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

Evaluating the multilingual and multicultural capabilities of Large Language Models (LLMs) is essential for their global utility. However, current benchmarks face three critical limitations: (1) fragmented evaluation dimensions that often neglect deep cultural nuances; (2) insufficient language coverage in subjective tasks relying on low-quality machine translation; and (3) shallow analysis that lacks diagnostic depth beyond simple rankings. To address these, we introduce **GaoYao**[1], a comprehensive benchmark with 182.3k samples, 26 languages and 51 nations/areas. First, GaoYao proposes a unified framework categorizing evaluation tasks into three cultural layers (General Multilingual, Cross-cultural, Monocultural) and nine cognitive sub-layers. Second, we achieve native-quality expansion by leveraging experts to rigorously localize subjective benchmarks into 19 languages and synthesizing cross-cultural test sets for 34 cultures, surpassing prior coverage by up to 111%. Third, we conduct an in-depth diagnostic analysis on 20+ flagship and compact LLMs. Our findings reveal significant geographical performance disparities and distinct gaps between tasks, offering a reliable map for future work. We release the benchmark[2].

1 Introduction

As Large Language Models (LLMs) increasingly serve a global user base, the ability to process diverse languages and navigate complex cultural contexts has become a critical measure of their inclusivity. However, the current landscape of multilingual evaluation is fraught with challenges that hinder a holistic understanding of model performance:

(1) Lack of Systematicity and Cultural Neglect. Many prominent benchmarks focus narrowly on single specific facets of language ability, such as factual knowledge (Romanou et al., 2025) or reading comprehension (Bandarkar et al., 2024). Consequently, they often overlook the deeper capabilities a model should possess (*e.g.*, cultural sensitivity), treating multilingualism merely as isolated evaluation points rather than interconnected dimensions rooted from cultural and cognitive sources.

(2) Limited Language Coverage and Quality in Subjective Tasks. Subjective tasks (*i.e.*, answers are open-ended) such as instruction following and multi-turn dialogue are predominantly assessed in English (Li et al., 2023; Zheng et al., 2023). Existing multilingual extensions often rely on automated machine translation (MT) or cover only a handful of languages (Zhang et al., 2024; Liu et al., 2024). This reliance on MT introduces "translationese" and fails to reflect native tongues, which can be trivial in objective tasks (*e.g.*, true/false) but is especially harmful for subjective evaluation.

(3) Lack of In-Depth Diagnostic Analysis. Existing studies often stop at superficial leaderboard rankings (Pomerenke et al., 2025; Liu et al., 2024), failing to reveal implications under performance variance. There is a scarcity of insights regarding how performance correlates with geographical regions, task types, or model architectures, leaving potential challenges and gaps untouched.

To address these challenges, we introduce **GaoYao**, a multilingual and multicultural benchmark emphasizing systematicity, authenticity, and analytical depth, which features three aspects:

(1) A Systematic Evaluation Framework. Grounded in cultural theory (Hall, 1976) and cognitive taxonomy (Anderson and Krathwohl, 2001), we propose a unified evaluation matrix. This framework categorizes capabilities into three layers: *General Multilingual Abilities* (universal concepts), *Cross-cultural Abilities* (culturally shared concepts with variance), and *Monocultural Abilities* (unique concepts in cultures). These are further expanded

[*] Equal contribution.

[1] GaoYao is derived from Chinese mythology, where he served as the first judicial officer, symbolizing fairness and comprehensiveness.

[2] https://github.com/lunyiliu/GaoYao

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 21364–21384

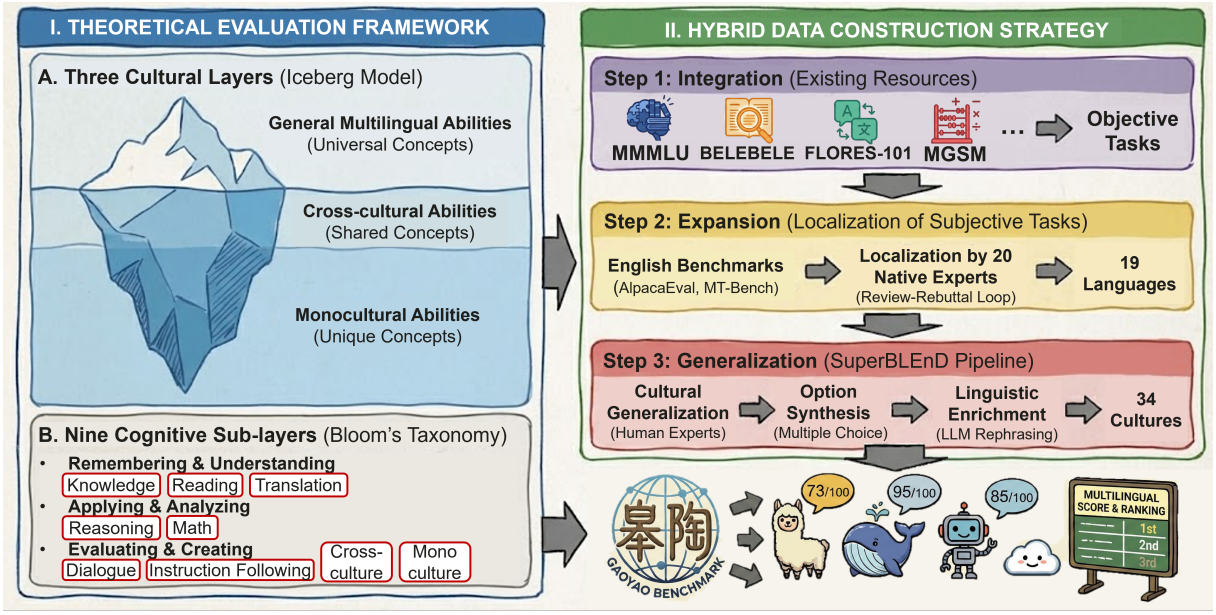


Figure 1: Illustration on design and construction of GaoYao. The benchmark is grounded in theoretical models of culture and cognition, and constructed through a hybrid strategy of integration, expansion and generalization.

into nine cognitive sub-layers, ranging from knowledge retention to creative writing.

(2) Native-Quality Data Expansion. We address the data scarcity by leveraging a team of multilingual experts to meticulously localize English evaluation sets in two critical sub-layers into 19 languages, ensuring native-level quality compared to machine-translated alternatives. Additionally, we generalize a cultural evaluation set to cover 34 cultures through a novel expert-verified synthesis pipeline. In Fig. 7, our three curated test sets are able to better reflect capability stratification of LLMs due to the native-level data quality.

(3) In-Depth Diagnostic Findings. We conduct a tiered evaluation of representative SOTA models and go beyond simple rankings. Our analysis reveals the severe "digital divide" across regions and the performance gap between mature and frontier tasks. Synthesizing these findings, we propose a *meta-finding* to guide the community: recommending strategic deployment methods for efficient model usage and advocating for equitable data construction to bridge the capability gaps.

