uniformly sampling instances in this reduced space. Because queries on similar topics tend to cluster in localized regions of the embedding map, this procedure helps reduce redundant samples during dataset construction.

Empirical validation. To validate this approach, we conduct a controlled duplicate-injection experiment. We first extract 7,715 English BENCHHUB samples categorized under mathematics. We then introduce 60 synthetic duplicates by prompting `gemini-2.5-flash` to generate identical copies and five near-duplicates for 10 randomly selected questions through paraphrasing or numerical perturbation. Finally, we apply the embedding-based projection and uniform sampling procedure described above.

We find that embedding-based sampling consistently limits duplicates to at most 0–1 per batch, even at large sample sizes. In contrast, random sampling frequently yields more than five duplicates once the sample size exceeds 1,250. Detailed results are shown in Figure 9.

C.2.2. CITATION REPORT GENERATOR

For each customized dataset, BENCHHUB generates a LaTeX report that summarizes source datasets, licenses, instance counts, and citation guidelines. It also provides BibTeX entries to help users directly incorporate the selected dataset into their projects while ensuring proper credit to the original dataset authors.

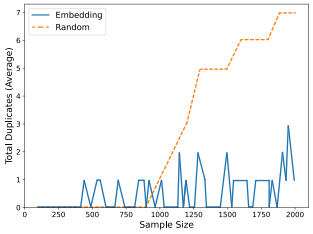


Figure 9. Average number of duplicates included in the sampling size when using the embedding-based method (Blue) and random sampling (Orange).

