Article:
Permadi, V.A., Tan, X., Moosavi, N.S. et al. (2026) No shortcuts to culture: Indonesian multi-hop question answering for complex cultural understanding. Transactions of the Association for Computational Linguistics, 14. pp. 1243-1265. ISSN: 2307-387X

https://doi.org/10.1162/tacl.a.726

No Shortcuts to Culture: Indonesian Multi-hop Question Answering for Complex Cultural Understanding

Vynska Amalia Permadi[1,2] **Xingwei Tan**[1] **Nafise Sadat Moosavi**[1] **Nikolaos Aletras**[1]
[1]School of Computer Science, University of Sheffield, United Kingdom
[2]Department of Informatics, Universitas Pembangunan Nasional "Veteran" Yogyakarta, Indonesia
{vpermadi1,xingwei.tan,n.s.moosavi,n.aletras}@sheffield.ac.uk

Abstract

Understanding culture requires reasoning across context, tradition, and implicit social knowledge, far beyond recalling isolated facts. Yet most culturally focused question answering (QA) benchmarks rely on single-hop questions, which may allow models to exploit shallow cues rather than demonstrate genuine cultural reasoning. In this work, we introduce ID-MoCQA, the first large-scale multi-hop QA dataset for assessing the cultural understanding of large language models (LLMs), grounded in Indonesian traditions and available in both English and Indonesian. We present a new framework that systematically transforms single-hop cultural questions into multi-hop reasoning chains spanning six clue types (e.g., commonsense, temporal, geographical). Our multi-stage validation pipeline, combining expert review and LLM-as-a-judge filtering, ensures high-quality question-answer pairs. Our evaluation across state-of-the-art models reveals substantial gaps in cultural reasoning, particularly in tasks requiring nuanced inference. ID-MoCQA provides a challenging and essential benchmark for advancing the cultural competency of LLMs.[1]

1 Introduction

Developing large language models (LLMs) that can truly understand unwritten social norms, diverse local traditions, and cultural knowledge is important for the development of systems that can effectively avoid cultural insensitivities or misunderstandings, reinforcing stereotypes, or causing offence (Myung et al., 2024; Pawar et al., 2025).

Recent research has focused on developing resources for assessing cultural knowledge of LLMs, especially on low-resource languages (Myung et al., 2024; Putri et al., 2024).

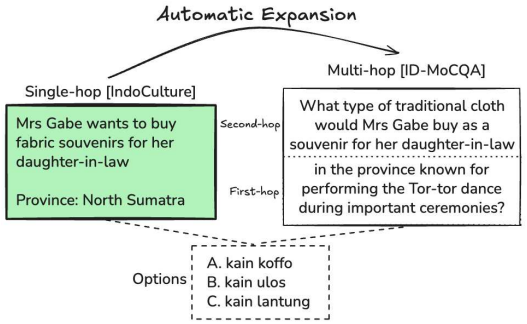


Figure 1: Single to multi-hop transformation from IndoCulture (Koto et al., 2024) to ID-MoCQA. Left: Original question about fabric souvenirs with origin province. Right: Our expansion requires first predicting the province (*North Sumatra*) through cultural clues (*Tor-tor dance*), then answering the question.

However, the majority of these benchmarks are built around single-hop question answering, where the answer can be retrieved directly from a single fact or cue. While effective for measuring factual knowledge, such setups often fail to probe whether models can reason through more complex, interrelated cultural knowledge concepts (Wang et al., 2024; Kim et al., 2024; Koto et al., 2024).

By contrast, multi-hop QA aims to evaluate deeper reasoning. Datasets such as HotpotQA (Yang et al., 2018), 2WikiMultiHopQA (Ho et al., 2020), and MuSiQue (Trivedi et al., 2022) challenge models to combine multiple pieces of evidence to reach an answer, reducing the likelihood of shortcut exploitation. Applying this multi-hop paradigm to the cultural domain is a natural next step, one that enables us to test whether models can interpret cultural clues, connect context, and infer appropriate practices.

In this work, we present a new framework to address this gap by transforming culturally grounded single-hop questions into two-hop QA instances that simulate realistic cultural reasoning (Figure

[1]Dataset is available at 🤗https://huggingface.co/datasets/vynsk/ID-MoCQA

Transactions of the Association for Computational Linguistics, vol. 14, pp. 1243–1265, 2026. https://doi.org/10.1162/TACL.a.726
Action Editor: Ellie Pavlick. Submission batch: 11/2025; Revision batch: 1/2026; Published 7/2026.

1). Our multi-hop structure tests whether LLMs understand not just cultural facts, but their contextual application: models must first identify relevant cultural context, then select practices appropriate to that context. To ensure correctness, we first prompt an LLM to add one intermediate reasoning step that connects to the original context while ensuring that the added step is relevant for answering the final multi-hop question. We further incorporate a multi-stage validation process that combines expert annotation and LLM-as-a-judge (Zheng et al., 2023) evaluation. This framework results in ID-MoCQA, the first large-scale multi-hop cultural QA dataset focused on a single national context: Indonesia. ID-MoCQA contains 15,590 questions, equally distributed across six clue types and two languages (English and Indonesian). Our extensive evaluation across a range of open, frontier, and region-specific LLMs reveals clear limitations in multi-hop cultural reasoning, even among top-performing models. Our contributions can be summarized as follows:

- We propose a new comprehensive framework for generating cultural multi-hop questions from existing single-hop data.
- We release **ID-MoCQA**, a human-verified dataset of 15,590 multi-hop questions in Indonesian and English about Indonesian culture generated by our proposed framework.
- We conduct extensive evaluation, including a diverse collection of open and frontier LLMs on **ID-MoCQA**, revealing persistent challenges in cultural multi-hop reasoning and establishing the dataset as a robust benchmark for future research.

2 Related Work

2.1 Cultural Competence in Sociolinguistics

Cultural competence is the ability to communicate and act appropriately within different communities and contexts, including knowing when, how, and to whom communications are suitable (Hymes, 1972; Byram, 1997). The same statement can be acceptable in one community but not in another, depending on factors such as social relationships, status, and context (Goffman, 1967; Brown and Levinson, 1987). Previous work has identified systematic patterns in cultural differences. Analysing high- and low-context communication shows whether meaning relies on shared cultural knowledge or on explicit verbal content (Hall, 1976). Cultural dimensions, including individualism-collectivism and power distance, also describe how cultures differ in communication expectations (Hofstede, 2011). These patterns may also guide behaviour in common social situations such as greetings, requests, and expressions of gratitude (Wierzbicka, 1994).

2.2 Cultural QA Benchmarks

Measuring and modeling culture in LLMs has emerged as a critical research area (Adilazuarda et al., 2024), with growing recognition that cultural understanding encompasses more than factual knowledge (Zhou et al., 2025). In response, specialized cultural QA benchmarks have been developed to evaluate LLM performance across diverse socio-cultural contexts. BLEnD (Myung et al., 2024) offers over 52,000 QA pairs on daily life and socio-cultural topics, spanning 16 countries and 13 languages. NativQA (Hasan et al., 2025) provides a semi-automatic framework for building culturally aligned QA datasets. It includes approximately 64,000 manually annotated and 55,000 automatically generated pairs across seven languages and 18 topics from nine regions. Complementing these approaches, WorldValuesBench (HRMCR) (Zhao et al., 2024) evaluates multicultural value understanding through scenarios inspired by the World Values Survey, while CultureAtlas (Fung et al., 2024) compiles culturally rich Wikipedia knowledge representing diverse sub-country regions and ethnolinguistic groups. INCLUDE (Romanou et al., 2025) addresses multilingual evaluation gaps with over 197,000 QA pairs from local exams in 44 languages, grounding evaluation in regional settings rather than English translations.

Recognizing the need for deeper cultural authenticity and representation, researchers have developed language-specific benchmarks that capture specific regional contexts and linguistic nuances in under-represented languages. For Indonesian, IndoCloze (Koto et al., 2022) offers short narratives assessing story comprehension and commonsense reasoning with causal and temporal understanding requirements. ID-CSQA (Putri et al., 2024) includes 9,000 culturally relevant commonsense QA pairs for Indonesian and Sundanese, created through LLMs and human anno-

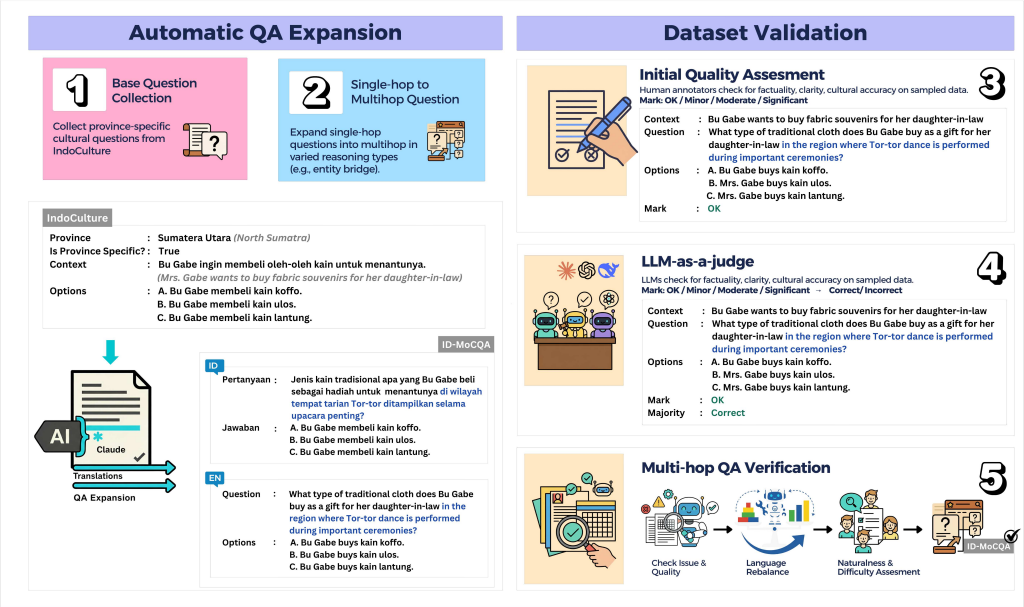


Figure 2: ID-MoCQA dataset creation pipeline. **Left (Automatic QA Expansion)**: (1) Collection of province-specific questions from IndoCulture; (2) Expansion to multi-hop questions using *Claude-3.7-Sonnet* with varied clue types, generating bilingual (Indonesian/English) versions. **Right (Dataset Validation)**: (3) Human assessment of factuality, clarity, and cultural accuracy; (4) Quality verification via LLM-as-a-judge; (5) Multi-hop verification to check quality, ensure language balance, and assess naturalness and difficulty.