Our contributions can be summarized as:

- We propose a systematic evaluation benchmark with 23.3M tokens, three cultural layers and nine cognitive sub-layers, addressing the fragmented nature of existing benchmarks.
- We expand critical instruction-following and dialogue evaluations to 19 languages and generalize cultural evaluation sets to 34 cultures via a rigorous human-in-the-loop method, filling the blank with high data quality and enabling better identification of LLM abilities.
- We provide a comprehensive capability landscape of existing multilingual LLMs through deep analysis, revealing several key insights which guide LLM usage and development.

In addition, we release all test sets and evaluation codes, providing the community with a reliable compass for future work on multilingual LLMs.

2 Methodology

As shown 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, coverage and 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

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 tasks into three major layers representing cultural deepening levels:

- **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 monsters while the eastern dragon (or loong) is revered as an auspicious symbol (Zhao, 1988). An LLM must discern these subtle 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 special greeting etiquette in India (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, the task design of GaoYao encompasses nine distinct sub-layers to ensure a rigorous assessment:

- **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:** This highest cognitive level encompasses: (1) Creative Tasks: *Instruction Following* and *Multi-turn Dialogue*, which demand creative writing to satisfy complex, open-ended user intents; (2) Evaluative Tasks: The advanced *Cross-cultural* and *Monocultural* assessments. Unlike simple factual retrieval, these tasks require the model to evaluate social nuances, discern cultural appropriateness among highly plausible distractors, and make value judgments aligned with cultural norms (Rystrøm et al., 2025).

2.2 Construction of GaoYao Benchmark

Guided by the theoretical 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. These datasets are mapped to our sub-layers as follows:

(1) *Knowledge & Reasoning*: Given their coverage on factual knowledge spanning various subjects from elementary-level knowledge up to advanced professional subjects, both INCLUDE (Romanou et al., 2025) and MMMLU (OpenAI, 2024) are integrated. Compared with INCLUDE, MMMLU focuses more on 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 cultures (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 (Li et al., 2023), validated by over 20k human judgments for general instruction following, and MT-BENCH (Zheng et al., 2023), 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 the 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.

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, expanding such coverage presents a unique challenge: direct translation retains source-culture concepts, while manual creation is costly. To address this, we generalized BLEND (Myung et al., 2024) into SUPERBLEND, expanding coverage from 16 to 34 cultures via a three-stage semi-automated pipeline incorporating rigorous human verification. SUPERBLEND focuses on evaluating understanding of cultural differences regarding everyday concepts such as festivals, food and sports. The pipelines are as follows (full annotations and technical details are in Appendix E):

Stage 1: Cultural Generalization. We curated a subset of high-quality templates from BLEND (inheriting answers for the original 16 cultures) and recruited native experts to provide authentic answers for 18 additional cultures based on lived experience. Answers underwent strict manual verification to eliminate invalid or toxic content (discarding ∼41.1% of raw data), ensuring high-quality ground truth even for questions with multiple valid answers (*e.g.*, accepting both "beer" and "carbonated drinks" for Malaysian nightclubs).

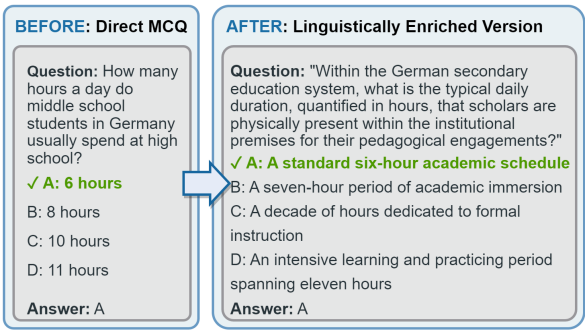


Figure 2: An example of the linguistic enrichment process (stage 3), which increases complexity of MCQs without altering the underlying cultural fact.

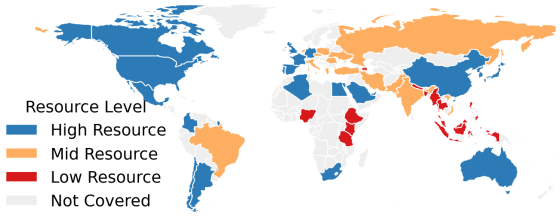


Figure 3: The language and culture coverage on the world map. Colors indicate resource popularity levels.

Stage 2: Option Synthesis. To enhance diversity, verified Q&A pairs from stage 1 were converted into multiple-choice questions (MCQs) by combining target answers with distractors from other cultures or plausible LLM-generated "dummy options". Options underwent strict verification to exclude low-quality cases such as hierarchical conflicts (*e.g.*, rejecting "Pepsi" as a distractor if the answer is "beer", as it falls under the valid category of "carbonated drinks").

Stage 3: Linguistic Enrichment. To enhance difficulty and prevent simple pattern matching, we utilized an LLM to rephrase question stems and options via techniques like syntactic restructuring and voice alternation. As shown in Fig. 2, this process ensures the benchmark tests deep cultural reasoning rather than superficial keyword recognition.

3 Experiment

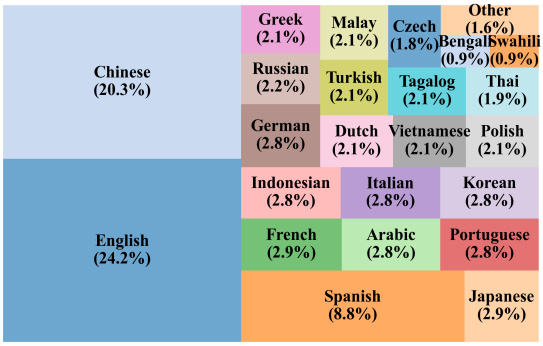
3.1 Experimental Setups

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.*, DeepSeek-V3.1 (DeepSeek, 2025)) inferred on standardized NPU computation nodes, and proprietary commercial models (*e.g.*, GPT-5 (OpenAI, 2025)) accessed via official APIs. The specific model versions and resource addresses are in Table 7. We make sure all evaluated LLMs are post-trained versions (*e.g.*, "instruct" or "chat" versions) with "thinking" disabled (except Fig. 8). Following Yang et al. (2025), we adopted only a random 10% subset of MMMLU due to its unproportionate volume. Section 3.1.1 and Section 3.1.2 further illustrates the setups. See a reliability analysis of GaoYao in Appendix C.

