LLMs as Cultural Archives:
Cultural Commonsense Knowledge Graph Extraction

Junior Cedric Tonga[1] **Chen Cecilia Liu**[2] **Iryna Gurevych**[1,2] **Fajri Koto**[1]
[1]Mohamed bin Zayed University of Artificial Intelligence
[2]Ubiquitous Knowledge Processing Lab (UKP Lab)
Department of Computer Science and Hessian Center for AI (hessian.AI)
Technische Universität Darmstadt
{junior.tonga, fajri.koto}@mbzuai.ac.ae

Abstract

Large language models (LLMs) encode rich cultural knowledge learned from diverse web-scale data, offering an unprecedented opportunity to model cultural commonsense at scale. Yet this knowledge remains mostly implicit and unstructured, limiting its interpretability and use. We present an iterative, prompt-based framework for constructing a Cultural Commonsense Knowledge Graph (CCKG) that treats LLMs as cultural archives, systematically eliciting culture-specific entities, relations, and practices and composing them into multi-step inferential chains across languages. We evaluate CCKG on five countries with human judgments of cultural relevance, correctness, and path coherence. We find that the cultural knowledge graphs are better realized in English, even when the target culture is non-English (e.g., Chinese, Indonesian, Arabic), indicating uneven cultural encoding in current LLMs. Augmenting smaller LLMs with CCKG improves performance on cultural reasoning and story generation, with the largest gains from English chains. Our results show both the promise and limits of LLMs as cultural technologies and that chain-structured cultural knowledge is a practical substrate for culturally grounded NLP.[1]

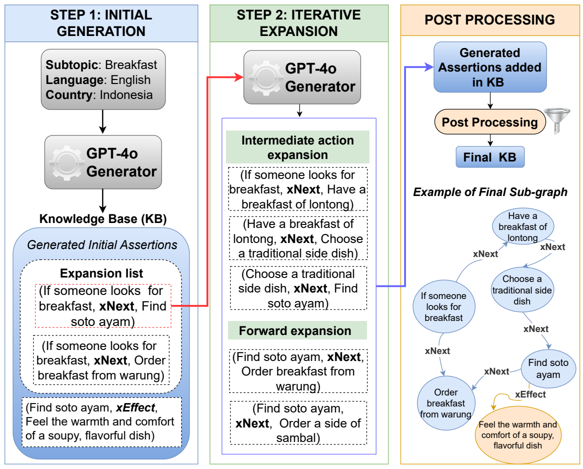

Figure 1: Application of our framework for constructing a partial Cultural Commonsense Knowledge Graph (CCKG) capturing culturally grounded reasoning about breakfast in Indonesia. Given an input prompt specifying the subtopic, language, country, and task-specific constraints, GPT-4o generates English *if–then* commonsense assertions $(action_i, relation, action_j)$ to form an initial knowledge base (KB). Assertions with relations (`xNext`, `oNext`) are iteratively expanded by re-prompting GPT-4o to generate **intermediate action expansions** that decompose $action_i$ into finer-grained steps leading to $action_j$ and **forward actions** occurring after $action_j$. In this example, only the first assertion in the expansion list is expanded for a single iteration. The resulting assertions are added to the KB, post-processed and composed into the final CCKG subgraph.

1 Introduction

Culture and commonsense reasoning are deeply intertwined, as culture shapes how people interpret everyday situations, social conventions, and causal regularities (Koto et al., 2024; Sadallah et al., 2025; Sap et al., 2020). Culture encompasses shared and learned values, norms, and practices that guide interpretation and action within a community (Hershcovich et al., 2022; Adilazuarda et al., 2024; Liu et al., 2025). Because commonsense is grounded in cultural experience rather than universal logic, what seems self-evident in one community can be unfamiliar, or even misleading, in another (Naous et al., 2024; Almheiri et al., 2025). While early computational approaches largely treated commonsense as culture-neutral (Bisk et al., 2020; Sap et al., 2019b), recent works have begun to recognize and model its cultural dimensions, highlighting the linguistic and cognitive variation that arises across communities (Koto et al., 2024; Sadallah et al., 2025).

We argue that language models can reason more

[1]Code available at `https://github.com/JuniorTonga/Cultural_Commonsense_Knowledge_Graph`

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics Volume 1: Long Papers, pages 6289–6309

appropriately across contexts when equipped with structured representations of culture, especially given its complexity and uneven representation in online data. Early efforts such as ATOMIC (Sap et al., 2019a) demonstrated the value of modeling inferential knowledge between everyday events (Hwang et al., 2021; Bosselut et al., 2019), yet these resources remain largely Western-centric and lack cross-cultural generalizability. Extending this idea to cultural contexts requires uncovering and organizing the implicit cultural knowledge within LLMs into interpretable, multi-step structures that reflect local norms and practices.

In this work, we study how far large language models serve as cultural archives—examining the extent and accuracy of the cultural information they encode. Since large language models are pretrained on culturally diverse corpora (Jiang et al., 2021; Sun et al., 2024; Brown et al., 2020), much of this knowledge already exists implicitly and can be systematically extracted in structured form. Specifically, we address three research questions: (1) *To what extent do LLMs encode cultural knowledge that aligns with real-world cultural relevance?* (2) *Which language best represents cultural knowledge when extracted from LLMs?* and (3) *Can the extracted cultural knowledge be used to enhance the cultural reasoning ability of smaller or weaker models?* While our initial hypothesis assumes that the language spoken by the culture provides the most authentic representation, our findings reveal an interesting contrast: English often serves as a more coherent medium for representing cultural knowledge graphs.

Prior work has constructed cultural knowledge bases from sources such as Wikipedia, CommonCrawl, and social media (Nguyen et al., 2024; Fung et al., 2024; Shi et al., 2024). While valuable, these approaches face two key limitations. First, they represent culture as isolated, static facts (Yin et al., 2022; Nguyen et al., 2024; Fung et al., 2024), overlooking the procedural and contextual nature of cultural practices. Many traditions consist of ordered sequences of actions—such as the proposal-to-marriage process in Indonesian weddings—and reducing them to atomic statements omits crucial context, limiting applications like reasoning and story generation. Second, most cultural knowledge bases are built in English, despite the fact that many cultural nuances are best expressed in native languages.

Our contributions can be summarized as follows:

- We propose an iterative, prompt-based framework for constructing **Cultural Commonsense Knowledge Graphs (CCKG)**, which extract multilingual, culture-specific *if–then* inferential knowledge chains from large language models. Following the ATOMIC-style formulation (Sap et al., 2019a), we compose these relations into multi-step cultural knowledge chains (Figure 1).
- Through extensive human evaluations of cultural relevance, correctness, and path coherence, we find that while native languages capture richer cultural detail, English extractions are more coherent and consistently preferred.
- We show that augmenting smaller or weaker LLMs with CCKG improves their cultural reasoning and story generation performance, highlighting the value of inferential cultural knowledge for developing culturally grounded NLP systems.

2 Related work

Knowledge Bases for Commonsense Reasoning. Early commonsense knowledge bases (KBs) such as WebChild (Tandon et al., 2014) linked nouns with descriptive adjectives to encode physical and perceptual properties, providing one of the first large-scale automatically constructed commonsense resources. ConceptNet (Speer et al., 2017) aggregated human-authored assertions but primarily captured lexical relations between words rather than inferential or situational knowledge. Quasimodo (Romero et al., 2019) further expanded coverage by mining commonsense assertions from QA forums and web text through large-scale automated extraction. ATOMIC (Sap et al., 2019a) marked a shift toward event-level reasoning, introducing large-scale *if–then* relations that capture causes, effects, and social intentions.

While these knowledge bases advanced commonsense modeling, they remain largely English-centric and culturally neutral. Recent work has begun integrating culture into knowledge representation: CANDLE (Nguyen et al., 2023) performs large-scale web extraction to probe factual knowledge across regional contexts, Mango (Nguyen et al., 2024) extracts cross-cultural facts from LLMs, CultureAtlas (Fung et al., 2024) constructs a cultural knowledge base from Wikipedia and Wikidata, and CultureBank (Shi et al., 2024) collects cultural descriptors from social media such as TikTok and Reddit. However, these efforts primarily focus on factual or descriptive knowledge—typically short, unstructured snippets without inferential chains or sequential reasoning.

In contrast, we treat large language models as cultural archives, leveraging their internalized knowledge to construct structured, inferential representations. We introduce the Cultural Commonsense Knowledge Graph (CCKG)–a graph-based resource that models *if–then* reasoning chains involving human actions, intentions, and consequences.

Evaluating Cultural Commonsense Knowledge. Early research on commonsense evaluation was predominantly conducted in English, often lacking strong cultural grounding and reflecting universal or Western-centric perspectives. This research spans physical commonsense (Bisk et al., 2020), which evaluates models' understanding of real-world dynamics and object relations, and social reasoning (Sap et al., 2019b), which tests their ability to infer human emotions, intentions, and social norms. Later studies extended this line of inquiry to numerical (Lin et al., 2020; Akhtar et al., 2023), temporal (Tan et al., 2023), and causal reasoning (Du et al., 2022).

More recently, research has begun to examine the cultural dimensions of commonsense reasoning, investigating how language models encode and generalize culturally grounded knowledge. Koto et al. (2024) introduced `IndoCulture`, a dataset for evaluating commonsense reasoning across eleven Indonesian provinces, while Sadallah et al. (2025) proposed a complementary benchmark for Arab culture. Other efforts have explored cultural variation beyond commonsense reasoning: Durmus et al. (2024) introduced `GlobalOpinionQA`, built on the World Values Survey (Haerpfer et al., 2022), to analyze cross-national differences in LLM-generated responses, and `CultureNLI` (Huang and Yang, 2023) examined entailment across Indian and American cultural contexts. In our downstream tasks (Section 4.2), we focus on `IndoCulture` and `ArabCulture`, as they directly target cultural commonsense reasoning and are manually curated with high-quality annotations.

