The GaoYao Benchmark: A Comprehensive Framework for Evaluating Multilingual and Multicultural Abilities of Large Language Models

Yilun Liu[1]*, **Chunguang Zhao**[1]*, **Mengyao Piao**[1], **Lingqi Miao**[1], **Shimin Tao**[1], **Minggui He**[1], **Chenxin Liu**[1], **Li Zhang**[1], **Hongxia Ma**[1], **Jiaxin Guo**[1], **Chen Liu**[1], **Liqun Deng**[1], **Jiansheng Wei**[1], **Xiaojun Meng**[1], **Fanyi Du**[1], **Daimeng Wei**[1], **Yanghua Xiao**[2]
[1] Huawei, China
[2] Fudan University, China
liuyilun3@huawei.com, zhaochunguang6@huawei.com

Abstract

The evaluation of multilingual and multicultural capabilities in large language models (LLMs) is pivotal for their global application and equitable development. However, existing benchmarks are often constrained by several critical shortcomings: a lack of systematic evaluation dimensions, limited language coverage, insufficient representation of existing models, and reliance on machine translation without rigorous human quality control. To address these gaps, we introduce **GaoYao** [2] , a comprehensive multilingual and multicultural benchmark designed to provide a holistic assessment of LLM abilities. Our framework systematically categorizes multilingual and multicultural capabilities into three evaluation dimensions—*General Multilingual Abilities* (knowledge Q&A, reasoning, comprehension, translation, instruction following, multi-turn dialogue), *Cross-cultural Abilities* (Cultural differences of common concepts), and *Monocultural Abilities* (unique concepts in a culture)—encompassing nine fine-grained sub-capabilities. We evaluate these sub-capabilities by integrating existing resources and, most notably, by significantly expanding test sets for instruction-following (AlpacaEval) and multi-turn dialogue (MT-bench) to 19 languages through meticulous curation by our team of language experts to ensure superior quality compared to purely machine-translated alternatives, surpassing existing efforts by 111% in number of supported languages. For cultural sub-capabilities, we enhance cultural coverage (by 88%) to 34 cultures through an expert-in-the-loop data synthesis process. We also conduct a tiered evaluation on 20+ state-of-the-art commercial and open-source models. GaoYao establishes a more systematic, extensive, and reliable benchmark, aiming to accurately map the landscape of multilingual capabilities in LLMs and guide their future development.

1 Introduction

The exponential growth of large language models (LLMs) has marked a transformative era in artificial intelligence, pushing the boundaries of language comprehension and generation across a multitude of tasks. As these models increasingly serve global user bases, their ability to understand and

*Equal contribution.

[2]GaoYao is derived from Chinese mythology, where he served as the first judicial officer, symbolizing fairness and comprehensiveness—qualities that align with our benchmark's goal of providing a just and thorough assessment of multilingual and multicultural abilities.

Preprint. Under review.

generate text in diverse languages and cultural contexts has become a critical measure of their utility and inclusivity. This underscores the paramount importance of a robust, comprehensive, and fair evaluation benchmark to accurately gauge and steer the development of multilingual and multicultural capabilities.

However, the current landscape of multilingual and multicultural evaluation is fraught with significant challenges that hinder a holistic understanding of model performance. A primary issue is the lack of systematicity in existing benchmarks. Many prominent efforts focus narrowly on one or a few facets of language ability, such as factual knowledge or culture-agnostic tasks, while neglecting the broader capability dimensions a model should possess (*e.g.*, instruction-following abilities) and the intricate interplay between language and culture. For instance, while INCLUDE [Romanou et al., 2025] offers extensive language coverage, it is confined to knowledge-based question and answering (Q&A). Similarly, Pomerenke et al. [2025] provide broad model and language coverage but omits crucial cultural dimensions. These fragmented approaches fails to provide a unified framework necessary for a complete diagnosis of a model's strengths and weaknesses.

Secondly, there is a severe limitation in language coverage for critical evaluation dimensions. Capabilities such as instruction following and multi-turn dialogue, which are central to the practical deployment of LLMs, are predominantly assessed only in English. Although high-quality benchmarks like ALPACAEVAL [Li et al., 2023] and MT-BENCH [Zheng et al., 2023] exhibit strong correlation with human judgment, their rigorous multilingual extensions, such as X-ALPACAEVAL (4 languages) [Zhang et al., 2024] and OMGEVAL (9 languages) [Liu et al., 2024], remain limited in supported languages. This leaves a vast majority of the world's languages underserved and their speakers' interaction patterns unevaluated, creating a significant gap in our understanding of true model globalization.

Furthermore, the scope of models evaluated by existing benchmarks is often incomplete. Many studies [Liu et al., 2024, Romanou et al., 2025] focus on a specific category of models, such as smaller open-source LLMs, or fail to keep pace with the rapid release of state-of-the-art (SOTA) commercial and large-scale LLMs like DeepSeek-V3.1 [DeepSeek, 2025] and Qwen3-235B-A22B [Yang et al., 2025]. This may result in an incomplete and outdated map of the LLM's multilingual capabilities, making it difficult for researchers and practitioners to make informed comparisons.

Finally, the data quality of many multilingual benchmarks is a cause for concern. A substantial number rely primarily on automated methods such as machine translation (MT) for dataset localization (from a main source language such as English to target languages) and data synthesis with LLMs, which can introduce hallucinations, cultural insensitivities, and a lack of natural fluency. According to Pomerenke et al. [2025], only a minority of multilingual benchmarks (around 35.7%) incorporate rigorous human translation and validation. This over-reliance on automation compromises the reliability and quality of the evaluations.

To address these challenges, we introduce GaoYao, a benchmark evaluating LLMs' multilingual and multicultural abilities that emphasizes systematicity, breadth, and quality. GaoYao features on the following aspects:

(1) Systematicity. Based on existing theories on cultures [Schein, 2010, Hall, 1976], we categorizes evaluation dimensions of GaoYao into three major layers: *General Multilingual Abilities* (handling consistent concepts across languages), *Cross-cultural Abilities* (navigating culturally-variant concepts), and *Monocultural Abilities* (understanding unique concepts in a culture). And from the perspective of cognition theories [Anderson and Krathwohl, 2001], the evaluation dimensions can be further expanded into nine sub-layers from memorizing knowledge to perform creative writing, thereby enabling a comprehensive evaluation matrix.

(2) More Languages. We address the language coverage gap by leveraging a team of full-time multilingual experts to meticulously translate and adapt critical English evaluation sets (*i.e.*, instruction-following and multi-turn dialogue) into over 19 languages, significantly surpassing previous efforts by 111% in number of supported languages.

(3) Comprehensive LLMs Coverage. Our model evaluation is tiered and exhaustive, encompassing the latest commercial and open-source models with both large-scaled and smaller-scaled sizes to provide a complete landscape of the field.

