Macaron: Controlled, Human-Written Benchmark for Multilingual and Multicultural Reasoning via Template-Filling

Alaa Elsetohy[1,*], **Sama Hadhoud**[1], **Haryo Akbarianto Wibowo**[1], **Chenxi Whitehouse**[2], **Genta Indra Winata**[3], **Fajri Koto**[1], **Alham Fikri Aji**[1,*]

[1]MBZUAI [2]Meta [3]Capital One

{alaa.elsetohy,alham.fikri}@mbzuai.ac.ae

[*]*Corresponding authors*

Abstract

Multilingual benchmarks rarely test reasoning over culturally grounded premises: translated datasets keep English-centric scenarios, while culture-first datasets often lack control over the reasoning required. We propose Macaron, a template-first benchmark that factorizes reasoning type and cultural aspect across question languages. Using 100 language-agnostic templates that cover 7 reasoning types, 22 cultural aspects, native annotators create scenario-aligned English and local-language multiple-choice questions and systematically derived True/False questions. Macaron contains 11,862 instances spanning 20 countries/-cultural contexts, 10 scripts, and 20 languages and dialects (including low-resource ones like Amharic, Yoruba, Zulu, Kyrgyz, and some Arabic dialects). In zero-shot evaluation of 21 multilingual LLMs, reasoning-mode models achieve the strongest performance (80.8% overall) and near-parity between English and local languages (ΔMC = -1.3%), while open-weight models degrade substantially in local languages (ΔMC = -6.8%) and often approach chance on T/F tasks. Culture-grounded mathematical and counting templates are consistently the hardest. The data can be accessed here https://huggingface.co/datasets/AlaaAhmed2444/Macaron.

1 Introduction

With the growing progress of multilingual LLMs, benchmarking them across languages and cultures is equally important. Existing benchmarks pursue this along two complementary directions, each with its own blind spot. Translation-parallel benchmarks enable controlled cross-lingual comparison but inherit English-centric scenarios (Conneau et al., 2018; Ponti et al., 2020; Artetxe et al., 2019; Lin et al., 2021; Singh et al., 2025; Xuan et al., 2025). Culture-first benchmarks provide locally salient content but lack explicit control over reasoning skills, and scaling them requires new questions from scratch for each culture, which drifts in scope and difficulty (Chiu et al., 2025; Myung et al., 2024; Sadallah et al., 2025; Hasan et al., 2025; Wibowo et al., 2024; Romero et al., 2024).

We propose Macaron, a template-first benchmark for multilingual multicultural reasoning that addresses these issues by separating, by construction, the three factors that prior evaluations conflate: the reasoning skill a question requires, the cultural aspect it probes, and the language in which it is presented. We design 100 language-agnostic templates tagged with 7 reasoning types and 22 cultural aspects, and recruit native annotators to instantiate them with culturally grounded content and produce scenario-aligned English–local versions. Because templates are reusable, extending Macaron to new cultures primarily requires instantiation and translation of the same template set, keeping structure and targeted reasoning stable.

From 1,977 bilingual MC scenarios spanning 20 languages and dialects, we derive aligned True/-False variants yielding 11,862 evaluation instances. Evaluations across 21 multilingual LLMs show that reasoning-mode models are strongest (80.8%) and nearly language-robust (Avg. ΔMC = -1.3), while open-weight models lag (58.0%) and degrade more in local languages (Avg.ΔMC = -6.8), with culture-grounded mathematical and counting questions consistently hardest.

Our contributions are:

1. A template-first framework that factorizes *reasoning type* and *cultural aspect* for controlled multilingual cultural reasoning.
2. Macaron: a scenario-aligned bilingual benchmark with MCQ and derived T/F variants across 20 cultural contexts.
3. An evaluation of 21 multilingual LLMs with analyses across languages, reasoning categories, and cultural aspects.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 47885–47906

2 Related Work

Reasoning and diagnostic evaluation. English-first benchmarks cover commonsense and plausibility reasoning (HellaSwag, WinoGrande, ARC, CROW) (Zellers et al., 2019; Sakaguchi et al., 2019; Clark et al., 2018; Ismayilzada et al., 2023) and exam-style reasoning (BIG-bench, MMLU, MMLU-Pro) (Srivastava et al., 2023; Hendrycks et al., 2021; Wang et al., 2024), with harder diagnostic subsets such as BBH (Suzgun et al., 2023). Controlled-structure datasets such as bAbI and CLUTRR (Weston et al., 2015; Sinha et al., 2019) motivate template-controlled evaluation. While these resources provide strong reasoning diagnostics, they are not designed to evaluate reasoning under culturally grounded premises.

Translation-parallel multilingual evaluation. A common multilingual strategy is to translate English-source datasets to many languages, enabling controlled cross-lingual comparison but inheriting source framing and assumptions. Examples include XNLI (Conneau et al., 2018), XCOPA (Ponti et al., 2020), XQuAD (Artetxe et al., 2019), and X–CSR (Lin et al., 2021). Global-MMLU and MMLU-ProX expand exam-style evaluation across languages and scripts while keeping instances parallel (Singh et al., 2025; Xuan et al., 2025). M3Exam extends this to a multilingual, multimodal, multilevel setting using real exam questions from multiple countries (Zhang et al., 2023), but like other translation-parallel benchmarks it does not systematically control reasoning type or pair English and local-language items over identical cultural scenarios.

Culture-grounded and regional benchmarks. Regional-sourcing benchmarks such as INCLUDE and MILU draw from local exams or region-specific materials (Romanou et al., 2025; Verma et al., 2025). Culture-first and native-query resources such as CulturalBench, BLEnD, ArabCulture, and NativQA/MultiNativQA emphasize locally salient content (Chiu et al., 2025; Myung et al., 2024; Sadallah et al., 2025; Hasan et al., 2025). NormAd is complementary for norm and etiquette judgments, but it is not a bilingual, scenario-aligned reasoning test (Rao et al., 2025). MultiNRC adds explicit reasoning categories alongside native-authored questions, but covers fewer languages and does not systematically cross reasoning types with cultural domains at scale (Fabbri et al., 2025). Rystrøm et al. (2025) show that multilingual capability and cultural alignment are distinct dimensions in LLMs, motivating benchmarks that explicitly separate the two, a goal our template-first design directly addresses. Veselovsky et al. (2025) show from an interpretability perspective that localized cultural knowledge is internally represented and controllable in LLMs, providing a complementary mechanistic view to our behavioral evaluation.

Disentangling language and culture. Closest in motivation to ours, Ying et al. (2025) propose a post-hoc framework for decomposing model scores along linguistic and cultural axes, but applied to pre-existing datasets, whereas Macaron enforces this separation by construction through native-annotated, scenario-aligned bilingual items.

Table 1 compares our benchmark with previous work. Macaron is the only benchmark to simultaneously satisfy all seven properties: culture-grounded, native-authored, bilingual-aligned, template-based, with a reasoning taxonomy, a culture taxonomy, and their joint coverage at the item level.

3 Data Curation

Our goal is to evaluate *multilingual, multicultural reasoning* in a controlled setting. We operationalize this as (i) multiple-choice question answering and (ii) binary True/False verification over the *same* culturally grounded scenarios as shown in Figure 1. The benchmark is designed to help disentangle three factors that are often confounded in multilingual evaluation: *language* (English vs. local input), *cultural grounding*, and *reasoning* (the inference required to answer).

3.1 Task Definition

Let $\mathcal{L}$ denote the set of local languages in the benchmark and $\mathcal{C}_{\text{ctx}}$ the set of cultural contexts (countries or regions). We construct *base annotations* as bilingual, culturally aligned multiple-choice items. A base annotation is a tuple

$$a = \left(q^{\text{en}}, A^{\text{en}}, q^{\ell}, A^{\ell}, R_a, C_a\right),$$

where q^{en} and q^{ℓ} are the English and local-language question texts, A^{en} and A^{ℓ} are the corresponding sets of four answer options with exactly one correct choice, and $\ell \in \mathcal{L}$ is the local language. We treat both reasoning and culture as explicit (potentially multi-label) metadata: $R_a \subseteq \mathcal{R}$ is the set of reasoning types targeted by the item, and $C_a \subseteq \mathcal{C}_{\text{aspect}}$ is the set of cultural aspects it probes.

Benchmark	Format	#Eval items	#Langs & Dialects	#Cultures/ Regions	Culture grounded	Native authored	Bilingual aligned	Template based	Reasoning taxonomy	Culture taxonomy	Reasoning × Culture
Translation-parallel benchmarks											
XNLI (Conneau et al., 2018)	NLI	5k/lang	15	—	×	×	✓	×	×	×	×
XCOPA (Ponti et al., 2020)	MCQ	500/lang	11	—	×	×	✓	×	×	×	×
Global-MMLU (Singh et al., 2025)	MCQ	14k/lang	42	42	×	×	✓	×	×	×	×
MMLU-ProX (Xuan et al., 2025)	MCQ	11.8k/lang	29	29	×	×	✓	×	×	×	×
Culture-grounded benchmarks											
INCLUDE (Romanou et al., 2025)	MCQ	197.2k	44	52	✓	✓	×	×	×	×	×
MILU (Verma et al., 2025)	MCQ	79.6k	11	11	✓	✓	×	×	×	×	×
BLEnD (Myung et al., 2024)	MCQ+Free	52.6k	13	16	✓	✓	✓	×	×	×	×
CulturalBench (Chiu et al., 2025)	MCQ+T/F	1.7k	1	45	✓	✓	×	×	×	✓	×
ArabCulture (Sadallah et al., 2025)	MCQ	3.5k	1	13	✓	✓	×	×	×	×	×
NativQA (Hasan et al., 2025)	QA	64k	7	9	✓	✓	×	×	×	×	×
WorldCuisines (Winata et al., 2025)	MCQ+QA	∼1M	30	—	✓	✓	✓	✓	×	×	×
MultiNRC (Fabbri et al., 2025)	Free-text	1k	3	3	✓	✓	✓	×	✓*	×	×
Ours											
Macaron	**MCQ+T/F**	**∼12k**	**20**	**20**	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison of Macaron with related multilingual and multicultural benchmarks across key design properties. **#Cultures/Regions**: number of distinct cultural contexts covered; dashes (—) indicate translation-parallel benchmarks with no distinct cultural grounding, and WorldCuisines where the cultural unit is cuisine rather than a discrete cultural context. **Bilingual aligned**: English vs Local language. **Reasoning × Culture**: reasoning type and cultural aspect are jointly tagged at the item level. *MultiNRC provides reasoning categories but at a much coarser granularity than Macaron's seven-type taxonomy.

From each base annotation, we derive four additional binary instances (True/False in English and in local language; Section 3.5), yielding **six aligned evaluation instances per cultural scenario**.

3.2 Reasoning and Cultural Taxonomies

Reasoning types. We define a taxonomy of seven reasoning types that commonly arise in culturally grounded questions: *mathematical* (numerical computation and comparison), *commonsense* (everyday plausibility and typical situations), *causal* (cause–effect relations), *temporal* (time, order, calendars), *logical* (deduction, implication, and analogy), *spatial* (geographic and spatial relations), and *multi-hop* (composition of two or more inference steps, e.g., symbol → religion → practice). Templates may be tagged with multiple reasoning types when solving requires more than one skill.