3 Cultural Commonsense Knowledge Graph (CCKG)

In this section, we detail our framework for constructing the Cultural Commonsense Knowledge Graph (CCKG) from LLMs.

3.1 Preliminaries

Let

$$G = (V, E)$$

be a directed labeled graph, where V denotes the set of actions and $E \subseteq V \times R \times V$ represents the set of labeled edges. Each edge $(A_i, R, A_j) \in E$ encodes an assertion of the form

$$A_i \xrightarrow{R} A_j,$$

interpreted as "if action A_i occurs, then a related action A_j follows," connected through relation R.

Actions. Actions are phrases that describe activities, events, or processes representing culturally grounded behaviors.

Assertion. An assertion is a triple $(A_i, R, A_j) \in E$ capturing culturally specific commonsense inferences as a conditional link between an initiating action A_i and a resulting action A_j via relation R, which defines their causal, motivational, or consequential connection.

Path. A path is an ordered sequence of assertions that connects an initiating action to a resulting action via their relations.

$$A_0 \xrightarrow{R_1} A_1 \xrightarrow{R_2} A_2 \xrightarrow{R_3} \dots \xrightarrow{R_k} A_k,$$

where each $(A_{i-1}, R_i, A_i) \in E$ for $i = 1, \dots, k$, A_0 is the initial action, A_k is the resulting or extended action, $A_1, \dots, A_{k-1}$ are intermediate actions, and $R_1, \dots, R_k$ specify the relations connecting them.

Relation. A relation connects two actions and defines the type of inferential link between them. We categorize five relation types (`xNext`, `xEffect`, `xNeed`, `oNext`, `oEffect`), summarized with illustrative examples in Table 1.

3.2 Knowledge Graph Construction

We designed an iterative, prompt-based method to generate CCKG. It can be constructed either in English or in the target country's native language. The overall construction procedure of the CCKG is summarized in Algorithm 1 and consists of two stages:

Initial Generation. Given a target language L (English or the native language of the country c), we prompt the LLM (using prompt 5 in the Appendix) to generate, for each subtopic s, possible

Relation	Definition	Example
`xNext`	What would x likely want to do after the action?	buys groceries → want to cook (If x buys groceries, x will want to cook.)
`xEffect`	What effects does the action have on x?	gives a gift → gets thanked (If x gives a gift, x gets thanked.)
`xNeed`	What does x need to do before the action?	cook a meal → gather ingredients (Before x can cook a meal, x needs to gather ingredients.)
`oNext`	What would others likely want to do after the action?	calls the police → dispatch officers (If x calls the police, others want to dispatch officers.)
`oEffect`	What effects does the action have on others?	insults someone → feel angry (If x insults someone, others will feel angry.)

Table 1: Types of relations between actions. Three are adopted from `ATOMIC` Sap et al. 2019a (`oEffect`, `xEffect`, `xNeed`) and two are newly introduced (`oNext`, `xNext`). In these relation types, x denotes the agent or primary person performing the action, while o refers to others who interact with or are affected by x's action. Text in brackets shows the "if $action_1$, then $action_2$" version.

cultural assertions guided by the predefined relation types. Each assertion is expressed as a triple (A_i, R, A_j), where A_i and A_j are actions and R is one of the relations defined in Table 1. The resulting assertions are stored in the knowledge base $\mathcal{K}_c^L$.

Iterative Expansion. Starting from assertions (A_i, R, A_j) with $R \in \{\texttt{oNext}, \texttt{xNext}\}$, the knowledge base is enriched over N user-specified number of expansions by elaborating paths. Expansion proceeds in two steps: **(i) Intermediate action expansion.** The LLM decomposes each (A_i, R, A_j) into a sequence of intermediate steps $A_i \xrightarrow{R_0} A_{i_1} \xrightarrow{R_1} A_{i_2} \xrightarrow{R_2} \cdots \xrightarrow{R_k} A_j$, adding intermediate triples to $\mathcal{K}_c^L$; **(ii) Forward expansion.** Using A_j as the new starting point, the LLM generates generates new continuations (A_j, R', A_k) where $R' \in \{\texttt{oNext}, \texttt{xNext}\}$, conditioned on the original context (A_i, R, A_j). To avoid redundant expansions, we maintain a set $\mathcal{U}$ containing all unique actions already present in $\mathcal{K}_c^L$. For each newly generated assertion (A_j, R', A_k), candidate actions A_k are matched against existing actions $\mathcal{U}$ via a similarity score, replacing A_k with $u \in \mathcal{U}$ if $\mathrm{sim}(A_k, u) > 0.8$. Otherwise, (A_j, R', A_k) is inserted as a novel assertion in $\mathcal{K}_c^L$ and included in the pool of candidates for further expansion. Because the number of assertions generated for each A_j may vary, we retain at most six assertions per list for expansion in the following iteration. This pruning strategy balances computational efficiency with knowledge coverage. Both steps of the iterative expansion leverage the prompt 6 in the Appendix to instruct the LLM to produce intermediate actions and forward actions.

4 Experiments

4.1 CCKG Extraction

Topic Taxonomy and Country Selection. We selected five countries: China, Indonesia, Japan, England, and Egypt, to ensure broad geographic coverage, cultural diversity, and the availability of native human evaluators. Their corresponding native languages are Chinese (CHI), Bahasa Indonesia (IND), Japanese (JAP), English (EN), and Modern Standard Arabic (MSA). We defined 11 daily-life topics comprising 65 fine-grained subtopics, adapted from Koto et al. (2024). These topics cover diverse aspects of everyday life, including food, weddings, art, habits, daily routines, family relationships, pregnancy and child-rearing, death, religious holidays, traditional games, and socio-religious practices. A full taxonomy of topics and subtopics is presented in Table 6 in the Appendix. While our current selection of countries is limited to this experimental setup, the proposed framework is general and can be extended to any cultural or linguistic context.

Evaluation Setup. To assess the quality of CCKG, we conduct a manual evaluation of assertions and their derived paths across three dimensions using binary labels (yes/no), following prior work (Nguyen et al., 2024; Bhatia and Shwartz, 2023). Specifically, we assess: (i) *Correctness* (**COR**): whether A_i and A_j are valid actions and the relation R accurately represents their connection (see Table 1); (ii) *Cultural relevance* (**CR**): whether the assertion is culturally specific to the

Algorithm 1: CCKG Construction for country c in language L

Input: Country c, language L, subtopics $\mathcal{S}$, expansion depth N
Output: Knowledge base $\mathcal{K}_c^L$
Init: $\mathcal{K}_c^L \leftarrow \emptyset, \mathcal{U} \leftarrow \emptyset, \mathcal{L}_0 \leftarrow \emptyset$; `// knowledge base, unique actions, expansion list`
1) Initial generation
foreach $s \in \mathcal{S}$ **do**
  Generate assertions $\mathcal{A}_s$ using `Prompt 5`, insert into $\mathcal{K}_c^L$;
  $\mathcal{U} \leftarrow \mathcal{U} \cup \{A_i, A_j \mid (A_i, R, A_j) \in \mathcal{A}_s\}$;
  $\mathcal{L}_0 \leftarrow \mathcal{L}_0 \cup \mathcal{A}_s$;
2) Iterative expansion
for $t = 1$ **to** N **do**
  $\mathcal{L}_t \leftarrow \emptyset$;
  foreach $(A_i, R, A_j) \in \mathcal{L}_{t-1}$ *with* $R \in \{\texttt{oNext}, \texttt{xNext}\}$ **do**
    `// Intermediate expansion`
    Generate decomposition $A_i \xrightarrow{R_0} A_{i_1} \xrightarrow{R_1} \dots \xrightarrow{R_k} A_j$ using `Prompt 6` ;
    Insert intermediate assertions into $\mathcal{K}_c^L$;
    Update $\mathcal{U} \leftarrow \mathcal{U} \cup \{A_{i_1}, \dots, A_{i_k}\}$;
    `// Forward expansion`
    Generate new assertions (A_j, R', A_k) with $R' \in \{\texttt{oNext}, \texttt{xNext}\}$ using Prompt 6 ;
    foreach (A_j, R', A_k) *generated* **do**
      $u^\star = \arg\max_{u \in \mathcal{U}} \mathrm{sim}(A_k, u)$, $\sigma = \max_{u \in \mathcal{U}} \mathrm{sim}(A_k, u)$;
      if $\sigma > 0.8$ **then**
        Replace A_k with $u^\star$, insert $(A_j, R', u^\star)$ into $\mathcal{K}_c^L$;
      else
        Insert (A_j, R', A_k) into $\mathcal{K}_c^L$;
        Retain at most 6 assertions per A_j in $\mathcal{L}_t$ for expansion;
        Add (A_j, R', A_k) to $\mathcal{L}_t$; `// candidate for next iteration`
        Update $\mathcal{U} \leftarrow \mathcal{U} \cup \{A_k\}$;

target country, as opposed to being universal or broadly cross-cultural; (iii) *Logical path coherence* (**LPC**): whether a sequence of actions forms a coherent, logically structured, and contradiction-free inferential chain.

The evaluation was conducted by expert annotators who are native speakers of the corresponding languages, possess at least a high-school diploma, and are proficient in both English and their native language (see §A in the Appendix for full eligibility criteria). For each country, we recruited two evaluators. To discourage careless responses, each evaluation set included five randomly embedded gold-standard samples for quality control, and annotators were required to correctly label at least four of them. All annotators were compensated at their country's minimum wage, and each task took approximately three hours to complete on average.