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{...}
```

C.2.3. AUTOMATIC EVALUATION PIPELINE

For better reproducibility, we adopt HRET (Lee et al., 2025) [5], enabling direct evaluations on BENCHHUB.

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

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.

D. 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.

D.1. Automated Categorization Process

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

D.1.1. TRAINING CATEGORIZER FOR ENGLISH AND KOREAN LANGUAGE

Table 2. SFT configuration details.

Hyperparameter	Value
Sequence Length	8,192
Learning Rate	2×10^{-5}
Global Batch (Effective)	256
Learning Rate Scheduler	Cosine Decay
Warmup Ratio	0.05
Training Epochs	3

We fine-tune the Qwen-2.5-7B-instruct models to automatically categorize the skill, subject, and cultural specificity type of a given sample. Table 2 shows the SFT configs of the categorizer model. In Table 3, we report accuracy by comparing model predictions against human-annotated labels. 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, subject, 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 (OpenAI et al., 2024) 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.

Table 3. Accuracy of the fine-tuned categorizer based on Qwen2.5-7B.

Sample-level Attribute	Accuracy
Subject	0.871
Skill	0.967
Cultural-specificity	0.986

For the target category, we adopt a binary classification scheme consistent with prior work such as Global MMLU (Singh et al., 2024). 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.

Table 3 shows the performance of the categorizer across three classification types. To evaluate categorization, we use existing datasets that already have fine-grained categories, which are similar to or more confined than our taxonomy. For the subject, we used a multi-label approach, whereas they used a single label; we considered it correct if the designated main category was included. Since the existing datasets often do not contain all three taxonomies we've defined, we compare the available category types in the evaluation. For example, MMLU defines only the subject category, while MATH datasets fall under the 'reasoning' (skill type). Both subject and skill showed high performance, while the target (to determine which culture/country the query is related to) scored lower. The target has a large number of possible culture options (all countries), making it difficult to train for all options. Therefore, instead of segmenting the target by country/culture, we enhance reliability by separating it into two labels: general and cultural, based on the initial label from categorize.

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.*, skill type), whether it's a question about a specific culture or a general question (*i.e.*, cultural specificity 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: ### **skill 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"

D.2. Reliability of Automated Categorization

D.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.

D.2.2. ENHANCING CATEGORIZATION ROBUSTNESS

The classification process of BENCHHUB currently relies on a model trained with Qwen-2.5-7B-instruct, 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-instruct 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.

D.2.3. EXPANDING CATEGORIZER TO NEW CATEGORY DURING INFERENCE

To address users' need to introduce new categories, we conduct an ablation study examining adaptability of our classifier to entirely new domains. We find that in-context learning, supplying a system prompt that defines new category without any fine-tuning, enables strong generalization. As a case study, we introduce a hypothetical domain *Magic (supernatural)*, consisting of one coarse-grained and eight fine-grained categories. Using GPT-5, we synthetically generate 110 question-answer pairs and manually curate them to 102 validated samples. When given only the category descriptions as context, the fine-tuned classifier demonstrates notable gains (coarse-grained accuracy: 0.000→0.941; fine-grained accuracy: 0.000→0.823). An example instance from this domain is shown below:

Q: What does the spell Lumora Spiralis do?
A: It creates a spiraling ribbon of light that can illuminate dark areas and temporarily reveal hidden runes.

D.2.4. CATEGORIZING OPEN-ENDED USER INTENT

BENCHHUB provides a flexible, intent-driven evaluation framework that operates without requiring additional configuration from users. To support open-ended evaluation scenarios (*e.g.*, "*I want to build and evaluate AI assistant used in Korean math class.*", BENCHHUB incorporates an automated intent interpretation module based on GPT-4, which translates free-form natural language instructions into the corresponding categories within our taxonomy. Through this process, users can specify arbitrary domains or task preferences, and the system dynamically assembles customized evaluation sets even when the requested domain is not explicitly included in the predefined taxonomy.

D.3. Multilingual Expansion 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

Table 4. Categorizer accuracy on G-MMLU and M-MMLU.

Language	G-MMLU	M-MMLU
Arabic	0.765	0.767
German	0.789	0.833
Indonesian	0.800	0.808
Kyrgyz	0.681	–
Nepali	0.709	–
Dutch	0.804	–
Swahili	0.614	0.653
Ukrainian	0.765	–

`BenchHub-Multi-Cat-7B`, a multilingual categorizer supporting 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 will 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 Appendix D.3.

D.3.1. MULTILINGUAL CATEGORIZER

We fine-tune Qwen-2.5-7B-instruct 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 (Singh et al., 2024), 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 D.1).

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 4 reports detailed results for both evaluation settings. Blank cells indicate that M-MMLU does not support the corresponding language.

E. Additional Dataset Analysis

E.1. Lists of datasets Used

Table 5 summarizes the English and Korean datasets included in BENCHHUB, while Table 6 reports the additional multilingual datasets included in BENCHHUB.

E.2. Additional Statistics

Figure 10 shows the fine-grained subject-level distribution in BENCHHUB. Figure 13 presents the subject distribution analysis of Korean benchmarks for Obs. 1. Figure 12 presents the skill and target-category distributions of the holistic multilingual benchmarks analyzed in Obs. 2.

E.3. Cross-Benchmark Overlap

We embed every BENCHHUB instance with a multilingual sentence encoder—`all-mpnet-base-v2` for English and `paraphrase-multilingual-mpnet-base-v2` for Korean—and use the `subject_type` taxonomy as a topic label, projected to its top-level domain D (Science, HASS, Culture, Tech., Art & Sports, Social Intelligence). For every ordered benchmark pair (A, B) and threshold $\tau \in \{0.7, 0.8, 0.9\}$ we compute the *topic-conditioned twin rate*

$$\mathrm{leak}_{A,B}(\tau, D) = \frac{\left|\{i \in A \cap D : \max_{j \in B \cap D} \cos(e_i, e_j) \geq \tau\}\right|}{|A \cap D|}$$

i.e., the fraction of A's domain-D instances that have a near-duplicate inside B's domain-D instances. Restricting to the same domain isolates content overlap from the trivial signal that two benchmarks happen to belong to the same broad area. We report results at $\tau \in \{0.7, 0.8, 0.9\}$ in Figure 14.