Cultural aspects. Following Adilazuarda et al. (2024), we treat culture as a broad, multifaceted concept that resists direct definition, and instead operationalize it through *proxies of culture*, specifically 22 *semantic proxies* that span the domains of everyday life a community shares. These aspects serve as our cultural taxonomy: *Agriculture*, *Brands and Commerce*, *Cities and Landmarks*, *Death and Funerals*, *Education*, *Events and Festivals*, *Famous People*, *Fashion and Media*, *Folklore and Folktales*, *Food and Cuisine*, *Language and Communication*, *Literature and Written works*, *Music and Art*, *Naming*, *Objects and Units*, *Politics and Governance*, *Relationships*, *Social Customs*, *Sports*, *Time*, *Transportation*, and *socio-religious aspects of life*. Each template is associated with at least one aspect, and some span multiple aspects. We provide an example template for each aspect in Appendix G (Table 10).

3.3 Template Framework

To systematically cover the reasoning×culture space, we design a set of 100 language-agnostic templates. Each template specifies:

- a question skeleton with typed slots (e.g., `[COUNTRY]`, `[PERSON]`, `[FOOD1]`), including constraints on valid slot values;
- metadata tags indicating the targeted reasoning type(s) and cultural aspect(s);
- an expected output format (four-option multiple choice with exactly one correct answer).

Templates are authored and iteratively refined by the dataset creators. During refinement, we remove culturally insensitive or non-portable designs, tighten slot constraints to prevent ambiguity, and ensure that the intended reasoning path is stable across cultural contexts. Each template also includes a True/False variant (Section 3.5).

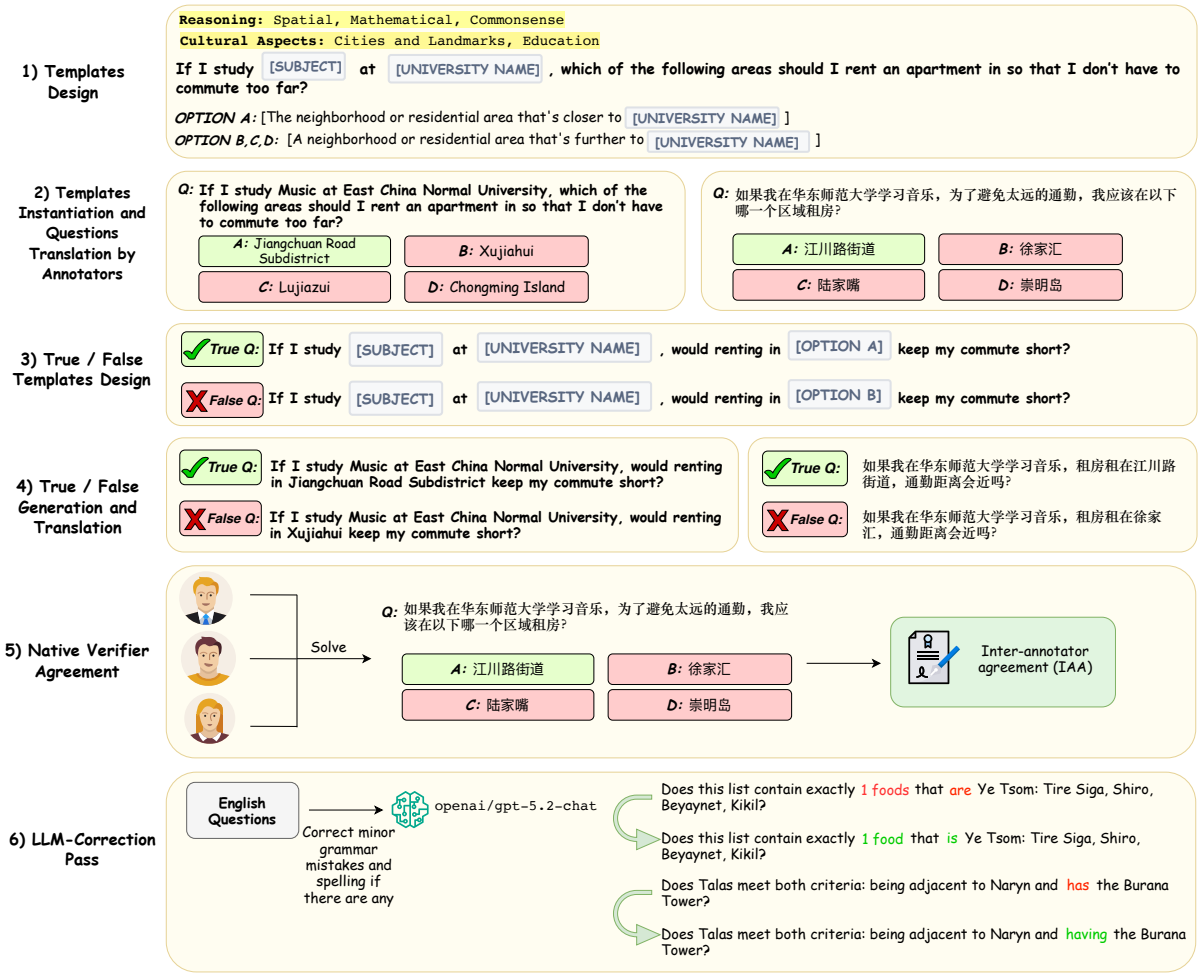

Figure 1: **Macaron Curation Pipeline.** We first design language-agnostic templates tagged with reasoning categories and cultural aspects. Native annotators instantiate each template with culturally grounded content to produce scenario-aligned English MCQs and translate them into local languages. From each MCQ, we derive aligned True/False statements by instantiating the template's binary child forms with the correct option (True) and a distractor (False), and translate these statements into the local language. Finally, we run an LLM proofreading pass to correct minor spelling/grammar issues while preserving the original meaning and leaving all local-language text unchanged.

3.4 Bilingual Annotation Pipeline

Annotators and onboarding For each cultural context, we recruit two annotators via the Upwork freelancing platform. Annotators are native speakers of the target local language and have substantial lived experience in the target context (e.g., having grown up and/or currently residing there). During onboarding, annotators receive some guidelines and complete a small pilot set of templates with feedback from the dataset creators. Annotator demographics (gender, age, education, residence duration) are reported in Appendix C; screenshots of the annotation platform are shown in Appendix I.

The guidelines emphasize:

1. **Cultural representativeness within templates.** Instantiate each assigned template with content that is locally appropriate and commonly recognizable to members of the target culture, aiming for diversity across everyday institutions and practices.

2. **Avoiding both stereotypes and obscure trivia.** Prefer culturally salient, everyday knowledge rather than tourist facts or internet stereotypes, while avoiding niche or hard-to-verify trivia that most locals would not know.

3. **Plausible within-context distractors.** Write distractors that are plausible *within the same cultural sphere* so items cannot be solved by eliminating obviously foreign options; ensure exactly one correct answer.

4. **Non-applicability and ambiguity flags.** Flag and skip templates that do not meaningfully apply to your cultural context.

Step 1: English multiple-choice instantiation Annotators are assigned a subset of templates. For each template, they:

1. fill required slots with culturally appropriate content based on lived experience and commonly shared local knowledge;
2. provide one correct option and three plausible distractors, ensuring exactly one option is correct for the target context.

Step 2: local-language translation After completing the English version, the same annotator translates the question and its options into their native language. For each base item, they produce local-language text q^ℓ and A^ℓ under the constraints that the translation must:

- preserve the underlying cultural content (same dish, institution, practice, person, etc.);
- preserve the reasoning structure and difficulty;
- allow only light adaptations for naturalness when needed, as long as the English and local versions still align.

Each item inherits the reasoning and cultural-aspect tags from the template metadata, yielding a bilingual base annotation $a = \left(q^{\text{en}}, A^{\text{en}}, q^\ell, A^\ell, R_a, C_a\right)$.

3.5 True/False Variant Generation

Starting from each base annotation, we construct four additional binary instances: True and False in English and in the local language. Concretely, for each base item we instantiate the template's binary child forms: a **True** variant by inserting the *correct* option, yielding a statement whose correct label is True; and a **False** variant by inserting a carefully chosen *distractor* option, yielding a statement whose correct label is False.

We generate these binary instances in both English and the local language, maintaining scenario-level alignment. The True and False variants share the same cultural scenario and reasoning requirements as the parent multiple-choice item. Thus, each base annotation yields six aligned instances: MC–EN, MC–L, T–EN, T–L, F–EN, and F–L.

3.6 Quality Control

To ensure cultural correctness, linguistic clarity, and consistency across annotators and cultural contexts, we apply a combination of human review and automated quality-control procedures.

Native verifier agreement. We conduct an independent verification step to assess the factual correctness of cultural content and the quality of local-language translations. For each cultural context, we recruit **three independent native verifiers**, distinct from the original annotators, and ask them to answer the translated MCQ version of every question in the dataset. Verifiers are native speakers with substantial lived experience in the target culture; their responses test both whether the cultural facts are accurate and whether the local-language phrasing is clear and unambiguous.

We compute **inter-annotator agreement** (IAA) as the proportion of items on which all three verifiers select the same answer. Table 6 (Appendix B) reports IAA scores across all 20 cultural contexts; scores range from 87.0% (South Africa) to 94.8% (China), with a mean of 90.9%, indicating high factual correctness and translation clarity across the dataset. For items where verifiers disagree with the originally labeled correct answer, we review the item through discussions with the annotators and verifiers; if consensus confirms the verifiers' answer is correct, we update the gold label accordingly while keeping the question text unchanged.

LLM-assisted English proofreading. Because questions are produced by instantiating shared templates with culture-specific content, small surface-level inconsistencies can arise in the English text across annotators and contexts (e.g., tense mismatches introduced by adapting a template from a generic present-tense form to a past event). To mitigate this template-instantiation noise without altering cultural content or reasoning difficulty, we run a deterministic LLM-assisted proofreading pass on *English* fields only (multiple-choice questions and options, and the English True/-False statements). For each English field, we query `openai/gpt-5.2-chat`[1] to correct *only* spelling and grammatical errors (including agreement, tense consistency, and capitalization), while preserving the original writing style and word choices; rephrasing or stylistic improvement is explicitly disallowed. All local-language text is left exactly as written by annotators.

[1] https://cdn.openai.com/gpt-5-system-card.pdf

Country	Language	Script	#Q
Egypt	Egyptian Arabic	Arabic	594
Philippines	Tagalog	Latin	594
India	Hindi	Devanagari	600
Ethiopia	Amharic	Ge'ez	588
Mexico	Mexican Spanish	Latin	594
Tunisia	Tunisian Arabic	Arabic	600
Greece	Greek	Greek	600
Brazil	Portuguese	Latin	600
Kyrgyzstan	Kyrgyz	Cyrillic	600
South Africa	Zulu	Latin	600
Italy	Italian	Latin	588
Thailand	Thai	Thai	594
Turkey	Turkish	Latin	600
Georgia	Georgian	Mkhedruli	594
China	Chinese	Han (Hans)	582
Indonesia	Indonesian	Latin	570
Yemen	Yemeni Arabic	Arabic	600
Nigeria	Yoruba	Latin	570
Morocco	Moroccan Arabic	Arabic	600
Japan	Japanese	Japanese	594
Total			**11,862**

Table 2: Dataset statistics and coverage. Small deviations across contexts reflect items removed in revisions.

3.7 Dataset Statistics

After quality control and expansion, the benchmark contains 11,862 total evaluation instances. Table 2 summarizes the distribution by cultural context (country), along with the associated local language and script. Appendix A provides additional breakdowns of template coverage across cultural aspects (Figure 5) and reasoning categories (Figure 6).

4 Experimental Setup

We evaluate 21 multilingual LLMs in zero-shot on both multiple-choice (MC) and True/False (T/F) formats, using paired English and local-language versions to isolate the effects of *language*, *cultural grounding*, and *reasoning type*.