Preliminary Experiments. To identify the most suitable model for our main experiments, we compare two strong candidates: GPT-4o (OpenAI et al., 2024) as a closed-source representative and Llama-3.3-70B-IT (Grattafiori et al., 2024) as its open-source counterpart. This preliminary study focuses on China and Indonesia, and evaluates only the first-stage extraction (Section 3.2), as the quality of the subsequent *iterative expansion* stage critically depends on these initial outputs. We randomly sample 100 assertions across 11 topics for each country and ask native evaluators from the corresponding countries to assess **CR** and **COR**.[2] As shown in Figure 3 (Appendix D.1), GPT-4o consistently outperforms Llama-3.3-70B, and we therefore adopt GPT-4o for all subsequent experiments.

Extraction and Evaluation. We apply Algorithm 1 using GPT-4o for each country, generating assertions in both English and the respective native language (temperature = 1, $N = 3$). To remove duplicates, we use sentence embeddings (`all-MiniLM-L6-v2`[3] for English and `stsb-xlm-r-multilingual`[4] for other languages). After filtering out duplicates and malformed assertions, 37,363 English (of 38,858) and 16,709 native-language (of 17,043) assertions remain. We then construct simple paths for each subtopic by treating every initial action A_i as a source node, resulting in 27,649 English and 6,571 native-language paths. Detailed dataset statistics are provided in Table 7 (Appendix).

Result. To assess how language choice affects CCKG quality, we compared graphs generated in English against those produced in the corresponding native languages. As shown in Table 2, English CCKG consistently outperform native-language versions across nearly all evaluation dimensions. On average, English versions achieve higher scores in correctness, cultural relevance, and logical path coherence, indicating that LLMs express cultural knowledge more accurately and coherently when

[2] **LPC** is excluded since this experiment involves only the initial extraction stage.
[3] `https://www.sbert.net/docs/sentence_transformer/pretrained_models.html`
[4] `https://huggingface.co/sentence-transformers/stsb-xlm-r-multilingual`

Country (Language)	CR	COR	LPC
England (EN)	**40.0**	**96.6**	**82.9**
China (EN)	**80.8**	**86.9**	**70.2**
China (CHI)	59.1	85.2	59.9
Egypt (EN)	**56.9**	**89.2**	**96.1**
Egypt (MSA)	13.4	82.8	89.9
Japan (EN)	**72.7**	**88.3**	42.9
Japan (JAP)	55.9	83.4	**63.9**
Indonesia (EN)	40.0	**81.0**	72.2
Indonesia (IND)	**42.1**	70.7	**73.9**

Table 2: Average percentage of positive annotations (*yes* labels) for Correctness (COR), Cultural Relevance (CR), and Logical Path Coherence (LPC) across two annotators. Bold values indicate higher scores between English and native-language CCKG for each country and criterion.

operating in English. This pattern holds across diverse linguistic families—including Arabic, Chinese, and Japanese—suggesting that English serves as a more stable representational medium for encoding culturally grounded reasoning. Native-language CCKG, while sometimes capturing localized nuances, tend to produce less coherent or less contextually grounded inferential chains, likely due to limited language-specific pretraining data. Overall, these findings highlight a key asymmetry in current multilingual LLMs: despite aiming to model local cultural reasoning, they still represent cultural commonsense most effectively through English.

4.2 Evaluation on Cultural Commonsense Reasoning

Dataset. We evaluate whether integrating LLMs with cultural inferential knowledge from CCKG—used as in-context exemplars—enhances their performance on tasks requiring culturally grounded reasoning. We use two human-constructed benchmarks: `ArabCulture` (Sadallah et al., 2025) and `IndoCulture` (Koto et al., 2024). `ArabCulture` covers cultural commonsense from 13 Arab countries (including Egypt) and is written in Modern Standard Arabic (MSA), while `IndoCulture` represents cultural reasoning across 11 Indonesian provinces in Bahasa Indonesia. Both datasets span diverse cultural domains and everyday life scenarios, and can be evaluated in two formats: (i) multiple-choice question answering (MCQA), where each instance presents three candidate completions with exactly one correct answer, and (ii) sentence completion tasks (i.e., open-ended generation). All evaluations are conducted in both English and the respective native languages of each country.

Models. We experimented with 13 models in total: base models Llama3.2-1B/3B, Llama3.1-8B (Grattafiori et al., 2024), Qwen2.5-0.5B/1.5B/3B/7B (Yang et al., 2025), and Gemma2-2B/9B (Team et al., 2024); and instruction-tuned models Llama3.1-8B-I, Gemma2-9B-I, and Qwen2.5-7B-I. All models are used for cultural commonsense question answering, while only the instruction-tuned models are used for generation tasks, including cultural commonsense completion and story generation.

Augmentation Methods. We perform in-context augmentation with relevant assertions (5-shot, ***-Asrt***) or paths (1-shot, ***-Path***), on both MCQA and sentence completion tasks. Here, we use SBERT embeddings (Reimers and Gurevych, 2019) with `stsb-xlm-r-multilingual`[5] for semantic search[6] to retrieve the most relevant assertions and paths.

As baselines, we include (i) zero-shot prompting without in-context augmentation, denoted as ***Base***, and (ii) chain-of-thought prompting (CoT; Wei et al. 2022) for MCQA. We also compare with in-context augmentation using `Mango` (5-shot, ***Mango***), a widely used LLM-extracted cultural commonsense knowledge base that provides factual assertions but does not model paths.

Evaluation Metrics. For MCQA, we report accuracy using the official evaluation scripts and generation parameters provided by each benchmark. For sentence completion, we evaluate using BERTScore-F1 (Zhang* et al., 2020) and sentence similarity (Corley and Mihalcea, 2005), computed between the LLM-generated text and the corresponding reference completion. All experiments use the original benchmark prompts.

Results on MCQA. Table 3 presents the MCQA accuracy across models and augmentation methods using English prompts, while results for Arabic and Indonesian prompts are provided in Appendix D.6 and D.7. Overall, integrating CCKG knowledge—either through assertions

[5] https://huggingface.co/sentence-transformers/stsb-xlm-r-multilingual

[6] https://github.com/UKPLab/sentence-transformers/blob/master/examples/applications/semantic-search/semantic_search.py

Models	IndoCulture							ArabCulture						
	Before Aug		After Aug					Before Aug		After Aug				
	Base	CoT	Mango	E-Asrt	E-Path	N-Asrt	N-Path	Base	CoT	Mango	E-Asrt	E-Path	N-Asrt	N-Path
Qwen2.5-0.5B	43.1	35.8	**43.5**	43.4	42.2	41.6	41.9	33.3	34.2	**34.3**	33.8	33.8	34.2	34.2
Qwen2.5-1.5B	43.8	**45.5**	45.0	45.2	44.8	44.5	41.8	39.4	43.4	43.5	45.2	41.4	**48.3**	42.7
Qwen2.5-3B	53.0	53.1	**54.3**	53.2	51.7	53.6	51.8	37.5	35.2	40.6	44.6	39.8	**45.7**	38.1
Qwen2.5-7B	58.5	53.4	59.7	60.1	59.1	**60.8**	58.6	49.3	40.6	52.6	53.6	46.5	**59.6**	51.1
Gemma2-2B	33.4	35.5	38.3	37.5	38.2	35.7	**39.4**	34.3	34.3	34.3	34.2	34.0	**34.6**	34.0
Gemma2-9B	65.2	53.2	64.8	64.6	65.0	65.7	**65.9**	**34.5**	34.3	34.3	34.3	34.3	34.3	34.3
Llama3.2-1B	46.9	44.5	48.9	50.9	51.0	**53.4**	53.1	**33.9**	33.8	33.7	33.8	33.8	33.8	33.4
Llama3.2-3B	49.4	41.3	49.2	49.6	49.4	50.0	**50.5**	33.3	**37.3**	34.0	34.0	33.0	34.4	31.9
Llama3.1-8B	32.7	**35.6**	32.7	32.7	32.7	32.7	32.7	34.9	35.3	34.5	34.8	34.2	**35.4**	34.7
Gemma2-9B-IT	57.9	39.1	57.9	57.3	59.1	59.4	**61.0**	**57.3**	34.3	47.0	43.7	46.8	42.9	49.3
Qwen2.5-7B-IT	66.2	63.5	65.3	66.1	66.9	66.3	**67.5**	48.7	34.2	46.5	47.8	49.3	**50.8**	49.2
Llama3.1-8B-IT	55.5	**59.0**	53.4	54.2	54.4	54.4	56.4	**49.3**	34.4	35.4	37.0	40.1	37.2	39.3
Avg Δ	NA	−3.8	0.6	0.8	0.7	1.0	**1.2**	NA	−4.6	−1.3	−0.7	−1.5	**0.4**	−1.2

Table 3: Accuracy comparison on MCQA across different methods. **E-Asrt**: English CCKG Assertions; **E-Path**: English CCKG Paths; **N-Asrt**: Native-language (Arabic or Indonesian) CCKG Assertions; **N-Path**: Native-language (Arabic or Indonesian) CCKG Paths. "Avg Δ" denotes the average improvement over the baseline. Best results per model are highlighted in bold.

or paths—consistently improves performance in `IndoCulture`. For `ArabCulture`, the improvement is not observed in Mango, but only in our CCKG assertion. The largest gains are observed when models are augmented with native-language assertions, suggesting that culturally grounded examples expressed in the original language are more effective in guiding model predictions. For instance, Qwen2.5-7B improves from 58.5% to 60.8% on `IndoCulture` and from 49.3% to 59.6% on `ArabCulture` when augmented with Indonesian and Arabic assertions, respectively. On average, native-language augmentation achieves the highest improvement (+1.2 points), outperforming English-based augmentation and Mango.

Smaller or base models benefit the most from in-context augmentation, implying that explicit inferential cues from CCKG compensate for their limited internalized cultural knowledge. Larger instruction-tuned models, such as Gemma2-9B-IT and Qwen2.5-7B-IT, show more modest or mixed gains, likely because they already encode general cultural knowledge, making additional context less impactful.