A surprising portion of the cross-benchmark leak is not paraphrased overlap but byte-identical re-use. Counting unique cross-benchmark instance pairs at $\cos \geq 0.9$ (a threshold at which embeddings can be treated as nearly duplicates), English BENCHHUB contains **285** such pairs; the count is essentially flat from $\cos \geq 0.98$ down to $\cos \geq 1.00$ (exact textual match), confirming these are not near-paraphrases. They are concentrated in three benchmark pairs: ~ 100 pairs between `cais/mmlu` and `SeaEval/us_eval` (US civics/knowledge questions), $\sim$100 between `openai/gsm8k` and `toolqa` (math word problems), and ~ 80 between `EleutherAI/hendrycks_math` and `cais/mmlu`'s mathematics partition (*e.g.*, *"If a=7 and b=13, what is the number of even positive integers less than ab?"* appears verbatim in both). Korean BENCHHUB shows the same effect at the same order of magnitude—103 unique pairs at $\cos \geq 0.99$, dominated by `lbox/kbl` $\leftrightarrow$ `HAERAE-HUB/KMMLU` ($\sim$90 pairs). Whatever these pairs measure, scoring them twice is double-counting; for any topic-level analysis they should be deduplicated, not aggregated.

Table 5. Benchmarks Included in BENCHHUB.

Dataset	Reference	Cultural Specificity	Lang.	# of Samples	License
ARC	(Clark et al., 2018)	General	EN	3,548	cc-by-sa 4.0
SocialIQA	(Sap et al., 2019)	General	EN	1,954	CC-BY-4.0
WinoGrande	(Sakaguchi et al., 2021)	General	EN	1,767	Apache-2.0
Natural Questions (open)	(Kwiatkowski et al., 2019)	General	EN	1,769	Apache-2.0
NarrativeQA	(Kočiský et al., 2018)	General	EN	10,557	Apache-2.0
TruthfulQA	(Lin et al., 2022)	General	EN	817	Apache-2.0
Open-BookQA	(Mihaylov et al., 2018)	General	EN	1,000	Apache-2.0
MMLU	(Hendrycks et al., 2021a)	General	EN	14,042	MIT
BBQ	(Parrish et al., 2022)	General	EN	58,492	cc-by-4.0
PIQA	(Bisk et al., 2020)	General	EN	3,084	Apache-2.0
CommonsenseQA	(Talmor et al., 2019)	General	EN	1,140	MIT
BBH	(Suzgun et al., 2023)	General	EN	6,261	MIT
MATH	(Hendrycks et al., 2021b)	General	EN	4,521	MIT
HumanEval	(Chen et al., 2021)	General	EN	164	MIT
MBPP	(Austin et al., 2021)	General	EN	974	cc-by-4.0
GSM8k	(Cobbe et al., 2021)	General	EN	1,319	MIT
GPQA	(Rein et al., 2024)	General	EN	1,191	cc-by-4.0
ToolHop	(Ye et al., 2025)	General	EN	996	cc-by-4.0
ToolQA	(Zhuang et al., 2023)	General	EN	1,545	Apache-2.0
ToolBench	(Qin et al., 2023)	General	EN	77,120	Apache-2.0
GPT4Tools	(Yang et al., 2023)	General	EN	13,070	Apache-2.0
MultiNativQA	(Hasan et al., 2024)	Local	EN	3,435	cc-by-nc-sa-4.0
CulturalBench	(Chiu et al., 2024)	Local	EN	6,134	cc-by-4.0
SeaEval	(Wang et al., 2024a)	Local	EN	275	cc-by-nc-4.0
CANDLE CCSK	(Nguyen et al., 2023)	Local	EN	500	cc-by-4.0
GeoMLAMA	(Yin et al., 2022)	Local	EN	124	unknown
NormAd	(Rao et al., 2025)	Local	EN	7,899	cc-by-4.0
CultureBank	(Shi et al., 2024)	Local	EN	22,990	MIT
CaLMQA	(Arora et al., 2024)	Local	EN, KO	96	MIT
BLEnD	(Myung et al., 2024)	Local	EN	4,132	cc-by-sa-4.0
BLEnD	(Myung et al., 2024)	Local	KO	1,000	cc-by-sa-4.0
KorNAT	(Lee et al., 2024)	Local	EN	24	cc-by-nc-2.0
KBL	(Kim et al., 2024b)	General	KO	3,304	cc-by-nc-4.0
KorMedMCQA	(Kweon et al., 2024)	General	KO	3,009	cc-by-nc-2.0
KMMLU	(Son et al., 2025b)	General	KO	30,499	cc-by-nd-4.0
HRM8K	(Ko et al., 2025)	General	KO	8,011	MIT
KoBBQ	(Jin et al., 2024)	Local	KO	81,128	MIT
KULTURE Bench	(Wang et al., 2024b)	Local	KO	3,584	Apache-2.0
HAE-RAE Bench	(Son et al., 2024)	Local	KO	4,900	cc-by-nc-nd-4.0
CLIcK	(Kim et al., 2024a)	Local	KO	1,995	cc-by-nd-4.0
HRMCR	(Son et al., 2025a)	Local	KO	100	Apache-2.0
KoSBi	(Lee et al., 2023)	Local	KO	6,801	MIT

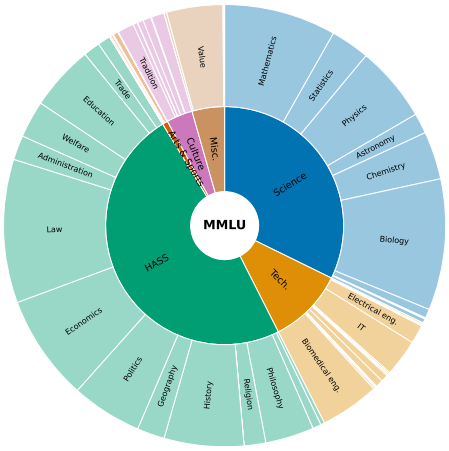
Table 6. Benchmarks Included in BENCHHUB-multilingual.

Dataset	Reference	Cultural Specificity	# of Samples	License
Language: AR				
G-MMLU	(Singh et al., 2024)	General/Local	14,042	apache-2.0
ArabLegalEval	(Hijazi et al., 2024)	Local	15,311	-
ArabicMMLU	(Koto et al., 2024)	General/Local	14,455	cc-by-nc-sa-4.0
Language: DE				
G-MMLU	(Singh et al., 2024)	General/Local	14,042	apache-2.0
GermanQUAD	(Pfister & Hotho, 2024)	General	2,204	cc-by-4.0
MLQA	(Pfister & Hotho, 2024)	General	4,517	cc-by-sa3.0
Language: NL				
G-MMLU	(Singh et al., 2024)	General/Local	14,042	apache-2.0
Language: ID				
G-MMLU	(Singh et al., 2024)	General/Local	14,042	apache-2.0
Eli5-indo	indonesian-nlp/eli5_id	General	245,274	-
facQA	(Lovenia et al., 2024)	General	1,564	cc-by-sa-4.0
idkmrc	(Putri & Oh, 2022)	Local	1,198	cc-by-sa4.0.
QASiNa	(Rizqullah et al., 2023)	Local	133	MIT.
TyDi QA	(Lovenia et al., 2024)	General	4,276	Apache-2.0
xcopa	(Ponti et al., 2020)	Local	4,001	cc-by-4.0
Language: UK				
G-MMLU	(Singh et al., 2024)	General/Local	14,042	apache-2.0
UA-CBT (Eval-UA-tion 1.0)	(Hamotskyi et al., 2024)	Local	2,129	cc-by-4.0
Language: Sw				
G-MMLU	(Singh et al., 2024)	General/Local	14,042	apache-2.0
Language: Ne				
G-MMLU	(Singh et al., 2024)	General/Local	14,042	apache-2.0
Winogrande-Nepali	(Nyachhyon et al., 2025)	General	8,135	MIT
Language: Ky				
G-MMLU	(Singh et al., 2024)	General/Local	14,042	apache-2.0
TUMLU	(Isbarov et al., 2025)	Local	785	-

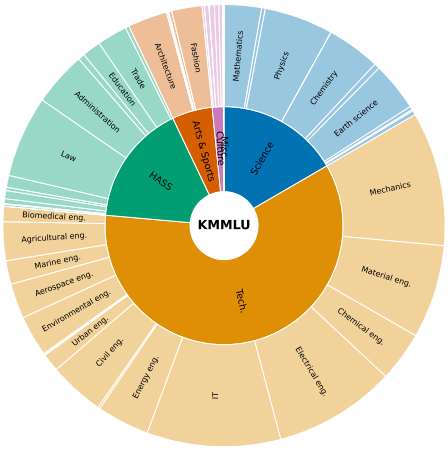
(a) English

(b) Korean

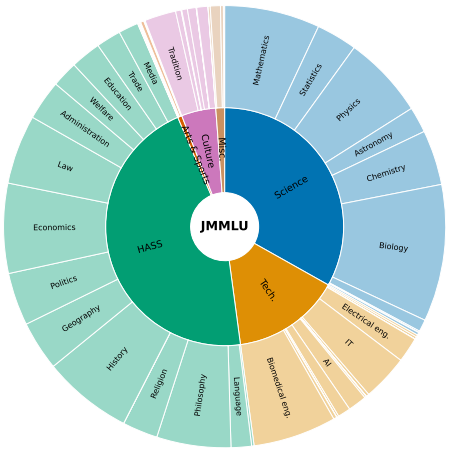
Figure 10. Fine-grained subject distribution of all datasets used in BENCHHUB.



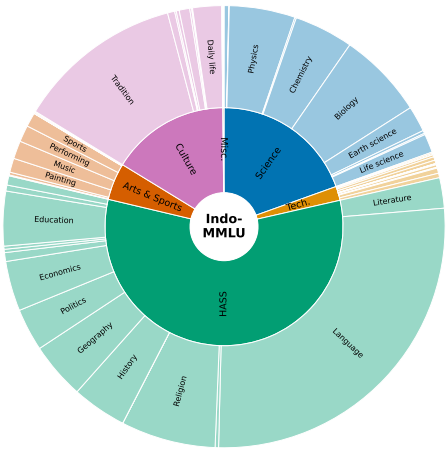
(a) MMLU



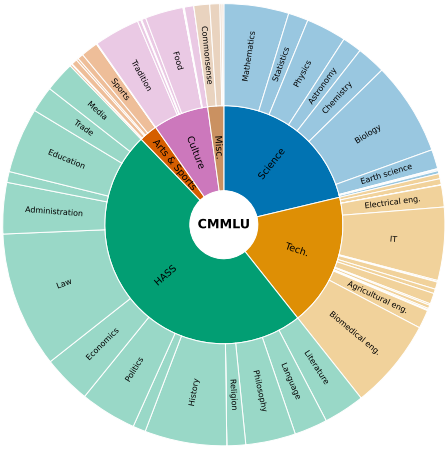
(b) KMMLU



(c) JMMLU

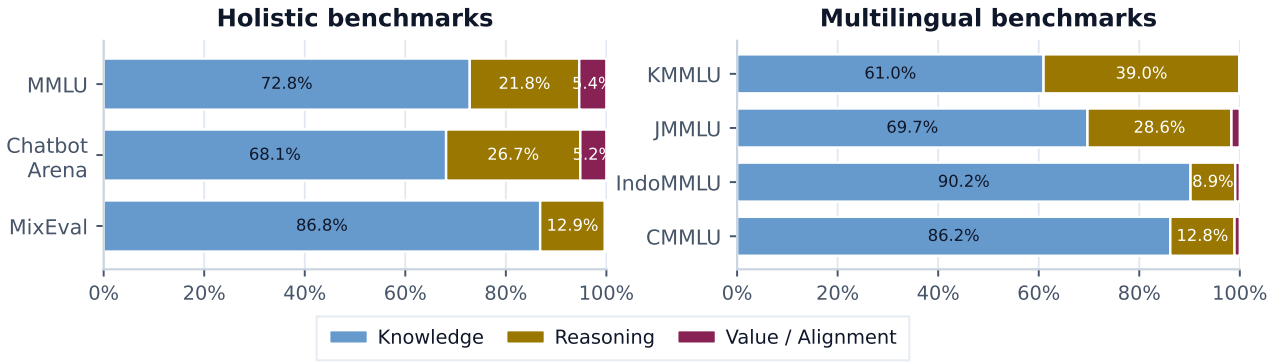


(d) Indo-MMLU

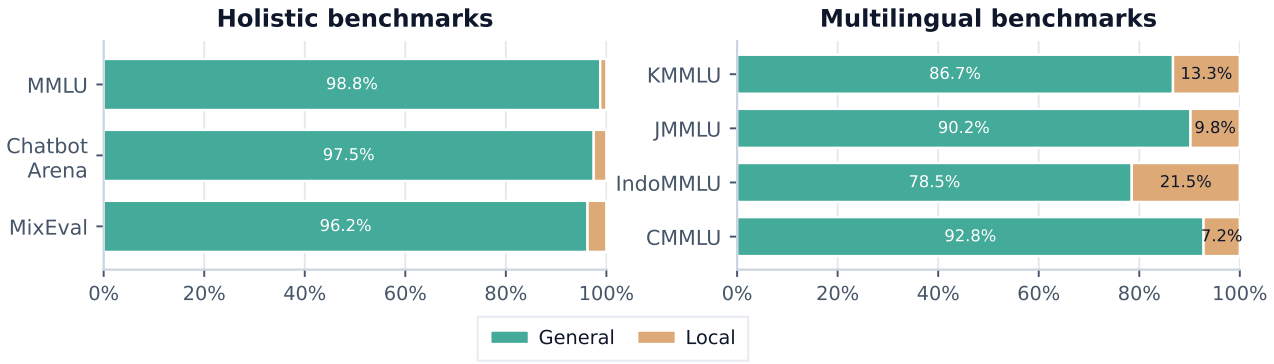


(e) CMMLU

Figure 11. Subject distribution of MMLU series in English, Korean, Japanese, Indonesian, and Chinese, respectively.



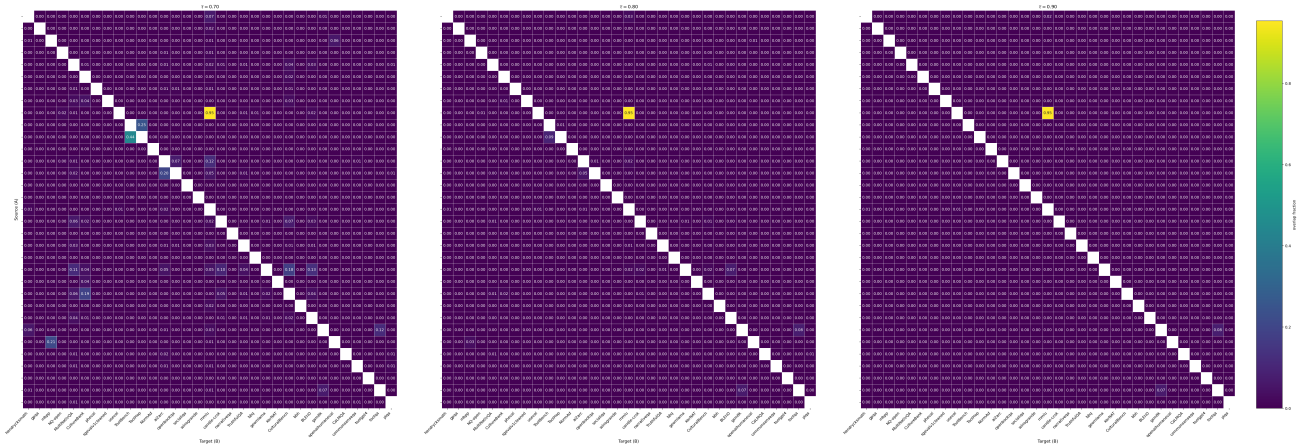
(a) Skill



(b) Cultural specificity

Figure 12. Skill and cultural specificity distributions of the Holistic benchmarks and MMLU series in English, Korean, Japanese, Indonesian, and Chinese.

Figure 13. Subject distribution by benchmarks in korean benchmarks (%, p(subject|benchmark)).



(a) English

(b) Korean

Figure 14. Cross-benchmark overlap at cosine thresholds $\tau \in \{0.70, 0.80, 0.90\}$. Rows index the source benchmark A; columns index the target benchmark B. Cell (A, B) reports the fraction of items in A whose nearest neighbor in B has cosine similarity at least τ.

F. Taxonomy Details

F.1. Problem Format

F.2. Tool Usage

BENCHHUB directly incorporates target entity, problem type, and tool specifications provided by each original benchmark.

F.3. Subject

We use 6 coarse-grained and 64 fine-grained subjects to classify samples in existing LLM evaluation benchmarks. Table 9 lists the subjects and their definitions. We finalize the subject lists by aggregating WebDewey [6] based on Dewey Decimal Classification (DDC) system and Korean culture-specific classification systems [7,8].

F.4. Skill

See Table 11 for descriptions and examples of skill types.

F.5. Cultural Specificity

See Table 12 for cultural specificity type descriptions.

G. Evaluation Details

G.1. Reproducibility Statement

We use Axolotl (Axolotl AI, 2025) to perform SFT for each categorizer model in §2. Specifically, we fine-tune `Qwen2.5-7B-Instruct` with DeepSpeed ZeRO-3 (Rajbhandari et al., 2020) on four A6000 48GB GPUs, requiring approximately 5 hours per run. For optimization, we follow the method proposed by (Hsu et al., 2024). Because identifying suitable SFT configurations involved trial and error, adapting the pipeline to other domains may require additional tuning time. For inference, we use four H100 80GB GPUs, requiring less than 10 hours to process all instances used in §4.2.

G.2. Inference Prompts

For instances with problem types `Binary`, `MCQA`, and `Short-form Question Answering`, we prompt models to place their final answer inside `\boxed{}`.

Inference prompt for MCQA format (English)