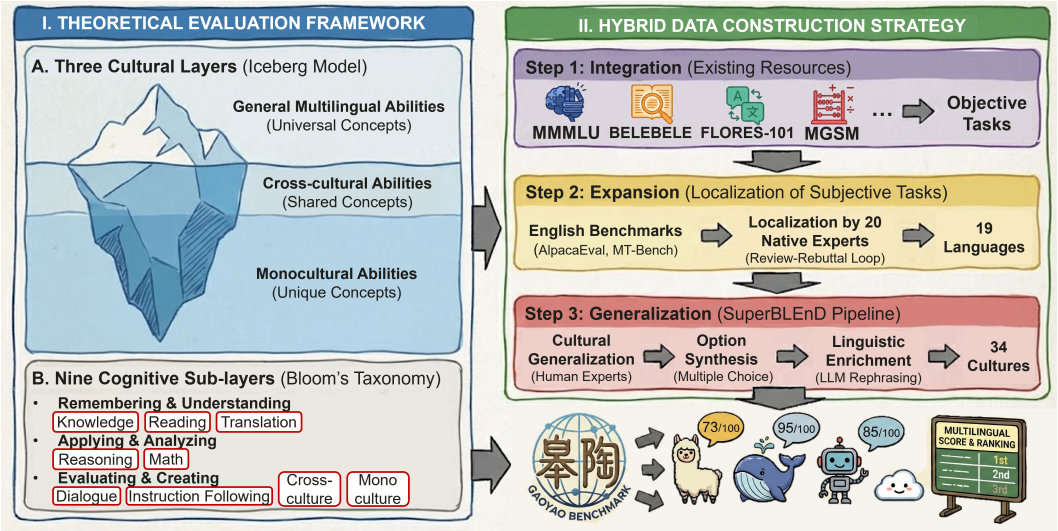


Figure 1: Illustration on design and construction of GaoYao. The benchmark is grounded in theoretical models of culture and cognition, and constructed through a hybrid strategy of integration, expansion and generalization.

(4) Data Quality. The data curation follows a rigorous human-in-the-loop process with a two-round review-rebuttal feedback loop to monitor the data quality for each language. All participated multilingual experts are professionals in fields such as translation, localization, proofreading, editing, copy-writing, technical writing, and linguistic testing. These measures ensure data quality of GaoYao, enabling accurate multilingual evaluation and multicultural perception.

To achieve this, we construct the GaoYao benchmark through a hybrid strategy of integration, expansion, and generalization: *(1) Integration.* For each sub-layer within our evaluation dimensions, we first seek for well-established open-source datasets aiming at evaluating the specific abilities. After selecting based on both quality and coverage, we incorporate subsets of them into GaoYao, such as INCLUDE [Romanou et al., 2025] for multilingual knowledge, BELEBELE [Bandarkar et al., 2024] for multilingual reading comprehension, and FLORES-101 [Goyal et al., 2022] for multilingual translation. *(2) Expansion.* As discussed above, we expand the language coverage of two vital English test sets into over 19 languages through rigorous manual translation. *(3) Generalization.* For the evaluation of cultural-related abilities, since direct translation still represent the culture of the source language, we developed a novel human-in-loop data synthesis procedure to acquire a more generalized cross-cultural test set. Specifically, we generalize the open-source BLEND [Myung et al., 2024] into SUPERBLEND, expanding its cultural coverage from 16 to 34 nations.

In summary, our contributions are as follows:

- We propose a systematic multilingual evaluation framework with three major layers and nine sub-layers, providing a comprehensive evaluation for LLMs' multilingual capabilities while addressing the fragmented nature of existing benchmarks.
- We address the challenge of limited language and cultural coverage for critical evaluation dimensions, expanding to 19 languages and generalizing to 34 cultures via a human-in-the-loop methodology, while ensuring high data quality.
- We evaluate a wide array of SOTA LLMs through a tiered approach, providing a map to the landscape of LLM multilingual and multicultural capabilities.

In addition, we release various assets including codes and datasets utilized in GaoYao benchmark.

2 Methodology

The construction of GaoYao is founded on a synthesis of established cognitive and cultural theories, executed through a rigorous, hybrid data construction strategy. As illustrated in Fig. 1, our framework

begins by defining a theoretical landscape of multilingual and multicultural capabilities critical in LLM evaluation. Guided by this theoretical structure, we employ a three-pronged approach—Integration, Expansion, and Generalization—to curate a comprehensive benchmark that addresses existing gaps in systematicity, language (or culture) coverage, and data quality. Section 2.1 details the theoretical underpinnings of our evaluation dimensions and Section 2.2 discusses the specific processes involved in constructing the GaoYao benchmark through these three strategies.

2.1 Layered Evaluation Dimensions for LLMs' Multilingual and Multicultural Capabilities

To ensure a systematic and holistic evaluation, GaoYao's evaluation dimensions are theoretically grounded in established models of human culture and cognition.

Theoretical Foundations of Three Major Layers Drawing inspiration from the Cultural Iceberg Model [Hall, 1976] and the Three-Layer Model of organizational culture [Schein, 2010], we posit that existing multilingual benchmarks often predominantly assess "surface-level" linguistic proficiencies while overlooking the deeper, implicit cultural contexts that shape communication. To address this, GaoYao categorizes capabilities into three major layers representing deepening levels of cultural embedment:

- **General Multilingual Abilities:** This layer corresponds to the "tip of the iceberg," focusing on universal concepts that remain consistent across languages (*e.g.*, applying target language to handle problems involving reasoning, knowledge or comprehension).
- **Cross-cultural Abilities:** Moving beneath the surface, this layer assesses the model's capacity to navigate shared concepts that manifest differently across cultures. For instance, while the lexical term "dragon" translates directly, its symbolic meaning varies drastically: Western dragons are typically depicted as malevolent, fire-breathing monsters to be slain, whereas the eastern dragon (or loong) is revered as an auspicious symbol of power, benevolence, and control over water [Zhao, 1988, Maguth and Wu, 2020]. An LLM must discern these sharp semantic and cultural divergences.
- **Monocultural Abilities:** The deepest layer evaluates the understanding of unique concepts exclusive to specific cultures, which often lack direct equivalents elsewhere. An example is the Chinese phenomenon of "Chunyun" (the massive Spring Festival travel rush [Zhu et al., 2021]), a culturally specific event laden with unique social implications. Another example is "Namaste", the greeting etiquette in India involving pressing palms together [Zhang et al., 2025].