In contrast, chain-of-thought prompting (CoT) performs worse than the baseline across nearly all models and datasets, with average drops of 3.8 points on `IndoCulture` and 4.6 points on `ArabCulture`. This suggests that cultural commonsense reasoning relies more on intuitive and context-dependent knowledge than on step-by-step logical reasoning—a finding consistent with prior observations that explicit reasoning often weakens culturally situated inference (Sadallah et al., 2025).[7]

Results on Sentence Completion. As shown in Table 4, incorporating context from CCKG generally improves both similarity and BERTScore-F1. Native-language assertions and paths yield the highest gains in similarity, with modest but consistent improvements in BERTScore, outperforming both the baseline without augmentation and Mango augmentation. For example, compared to the baseline, Llama3.1-8B-IT improves similarity scores from 32.6% to 36.0% on IndoCulture and from 29.9% to 33.5% on ArabCulture. Qwen2.5-7B-IT shows comparable gains, increasing from 39.6% to 43.0% and from 38.8% to 42.4%, respectively. BERTScore changes are modest, with Llama3.1-8B-IT exhibiting slight drops (65.0% to 64.5% on ArabCulture, 70.2% to 70.1% on IndoCulture), while Qwen2.5-7B-IT shows marginal improvements (72.3% to 72.6% on ArabCulture) compared to the baseline. Similar trends are observed with prompts in native language (see §D.7).

4.3 Evaluation on Story Generation

To further examine the usefulness of CCKG paths, we conducted a free-form short story generation

[7] Similar trends are also observed with native prompts (§D.6).

Models	Before Aug	+Mango	+CCKG Methods			
			E-Asrt	E-Path	N-Asrt	N-Path
IndoCulture w/ Sentence Similarity Score						
Gemma2-9B-IT	32.0	33.3	34.4	34.6	**35.5**	35.4
Qwen2.5-7B-IT	39.6	41.7	42.5	42.6	42.5	**43.0**
Llama3.1-8B-IT	32.6	33.8	34.3	34.9	36.1	**36.0**
Avg	34.8	36.3	37.1	37.3	38.0	**38.1**
IndoCulture w/ Avg BERT Score F1						
Gemma2-9B-IT	71.2	71.0	71.3	71.3	**71.5**	71.4
Qwen2.5-7B-IT	72.3	72.3	**72.6**	72.5	**72.6**	72.5
Llama3.1-8B-IT	**70.2**	69.6	69.9	70.0	70.1	70.0
Avg	71.3	71.0	71.3	71.2	**71.4**	71.3
ArabCulture w/ Sentence Similarity Score						
Gemma2-9B-IT	33.9	33.0	36.5	32.9	**37.2**	34.1
Qwen2.5-7B-IT	38.8	39.2	42.0	35.3	**42.5**	37.6
Llama3.1-8B-IT	29.9	29.1	31.2	26.6	**33.5**	29.4
Avg	34.2	33.8	36.5	31.6	**37.7**	33.7
ArabCulture w/ Avg BERT Score F1						
Gemma2-9B-IT	68.6	68.4	68.7	68.6	**68.7**	68.6
Qwen2.5-7B-IT	67.8	67.4	67.4	65.7	**67.8**	66.5
Llama3.1-8B-IT	**65.0**	63.5	63.0	63.2	64.5	64.0
Avg	67.0	66.4	66.3	65.8	**67.0**	66.4

Table 4: Sentence similarity and BERT scores for sentence completion task. **E-Asrt**: CCKG English Assertions, **E-Path**: CCKG English Paths, **N-Asrt**: CCKG Native-language Assertions, **N-Path**: CCKG Native-language Paths. Best results per row are bolded.

task covering 25 randomly selected subtopics (see prompts in §E.2). Stories were generated in both English and the native languages of Egypt, China, and Indonesia—chosen based on the availability of qualified human evaluators. We compared three setups: baseline zero-shot prompting, in-context inference with *+Mango* (5-shot assertions from Mango), and *+CCKG* (1-shot paths retrieved from CCKG). Relevant assertions or paths were selected using SBERT embeddings, following the same retrieval procedure described in §4.2. For CCKG, we focus on path-based augmentation here, as story generation naturally benefits from sequential and causal structure.

Evaluation Metrics. For the evaluation, we primarily relied on human judgments. Two annotators rated each story on a 1–10 Likert scale along three dimensions: 1) *Cultural relevance* (**CR**), which measures how accurately the story reflects the traditions, customs, values, and social norms of the country; 2) *Fluency* (**FL**), which assesses grammatical correctness, sentence structure, vocabulary, and readability; and 3) *Coherence* (**CO**), which analyzes the logical flow, clarity, and consistency of events and character actions. As a complementary analysis, we also employed LLM-as-a-Judge (Qiu et al., 2025; Li et al., 2024), using GPT-4o (prompts in Appendix E.2) with the same evaluation criteria and examining its correlation with human judgments.

	China			Indonesia			Egypt		
	CR	FL	CO	CR	FL	CO	CR	FL	CO
Llama3.1-8B-IT	6.3	8.4	7.6	6.9	8.1	7.7	6.3	8.9	8.7
+Mango	7.0	8.4	7.6	7.0	8.3	8.0	7.0	9.1	8.9
+CCKG	**7.3**	**9.0**	**8.2**	**7.7**	**8.4**	**8.3**	**8.7**	**9.1**	**9.5**
Qwen2.5-7B-IT	7.0	8.7	7.9	6.6	7.5	7.2	7.0	8.9	8.8
+Mango	6.9	8.9	7.9	6.8	8.3	7.7	7.3	8.9	8.8
+CCKG	**7.3**	**8.9**	**8.5**	**7.3**	**8.3**	**7.9**	**8.9**	**9.0**	**9.2**
Gemma2-9B-IT	6.5	8.8	7.7	7.6	8.7	8.4	6.5	8.5	8.4
+Mango	6.9	9.0	7.8	6.9	8.5	8.1	6.8	8.6	8.6
+CCKG	**7.8**	**9.2**	**8.7**	**8.0**	**8.9**	**8.8**	**7.8**	**8.8**	**8.7**

Table 5: Aggregated annotator scores for English story generation, comparing Base, **+Mango**, and **+CCKG**. **CR**: Cultural relevance, **FL**: Fluency, **CO**: Coherence. Best scores are bolded; inter-annotator correlation is strongest for CR (0.72), moderate for CO (0.34), and weak for FL (0.26) (Appendix D.3).

Results. Table 5 summarizes the aggregated human evaluation scores for English story generation. Across all nine model–country pairs, incorporating CCKG paths consistently improves story quality, with the largest gains observed in *Cultural Relevance*—averaging a +1.4 increase for Llama models over the baseline across three countries. *Fluency* and *Coherence* also show steady, smaller improvements, suggesting that path-based augmentation helps models produce more logically structured and contextually grounded narratives. Results for native-language story generation show greater variation and are detailed in Appendix D.2.

Figure 2 further shows the relative improvements in story quality when augmenting with CCKG over the baseline, across three evaluation metrics for each country and each model. Overall, the benefits of CCKG paths are more evident in English story generation, with Indonesian story generation using Llama3.1-8B-IT being a notable exception. Full per-metric results are provided in Appendix D.8.

LLM-as-a-Judge aligns with human ratings on cultural relevance in native languages but only moderately in English. We observe moderate correlations between the LLM judge and human evaluators on *cultural relevance* (average 0.4) in English story generation, but stronger correlations (average 0.8) in native languages. Interestingly,

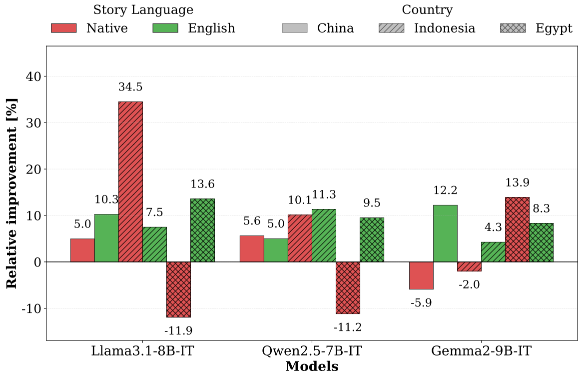

Figure 2: Relative improvement from **+CCKG** over the baseline in *Native* vs. *English* story generation. Bars show percentage lift in average human scores (Cultural relevance, Fluency, Coherence); numbers above bars indicate gains in percentage points.

negative correlations are found for English generations in *Fluency* (–0.1) and *Coherence* (0.0), but strong positive correlations for generations in native languages (0.9 and 0.8, respectively). One possible explanation is that English evaluations may be more sensitive to stylistic variation—for example, differing preferences for concise versus elaborative writing. In addition, culturally specific or tradition-related expressions often sound more natural and authentic in their native languages than in English translations, which may further influence evaluators' preferences. Full LLM evaluation results and correlations with human judgments are provided in Section D.4 in the Appendix.

5 Conclusion

This paper explores LLMs as cultural technologies and knowledge extractors through **CCKG**, a framework for constructing multilingual cultural commonsense knowledge chains that extend inferential reasoning beyond static, English-centric resources. By modeling culture as procedural and sequential rather than as isolated facts, **CCKG** captures the flow of cultural practices across languages. Human evaluations show that while native languages convey richer cultural depth, English outputs are generally more coherent and preferred. Empirically, augmenting LLMs with **CCKG** improves performance on cultural commonsense reasoning and story generation.

Limitations

Our method for constructing the CCKG relies on prompting, which makes it sensitive to the specific prompt formulations. Consequently, some degree of prompt tuning may be required when applying the approach to new models. Nevertheless, we ran experiments with two state-of-the-art language models: a closed-source model (GPT-4o) and an open-source model (Llama-3.3-70B-Instruct). We successfully extracted CCKG from both, demonstrating that our method is robust across different model types.