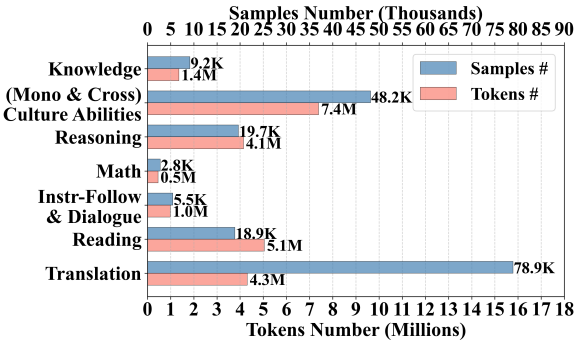
3.1.1 Statistics of Test Sets in GaoYao

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 4 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 relatively balanced, with each remaining language constituting roughly 1%-3% of the total volume in sample level. According to Joshi et al. (2020), the languages include nine low-resource and ten mid-resource varieties (See Appendix B).

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



(a) Sample distribution by language



(b) Distribution by evaluation dimensions

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

	General Multilinguality							Culture Abilities		
openPangu-Ultra-MoE-718B-V1.1	92.11 (1)	83.92 (2)	74.71 (7)	41.77 (4)	33.83 (7)	88.79 (8)	88.36 (1)	93.84 (1)	69.34 (5)	96.51 (3)
Qwen3-235B-A22B	90.47 (5)	77.76 (7)	77.61 (4)	50.01 (3)	50.49 (5)	93.44 (4)	88.32 (2)	93.22 (3)	69.31 (6)	98.66 (1)
Qwen3-VL-235B-A22B	92.0 (2)	79.49 (6)	78.79 (3)	58.72 (2)	61.46 (2)	93.52 (3)	88.19 (4)	93.2 (4)	70.87 (2)	97.46 (2)
DeepSeek-V3.1	91.49 (4)	81.16 (4)	76.2 (6)	27.17 (7)	35.8 (6)	89.25 (7)	88.07 (5)	92.09 (7)	67.12 (7)	96.03 (4)
DeepSeek-R1	92.0 (2)	81.97 (3)	81.7 (2)	76.44 (1)	77.14 (1)	93.88 (2)	50.76 (8)	93.09 (5)	70.58 (3)	93.76 (6)
Llama-3.1-405B	89.45 (8)	76.3 (8)	71.37 (8)	5.48 (8)	14.86 (8)	92.27 (6)	84.56 (6)	87.51 (8)	65.79 (8)	79.56 (8)
GLM-4.6	90.22 (6)	84.15 (1)	81.8 (1)	37.67 (6)	57.51 (3)	94.09 (1)	70.88 (7)	93.74 (2)	72.95 (1)	96.02 (5)
Kimi-k2	89.67 (7)	80.15 (5)	76.99 (5)	39.21 (5)	51.32 (4)	92.72 (5)	88.22 (3)	92.44 (6)	70.07 (4)	91.63 (7)
	Math	Reason-ing	Know-ledge	Instruct Follow	Dialogue	Reading	Transla-tion	Cross Cul ture (SA)	Cross Cul ture (SB)	Mono Cul ture (CS)

(a) Open-Source Models

	General Multilinguality							Culture Abilities		
Doubao-Seed-1.6	92.69 (3)	83.11 (6)	80.51 (5)	62.92 (2)	62.34 (1)	94.24 (4)	86.19 (7)	91.11 (5)	70.52 (8)	95.69 (1)
Qwen-max	87.96 (9)	72.21 (9)	73.53 (9)	4.06 (9)	50.0 (5)	90.92 (9)	85.42 (9)	87.66 (9)	68.71 (9)	92.38 (6)
Gemini-2.5-Pro	92.07 (6)	88.77 (2)	86.73 (1)	74.21 (1)	62.23 (2)	95.35 (1)	88.27 (1)	95.94 (1)	74.87 (2)	95.2 (3)
Claude-Sonnet-4.5	93.64 (1)	86.66 (3)	83.5 (3)	27.56 (4)	50.99 (4)	94.61 (3)	87.82 (3)	89.08 (7)	73.72 (4)	95.11 (4)
Grok-3	91.67 (7)	81.18 (7)	78.19 (8)	15.17 (6)	35.64 (7)	93.21 (7)	87.73 (4)	94.34 (3)	74.53 (3)	95.68 (2)
o3	92.11 (5)	89.22 (1)	85.43 (2)	29.59 (3)	55.59 (3)	94.94 (2)	88.04 (2)	91.16 (4)	75.78 (1)	84.36 (8)
o4-mini	93.53 (2)	84.25 (4)	81.87 (4)	16.96 (5)	36.84 (6)	94.12 (5)	87.72 (5)	89.04 (8)	71.82 (7)	89.04 (7)
GPT-5-chat	92.44 (4)	83.56 (5)	79.3 (6)	9.65 (7)	35.31 (8)	93.54 (6)	85.87 (8)	95.35 (2)	72.96 (5)	93.33 (5)
GPT-4o	90.15 (8)	79.46 (8)	79.23 (7)	7.33 (8)	19.85 (9)	92.37 (8)	87.46 (6)	90.2 (6)	72.61 (6)	83.15 (9)
	Math	Reason-ing	Know-ledge	Instruct Follow	Dialogue	Reading	Transla-tion	Cross Cul ture (SA)	Cross Cul ture (SB)	Mono Cul ture (CS)

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

	General Multilinguality							Culture Abilities		
Qwen3-14B	84.11(1)	66.1(1)	68.26(1)	10.55(2)	19.52(1)	89.39(1)	87.3(1)	92.4(1)	68.29(1)	92.72(1)
Qwen3-8B	79.6(3)	61.55(3)	64.08(3)	9.87(3)	16.89(2)	84.81(3)	86.47(2)	87.81(2)	57.54(3)	91.61(2)
Gemma-3-12B-IT	79.75(2)	63.2(2)	64.93(2)	13.43(1)	13.87(3)	88.77(2)	59.23(5)	85.82(3)	60.43(2)	89.97(3)
Llama-3.1-8B	68.15(4)	49.71(4)	52.41(5)	3.51(5)	10.42(5)	75.76(6)	83.49(3)	80.75(5)	55.54(4)	75.34(6)
Llama-3-8B	58.04(6)	45.56(6)	53.0(4)	4.36(4)	11.95(4)	79.49(4)	58.12(6)	83.13(4)	52.91(5)	82.22(4)
Minstral-8B-Instruct	62.04(5)	46.36(5)	51.24(6)	3.47(6)	8.99(6)	79.16(5)	75.51(4)	80.56(6)	49.19(6)	76.27(5)
	Math	Reason-ing	Know-ledge	Instruct Follow	Dialogue	Reading	Transla-tion	Cross Cul ture (SA)	Cross Cul ture (SB)	Mono Cul ture (CS)

(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.

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.

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 6):

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 compares 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 (Yang et al., 2025) as the reference anchor.

All scores (*e.g.*, accuracy, win rate) are displayed at the scale of 0-100 for clearer viewing. The results are aggregated along specific axes, *e.g.*, *Task Dimension* (averaging across all languages for a specific evaluation sub-layer) for Finding 1&3 and *Language Dimension* (averaging across all sub-layers for a specific language) for Finding 2.