Deriving Nine Sub-layers via Cognitive Taxonomy Within these major cultural layers, we further ensure a comprehensive evaluation matrix by structuring tasks according to Bloom's Taxonomy of cognitive domains [Anderson and Krathwohl, 2001]. This taxonomy categorizes human thought processes into six categories along a gradient of complexity, ranging from basic remembering to complex creation. Inspired by the six cognitive levels, we designed nine distinct sub-capabilities within GaoYao to ensure a scientifically rigorous and comprehensive assessment of LLMs' multilingual capabilities:

- **Remembering & Understanding:** Reflected by tasks of multilingual *Knowledge* Q&A, *Reading* Comprehension, and *Translation*.
- **Applying & Analyzing:** Assessed through *Reasoning* tasks and *Math* problem solving.
- **Evaluating & Creating:** Evaluated via high-level subjective tasks including *Instruction Following* and *Multi-turn Dialogue*, which often involve creative writing to satisfy user desires, as well as the aforementioned deeper *Cross-cultural* and *Monocultural* understanding tasks, which involve making appropriate cultural judgments.

2.2 Construction of GaoYao Benchmark

Guided by the theoretically layered framework above, we construct the GaoYao benchmark through a hybrid strategy combining the integration of established resources (for seven of the nine sub-layers), the linguistic expansion of high-value under-served benchmarks (two most critical sub-layers:

instruction following and multi-turn dialogues), and the generalization of cultural data through human-in-loop synthesis pipelines (the cross-cultural layer).

2.2.1 Integration of Existing Test Sets

For several of the defined cognitive sub-layers, particularly those related to objective knowledge and reasoning, the research community has already established high-quality open-source benchmarks. Rather than reinventing these, we conducted literature review and quality checks to select and integrate some of the most widely-verified and robust datasets into GaoYao, ensuring a complete coverage of our defined evaluation sub-layers. The integrated resources are mapped to our sub-layers as follows:

(1) *Knowledge & Reasoning*: Both INCLUDE [Romanou et al., 2025] and MMMLU [OpenAI, 2024] are integrated given their coverage on factual knowledge spanning various subjects from elementary-level knowledge up to advanced professional subjects. Compared with INCLUDE, MMMLU focuses more on evaluating reasoning abilities, *i.e.*, how LLMs apply these knowledge to solve practical problems.

(2) *Reading*: We incorporate BELEBELE [Bandarkar et al., 2024] for evaluating multilingual reading comprehension capabilities given its native passage coverage and rigorous quality assurance.

(3) *Translation*: FLORES-101 [Goyal et al., 2022] provides a widely-recognized standard for assessing MT across numerous language pairs.

(4) *Math*: MGSM [Shi et al., 2023] is also a widely-used dataset to evaluate multilingual mathematical reasoning capabilities.

(5) *Cross-culture & Monoculture*: Since the research community has only recently begun to rigorously define and evaluate the multicultural capabilities of LLMs [Rystrøm et al., 2025], open-source resources remain scarce. We leverage two recent datasets: SAGE [Guo et al., 2025] for identifying cultural differences in shared concepts (cross-culture) and CULTURESCOPE [Zhang et al., 2025] for understanding unique cultural concepts (monoculture). While both datasets delve deeply into culture-specific concepts and employ rigorous design procedures, their coverage is restricted to Chinese and Spanish. To supplement this, we constructed a cross-cultural evaluation set spanning 34 nations (see Section 2.2.3).

2.2.2 Expansion of Language Coverage for ALPACAEVAL and MT-BENCH

Instruction following and multi-turn dialogue represent critical capabilities reflecting an LLM's practical utility and "human-likeness." However, existing multilingual benchmarks heavily prioritize objective tasks, leaving these subjective, open-ended abilities predominantly evaluated only in English. To close this significant gap, we selected two widely recognized English benchmarks: ALPACAEVAL, validated by over 20k human judgments for general instruction following, and MT-BENCH, designed with challenging multi-turn questions across intent categories such as role playing and creative writing. We then expanded their coverage to over 19 languages, denoting as S-ALPACAEVAL and S-MT-BENCH, respectively.

This expansion was not a simple translation task but a rigorous localization effort. From the language service center of a top-tier corporation, we recruited a team of 20 native-speaker professionals with expertise in translation, localization, and linguistic testing. The team dedicated a total of 175 person-days to this development. To ensure the highest quality, a strict review-rebuttal feedback loop was implemented for each language. Third-party reviewers continuously inspected samples during annotation. Disagreements triggered a discussion phase where annotators either revised their work based on reviewer concerns or provided justifications to persuade the reviewer to unflag the sample.

Crucially, there is a localization process to make sure every user question is linguistically feasible, which can hardly be guaranteed using MT. For instance, constrained English instruction like "list items starting with the letter A" will be invalid if being translated literally to a language without letter A. Thus, such instructions were manually adapted or reconstructed to suit the phonetic and script characteristics of the target language while ensuring the cognitive task remained equivalent across languages.

2.2.3 Generalization of Cross-cultural Evaluation (SUPERBLEND)

As discussed in Section 2.2.1, existing cultural evaluation sets are limited in its culture coverage. However, unlike the two multilingual abilities in Section 2.2.2, expanding coverage of existing cultural evaluation sets presents a unique challenge: direct translation of cultural QA pairs retains the cultural perspective of the source language (especially for the answer parts) which fails to test true cross-cultural generalizability, while a purely manual reconstruction is prohibitive in costs. To address this, we developed a three-stage semi-automated data expansion procedure incorporating human verification, where native speakers provide initial seeds in the first stage and inspect quality in subsequent diversifying stages to ensure both high accuracy and linguistic diversity. The resulting evaluation set, denoted as SUPERBLEND (an generalization of the BLEND dataset [Myung et al., 2024]), evaluates LLMs' understanding of cultural differences regarding everyday concepts across 34 nations (up from 16 in BLEND).

Stage 1: Generalization Q&A Seeds to More Cultures Starting from the question templates released by BLEND, which cover culturally shared topics ranging from festivals and food to sports, we first selected a high-quality subset. This filtration process excluded templates which were not universally applicable or were deemed sensitive in specific cultural contexts. To ensure cultural authenticity and factual accuracy, answers to the selected questions were built by native members of corresponding cultures. These annotators come from the cooperated language service center as described in Section 2.2.2. Of the 34 cultures SuperBLEnD covers, data for 16 was inherited from BLEND. For each of the 18 newly expanded cultures, three native annotators are assigned. The instruction asks annotators to provide 1-3 concise answers to each question based strictly on their personal life experiences within that cultural context, unassisted by AI or search engines. To prevent forced fabrication, annotators could mark questions as "not applicable" or "no clear answer."

Q&A pairs across all 34 cultures underwent rigorous manual verification to eliminate duplicate (*e.g.*, merging similar answers), invalid, or toxic content, discarding approximately 41.1% of the raw data. The final collection contains an average of 2.17 high-quality answers per question template. Note that many cultural questions accept multiple correct answers. For example, both "beer" and "carbonated drinks" are valid, verified responses to the question: "What do young people in Malaysia usually drink at nightclubs?"