tation. COPAL-ID (Wibowo et al., 2024), crafted by native speakers, focuses on natural causal reasoning in both standard and Jakartan Indonesian to capture local context. IndoCulture (Koto et al., 2024) is a benchmark developed through collaborative discussions with Indonesian natives, ensuring comprehensive coverage of diverse cultural aspects from 11 provinces across 6 islands of the Indonesian archipelago. Each province represents distinct ethnic groups, regional languages, and religious practices. Korean benchmarks include KULTURE Bench (Wang et al., 2024), featuring cultural news, idioms, and poetry, and CLIcK (Kim et al., 2024), offering 1,995 QA pairs from official exams and textbooks with fine-grained cultural and linguistic knowledge annotations. CAMeL (Naous et al., 2024) provides an Arabic benchmark contrasting Arab and Western cultures across tasks such as story generation and sentiment analysis. More recently, CulturalBench (Chiu et al., 2025) introduced 1,696 human-verified questions covering 45 global regions through a human-AI collaborative approach, which represents an advance in robust cultural knowledge evaluation. Despite these advances, all existing cultural QA datasets focus exclusively on single-hop questions.

2.3 Automatic QA Data Construction

LLMs have demonstrated potential for QA dataset generation through prompting strategies. CulturePark (Li et al., 2024) uses LLMs to generate diverse single-hop cross-cultural reasoning questions at scale. Shah et al. (2024) introduce planned query guidance using few-shot examples to enable systematic multi-hop reasoning over knowledge graphs. Cultural applications include ID-CSQA for Indonesian commonsense reasoning (Putri et al., 2024), NativQA for multilingual cultural alignment (Hasan et al., 2025), and WikiQA-IS for Icelandic cultural knowledge (Arnardóttir et al., 2025). However, the intersection of cultural authenticity and multi-hop complexity presents unique challenges. Ensuring both cultural accuracy and valid reasoning structures requires careful methodology, particularly when dealing with culture-specific knowledge that may be underrepresented in LLM training data.

3 Multi-hop QA Generation Framework

Our aim is to expand single-hop cultural questions to multi-hop. Figure 2 presents our com-

prehensive framework for building ID-MoCQA, which consists of two main components: (1) *Automatic QA Expansion* (Steps 1–2), which systematically transforms single-hop questions into multi-hop questions through LLM-guided generation; and (2) *Dataset Validation* (Steps 3–5), which implements a multi-stage quality assurance process combining human expertise with LLM verification to ensure dataset reliability.

3.1 Base Single-hop Question Collection

First, we derive our initial single-hop QA pairs from the IndoCulture dataset (Figure 2, Step 1). Each pair has a label distinguishing between province-specific and general cultural elements (*True* or *False*), indicating whether the cultural element uniquely pertains to the province. Figure 1 (top left) shows an example of a single-hop instance with its associated location. We exclusively select instances marked as *True*, representing practices unique to particular provinces, so we can use the province names as a first-hop link (i.e., *region where Tor-tor dance is performed during important ceremonies*). This yields 1,847 province-specific QA pairs, which serve as the foundation for multi-hop expansion.

3.2 From Single-hop to Multi-hop

Our framework transforms the manually curated high-quality single-hop questions from IndoCulture (Figure 2, Step 2). We build the first-hop by converting the province information into clues that require geographical, temporal, commonsense, or other cultural reasoning. This design introduces multi-step reasoning into the question while preserving the cultural authenticity of IndoCulture. In IndoCulture, the input consists of the province name, context (i.e., *Mrs. Gabe wants to buy fabric souvenirs for her daughter-in-law*), and options (i.e., *A. Mrs Gabe buys kain koffo; B. Mrs. Gabe buys kain ulos; C. Mrs. Gabe buys kain lantung.*). To create more challenging questions that test multi-step cultural reasoning, our expansion process consists of the following steps: (1) first-hop link type creation; and (2) bilingual multi-hop question generation, where we simultaneously generate culturally authentic questions in both Indonesian and English using an LLM. Figure 2 (bottom left) shows our overall process for expanding single-hop questions to multi-hop.

First-hop question type (clue type). Following Mavi et al. (2024), we design six types of cultural clues: commonsense, comparison, entity, geographical, intersection, and temporal. For each type, we develop specific transformation guidelines and create distinct prompts with tailored instructions and few-shot examples (Appendix A). These cultural clues are automatically generated by prompting LLMs to produce reasoning-based multi-hop questions. To answer the transformed question, models must first determine which province the cultural clues refer to. Table 1 summarises the key principles and provides examples for each question type. For example, the ENTITY clue type uses prompts designed to identify provinces through specific cultural items such as the *Tor-tor dance*, which is a unique traditional dance from *North Sumatra*. *The province serves as a first-hop entity while the original IndoCulture context becomes the second-hop cultural question*.

Bilingual multi-hop question generation. To enable broader accessibility to our data, we perform multi-hop question generation through two sequential sub-processes for each clue type, simultaneously generating culturally authentic questions in both Indonesian and English:

- **Statement to question conversion:** We convert each original context statement into a question while removing direct province mentions (if any). For example, given the example in Figure 2, this is transformed to *What type of traditional cloth does Bu Gabe buy as a gift for her daughter-in-law*.
- **First-hop link type integration:** We add first-hop clues based on the selected clue type that require reasoning to identify the target province. These use only indirect cultural references without mentioning provinces, cities, or regencies directly. The final result combines both steps: *What type of traditional cloth does Bu Gabe buy as a gift for her daughter-in-law in the region where Tor-tor dance is performed during important ceremonies?*

We generate questions in Indonesian and English using *Claude-3.7-Sonnet* (Anthropic, 2025) with temperature $= 1$. We translate the text into the target languages while keeping the culture-specific terms (e.g., *Rumoh Aceh*) unchanged. The

Type	Key Rule / Focus	Example
Entity	Use only exact entity names (people, historical figures, cultural artifacts) uniquely associated with one province. Avoid geographical features and compound entities with "and".	What musical instrument is central to wedding ceremonies in the province where Cut Nyak Dhien led resistance against Dutch colonialism?
Geographical	Use only geographical features located exclusively in target province. Never use cross-boundary features (rivers, mountain ranges, watersheds).	What traditional fabric is produced in the region where the Derawan Islands are located?
Temporal	Use significant events with specific dates/time periods that uniquely identify exactly one province. Verify historical accuracy of all temporal references.	What traditional dance is performed during harvest festivals in the province where the Majapahit Empire had its capital from 1293 to 1527?
Commonsense	Begin with "If..." scenario (max 2 sentences) using descriptive language to uniquely identify province through distinctive cultural attributes without naming popular items.	If A is female and inherits her mother's property in a western Indonesian province with matrilineal traditions based on Adat Perpatih customary law, what traditional dish would she serve to welcome visitors?
Comparison	Use single, precise comparison (rankings, superlatives, relative metrics) with verifiable data that identifies exactly one province. Specify year for population data.	What traditional ceremony is practised in the Indonesian province with the third highest number of UNESCO World Heritage Sites?
Intersection	First statement [S1] identifies multiple provinces (3-5) using general categories. Second statement [S2] narrows to exactly province. Both conditions must be distinct attributes.	What special dish is prepared during Eid celebrations in the Indonesian province with BOTH active volcanoes AND the largest Buddhist temple in the world?

Table 1: Examples of multi-hop prompt types and their key principles. Blue text represents the first-hop clues that suggest the provinces, and the black text represents the original IndoCulture question.

Mark	Count	Percentage %
OK	**1,712**	**57.07%**
Minor	242	8.07%
Moderate	260	8.67%
Significant	786	26.20%
Total	3,000	100.00%

Table 2: Distribution of quality marks from manual verification of 3,000 randomly sampled multi-hop instances.

prompt template is shown in Appendix A.1. We apply this process to 1,847 IndoCulture questions across six clue types in both languages, yielding 22,164 instances.

4 Dataset Validation

4.1 Initial Quality Assessment

As in the first validation stage (Figure 2, Step 3), we first conduct manual verification on 3,000 randomly sampled instances (in both languages) from our dataset. We reviewed each question and classified them into four quality categories. **OK:** Questions have no substantive issues, or only negligible stylistic flaws (e.g., occasional repeated terms, inconsistencies in punctuation, or missing people's names) that do not affect meaning or clarity. **Minor:** Questions contain slightly unnatural yet understandable translations, including suggestions for improvements in format, tone, or clarity. **Moderate:** Questions where cultural clues are ambiguous and could apply to multiple provinces, making the intended answer uncertain. **Significant:** Questions include factually incorrect statements, leak the correct answer, or are incomprehensible.