3.2 Results and Findings

Finding 1: Performance Differentiation Among Flagship and Compact 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 multilingual capability profiles rather than a uniform dominance:

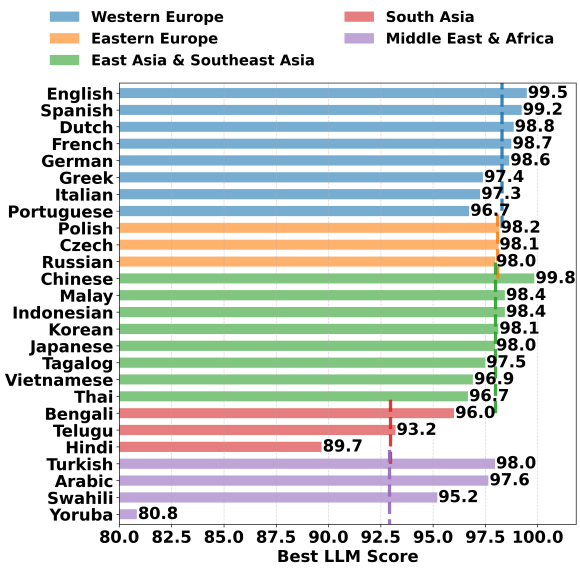


Figure 6: Impact of geography on **best** performance achieved by LLMs across languages. Vertical dashed lines represent group averages.

- *Logic & Culture Specialist:* openPangu-Ultra-MoE-718B-V1.1 (Ascend Tribe, 2025) exhibits exceptional strength in *Math* (#1) and *Reasoning* (#2) among open-source LLMs, while simultaneously securing among top ranks in the *Cross-culture* sub-layer. This correlation suggests its rigorous logical training may facilitate understanding of complex cultural frameworks.

- *Knowledge Heavyweights:* Gemini-2.5-Pro (Comanici et al., 2025) demonstrates dominance in *Knowledge*, *Reading* and *Translation*, suggesting a pre-training corpus with extensive informational breadth.

- *Interaction Specialists:* DeepSeek-R1 (DeepSeek-AI, 2025) leads both open-source and closed-source models in *Instruction Following* and *Dialogue*, reflecting a post-training strategy optimized for conversational utility and complex user constraints.

Fig. 5(c) illustrates the performance of compact models ($<$ 20B parameters). We observed a possibility of saturation for open-source benchmark: A crucial finding is the discrepancy in performance gaps. On established benchmarks like BELEBELE and INCLUDE, the compact Qwen3-14B (Yang et al., 2025) 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 some popular open-source benchmarks may have become insensitive in identifying multilingual capabilities (a phenomenon called benchmark saturation) (Akhtar et al., 2026), whereas GaoYao's fresh, expert-localized data exposes the true gap between compact and flagship models. Finding 3 further investigates it.

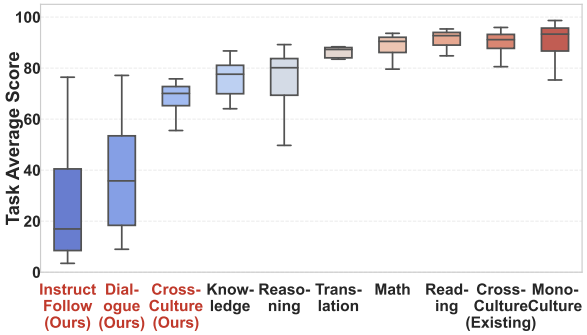


Figure 7: Distribution of model performance across tasks (*i.e.*, sub-layers) using standard box plot. Height of boxes and whiskers indicate the performance divergence among models, while the horizontal line is the median score. Datasets constructed in this work are in **Red**.

Finding 2: Digital Divide Among Language Geography and Resource-Levels As shown in Fig. 6 and Fig. 9, by analyzing LLMs' best performances through a geopolitical and resource-level lens (*i.e.*, aggregating highest scores among all LLMs by languages), a persistent "digital divide" is revealed. 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 is consistent with resource levels in Fig. 9: High > Medium > Low popularity across maximum, mean, and minimum scores, underscoring that current multilingual progress is uneven and largely driven by data volume rather than universal linguistic transfer.

Finding 3: Capability Stratification Between Solved and Frontier Tasks Fig. 7 presents a boxplot analysis of model performance distributions, revealing a severe stratification in current LLM capabilities for different tasks (*i.e.*, sub-layers). Objective tasks such as *Reading* and *Math* exhibit high median scores ($>$ 85) with compressed boxes, indicating a closed gap between flagship and compact models due to their standardized patterns. In contrast, subjective tasks (especially for *Instruc-*

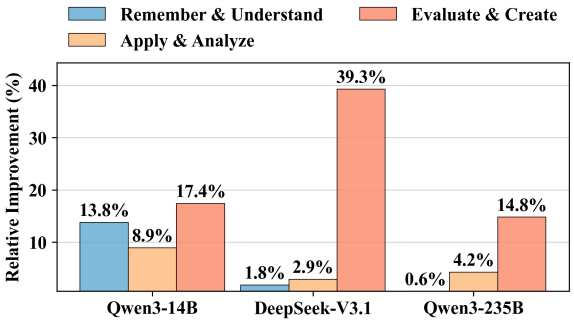


Figure 8: Multilingual performance gain by "thinking" mode ($\frac{\text{Think}-\text{Base}}{\text{Base}} * 100\%$) across Bloom cognitive layers.

tion Following and *Dialogue* with expert-localized datasets in GaoYao) display significantly lower medians and elongated interquartile ranges (box and whisker heights). This high variance confirms that these tasks serve as high-sensitivity discriminators due to their advanced requirements on creating and human-likeness, effectively exposing the capability frontier where flagship models significantly outperform average open-source models.

In addition, a comparison within cultural tasks demonstrates the value of the SUPERBLEND dataset. While the two existing cultural test sets (*i.e.*, SAGE and CULTURESCOPE) show signs of saturation (medians ≈ 90), our synthesized dataset for the cross-cultural layer reveals a significant drop in median score and relatively wide dispersion. This proves that the construction strategy in Section 2.2.3 successfully elevates the challenge from simple knowledge retrieval to rigorous cultural reasoning on authentic experiences, leading to a more precise test set for cultural capabilities.