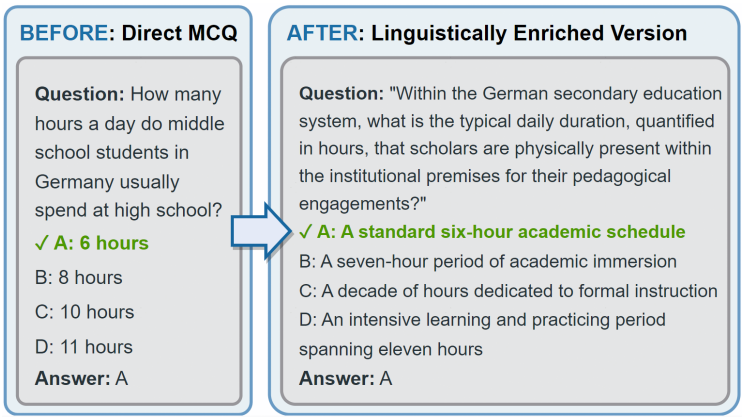


Figure 2: An example of the linguistic enrichment process (stage 3), which increases complexity of MCQs without altering the underlying cultural fact.

Stage 2: Option Synthesis For ease of evaluation and to enhance diversity, following Myung et al. [2024], we generalize each verified seed question into a series of multiple-choice questions (MCQs). The volume of generated MCQs is dynamically adjusted based on the answer set size, ensure that all verified answers are comprehensively covered as correct options while upsampling (with different distractors) questions with fewer answers as a balance. For each MCQ, we synthesized options by combining the correct answer for the target country with three wrong answers as distractors derived from other countries. In cases where insufficient real-world distractors were available, an LLM was employed to generate plausible but incorrect "dummy options" that exist in reality but do not answer the specific question (Appendix A.1).

Synthesized MCQs underwent automated and human verification to ensure safety, answer uniqueness, and the exclusion of synonyms or hierarchical conflicts. For instance, in the "Malaysia nightclub"

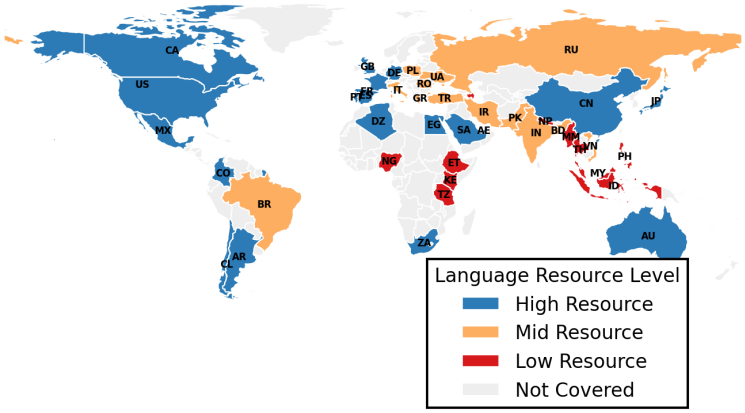


Figure 3: The language and culture coverage on the world map. Colors indicate resource popularity levels.

scenario, "Pepsi" is an unsuitable distractor when the target answer is "beer." Because "Pepsi" falls under the category of "carbonated drinks" (another acceptable cultural answer), its inclusion introduces a hierarchical relationship that compromises the MCQ's validity.

Stage 3: Linguistic Enrichment To further enhance linguistic diversity and reasoning difficulty, the finalized MCQs underwent a rephrasing stage. We utilized an LLM prompted (Appendix A.2) to rewrite both the question stem and options, employing techniques like paraphrasing, voice alternation, and syntactic restructuring without altering the core semantic meaning or named entities. This ensures the benchmark tests cultural knowledge rather than simple pattern matching. Fig. 2 illustrates an example of this enrichment process, showing how a straightforward MCQ is transformed into a more complex version while retaining the same cultural core.

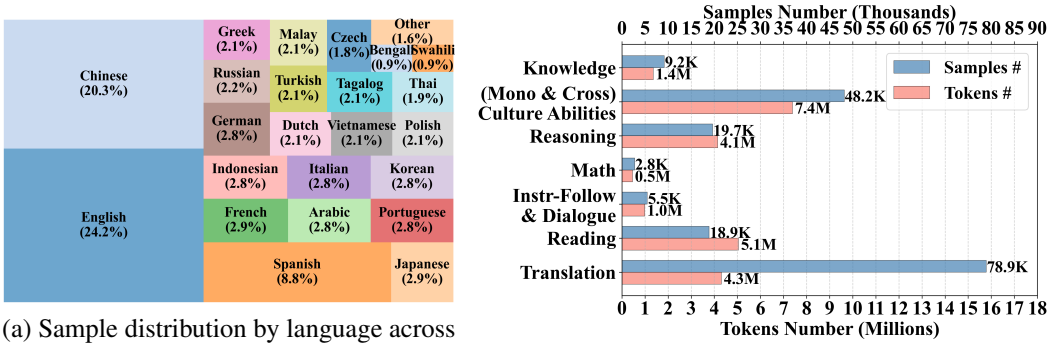
3 Experiment

3.1 Experimental Setups

To empirically validate GaoYao's efficacy in mapping the global LLM landscape, we conducted a tiered evaluation across a spectrum of models representing the current SOTA. Our selection encompasses both open-weights models (*e.g.*, Qwen3 series, DeepSeek-V3.1) inferred on standardized NPU computation nodes using the MindSpore [Chen, 2021] and PyTorch [Paszke et al., 2019] frameworks, and proprietary commercial models (*e.g.*, Doubao-seed-1.6, Qwen-max) accessed via official APIs. The specific model versions and resource addresses are in Table 4. We make sure all evaluated LLMs are post-trained versions (*e.g.*, "instruct" or "chat" versions). Following Yang et al. [2025], we adopted only a random 10% subset of MMMLU due to its unproportionate volume with other test sets. Also, to control cost, the proposed S-ALPACAEVAL and S-MT-BENCH were curated based on a random 30% subset of the English source datasets. Section 3.1.1 and Section 3.1.2 further illustrates the setups.

3.1.1 Statistics of Test Sets in GaoYao

As shown in Fig. 3, the dataset spans 26 languages distributed across 51 nations/areas (34 of them are culturally represented as discussed in Section 2.2.3). The distribution encompasses five geopolitical clusters: Western Europe, Eastern Europe, East Asia & Southeast Asia, Middle East & Africa, and South Asia (See Table 2 for detailed statistics). A crucial design principle of GaoYao is the mitigation of resource bias; as illustrated in Fig. 4(a), excluding the three dominant lingua francas, the distribution is meticulously balanced, with each remaining language constituting roughly 3% of the total volume. Classified by the framework of Joshi et al. [2020], the languages include nine low-resource and ten mid-resource varieties (detailed in Appendix B).



(a) Sample distribution by language across all test sets of GaoYao

(b) Distribution by evaluation dimensions

Figure 4: Distribution statistics of test sets in GaoYao (a) by languages and (b) by evaluation sub-layers.