Open-weight models use log-probability scoring; API-only models, including thinking models, output a structured JSON answer on the final line (Appendix F). Full scoring details and validation against generation-based scoring are in Appendix D. Answer extraction error rates for generation-based models are near zero across all models and do not meaningfully affect conclusions (Appendix E).

5 Results and Discussion

Table 3 reports overall performance on scenario-aligned English and local-language instances in both multiple-choice (MC) and True/False formats, grouped by model category. We additionally report cross-lingual gaps ΔMC and ΔTF, computed as (Local $-$ English), where negative values indicate

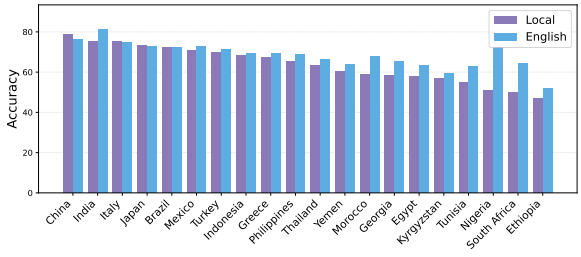

Figure 2: Average performance in English vs. local language across cultural contexts.

degraded performance in the local language. Per-language and per-script breakdowns are provided in Appendix H.

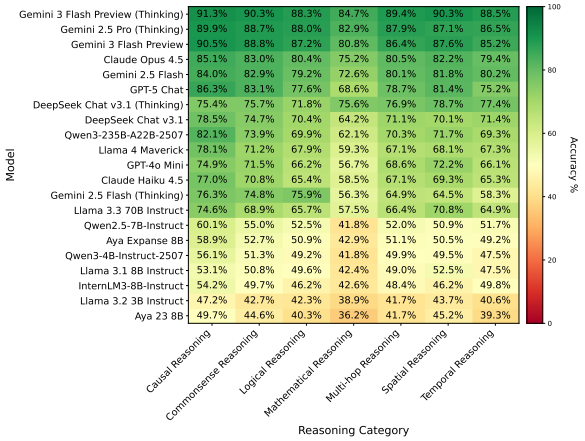

Figure 3: Accuracy by reasoning type across models.

Open-weight models show larger English–local gaps and reduced reliability on T/F. Closed-source thinking models achieve the highest overall accuracy (**80.8%** on average), outperforming closed (standard) models (**74.8%**) and open-weight models (**58.0%**). The English–local gap widens as model capacity decreases: thinking models are near-parity (Avg. ΔMC $= -1.3$), whereas open-weight models exhibit much larger drops (Avg. ΔMC $= -6.8$), particularly at 3B–8B scale. While most models are above random baseline in MC, many open-weight models cluster around $\sim$50–55% on T/F, indicating limited reliability in binary verification of culturally grounded statements.

Paired True/False accuracy exposes scenario-level verification. Each MC scenario yields a True/False pair (Section 3.5); we count a scenario correct only if the model answers both correctly. As shown in Table 5, paired accuracy is substantially lower than per-question T/F accuracy across all categories, suggesting single T/F performance is inflated by shallow response tendencies rather

Category	Model	MC-EN	MC-L	ΔMC	TF-EN	TF-L	ΔTF	Overall
Closed (thinking)	Gemini 3 Flash Preview (Google, 2025) (thinking)	89.5%	89.1%	-0.4%	84.5%	82.7%	-1.8%	86.5%
	Gemini 2.5 Pro (Comanici et al., 2025) (thinking)	87.5%	88.3%	+0.8%	81.6%	81.2%	-0.4%	84.7%
	Gemini 2.5 Flash (Comanici et al., 2025) (thinking)	70.1%	65.9%	-4.2%	74.6%	74.1%	-0.5%	71.2%
Average (Closed (thinking))		82.4%	81.1%	-1.3%	80.2%	79.3%	-0.9%	**80.8%**
Closed (standard)	Gemini 3 Flash Preview (Google, 2025) (standard)	86.8%	87.0%	+0.2%	81.9%	79.7%	-2.2%	83.9%
	Claude Opus 4.5 (Anthropic, 2025)	81.3%	80.2%	-1.1%	76.9%	75.2%	-1.7%	78.4%
	GPT-5 Chat (OpenAI, 2025)	79.0%	78.2%	-0.8%	77.1%	73.5%	-3.6%	77.0%
	Gemini 2.5 Flash (Comanici et al., 2025)	80.2%	80.2%	+0.0%	71.9%	70.5%	-1.4%	75.7%
	Claude Haiku 4.5 (Anthropic, 2025)	70.5%	63.8%	-6.7%	68.5%	64.8%	-3.7%	66.9%
	GPT-4o-mini (OpenAI et al., 2024)	70.6%	65.3%	-5.3%	67.0%	64.2%	-2.8%	66.8%
Average (Closed (standard))		78.1%	75.8%	-2.3%	73.9%	71.3%	-2.6%	**74.8%**
Open-weight	DeepSeek-Chat v3.1 (DeepSeek-AI et al., 2025) (thinking)	76.2%	75.6%	-0.6%	76.7%	71.7%	-5.0%	75.1%
	DeepSeek-Chat v3.1 (DeepSeek-AI et al., 2025)	74.0%	68.4%	-5.6%	67.9%	64.1%	-3.8%	68.6%
	Qwen3-235B-A22B (Yang et al., 2025)	73.3%	68.3%	-5.0%	67.3%	65.9%	-1.4%	68.7%
	Llama 3.3-70B (Grattafiori et al., 2024)	70.2%	62.6%	-7.6%	67.5%	62.0%	-5.5%	65.6%
	Llama 4 Maverick (Meta, 2025)	68.7%	67.5%	-1.2%	64.1%	62.1%	-2.0%	65.6%
	Llama 3.1-8B (Grattafiori et al., 2024)	54.2%	43.4%	-10.8%	56.7%	53.4%	-3.3%	51.9%
	Qwen3-4B (Yang et al., 2025)	52.6%	45.6%	-7.0%	55.2%	53.7%	-1.5%	51.8%
	InternLM3-8B (InternLM Team, 2025)	54.8%	40.9%	-13.9%	55.8%	52.2%	-3.6%	50.9%
	Qwen2.5-7B (Qwen et al., 2025)	57.0%	46.9%	-10.1%	52.6%	52.6%	+0.0%	52.3%
	Aya Expanse-8B (Dang et al., 2024)	52.7%	48.7%	-4.0%	51.7%	52.5%	+0.8%	51.4%
	Llama 3.2-3B (Grattafiori et al., 2024)	47.4%	36.3%	-11.1%	54.9%	51.6%	-3.3%	47.6%
	Aya-23-8B (Aryabumi et al., 2024)	43.8%	39.2%	-4.6%	50.4%	50.5%	+0.1%	46.0%
Average (Open-weight)		60.4%	53.6%	-6.8%	60.1%	57.7%	-2.4%	**58.0%**
Average (All)		68.6%	63.9%	-4.7%	66.9%	64.7%	-2.2%	**66.0%**

Table 3: Zero-shot accuracy (%) of 21 LLMs on `Macaron`, grouped by model category. **MC-EN/MC-L**: multiple-choice accuracy in English and local language; **TF-EN/TF-L**: True/False accuracy; **ΔMC/ΔTF**: cross-lingual gap (Local−English), where negative values indicate degraded local-language performance. **Overall** averages all four scores.

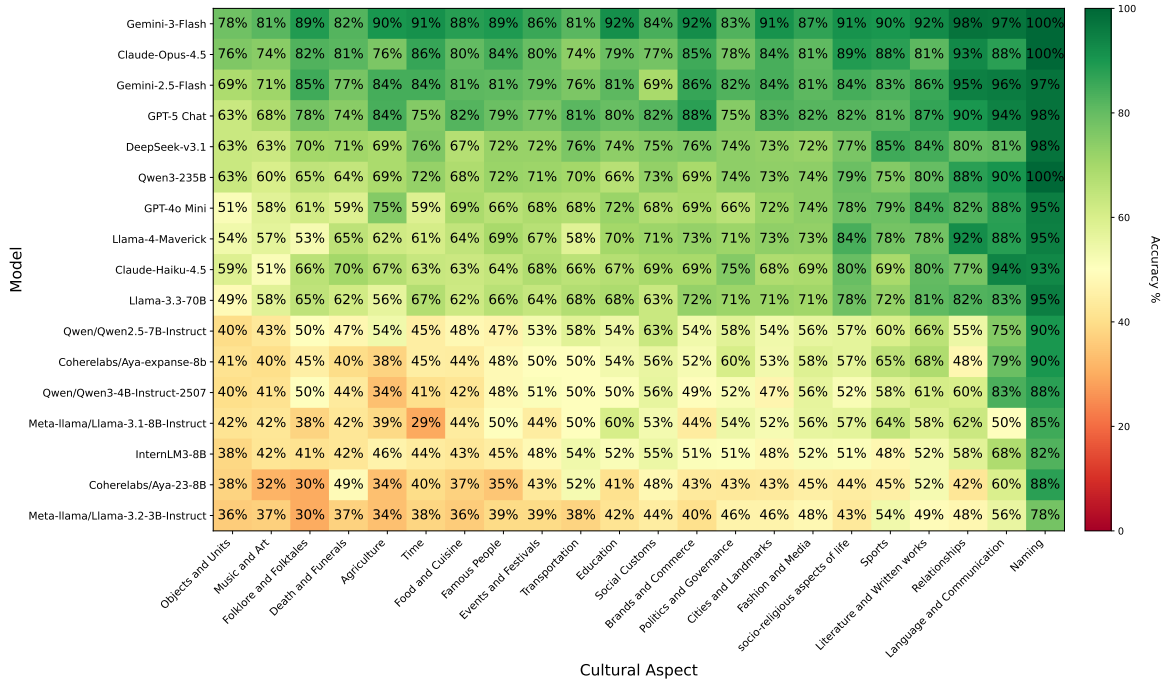

Figure 4: Accuracy (%) by cultural aspect across models. Each cell aggregates over evaluation instances whose templates are tagged with the corresponding aspect (multi-label; a single instance may contribute to multiple aspects). Aspects are ordered by mean accuracy (hard → easy).

than genuine cultural knowledge.

English outperforms local except for China, while the largest gaps concentrate in lower-resource languages. Figure 2 shows China as nearly the only case where local slightly exceeds English, plausibly influenced by the strong presence of Chinese-focused models (e.g., multiple Qwen variants) in our evaluation. In contrast,

Top-5 Easiest Templates (avg. over models)		Top-5 Hardest Templates (avg. over models)	
Acc	**Template prompt**	**Acc**	**Template prompt**
92.51%	`If my friend has the last name [LAST NAME], which country is most likely their birthplace?`	35.76%	`Among all the provinces in [SET OF PROVINCES/LOCATION], how many provinces have an area smaller than [PROVINCE]?`
85.66%	`Which of the following authors is traditionally associated with [CULTURE/REGION]'s literature?`	37.52%	`[GIFT 1], [GIFT 2], [GIFT 3], and [GIFT 4]. How many of these are suitable gifts to give when attending a [NATIONALITY] friend's wedding?`
85.44%	`If I am visiting [PLACE] in [MONTH], what kind of clothes should I pack?`	38.31%	`[MATERIAL 1], [MATERIAL 2], [MATERIAL 3], [MATERIAL 4]. How many of these are not typically used when making [TRADITIONAL ITEM]?`
84.98%	`In [NATIONALITY OR CULTURE] attire, what determines the [GARMENT FEATURE]?`	39.61%	`Among all [POLITICAL POSITION] in [COUNTRY] up until [YEAR], how many of them are [CONDITION]?`
84.36%	`Which of these political systems was/is traditionally practiced in [REGION] during the [TIME PERIOD]?`	42.30%	`[FOOD 1], [FOOD 2], [FOOD 3], [FOOD 4], how many of these are [SOME CLASSIFICATION]?`

Table 4: Template difficulty extremes: easiest prompts are simple cultural associations, while hardest prompts are constraint-heavy exact-count questions.