Finding 4: Uneven Gain by "Thinking" in Multilingual Tasks The paradigm of inference-time reasoning (or "thinking") has been proven effective in massive fields (DeepSeek-AI, 2025; Liu et al., 2025a; He et al., 2025). In Fig. 8, we investigated the impact of "thinking" in multilingual context. By aggregating task scores across Bloom's cognitive taxonomy (as described in Section 2.1), two divergent behaviors are revealed:

(1) *Selective Gain for Flagships.* For flagship models (DeepSeek-V3.1, Qwen3-235B), the "Thinking" mode acts as a specialized tool. In the *Remembering & Understanding* layer (*e.g.*, Translation, Knowledge), gains are marginal, suggesting the bottleneck is mainly multilingual knowledge for retrieval-heavy tasks. However, in the *Evaluating & Creating* layer (*e.g.*, Instruction Following), we observe significant gains, probably due to a more comprehensive considerations of the constraints in users' instructions.

(2) *Universal Gain for Compact Models.* In contrast, the compact Qwen3-14B benefits universally, achieving significant gains even in basic understanding tasks. This suggests that "Thinking" effectively compensates for the limited parameter capacity of smaller models, allowing them to punch above their weight.

Meta-Finding: From Benchmarking to Guidance Transcending individual metrics, our findings (F1-F4) combine into a meta-finding that guides multilingual LLM utility and development:

Strategic Deployment (Usage). There is no "one-fits-all" model. Users should choose flagship models wisely based on their features (the interaction specialist or the knowledge heavyweights, per F1) and adopt a dynamic strategy for "thinking" mode (F4): deploy compact models with "thinking" enabled for better performance in resource-constrained inference, while reserving flagship models for complex creative tasks to balance cost and performance.

Equitable Construction (Development). The geographic performance cliffs (F2) and task stratification (F3) reveal the training data gap both for low-resource areas and highly subjective tasks. Future data development must pivot from English-centric translation to authentic regional curation. We recommend leveraging the *human-in-the-loop generalization* pipeline (in Section 2.2.3) to efficiently fill the training data voids, ensuring both language equity and authenticity.

4 Discussion

4.1 Necessity of the Hierarchical Framework

To further examine whether the three-layer framework reveals meaningful capability distinctions beyond an aggregation of existing benchmarks, we conduct a rank-based transfer analysis across the three major layers in GaoYao: General multilingual, Cross-cultural, and Monocultural abilities. Following the model set in Fig. 5, we rank the 23 evaluated models by their average scores within each layer and compute Spearman's rank correlation between the General multilingual ranking and the rankings of the two deeper cultural layers.

The results show only modest transfer from general multilingual capability to deeper cultural capabilities: the correlation is higher for Cross-cultural

Model	General	Cross-cultural	Monocultural
Gemini-2.5-Pro	#1	#1	#8
Doubao-Seed-1.6	#2	#14	#6
Qwen3-235B-A22B	#9	#11	#1
DeepSeek-V3.1	#15	#16	#4

Table 1: Representative model rankings across the three major layers in GaoYao. Higher general multilingual ranking does not necessarily transfer to deeper cultural layers.

tasks (Spearman's $\rho = 0.74$) but drops for Monocultural tasks ($\rho = 0.61$). This decline is important because monocultural evaluation requires models to recognize culturally unique concepts and norms rather than solve language-general problems. As shown in Table 1, Gemini-2.5-Pro ranks first in the General Multilingual and Cross-cultural layers but drops to eighth in the Monocultural layer, while DeepSeek-V3.1 rises from fifteenth in General Multilingual to fourth in Monocultural. Similarly, Qwen3-235B-A22B ranks ninth in General Multilingual but first in Monocultural. These rank shifts indicate that strong surface-level fluency or general reasoning does not guarantee deep cultural nuance.

This analysis validates the necessity of GaoYao's hierarchical framework. If all tasks were collapsed into a single multilingual score, these capability decouplings would be obscured. By separating General Multilingual, Cross-cultural, and Monocultural layers, GaoYao can diagnose where a model's multilingual competence genuinely transfers and where culturally grounded evaluation remains a distinct frontier.

4.2 Necessity of Linguistic Enrichment in SUPERBLEND

We also conduct an ablation study to verify whether the option synthesis and linguistic enrichment strategies in SUPERBLEND make the benchmark more discriminative. We compare model accuracy on the original BLEND setting without our enrichment against SUPERBLEND on the same 16 overlapping cultures. This controls for culture coverage, so the main difference is whether our synthesized distractors and enriched phrasings are applied.

Table 2 shows that all three representative models experience accuracy drops after applying our synthesis and enrichment strategies. The decrease is relatively modest for larger flagship models, such as Qwen3-235B (-4.51) and GPT-5-chat (-8.07), but significantly more substantial for the compact

Model	BLEND	SUPERBLEND	Δ
Qwen3-235B-A22B	72.57	68.06	**-4.51**
Qwen3-8B	78.06	57.25	**-20.81**
GPT-5-chat	78.45	70.38	**-8.07**

Table 2: Ablation of SUPERBLEND on the 16 cultures overlapping with BLEND. Scores are average accuracy.

Qwen3-8B (-20.81). This pattern suggests that the enriched benchmark removes shortcuts based on surface forms and keyword associations, forcing models to resolve the underlying cultural context among plausible distractors.

The ablation also restores a more expected capability hierarchy. Without enrichment, Qwen3-8B unexpectedly outperforms Qwen3-235B-A22B on the original BLEND subset (78.06 vs. 72.57), suggesting that the original format may permit shortcut exploitation. After enrichment, Qwen3-235B-A22B becomes clearly more robust than Qwen3-8B (68.06 vs. 57.25). Therefore, linguistic enrichment is not merely a stylistic transformation; it improves diagnostic validity by making SUPERBLEND better distinguish culturally grounded reasoning from shallow pattern matching.

5 Conclusion

In this work, we introduced GaoYao, a holistic benchmark designed to map the full spectrum of multilingual intelligence. Unlike fragmented prior efforts, GaoYao establishes a unified framework covering three cultural layers and nine cognitive dimensions. Through a rigorous *human-in-the-loop* construction pipeline, we addressed the critical scarcity of high-quality resources for subjective and cultural tasks, proving that authentic evaluation requires native expertise rather than automated translation. GaoYao stands as a robust alternative to English-centric evaluations, guiding the field to move beyond surface-level linguistic fluency towards deep cultural alignment and equitable global access.

Future work include expanding coverage of domains (*e.g.*, agent abilities) and languages, and developing a dynamic leaderboard to keep up with the latest iterations of models.

6 Limitations

Despite our comprehensive efforts, GaoYao has several limitations that outline future directions:

(1) Domain and Task Coverage: Currently, GaoYao focuses primarily on general multilingual and multicultural capabilities. We do not currently cover specialized vertical domains (*e.g.*, legal, medical, financial) or agentic capabilities (*e.g.*, tool use, API calling) in multilingual contexts. However, we argue that a robust general-purpose multilingual foundation is a prerequisite for these specialized abilities. Constructing high-quality benchmarks for such specific domains requires advanced expertise that falls outside the scope of this foundational work.