Fig. 4(b) shows the distribution of test set sizes across the nine evaluation sub-layers of the GaoYao benchmark, as introduced in Section 2.1. The distribution of samples and tokens varies distinctively across sub-layers due to their inherent task characteristics. While *Translation* comprises the highest volume of samples, it accounts for a relatively modest share of total tokens, reflecting the sentence-level brevity typical of the FLORES-101 dataset. In contrast, sub-layers such as *Reasoning* and *Reading* exhibit a significantly higher token-to-sample ratio, as these domains necessitate extensive context to define complex problem spaces. Additionally, the cultural layers (*Monoculture* and *Cross-culture*) represent a relatively large token count to ensure sufficient depth to capture cultural nuances, reflecting GaoYao's emphasis on cultural evaluation of LLMs.

3.1.2 Evaluation Approaches

The evaluation protocol for each sub-dataset can be divided into two categories (details on metrics, judges and calculation methods are in Table 3):

Objective Evaluation: For question type with deterministic outputs (*e.g.*, MCQ, calculation problems), we utilize standardized prompt templates [OpenAI, 2024, Romanou et al., 2025, Bandarkar et al., 2024, Goyal et al., 2022] and rule-based extraction with regular expressions to parse answers from LLMs' responses. To ensure reproducibility, all pre-processing and post-processing scripts have been released.

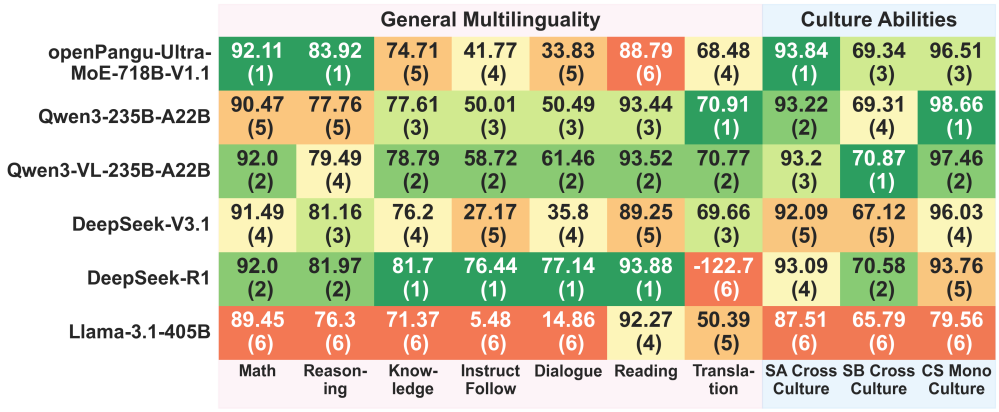
Subjective Evaluation: For open-ended tasks (*e.g.*, Q&A), we adopt the widely-used "LLM-as-Judge" paradigm[Li et al., 2023, Zheng et al., 2023], where the judge model compare response from a candidate model with a reference response based on specific dimensions and concludes with "win", "lose" or "tie". We standardized on DeepSeek-v3.1 as the judge due to its superior reasoning abilities. The primary metric is *Win Rate* against the reference responses. For datasets lacking inherent references (S-ALPACAEVAL, S-MT-BENCH), we introduced Qwen3-235B-A22B as the reference anchor.

All scores (*e.g.*, accuracy, win rate) are displayed at the scale of 0-100 (except for Comet where we simply multiplied the original output Comet-20 scores by 100). We aggregate results along two axes: *Task Dimension* (averaging across all languages for a specific evaluation sub-layer) and *Language Dimension* (averaging across all sub-layers for a specific language), enabling a multi-view diagnosis of model capabilities.

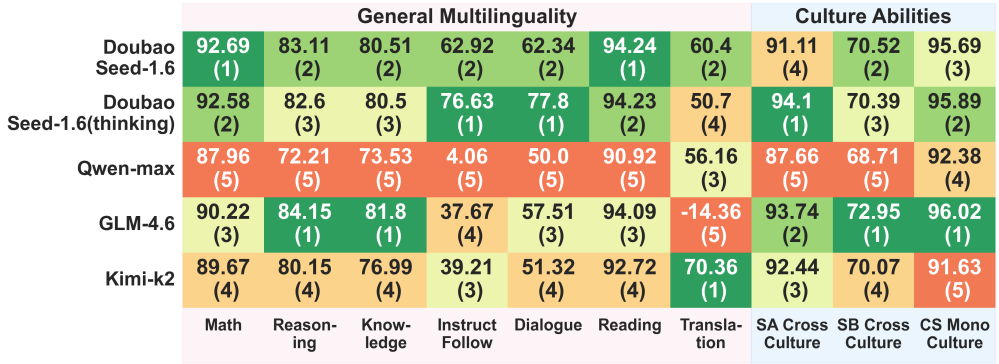
3.2 Multilingual LLM Leaderboard

3.2.1 Flagship Models

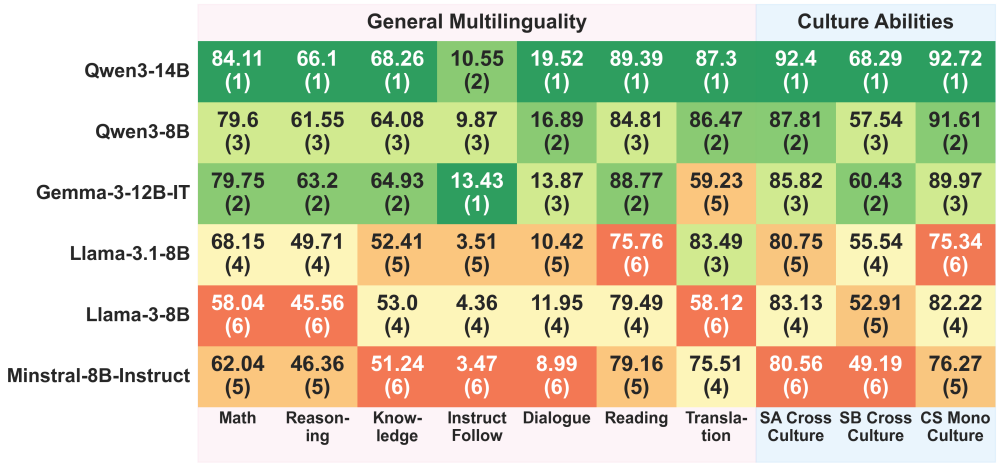
Fig. 5(a) and (b) present the landscape of flagship models, including open-source leaders and closed-source commercial APIs. The results reveal distinct capability profiles rather than a uniform dominance:



(a) Open-Source Models



(b) Closed-Source (API-based) Models



(c) Compact Models ($<$ 20B)

Figure 5: Performance heatmaps across nine evaluation sub-layers. Scores are averaged across all languages. Numbers in parentheses indicate rank within the group. Backgrounds: Pink (General Multilingual), Blue (Cultural Abilities). SA, SB and CS represents specific datasets: SAGE, SUPERBLEND and CULTURESCOPE.

