Cultural Benchmark Annotation Guideline

XXX

October 15, 2025

1 Before Annotation

1.1 Sample Format

A sample consists of five fields: topic, scenario, question, answer, and explanation.

[Topic]
Commerce
[Scenario]
I run an online flower shop and plan to launch a promotion at the end of August 2025.
[Question]
Which type of flower should I feature?
[Answer]
Roses
[Explanation]
August 29 is Qixi Festival (Chinese Valentine's Day).

Figure 1: Chinese case translated into English (1).

[Topic]
Travel
[Scenario]
I was traveling in Guoluo recently, and whenever the locals saw me they stuck out their tongues.
[Question]
What was their intention?
A. Asking me for water
B. Showing friendliness
C. Showing dislike
D. No particular meaning
[Answer]
B
[Explanation]
Guoluo is part of a Tibetan autonomous prefecture, and sticking out one's tongue is a gesture of highest respect there.

Figure 2: Chinese case translated into English (2).

Topic Select the most appropriate topic from the predefined topic list (16 topics in Tab. 1, the complete set of seed examples is in **data_creation/cultural_topics.xlsx**) for the created sample.

Topic	Description	Example 1	Example 2	Example 3
Belief	Systems of conviction that shape values, rituals, institutions, life-cycle events, and views on existence—covering religious faith, spiritual practice, secular ethics, and cultural traditions (e.g., funerary customs and ideas of an afterlife).	Typical length and order of a wedding ceremony	Dietary restrictions during major religious holidays	Whether to pull the lever in the classic trolley-problem dilemma
Commerce	Buying, selling, marketing, and payment of goods and services—from daily necessities to luxury fashion—across bricks-and-mortar shops, e-commerce sites, and mobile wallets.	Typical opening hours for supermarkets	Return policy for online purchases	Legal limits on alcohol sales in retail stores
Education	Formal and informal learning, teaching, research, and skill-building for all ages, settings, and disciplines.	Courses normally taken in middle school	National university-entrance-exam format	Grading scale used in secondary schools
Entertainment	Media, arts, sports, games, performances, hobbies, and events created for leisure and enjoyment.	Popular sport clubs	National mascots or iconic cartoon characters	Gambling age and casino legality
Finance	Earning, saving, budgeting, investing, insuring, transferring, and distributing wealth during life and after death.	Color that signals a stock-price rise or fall on trading screens	Common payment methods in everyday shopping	Typical tax-filing deadline for individuals
Food	Agriculture, sourcing, processing, cooking, nutrition, beverages, and dining culture from farm to table.	Typical breakfast foods	Is tipping expected in restaurants?	Common allergens that must be listed on packaged food
Government	Public policy, legislation, courts, law enforcement, defense, emergency response, and civic administration.	Highway speed limits	Emergency number to call when lost in the mountains	Length of mandatory military or civil service
Habitat	Homes, buildings, infrastructure, utilities, urban planning, ecosystems, weather patterns, and sustainability practices.	Typical home-heating system	Floor-numbering convention in multi-story buildings	Recycling rules for household waste
Health	Physical, mental, and emotional well-being—prevention, treatment, fitness, wellness, palliative, and end-of-life care.	Standard childhood-vaccination schedule	Prescription vs. over-the-counter drug availability	Legal age of consent for medical decisions
Heritage	Past events, living traditions, festivals, monuments, and other cultural inheritances—and their study, preservation, and commemoration.	Date and rituals of New-Year celebrations	Historic event marked by a public holiday	Customs from a particular historical period
Language	Official and minority languages, scripts, dialects, idioms, emotional nuance, politeness levels, sign language, literacy, and translation norms.	Order of family and given names on official documents	Appropriate greetings and honorifics in business	Meaning and proper use of a common proverb
Pets	Care, health, training, companionship, and welfare of domesticated animals.	Rules for bringing pets on public transport	Mandatory rabies vaccination for dogs	Cultural status of certain animals
Science	Systematic inquiry into the natural world and its applications—research, engineering, technology, and innovation.	Unit used to state distance between two cities	Standard format for writing dates	Whether smartphones support dual-SIM use
Social	Family, friendships, romance, community networks, demographics, and social issues.	Table etiquette at family gatherings	Meaning of two women holding hands in public	Typical blind-dating process
Travel	Planning, transport, logistics, accommodation, tourism, and movement of people or goods.	Information needed before booking a city trip	Visa rules for a 90-day tourist stay	Cost of popular tourist attractions
Work	Careers, labor markets, workplaces, productivity tools, and professional development.	Statutory length of paid annual leave	Legal steps for ending an employment contract	Region-specific unique occupations

Table 1: Cultural topics with concise descriptions and illustrative examples.