Automatically extracting cultural commonsense knowledge from LLMs carries the potential risk of reproducing stereotypes. In this work, we did not focus on detecting or evaluating such biases. However, our human evaluators reviewed a subset of the extracted content for quality analysis (see Section D.5 in Appendix), and the majority of the items were not flagged as stereotypical or harmful material. We plan to conduct a more in-depth investigation of this issue in future work.

In this work, we focus on a limited set of high(er)-resource cultures, reflecting both the accessibility to human evaluators and the assumption that LLMs have already acquired substantial knowledge about these cultures during pre-training. We further evaluate culture at the country level, which we plan on extending to more fine-grained levels in the future.

We use data extracted from LLMs as a research prototype and as an exploratory foundation for the concept of LLMs as Cultural Archives. Importantly, this extracted content should not be viewed as a formal dataset. We advise against its use in production systems without careful consideration of both the potential benefits—such as enabling more culturally aware technologies—and the corresponding challenges and risks, including the possible reinforcement of stereotypes or other unintended biases.

Acknowledgments

We thank the supercomputing center at MBZUAI for providing resources for this research. We also gratefully acknowledge MBZUAI's support in funding the hiring of human annotators. This work was further supported by the LOEWE Distinguished Chair "Ubiquitous Knowledge Processing", LOEWE initiative, Hesse, Germany (Grant Number: LOEWE/4a//519/05/00.002(0002)/81).

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.

Mubashara Akhtar, Abhilash Shankarampeta, Vivek Gupta, Arpit Patil, Oana Cocarascu, and Elena Simperl. 2023. Exploring the numerical reasoning capabilities of language models: A comprehensive analysis on tabular data. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15391–15405, Singapore. Association for Computational Linguistics.

Saeed Almheiri, Rania Elbadry, Mena Attia, Chenxi Wang, Preslav Nakov, Timothy Baldwin, and Fajri Koto. 2025. Cross-cultural transfer of commonsense reasoning in LLMs: Evidence from the Arab world. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4593–4614, Suzhou, China. Association for Computational Linguistics.

Mehar Bhatia and Vered Shwartz. 2023. GD-COMET: A geo-diverse commonsense inference model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7993–8001, Singapore. Association for Computational Linguistics.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7432–7439. AAAI Press.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, pages 1877–1901. Curran Associates, Inc.

Courtney Corley and Rada Mihalcea. 2005. Measuring the semantic similarity of texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 13–18, Ann Arbor, Michigan. Association for Computational Linguistics.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-CARE: a new dataset for exploring explainable causal reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 432–446, Dublin, Ireland. Association for Computational Linguistics.

Esin Durmus, Karina Nguyen, Thomas Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2024. Towards measuring the representation of subjective global opinions in language models. In *First Conference on Language Modeling*.

Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively multi-cultural knowledge acquisition & LM benchmarking. *arXiv preprint*, abs/2402.09369.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, and Aston Zhang et al. 2024. The llama 3 herd of models. *arXiv preprint*, abs/2407.21783.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, Bjorn Puranen, et al. 2022. World values survey: Round seven–country-pooled datafile version 5.0.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Jing Huang and Diyi Yang. 2023. Culturally aware natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP*

2023, pages 7591–7609, Singapore. Association for Computational Linguistics.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 6384–6392. AAAI Press.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

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.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *arXiv preprint*, abs/2412.05579.

Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. Birds have four legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-Trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6862–6868, Online. Association for Computational Linguistics.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025. Culturally aware and adapted NLP: A taxonomy and a survey of the state of the art. *Transactions of the Association for Computational Linguistics*, 13:652–689.

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.

Tuan-Phong Nguyen, Simon Razniewski, Aparna S. Varde, and Gerhard Weikum. 2023. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, pages 1907–1917. ACM.

Tuan-Phong Nguyen, Simon Razniewski, and Gerhard Weikum. 2024. Cultural commonsense knowledge for intercultural dialogues. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 1774–1784, New York, NY, USA. Association for Computing Machinery.

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

Haoyi Qiu, Alexander Fabbri, Divyansh Agarwal, Kung-Hsiang Huang, Sarah Tan, Nanyun Peng, and Chien-Sheng Wu. 2025. Evaluating cultural and social awareness of LLM web agents. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3978–4005, Albuquerque, New Mexico. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Julien Romero, Simon Razniewski, Koninika Pal, Jeff Z. Pan, Archit Sakhadeo, and Gerhard Weikum. 2019. Commonsense properties from query logs and question answering forums. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 1411–1420, New York, NY, USA. Association for Computing Machinery.

Abdelrahman Sadallah, Junior Cedric Tonga, Khalid Almubarak, Saeed Almheiri, Farah Atif, Chatrine Qwaider, Karima Kadaoui, Sara Shatnawi, Yaser Alesh, and Fajri Koto. 2025. Commonsense reasoning in Arab culture. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7695–7710, Vienna, Austria. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019a. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3027–3035. AAAI Press.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019b. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the*

9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Maarten Sap, Vered Shwartz, Antoine Bosselut, Yejin Choi, and Dan Roth. 2020. Commonsense reasoning for natural language processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 27–33, Online. Association for Computational Linguistics.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula, and Diyi Yang. 2024. CultureBank: An online community-driven knowledge base towards culturally aware language technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4996–5025, Miami, Florida, USA. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Kai Sun, Yifan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2024. Head-to-tail: How knowledgeable are large language models (LLMs)? A.K.A. will LLMs replace knowledge graphs? 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 311–325, Mexico City, Mexico. Association for Computational Linguistics.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023. Towards benchmarking and improving the temporal reasoning capability of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14820–14835, Toronto, Canada. Association for Computational Linguistics.

Niket Tandon, Gerard de Melo, Fabian M. Suchanek, and Gerhard Weikum. 2014. Webchild: harvesting and organizing commonsense knowledge from the web. *Proceedings of the 7th ACM international conference on Web search and data mining*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, and Thomas Mesnard et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint*, abs/2408.00118.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *arXiv preprint*, abs/2412.15115.

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. GeoMLAMA: Geo-diverse commonsense probing on multilingual pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2039–2055, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Annotator Criteria

We recruited two annotators per country (ten in total) and applied strict eligibility requirements to ensure cultural authenticity and linguistic proficiency. Annotators were required to meet the following criteria:

- Native speaker of the specified local language and proficient in English (speaking and comprehension).
- Resided in the country for at least ten years.
- Demonstrated deep familiarity with the country's culture.
- Both parents are also natives residing in the same country.
- Minimum qualification of senior high school graduation (higher degrees preferred).

Among the ten annotators, four held a Bachelor's degree, three a Master's degree, two a Ph.D., and one a postdoctoral qualification. To discourage careless responses, five gold-standard samples were randomly embedded in each evaluation set for quality control, and annotators were required to correctly label at least four of them. Annotators were compensated at their country's minimum wage, and the task took approximately three hours on average.

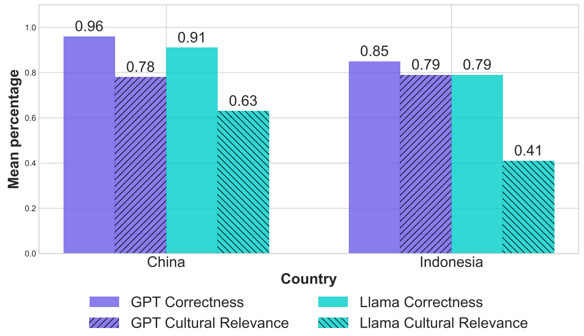

Figure 3: Performance of Llama3.3-70B-IT and GPT-4o on initial generation data to determine the optimal model for CCKG generation.

B Topic Diversity

For our study, we defined 11 daily-life topics encompassing 65 fine-grained subtopics (see Table 6). Each subtopic was translated into the native languages of the 4 target countries (except England) by native speakers, enabling both English and native-language CCKG generation settings. For Japan and China, additional care was taken during translation to preserve culturally specific nuances.

B.1 Data Statistics

Table 7 shows the dataset statistics of CCKG.

C Software

In this paper, we use the Huggingface Transformers library for experiments with cultural commonsense QA. For all the free-form generation tasks, we use the APIs provided by OpenRouter[8].

D Additional Results

D.1 Model Selection

Figure 3 presents the evaluation results of Llama-3.3-70B-instruct and GPT-4o models on the initial CCKG generation. Overall, the GPT-4o model shows better generation quality measured by cultural relevance and correctness.

D.2 Native Story Generation Results

In native-language story generation, results vary more substantially, reflecting the interaction between language and model alignment (see Table 8). CCKG provides the largest benefit when baseline cultural relevance is weaker—e.g., Llama/Indonesia +1.6 CR ($4.7 \rightarrow 6.3$), Qwen/Indonesia +0.7 ($5.8 \rightarrow 6.5$), and Gemma/Egypt +0.7 ($4.7 \rightarrow 5.4$)—but has little or even negative effect where performance is already strong or misaligned (e.g., Llama/Egypt -0.0, Qwen/Egypt -0.2). Qwen's Chinese baseline is already high (7.68 cultural relevance) but still benefits from CCKG ($\rightarrow 8.5$, $+0.8$). Fluency and coherence remain largely stable, though dips appear in some cases (e.g., Gemma/Indonesia fluency $8.3 \rightarrow 7.4$), underscoring the importance of language–model fit.

D.3 Pearson Correlation Between Annotators in Story Generation

Table 9 reports Pearson correlation scores between the two human annotators, broken down by country, story generation setting, and evaluation criterion. We observe strong agreement in native-language evaluations across all three dimensions—cultural relevance, fluency, and coherence (e.g., Indonesian: 0.9 / 0.8 / 0.7; Chinese: 0.9 / 0.9 / 0.9; MSA: 0.9 / 0.9 / 0.9). In contrast, agreement in English is more variable: correlations are moderate for cultural relevance but substantially weaker for fluency and coherence in the Indonesian and Egyptian sets (e.g., Indonesia-EN: 0.2; Egypt-EN fluency: 0.0), while Chinese-English shows moderate consistency (0.6–0.9). This discrepancy suggests that evaluating stories in English introduces greater variability for fluency and coherence. Two factors likely contribute. First, differences in dialect and stylistic preferences may shape judgments: one annotator may prefer concise, direct English, while another favors more elaborate phrasing or richer descriptive detail. Second, culturally specific expressions and tradition-related terms often sound more natural and authentic in their native languages than in English translation. When such concepts are rendered in English, they may lose nuance or appear less idiomatic, leading to divergent impressions of fluency or coherence.