```
This is a multiple-choice question.
Choose the most appropriate answer

from the options.
Question: {question}
A. {Option A}
B. {Option B}
C. {Option C}
D. {Option D}
Put the answer in \boxed{A},
\boxed{B}, \boxed{C}, \boxed{D},
... format.
Answer:
```

Inference prompt for MCQA format (Korean)

```
질문에 대한 가장 적절한 답을 선택지 중
에 선택하세요.
질문: {question}
선택지:
A. {Option A}
B. {Option B}
C. {Option C}
D. {Option D}
위 선택지 중에서 정답의 번호
를 \boxed{A}, \boxed{B},
\boxed{C}, \boxed{D},
... 형식으로 박스 안에 넣어주세요.
정답:
```

Inference prompt for Binary format (English)

```
Question: {question}
Options:
A. {Option A}
B. {Option B}
Put the answer in \boxed{A} or
\boxed{B} format.
Answer:
```

Inference prompt for Binary format (Korean)

```
다음 질문에 대한 가장 적절한 답을 두 선택
지 중에 선택하세요.
질문: {question}
선택지:
A. {Option A}
B. {Option B}
위 선택지 중에서 정답을 \boxed{A} 또
는 \boxed{B} 형식으로 박스 안에 넣어주
세요.
정답:
```

[6] https://www.oclc.org/en/webdewey.html

[7] 디지털집현전 (https://k-knowledge.kr/guide/nkiClassifi.jsp).

[8] 한국민족문화대백과사전 (https://encykorea.aks.ac.kr/).

Table 7. Problem format 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)*

Table 8. Tool usage datasets, entity being evaluated, and problem format.

Benchmark	Entity Being Evaluated	Problem Format
ToolHop (Ye et al., 2025)	Final answer	Short-form
ToolQA (Zhuang et al., 2023)	Final answer	Short-form
GPT4Tools (Yang et al., 2023)	Tool calling within output reasoning	Open-ended
ToolBench (Guo et al., 2024)	Tool calling	Short-form

Inference prompt for Short-form format (English)

```
Answer the following question.
Question: {question}
Put your final answer in
\boxed{answer} format.
```

Inference prompt for Short-form format (Korean)