Category	T/F (avg. EN/L)	Paired T/F	Drop (pp)
Closed (thinking)	78.4%	62.1%	16.3
Closed (standard)	71.7%	48.4%	23.3
Open-weight	56.6%	18.2%	38.4
Average (All)	**65.8%**	**36.6%**	**29.2**

Table 5: True/False accuracy (averaged over English and local vs. paired True/False accuracy by model category. **Drop** is the difference in percentage points.

the English–local gap is substantially larger for Amharic, Yoruba, Zulu, and Arabic dialects, highlighting that cross-lingual brittleness concentrates in lower-coverage languages and dialects.

Mathematical reasoning is hardest in cultural contexts; causal and temporal reasoning are easier. In Figure 3, Mathematical Reasoning is the lowest-accuracy category for 20/21 models, while Causal (often alongside Commonsense) is typically highest. We attribute this to a double burden: culture-grounded math questions require both retrieving long-tail, locale-specific numeric facts and performing exact composition (e.g., counting or aggregation), so either retrieval or calculation errors can flip the answer. Moreover, such numeric facts are often sparse in training data and region-specific, raising the risk of confident but incorrect answers. In contrast, causal and commonsense questions are often supported by broadly shared everyday knowledge that is better covered in pretraining corpora.

Cultural aspect difficulty is consistent across models. Figure 4 reports a model×aspect heatmap and reveals a stable hard → easy ordering across model families. Averaged over models, *Naming* is the easiest aspect (92.5%), followed by *Language and Communication* (80.6%), whereas *Objects and Units* (52.9%) and *Music and Art* (54.0%) are the hardest, yielding a ∼40-point spread. Most remaining aspects form a broad middle band (roughly 64–70% mean accuracy). Beyond mean difficulty, the heatmap highlights where robustness gaps are largest: *Time* varies from 91% (strongest model) down to 29% (smaller open-weight models), with similarly large spreads across other aspects.

"How many" templates are the main failure mode. Table 4 shows that the easiest templates are mostly single-step cultural associations (e.g., last name → likely birthplace at 92.51%), while the hardest are uniformly "How many…" prompts that require enumerating a culturally grounded set, applying a constraint (often with negation/comparison), and producing an exact count (down to 35.76%). This pattern is consistent with our earlier finding that mathematical reasoning remains the weakest capability in culturally situated scenarios.

6 Conclusion

We introduce a template-first benchmark for multilingual, multicultural reasoning across 20 cultural contexts, languages and dialects, and 10 scripts (11,862 total instances). Our dataset separates language, cultural grounding, and reasoning type using scenario-aligned multiple-choice and True/False items. Zero-shot evaluations show that closed reasoning models achieve near-parity between English and local inputs, while open-weight models lag with significant performance drops. This benchmark supports diagnostic evaluation to motivate more culturally robust model development.

Limitations

Despite its breadth, the benchmark has several limitations. First, **coverage** is necessarily coarse: we include 20 cultural contexts with one primary local language each, which cannot capture within-country cultural diversity, minority languages, or finer-grained dialect continua. As a result, performance within a single country or language should not be interpreted as representative of all local varieties or communities. Second, the **task format** is intentionally controlled: while multiple-choice and binary verification enable precise alignment and diagnostic evaluation, they do not reflect open-ended dialogue, interactive reasoning, tool use, or real-world information access. Consequently, the benchmark measures culturally grounded reasoning under constrained conditions rather than end-to-end performance in realistic deployment settings.

Ethical Considerations

`Macaron` is a human-written benchmark designed to evaluate multilingual reasoning over culturally grounded premises in a controlled, template-first setting.

Annotator compensation. We recruited two annotators per cultural context via Upwork and compensated them at a fixed rate of US$9 per 10 completed template instantiations, where a completion consists of writing the English multiple-choice item and translating the question, options, and T/F variants into the local language. We additionally recruited three independent native verifiers per cultural context to assess factual correctness and translation quality, compensated at a rate of US$8 per hour. All annotators and verifiers were anonymized, and no personally identifiable information about them is included in the dataset or released publicly.

Cultural sensitivity and representational harms. Because items are cultural-based, there is a risk of stereotyping, oversimplifying a country/region into a single culture, or encoding contested practices as universal. We mitigate this through (i) iterative template refinement to remove culturally insensitive/non-portable designs and reduce ambiguity, (ii) annotator guidelines that emphasize culturally representative everyday knowledge while avoiding stereotypes and obscure trivia, and (iii) plausible within-context distractors to reduce "foreign-option elimination." Annotators may also flag templates as non-applicable when they do not meaningfully transfer.

Coverage is coarse (one primary local language per cultural context), so results should not be interpreted as measuring within-country diversity or dialect variation. As with any benchmark, `Macaron` can be misused for overfitting or for simplistic "ranking" of languages/cultures; we recommend using it as a diagnostic tool and reporting results with the above coverage and format constraints in mind.

LLM assistance in writing. We used an LLM as a writing aid (e.g., to improve clarity and correct minor grammar issues) while drafting this manuscript. All technical content, experimental design, analyses, and conclusions were produced and verified by the authors, who take full responsibility for the final paper. We did not provide any sensitive or personally identifying information to the model.

References

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Ashutosh Dwivedi, Laxmaan Khan, Maria O'Brien, Subhadarshi Ghosh, Rohan Saxena, Dominik Schneider, and 1 others. 2024. Towards measuring and modeling "culture" in LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Anthropic. 2025. Claude opus 4.5. https://platform.claude.com/docs/en/about-claude/models/whats-new-claude-4-5.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. Introduces the XQuAD dataset.

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

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, and 1 others. 2025. Culturalbench: A robust, diverse and challenging benchmark for measuring lms' cultural knowledge through human–ai red-teaming. In *Proceedings of the 63rd Annual Meeting of the Association for*

Computational Linguistics, Vienna, Austria. Association for Computational Linguistics.

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

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Alexander R. Fabbri, Diego Mares, Jorge Flores, Meher Mankikar, Ernesto Hernandez, Dean Lee, Bing Liu, and Chen Xing. 2025. Multinrc: A challenging and native multilingual reasoning evaluation benchmark for llms. *Preprint*, arXiv:2507.17476.

Google. 2025. Gemini 3 flash. https://blog.google/products-and-platforms/products/gemini/gemini-3-flash/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Md. Arid Hasan, Maram Hasanain, Fatema Ahmad, Sahinur Rahman Laskar, Sunaya Upadhyay, and 1 others. 2025. Nativqa: Multilingual culturally-aligned natural query for llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, Vienna, Austria. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

InternLM Team. 2025. InternLM3-8B model card. https://internlm.readthedocs.io/en/latest/model_card/InternLM3.html.

Mete Ismayilzada, Debjit Paul, Syrielle Montariol, Mor Geva, and Antoine Bosselut. 2023. Crow: Benchmarking commonsense reasoning in real-world tasks. *Preprint*, arXiv:2310.15239.

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. 2021. Common sense beyond english: Evaluating and improving multilingual language models for commonsense reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

Meta. 2025. Llama 4. https://www.llama.com/models/llama-4/.

Junho Myung and 1 others. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*.

OpenAI. 2025. GPT-5 system card. https://cdn.openai.com/gpt-5-system-card.pdf.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Online. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Abhinav Sukumar Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2025. Normad: A framework for measuring the cultural adaptability of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, page 2373–2403. Association for Computational Linguistics.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Zeming Chen, Mohamed A. Haggag, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Danylo Boiko, Michael Chang, Jenny Chim, Gal Cohen, Aditya Kumar Dalmia, Abraham Diress, Sharad Duwal, and 38 others. 2025. INCLUDE: Evaluating multilingual language understanding with regional knowledge. In *The Thirteenth International Conference on Learning Representations*.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, and 57 others. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *Preprint*, arXiv:2406.05967.

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

Abdelrahman Sadallah, Junior Cedric Tonga, Khalid Almubarak, Saeed Almheiri, Farah Atif, and 1 others. 2025. Commonsense reasoning in arab culture. *arXiv preprint*, arXiv:2502.12788.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale. *Preprint*, arXiv:1907.10641.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Sebastian Ruder, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, and 5 others. 2025. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *Preprint*, arXiv:2412.03304.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. 2019. CLUTRR: A diagnostic benchmark for inductive reasoning from text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4506–4515, Hong Kong, China. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 432 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Preprint*, arXiv:2206.04615.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2023. Challenging big-bench tasks and whether chain-of-thought can solve them. In *ACL (Findings)*, pages 13003–13051.

Sshubam Verma, Mohammed Safi Ur Rahman Khan, Vishwajeet Kumar, Rudra Murthy, and Jaydeep Sen. 2025. MILU: A multi-task Indic language understanding benchmark. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10076–10132, Albuquerque, New Mexico. Association for Computational Linguistics.

Veniamin Veselovsky, Berke Argin, Benedikt Stroebl, Chris Wendler, Robert West, James Evans, Thomas L. Griffiths, and Arvind Narayanan. 2025. Localized cultural knowledge is conserved and controllable in large language models. *Preprint*, arXiv:2504.10191.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Preprint*, arXiv:2406.01574.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *Preprint*, arXiv:1502.05698.

Haryo Wibowo, Erland Fuadi, Made Nityasya, Radityo Eko Prasojo, and Alham Aji. 2024. COPAL-ID: Indonesian language reasoning with local culture and nuances. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1404–1422, Mexico City, Mexico. Association for Computational Linguistics.

Genta Indra Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Wang Yutong, Adam Nohejl, Ubaidillah Ariq Prathama, Nedjma Ousidhoum, Afifa Amriani, and 1 others. 2025.

Worldcuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3242–3264.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, and 13 others. 2025. Mmlu-prox: A multilingual benchmark for advanced large language model evaluation. *Preprint*, arXiv:2503.10497.

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

Jiahao Ying, Wei Tang, Yiran Zhao, Yixin Cao, Yu Rong, and Wenxuan Zhang. 2025. Disentangling language and culture for evaluating multilingual large language models. *Preprint*, arXiv:2505.24635.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *Preprint*, arXiv:1905.07830.

Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Preprint*, arXiv:2306.05179.

A Data Statistics in More Detail

This section provides additional breakdowns of template coverage across cultural aspects and reasoning categories, supplementing Table 2 in the main paper.

A.1 Cultural-aspect coverage

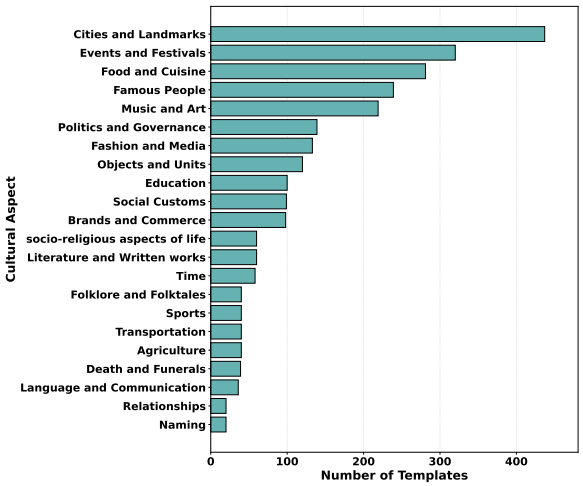

Figure 5: **Distribution of cultural-aspect tags in the benchmark.** Bars report the number of *template instantiations* tagged with each of the 22 cultural aspects. Aspects are not mutually exclusive, so a single item may contribute to multiple bars.