We found that 57.07% of the sampled questions meet acceptable standards (OK), 26.20% contain significant errors. Analysis by clue type shows that INTERSECTION and COMPARISON questions demonstrate higher rates of "significant" issues (46.8% and 68.0%, respectively), indicating that *Claude-3.7-Sonnet* struggles more with generating these question types. COMPARISON questions show particularly low quality rates, with only 25.4% marked as OK. Common issues include incorrect factual statements in COMPARISON criteria, e.g., a question about *the province with the third largest area of wetland rice cultivation in Kalimantan* incorrectly refers to South Kalimantan (second in agricultural land among Kalimantan provinces according to 2024 data).

4.2 LLM-as-a-Judge

While manual verification provides insights into data quality, evaluating manually all 22,164 instances is not feasible. Hence, we implement an LLM-as-a-judge (Zheng et al., 2023) using three frontier models (Figure 2, step 4): *GPT-4o* (OpenAI, 2024), *Claude-3.7-Sonnet*, and *DeepSeek-V3* (DeepSeek-AI, 2024). The evaluation assesses key aspects of question quality including factual accuracy, structural coherence, and linguistic quality (guidelines in Appendix B).

Multi-Annotator Validation. To assess the reliability of the LLM-as-a-Judge, we conduct a second validation round with two independent annotators (native Indonesian speakers who have lived in multiple Indonesian provinces) on the same subset as in §4.1. The annotators validated the questions using the same scale and guidelines as the LLM-as-a-judge. When they disagree, a third annotator provides the final judgment.

Inter-annotator Agreement. Table 3 presents Cohen's κ for each question type, computed fol-

Question Type	Cohen's κ
Geographical	0.75
Entity	0.68
Temporal	0.55
Commonsense	0.48
Comparison	0.42
Intersection	0.35
Average	**0.54**

Table 3: Inter-annotator agreement (Cohen's κ) across question types on sample questions.

lowing the survey of agreement metrics by Artstein and Poesio (2008). The average κ across question types is 0.54, with individual types ranging from 0.35 to 0.75. According to the standard benchmarks proposed by Landis and Koch (1977), these scores indicate fair to moderate agreement. GEOGRAPHICAL questions have the highest agreement score, while INTERSECTION questions the lowest. This variation indicates that cultural reasoning complexity varies by category, with location-specific and entity-identification tasks proving easier for consistent annotation than INTERSECTION or COMPARISON tasks. We analyse the disagreements between human annotators (see Appendix G.1). We find that the disagreements often reflect difference perspectives in judgment rather than annotation errors. As Fleisig et al. (2023) note, when annotators disagree on subjective judgments, this often reflects genuine differences in interpretation.

Validating LLM-as-a-Judge with human annotations. We convert human and LLM ratings into binary labels: *OK* and *Minor* to *Acceptable*, *Moderate* and *Significant* to *Unacceptable*. First, we evaluate the LLM judge's decisions against the human gold standard on the dual-annotated instances. The automated filtering achieves a precision of 0.78 and a recall of 0.82. This indicates that while the LLM judge identifies most acceptable questions (high recall), approximately 22% of accepted instances may contain false positives and false negatives. We calculate Intraclass Correlation Coefficient (ICC) to quantify the consistency of observations within groups. We use a two-way random effects model for absolute agreement. As shown in Table 4, the LLM judge achieves moderate agreement with human annotators, with higher agreement on acceptable than unacceptable instances. This pattern, combined with the precision and recall scores, shows that the LLM judge effectively identifies high-quality questions but shows more variation when detecting problematic instances.

Quality Class	LLM vs Ann.1	LLM vs Ann.2
Acceptable	0.75	0.72
Unacceptable	0.66	0.63
Overall	0.71	0.68

Table 4: Intraclass Correlation Coefficient (ICC) by quality class.

While automated quality assessment has limitations, the LLM-as-a-Judge provides a practical solution for filtering large-scale generated data when comprehensive manual annotation is infeasible. Based on the validation results, we apply the following filtering rules: instances receiving majority votes of *Acceptable* from at least two of the three LLMs are retained, while any instance marked as *Significant* by any single LLM is automatically rejected. This process filtered the dataset to 12,939 instances in both Indonesian and English. Given the observed false positive rate, we implement additional structure-based verification in the next validation stage to further validate the annotations.

4.3 Question Structure Verification

As part of the final validation stage (Figure 2, Step 5), we implement a two-stage verification process (Appendix C).

Phase 1: Issue Detection. The first phase uses *Claude-3.7-Sonnet* to identify and correct two specific structural issues. We ask *Claude-3.7-Sonnet* to assess whether the multi-hop question contains phrases directly copied from the provided answer options. If it does, it rewrites the copied text. To detect invalid province name (`contains_province`), we ask *Claude-3.7-Sonnet* to examine whether province names appear as geographical location references (e.g., "from Bali"). If it does, we replace them with indirect references while preserving cultural terminology (e.g., "Batik Aceh").

Phase 2: Quality Assessment. For the questions verified by the previous phase, *Claude-3.7-Sonnet* assesses whether they meet multi-hop requirements based on a two-step reasoning structure, sequential logic, and cultural question alignment. *Claude-3.7-Sonnet* also suggests which

type of revision is needed. If a question needs only minor improvements (i.e., grammar, clarity, formatting, capitalization), it will receive automated refinements. If a question needs a fundamental restructuring of the reasoning flow, it will be flagged as "[NEEDS MAJOR REVISION]" and removed from the dataset. Less than 1% of the questions are removed due to their failure to meet multi-hop requirements in the final assessment. Manual verification of the "[NEEDS MAJOR REVISION]" questions confirm that they are not the desired multi-hop questions.

4.4 Question Language Rebalance

Continuing Step 5 of our framework (Figure 2), we identify questions that only exist in one language based on the ID-type pairs. If an ID-type pair appears only in English, we ask *Claude-3.7-Sonnet* to translate the question into Indonesian, and vice versa for questions in Indonesian. We ensure the translation preserves cultural terms, proper nouns, and traditional item names.

4.5 Naturalness and Difficulty Assessment

In the final part of Step 5 (Figure 2), we perform an evaluation of linguistic naturalness and cognitive difficulty across all questions. Three native Indonesian speakers independently assess the questions following specific guidelines (Appendix D).

Naturalness. Annotators rated both Indonesian and English versions on a three-point scale: *Natural*, *Acceptable*, or *Unnatural*. Using majority voting of their ratings, we identify questions that need revision. About 8% of Indonesian questions were rated *Unnatural* due to translation errors or incorrect subjects/names, while 3% were *Acceptable* due to minor grammar or translation issues. For English questions, about 7% were flagged as *Unnatural* and 2% as *Acceptable*. All questions rated as *Unnatural* are manually revised.

Cognitive Difficulty. On average, 44.8% of the questions were rated as *Hard*, 25.9% as *Moderate*, and 29.2% as *Easy*. This highlights the challenging nature of our dataset.

4.6 Final Dataset

Following our verification framework in §4.1 to §4.5, ID-MoCQA contains 15,590 instances evenly distributed across Indonesian and English with 7,795 instances per language. The distribution of question types is uneven due to the difficulties in generating high-quality questions for different categories, as shown in Table 6. COMPARISON questions have the lowest verification success rate, representing the smallest category with only 730 instances per language. These questions require generating superlatives or performing differential analysis between cultural elements, but are frequently marked as "Significant" during LLM-as-a-judge evaluation due to factual inaccuracies. The generated questions often contain incorrect ranking statements or unverifiable comparative assertions about cultural practices across provinces.

Topic	#Q	Province	#Q
Food	1,335	West Sumatra	1,072
Wedding	1,175	Papua	891
Art	1,025	North Sumatra	850
Family relations	695	Aceh	808
Pregnancy & kids	667	South Kalimantan	783
Socio-religious	646	Bali	747
Religious holiday	497	South Sulawesi	714
Death	431	Central Java	653
Daily activities	384	West Java	582
Traditional games	330	East Java	469
Fisheries & trade	311	East Nusa Tenggara	226
Agriculture	299		

Table 5: Questions across topics and provinces.

Each question requires sequential reasoning: first identifying the target province through cultural clues, then answering the province-specific cultural question. Questions span 11 Indonesian provinces across 6 islands, covering 12 cultural topics. Table 5 presents the distribution across topics and provinces.