- **Logic & Culture Specialist:** OpenPangu-Ultra-MoE-V1.1 exhibits exceptional strength in *Reasoning* and *Math*, while simultaneously securing the top rank in the *Cross-culture* sub-layer. This correlation suggests that rigorous logical training may facilitate the structured understanding of complex cultural frameworks.
- **Knowledge Heavyweights:** DeepSeek-R1 demonstrates dominance in *Knowledge* and *Reading*, suggesting a pre-training corpus with extensive informational breadth.
- **Interaction Specialists:** Doubao-Seed-1.6 leads both open-source and closed-source models in *Instruction Following* and *Dialogue*, reflecting a post-training strategy optimized for conversational utility and adherence to complex user constraints.

3.2.2 Compact Models

We have also conducted extensive experiments for compact models—which are particularly amenable to efficient adaptation and deployment within the MindSpore framework. Fig. 5(c) illustrates the performance of compact models ($<$ 20B parameters). We observe three critical trends:

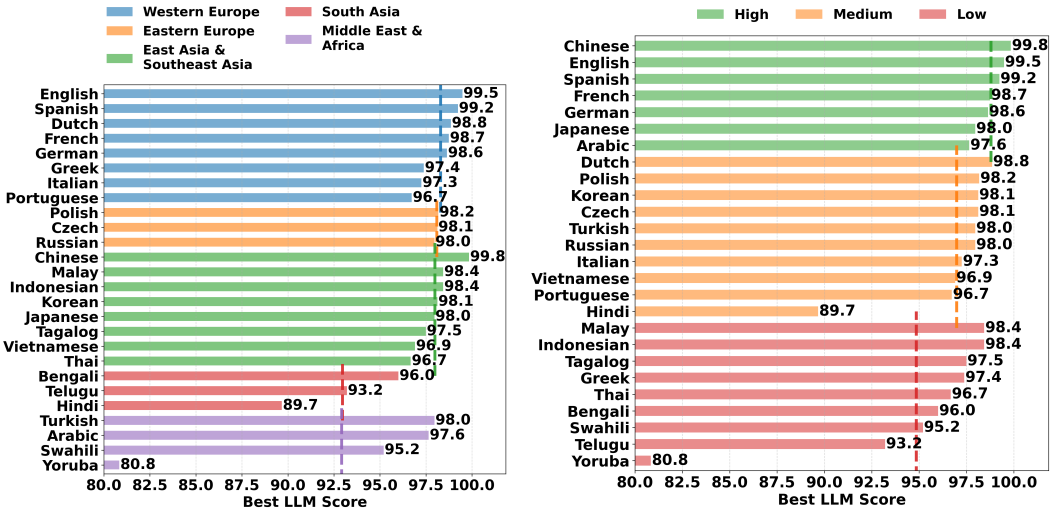
Validation of Scaling Laws: Across all multilingual tasks, performance scales predictably with parameter count (*e.g.*, 8B $<$ 14B), reaffirming that model capacity remains a bottleneck for multilingual generalization.

Efficacy of "Thinking" Modes: The integration of Chain-of-Thought (CoT) reasoning significantly boosts performance. For instance, Qwen3-8B (think) outperforms its standard counterpart across complex reasoning tasks and even larger models (*e.g.*, Gemma-3-12B), effectively enabling a smaller model to punch above its weight class.

Possibility of Open-source Benchmark Contamination: A crucial finding is the discrepancy in performance gaps. On established benchmarks like BELEBELE and INCLUDE, the compact Qwen3-14B achieves near-parity with the massive Qwen3-235B. However, on **GaoYao's newly constructed subjective sets** (S-ALPACAEVAL and S-MT-BENCH), a significant gap persists. This suggests that popular open-source benchmarks may have been internalized during training (contamination), whereas GaoYao's fresh, expert-localized data successfully exposes the true gap in generalization ability between compact and flagship models.

3.2.3 Analysis by Language Group

As shown in Fig. 6, by analyzing SOTA LLMs' performances through a geopolitical lens (*i.e.*, aggregating best scores among all LLMs by languages), a persistent "digital divide" is revealed.



(a) Peak performance by geographical region (b) Peak performance by resource popularity

Figure 6: Impact of (a) geography and (b) resource popularity on best performance achieved by LLMs. Vertical dashed lines represent group averages.

Performance on a certain language is strongly correlated with geographic attributes and resource availability: Western European languages consistently score highest, while low-resource languages in South Asia and Africa lag significantly. This hierarchy—High > Medium > Low popularity—is consistent across maximum, median, and minimum scores, underscoring that current multilingual progress is uneven and largely driven by data volume rather than universal linguistic transfer.

3.3 Reliability Analysis of GaoYao

A robust benchmark must strike a balance: it should reflect current user needs (ecological validity) while probing capabilities that users may not yet explicitly request but are essential for advanced intelligence (comprehensive coverage). To assess this, we compared the topic distribution of GaoYao against 140k authentic user queries sampled from *LMSYS Chatbot Arena* [Chiang et al., 2024]. We employed a multilingual tagging model to categorize both datasets and calculated the Tag Semantic Alignment (TSA) score, *i.e.*, the average semantic similarity between tags extracted from two groups.

Table 1 presents the results. We observe a **High Alignment in Utility Tasks**: The *Instruction Following* (TSA 0.89) and *Dialogue* (TSA 0.79) sub-layers show strong correlation with real-world queries, where tags like "IT Technology" and "Creative Writing" dominate. This confirms that our localization of subjective benchmarks accurately captures the core interaction patterns of global users.

Conversely, we see **Low Alignment in Specialized Domains**: Layers like *Math* (0.03) and *Cross-Cultural SA* (0.13) show lower alignment. This is likely because real-world users currently employ LLMs less frequently for complex tasks like advanced logic proofs or nuanced cultural philosophy. However, these "long-tail" capabilities are precisely where SOTA models differentiate themselves. By including these low-TSA but high-value domains, GaoYao prevents overfitting to "average" user behavior and ensures models are also evaluated on the frontier of capability.

	Sub-layer	**TSA**	**Top-10 Frequent Tags**
1	Cross-Cultural (SA)	0.1323	Educational Philosophy, Philosophy/Religion, Values, Higher Education, Language/Writing, Humanities, Workplace Life, Politics
2	Cross-Cultural (SB)	0.4548	Workplace Life, Social Customs, Food & Cooking, IT Technology, Higher Education, Sports, Artifacts, AI
3	Mono-Cultural (CS)	0.2116	Workplace Life, Social Customs, Educational Philosophy, Humanities, Language/Writing, Business Management, Values, Banking
4	Math	0.0319	Applied Math, Logic, Algebra, Business Management, Probability, Operations Research, Education Info
5	Reasoning	0.3382	Philosophy/Religion, Civil Law, Politics, World History, Business Management, Economics, Constitution, Clinical Medicine
6	Knowledge	0.3297	World History, Geography, Business Management, Physiology, Ecology, Economics, Zoology, Politics
7	Instruct Follow	0.8955	IT Technology, Food & Cooking, Language/Writing, Literature, Movies/TV, AI, Music, Games, Business Management
8	Dialogue	0.7980	IT Technology, Literature, Algebra, Language/Writing, Business Management, Probability, Movies/TV, AI, Physics
9	Reading	0.5046	World History, Geography, IT Technology, Politics, Travel, Zoology, Transportation, Social Customs
10	Translation	0.2117	Artifacts, Values, Language/Writing, Regional Characteristics, Zoology, Geography, World History, Politics

Table 1: Tag Semantic Alignment (TSA) scores comparing GaoYao layers with real-world user queries (LMArena). High TSA indicates alignment with common daily usage; low TSA indicates specialized or long-tail capabilities.