(2) Static Nature of Benchmarking: The landscape of LLMs evolves at an unprecedented pace, and a static publication inevitably lags behind the release of the very latest models. To address this, we have open-sourced all test sets and evaluation code, ensuring full transparency and enabling third-party model developers to easily verify their own systems against GaoYao. Furthermore, we plan to launch and maintain a dynamic online leaderboard to continuously track the community's progress.

(3) Scalability of Human-in-the-loop Pipeline: Our insistence on native expert curation and verification ensures high data quality but inherently limits scalability compared to fully automated pipelines. Expanding to hundreds of low-resource languages using this rigorous standard is resource-intensive. Yet, we believe that in the current era of ubiquitous machine-generated content, establishing a high-quality, human-verified gold standard for a representative set of languages is more critical than broad but low-quality coverage.

(4) Task and Language Imbalance: GaoYao prioritizes high-quality and expert-verified coverage over perfectly uniform coverage across all sublayers. As a result, some integrated resources naturally differ in language scope; for example, MGSM covers 10 languages, while cultural resources such as SAGE and CULTURESCOPE are limited to 2 languages/cultures. This imbalance reflects the current scarcity of reliable multilingual and multicultural evaluation resources rather than an assumption that all languages are equally represented in every task. Future versions will expand under-covered task-language pairs while preserving the same native-expert verification standard.

7 Ethical Considerations

We reveal the following ethical considerations of GaoYao:

(1) Benchmark Usage and Contamination: We release GaoYao to facilitate the assessment of LLMs. We explicitly discourage the inclusion of our test sets into model training corpora (contamination), which would render the evaluation validity null. We urge the community to treat this benchmark as a diagnostic tool rather than a leaderboard to be gamed.

(2) Annotator Fair Compensation and Well-being: All data annotators and linguistic experts involved in the localization and SUPERBLEND synthesis processes were full-time employees of professional language service providers and participated as part of their regular duties. They received standard professional salaries above local minimum wage, were informed about the intended use of the data, and were not recruited through unpaid labor or low-paid crowdsourcing. No personally identifiable information was collected.

(3) Mitigation of Cultural Stereotypes: Constructing cultural benchmarks carries a risk of reinforcing stereotypes. To mitigate this, we implemented a careful review process where native experts explicitly screened for offensive content, harmful generalizations, or political sensitivity as specified in Appendix E. While we strive for neutrality, we acknowledge that cultural data may still reflect the subjective perspectives of the annotators.

References

Mubashara Akhtar, Anka Reuel, Prajna Soni, Sanchit Ahuja, Pawan Sasanka Ammanamanchi, Ruchit Rawal, Vilém Zouhar, Srishti Yadav, Chenxi Whitehouse, Dayeon Ki, et al. 2026. When ai benchmarks plateau: A systematic study of benchmark saturation. *arXiv preprint arXiv:2602.16763*.

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

Anthropic. 2026. Introducing claude sonnet 4.5. Accessed: 2026-01-06.

Ascend Tribe. 2025. openpangu-ultra-moe-718b-v1.1. https://ai.gitcode.com/ascend-tribe/openPangu-Ultra-MoE-718B-V1.1.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei

Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, and et al. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R Costa-Jussa, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. 2019. Findings of the 2019 conference on machine translation (wmt19). ACL.

ByteDance. 2026. SEED 1.6. Accessed: 2026-01-05.

Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth Heafield. 2024. Monolingual or multilingual instruction tuning: Which makes a better alpaca. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1347–1356.

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. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

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

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. In *arXiv preprint arXiv:2501.12948*.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. In *First Conference on Language Modeling*.

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

Shiwei Guo, Sihang Jiang, Qianxi He, Yanghua Xiao, Jiaqing Liang, Bi Yude, Minggui He, Shimin Tao, and Li Zhang. 2025. Do large language models truly understand cross-cultural differences? *Preprint*, arXiv:2512.07075.

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

Minggui He, Yilun Liu, Shimin Tao, Yuanchang Luo, Hongyong Zeng, Chang Su, Li Zhang, Hongxia Ma, Daimeng Wei, Weibin Meng, et al. 2025. R1-t1: Fully incentivizing translation capability in llms via reasoning learning. *arXiv preprint arXiv:2502.19735*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, and et al. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. 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.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, et al. 2024. Findings of the wmt24 general machine translation shared task: The llm era is here but mt is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46.

Wen Lai, Mohsen Mesgar, and Alexander Fraser. 2024. Llms beyond english: Scaling the multilingual capability of llms with cross-lingual feedback. *arXiv preprint arXiv:2406.01771*.

Hongyu Li, Liang Ding, Meng Fang, and Dacheng Tao. 2024. Revisiting catastrophic forgetting in large language model tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4297–4308.

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

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

Yilun Liu, Ziang Chen, Song Xu, Minggui He, Shimin Tao, Weibin Meng, Yuming Xie, Tao Han, Chunguang Zhao, Jingzhou Du, et al. 2025a. R-log: Incentivizing log analysis capability in llms via reasoning-based reinforcement learning. *arXiv preprint arXiv:2509.25987*.

Yilun Liu, Chunguang Zhao, Xinhua Yang, Hongyong Zeng, Shimin Tao, Weibin Meng, Minggui He, Chang Su, Yan Yu, Hongxia Ma, et al. 2025b. Midb: Multilingual instruction data booster for enhancing multilingual instruction synthesis. *arXiv preprint arXiv:2505.17671*.

Meta AI. 2024. Introducing Llama 3.1: Our most capable models to date. Accessed: 2026-01-05.

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

Taishi Nakamura, Mayank Mishra, Simone Tedeschi, Yekun Chai, Jason T Stillerman, Felix Friedrich, Prateek Yadav, Tanmay Laud, Vu Minh Chien, Terry Yue Zhuo, et al. 2025. Aurora-m: Open source continual pre-training for multilingual language and code. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 656–678.

ASR Omnilingual, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, et al. 2025. Omnilingual asr: Open-source multilingual speech recognition for 1600+ languages. *arXiv preprint arXiv:2511.09690*.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and et al. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

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

OpenAI. 2025. Gpt-5 system card.

OpenAI. 2025. Introducing o3 and o4-mini. Accessed: 2026-01-06.

OpenAI. 2026. Introducing gpt-5. Accessed: 2026-01-06.

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

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

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

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

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

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

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

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, Zhuofu Chen, Jialei Cui, Hao Ding, and et al. 2025b. Kimi k2: Open agentic intelligence. *Preprint*, arXiv:2507.20534.

Qwen Team. 2025. Qwen3-max: Just scale it.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

xAI. 2025. Grok 3 beta — the age of reasoning agents. Accessed: 2026-01-06.

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

Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. 2023. Benchmarking machine translation with cultural awareness. *arXiv preprint arXiv:2305.14328*.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, and et al. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *Preprint*, arXiv:2508.06471.