A.2 Reasoning-category coverage

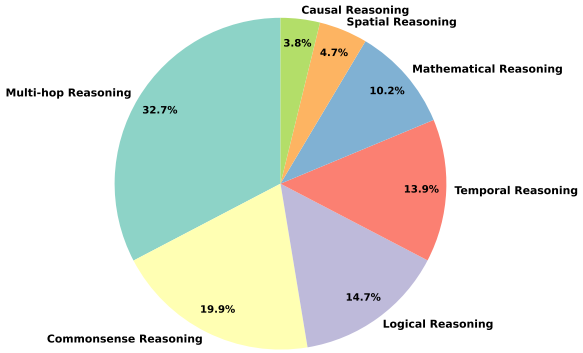

Figure 6: **Distribution of reasoning-category tags in the template set.** Percentages are normalized over tag assignments (multi-label items contribute multiple assignments).

B Verifier Agreement

To verify the factual correctness of cultural content and the clarity of local-language translations, we recruited three independent native verifiers per cultural context, distinct from the original annotators. Each verifier independently answered the MCQ version of every item in the local language, without access to the original annotations. Inter-annotator agreement (IAA) is defined as the proportion of

items on which all three verifiers selected the same answer. When verifiers disagreed, we applied majority voting to determine the accepted answer and flagged the item for further review; items identified by a majority as factually incorrect or ambiguous were revised accordingly.

Table 6 reports per-context IAA scores. Scores range from 87.0% (South Africa) to 94.8% (China), with a mean of 90.9%, confirming high factual accuracy and translation clarity across all 20 cultural contexts.

Country	IAA (%)	Country	IAA (%)
Brazil	92.00	Japan	91.80
China	94.80	Kyrgyzstan	92.50
Egypt	91.30	Mexico	93.43
Ethiopia	89.10	Morocco	90.46
Georgia	92.30	Nigeria	92.80
Greece	90.21	Philippines	88.00
India	91.30	South Africa	87.00
Indonesia	90.00	Thailand	88.60
Italy	88.97	Tunisia	89.70
Turkey	90.11	Yemen	89.80

Table 6: Inter-annotator agreement (IAA) across all 20 cultural contexts, defined as the proportion of items on which all three independent native verifiers selected the same answer.

C Annotator Demographics

Table 7 reports demographic information for the two annotators recruited per cultural context via Upwork, including gender, age group, duration of residence in the target culture, and education level. Annotators were required to be native speakers of the target language with substantial lived experience in the corresponding cultural context.

D Log-Probability vs. Generation Scoring

For open-weight models we primarily report log-probability scoring, which selects the answer option with the highest token-level log-likelihood. To confirm this choice does not introduce systematic bias, we re-evaluated all open-weight models using direct generation and compared results. Table 8 reports both protocols side by side. Differences are small and go in both directions, confirming the absence of systematic bias. The most notable exception is `meta-llama-llama-3.1-8b-instruct`, which drops substantially under generation in MC-EN (54.19 → 38.75) — a result consistent with smaller models being more prone to output-format failures rather than knowledge failures. This validates our choice to use log-probability scoring for open-weight models.

Model	Mode	MC-EN	MC-L	TF-EN	TF-L
coherelabs-aya-23-8b	logprob	43.79	39.15	50.35	50.45
	gen	40.00	36.00	50.00	50.00
coherelabs-aya-expanse-8b	logprob	52.67	48.74	51.74	52.47
	gen	54.02	47.75	56.73	54.91
meta-llama-llama-3.1-8b-instruct	logprob	54.19	43.39	56.66	53.36
	gen	38.75	46.25	53.12	51.88
meta-llama-llama-3.2-3b-instruct	logprob	47.38	36.33	54.87	51.64
	gen	44.86	33.87	47.75	49.64
qwen-qwen2.5-7b-instruct	logprob	56.96	46.92	52.65	52.62
	gen	55.54	48.15	58.27	55.49
qwen-qwen3-4b-instruct-2507	logprob	52.62	45.56	55.25	53.68
	gen	53.16	46.13	55.16	53.95

Table 8: Log-probability vs. generation scoring for open-weight models (accuracy, %). Differences are small and unsystematic, supporting log-probability scoring as the primary protocol.

E Answer Extraction Error Rates

For generation-based models, Table 9 reports three error types: (i) **Empty Output** — the model produced no output; (ii) **Answer Not Extracted** — the model produced output but no valid answer letter could be parsed; (iii) **Correct but Mis-scored** — the model's response contained the correct answer but was marked wrong due to an extraction failure. Error rates are near zero across all models, confirming that extraction failures do not meaningfully affect benchmark scores or conclusions.

Model	Empty Output	Not Extracted	Correct but Mis-scored
meta-llama-llama-3.2-3b-instruct	0.00	3.56	0.37
meta-llama-llama-3.1-8b-instruct	0.00	0.14	1.27
coherelabs-aya-23-8b	0.07	0.35	0.01
anthropic-claude-haiku-4.5	0.00	0.14	0.00
deepseek-deepseek-chat-v3.1	0.00	0.11	0.00
meta-llama-llama-3.3-70b-instruct	0.00	0.07	0.00
openai-gpt-5-chat	0.00	0.04	0.00
google-gemini-3-flash-preview	0.00	0.03	0.00
anthropic-claude-opus-4.5	0.00	0.02	0.00
qwen-qwen2.5-7b-instruct	0.00	0.02	0.00
meta-llama-llama-4-maverick	0.00	0.02	0.00
qwen-qwen3-4b-instruct-2507	0.00	0.01	0.00
google-gemini-2.5-flash	0.00	0.01	0.00
coherelabs-aya-expanse-8b	0.00	0.00	0.00
qwen-qwen3-235b-a22b-2507	0.00	0.00	0.00
openai-gpt-4o-mini	0.00	0.00	0.00

Table 9: Answer extraction error rates (%) for generation-based models. The "Correct but Mis-scored" rate reaches at most 1.27% (Llama 3.1 8B), confirming extraction failures do not significantly affect conclusions.

	EG	PH	IN	ET	MX	TN	GR	BR	KG	ZA	IT	TH	TR	GE	CN	ID	YE	NG	MA	JP
No. of annotators	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Gender (%)																				
Female	100	100	50	0	100	100	50	50	50	50	50	100	50	100	50	0	50	50	50	50
Male	0	0	50	100	0	0	50	50	50	50	50	0	50	0	50	100	50	50	50	50
Age (%)																				
20–24	50	50	0	50	0	0	0	0	0	0	0	0	0	0	0	0	50	0	50	0
25–35	50	50	50	50	50	100	0	100	100	50	50	50	100	0	100	100	50	100	0	100
36–45	0	0	50	0	0	0	100	0	0	50	0	50	0	100	0	0	0	0	50	0
46+	0	0	0	0	50	0	0	0	0	0	50	0	0	0	0	0	0	0	0	0
Duration of residence (years) (%)																				
Native / whole life	0	100	100	100	100	100	100	100	100	100	50	100	50	100	0	0	0	100	100	100
$\geq$20 years	100	0	0	0	0	0	0	0	0	0	50	0	0	0	100	100	100	0	0	0
10–19 years	0	0	0	0	0	0	0	0	0	0	0	0	50	0	0	0	0	0	0	0
$<$10 years	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Education level (%)																				
High school or below	0	0	0	0	0	0	50	0	0	50	0	0	0	0	0	0	0	0	0	0
Diploma / college	0	0	0	0	0	0	0	0	0	50	0	0	0	0	0	0	0	0	0	0
Bachelor's	0	100	100	100	50	0	0	50	100	0	0	0	100	50	0	0	100	50	50	50
Master's / postgrad	100	0	0	0	50	100	50	50	0	0	100	100	0	50	0	0	0	50	50	50
Doctorate	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	100	0	0	0	0

Table 7: **Annotator demographics by cultural context (2 annotators per context).** Percentages are multiples of 50% due to $n = 2$. Abbreviations: EG=Egypt, PH=Philippines, IN=India, ET=Ethiopia, MX=Mexico, TN=Tunisia, GR=Greece, BR=Brazil, KG=Kyrgyzstan, ZA=South Africa, IT=Italy, TH=Thailand, TR=Turkey, GE=Georgia, CN=China, ID=Indonesia, YE=Yemen, NG=Nigeria, MA=Morocco, JP=Japan.

F Evaluation Prompts

We evaluate each model under two prompting regimes. For **standard (non-thinking)** models, we request a direct answer without eliciting explanations. For **thinking-capable** models, we allow step-by-step reasoning but enforce a strict final-line JSON output format for automatic parsing.

F.1 Non-thinking models: Multiple-choice prompt

Non-Thinking Models: MCQ Prompt

```
Question: {question}

A) {option1}
B) {option2}
C) {option3}
D) {option4}

Answer directly without any
    explanation or reasoning.
Output a single line containing only
    a JSON object of the form:
{"answer": "X"}

Replace X with the correct letter
    (A, B, C, or D).
Do not include anything else.
```

F.2 Non-thinking models: True/False prompt

Non-Thinking Models: True/False Prompt

```
Statement: {question}

Answer directly without any
    explanation or reasoning.
Output a single line containing only
    a JSON object of the form:
{"answer": "ANSWER"}

Replace ANSWER with either "T" for
    True or "F" for False.
Do not include anything else.
```

F.3 Thinking models: Multiple-choice prompt

Thinking Models: MCQ Prompt

```
Question: {question}

A) {option1}
B) {option2}
C) {option3}
D) {option4}

First, you may think step by step to
    solve the question.
At the very end, output a single
    line containing only a JSON
    object of the form:
{"answer": "X"}
```

```
Replace X with the correct letter
    (A, B, C, or D).
Do not include anything else on that
    final line.
```

F.4 Thinking models: True/False prompt

Thinking Models: True/False Prompt

```
Statement: {question}

First, you may think step by step.
At the very end, output a single
    line containing only a JSON
    object of the form:
{"answer": "ANSWER"}

Replace ANSWER with either "T" for
    True or "F" for False.
Do not include anything else on that
    final line.
```

F.5 Data verification prompt (GPT-5.2-chat)

To improve linguistic quality while preserving meaning, we run an automated proofreading pass using GPT-5.2-chat that is restricted to spelling and grammar fixes only:

Data Verification Prompt (GPT-5.2-chat)

```
You are a careful proofreader. Your
    task is to check this {context}
    for ONLY grammatical and
    spelling mistakes.

**Instructions:**
1. Fix ONLY spelling errors and
    grammatical mistakes
2. Preserve the exact writing style,
    tone, and word choices
3. Keep all cultural references and
    proper nouns exactly as they are
4. Do NOT rephrase or improve the
    writing - only correct errors
5. If there are no errors, return
    the text exactly as-is

**Common errors to check for:**
- **Spelling mistakes**: "recieve"
    -> "receive", "occured" ->
    "occurred"
- **Plural/singular agreement**: "1
    foods" -> "1 food", "2 dish" ->
    "2 dishes"
- **Subject-verb agreement**: "they
    was" -> "they were", "he don't"
    -> "he doesn't", "that are" with
    singular -> "that is"
- **Verb tense for past events**:
    "Is X established in 1950?" ->
    "Was X established in 1950?" (BE
    CONSISTENT across similar
    questions)
- **Pronoun agreement**: Singular
    "their" is acceptable as
    gender-neutral. Keep it unless
    clearly wrong. Be CONSISTENT
    within the same question set.
- **Unnecessary apostrophes**:
    "apple's" (plural) -> "apples"
- **Capitalization**: Only fix clear
    errors like "i" -> "I" or
    well-known proper nouns
    (countries, famous people). DO
    NOT automatically capitalize
    cultural terms, food names, or
    words that might be
    intentionally lowercase.
- **Article errors**: "a apple" ->
    "an apple", "an hour" not "a
    hour"

**Text to check:**
{text}

**Response format:**
If there are errors:
{
  "has_errors": true,
  "corrected": "the corrected text
    here",
  "explanation": "brief explanation
    of what was fixed (e.g., 'Fixed
    plural/singular agreement')"
}

If no errors:
{
  "has_errors": false,
  "corrected": "{text}",
  "explanation": "No errors found"
}

Respond ONLY with valid JSON,
    nothing else.
```

G Example Templates by Cultural Aspect

Table 10 lists one representative template prompt per cultural aspect, illustrating how each aspect is operationalized across the 100 language-agnostic templates.