D.4 LLMs-as-Judge for Story Generation Evaluation

Table 12 reports Pearson correlations between GPT-4o scores and human judgments. GPT-4o aligns well with human ratings when inter-annotator agreement is high, showing strong correlations in native languages (Indonesian: 0.7 / 0.8 / 0.7; Chinese: 0.7 / 0.8 / 0.7; MSA: 0.9 / 0.9 / 0.9). In contrast, correlations in English are moderate or even negative (e.g., Egypt–EN fluency: -0.6; Indonesia–EN coherence: 0.0).These unstable cor-

[8]https://openrouter.ai

Topics	Subtopics
Food	Breakfast, lunch, dinner, traditional foods and beverages, cutlery, cooking ware, fruit, food souvenirs, snacks
Wedding	Wedding location, wedding food, wedding dowry, traditions before marriage, traditions when getting married, traditions after marriage, men's wedding clothes, women's wedding clothes, songs and activities during the wedding, invited guests at a wedding, gift brought to weddings, food at a wedding
Habits	Eating habit, greetings habits, financial habits (saving, debit/credit), punctuality habit, cleanliness habit, shower time habit, transportation habit, popular sports
Art	Musical instruments, folks songs, traditional dances, use of art at certain events, poetry or similar literature
Daily activities	Morning activities, afternoon activities, evening activities, leisure and relaxation activities, household activities (cleaning, home management)
Family relationship	Relationships within the main family, relationships in the extended family, relations with society/neighbors, clan/descendant system
Pregnancy and kids	Traditions during pregnancy, traditions after birth, how to care for a newborn baby, how to care for toddlers, how to care for teenagers, parents and children interactions as adults
Death	When death occurs, the process of dealing with a corpse, traditions after the body is buried, the clothes of the mourners, inheritance matters
Religious holiday	Traditions before religious holidays, traditions leading up to religious holidays, traditions during religious holidays, traditions after holidays
Traditional games	Traditional game types
Socio-religious aspects of life	Regular religious activities, mystical things, traditional ceremonies, lifestyle, self care, traditional medicine, traditional sayings

Table 6: Overview of topics and their associated subtopics.

Country (Language)	Unique Nodes	Unique Paths	Total Assertions	Avg Path Length	Eval Assertions
England (EN)	7698	6174	8693	11.47	396
Indonesia (EN)	6267	3877	6905	10.72	355
Indonesia (IND)	5946	2398	6300	7.35	220
China (EN)	6082	3923	6721	9.58	335
China (CHI)	3051	1179	3059	5.48	297
Japan (EN)	6713	7581	7565	18.23	451
Japan (JAP)	2663	1057	2629	4.07	220
Egypt (EN)	6601	6094	7479	17.40	393
Egypt (MSA)	4527	1937	4721	6.33	276

Table 7: Dataset statistics across countries and languages. The Eval Assertions column shows the number of assertion samples to evaluate for 50 paths. The number of edges is the same as the number of assertions.

	China			Indonesia			Egypt		
Model	CR	Fl	CO	CR	Fl	CO	CR	Fl	CO
Llama3.1-8B-IT	3.6	3.8	3.4	4.7	4.9	4.6	1.9	2.1	1.9
+Mango	3.5	3.2	3.0	5.4	5.9	5.5	2.0	1.8	1.7
+CCKG	**3.9**	**3.7**	**3.8**	**6.3**	**6.5**	**6.4**	1.9	1.6	1.7
Qwen2.5-7B-IT	7.7	6.8	8.2	5.8	6.2	6.1	3.0	3.0	3.3
+Mango	8.5	7.5	8.1	5.4	6.4	6.2	2.6	2.4	2.5
+CCKG	**8.5**	**7.5**	7.9	**6.5**	**6.9**	**6.6**	**2.8**	**2.6**	**2.8**
Gemma2-9B-IT	7.4	7.4	7.9	6.6	8.3	7.4	4.7	5.1	5.1
+Mango	7.7	7.4	7.7	6.1	7.2	6.4	4.9	5.2	5.6
+CCKG	7.4	6.8	7.2	**7.2**	7.4	**7.3**	**5.4**	**5.4**	**6.1**

Table 8: Aggregated annotator scores across models and countries for native story generation, comparing the base setting (no augmentation) with two augmented settings: Mango and CCKG. Best scores are bolded.

relations likely reflect the same factors underlying lower inter-annotator agreement in English (see §D.3)—notably differences in stylistic expectations and the fact that culturally specific or tradition-related expressions often sound more natural in their native languages than in English. When such expressions are translated or adapted into English, they may lose nuance or feel less idiomatic, introducing greater variability in perceived fluency and coherence and making English evaluations overall less consistent. Table 11 and 10 report the scores obtained using GPT-4o in the LLMs-as-judge setting for the evaluation of native and English stories, respectively.

D.5 Assessment of Cultural Generalization and Stereotyping

We conducted a qualitative assessment by two native speakers of 30 assertions related to the culturally sensitive topics of *wedding* and *death* across four countries, namely Egypt, China, Indonesia, and Japan, in both English and native-language.

Story Generation	**Cultural Relevance**	**Fluency**	**Coherence**
Indonesia			
Indonesian	0.9	0.8	0.7
English	0.4	0.2	0.2
China			
Chinese	0.9	0.9	0.9
English	0.9	0.6	0.7
Egypt			
MSA	0.9	0.9	0.9
English	0.9	0.0	0.2

Table 9: Pearson correlation coefficients between Annotator 1 and Annotator 2 for the story-generation task across English and native-language settings.

	China			Indonesia			Egypt		
	CR	Fl	CO	CR	Fl	CO	CR	Fl	CO
Llama3.1-8B-IT	6.0	**7.6**	**4.8**	6.2	7.1	4.5	5.9	**7.3**	4.6
+Mango	6.3	6.6	4.5	6.7	**6.9**	**4.8**	**6.6**	7.2	**4.8**
+CCKG	**7.9**	**6.1**	4.2	**6.8**	6.7	4.5	**6.6**	6.9	**4.8**
Qwen2.5-7B-IT	6.7	**8.2**	**7.3**	6.3	**7.6**	5.8	5.8	**7.8**	5.6
+Mango	6.8	8.1	4.9	6.8	7.2	4.8	6.6	7.5	5.2
+CCKG	**7.4**	8.0	6.4	**6.9**	7.5	**5.9**	**6.7**	7.4	**5.8**
Gemma2-9B-IT	**7.7**	7.7	**6.1**	7.5	**8.1**	**5.8**	6.5	**8.3**	**5.5**
+Mango	7.6	**8.0**	5.4	**7.7**	7.8	5.2	**7.0**	7.7	4.8
+CCKG	7.6	7.2	5.2	7.5	**8.1**	**5.8**	6.9	7.7	5.2

Table 10: GPT-4o as Judge scores across models and countries for English story generation, comparing the base setting (no augmentation) with two augmented settings: Mango and CCKG. Best scores are bolded.

For each setting, we applied the following evaluation instruction:

> *Assess whether the text contains overly generalized, stereotypical, or unfair assumptions about individuals within the target culture, including but not limited to those based on gender roles, religion, ethnicity, regional affiliation, or socioeconomic class.*

Table 13 reports the average percentage of annotator judgments (“Yes”, “No”, and “Unsure”). Across most countries and languages, the majority of extracted assertions were judged as not containing overly generalized or stereotypical assumptions. In our extractions, we also observed that the model frequently uses modal language such as ‘might‘ and ‘will likely‘ to describe plausible reactions and responses of people in the given situations. Nonetheless, stereotypes and bias in LLMs are serious concerns, so we still recommend that future

	China			Indonesia			Egypt		
	CR	Fl	CO	CR	Fl	CO	CR	Fl	CO
Llama3.1-8B-IT	5.6	**7.0**	4.8	5.4	5.5	4.4	1.6	**1.8**	**1.6**
+Mango	**7.2**	6.9	**5.6**	6.3	**5.8**	**4.7**	**1.8**	**1.8**	1.5
+CCKG	**7.2**	6.0	5.2	**6.4**	5.6	4.6	**1.8**	1.6	1.5
Qwen2.5-7B-IT	6.6	7.2	6.8	6.2	6.9	6.4	3.9	**3.4**	3.2
+Mango	7.5	**7.9**	7.2	**7.2**	7.4	7.1	4.3	2.9	**3.0**
+CCKG	**7.6**	7.8	**7.7**	**7.2**	**7.7**	**7.5**	**4.5**	2.7	**3.0**
Gemma2-9B-IT	**8.6**	**9.6**	**8.8**	8.2	9.3	8.4	5.8	**5.2**	5.1
+Mango	8.4	8.6	8.0	**8.5**	**8.9**	8.1	**6.4**	5.0	5.1
+CCKG	8.0	8.5	8.3	8.3	**8.9**	**8.5**	6.3	5.0	**5.3**

Table 11: GPT-4o as Judge scores across models and countries for native story generation, comparing the base setting (no augmentation) with two augmented settings: Mango and CCKG. Best scores are bolded.

Story Generation	Cultural Relevance	Fluency	Coherence
Indonesia			
Indonesian	0.7	0.8	0.7
English	0.4	0.1	0.0
China			
Chinese	0.7	0.8	0.7
English	0.5	0.1	0.1
Egypt			
MSA	0.9	0.9	0.9
English	0.4	−0.6	−0.1

Table 12: Pearson correlation between LLM evaluations and human annotators (Annotator 1 and Annotator 2) for the story-generation task across English and native-language settings.

work include more detailed and large-scale audits on more topics.