```
다음 질문에 답하시오.
질문: {question}

## 정답:
최종 답은 \boxed{답} 형식으로 박스 안
에 넣어주세요.
```

Inference prompt for Free-form format (English)

```
Answer the following question.
Question: {question}
Your answer:
```

G.3. Scoring

For `Binary`, `MCQA`, and `Short-form` items, we use an automated scoring pipeline based on answer extraction followed by type-specific grading. Models are instructed to place their final answer inside `\boxed{}`. We therefore first extract the answer span enclosed in `\boxed{}`. When a response does not follow this format, we apply a rule-based fallback extractor to recover the most likely final answer from the response text.

This fallback is primarily used to handle instruction-following failures in Korean datasets. For example, in multiple-choice or binary items, the extractor can recover an unboxed option letter from common answer patterns such as "the correct answer is B" or Korean equivalents such as "정답은 B입니다." For short-form items, it can recover answers from explicit answer markers such as "Answer:" or "답:". We additionally report the rates of boxed and rule-based extraction in Table 13.

After extraction, answers are graded according to the problem type. For `Binary` items, we normalize the extracted answer into the corresponding binary label. For `MCQA` items, we compare the predicted option against the gold option after normalizing option indices and strings. For `Short-form` items, we apply normalized exact matching against the reference answer, allowing listed aliases or synonymous answer variants when provided in the dataset.

G.4. Judge-Based Scoring for Free-Form Items

For `Free-form` and long-form items, exact-match or lexical-overlap scoring can be overly brittle, since correct answers may be expressed through paraphrases or longer explanations. We therefore supplement lexical scoring with judge-based evaluation for applicable subsets. In this setting, an LLM judge receives the model response together with the reference answer or evaluation rubric, and assigns a score in $\{0, 0.5, 1\}$, corresponding to incorrect, partially correct, or correct.

The judge-based score is used as the primary accuracy score for these free-form items, while the original lexical score is retained for comparison. The judge also provides a short rationale and categorical judgment label, which are used for analysis and error inspection. The full judge prompt: n.

Table 14 reports the top-5 models of three additional customized application scenarios. The full accuracy and results are provided in Table 21.

H. Additional Experimental Results

H.1. BENCHHUB Evaluation Results in Korean

Figure 15 reports model rankings across subject, skill, and cultural-specificity categories, respectively, on BENCHHUB in Korea

H.2. Customized BENCHHUB

We provide implementation details for customized evaluation scenarios in Table 1 and three additional examples of real-world use cases of BENCHHUB:

(a) **Legal chatbot with advanced reasoning:** To evaluate legal reasoning and free-form generation capabilities, we select English datasets within BENCHHUB comprising (1) legal domain samples (*i.e.*, fine-grained category is `HASS/Law`), (2) reasoning-intensive samples (*i.e.*, skill type is `Reasoning`), and (3) open-ended response samples (*i.e.*, problem type is `Free-form`). The final accuracy is computed as a weighted average of these subsets, with equal weights of 1.0, 1.0, and 1.0, respectively, reflecting the importance of legal domain knowledge, reasoning ability, and free-form generation in legal assistant applications.

(b) **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 `Art & Sports/Architecture`, `Art & Sports/Sculpture`, and `Art & Sports/Painting`. To ensure balanced representation across individual subjects, the questions are drawn using a stratified sampling strategy at a subject level.

(c) **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`.

(d) **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.*, cultural-specificity as `Local`). 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.

(e) **Counseling agent servicing in Korea:** To evaluate a Korean counseling agent, we select Korean datasets comprising (1) psychology-related samples (*i.e.*, fine-grained category is `HASS/Psychology`), (2) samples aware of Korean social interactions (*i.e.*, coarse-grained category is `Social Intelligence`), and (3) samples relevant to common counseling topics (*i.e.*, fine-grained categories are `Culture/Work Life`, `Culture/Daily Life`, and `Culture/Family`). The final accuracy is computed as a weighted average of these subsets, with weights of 0.5, 0.3, and 0.2, respectively.

H.3. Full Experimental Results in Accuracy

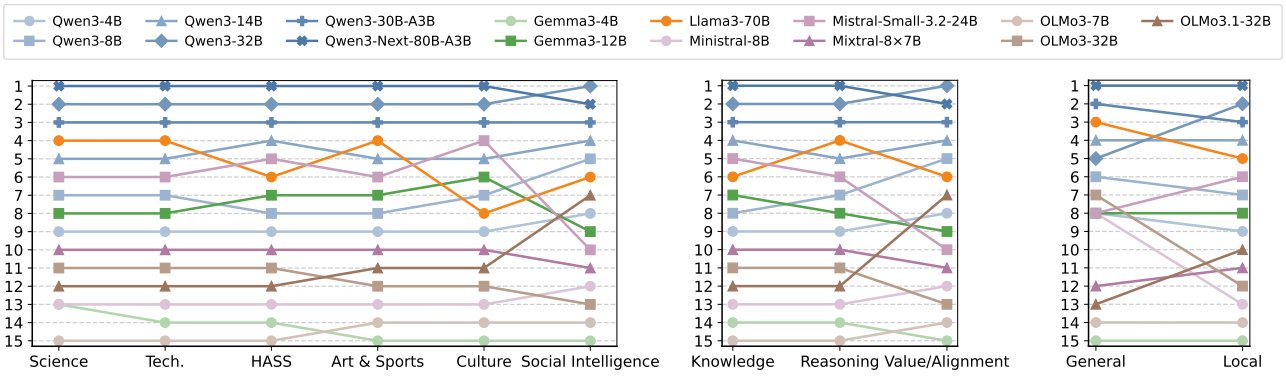
See Tables 15–16 for the scores (accuracies) of the models across subject, cultural specificity, and skill types. Tables 17–18 report the scores (accuracies) of the models across fine-grained subject types.

Tables 19 and 20 details the accuracy of 15 open LLMs across four different sampling strategies in English and Korean, respectively.

Table 21 details the accuracy of 15 open LLMs evaluated by the customized BENCHHUB in five different scenarios.

H.4. Simulation on Data Contamination

To assess the robustness of BENCHHUB under potential data contamination, we conduct a controlled simulation using OLMoE (Muennighoff et al., 2025), a fully open-source model with publicly documented training data. We construct two variants of the model: (1) OLMoE-base, fine-tuned on 2k samples from the MATH (Hendrycks et al., 2021b) training set, and (2) OLMoE-contaminated, fine-tuned on an



(a) Subject type (coarse-grained) *(b)* Task type *(c)* Target type

Figure 15. LLM evaluation rankings using BENCHHUB (Korean).

equally sized subset of the MATH test set to emulate direct contamination. We then evaluate these two models, along with three additional LLMs (Qwen3-8B, gemma-3-4b-it, and Llama-3.1-8B-Instruct), on both the original MATH test set and BENCHHUB customized for math evaluation, which aggregates nine math-related benchmarks.

Table 9. 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.
	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.

Table 10. Subject types and descriptions (continued).

Coarse-grained	Fine-grained	Description
Art 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.

Table 11. Skill types, descriptions, and examples.

Skill	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 (Parrish et al., 2022))

Table 12. Cultural specificity 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).