Scenario Construct a plausible real-world situation, withholding any explicit hints that would let a model solve the task without relying on relevant cultural knowledge.

Question Ensure the query arises naturally from the scenario and cannot be answered correctly without an understanding of the relevant cultural knowledge.

Answer To facilitate automatic evaluation, answers should be objective and as brief as possible. If an objective free-form answer is impractical, convert the question to a four-option multiple-choice format (A–D) and return only the chosen letter.

Explanation When appropriate, supply the cultural or domain knowledge that supports the answer.

1.2 Sample Requirements

1.2.1 Using native languages

Each annotator composes samples in the dominant language of the related culture, e.g., English in the United States and Mandarin Chinese in mainland China.

1.2.2 Culture-related

Cultural knowledge includes but not limited to local vocabulary, social norms, cultural commonsense, regulations, and domain-specific knowledge. Generic trivia (e.g., math puzzles or textbook facts) is out of scope.

1.2.3 Factually correct

1.2.4 Objective and brief answer

1.2.5 Challenging

The benchmark is required to be challenging for LLMs. Human annotators are required to try to apply one or more of these techniques to enhance sample difficulty.

Long-Tail Swap Common entities are replaced with rarer or more specific ones—for instance, substituting the general location "Hong Kong" with "MacLehose Trail," a lesser-known hiking route within the region.

More/Less Context Additional situational details are introduced, requiring the answer to hinge on conditional, multi-step reasoning (e.g., determining if a traveler has a prior visa). Conversely, unnecessary context that could inadvertently provide hints to LLMs can be removed to increase the challenge.

Compositional Example Two independent knowledge points are combined into a single query—for example, merging the entry requirements for both Hong Kong and Bangkok—forcing the model to engage in compositional reasoning.

2 Start Annotation

2.1 Data Modification

You are provided with data generated by the LLM based on cultural datasets, following predefined schemas. The table below lists the important fields:

Field	Description
`source_excerpt`	Quoted passage from the original source, used as context for generation
`topic`	One of the predefined 16 cultural topics used in the dataset
`scenario`	A narrative context in which the question is framed or asked
`question`	The formulated question based on the scenario
`answer`	The correct answer (answer key) to the question
`explanation`	Concise justification or reasoning for the answer

Table 2: Schema fields in the provided dataset. These enforce consistency and structure across all data items.

Annotation Instructions:

1. **File Identification:**
   - Copy the `.xlsx` file you are working on to create a new instance for annotation.
   - Append your ID to the filename to ensure traceability of the annotated document.
   - For example, rename the file from `ar_ma_gpt4o` to `ar_ma_gpt4o_linpq`.
2. **Annotation Tasks:**
   - For each row in the `.xlsx` file, modify **COLUMN M** as part of your annotation, specifying whether to **Accept**, **Revise**, or **Reject** the item.
   - If choosing **Revise**, update the relevant components (scenario, question, answer, explanation) to reflect the corrections based on `source_excerpt`.
   - Annotation options and their criteria:
     - **Accept**: Select this option if the generated sample (scenario, question, answer, explanation) is correctly transformed according to `source_excerpt`.
     - **Revise**: Choose this option if the generated sample requires modifications based on `source_excerpt`.
     - **Reject**: Use this option if revising is impossible for the following reasons:
       * `source_excerpt` contains incorrect information.

* `source_excerpt` is subjective and not objective.
* The generated sample or transformation cannot be reasonably revised.

2.2 Data Creation

Some important fields and their descriptions are listed below:

Field	Description
`topic`	One of the 16 predefined cultural topics
`scenario`	A narrative context in which the question is framed or asked
`question`	The formulated question corresponding to the scenario
`answer`	The correct answer or answer key to the question
`explanation`	A concise justification or reasoning for the answer

Table 3: Metadata fields for each dataset item. These fields ensure consistency and adherence to predefined schemas.

You can create the data using the following sources:

- **Personal experience:** Draw examples from your own perspective or cultural knowledge.
- **Local online forums:** Use credible forums or community discussions as inspiration.
- **Topic list with examples:** Refer to the predefined topics and examples provided in `data_creation/cultural_topics.xlsx`.
- **Multilingual data:** Leverage annotated data from other languages available in `data_creation/annotated_data.csv`.

When creating your own dataset:

- Save the file as a `.xlsx` and name it using the format: `LanguageCode_CountryCode_Id.xlsx`.
- For example: `zh_cn_linpq.xlsx`.

Keep in mind the following guidelines when creating samples:

- Ensure the samples adhere to the specified requirements (§1.2).
- Focus on creating challenging examples to test model performance effectively.
- It is not necessary to maintain a balanced distribution across topics.