H Detailed Benchmark Results

This section reports accuracy on *local-language* evaluation instances (MC–L, T–L, F–L), aggregated either (i) by writing system (macro-averaging across languages that share a script) or (ii) by benchmark language/cultural context.

I Annotation Platform

Figure 7 shows the onboarding flow presented to annotators, including requirements for cultural authenticity and guidance for handling culturally spe-

Cultural Aspect: Example Templates (I)		Cultural Aspect: Example Templates (II)	
Aspect	**Template prompt**	**Aspect**	**Template prompt**
Food and Cuisine	`Which dish would [FAMOUS PERSON] probably not recognize from their childhood?`	Sports	`In [CULTURE SPORT] tradition, what happens when [CONDITION]?`
Music and Art	`Which traditional musical instrument from [COUNTRY/REGION] has the earliest recorded history?`	Fashion and Media	`I couldn't stop laughing watching a/an [CULTURE/REGION] series with [ACTOR NAME]. Which of the following is most likely the name of the series?`
Cities and Landmarks	`Among all the provinces in [SET OF PROVINCES/LOCATION], how many provinces have an area smaller than [PROVINCE]?`	Transportation	`If I live in [LOCATION 1/RESIDENTIAL AREA 1] and I want to go to [LOCATION 2/RESIDENTIAL AREA 2], how much time would it take on average if I traveled by [TRANSPORTATION METHOD]?`
Famous People	`Who among these [NATIONALITY] [FAMOUS PEOPLE TYPE] does NOT share the key trait of [COMMON TRAIT]?`	Education	`Which [ACADEMIC PERIOD] would a [AGE]-year-old typically be in according to [COUNTRY]'s education system?`
Politics and Governance	`What is the name of the first child of the [Nth] president/leader of [COUNTRY]?`	Agriculture	`In [REGION] during [MONTH], which crop is typically being [AGRICULTURAL ACTIVITY]?`
Events and Festivals	`Which of the following special days is the closest to [EVENT]?`	Naming	`If my friend has the last name "[LAST NAME]", which country is most likely their birthplace?`
Objects and Units	`In [CULTURE] traditional measurements, how many [UNIT] equal one [LARGER UNIT]?`	Folklore and Folktales	`In [CULTURE]'s folk tales, which character would be considered out of place if it appears alongside [CHARACTER]?`
Socio-religious Aspects of Life	`According to [NATIONALITY] cultural superstition, what should one do after [ACTION] to avoid bad luck?`	Brands and Commerce	`Among these local brands in [CULTURE/REGION], which one would a typical middle-income person be most likely to use?`
Language and Communication	`In [COUNTRY], which age group or social category is LEAST likely to use the expression "[COMMON PHRASE]" to describe [MEANING OF EXPRESSION TO THEM]?`	Death and Funerals	`In [CULTURE/REGION], how many days after death is [RITUAL/EVENT] traditionally performed?`
Social Customs	`In [COUNTRY/REGION], when [CONDITION/ACTIVITIES], which of the following actions is considered a taboo?`	Time	`If you convert [DATE] in the [CALENDAR SYSTEM] calendar to the Gregorian calendar in [YEAR], which month would it fall in?`
Relationships	`If I call my [NATIONALITY] father [TERM], would my children call him [OPTION A]?`	Literature and Written works	`What age-appropriate book can I buy for my [AGE]-year-old son?`

Table 10: Cultural aspects covered in the benchmark, with one example template per aspect.

cific items. Figure 8 and Figure 9 show the main annotation interface used to instantiate templates in English, translate them into the local language, and generate verification statements.

Category	Model	Arabic	Cyrillic	Devanagari	Georgian	Ge'ez	Greek	Han	Japanese	Latin	Thai
Closed (thinking)	Gemini 3 Flash Preview (thinking)	88.7%	85.0%	97.0%	90.4%	87.5%	88.5%	89.7%	85.9%	90.0%	87.9%
	Gemini 2.5 Pro (thinking)	86.3%	85.0%	95.5%	90.9%	80.5%	84.0%	87.1%	85.5%	90.0%	83.8%
	DeepSeek V3.1 (thinking)	67.9%	67.5%	94.0%	76.3%	55.5%	82.5%	85.1%	79.3%	79.0%	75.3%
	Gemini 2.5 Flash (thinking)	67.4%	67.5%	76.5%	60.1%	70.5%	64.0%	70.1%	76.3%	66.8%	71.2%
Average (Closed thinking)		77.6%	76.2%	90.8%	79.4%	73.5%	79.8%	83.0%	81.8%	81.4%	79.6%
Closed (standard)	Gemini 3 Flash Preview	85.1%	82.0%	96.5%	91.4%	81.5%	88.0%	88.1%	84.8%	87.2%	88.9%
	Claude Opus 4.5	75.6%	74.0%	94.0%	80.3%	67.0%	81.5%	87.1%	82.3%	83.9%	72.9%
	GPT-5 Chat	75.9%	72.5%	89.5%	79.3%	50.0%	81.5%	86.1%	80.8%	81.3%	79.3%
	Gemini 2.5 Flash	75.1%	71.0%	95.5%	81.8%	64.5%	78.5%	86.6%	83.8%	82.5%	78.3%
	DeepSeek V3.1	61.3%	60.5%	87.5%	65.7%	46.5%	71.5%	84.5%	78.3%	76.6%	73.2%
	Claude Haiku 4.5	62.8%	57.5%	85.5%	64.1%	39.0%	71.0%	83.0%	77.8%	70.1%	59.1%
	GPT-4o mini	64.9%	58.5%	81.5%	64.6%	38.0%	67.0%	79.9%	76.8%	70.7%	68.2%
Average (Closed standard)		71.5%	68.0%	90.0%	75.3%	55.2%	77.0%	85.0%	80.7%	78.9%	74.3%
Open-weight	Qwen3-235B	67.2%	58.0%	85.5%	64.1%	45.0%	72.0%	84.0%	80.3%	73.4%	70.7%
	Llama 3.3 70B	59.0%	60.5%	79.5%	58.6%	41.5%	73.5%	79.9%	78.3%	69.7%	62.6%
	Llama 4 Maverick	54.4%	61.0%	84.5%	66.7%	49.5%	70.0%	80.9%	67.6%	70.0%	71.7%
	Llama 3.1 8B	42.6%	45.5%	61.5%	44.4%	29.0%	55.0%	65.5%	58.1%	50.1%	47.0%
	Qwen3-4B	41.7%	36.5%	63.0%	42.4%	26.5%	52.0%	75.3%	62.1%	51.0%	50.5%
	InternLM3-8B	42.0%	34.0%	53.0%	39.9%	36.0%	44.5%	77.8%	58.1%	49.8%	50.0%
	Qwen2.5-7B	47.5%	42.0%	56.5%	40.4%	34.5%	53.0%	73.2%	70.7%	53.3%	53.0%
	Aya Expanse 8B	47.4%	36.5%	66.5%	39.4%	31.5%	60.5%	59.8%	62.1%	53.2%	43.4%
	Llama 3.2 3B	35.3%	31.5%	52.5%	31.3%	37.5%	46.5%	51.5%	49.0%	44.8%	38.4%
	Aya 23 8B	41.1%	33.5%	50.0%	31.8%	24.0%	46.5%	54.1%	53.0%	42.0%	37.4%
Average (Open-weight)		47.8%	43.9%	65.2%	45.9%	35.5%	57.4%	70.2%	63.9%	55.7%	52.5%
Average (All)		61.4%	58.1%	78.4%	62.1%	49.3%	68.2%	77.6%	72.9%	68.4%	64.9%

Table 11: Local-language accuracy (%) by writing system. Each cell reports a model's overall accuracy on local-language evaluation instances (MC–L, T–L, F–L) aggregated over all benchmark contexts sharing a given script. For scripts used by multiple languages (e.g., Arabic, Latin), scores are macro-averaged across those languages. *Average* rows report the mean over models within each group; **Average (All)** averages over all models.