4 Conclusion

The evolution of LLMs into global tools demands evaluation frameworks that transcend mono-centric paradigms. In this work, we introduced **GaoYao**, a benchmark systematically designed to assess the "Iceberg" of multilingual intelligence—from surface-level translation to deep cultural nuance. Through a hybrid strategy of *Integration, Expansion, and Generalization*, we constructed a 26-language suite that significantly broadens the scope of multilingual evaluation. Our experiments highlight that while the scaling law persists, the "digital divide" remains a critical barrier for low-resource languages. Furthermore, our reliability analysis confirms that GaoYao successfully balances practical relevance with rigorous capability testing. Notably, the majority of our evaluation experiments have been validated across both MindSpore and PyTorch frameworks, ensuring reproducibility and platform independence. We release GaoYao to the community as a diagnostic compass, guiding the development of models that are not only powerful but truly globally inclusive.

References

Lorin W Anderson and David R Krathwohl. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: complete edition*. Addison Wesley Longman, Inc., 2001.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.44. URL https://aclanthology.org/2024.acl-long.44/.

Lei Chen. *Deep learning and practice with mindspore*. Springer Nature, 2021.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*, 2024.

DeepSeek. Deepseek-v3.1 release. https://api-docs.deepseek.com/news/news250821, 2025.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538, 2022.

Shiwei Guo, Sihang Jiang, Qianxi He, Yanghua Xiao, Jiaqing Liang, Bi Yude, Minggui He, Shimin Tao, and Li Zhang. Do large language models truly understand cross-cultural differences?, 2025. URL https://arxiv.org/abs/2512.07075.

Edward T Hall. *Beyond culture*. Anchor, 1976.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, 2020.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval, 5 2023.

Yang Liu, Meng Xu, Shuo Wang, Liner Yang, Haoyu Wang, Zhenghao Liu, Cunliang Kong, Yun Chen, Maosong Sun, and Erhong Yang. Omgeval: An open multilingual generative evaluation benchmark for large language models. *arXiv preprint arXiv:2402.13524*, 2024.

Brad M Maguth and Gloria Wu. What is the difference between the chinese dragon and its depiction in the west? In *Inquiry-Based Global Learning in the K–12 Social Studies Classroom*, pages 27–43. Routledge, 2020.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, et al. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems*, 37:78104–78146, 2024.

OpenAI. Multilingual massive multitask language understanding (mmmlu). `https://huggingface.co/datasets/openai/MMMLU`, 2024.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

David Pomerenke, Jonas Nothnagel, and Simon Ostermann. The ai language proficiency monitor–tracking the progress of llms on multilingual benchmarks. *arXiv preprint arXiv:2507.08538*, 2025.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Zeming Chen, Mohamed A Haggag, Alfonso Amayuelas, et al. Include: Evaluating multilingual language understanding with regional knowledge. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jonathan Rystrøm, Hannah Rose Kirk, and Scott Hale. Multilingual!= multicultural: Evaluating gaps between multilingual capabilities and cultural alignment in llms. *arXiv preprint arXiv:2502.16534*, 2025.

Edgar H Schein. *Organizational culture and leadership*, volume 2. John Wiley & Sons, 2010.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*, 2023.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Jinghao Zhang, Sihang Jiang, Shiwei Guo, Shisong Chen, Yanghua Xiao, Hongwei Feng, Jiaqing Liang, Minggui HE, Shimin Tao, and Hongxia Ma. Culturescope: A dimensional lens for probing cultural understanding in llms. *arXiv preprint arXiv:2509.16188*, 2025.

Zhihan Zhang, Dong-Ho Lee, Yuwei Fang, Wenhao Yu, Mengzhao Jia, Meng Jiang, and Francesco Barbieri. Plug: Leveraging pivot language in cross-lingual instruction tuning. In *Proceedings of the 62th Annual Meeting of the Association for Computational Linguistics*. ACL, 2024.

Qiguang Zhao. *A study of dragonology, East and West*. University of Massachusetts Amherst, 1988.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

Ruoxin Zhu, Yujing Wang, Diao Lin, Michael Jendryke, Mingxia Xie, Jianzhong Guo, and Liqiu Meng. Exploring the rich-club characteristic in internal migration: Evidence from chinese chunyun migration. *Cities*, 114:103198, 2021.

A Prompts

A.1 Prompt for Dummy Option Generation

> Provide {3 - n} dummy option(s) that makes sense to be the answer(s) of the given question, and has to exist in real-life (non-fiction), but is totally different from the given answers without any explanation. Make sure that the options are different from each other, and cannot be an answer from any country. Provide as JSON format: {"dummy_options":[]}

A.2 Prompt for MCQ Rephrasing

> I will give you a multiple-choice cultural question. Your task: refine the wording of both the stem and the option as I required. Goal: raise the overall difficulty and enrich the phrasing while keeping the underlying concepts intact. Recommended techniques: paraphrasing, expansion, morphological variation, idioms and figurative language, voice alternation (active $\leftrightarrow$ passive), syntactic restructuring, etc. Requirements: 1. You must not change the semantic meaning of the original stem and options. 2. Do not alter any entities or proper nouns (e.g., personal names, company names, sport names, countries/region names, festival names). 3. Do not add a country or regional name (or its adjective) to the options unless the answer itself is a country or region. 4. Ensure the index of the original correct answer stays the same. 5. Use English. 6. Keep capitalization consistent across the options; capitalizing the first letter of each option is recommended. 7. Follow the specified output format exactly, including JSON punctuation. Output format: [...]

B Detailed Information of Languages Supported by GaoYao

Table 2 presents the languages in our dataset, mapping full names to short codes, and detailing their geographical and resource-level information. Resource levels are determined using the Joshi et al. [2020] taxonomy, which rates languages from a set of 2,485 on a scale reflecting resource availability. According to this scale, we classify languages scoring 5 as high-resource (*e.g., Chinese*), a score of 4 as mid-resource (*e.g.*, Turkish), and a score of $\leq$3 as low-resource.