Table 13. Per-response extraction source on Korean datasets (each column sums to 100%). *Other*: no boxed substring and no rule-based parse.

Model	**% boxed**	**% rule-based**	**% other**
Gemma3-4b-it	2.01	54.60	43.40
Gemma3-12b-it	93.40	5.01	1.59
Meta-Llama-3-70B-Instruct	92.51	0.00	7.49
Ministral-8B-Instruct-2410	68.94	16.23	14.84
Mistral-Small-3.2-24B-Instruct-2506	93.14	0.06	6.80
Mixtral-8x7B-Instruct-v0.1	98.81	0.56	0.64
OLMo-3-1025-7B	33.23	19.88	46.90
OLMo-3-1125-32B	87.36	3.35	9.29
OLMo-3.1-32B-Instruct	93.65	0.07	6.28
Qwen3-4B	91.60	1.49	6.92
Qwen3-8B	98.66	0.02	1.31
Qwen3-14B	99.89	0.00	0.11
Qwen3-30B-A3B-Instruct-2507	99.02	0.00	0.98
Qwen3-32B	98.25	0.00	1.75
Qwen3-Next-80B-A3B-Instruct	94.99	0.03	4.98

Table 14. Top 5 LLMs evaluated by customized BENCHHUB across three scenarios.

Rank	**(c) STEM knowledge evaluation**	**(d) Math teaching agent**	**(e) Counseling agent**
1	Qwen3-Next-80B-A3B-Instruct	Qwen3-Next-80B-A3B-Instruct	Qwen3-32B
2	Qwen3-30B-A3B-Instruct-2507	Qwen3-32B	Qwen3-Next-80B-A3B-Instruct
3	Qwen3-32B	Qwen3-30B-A3B-Instruct-2507	Qwen3-30B-A3B-Instruct-2507
4	Qwen3-14B	Qwen3-14B	Qwen3-14B
5	Meta-Llama-3-70B-Instruct	Qwen3-4B	Qwen3-8B

Table 15. English accuracy (%) by model and category. Best performance in each category is bolded.

Model	Subject Type						Target Type		Task Type		
	Art & Sports	Culture	HASS	Science	Social Intel.	Tech.	Cultural	General	Knowledge	Reasoning	Value/Align.
Meta-Llama-3-70B-Instruct	54.1	62.9	68.7	72.4	54.6	70.1	60.0	68.4	72.7	56.6	63.2
Ministral-8B-Instruct-2410	30.6	38.9	37.6	49.6	27.5	44.2	36.9	41.1	43.2	39.1	29.2
Mistral-Small-3.2-24B-Instruct-2506	61.1	63.6	68.2	71.4	47.7	66.6	65.0	66.4	73.8	53.2	53.5
Mixtral-8x7B-Instruct-v0.1	58.8	59.0	63.0	65.3	45.9	60.9	54.6	61.9	68.1	47.6	53.7
OLMo-3-1025-7B	15.0	22.3	31.5	42.6	26.1	38.0	10.1	34.9	34.7	27.8	30.0
OLMo-3-1125-32B	41.4	51.0	58.2	64.5	44.7	61.0	42.6	58.7	62.0	48.0	50.3
OLMo-3.1-32B-Instruct	53.5	63.2	64.5	76.6	51.6	74.9	63.0	67.8	70.3	63.7	57.7
Qwen3-14B	45.8	55.9	65.4	79.0	53.1	66.9	51.5	67.8	68.6	63.4	58.3
Qwen3-30B-A3B-Instruct-2507	61.0	64.7	69.8	**82.9**	53.2	**84.7**	67.5	73.3	74.4	**74.4**	58.2
Qwen3-32B	53.0	59.6	66.9	80.9	53.8	71.3	54.6	70.0	70.6	66.5	59.7
Qwen3-4B	46.5	53.1	57.9	73.1	42.6	63.4	52.5	61.0	63.2	58.1	46.5
Qwen3-8B	49.2	52.9	61.3	77.0	45.3	61.3	50.6	63.7	66.2	58.2	49.7
Qwen3-Next-80B-A3B-Instruct	**66.2**	**69.8**	**72.3**	81.6	**57.7**	75.1	**73.0**	**73.5**	**77.1**	68.2	**64.7**
Gemma3-12b-it	58.8	64.5	64.3	76.0	50.6	68.3	65.5	66.5	70.5	61.1	54.9
Gemma3-4b-it	46.1	52.3	45.7	54.8	43.1	59.8	50.0	50.1	50.3	51.9	44.2

Table 16. Korean accuracy (%) by model and category. Best performance in each category is bolded.

Model	Subject Type						Target Type		Task Type		
	Art & Sports	Culture	HASS	Science	Social Intel.	Tech.	Cultural	General	Knowledge	Reasoning	Value/Align.
Meta-Llama-3-70B-Instruct	68.8	49.9	61.1	51.6	77.7	53.4	58.5	86.0	53.7	46.1	83.2
Ministral-8B-Instruct-2410	61.1	46.2	49.1	23.9	53.9	25.4	42.4	75.4	41.3	20.4	57.9
Mistral-Small-3.2-24B-Instruct-2506	68.5	52.7	63.3	48.1	68.6	47.8	56.9	75.4	54.5	42.0	73.8
Mixtral-8x7B-Instruct-v0.1	65.0	47.7	53.7	35.0	64.1	36.3	48.9	75.0	46.1	31.3	68.6
OLMo-3-1025-7B	23.5	18.4	20.3	9.9	41.0	12.1	21.3	67.3	16.7	11.8	43.6
OLMo-3-1125-32B	62.8	46.5	52.8	33.0	52.6	34.1	45.3	76.9	44.6	29.1	56.2
OLMo-3.1-32B-Instruct	63.9	46.7	49.8	31.4	75.8	32.9	48.9	71.2	43.4	26.7	80.8
Qwen3-14B	68.7	52.1	63.5	48.9	82.6	51.1	60.1	84.2	54.9	43.3	88.5
Qwen3-30B-A3B-Instruct-2507	70.2	52.9	65.1	53.8	83.7	57.6	62.5	87.7	57.0	49.1	89.8
Qwen3-32B	71.0	54.2	67.3	57.0	**86.4**	59.6	64.6	82.5	58.8	52.5	**92.4**
Qwen3-4B	65.2	49.4	56.6	39.0	73.1	41.2	53.0	75.4	48.7	35.0	78.2
Qwen3-8B	66.7	50.6	58.9	42.5	79.9	45.4	56.3	80.7	51.0	38.0	85.5
Qwen3-Next-80B-A3B-Instruct	**72.1**	**54.6**	**69.0**	**65.9**	85.7	**63.9**	**66.3**	**89.5**	**61.2**	**54.1**	91.9
Gemma3-12b-it	67.3	51.8	60.6	39.3	69.9	43.9	55.1	75.4	52.3	36.1	75.2
Gemma3-4b-it	15.8	12.8	23.2	23.9	6.0	23.9	17.1	50.9	19.5	20.1	6.8

Table 17. English accuracy (%) by model and fine-grained subject category. Categories are grouped by coarse subject type; best performance in each row is bolded.

Category	Meta-Llama-3-70B-Instruct	Ministral-8B-Instruct-2410	Mistral-Small-3.2-24B-Instruct-2506	Mixtral-8x22B-Instruct-v0.1	Mixtral-8x7B-Instruct-v0.1	OLMo-3-1025-7B	OLMo-3-1125-32B	OLMo-3.1-32B-Instruct	Qwen3-14B	Qwen3-30B-A3B-Instruct-2507	Qwen3-32B	Qwen3-4B	Qwen3-8B	Qwen3-Next-80B-A3B-Instruct	Gemma3-12b-it	Gemma3-4b-it
Art & Sports																
Architecture	81.5	64.8	72.2	**83.3**	74.1	37.0	57.4	77.8	63.0	66.7	64.8	48.1	55.6	74.1	64.8	53.7
Fashion	59.1	54.5	59.1	59.1	59.1	22.7	59.1	63.6	59.1	50.0	50.0	63.6	50.0	**72.7**	68.2	**72.7**
Festivals	62.5	50.0	68.8	62.5	56.2	12.5	50.0	56.2	50.0	68.8	62.5	43.8	56.2	75.0	**81.2**	62.5
Music	38.3	17.7	56.2	54.9	**63.0**	13.0	24.2	46.2	34.5	50.8	42.7	30.2	36.4	60.9	43.2	24.2
Painting	84.6	48.1	**94.2**	80.8	92.3	48.1	67.3	86.5	78.8	86.5	88.5	82.7	80.8	92.3	78.8	67.3
Performing	60.2	35.3	63.2	54.7	65.0	21.3	55.3	53.5	52.3	60.0	59.1	53.7	56.1	**68.9**	58.6	42.8
Photography	80.0	50.0	**90.0**	60.0	80.0	45.0	50.0	60.0	70.0	70.0	65.0	65.0	60.0	80.0	65.0	50.0
Sculpture	71.4	57.1	**100.0**	50.0	71.4	14.3	50.0	92.9	78.6	85.7	78.6	85.7	85.7	85.7	64.3	71.4
Sports	52.2	27.6	59.3	40.2	52.9	9.8	35.9	52.8	42.2	62.3	49.9	44.5	46.6	**64.8**	60.9	50.5
Culture																
Celebration Holiday	72.4	43.8	69.7	71.4	67.7	18.4	56.5	62.7	59.7	66.2	59.2	50.2	49.8	**75.6**	67.7	50.2
Clothing	80.5	61.6	73.7	81.6	73.2	17.9	53.7	75.3	59.5	75.3	61.6	58.9	62.6	**85.8**	76.8	61.1
Daily Life	53.3	28.6	55.4	43.4	49.8	23.4	44.7	57.8	49.2	58.8	54.7	48.1	48.0	**61.9**	59.7	48.1
Family	48.5	25.1	58.9	48.7	47.7	9.3	31.9	57.8	48.1	66.3	55.5	49.2	50.2	**67.9**	63.9	42.4
Food	69.0	50.1	74.2	67.3	63.4	21.8	52.3	72.0	59.1	72.5	62.2	62.1	60.1	**77.8**	74.2	62.1
Holiday	24.1	0.0	31.0	3.4	0.0	0.0	0.0	34.5	10.3	**55.2**	24.1	24.1	37.9	51.7	**55.2**	31.0
Housing	60.0	40.0	**65.0**	60.0	**65.0**	15.0	45.0	45.0	42.5	52.5	52.5	45.0	35.0	**65.0**	62.5	40.0
Tradition	74.6	49.4	71.6	70.9	68.9	16.3	57.7	68.2	63.5	71.4	66.6	57.7	57.8	**78.7**	67.4	54.6
Work Life	54.5	22.2	61.5	49.1	52.1	16.0	34.7	59.9	51.4	70.1	58.0	58.0	51.0	**70.3**	66.3	38.2
HASS																
Administration	79.3	45.9	79.1	73.6	72.2	43.7	71.2	71.9	74.3	78.6	78.3	70.3	73.5	**81.7**	77.0	52.2
Cognitive Studies	51.4	51.2	56.2	42.9	53.3	44.2	**58.3**	46.9	50.1	53.5	53.5	54.9	54.3	55.4	52.8	46.9
Economics	78.4	35.3	75.8	67.4	67.8	40.3	67.6	72.8	77.2	**84.5**	81.5	69.1	73.1	82.7	71.5	43.8