Semantic and Lexical Analysis. We conducted automated linguistic analysis using GPT-4o-mini to extract part-of-speech tags, named entities, temporal expressions, and Indonesian cultural terms from all 7,795 English questions (Appendix E). Analysis was performed on the English version to ensure consistent part-of-speech extraction, as Indonesian cultural terms are preserved identically across both language versions. Questions in ID-MoCQA average 24.3 words, with 12,381 adjectives (1.59 per question), 54,297 nouns (6.97 per question), and 14,967 verbs (1.92 per question). COMMONSENSE questions average 30.9 words with 2.5 adjectives per question, supporting their conditional scenario structures, while GEOGRAPHICAL questions average 18.9 words with 0.8 adjectives per question, reflecting more direct reference patterns. The data includes 1,274 unique person names and 1,447 unique location

Type	#QA	Avg. Culture	Adj. /Q	N. /Q	V. /Q	Unique Entities Pers.	Loc.	ID
Commonsense	1,424	30.9	2.5	8.7	2.4	391	223	680
Comparison	730	21.5	1.6	5.8	1.7	171	96	231
Entity	1,447	20.6	0.9	5.9	1.6	496	246	440
Geographical	1,508	18.9	0.8	5.4	1.5	327	274	294
Intersection	1,279	27.2	2.2	7.8	2.1	369	234	633
Temporal	1,407	26.1	1.6	7.4	2.0	377	280	419
Overall	7,795	**24.3**	**1.59**	**6.97**	**1.92**	**1,274**	**1,447**	**2,398**

Table 6: Overview of lexical and cultural characteristics across question types. The table shows the number of QA pairs, the average cultural length (Avg. Culture), and the average number of adjectives, nouns, and verbs per question. The final columns list the counts, including persons (Pers.), locations (Loc.), and identification terms (ID) in each question type.

names. ENTITY questions reference 496 unique persons, while TEMPORAL questions cite 280 unique locations. ID-MoCQA contains 2,398 unique Indonesian cultural terms appearing 8,499 times (1.09 per question), preserved in their original language across both versions. COMMONSENSE and INTERSECTION questions contain 680 and 633 unique terms respectively, while GEOGRAPHICAL contains 294. Approximately 38.1% of questions (2,972) include temporal expressions spanning historical periods, contemporary events, and cultural calendars. Table 6 presents complete statistics across question types.

5 Experimental Setup

5.1 Models

Frontier LLMs: *GPT-5* (OpenAI, 2025), *DeepSeek-V3*, and *Claude-3.7-Sonnet*.

Multilingual open models: *Gemma2-27B-Instruct* (Team, 2024a), *Llama3.3-70B-Instruct* (Meta, 2024), *Llama3.1-8B-Instruct* (Grattafiori et al., 2024), *Qwen2.5-72B-Instruct* and *Qwen2.5-7B-Instruct* (Team, 2024b).

Region-specific open models: *Merak-7B* (Ichsan, 2023) and *SeaLLM-7B* (Nguyen et al., 2024) are trained on Indonesian and are the top two performing models on the IndoCulture. This inclusion allows us to assess whether regional specialisation provides advantages for the multi-hop reasoning in ID-MoCQA.

Given a multi-hop cultural question, a model needs to first identify the relevant Indonesian province based on the clues, then answer a province-specific cultural question about that region. Each question provides three options, which are from the original IndoCulture dataset for the final answer and requires open-ended text generation for province identification. All experiments are conducted using prompts designed to obtain both province identification and final answer selection (Appendix F).

5.2 Human Baseline

We recruit three university graduates (different to the original annotators), who are native Indonesian speakers, to answer all 7,795 ID-MoCQA questions. The guidelines are shown in Appendix D. The participants need to identify the target province first, then select one of the three options without access to external tools.

6 Results and Analysis

6.1 Human Performance

Humans achieve 70.0% multi-hop accuracy, with individual performance ranging from 66.6% to 75.3%. First-hop accuracy is 95.1%. The 25.1% gap between first-hop and multi-hop shows that identifying the location is much easier.

Difficulty ratings and performance. The difficulty assessments align with their performance. On average, 44.8% of questions were rated as Hard, corresponding to the 30% failure rate in multi-hop accuracy. Individual difficulty perceptions varied considerably, with ratings ranging from 32.3% to 53.1% for Hard questions. The native speaker who rated the most questions as Hard achieved 68.1% multi-hop accuracy, while the one who rated the fewest as Hard achieved 75.3% multi-hop accuracy.

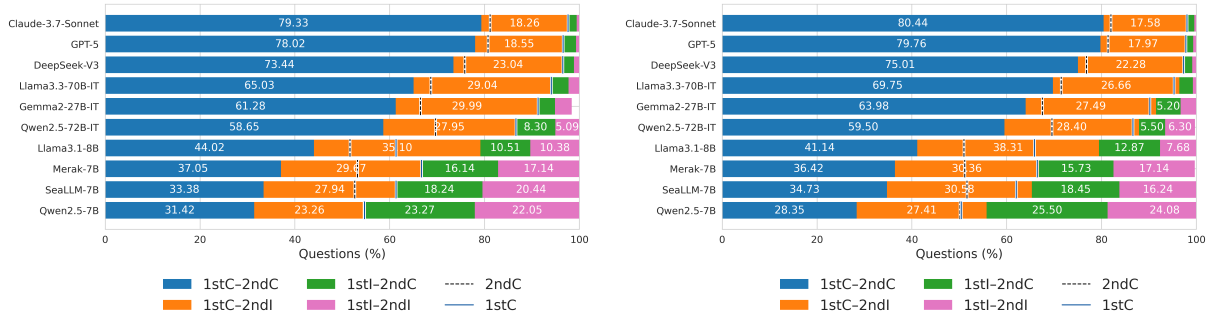
6.2 Zero-shot Results

Frontier LLMs outperform humans. Table 7 shows *GPT-5* and *Claude-3.7-Sonnet* achieve the highest multi-hop accuracy in both languages, surpassing the human baseline by over 10% in Indonesian. *DeepSeek-V3* follows closely, also outperforms humans.

Geographic knowledge differences explain the gap. Frontier LLMs outperform humans across all 11 provinces, but the gap varies depending on the status of those provinces. Bali, with its distinct Hindu culture and prominent role in the tourism industry, along with West Java and Central Java on Java island, which is the most populous island and

	Model	Comm.	Comp.	Entity	Geo.	Inter.	Tempo.	Overall
EN	Claude-3.7-Sonnet	**81.32**	79.86	**80.51**	**81.76**	81.24	**81.59**	**81.15**
	DeepSeek-V3	76.19	73.01	75.33	76.66	77.56	74.84	75.81
	GPT-5	79.85	**82.60**	79.82	80.90	**81.31**	80.95	80.74
	Llama3.3-70B-IT	69.73	66.58	68.00	69.10	67.94	69.44	68.65
	Qwen2.5-72B-IT	67.49	66.44	67.31	66.25	67.08	66.95	66.95
	Gemma2-27B-IT	65.87	65.21	66.76	68.37	67.32	64.61	66.47
	Llama3.1-8B-IT	53.93	54.52	55.15	55.84	54.42	53.16	54.52
	Qwen2.5-7B-IT	54.56	55.07	54.94	57.16	52.46	53.73	54.69
	Merak-7B	52.81	52.60	54.73	54.38	50.82	53.16	53.19
	SeaLLM-7B	51.47	51.23	51.14	52.79	51.21	51.60	51.62
ID	Claude-3.7-Sonnet	**82.23**	80.55	**80.72**	**83.16**	**82.17**	82.30	**81.98**
	DeepSeek-V3	77.46	75.89	75.88	76.66	77.72	77.04	76.83
	GPT-5	81.18	**82.47**	79.89	81.43	81.24	**82.59**	81.37
	Llama3.3-70B-IT	72.96	69.73	70.97	72.75	70.91	70.65	71.49
	Qwen2.5-72B-IT	70.01	68.77	69.59	69.69	69.59	69.23	69.54
	Gemma2-27B-IT	67.49	64.25	68.35	68.63	68.73	66.17	67.53
	Llama3.1-8B-IT	57.58	54.93	57.50	59.48	58.33	56.43	57.60
	Qwen2.5-7B-IT	54.14	53.15	54.39	54.38	51.84	56.43	54.18
	Merak-7B	48.42	51.10	52.66	52.19	50.20	52.24	51.14
	SeaLLM-7B	49.58	51.51	51.07	52.45	50.12	51.17	50.97
	Human (n=3)	69.85±4.6	70.05±3.9	69.89±4.4	70.53±4.2	70.45±4.7	69.23±6.0	69.99±4.6

Table 7: Accuracy (%) across multi-hop clue types and overall in English and Indonesian.



(a) English

(b) Indonesian

Figure 3: Breakdown of model predictions (%) by first-hop (province-level) and second-hop (final answer) correctness for English and Indonesian.

location of the capital, are more familiar to most Indonesians. In these provinces, humans score 84% on average while models score 86%. However, when the questions are about provinces (e.g., Papua and Aceh) that are away from the economic and political centers, human performance drops to 65% while frontier LLMs maintain 77%. This pattern suggests that LLMs' training data provides more balanced coverage of regional cultural knowledge than individuals' lived experience.[2]

Larger models perform better than smaller models in Indonesian. *GPT-5* and *Claude-3.7-Sonnet* achieve the highest performance, and they both perform better in Indonesian than English. *DeepSeek-V3* and other larger models follow the same pattern. However, smaller models show inconsistent language preferences. *Merak-7B* and *SeaLLM-7B* perform worse in Indonesian in most clue types despite being fine-tuned on Indonesian Wikipedia. *Qwen2.5-7B-IT* shows a similar trend with slightly lower performance in Indonesian. In contrast, *Llama3.1-8B-IT* achieves approximately 3% higher accuracy in Indonesian (57.60 vs. 54.52). *Merak-7B* and *SeaLLM-7B* achieve ∼53% accuracy in IndoCulture's single-hop questions, but drop to 51.14% and 50.97% respectively in ID-MoCQA's multi-hop questions. This shows that although specialized models can handle single-hop cultural questions, the additional complexity of multi-hop reasoning poses challenges to smaller models, even in their target languages. The performance gap between larger and smaller models increases for complex reasoning tasks, regardless of language specialization.

[2]To verify this gap reflects genuine knowledge differences rather than dataset biases, we analyze answer position distribution and question length effects. Human errors spread evenly across answer options with no position bias, and question length shows no correlation with the gap.