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

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

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

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

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

Shaolin Zhu, Leiyu Pan, Dong Jian, and Deyi Xiong. 2025. Overcoming language barriers via machine translation with sparse mixture-of-experts fusion of large language models. *Information Processing & Management*, 62(3):104078.

Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Extrapolating large language models to non-english by aligning languages. *arXiv preprint arXiv:2308.04948*.

Vilém Zouhar, Pinzhen Chen, Tsz Kin Lam, Nikita Moghe, and Barry Haddow. 2024. Pitfalls and outlooks in using COMET. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1272–1288, Miami, Florida, USA. Association for Computational Linguistics.

A Related Work

A.1 Evaluation of Multilingual and Multicultural Capabilities

Evaluating the capabilities of LLMs across diverse languages and cultures is critical for their global deployment. Traditional evaluations primarily focus on MT and objective understanding tasks. Goyal et al. (2022) introduced FLORES-101, a benchmark covering massive languages for MT, while the WMT shared tasks (Kocmi et al., 2024; Barrault et al., 2019) remain the standard for evaluating translation quality using metrics like COMET (Rei et al., 2020; Zouhar et al., 2024). Beyond translation, recent benchmarks have expanded to broader understanding capabilities. BELEBELE (Bandarkar et al., 2024) evaluates parallel reading comprehension across 122 language variants, and INCLUDE (Romanou et al., 2025) assesses multilingual understanding with regional knowledge. Similarly, Pomerenke et al. (2025) track LLM progress on various multilingual benchmarks.

However, existing evaluations often exhibit limitations in scope and depth. Many benchmarks rely heavily on objective formats such as multiple-choice questions (MCQs), neglecting subjective generation tasks that better reflect real-world usage. While ALPACAEVAL (Li et al., 2023) and MT-BENCH (Zheng et al., 2023) provide robust evaluations for instruction following and dialogue, they are predominantly English-centric. Multilingual extensions like OMGEVAL (Liu et al., 2024) and X-ALPACAEVAL (Zhang et al., 2024) exist but often suffer from limited language coverage. Furthermore, cultural evaluation remains underexplored. Yao et al. (2023) and Zhang et al. (2025) have initiated efforts to probe cultural awareness, and Myung et al. (2024) introduced BLEND for everyday cultural knowledge. Yet, these datasets often suffer from limited culture coverages, and the problem of relying on direct translation which retains source-culture bias still exists (Rystrøm et al., 2025; Guo et al., 2025).

A.2 Comparison with Existing Benchmarks.

As summarized in Table 3, existing benchmarks typically exhibit a trade-off between breadth and depth. Massive-scale benchmarks (*e.g.*, FLORES-101 (Goyal et al., 2022), BELEBELE (Bandarkar et al., 2024)) cover over 100 languages but are restricted to objective tasks (Translation and Reading). Conversely, benchmarks focusing on subjective tasks (*e.g.*, X-ALPACAEVAL (Zhang et al., 2024)) are severely limited in language coverage (< 10). Crucially, even dedicated cultural benchmarks like SAGE (Guo et al., 2025) and CULTURESCOPE (Zhang et al., 2025) are constrained to specific bilingual dyads (2 languages, 2 cultures). Compared with these existing works, GaoYao establishes a more comprehensive framework. It in-

Benchmark	Focus Domain	Task Diversity	# Total Langs	# Subj. Langs†	# Cultures
FLORES-101 (Goyal et al., 2022)	Translation	Single	101	-	-
BELEBELE (Bandarkar et al., 2024)	Reading	Single	**122**	-	-
INCLUDE (Romanou et al., 2025)	Knowledge	Single	44	-	-
X-ALPACAEVAL (Zhang et al., 2024)	Instruction	Single	4	4	-
OMGEVAL (Liu et al., 2024)	Instruction	Single	9	9	-
SAGE (Guo et al., 2025)	Culture	Single	2	2	2
CULTURESCOPE (Zhang et al., 2025)	Culture	Single	2	2	2
GaoYao (Ours)	**Comprehensive**	**All (9 Sub-layers)**	26	**19**	**34**

†*Subj. Langs* refers to languages supported for open-ended subjective tasks (*e.g.*, Instruction Following & Multi-turn Dialogue).

Table 3: Comparison between GaoYao and other representative benchmarks. GaoYao distinguishes itself by its comprehensive scope, specifically surpassing others in subjective task coverage (19 languages) and cultural breadth (34 cultures) with rigorous human verification.

tegrates objective tasks with rigorously localized subjective tasks across 19 languages. Moreover, it significantly expands cultural evaluation through SUPERBLEND, a human-verified dataset covering 34 cultures, addressing the critical gap in deep, authentic multicultural assessment.

A.3 LLMs for Multilingual and Multicultural Tasks

The multilingual abilities of early LLMs are relatively underdeveloped (Lai et al., 2024) due to the predominance of English in their pretraining data, such as the early LLaMA series (Touvron et al., 2023a), *e.g.*, the ratio of non-English languages in pretraining corpus of LLaMA-2 (Touvron et al., 2023b) is merely around 2%. To improve the multilingual abilities of existing foundation LLMs, researchers have explored specialized tuning strategies. Many methods fouces on improving the quantity and quality of post-training using curated multilingual dataset and refined training strategies (Chen et al., 2024; Zhu et al., 2023; Zhang et al., 2024), while others leveraged continual pre-training (Fujii et al., 2024; Nakamura et al., 2025) and sparse Mixture-of-Experts architectures (Zhu et al., 2025) to mitigate catastrophic forgetting (Li et al., 2024).

More recently, the focus has shifted towards monolingual capabilities into multilingual tasks. The ratio of multilingual data in pre-training corpus of recent LLMs is significantly growing (Yang et al., 2025; DeepSeek, 2025). In specific tasks (*e.g.*, automatic speech recognition), the supported number of languages reaches 1600+ (Omnilingual et al., 2025). Despite the dramatic expansion of language coverage, the community has raised questions on the native level of responses in subjective tasks (Liu et al., 2025b) and authentic cultural experience for global users (Rystrøm et al., 2025), where existing LLMs still lag behind. The three curated subjective and cultural test sets in GaoYao can serve as a pioneering step for bridging the gap for existing multilingual LLM to achieve truly equitable AI access.

B Detailed Information of Languages Supported by GaoYao

Table 4 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. Fig. 9 displays the impact on LLM's best scores caused by resource popularity of languages, suggesting that LLM capabilities for languages with lower resources are relatively underdeveloped compared with high-resource ones.