D.6 MCQA Results with Native Prompts

Table 14 reports accuracy scores on the ArabCulture and IndoCulture benchmarks for the MCQA task using native prompts. The results follow the same trends observed with English prompts, as discussed in the main text.

D.7 Sentence Completion Results with Native Prompts

Table 15 presents BERT F1 and sentence similarity scores on the ArabCulture and IndoCulture benchmarks for the sentence completion task with native prompts, again showing the same trends as with English prompts.

	No (%)	Yes (%)	Unsure (%)
Egypt (EN)	100.0	0	0.0
Egypt (MSA)	93.3	3.3	3.3
Indonesia (IND)	90.0	5.0	5.0
Indonesia (EN)	98.3	1.7	0.0
Japan (EN)	91.7	8.3	0.0
Japan (JAP)	80.0	20.0	0.0
China (EN)	100.0	0.0	0.0
China (CHI)	100.0	0.0	0.0

Table 13: Average percentage of human judgments assessing whether extracted assertions related to *wedding* and *death* contain overly generalized or stereotypical cultural assumptions.

D.8 Results by Criterion: English vs. Native Story Generation

Figure 4 illustrates the average score for each evaluation metric, broken down by country and model.

E Prompts

This section presents all the prompts used in our experiments. All the translations were produced by native speakers of the country.

E.1 Prompts for CCKG

Figure 5 illustrates the prompt employed during the initial generation phase of our algorithm, while Figure 6 shows the prompt used in the iterative expansion phase. Both prompts are applied verbatim when constructing CCKG in English. For cases where CCKG is generated in a non-English language, the same prompts are translated into the target language. In our experiments, this includes Modern Standard Arabic (MSA), Chinese, Japanese, and Bahasa Indonesian.

E.2 Prompts for Story Generation

Figure 10 presents the base and augmentation prompts used for story generation in both English and the native languages. The variables `SUBTOPIC`, `COUNTRY`, `LANGUAGE`, and the assertions are replaced with their corresponding values. When the story is generated in English, the variable `LANGUAGE` is set to "English"; otherwise, it is set to the respective native language.

E.3 Prompts for Evaluation with LLMs-as-Judge

Figures 9, 8, and 7 show the prompts used to evaluate cultural relevance, fluency, and coherence, re-

Models	IndoCulture							ArabCulture						
	Before Aug		After Aug					Before Aug		After Aug				
	Base	CoT	Mango	E-Asrt	E-Path	N-Asrt	N-Path	Base	CoT	Mango	E-Asrt	E-Path	N-Asrt	N-Path
Qwen2.5-0.5B	40.8	37.3	40.9	**41.6**	38.5	40.6	39.0	34.3	34.3	34.3	34.4	34.3	34.2	**34.5**
Qwen2.5-1.5B	38.8	33.7	44.7	**44.8**	42.1	40.5	38.6	**38.8**	35.2	35.5	37.5	36.2	37.2	36.6
Qwen2.5-3B	52.5	53.1	**54.8**	53.9	52.7	51.9	51.0	37.5	36.4	37.9	36.9	36.7	**39.3**	37.8
Qwen2.5-7B	58.5	56.6	59.4	59.2	**59.8**	59.3	59.5	**47.9**	35.4	38.7	40.5	39.2	37.5	38.7
Gemma2-2B	34.5	35.9	38.1	**40.0**	38.6	39.1	36.6	34.3	34.3	34.3	34.3	34.3	34.3	34.3
Gemma2-9B	65.4	42.2	66.6	65.6	**67.5**	65.7	67.4	35.6	36.6	35.5	37.3	35.3	**38.1**	34.8
Llama3.2-1B	56.7	56.0	54.8	55.6	55.0	55.7	**57.1**	33.9	**34.2**	33.8	34.1	34.1	33.8	34.1
Llama3.1-8B	32.7	**35.6**	32.7	32.7	32.7	32.7	32.7	34.3	34.3	34.2	34.3	34.2	34.0	34.2
Llama3.2-3B	47.1	44.1	46.9	47.4	46.7	**48.3**	47.3	34.3	34.1	**34.4**	**34.4**	34.3	34.3	34.3
Gemma2-9B-IT	57.4	56.7	57.1	56.2	57.8	**58.5**	58.1	34.3	**34.8**	34.4	34.3	34.3	34.3	34.3
Qwen2.5-7B-IT	66.1	**67.2**	64.6	65.5	65.6	65.4	65.7	**39.5**	34.3	35.6	36.8	38.0	35.8	36.2
Llama3.1-8B-IT	53.4	48.9	55.2	55.6	53.6	55.7	**55.9**	37.9	34.2	34.3	34.3	**34.4**	34.3	34.4
Avg Δ	NA	-3	1	**1.2**	0.6	0.8	0.4	NA	-2.1	-1.7	-1.2	-1.5	-1.3	-1.5

Table 14: Accuracy comparison on MCQA across different methods with native prompts. **E-Asrt**: CCKG English Assertions, **E-Path**: CCKG English Paths, **N-Asrt**: CCKG Native-language (in Arabic or Indonesian) Assertions, **N-Path**: CCKG Native-language (in Arabic or Indonesian) Paths. "Avg Δ" denotes the average improvement over the baseline. Best results per model are bolded.

Models	Before Aug	+Mango	+CCKG Methods			
			E-Asrt	E-Path	N-Asrt	N-Path
ArabCulture - Avg Sentence Similarity						
Gemma2-9B-IT	31.6	32.1	35.0	32.7	**35.4**	33.2
Qwen2.5-7B-IT	40.2	42.3	43.9	41.3	**44.7**	41.9
Llama3.1-8B-IT	27.3	**32.4**	30.3	28.0	28.5	27.2
Avg	33.0	35.6	36.4	34.0	**36.2**	34.1
IndoCulture - Avg Sentence Similarity						
Gemma2-9B-IT	31.3	33.2	34.0	34.0	34.6	**35.0**
Qwen2.5-7B-IT	39.8	41.1	**42.0**	41.7	41.5	34.4
Llama3.1-8B-IT	31.1	33.6	33.8	34.4	35.0	**41.6**
Avg	34.1	35.9	36.6	36.7	**37.1**	37.0
ArabCulture - Avg BERT Score F1						
Gemma2-9B-IT	68.1	68.2	68.4	68.4	**68.6**	68.3
Qwen2.5-7B-IT	68.3	68.7	68.8	67.7	**69.5**	68.7
Llama3.1-8B-IT	65.0	**66.0**	65.8	65.0	65.6	65.3
Avg	67.1	67.6	67.7	67.0	**67.9**	67.4
IndoCulture - Avg BERT Score F1						
Gemma2-9B-IT	70.5	70.8	71.1	71.1	71.3	**71.4**
Qwen2.5-7B-IT	71.9	71.9	71.9	71.9	**72.0**	70.1
Llama3.1-8B-IT	69.6	70.0	70.2	70.0	70.2	**72.1**
Avg	70.7	70.9	71.1	71.0	**71.2**	**71.2**

Table 15: Performances for sentence similarity and BERT score F1 with native prompts. **E-Asrt**: CCKG English Assertions, **E-Path**: CCKG English Paths, **N-Asrt**: CCKG Native-language (in Arabic or Indonesian) Assertions, **N-Path**: CCKG Native-language (in Arabic or Indonesian) Paths. Best results per row are bolded.

spectively. In all cases, the `story` and `country` variables are replaced with their corresponding values. The same prompts were also provided to human annotators for the human evaluation of stories generated in both English and native languages.

E.4 Prompts for MCQA and SENTENCE COMPLETION

For MCQA and sentence completion tasks, we use the original benchmark prompts.

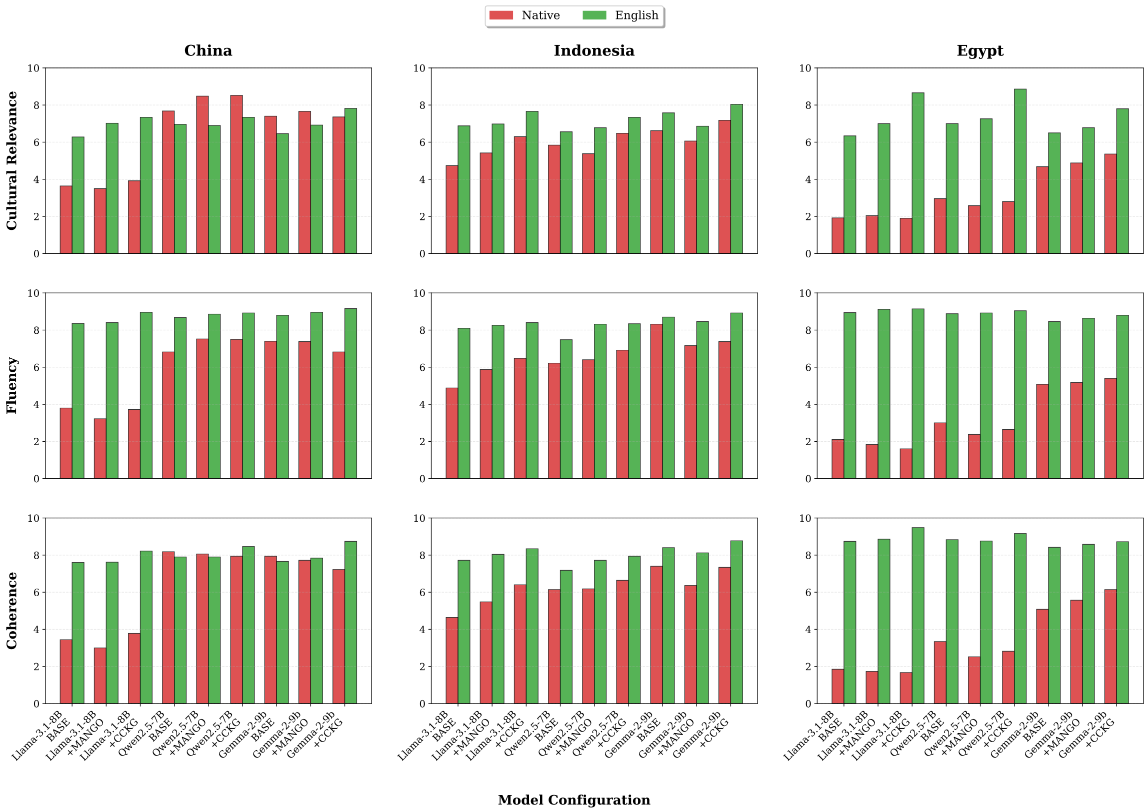

Figure 4: Average scores per evaluation metric for each country and model.