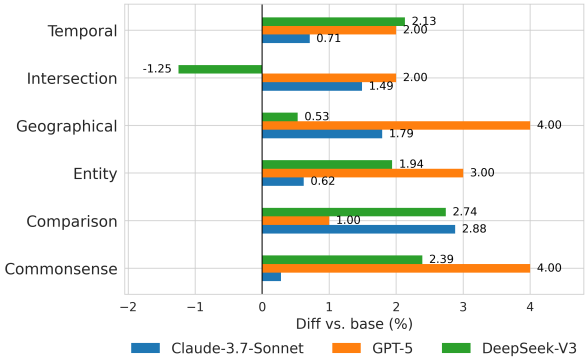
Performance on clue types varies by both model type and language. Table 7 reveals that no single clue type is universally easier or harder for all models. Each model shows a distinct profile of strengths and weaknesses. For instance, while COMPARISON represents the most challenging type for *Claude-3.7-Sonnet*, *DeepSeek-V3*, and *Llama3.3-70B-IT*, *GPT-5* achieves their peak performance on this type in English. Smaller models show more diverse patterns: *Qwen2.5-7B-IT* peaks on GEOGRAPHICAL, while *Merak-7B* peaks on ENTITY. Similarly, *Merak-7B* struggles most on INTERSECTION in English, yet *DeepSeek-V3* shows its highest accuracy on this type. In contrast, *Llama3.1-8B-IT* struggles most with TEMPORAL, while *SeaLLM-7B* peaks on GEOGRAPHICAL. This divergence indicates that different models have developed distinct reasoning capabilities.

The pattern shifts in Indonesian. *Merak-7B*'s lowest-performing type shifts from INTERSECTION in English to COMMONSENSE in Indonesian, dropping below 49% accuracy. *SeaLLM-7B* shifts from ENTITY to COMMONSENSE as its weakest type. *GPT-5* maintains ENTITY as its weakest type in both languages, while *Qwen2.5-72B-IT* shifts from GEOGRAPHICAL in English to TEMPORAL in Indonesian. Their peak performance types also change: *GPT-5* moves from COMPARISON to TEMPORAL as its strongest type, and *Gemma2-27B-IT* shifts from GEOGRAPHICAL to INTERSECTION.

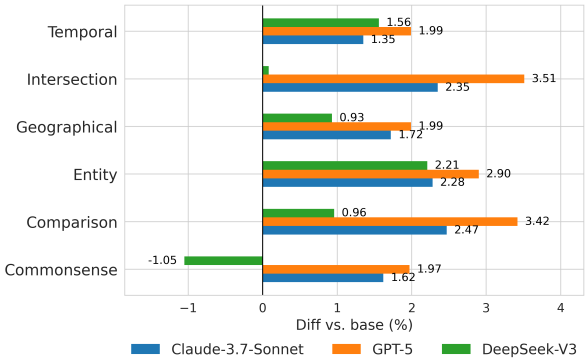
Frontier LLMs excel at province prediction but not so good at selecting final answers. Figure 3 shows frontier models achieve over 96% first-hop accuracy but are 18% to 23% lower when considering both steps. Correct first-hop with incorrect second-hop occurs six to ten times more than the opposite (under 3%), and both remain below 1.2%. This contrast indicates models rarely achieve correct cultural answers without accurate province identification. Smaller models show even larger variation in gaps, and face difficulties in both steps.

6.3 Chain-of-Thought Results

To evaluate how in-context reasoning influence LLMs on the ID-MoCQA questions, we tested the three frontier LLMs using Zero-shot Chain-of-Thought (CoT) prompting (Kojima et al., 2022) by adding ***"Let's think step by step"*** to the inputs. Appendix F shows the full prompt. CoT results



(a) English



(b) Indonesian

Figure 4: Improvement (%) from CoT over Non-CoT prompting across models and question types.

in mixed improvements, with *GPT-5* showing the largest overall gains (averaging 2.67% in English, 2.63% in Indonesian), followed by *Claude-3.7-Sonnet* (1.97% in Indonesian, 1.30% in English) and *DeepSeek-V3* (1.41% in English, 0.78% in Indonesian). These improvements suggest that CoT can aid cultural inference tasks, aligning with Romanou et al. (2025).

Figures 4a and 4b reveal variations in CoT performance across question types, models, and languages. *GPT-5* has the largest and most consistent gains, reaching up to 4.00% on both GEOGRAPHICAL and COMMONSENSE in English, and 3.51% on INTERSECTION in Indonesian. However, the negative improvements indicate that CoT can be counterproductive for certain model-task-language combinations, introducing noise or misaligned reasoning. The variation of effectiveness between languages and models suggests that cultural reasoning structures may be represented differently across languages (Shi et al., 2023). Overall, while CoT prompting provides benefits for some models, the inconsistent gains and notable negative cases indicate that cultural reasoning re-

Cultural Domain	Question Context	LLMs Answers	Correct Answer	Bias Explanation
Food	Aceh casual dining outside home	*Kuah beulangong* (ceremonial curry)	*Sate matang* (grilled meat)	Selected elaborate ceremonial dish over practical everyday food appropriate for casual context
Pregnancy and kids	Aceh *mee boh kayee* ceremony	8th month pregnancy	3rd month pregnancy	Selected 8th month as approximation to common 7th month traditions rather than regional 3rd month practice
Wedding	West Sumatra proposal tradition	Groom's family gives *uang adat* to bride	Bride's family gives *uang adat* to groom	Applied patriarchal payment direction instead of matrilineal practice where bride's family initiates and pays
Fisheries and trade	Livestock distribution	Distribute to relatives and neighbors	Sell at *wosi* market per kg	Applied "traditional equals communal" stereotype, missing local market-based practice
Wedding	West Sumatra Koto Gadang style wedding	*Suntiang* headdress	Cloth head covering	Selected widely documented Minangkabau headdress over regional specific practice of mentioned Koto Gadang traditions
Pregnancy and kids	Bali placenta burial in house yard	Baby recognizes their house	Baby is always protected	Applied literal practical reasoning over spiritual belief
Food	Central Java gudeg Solo taste	Sweet taste	Savory taste	Confused regional variants, selecting Yogyakarta gudeg characteristics over Solo (Central Java) gudeg characteristics
Death	North Sumatra post-burial family ritual	Hold prayers for 7 nights at home	Make pilgrimage to scatter flowers on grave	Applied Islamic mourning tradition (7-night prayers), reflecting Indonesia's Muslim majority, over Batak Christian practice of immediate grave visitation with flowers, showing religious framework override

Table 8: Examples contrasting knowledge prominence versus contextual correctness in model selection across Indonesian cultural domains. All examples show systematic failures where all three models (Claude-3.7-Sonnet, DeepSeek-V3, GPT-5) selected the same incorrect answer.

mains challenging and cannot be uniformly solved with zero-shot CoT.

6.4 Qualitative Analysis

We observe that models favor well-documented practices over situationally appropriate ones (Table 8). When questions explicitly describe casual dining contexts, models select *kuah beulangong* (elaborate ceremonial curry) over *sate matang* (everyday grilled meat). This bias extends to pregnancy ceremonies: models choose 8th-month rituals, likely as approximation to widely documented 7th-month traditions, rather than Aceh's regional 3rd-month *mee boh kayee* ceremony. Papua failures reveal models' *traditional equals communal* stereotypes about indigenous practices. When questions describe pig slaughter in contexts with *bakar batu* stone cooking traditions, models correctly associate *bakar batu* with Papua and successfully identify the province. However, because *bakar batu* is a traditional cultural practice, models then apply *traditional practice equals communal sharing* logic, expecting pigs to be distributed freely to relatives and neighbors. The correct answer is that pigs are sold at *wosi* markets per kilogram, reflecting common practice among locals. Models correctly identify Papua but then misunderstand how cultural practices happen in day-to-day local customs.

7 Conclusion

We proposed a framework for expanding single-hop cultural questions into multi-hop questions targeting Indonesian culture. Our resulting multi-hop QA dataset, **ID-MoCQA**, consists of 15,590 multi-hop questions in Indonesian and English. Our systematic evaluation across ten open-weight and frontier LLMs shows that they struggle with the multi-hop cultural questions. They tend to choose the most well-known cultural information, regardless of whether it is suitable for the specific situation. In the future, we will explore debiasing methods (Ko et al., 2020; Zheng et al., 2024) to mitigate LLMs' preference towards prominent culture. Preference-tuning approaches might also help alleviate LLMs' biases and steer them towards local cultural practices.

Acknowledgments

VA is supported by Indonesia Endowment Fund for Education (LPDP), under the Ministry of Finance, Indonesia. XT and NA are supported by the EPSRC [grant number EP/Y009800/1], through funding from Responsible AI UK (KP0016) as a Keystone project. We also acknowledge IT Services at the University of Sheffield for the provision of services for High Performance Computing.

References

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in

LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.

Anthropic. 2025. Claude 3.7 sonnet and claude code. Large language model.

Þórunn Arnardóttir, Elías Bjartur Einarsson, Garðar Ingvarsson Juto, Þorvaldur Páll Helgason, and Hafsteinn Einarsson. 2025. WikiQA-IS: Assisted benchmark generation and automated evaluation of Icelandic cultural knowledge in LLMs. In *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, pages 64–73, Tallinn, Estonia. University of Tartu Library, Estonia.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Comput. Linguist.*, 34(4):555–596.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Number 4 in Studies in Interactional Sociolinguistics. Cambridge University Press, Cambridge.