C 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 5 presents the results. We observe a

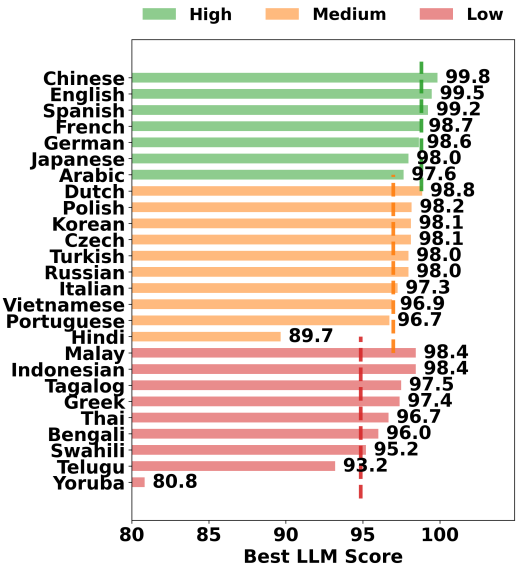


Figure 9: Impact of language resource popularity on best performance achieved by LLMs. Vertical dashed lines represent group averages. The taxonomy on language resource-levels is based on Joshi et al. (2020).

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.

D Data Card for Human Annotation

Annotators were selected based on native proficiency and professional experience in translation, localization, proofreading, editing, copy-writing, technical writing, or linguistic testing. The annotation guidelines emphasized meaning-preserving localization rather than literal translation, especially for instructions involving scripts, phonetics, culturally bound terms, or pragmatic conventions. For SUPERBLEND, annotators were instructed to answer from lived cultural experience without using AI systems or search engines, and to mark a question as "not applicable" when no clear cultural consensus existed. All annotations followed a review-rebuttal protocol: third-party reviewers flagged questionable cases, after which annotators either revised the answer or provided a justification for keeping it. Annotators were full-time employees of the partner language service center and participated as part of their regular professional duties, receiving standard salaries above local minimum wage rather than unpaid or crowdsourced compensation.

E Technical Details in the Curation of SUPERBLEND

As discussed in Section 2.2.1, existing cultural evaluation sets are limited in their 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, is a generalization of the BLEND dataset and 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 Su-

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 4: Language mapping: codes, geographical groups, and resource levels.

	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 5: 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.

perBLEnD 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, discarding approximately 41.1% of the raw data. The removed cases mainly fall into four categories. *Semantic redundancy* covers near-duplicate answers that express the same concept, such as merging "Mum" and "Mother" into a single normalized answer. *Context errors* refer to answers that are linguistically plausible but culturally or semantically incorrect in the target context; for example, "Wochenbett" was rejected for a German question about postpartum recovery locations because it refers to the puerperium period rather than a physical place. *Hierarchical conflicts* occur when a distractor or answer overlaps with another valid answer at a different granularity; for instance, "Pepsi" is unsuitable when "carbonated drinks" is also a valid answer. *Safety issues* include sensitive or potentially toxic responses that are inappropriate for a public benchmark. 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?"

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, ensuring 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 with the following prompt:

> 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":[]}

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" 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 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. The prompt is as follows:

> 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: [...]

This transforms straightforward MCQs into more complex versions while retaining the same cultural core, ensuring the benchmark tests cultural knowledge rather than simple pattern matching

F Disclosure of Generative AI Usage

The technology of generative AI is partially involved in this paper for the following three scenarios: (1) polishing texts for language fluency, (2) aiding the curation of SuperBLEnD dataset as specified in Appendix E and (3) aiding in plotting art elements in figures and diagrams (*e.g.*, the iceberg icon in Fig. 1).

G Detailed Experimental Setups

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 = $\frac{\#\text{correct}}{\#\text{all}}$
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	wmt22 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. weighted 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. weighted sum score is used as final score.
S-MT-Bench	QA	Subj.	Win Rate	Deep Seek V3.1	Qwen3-235B-A22B	1. Judge comparison (multi-turn averaged); 2. Win Rate = $\frac{\#\text{win}+\#\text{tie}/2}{\#\text{all}}$

Table 6: 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(Ascend Tribe, 2025)	openPangu-Ultra-MoE-718B-V1.1	https://ai.gitcode.com/ascend-tribe/openPangu-Ultra-MoE-718B-V1.1
DeepSeek-V3.1(DeepSeek, 2025)	DeepSeek-v3.1-250821	https://huggingface.co/deepseek-ai/DeepSeek-V3.1
Qwen3-235B-A22B(Yang et al., 2025)	Qwen3-235B-A22B-Instruct-2507	https://huggingface.co/Qwen/Qwen3-235B-A22B-Instruct-2507
Qwen3-VL-235B-A22B(Bai et al., 2025)	Qwen3-VL-235B-A22B-Instruct	https://huggingface.co/Qwen/Qwen3-VL-235B-A22B-Instruct
DeepSeek-R1(DeepSeek-AI, 2025)	DeepSeek-R1	https://huggingface.co/deepseek-ai/DeepSeek-R1
Llama-3.1-405B(Meta AI, 2024)	Llama-3.1-405B-Instruct	https://huggingface.co/meta-llama/Llama-3.1-405B-Instruct
GLM-4.6(Zeng et al., 2025)	GLM-4.6	https://chatglm.cn
Kimi-k2(Team et al., 2025b)	kimi-k2-250711	https://www.kimi.com
	Closed-Source Commercial Models	
Doubao-seed-1.6(ByteDance, 2026)	doubao-seed-1-6-250615	https://www.doubao.com/chat
Qwen-max(Team, 2025)	Qwen-max	https://chat.qwen.ai/
Gemini-2.5-Pro(Comanici et al., 2025)	Gemini-2.5-Pro	https://aistudio.google.com
Claude-Sonnet-4.5(Anthropic, 2026)	claude-sonnet-4-5-20250929	https://claude.ai
Grok-3(xAI, 2025)	Grok-3	https://grok.com
o3(OpenAI, 2025)	o3	https://chatgpt.com
o4-mini(OpenAI, 2025)	o4-mini	https://chatgpt.com
GPT-5-chat(OpenAI, 2026)	GPT-5-chat	https://chatgpt.com/
GPT-4o(OpenAI et al., 2024)	GPT-4o	https://chatgpt.com
	Compact Models (<20B)	
Qwen3-14B(Yang et al., 2025)	Qwen3-14B	https://huggingface.co/Qwen/Qwen3-14B
Qwen3-8B(Yang et al., 2025)	Qwen3-8B	https://huggingface.co/Qwen/Qwen3-8B
Gemma-3-12B-IT(Team et al., 2025a)	gemma-3-12b-it	https://huggingface.co/google/gemma-3-12b-it
Llama-3.1-8B(Meta AI, 2024)	Llama-3.1-8B-Instruct	https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct
Llama-3-8B(Meta AI, 2024)	Llama-3-8B-Instruct	https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
Ministral-8B-Instruct(Jiang et al., 2024)	Ministral-8B-Instruct-2410	https://huggingface.co/mistralai/Ministral-8B-Instruct-2410

Table 7: Detailed specifications of models evaluated in GaoYao.