System role

You are a cultural commonsense knowledge extraction assistant

User role

Generate as many if-then cultural commonsense knowledge statements as possible for {location} related to {sub_topic}. Each statement should follow this structure: "If [action], then [knowledge]." The knowledge should describe what action is likely to follow or precede the initial action, reflecting cultural habits and practices associated with {sub_topic} in {location}.

Let's think step by step to generate the chain of action between the action and the knowledge. The knowledge must be classified into one of the following relation types:

- xEffect: What effects does the action have on x?
- xNext: What would x likely want to do after the action?
- xNeed: What does x need to do before the action?
- oNext: What would others likely want to do after the action?
- oEffect: What effects does the action have on others?

In the relation types above, 'x' represents the agent or primary person performing the action, while 'o' refers to others who interact with or are affected by the action of 'x'.

Format your response strictly as a JSON array of objects, following this precise format without any additional text or explanation:

```
[
    {{
        "action": "action",
        "knowledge": "knowledge",
        "relation_type": "relation type",
        "result": "Complete sentence in {language}, using If and Then"
    }}, ...
]
```

Ensure that the "action," "knowledge," and the entire "result" sentences are generated in {language}, while the "relation_type" should be one of the specified options above.

Figure 5: Prompt used in the initial generation step of CCKG. The variables `sub_topic`, `location`, and `language` are replaced with their corresponding values (sub-topic, country, and language), expressed in English when the KB is generated in English and in the native language otherwise.

System role

You are a cultural commonsense knowledge extraction assistant

User role

Your task is to generate as many culturally sensitive commonsense knowledge events as possible for {location}, focusing on {sub_topic}, based on an initial event comprising an action and a knowledge. Each event must follow the structure: "If [action], then [knowledge]," and should accurately reflect the cultural habits, practices, traditions, and customs related to {sub_topic} in {location}.

Initial event: {initial_event}
initial action: {init_action}
initial knowledge: {init_knowledge}

Two types of generations are requested:

1. Next steps generation
- Start from the entire initial event
- use the initial knowledge as the starting point to generate all new knowledges
- Ensure that each new knowledge logically derives from the initial event as a whole, taking into account both the initial action and the initial knowledge.
- Each new knowledge must logically align with the initial action and initial knowledge, showing their natural progression. Specifically, focus on the relationship types listed below.
- Format the new knowledge as conditional statements, starting with: “if {init_knowledge}, then…”
- Each new knowledge must be classified into one of these relation types:
  *xNext: What would x likely want to do after the action?
  *oNext: What would others likely want to do after the action?
  In the relation types above, 'x' represents the agent or primary person performing the action, while 'o' refers to others who interact with or are affected by the action of 'x'.

2. Intermediate steps generation
- Start from the initial action from the initial event.
- Create a chain of new knowledge (A_1, A_2, .., A_i) leading to the initial knowledge.
- Use a stepwise approach:
  *Start with: if {init_action}, then A_1
  *then: “if A_1, then A_2”
  *generate new knowledge (A_i) that logically and culturally connects to the previous step. Continue iterating until you arrive at the final step, where the generated knowledge aligns with the initial knowledge.
  *Please respect the stepwise approach
- Each step must be classified into one of the relation types listed above.
- The chain should logically explain how someone would arrive at the initial knowledge in a culturally sensitive way.

Other requirements:

1. Cultural relevance
   Ensure all steps (new knowledge) are deeply rooted in and reflective of the culture in {location}, while being closely associated with the sub-topic {sub_topic} in {location}.

2. Output format and output language
   Entire event sentence, action, knowlege in both next steps and intermediate steps must be written in {language} language. Your response must be formatted strictly as a JSON array of objects without any additional text or explanation and organized into two categories: next steps and intermediate steps:

```
[
    {{
        "intermediate_steps": [
            {{ "action": " action", "knowledge": "knowledge", "relation_type": "relation", "event": "If action, then knowledge" }}
                ],
        "next_steps": [
            {{ "action": " action", "knowledge": "knowledge", "relation_type": "relation", "event": "If action, then knowledge" }}
                ]
    }}
]
```

Figure 6: Prompt used in the iterative expansion step of CCKG. The variables `sub_topic`, `location`, and `language` are replaced with the corresponding values, expressed in English when the KB is generated in English and in the native language otherwise. The variables `initial_event`, `init_action`, and `init_knowledge` are instantiated from the initial assertion: for an assertion “if *action_1*, then *action_2*,” `initial_event` is the full assertion, `init_action` is *action_1*, and `init_knowledge` is *action_2*.

Coherency Evaluation Prompt

Analyze the coherence of the following story by evaluating its logical flow, structural clarity, and consistency in events and character actions.

The story should be written in {LANGUAGE}.

Story:

{story}

Rate the story on a scale from 1 to 10 using the following guidelines:

- 10: Completely coherent, with a logical flow, well-structured events, and clear cause-effect relationships.
- 9: Very strong coherence, with only a minor issue that doesn't significantly impact the story's logic.
- 8: Mostly coherent, with some small inconsistencies or minor logical gaps.
- 7: Generally makes sense, but has a few unclear transitions or minor plot inconsistencies.
- 6: Somewhat coherent but contains multiple small logical flaws or confusing elements.
- 5: Moderately coherent; the main ideas are understandable, but there are noticeable inconsistencies.
- 4: Limited coherence, with frequent plot holes, unclear transitions, or inconsistent character actions.
- 3: Poor coherence; the story is often illogical or difficult to follow.
- 2: Very poor coherence. The plot is highly inconsistent or nonsensical.
- 1: No coherence. The story is completely disjointed and does not make logical sense.

Provide only the numerical score

Output format (no extra text):

Coherence Score: [1-10]

Figure 7: Prompt used to evaluate coherence in the story generation task.

Fluency Evaluation Prompt

Assess the fluency of the following story in terms of grammatical accuracy, sentence structure, vocabulary choice, and readability.

The story should be written in {LANGUAGE}.

Story:

{story}

Rate the story on a scale from 1 to 10 using the following guidelines:

- 10: Exceptionally fluent, with no grammatical errors, awkward phrasing, or unnecessary repetition. Feels entirely natural.
- 9: Very fluent, with only one or two minor errors that do not affect readability.
- 8: Strong fluency, but with a few small grammar issues or slightly awkward phrasing.
- 7: Mostly fluent, but contains occasional awkward sentences or minor grammatical errors.
- 6: Somewhat fluent but has noticeable grammar issues, unnatural phrasing, or occasional repetition.
- 5: Moderately fluent; there are multiple errors or awkward sentences that slightly hinder readability.
- 4: Limited fluency, with several grammar mistakes and clunky sentence structure that affects understanding.
- 3: Poor fluency, with frequent grammatical errors, awkward wording, and significant readability issues.
- 2: Very poor fluency. The text is difficult to read, with major grammar mistakes and repetitive phrasing.
- 1: Completely unreadable due to excessive errors and poor language structure.

Provide only the numerical score.

Output format (no extra text):

Fluency Score: [1-10]

Figure 8: Prompt used to evaluate fluency in the story generation task.

Cultural relevance Evaluation Prompt

Evaluate the cultural relevance of the following story based on how accurately it represents the traditions, customs, values, and social norms of {COUNTRY}.

Story:

{story}

Rate the story on a scale from 1 to 10 using the following guidelines:

- 10: Perfectly represents the cultural context with deep authenticity. All customs, language, behaviors, and references are accurate.
- 9: Almost perfect representation, with minor details that might be slightly inaccurate but do not misrepresent the culture.
- 8: Strong cultural accuracy but with a few notable inconsistencies or generalizations.
- 7: Mostly aligns with the culture, but has some incorrect elements that a native reader would recognize as imprecise.
- 6: Somewhat culturally relevant, but there are multiple noticeable inaccuracies or stereotypical representations.
- 5: Moderately relevant but contains a mix of accurate and inaccurate cultural elements. Some details feel generic.
- 4: Limited cultural accuracy. Several key aspects of the culture are misrepresented or omitted.
- 3: Poor cultural alignment. The story contains serious inaccuracies or misuses cultural elements.
- 2: Very poor representation. The culture is barely recognizable or is misrepresented in a way that may be misleading.
- 1: No cultural relevance. The story does not reflect the intended culture at all.

Provide only the numerical score.

Output format (no extra text):

Cultural Relevance Score: [1-10]

Figure 9: Prompt used to evaluate cultural relevancy in the story generation task.

Baseline prompt

Write a 5-sentence narrative story about {SUBTOPIC} set in {COUNTRY}. The story should be written in {LANGUAGE}.

Do not output anything else except the 5-sentence in {LANGUAGE}

Augmentation prompt

Write a 5-sentence narrative story about {SUBTOPIC} set in {COUNTRY}. The story should be written in {LANGUAGE}.

You may consider this additional cultural information of the country: {assertions}

Do not output anything else except the 5-sentence in {LANGUAGE}

Figure 10: Story generation prompts.