Michael Byram. 1997. *Teaching and Assessing Intercultural Communicative Competence*. Multilingual Matters. Multilingual Matters.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 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 (Volume 1: Long Papers)*, pages 25663–25701, Vienna, Austria. Association for Computational Linguistics.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, Singapore. Association for Computational Linguistics.

Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively multi-cultural knowledge acquisition & lm benchmarking. *Preprint*, arXiv:2402.09369.

Erving Goffman. 1967. *Interaction Ritual: Essays on Face-to-Face Behavior*. Anchor Books, Garden City, NY.

Aaron Grattafiori et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Edward T. Hall. 1976. *Beyond Culture*. Anchor Press, Garden City, NY.

Md. Arid Hasan, Maram Hasanain, Fatema Ahmad, Sahinur Rahman Laskar, Sunaya Upadhyay, Vrunda N Sukhadia, Mucahid Kutlu, Shammur Absar Chowdhury, and Firoj Alam. 2025. NativQA: Multilingual culturally-aligned natural query for LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14886–14909, Vienna, Austria. Association for Computational Linguistics.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Geert Hofstede. 2011. Dimensionalizing cultures: The Hofstede model in context. *Online Readings in Psychology and Culture*, 2(1).

D.H. Hymes. 1972. On communicative competence. In J.B. Pride and J. Holmes, editors, *Sociolinguistics: Selected Readings*, pages 269–293. Penguin Books.

Muhammad Ichsan. 2023. Merak-7b: The LLM for Bahasa Indonesia. `https://huggingface.co/Ichsan2895/Merak-7B-v5-PROTOTYPE1`. Hugging Face Repository.

Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024. CLIcK: A benchmark dataset of cultural and

linguistic intelligence in Korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3335–3346, Torino, Italia. ELRA and ICCL.

Miyoung Ko, Jinhyuk Lee, Hyunjae Kim, Gangwoo Kim, and Jaewoo Kang. 2020. Look at the first sentence: Position bias in question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1109–1121, Online. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Fajri Koto, Timothy Baldwin, and Jey Han Lau. 2022. Cloze evaluation for deeper understanding of commonsense stories in Indonesian. In *Proceedings of the First Workshop on Commonsense Representation and Reasoning (CSRR 2022)*, pages 8–16, Dublin, Ireland. Association for Computational Linguistics.

Fajri Koto, Rahmad Mahendra, Nurul Aisyah, and Timothy Baldwin. 2024. IndoCulture: Exploring geographically influenced cultural commonsense reasoning across eleven Indonesian provinces. *Transactions of the Association for Computational Linguistics*, 12:1703–1719.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024. Culturepark: Boosting cross-cultural understanding in large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 65183–65216. Curran Associates, Inc.

Vaibhav Mavi, Anubhav Jangra, and Adam Jatowt. 2024. Multi-hop question answering. *Found. Trends Inf. Retr.*, 17(5):457–586.

Meta. 2024. Model cards & prompt formats llama 3.3.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, Víctor Gutiérrez-Basulto, Yazmín Ibáñez García, Hwaran Lee, Shamsuddeen Hassan Muhammad, Kiwoong Park, Anar Sabuhi Rzayev, Nina White, Seid Muhie Yimam, Mohammad Taher Pilehvar, Nedjma Ousidhoum, Jose Camacho-Collados, and Alice Oh. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. In *Advances in Neural Information Processing Systems*, volume 37, pages 78104–78146. Curran Associates, Inc.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2024. SeaLLMs - large language models for Southeast Asia. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 294–304, Bangkok, Thailand. Association for Computational Linguistics.

OpenAI. 2024. GPT-4o mini: Advancing cost-efficient intelligence. OpenAI Index.

OpenAI. 2024. Gpt4o.

OpenAI. 2025. Gpt5.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2025. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics*, pages 1–96.

Rifki Afina Putri, Faiz Ghifari Haznitrama, Dea Adhista, and Alice Oh. 2024. Can LLM generate culturally relevant commonsense QA data?

case study in Indonesian and Sundanese. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20571–20590, Miami, Florida, USA. 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, Daniil Dzenhaliou, Daniel Fernando Erazo Florez, Fabian Farestam, Joseph Marvin Imperial, Shayekh Bin Islam, Perttu Isotalo, Maral Jabbarishiviari, Börje F. Karlsson, Eldar Khalilov, Christopher Klamm, Fajri Koto, Dominik Krzemiński, Gabriel Adriano de Melo, Syrielle Montariol, Yiyang Nan, Joel Niklaus, Jekaterina Novikova, Johan Samir Obando Ceron, Debjit Paul, Esther Ploeger, Jebish Purbey, Swati Rajwal, Selvan Sunitha Ravi, Sara Rydell, Roshan Santhosh, Drishti Sharma, Marjana Prifti Skenduli, Arshia Soltani Moakhar, Bardia soltani moakhar, Ayush Kumar Tarun, Azmine Toushik Wasi, Thenuka Ovin Weerasinghe, Serhan Yilmaz, Mike Zhang, Imanol Schlag, Marzieh Fadaee, Sara Hooker, and Antoine Bosselut. 2025. INCLUDE: Evaluating multilingual language understanding with regional knowledge. In *The Thirteenth International Conference on Learning Representations*.

Mili Shah, Joyce Cahoon, Mirco Milletari, Jing Tian, Fotis Psallidas, Andreas Mueller, and Nick Litombe. 2024. Improving LLM-based KGQA for multi-hop question answering with implicit reasoning in few-shot examples. In *Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024)*, pages 125–135, Bangkok, Thailand. Association for Computational Linguistics.

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

Gemma Team. 2024a. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Qwen Team. 2024b. Qwen2.5: A party of foundation models.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Xiaonan Wang, Jinyoung Yeo, Joon-Ho Lim, and Hansaem Kim. 2024. KULTURE bench: A benchmark for assessing language model in Korean cultural context. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 914–927, Tokyo, Japan. Tokyo University of Foreign Studies.

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.

Anna Wierzbicka. 1994. Cultural scripts: A semantic approach to cultural analysis and cross-cultural communication. *Pragmatics & Cognition*, 2(2):153–183.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. 2024. WorldValuesBench: A large-scale benchmark dataset for multi-cultural value awareness of language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language*

Resources and Evaluation (LREC-COLING 2024), pages 17696–17706, Torino, Italia. ELRA and ICCL.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Naitian Zhou, David Bamman, and Isaac L. Bleaman. 2025. Culture is not trivia: Sociocultural theory for cultural NLP. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25869–25886, Vienna, Austria. Association for Computational Linguistics.

A Multi-Hop Question Prompt Guidelines

A.1 Sample Full Prompt

Sample Full Prompt: Entity

INSTRUCTIONS *Context Conversion:*

- Convert each "context" statement into a culture-related, province-specific question answerable with the provided "options".
- **DO NOT** mention the province name in the question.
- Keep the original topic intact.

Multi-Hop Question Generation:

- Generate multi-hop questions requiring reasoning to identify the province:
  - **DO NOT** mention the province, cities, or regencies directly or indirectly — use only indirect **CULTURAL** clues.
  - Ensure the correct answer remains consistent with the original options.
  - Avoid repetitive use of similar entities across different questions.

Clue Type: Entity

- Connect the province through a person/thing.
- Use **ONLY exact entity names** (people, historical figures, cultural artifacts).
- Avoid geographical features like lakes, rivers, mountains.
- **DO NOT** use compound entities connected with "and" — focus on one clear, specific entity.
- Ensure the entity is uniquely associated with only this province.
- **NEVER** translate, replace, or use synonyms for proper names, historical events, cultural artifacts, or other essential cultural elements.

Examples:

- "What musical instrument is commonly featured to wedding ceremonies in the province where Cut Nyak Dhien led resistance against Dutch colonialism?"
- "What traditional dance is performed in the province where former President Susilo Bambang Yudhoyono spent his childhood?"
- "What ancient temple complex can be found in the province where coffee variety 'Toraja' originates?"
- "What ancient fortress can be visited in the province where *pala* was once worth more than gold to European traders?"
- "What traditional harvest celebration performed to express gratitude for blessings and abundance takes place in the province where *rendang* was named the world's most delicious food by CNN?"

Language Output:

- Provide both English and Indonesian versions of the question.

A.2 Clue Types and Structural Templates

Entity, Geographical, Temporal, and **Commonsense** clue type use the templates in Appendix A.1, varying only in reasoning specifications and few-shot examples. (**COMPARISON, INTERSECTION**) require enhanced prompts with verification procedures due to their factual complexity. Table 9 shows the details.

Comparison Embedded verification requires confirming comparative claims against empirical data before question finalization:

```
VERIFY:
- Claim: [exact comparative metric used]
- Data: [provinces with values, showing why the claim identifies exactly one province]
- Data source/year: [specify year for population data or source for other metrics]
- Unique? [YES/NO]
```

Failed verification triggers iterative claim revision and re-verification until unique identification achieved.

Intersection Structured step-by-step verification protocol:

```
S1: [Brief condition description] → [List AT LEAST 3 provinces with minimal proof]
S2: [Brief condition description] → [List provinces with minimal proof]
Intersection: [Expected single province]
```

Ensures S1 identifies multiple provinces, S2 narrows to exactly one target province, and intersection produces intended result.