C Detailed Experimental Setups

Language	Code	Geographical Group	Resource Level
English	EN	Western Europe	High Resource
French	FR	Western Europe	High Resource
German	DE	Western Europe	High Resource
Spanish	ES	Western Europe	High Resource
Dutch	NL	Western Europe	Medium Resource
Italian	IT	Western Europe	Medium Resource
Portuguese	PT	Western Europe	Medium Resource
Greek	EL	Western Europe	Low Resource
Czech	CS	Eastern Europe	Medium Resource
Polish	PL	Eastern Europe	Medium Resource
Russian	RU	Eastern Europe	Medium Resource
Chinese	ZH	East Asia & Southeast Asia	High Resource
Japanese	JA	East Asia & Southeast Asia	High Resource
Korean	KO	East Asia & Southeast Asia	Medium Resource
Vietnamese	VI	East Asia & Southeast Asia	Medium Resource
Indonesian	ID	East Asia & Southeast Asia	Low Resource
Malay	MS	East Asia & Southeast Asia	Low Resource
Tagalog	TL	East Asia & Southeast Asia	Low Resource
Thai	TH	East Asia & Southeast Asia	Low Resource
Arabic	AR	Middle East & Africa	High Resource
Turkish	TR	Middle East & Africa	Medium Resource
Swahili	SW	Middle East & Africa	Low Resource
Yoruba	YO	Middle East & Africa	Low Resource
Hindi	HI	South Asia	Medium Resource
Bengali	BN	South Asia	Low Resource
Telugu	TE	South Asia	Low Resource

Table 2: Language mapping: codes, geographical groups, and resource levels.

Data source	Task Type	Eval. Type	Metric	Judge Model	Ref. Source	Calculation Method
S-AlpacaEval	QA	Subj.	Win Rate	Deep Seek V3.1	Qwen3-235B	1. Judge compares candidate vs. reference (correctness, richness, comprehensiveness, etc.); 2. Win Rate $= \frac{\#\text{win}+\#\text{tie}/2}{\#\text{all}}$
Belebele	MCQ	Obj.	Accuracy	Rule-based	Human (Open Source)	1. Reading comprehension, 4-option regex match (A-D); 2. Accuracy $= \frac{\#\text{correct}}{\#\text{all}}$
INCLUDE	MCQ	Obj.	Accuracy	Rule-based	Human (Open Source)	1. Encyclopedic knowledge, 4-option regex match (A-D); 2. Accuracy $= \frac{\#\text{correct}}{\#\text{all}}$
M3Exam	MCQ	Obj.	Accuracy	Rule-based	Human (Open Source)	1. Encyclopedic knowledge, 4-option regex match (1-4, A-D); 2. Accuracy $= \frac{\#\text{correct}}{\#\text{all}}$
SuperBLEnD	MCQ	Obj.	Accuracy	Rule-based	Human and LLM Hybrid	1. Regional culture knowledge, 4-option regex match (A-D); 2. Accuracy $= \frac{\#\text{correct}}{\#\text{all}}$
MGSM	Math	Obj.	Accuracy	Rule-based	Human (Open Source)	1. Math reasoning, regex match for integer answers; 2. Accuracy $= \frac{\#\text{correct}}{\#\text{all}}$
MMMLU	MCQ	Obj.	Accuracy	Rule-based	Human and LLM Hybrid (Open Source)	1. Knowledge QA, 4-option regex match (A-D). Uses LLM if regex fails; 2. Accuracy $= \frac{\#\text{correct}}{\#\text{all}}$
Flores-101	Translation	Obj.	Comet	wmt20-comet-da	Human Translated Wiki	1. Translation task; 2. Comet Score
SAGE	MCQ + T/F + QA	Subj. + Obj.	Mixed	Deep Seek V3.1	Qwen3 max	1. MCQ and T/F uses accuracy as score; 2. QA use LLM to recognize culture points mentioned in the answer; 3. weighed sum score is used as final score.
CultureScope	MCQ + T/F + QA	Subj. + Obj.	Mixed	Deep Seek V3.1	Human Expert	1. MCQ and T/F uses accuracy as score; 2. QA use LLM to recognize culture points mentioned in the answer; 3. weighed sum score is used as final score.
S-MT-Bench	QA	Subj.	Win Rate	Deep Seek V3.1	Qwen3-235B	1. Judge comparison (multi-turn averaged); 2. Win Rate $= \frac{\#\text{win}+\#\text{tie}/2}{\#\text{all}}$

Table 3: Summary of evaluation methodologies. Task types include Multiple Choice Questions (MCQ), True/False (T/F), and Open-ended Q&A (QA). Evaluation types distinguish between subjective (Subj.) LLM-judged approaches and objective (Obj.) rule-based approaches. MGSM only generate Integer as final answer.

Model	Model Version	Resource Address
Flagship Open-Source Models		
openPangu-Ultra-MoE-718B-V1.1	openPangu-Ultra-MoE-718B-V1.1	https://ai.gitcode.com/ascend-tribe/openPangu-Ultra-MoE-718B-V1.1
DeepSeek-V3.1	DeepSeek-v3.1-250821	https://huggingface.co/deepseek-ai/DeepSeek-V3.1
Qwen3-235B-A22B	Qwen3-235B-A22B-Instruct-2507	https://huggingface.co/Qwen/Qwen3-235B-A22B-Instruct-2507
Qwen3-VL-235B-A22B	Qwen3-VL-235B-A22B-Instruct	https://huggingface.co/Qwen/Qwen3-VL-235B-A22B-Instruct
DeepSeek-R1	DeepSeek-R1	https://huggingface.co/deepseek-ai/DeepSeek-R1
Llama-3.1-405B	Llama-3.1-405B-Instruct	https://huggingface.co/meta-llama/Llama-3.1-405B-Instruct
Closed-Source Commercial Models		
Doubao-seed-1.6	doubao-seed-1-6-250615	https://www.doubao.com/chat/
Doubao-seed-1.6 (thinking)	doubao-seed-1-6-thinking-250715	https://www.doubao.com/chat/
Qwen-max	Qwen-max	https://chat.qwen.ai/
GLM-4.6	GLM-4.6	https://chatglm.cn
Kimi-k2	kimi-k2-250711	https://www.kimi.com
Compact Models (<20B)		
Qwen3-14B (think)	Qwen3-14B	https://huggingface.co/Qwen/Qwen3-14B
Qwen3-8B	Qwen3-8B	https://huggingface.co/Qwen/Qwen3-8B
DeepSeek-R1-14B (think)	DeepSeek-R1-Distill-Qwen-14B	https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-14B
Gemma-3-12B-IT	gemma-3-12b-it	https://huggingface.co/google/gemma-3-12b-it
Llama-3.1-8B	Llama-3.1-8B-Instruct	https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
Ministral-8B-Instruct	Ministral-8B-Instruct-2410	https://huggingface.co/mistralai/Ministral-8B-Instruct-2410

Table 4: Detailed specifications of models evaluated in GaoYao.