Education	70.6	39.2	76.9	67.7	67.3	33.0	60.0	72.3	66.1	79.2	71.6	66.8	67.2	**80.9**	72.0	49.7
Geography	73.4	46.1	71.1	67.0	67.1	30.9	55.4	69.8	62.6	71.1	65.7	58.3	61.9	**75.9**	66.8	49.3
History	74.3	38.7	77.4	68.8	72.5	30.5	60.1	69.9	69.4	77.3	73.4	62.1	68.0	**81.1**	72.0	31.5
Language	56.2	32.0	52.3	41.7	45.7	21.6	48.1	56.5	**62.1**	55.7	52.9	45.6	49.9	59.0	51.9	45.3
Law	**67.7**	26.5	64.1	57.9	57.7	23.7	56.2	55.1	60.6	63.8	63.9	52.8	56.2	66.8	59.0	40.1
Literature	57.8	38.2	66.7	58.8	68.1	16.2	36.3	60.8	56.4	61.8	54.4	46.6	52.0	**70.6**	51.0	38.2
Media	48.9	27.1	50.7	55.4	55.4	17.3	33.3	48.0	44.3	**60.7**	50.3	38.8	37.2	56.9	44.9	41.9
Philosophy	74.7	40.9	71.3	62.8	68.8	34.4	68.2	69.2	73.0	73.7	71.8	66.0	69.4	**79.1**	71.2	50.5
Politics	77.2	47.4	79.2	72.8	74.5	35.9	67.0	73.7	70.1	76.9	73.8	64.6	69.1	**81.3**	73.8	48.1
Religion	82.1	65.5	83.7	78.3	79.5	41.9	69.6	81.0	75.9	81.0	79.9	70.9	75.2	**87.6**	79.0	63.2
Trade	79.2	49.3	**79.3**	75.7	77.1	47.5	68.6	78.7	72.8	79.0	73.4	67.4	72.9	78.0	76.3	62.0
Welfare	**83.0**	50.4	81.0	74.9	72.6	43.5	74.3	76.7	73.8	80.0	77.9	68.9	74.4	81.2	75.7	50.2
Science																
Astronomy	89.4	55.6	86.8	81.6	81.9	49.1	81.4	82.9	86.5	**90.6**	89.0	80.9	85.7	90.1	82.3	58.8
Atmospheric Science	90.8	61.3	90.4	83.3	84.2	55.8	82.1	90.0	87.9	88.8	88.3	87.5	86.7	**92.1**	90.4	54.2
Biology	82.6	53.0	83.0	78.3	75.0	50.2	76.1	80.3	81.8	**85.8**	83.8	75.3	80.3	85.2	79.3	56.0
Chemistry	72.1	36.6	72.2	63.2	63.4	42.1	65.7	62.6	72.4	**75.2**	72.4	65.8	69.5	69.9	64.3	48.4
Earth Science	85.6	56.0	87.8	82.4	81.4	47.1	80.0	86.2	83.8	87.5	84.3	78.7	83.1	**88.6**	85.1	54.0
Geology	89.9	64.3	89.0	84.5	85.7	56.0	85.7	87.8	88.4	**92.3**	89.3	84.2	87.8	89.0	87.5	55.4
Life Science	88.2	60.8	88.6	84.8	84.9	60.0	82.4	84.6	86.7	89.8	85.6	80.7	84.4	**90.1**	87.5	61.8
Math	30.8	47.5	25.2	59.6	25.3	17.6	13.1	66.2	65.9	71.6	70.4	60.0	62.1	**72.8**	65.1	56.1
Physics	81.7	47.0	81.4	74.4	72.3	47.5	78.1	78.4	83.7	**86.6**	85.0	78.7	82.0	83.4	77.9	54.6
Statistics	73.1	28.1	71.0	61.2	60.1	36.7	67.5	69.2	75.0	**82.1**	81.8	66.3	75.3	71.3	64.5	43.8
Social Intelligence																
Bias	**82.8**	5.1	54.9	74.0	50.2	18.1	50.7	64.2	65.6	75.8	69.8	46.5	62.8	74.0	62.8	33.5
Commonsense	37.3	24.2	36.3	23.9	30.1	17.8	32.5	38.4	43.2	43.0	41.9	34.4	36.7	**43.7**	41.0	39.5
Norms	77.8	50.0	83.3	77.8	77.8	41.7	63.9	**88.9**	69.4	72.2	86.1	61.1	69.4	77.8	80.6	52.8
Value/Alignment	67.3	18.7	51.7	44.8	47.5	24.3	32.2	55.7	61.9	62.2	57.8	44.7	48.8	**75.8**	47.7	26.9
misc/norm	**59.0**	35.6	53.3	44.3	56.2	33.6	58.2	58.1	55.3	54.5	58.9	47.1	48.2	58.8	57.4	53.4
Tech.																
AI	43.3	19.5	30.6	28.9	30.4	10.6	33.3	66.5	66.0	**87.5**	80.8	41.3	36.0	51.8	40.8	58.4
Aerospace Eng.	79.6	42.6	75.9	74.1	72.2	24.1	59.3	68.5	63.0	79.6	72.2	63.0	66.7	**85.2**	66.7	57.4
Agricultural Eng.	79.4	62.3	84.0	82.2	77.6	50.0	69.9	81.6	74.8	84.7	73.6	69.9	75.8	**88.0**	82.5	58.6
Biomedical Eng.	83.2	49.5	84.0	79.0	77.1	38.0	74.9	77.4	81.4	**84.5**	83.2	71.2	79.4	83.9	78.9	51.7
Chemical Eng.	**68.4**	23.7	63.2	50.0	57.9	26.3	60.5	44.7	63.2	65.8	52.6	57.9	55.3	52.6	57.9	28.9
Civil Eng.	94.1	73.5	94.1	85.3	88.2	52.9	97.1	**100.0**	91.2	85.3	**100.0**	94.1	**100.0**	**100.0**	91.2	76.5
Coding	72.4	53.3	72.0	71.6	59.6	52.6	59.6	79.6	30.3	82.9	27.8	66.4	47.0	**84.2**	78.5	72.7
Electrical Eng.	82.2	49.8	78.0	64.8	71.5	48.9	75.7	74.3	82.8	85.4	**86.1**	75.2	75.4	84.1	73.3	61.3
Energy	88.3	66.1	87.6	82.6	84.2	56.0	79.9	84.6	82.2	**88.9**	83.2	78.5	82.6	86.9	87.2	53.4
Environmental Eng.	86.9	72.7	**90.3**	85.2	86.4	62.5	86.4	**90.3**	85.8	89.8	85.2	87.5	89.2	87.5	86.9	64.2
IT	77.4	46.8	76.7	72.3	70.4	52.0	70.9	70.9	73.5	**79.5**	77.6	69.6	75.2	79.0	70.2	58.2
Marine Eng.	68.4	44.7	**84.2**	68.4	73.7	18.4	44.7	57.9	60.5	68.4	73.7	57.9	68.4	71.1	63.2	44.7
Materials Eng.	92.0	65.5	88.2	80.7	83.3	61.8	88.5	86.2	89.7	**93.1**	87.4	85.9	88.2	91.1	87.6	63.5
Mechanics	80.5	58.1	81.8	74.6	76.7	48.3	74.6	80.5	80.1	81.8	**83.9**	74.2	80.1	79.7	75.8	55.9
Nuclear Eng.	80.0	35.0	**85.0**	**85.0**	75.0	30.0	55.0	65.0	75.0	**85.0**	**85.0**	80.0	65.0	80.0	75.0	55.0
Urban Eng.	**82.0**	40.0	80.0	65.0	65.0	34.0	64.0	73.0	65.0	72.0	70.0	58.0	64.0	78.0	78.0	53.0

Table 18. Korean accuracy (%) by model and fine-grained subject category. Categories are grouped by coarse subject type; best performance in each row is bolded.

Category	Meta-Llama-3-70B-Instruct	Ministral-8B-Instruct-2410	Mistral-Small-3.2-24B-Instruct-2506	Mixtral-8x7B-Instruct-v0.1	OLMo-3-1025-7B	OLMo-3-1125-32B	OLMo-3.1-32B-Instruct	Qwen3-14B	Qwen3-30B-A3B-Instruct-2507	Qwen3-32B	Qwen3-4B	Qwen3-8B	Qwen3-Next-80B-A3B-Instruct	Gemma3-12b-it	Gemma3-4b-it
Art & Sports															
Architecture	58.4	28.2	54.6	34.5	10.9	34.0	37.4	50.8	60.9	58.4	45.0	47.5	**65.1**	45.4	17.6
Fashion	43.8	17.6	37.1	33.3	7.6	24.8	29.5	41.9	50.5	51.4	33.3	36.2	**53.3**	39.5	23.3
Festivals	**97.3**	**97.3**	**97.3**	**97.3**	27.0	**97.3**	**97.3**	**97.3**	**97.3**	**97.3**	**97.3**	**97.3**	**97.3**	**97.3**	13.5
Music	89.7	85.3	90.6	88.2	28.3	85.0	87.6	90.6	90.6	91.2	86.4	88.2	**92.6**	89.7	18.3
Painting	91.4	89.4	**96.7**	93.3	38.5	92.0	91.0	93.7	94.1	**96.7**	91.4	92.7	96.0	93.1	24.4
Performing	94.9	93.3	95.5	95.2	36.1	94.2	94.6	96.5	95.7	96.2	94.2	95.5	**96.7**	95.9	16.1
Photography	**93.6**	80.9	83.0	85.1	23.4	85.1	80.9	89.4	89.4	89.4	83.0	85.1	91.5	89.4	12.8
Sculpture	77.8	66.7	77.8	66.7	33.3	72.2	77.8	66.7	66.7	83.3	66.7	66.7	77.8	**88.9**	11.1
Sports	5.5	5.5	5.8	6.0	1.0	4.2	4.2	6.3	5.2	**6.5**	5.0	5.5	6.0	5.0	1.8
Culture															
Celebration Holiday	82.6	82.2	82.6	**82.8**	35.7	81.9	82.3	**82.8**	82.6	82.7	82.7	82.7	82.5	82.7	9.6
Clothing	58.5	34.0	49.1	32.1	3.8	35.8	35.8	50.9	60.4	54.7	37.7	49.1	**69.8**	47.2	28.3
Daily Life	24.9	23.3	24.8	23.8	8.2	23.2	23.9	25.0	25.1	**25.4**	23.9	24.4	25.2	24.6	4.6
Family	13.8	12.9	13.8	11.9	1.3	12.5	12.9	13.8	13.2	14.8	10.9	12.2	**15.8**	13.8	1.9
Food	7.1	6.1	7.4	6.1	1.9	5.4	5.2	6.5	6.7	**8.2**	6.3	6.9	6.9	6.7	3.2
Holiday	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Housing	61.7	22.2	48.1	34.6	12.3	33.3	54.3	46.9	59.3	63.0	43.2	54.3	**67.9**	48.1	11.1
Tradition	60.9	56.2	67.5	58.6	24.1	56.5	56.0	65.9	67.0	69.1	61.3	63.4	**70.4**	65.7	20.3
Work Life	69.3	66.7	69.6	68.1	26.4	67.7	67.9	70.3	70.6	**70.8**	69.3	68.9	69.8	69.1	9.9
HASS															
Administration	67.6	50.3	63.9	55.7	22.6	55.2	49.4	68.1	71.7	75.0	60.5	62.7	**76.0**	65.6	24.0
Economics	70.0	54.7	70.4	62.1	14.3	56.7	56.2	70.4	71.9	76.8	65.5	66.5	**78.3**	66.0	12.8
Education	40.7	35.5	39.7	37.1	16.1	35.8	35.7	41.7	41.5	**42.5**	38.7	39.1	41.9	39.0	6.1