Entity	**Geographical**
Connect the province through a person/thing **IMPORTANT requirements:** • Use ONLY exact entity names (people, historical figures, cultural artifacts) • Avoid geographical features like lakes, rivers, mountains • Do NOT use compound entities connected with "and" – focus on one clear, specific entity • Ensure the entity is uniquely associated with only this province • NEVER translate, replace, or use synonyms for proper names, historical events, cultural artifacts, or other essential cultural elements	Connect the province through the location entity **IMPORTANT requirements:** • Use ONLY geographical features that are located EXCLUSIVELY in the target province • NEVER use geographical features that cross provincial boundaries (rivers, mountain ranges, watersheds, etc.) • NEVER translate, replace, or use synonyms for proper names, historical events, cultural artifacts, or other essential cultural elements
Temporal	**Commonsense Reasoning**
Connect the province through a temporal event (historical or contemporary) **IMPORTANT requirements:** • Use significant events with SPECIFIC dates or time periods (historical OR recent) • Temporal references can include, but are not limited to: Ancient & Pre-Colonial Events, Colonial Period, Independence Era, Modern Historical Events, Natural Disasters, Contemporary Developments, Cultural Milestones, Political Changes, Economic Transformations • Ensure the temporal event is uniquely associated with only this province • Include clear dates or time periods (years, centuries, decades) • NEVER translate, replace, or use synonyms for proper names, historical events, cultural artifacts, or other essential cultural elements	Formulate a single integrated question that: 1. Begins with "If" followed by a concise scenario related to the original context 2. The scenario should uniquely identify the province through distinctive cultural attributes WITHOUT naming it 3. Then directly asks about the cultural element from the original context/options 4. The question should require connecting the scenario (first hop) to the cultural element (second hop) **IMPORTANT:** • Keep the scenario part of the question BRIEF and CONCISE (no more than 2 sentences) • The scenario must provide enough cultural context to uniquely identify the province • Use descriptive language instead of naming popular/familiar items of the province
Comparison	**Intersection**
Refer the province through comparison **IMPORTANT requirements:** • Each comparison MUST identify EXACTLY ONE province • Use diverse formats: Numeric rankings, Range comparisons, Superlatives, Relative metrics • Avoid compound comparisons with "and" – use single, precise comparisons • Use concrete, verifiable metrics (area, number of temples, cultural sites, etc.) • When using population data, ALWAYS specify the year • AVOID using temperature or climate metrics as these can fluctuate seasonally • NEVER translate, replace, or use synonyms for proper names, historical events, cultural artifacts, or other essential cultural elements	Find the province meeting multiple conditions **IMPORTANT requirements:** • The first statement (S1) MUST identify MULTIPLE provinces (at least 3, MAXIMUM 5) • The second statement (S2) must narrow down to EXACTLY ONE province (the target province) • Both conditions must be DISTINCT cultural, geographical, or historical features • Check intersection: Does exactly ONE province match both C1 and C2? • ONLY after verification passes, formulate combined question. • NEVER translate, replace, or use synonyms for proper names, historical events, cultural artifacts, or other essential cultural elements

Table 9: Prompts of all the clue types.

B LLM-as-a-Judge Evaluation Criteria

Our LLM-as-a-judge framework employs eight criteria, each scored 0-2 (16-point maximum), where 2 indicates the highest quality and 0 the lowest.

B.1 Province Identification and Structural Quality Criteria

The first four evaluate provincial clue accuracy, conciseness, cultural alignment, and reasoning structure.

Criterion	Score	Description
Provincial Specificity & Factual Accuracy	2	Clues uniquely identify one province with accurate information
	1	Clues apply to 2-3 provinces with minor ambiguity, no factual errors
	0	Clues contain factual errors or hallucinations
Redundancy of Province Clues	2	Concise, non-repetitive cultural or geographic clues
	1	Minor redundancy present, core clue remains meaningful
	0	Excessive repetition reduces reasoning complexity
Reference Alignment & Rephrasing Quality	2	Accurately reflects original context with different wording
	1	Minor deviations but maintains essential cultural elements
	0	Significantly alters or misrepresents cultural information
Multi-hop Structure	2	Clear two-step reasoning: province identification, then cultural question
	1	Attempts two-step process but partially combines steps
	0	Fails to create proper two-step structure

Table 10: Province identification and structural quality criteria.

B.2 Answer Quality and Presentation Criteria

Four criteria assess reasoning necessity, answer alignment, clarity, and language quality.

Criterion	Score	Description
Answer Discrimination	2	Requires both reasoning steps for correct answer
	1	Can narrow to 2 options after one step
	0	Correct answer identifiable without province identification
Answer Quality	2	Perfect alignment with original options and reasoning process
	1	Minor inconsistencies but remains functionally appropriate
	0	Significantly alters context, making options inappropriate
Question Clarity	2	Clearly articulated and easily understandable
	1	Minor clarity issues but remains interpretable
	0	Confusing phrasing or disorganised structure
Language Quality	2	Grammatically correct with natural phrasing
	1	Minor grammar issues but understandable
	0	Significant grammar problems hindering understanding

Table 11: Answer quality and presentation criteria. **Answer Discrimination** verifies both reasoning steps required (prevents shortcuts). **Answer Quality** ensures multi-hop-option compatibility and reasoning structure alignment. **Question Clarity** assesses comprehension, coherence, and organization. **Language Quality** evaluates grammar and naturalness (both languages) while preserving Indonesian cultural terms.

C Dataset Verification Pipeline: Prompt Engineering

C.1 Phase 1: Issue Detection

Identifies and corrects option text copying and incorrect province name usage as location references.

```
1. OPTION TEXT COPYING:
   - Does the MHQA contain phrases copied from the options? Mark as true/false.
2. PROVINCE NAME USAGE:
   - Is the province name used specifically as a location in the MHQA?
   - Cultural terms (e.g., 'Rumoh Aceh') must NOT be marked.
   - Only flag location usage (e.g., 'in Aceh province', 'from Bali').
REVISION GUIDELINES:
```

```
- If copying detected: Reword to avoid option text
- If province location detected: Use indirect references
```

Cultural terms (e.g., 'Rumoh Aceh') versus location references (e.g., 'in Aceh province') distinction maintains authenticity while eliminating geographic shortcuts.

C.2 Phase 2: Quality Assessment

Evaluates multi-hop structure integrity. CRITICAL: Prevent reintroduction of location-based province references.

```
CRITICAL RULE: DO NOT use province names as locations in revisions
EVALUATE:
- Proper multi-hop question? (identify province → cultural question)
- Clear sequential reasoning?
- Correct grammar for the language used?
IMPROVE:
1. Two-step structure: province identification → cultural question
2. Use indirect province references only
3. Preserve cultural terms exactly ('Rumoh Aceh', 'Soto Aceh')
4. Fix grammar and maintain natural flow
DECISION:
- Minor fixes needed: REVISE
- Major restructuring needed: mark "[NEEDS MAJOR REVISION]"
```

D Manual Evaluation Guideline

Three native Indonesian graduate students evaluate ID-MoCQA question quality and difficulty through (1) linguistic naturalness assessment, (2) multi-hop question answering, and (3) cognitive difficulty rating. External sources and AI assistance are prohibited. Annotators receive: **Context** (IndoCulture premise), **ID-MoCQA** bilingual questions, and **Options** (three choices A/B/C in both languages).

D.1 Task 1: Naturalness

Rate linguistic and cultural naturalness on a 3-point scale:

- **Natural**: Fluent, grammatically correct, culturally accurate, and sounds authentic.
- **Acceptable**: Understandable with minor issues (slight awkwardness, minor grammar errors, somewhat unnatural phrasing).
- **Unnatural**: Major grammatical errors, very awkward phrasing, culturally incorrect references, or incomprehensible.

D.2 Task 2: Multi-hop Question Answering

Objective: Answer each question through a two-step reasoning process that connects cultural context to the correct answer.

Annotators perform the following steps:

1. **Identify the target province:** Use the provided cultural clues to determine which Indonesian province is being referenced.
2. **Select the correct answer:** Choose one option (A, B, or C) that correctly answers the question based on the identified province.

D.3 Task 3: Difficulty Assessment

Objective: Assess the cognitive difficulty required to answer each question.

Annotators assign one difficulty label to each question based on the reasoning complexity and the rarity of cultural knowledge required:

- **Easy:** The question involves widely known cultural facts and can be answered with minimal reasoning or common knowledge.
- **Moderate:** The question requires moderate reasoning or specific cultural knowledge that may not be universally familiar.
- **Hard:** The question requires specialized regional cultural knowledge, uncommon facts, or complex multi-hop reasoning to derive the correct answer.

E Semantic Analysis

We extracted lexical and semantic features from all English questions using *GPT-4o-mini* (OpenAI, 2024) (temperature=0). English questions were analyzed for two reasons: (1) current LLMs provide more reliable part-of-speech tagging and named entity recognition for English; (2) Indonesian cultural terms (traditional items, ceremonies, place names) remain identical across both language versions, ensuring cultural authenticity is captured regardless of analysis language. The bilingual dataset's parallel structure ensures lexical patterns in English reflect the same cultural content as Indonesian versions.

- **Lexical features**: Adjectives, nouns, and verbs.
- **Named entities**: Person names and location names.
- **Temporal expressions**: Time references (historical periods, dates, contemporary events).
- **Indonesian cultural terms**: Culture-specific words preserved in original language.
- **Word count**: Total words per question.

Prompt Format: Semantic Analysis

```
Analyze this Indonesian culture question.
Question: "{question}"
Extract (return ONLY valid JSON)
```

F Model Evaluation Prompts

F.1 Evaluation Configuration

Models are evaluated with temperature=0 (except *GPT-5*: default=1.0, no temperature support) via APIs in zero-shot settings. The Zero-shot prompt asks for structured output (province + answer) without explanations. The CoT prompt requests step-by-step reasoning before structured output.