Category	Model	Amh.	Pt-BR	Zho.	Egy-Ar	Geo.	Gre.	Hin.	Ind.	Ita.	Jpn.	Kyr.	Mx-Es	Mor-Ar	Tgl.	Tha.	Tun-Ar	Tur.	Yem-Ar	Yor.	Zul.
Closed (thinking)	Gemini 3 Flash Preview (thinking)	87.5%	94.0%	89.7%	92.9%	90.4%	88.5%	97.0%	93.2%	90.3%	85.9%	85.0%	91.9%	90.5%	88.9%	87.9%	83.0%	94.0%	88.5%	88.9%	78.5%
	Gemini 2.5 Pro (thinking)	80.5%	93.3%	87.1%	89.9%	90.9%	84.0%	95.5%	90.8%	89.3%	85.5%	85.0%	90.4%	88.0%	88.4%	83.8%	80.0%	94.0%	87.5%	83.7%	84.5
	DeepSeek V3.1 (thinking)	55.5%	88.5%	85.1%	73.2%	76.3%	82.5%	94.0%	84.2%	69.4%	79.3%	67.5%	84.3%	71.0%	79.3%	75.3%	52.5%	85.0%	75.0%	75.3%	66.0%
	Gemini 2.5 Flash (thinking)	70.5%	65.0%	70.1%	66.2%	60.1%	64.0%	76.5%	69.5%	62.8%	76.3%	67.5%	73.2%	69.5%	66.7%	71.2%	65.0%	77.5%	69.0%	59.5%	60.5%
Average (Closed thinking)		73.5%	85.2%	83.0%	80.5%	79.4%	79.8%	90.8%	84.4%	78.0%	81.8%	76.2%	85.0%	79.8%	80.8%	79.6%	70.1%	87.6%	80.0%	76.8%	68.3%
Closed (standard)	Gemini 3 Flash Preview	81.5%	90.0%	88.1%	83.8%	91.4%	88.0%	96.5%	94.2%	88.3%	84.8%	82.0%	87.4%	88.0%	84.8%	88.9%	82.0%	90.5%	86.5%	83.2%	79.0%
	Claude Opus 4.5	67.0%	89.6%	87.1%	71.1%	80.3%	81.5%	94.0%	91.6%	87.2%	82.3%	74.0%	86.9%	75.3%	78.8%	72.9%	76.5%	86.0%	79.6%	76.1%	75.0%
	GPT-5 Chat	50.0%	87.0%	86.1%	74.7%	79.3%	81.5%	89.5%	81.6%	82.7%	80.8%	72.5%	83.3%	77.0%	79.8%	79.3%	74.0%	83.5%	78.0%	78.9%	73.5%
	Gemini 2.5 Flash	64.5%	83.5%	86.6%	74.2%	81.8%	78.5%	95.5%	86.8%	89.8%	83.8%	71.0%	86.4%	79.5%	80.3%	78.3%	76.5%	84.0%	70.0%	78.4%	71.0%
	DeepSeek V3.1	46.5%	84.5%	84.5%	59.1%	65.7%	71.5%	87.5%	77.4%	86.7%	78.3%	60.5%	77.3%	64.0%	75.8%	73.2%	56.5%	79.0%	65.5%	70.0%	62.0%
	Claude Haiku 4.5	39.0%	77.5%	83.0%	62.1%	64.1%	71.0%	85.5%	68.9%	81.1%	77.8%	57.5%	71.2%	64.0%	69.7%	59.1%	57.0%	72.5%	68.0%	61.8%	58.0%
	GPT-4o mini	38.0%	75.5%	79.9%	63.6%	64.6%	67.0%	81.5%	71.1%	82.1%	76.8%	58.5%	72.7%	66.0%	70.2%	68.2%	62.5%	75.0%	67.5%	60.5%	58.5%
Average (Closed standard)		55.2%	83.9%	85.0%	69.8%	75.3%	77.0%	90.0%	81.7%	85.4%	80.7%	68.0%	80.7%	73.4%	77.1%	74.3%	69.3%	81.5%	73.6%	72.7%	68.1%
Open-weight	Qwen3-235B	45.0%	77.5%	84.0%	67.7%	64.1%	72.0%	85.5%	75.8%	80.6%	80.3%	58.0%	79.8%	62.5%	76.3%	70.7%	67.0%	75.5%	71.5%	62.6%	59.0%
	Llama 3.3 70B	41.5%	71.5%	79.9%	59.1%	58.6%	73.5%	79.5%	73.7%	80.6%	78.3%	60.5%	70.2%	60.0%	72.2%	62.6%	54.5%	72.5%	62.5%	61.1%	56.0%
	Llama 4 Maverick	49.5%	75.0%	80.9%	55.6%	66.7%	70.0%	84.5%	70.8%	78.1%	67.6%	61.0%	70.7%	65.5%	74.7%	71.7%	61.5%	71.0%	35.0%	56.8%	62.5%
	Llama 3.1 8B	29.0%	42.5%	65.5%	40.4%	44.4%	55.0%	61.5%	52.6%	60.7%	58.1%	45.5%	54.5%	44.5%	46.5%	47.0%	41.5%	55.5%	44.0%	45.3%	43.0%
	Qwen3-4B	26.5%	64.0%	75.3%	43.4%	42.4%	52.0%	63.0%	44.7%	64.3%	62.1%	36.5%	52.5%	46.0%	45.5%	50.5%	40.0%	48.0%	37.5%	43.2%	46.0%
	InternLM3-8B	36.0%	46.5%	77.8%	33.8%	39.9%	44.5%	53.0%	47.9%	63.3%	58.1%	34.0%	57.1%	46.0%	49.0%	50.0%	45.0%	46.0%	43.0%	45.8%	42.5%
	Qwen2.5-7B	34.5%	62.5%	73.2%	50.0%	40.4%	53.0%	56.5%	47.9%	62.2%	70.7%	42.0%	61.6%	48.0%	51.5%	53.0%	45.5%	50.0%	46.5%	46.8%	43.5%
	Aya Expanse 8B	31.5%	60.0%	59.8%	44.4%	39.4%	60.5%	66.5%	51.6%	66.8%	62.1%	36.5%	55.1%	50.0%	50.0%	43.4%	45.0%	55.5%	50.0%	45.3%	41.5%
	Llama 3.2 3B	37.5%	44.0%	51.5%	36.9%	31.3%	46.5%	52.5%	42.1%	54.1%	49.0%	31.5%	48.0%	36.5%	46.0%	38.4%	32.5%	44.5%	35.5%	37.4%	42.5%
	Aya 23 8B	24.0%	45.5%	54.1%	34.8%	31.8%	46.5%	50.0%	42.6%	54.1%	53.0%	33.5%	54.5%	43.5%	37.4%	37.4%	42.0%	40.5%	44.0%	35.8%	25.5%
Average (Open-weight)		35.5%	58.9%	70.2%	46.6%	45.9%	57.4%	65.2%	55.0%	66.5%	63.9%	43.9%	60.4%	50.3%	54.9%	52.5%	47.5%	55.9%	46.9%	48.0%	46.2%
Average (All)		49.3%	72.3%	77.6%	60.8%	62.1%	67.2%	78.4%	68.4%	73.0%	72.9%	57.2%	71.7%	61.9%	65.4%	62.7%	59.0%	70.5%	62.1%	61.7%	57.2%

Table 12: Local-language accuracy (%) by benchmark language/cultural context. Each cell reports a model's overall accuracy on local-language instances (MC–L, T–L, F–L) for the corresponding context, aggregated over all templates. Missing entries (–) are ignored in averages. *Average* rows report the mean over models within each group; **Average (All)** averages over all models. Column headers are abbreviated: Amh.=Amharic, Pt-BR=Brazilian Portuguese, Zho.=Chinese, Egy-Ar=Egyptian Arabic, Geo.=Georgian, Gre.=Greek, Hin.=Hindi, Ind.=Indonesian, Ita.=Italian, Jpn.=Japanese, Kyr.=Kyrgyz, Mx-Es=Mexican Spanish, Mor-Ar=Moroccan Arabic, Tgl.=Tagalog, Tha.=Thai, Tun-Ar=Tunisian Arabic, Tur.=Turkish, Yem-Ar=Yemeni Arabic, Yor.=Yoruba, Zul.=Zulu.

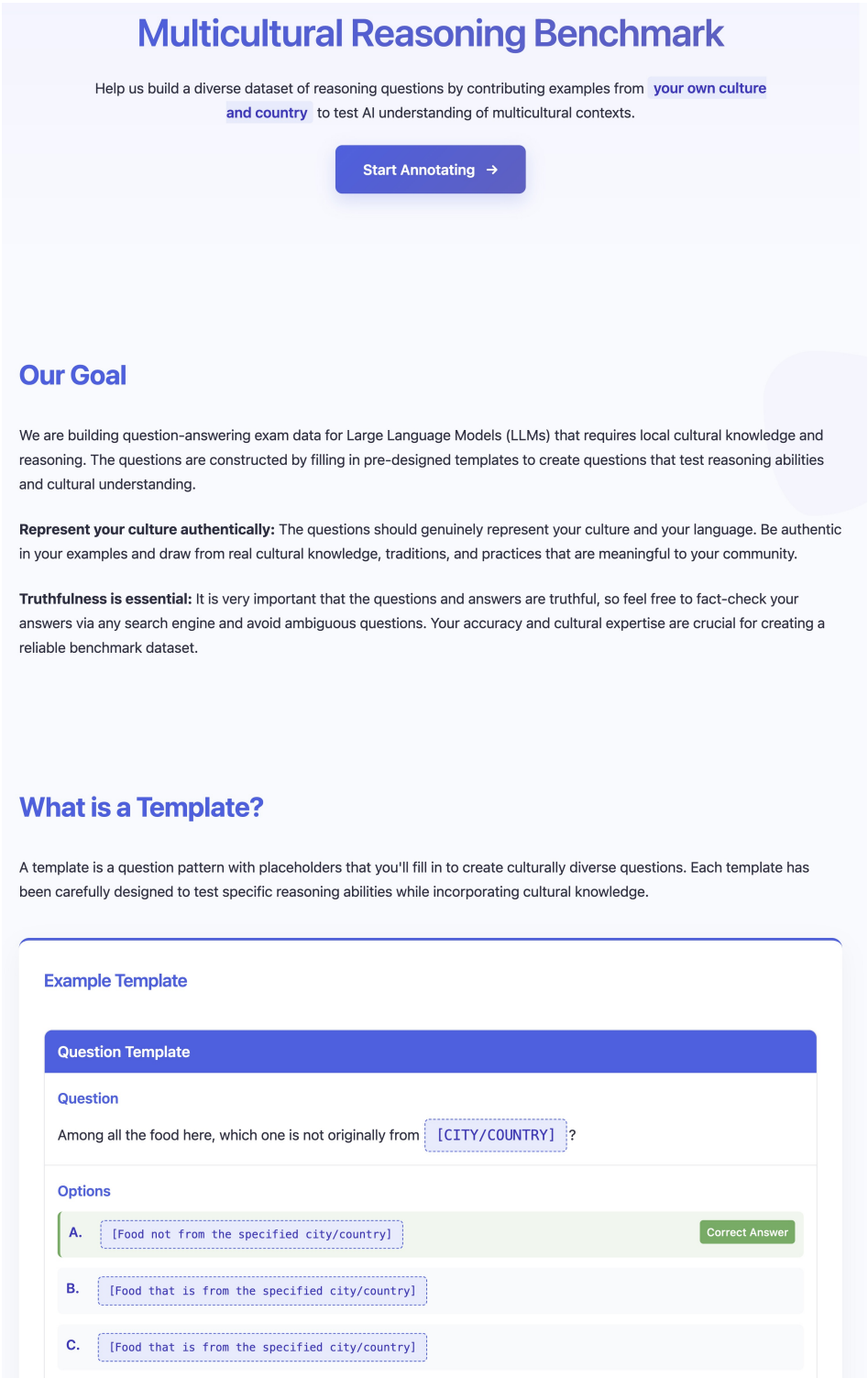

Figure 7: Onboarding landing page (Part 1 of 3).