Family	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Geography	62.3	40.6	59.1	45.9	10.1	42.1	38.4	58.8	60.1	63.2	47.8	54.1	**70.1**	57.5	9.7
History	54.9	44.0	61.0	47.5	17.3	45.2	39.9	54.5	60.4	61.0	46.2	52.9	**65.6**	52.5	24.7
Language	44.9	37.5	62.0	38.5	25.0	40.2	38.0	59.3	58.7	62.1	52.6	49.7	**64.5**	54.8	41.4
Law	53.0	35.5	51.4	42.7	14.0	42.1	38.2	53.0	55.4	**57.4**	45.6	47.6	57.2	49.3	28.4
Literature	66.5	46.9	70.3	60.9	28.0	59.3	53.5	72.9	74.5	80.4	61.3	64.0	**82.2**	67.3	36.3
Media	95.2	92.0	95.1	92.5	32.1	92.3	92.6	94.7	95.1	94.4	92.8	93.6	**95.3**	94.1	9.7
Philosophy	66.5	50.3	73.3	60.5	38.7	59.1	55.4	72.3	79.1	77.5	69.6	73.3	**81.2**	72.8	33.5
Politics	63.7	53.6	69.5	56.5	24.0	54.5	53.1	68.9	69.2	72.2	59.4	64.1	**75.5**	66.9	24.1
Religion	83.9	84.7	92.7	91.1	34.1	85.9	83.0	88.3	87.6	90.5	83.9	88.3	**93.4**	88.3	30.7
Trade	76.4	63.0	73.2	73.2	13.4	63.0	58.3	85.8	81.9	**86.6**	68.5	67.7	85.8	64.6	22.0
Welfare	66.2	43.6	61.1	51.3	17.1	50.4	50.9	65.0	68.4	**70.9**	56.0	62.0	67.5	61.5	23.1
Science															
Astronomy	**66.7**	**66.7**	**66.7**	50.0	33.3	50.0	50.0	50.0	**66.7**	**66.7**	50.0	50.0	**66.7**	**66.7**	0.0
Atmospheric Science	**60.0**	30.0	50.0	30.0	20.0	30.0	30.0	20.0	40.0	50.0	40.0	20.0	**60.0**	40.0	0.0
Biology	**70.0**	46.7	67.8	43.3	18.9	37.8	43.3	53.3	64.4	58.9	53.3	51.1	67.8	60.0	35.6
Chemistry	62.3	29.5	60.7	47.5	6.6	42.6	34.4	68.9	68.9	67.2	41.0	52.5	**78.7**	49.2	29.5
Earth Science	43.8	20.4	40.4	32.8	7.7	29.4	29.4	43.4	43.8	54.5	31.1	31.9	**60.9**	33.2	17.4
Geology	**75.0**	35.7	53.6	35.7	14.3	53.6	28.6	71.4	57.1	53.6	42.9	64.3	**75.0**	50.0	28.6
Life Science	44.4	22.2	33.3	29.6	3.7	22.2	22.2	40.7	40.7	40.7	18.5	40.7	**48.1**	37.0	22.2
Math	40.0	25.5	41.8	32.7	14.5	30.9	25.5	41.8	52.7	54.5	45.5	40.0	**58.2**	32.7	16.4
Physics	51.4	18.4	50.3	34.6	9.5	36.9	34.6	52.5	62.6	63.1	45.8	49.2	**74.3**	37.4	27.4
Statistics	50.0	16.7	33.3	16.7	0.0	33.3	16.7	50.0	66.7	**83.3**	66.7	66.7	66.7	33.3	0.0
Social Intelligence															
Bias	83.6	58.0	73.7	68.9	44.1	56.5	81.6	88.8	90.0	**92.7**	78.6	86.0	92.1	75.2	6.3
Value/Alignment	20.0	0.0	**60.0**	**60.0**	0.0	40.0	40.0	**60.0**	**60.0**	**60.0**	**60.0**	**60.0**	**60.0**	**60.0**	**60.0**
Tech.															
AI	68.8	72.9	**89.6**	72.9	25.0	70.8	60.4	77.1	83.3	83.3	85.4	66.7	83.3	72.9	20.8
Aerospace Eng.	57.7	30.8	53.8	52.0	7.7	42.3	42.3	57.7	65.4	61.5	38.5	61.5	**69.2**	42.3	19.2
Agricultural Eng.	45.2	16.8	30.3	22.1	7.8	24.0	26.0	41.3	36.8	48.4	32.9	29.7	**49.0**	30.3	19.4
Biomedical Eng.	62.0	32.3	58.5	39.7	14.4	35.9	33.9	61.7	66.7	66.8	46.7	53.8	**70.1**	54.4	31.5
Chemical Eng.	37.1	21.0	48.4	33.9	12.9	37.1	37.1	53.2	54.8	51.6	40.3	48.4	**67.7**	38.7	19.4
Civil Eng.	42.3	15.5	31.4	31.0	5.9	29.3	23.8	39.3	47.7	50.2	33.9	32.2	**55.6**	29.3	19.7
Coding	**100.0**	**100.0**	50.0	0.0	0.0	50.0	0.0	**100.0**	**100.0**	50.0	50.0	50.0	**100.0**	50.0	50.0
Electrical Eng.	53.0	21.4	40.9	31.7	13.8	31.8	33.6	42.2	52.1	55.8	36.8	41.8	**57.1**	38.4	23.3
Energy	40.6	20.8	38.2	27.1	12.1	27.1	31.4	32.9	40.1	46.9	38.6	35.7	**50.2**	32.4	16.9
Environmental Eng.	41.9	14.8	30.6	27.9	11.4	20.5	21.8	31.4	44.5	48.0	34.5	34.9	**51.5**	37.1	18.8
IT	73.1	43.6	66.2	58.7	14.6	58.9	54.7	70.1	74.2	76.9	62.7	63.2	**77.9**	64.7	28.9
Marine Eng.	53.7	22.1	50.5	41.1	9.5	41.1	31.6	40.0	51.6	53.7	33.7	41.1	**63.2**	37.9	21.1
Materials Eng.	46.9	17.5	40.5	29.0	7.7	26.1	27.2	44.3	53.3	51.8	33.0	37.4	**62.8**	35.6	23.5
Mechanics	42.5	16.0	39.8	28.3	12.7	24.7	24.1	43.5	50.6	52.6	31.9	37.5	**58.7**	35.2	17.3
Nuclear Eng.	50.0	0.0	10.0	50.0	0.0	30.0	0.0	40.0	60.0	**90.0**	50.0	20.0	80.0	10.0	40.0
Urban Eng.	47.6	23.1	41.5	29.3	11.6	31.3	25.9	48.3	49.0	44.2	40.8	35.4	**53.1**	44.2	27.2

Table 19. Evaluation results of 15 open LLMs in English across four different sampling strategies.

Model	Random	Stratified	Chatbot Arena	MixEval
Qwen3-4B	0.600	0.559	0.592	0.577
Qwen3-8B	0.624	0.577	0.601	0.598
Qwen3-14B	0.662	0.609	0.638	0.635
Qwen3-32B	0.683	0.642	0.672	0.660
Qwen3-30B-A3B-Instruct-2507	0.726	0.693	0.744	0.707
Qwen3-Next-80B-A3B-Instruct	0.734	0.704	0.727	0.720
gemma-3-4b-it	0.500	0.502	0.525	0.490
gemma-3-12b-it	0.664	0.637	0.658	0.647
Meta-Llama-3-70B-Instruct	0.674	0.638	0.668	0.667
Ministral-8B-Instruct-2410	0.406	0.379	0.401	0.386
Mistral-Small-3.2-24B-Instruct-2506	0.662	0.629	0.658	0.661
Mixtral-8x7B-Instruct-v0.1	0.612	0.587	0.607	0.610
OLMo-3-1025-7B	0.325	0.292	0.321	0.307
OLMo-3-1125-32B	0.571	0.532	0.568	0.562
OLMo-3.1-32B-Instruct	0.673	0.640	0.678	0.653

Table 20. Evaluation results of 15 open LLMs in Korean across four different sampling strategies.

Model	Random	Stratified	Chatbot Arena	MixEval
Qwen3-4B	0.557	0.550	0.497	0.533
Qwen3-8B	0.577	0.570	0.521	0.550
Qwen3-14B	0.617	0.607	0.566	0.595
Qwen3-32B	0.671	0.666	0.633	0.650
Qwen3-30B-A3B-Instruct-2507	0.642	0.639	0.606	0.623
Qwen3-Next-80B-A3B-Instruct	0.706	0.705	0.673	0.684
Gemma3-4b-it	0.190	0.190	0.218	0.222
Gemma3-12b-it	0.564	0.555	0.515	0.555
Meta-Llama-3-70B-Instruct	0.615	0.622	0.579	0.598
Ministral-8B-Instruct-2410	0.455	0.452	0.391	0.447
Mistral-Small-3.2-24B-Instruct-2506	0.602	0.597	0.559	0.596
Mixtral-8x7B-Instruct-v0.1	0.487	0.490	0.443	0.481
OLMo-3-1025-7B	0.265	0.248	0.196	0.218
OLMo-3-1125-32B	0.473	0.476	0.434	0.480
OLMo-3.1-32B-Instruct	0.501	0.493	0.423	0.461

Table 21. Evaluation results of 15 open LLMs using customized BENCHHUB across five use cases.

Model	(a)	(b)	(c)	(d)	(e)
Qwen3-4B	0.553	0.677	0.555	0.413	0.593
Qwen3-8B	0.556	0.690	0.584	0.401	0.637
Qwen3-14B	0.580	0.704	0.629	0.420	0.658
Qwen3-32B	0.605	0.795	0.687	0.468	0.684
Qwen3-30B-A3B-Instruct-2507	0.707	0.739	0.703	0.451	0.665
Qwen3-Next-80B-A3B-Instruct	0.698	0.796	0.717	0.469	0.678
Gemma3-4b-it	0.509	0.177	0.400	0.086	0.059
Gemma3-12b-it	0.620	0.758	0.586	0.372	0.576
Meta-Llama-3-70B-Instruct	0.607	0.759	0.622	0.408	0.626
Ministral-8B-Instruct-2410	0.405	0.614	0.366	0.324	0.468
Mistral-Small-3.2-24B-Instruct-2506	0.585	0.764	0.587	0.401	0.568
Mixtral-8x7B-Instruct-v0.1	0.529	0.648	0.498	0.354	0.534
OLMo-3-1025-7B	0.263	0.276	0.264	0.153	0.306
OLMo-3-1125-32B	0.480	0.661	0.484	0.344	0.460
OLMo-3.1-32B-Instruct	0.628	0.687	0.543	0.325	0.609

Table 22. Model ranking and accuracy across BENCHHUB customized for math evaluation and MATH (Hendrycks et al., 2021b) for simulation study on data contamination

Rank	BENCHHUB (Ours)	MATH (Hendrycks et al., 2021b)
1	Qwen/Qwen3-8B (0.35)	OLMoE-Contaminated (0.84)
2	gemma-3-4b-it (0.30)	Qwen/Qwen3-8B (0.31)
3	Llama-3.1-8B-Instruct (0.26)	gemma-3-4b-it (0.28)
4	OLMoE-Contaminated (0.22)	Llama-3.1-8B-Instruct (0.13)
5	OLMoE-base (0.16)	OLMoE-base (0.10)