F.2 Zero-shot Evaluation Prompt Structure

Prompt Format: Indonesian

```
Jawab HANYA dengan format: PROVINSI: [nama provinsi] JAWABAN: [huruf]. Tanpa penjelasan.
Pertanyaan: {pertanyaan}
Pilihan: A. {opsi_a} B. {opsi_b} C. {opsi_c}
```

Prompt Format: English

```
ONLY respond with format: PROVINCE: [province name] ANSWER: [letter]. No explanations.
Question: {question}
Options: A. {option_a} B. {option_b} C. {option_c}
```

F.3 Zero-shot Chain of Thought (CoT) Evaluation Prompt Structure

CoT Prompt Format: Indonesian

```
Pertanyaan: {pertanyaan}
Pilihan: A. {opsi_a} B. {opsi_b} C. {opsi_c}

Mari berpikir langkah demi langkah untuk menjawab pertanyaan ini.
Pertama, prediksi provinsi Indonesia yang paling terkait dengan pertanyaan ini.
Kemudian, analisis pertanyaan dan pilihan untuk menentukan jawaban yang benar.
Analisis: [Jelaskan proses pemikiran langkah demi langkah]
Setelah memberikan penjelasan, akhiri jawaban dengan format:
Provinsi: [nama provinsi]
Jawaban: [A/B/C]
```

CoT Prompt Format: English

```
Question: {question}
Options: A. {option_a} B. {option_b} C. {option_c}

Let's think step by step to answer this question.
First, predict which Indonesian province this question is most associated with.
Then, analyse the question and options to determine the correct answer.
Analysis: [Explain your step-by-step thinking process]
After providing your explanation, end your answer with this format:
Province: [province name]
Answer: [A/B/C]
```

G Additional Analysis

G.1 Human-Human Disagreement Examples

During the multi-annotator validation process §4.2, we observed several patterns in cases where human annotators disagreed on question quality. Table 12 presents two examples illustrating the types of ambiguities that led to disagreement and how they were resolved through majority vote. These cases demonstrate that disagreements often stem from legitimate differences in judgment criteria rather than annotation errors, particularly regarding: (1) the level of temporal and statistical specificity required for comparative claims, and (2) whether certain question structures might reveal answers.

G.2 Distribution of Culture-Specific Terms Across Provinces

As discussed in §4.6, provincial distribution shows variation in cultural-linguistic specificity. East Nusa Tenggara (226 questions) has the highest density of culture-specific terms preserved in local language (0.92 terms per question), suggesting its cultural practices rely heavily on local terminology. In contrast, West Sumatra, despite having the most questions (1,072), shows lower cultural term density (0.65 terms per question), indicating its cultural practices may be more well known or describable with general Indonesian vocabulary. This pattern appears across other provinces: East Java (469 questions, 0.81 terms per question), Central Java (653 questions, 0.81 terms per question), and Aceh (808 questions, 0.76 terms per question) maintain higher cultural-linguistic specificity than the three most-represented provinces (West Sumatra, Papua, and North Sumatra), which average 0.68 terms per question.

G.3 Error patterns of human performance

Analysis of the human baseline reported in §6.1 reveals systematic patterns across topics and provinces. Food-related questions accounted for 15-16% of all errors, followed by Wedding (15-16%) and Art-related questions (12-16%). Province-level error rates varied substantially: West Sumatra showed error rates of 32-42% across participants, followed by South Sulawesi (31-45%) and Papua (29-41%). In contrast, West Java showed the lowest error rates at 9-15%, followed by Bali (13-17%) and Central Java (16-27%). Notably, provinces with the lowest error rates are among Indonesia's most well-known regions both domestically and internationally.

G.4 Performance Consistency Across Clue Types

Beyond the overall accuracy patterns shown in Table 7, performance balance across clue types varies more by individual model characteristics than by scale alone. *GPT-5* shows the narrowest performance range at approximately 2.8 percentage points in English and 2.7 points in Indonesian between its best and worst types, maintaining consistency across all six reasoning categories. *Claude-3.7-Sonnet* demonstrates similarly tight ranges of 1.9 points in English and 2.6 points in Indonesian, indicating highly balanced capabilities. *Llama3.3-70B-IT* shows moderate ranges around 3.2 points in both languages. However, some smaller models display comparable stability: *SeaLLM-7B* shows ranges of approximately 1.7 points in English and 2.3 points in Indonesian, *Qwen2.5-7B-IT* shows approximately 4.7 points in English and 3.3 points in Indonesian, and *Merak-7B* demonstrates ranges around 3.9 points in English and 4.2 points in Indonesian. These patterns indicate that frontier models develop more balanced reasoning capabilities, while some smaller models show greater variability.

Component	Details
Example 1: Temporal Specificity in Comparative Claims	
Question	After grandfather Ali was buried, what activity did Ali perform for 7 or 10 days in the region with the highest percentage of Muslim population in Indonesia?
Options	A. Ali holds *tahlilan* at home B. Ali goes to work immediately C. Ali recites *yasin* prayers at the gravesite for 7 or 10 nights
Correct Answer	C
Annotator 1	*Minor*
Annotator 2	*Moderate*
Annotator 3	*Moderate* – Without explicit official statistic data (and specificity mentioned year), comparative claim lacks verifiability. Though Aceh's Sharia law provides cultural signal, statistical claim needs citation
Final Decision	**Moderate** (majority vote: 2/3)
Rationale	Comparative statistical claims should reference specific, verifiable data to avoid ambiguity
Example 2: Answer Discrimination and Structural Issues	
Question	If Budi lives in a region that formally implements Islamic Sharia law with qanun (special regional regulations) governing public behavior and has Wilayatul Hisbah as Sharia police, what traditional beverage would he buy as a souvenir for his father?
Options	A. Budi buys *Emping Melinjo* B. Budi buys *kopi Gayo* C. Budi buys *Kue Bhoi*
Correct Answer	B
Annotator 1	*Minor*
Annotator 2	*Moderate*
Annotator 3	*Minor* – Though only one option is a beverage (option B), question retains complexity by testing both province identification (Sharia law) and cultural knowledge (traditional beverage)
Final Decision	**Minor** (majority vote: 2/3)
Rationale	Question retains sufficient cultural reasoning complexity despite having only one beverage option, as it requires identifying the region through legal/cultural markers before selecting the appropriate souvenir

Table 12: Examples of human annotator disagreements and the resolution through majority vote.

G.5 Multi-Hop Reasoning Error Patterns

Figure 3 reveals that performance gaps widen across model scales. *Llama3.3-70B-IT*, *Qwen2.5-72B-IT*, and *Gemma2-27B-IT* show 28-30 point gaps with incorrect first-hop but correct second-hop reaching up to 8.3%. Smaller models show even larger variation in gaps (16-35 points across both languages), with *Llama3.1-8B-IT* at 35 points in English and *Qwen2.5-7B-IT* showing incorrect first-hop but correct second-hop at 23.3% and both incorrect at 22.1%. *Merak-7B* shows approximately 30-point gaps with both incorrect reaching 17-18% and first-hop accuracy around 66-67% despite Indonesian training. *SeaLLM-7B* demonstrates smaller gaps (16-20 points) but lower overall first-hop accuracy (51-53%) and higher both-incorrect rates (18-20%). These patterns indicate smaller models face difficulties at both reasoning steps, with elevated reverse error rates suggesting occasional reliance on alternative reasoning pathways that do not depend on accurate province identification.

Cross-linguistic comparison reveals that language effects vary by model category. Frontier models show minimal changes (0.6-0.8 point decreases in first-hop correct but second-hop incorrect), while *Llama3.3-70B-IT* increases both-correct by 3.6 points in Indonesian, demonstrating target-language presentation specifically reduces cultural reasoning errors. In contrast, *Merak-7B*'s both-correct declines 1.6 points despite language-specific training. Overall, first-hop to both-correct gaps widen as model performance decreases (frontier: 18-23 points; 70B models: 28-30 points; smaller models: 16-35 points),

suggesting that weaker models accumulate errors across the reasoning chain rather than failing at specific steps.

G.6 Qualitative Error Analysis

We examined the failure cases of *GPT-5*, *Claude-3.7-Sonnet*, and *DeepSeek-V3* models reported in §6.4. In most of the cases, all three models choose the same incorrect answer, indicating shared systematic biases. Topics like death ceremonies, traditional games, and art forms exhibit substantially higher same-wrong-answer rates than daily activities.

Models consistently demonstrate strong province identification (averaging 96.5%) but struggle with cultural reasoning within correctly identified contexts. West Sumatra exemplifies this pattern: models recognize matrilineal cultural markers yet systematically apply patriarchal logic. In the *bajapuik* wedding tradition, the bride's family pays *uang japuik* to the groom's family, reflecting matrilineal practice. However, models incorrectly expect the groom's family to pay, following widespread patriarchal dowry patterns. Central Java demonstrates unique identification challenges (89% accuracy), driven by cultural similarity rather than geographic proximity. Models frequently confuse Central Java with other Javanese provinces (West Java, East Java, Yogyakarta) that share gamelan music, batik arts, and court traditions. In contrast, Papua, though geographically isolated, demonstrates strong identification. Models struggle to distinguish provinces sharing similar cultural features. Javanese regions exemplify this: despite individual prominence, models confuse those sharing gamelan music, batik arts, and court traditions.