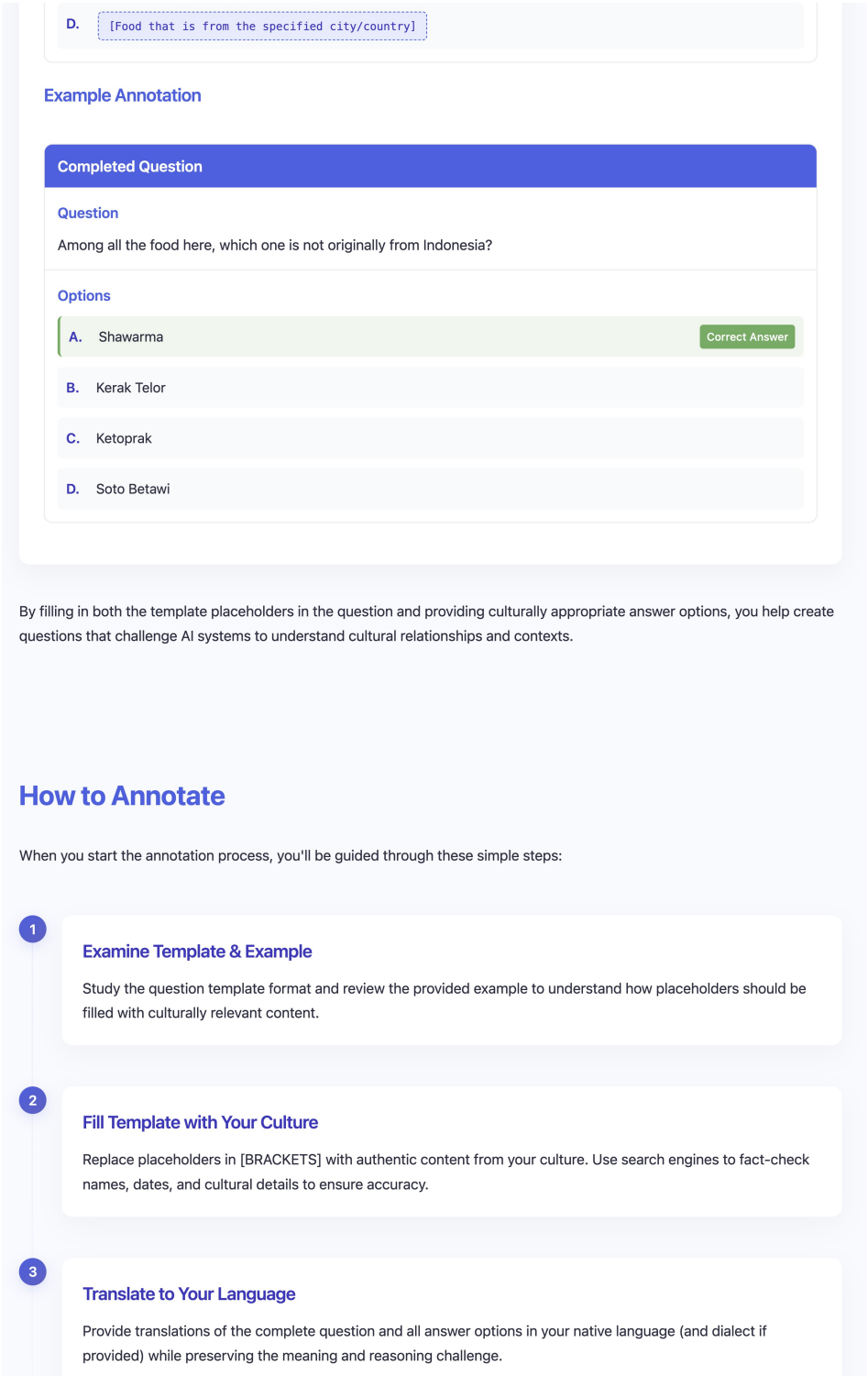

Figure 7: Onboarding landing page (Part 2 of 3).

4

Review and Submit

Double-check your work for accuracy and authenticity, then submit your completed annotation to contribute to the multicultural reasoning dataset.

Cultural Translation Guidelines

When filling placeholders in the English version of the questions with Culturally-Specific Items (CSIs), it's important to preserve authenticity. Follow these guidelines:

✅ CSI with Possible Translation

When there's a culturally-equivalent term that conveys similar meaning, you may translate:

Examples:

- **Well-known landmarks:** "Great Wall of China", "Eiffel Tower" (internationally recognized names)
- **Geographic features:** "Mount Jais", "Nile River" (use English descriptor + local name)
- **Common concepts:** Mexican "tianguis" → "flea market", "Fiesta" → "festival" (when equivalent concept exists)

❌ CSI that Should Stay Original

When translation would lose cultural meaning or authenticity, keep the original term:

Examples:

- **Traditional foods:** "Hawawshi", "Biryani", "Sushi", "Kimchi" (not generic descriptions like "meat sandwich")
- **Local places:** "Tahrir Square" not "Liberation Square" (even though "tahrir" translates to "liberation", keep the name as locals and visitors know it from maps and tourism)
- **Musical instruments:** "Tabla", "Balalaika" (different from generic "drum" or "string instrument")

Golden Rule: Use the name that best preserves cultural authenticity and would be recognized by both locals and international visitors. When in doubt, keep it authentic!

Figure 7: Onboarding landing page (Part 3 of 3).

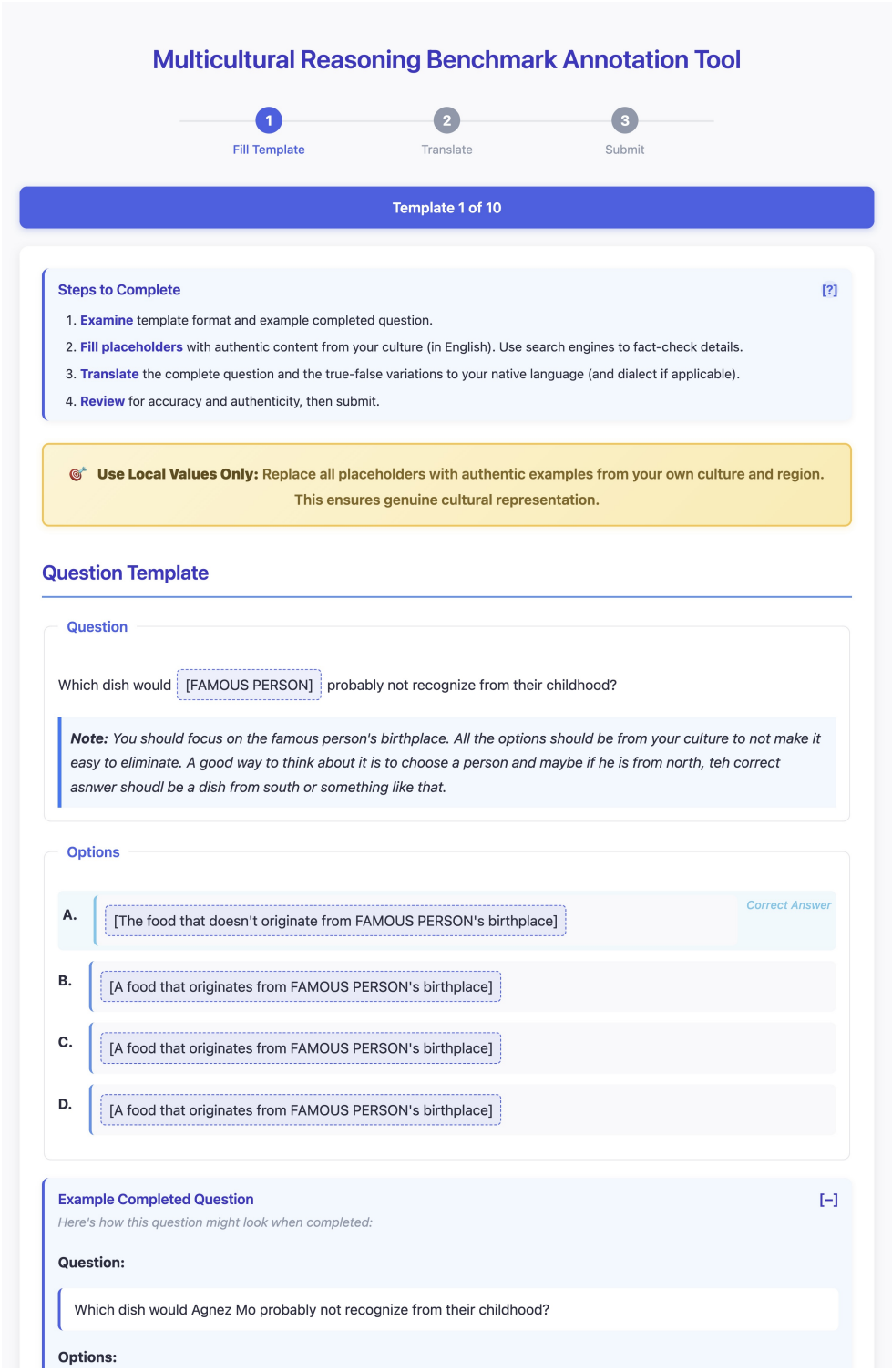

Figure 8: Annotation interface (Part 1 of 2): **Contextualization**.

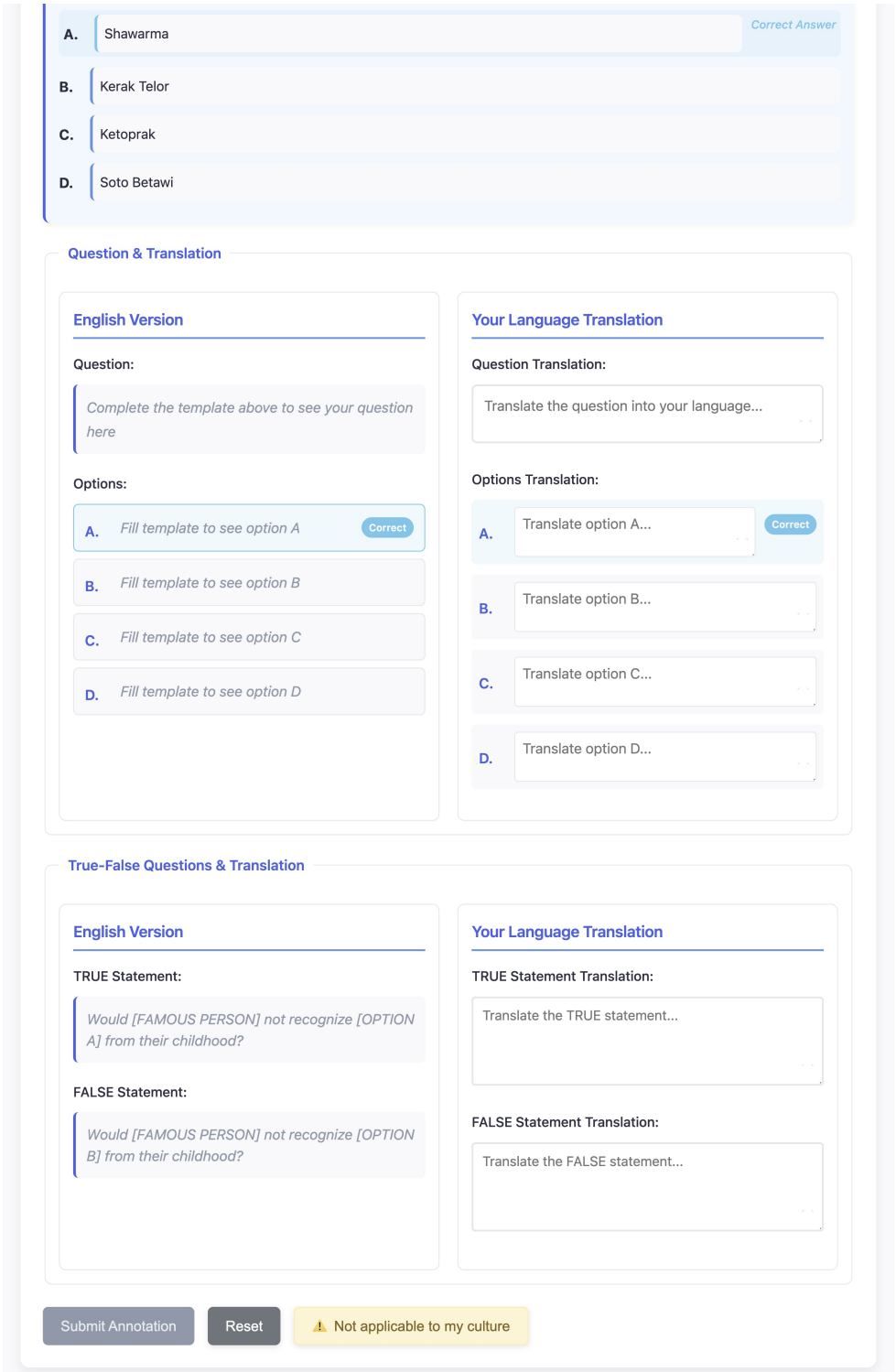

Figure 9: Annotation interface (Part 2 of 2): **Instantiation and translation**.