CultureTalk: Towards Multi-turn Cultural Competence in Large Language Models

Minzhi Li[*σν] **Jiayu Han**[σ] **Oleg Poliannikov**[σ]
Vipul Agarwal[σ] **Dezhi Hong**[σ] **Mingxian Wang**[σ]
[σ]Amazon [ν]National University of Singapore

Abstract

As conversational AI systems increasingly engage users through multi-turn dialogue across diverse cultural contexts, cross-cultural competence, the ability to recognize, respect, and adapt to cultural differences, is essential for effective human-AI interaction. Existing evaluations largely rely on single-turn settings, overlooking the complexities of real-world, multi-turn conversations where cultural cues often emerge gradually. To address this gap, we introduce *CultureTalk*[1], a benchmark for evaluating *multi-turn cultural reasoning* in conversational AI. The dataset contains 1,741 culturally grounded dialogues derived from human-verified cultural knowledge resources and evaluates models' ability to accumulate cultural context across turns and remain robust to topic shifts. Our evaluation reveals substantial performance gaps between single- and multi-turn settings, highlighting current models' limitations in sustained culturally grounded context understanding. We further show that fine-tuning on augmented single-turn conversational data and inference scaling significantly improve multi-turn cultural reasoning, yielding a 64.7% relative gain in cultural knowledge entailment with a fine-tuned Qwen3-8B model.

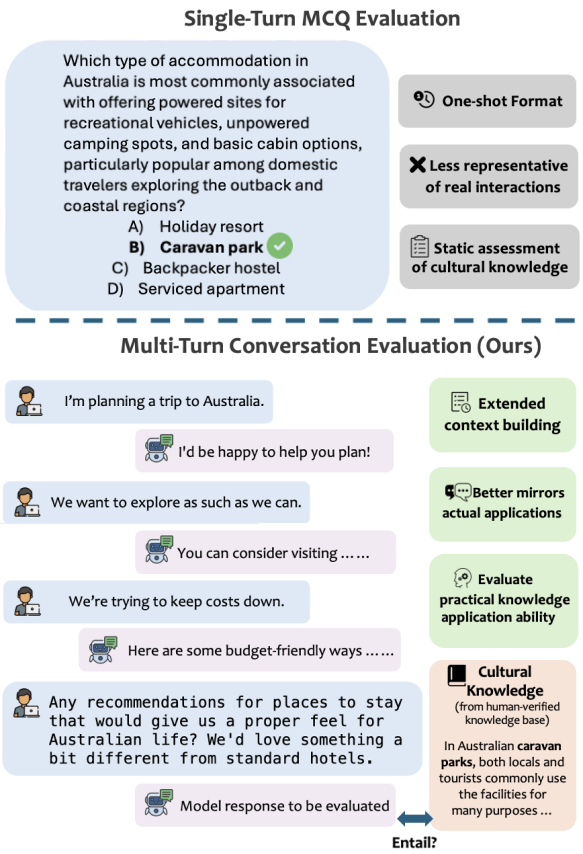

Figure 1: Comparison of single-turn and multi-turn cultural capability assessment. The final turn in the figure is not a real system response.

1 Introduction

The widespread deployment of language models in user-facing applications, from content recommendations (Bao et al., 2023) to customer support (Pandya and Holia, 2023) and search systems (Xiong et al., 2024), has made cultural competence increasingly essential. However, due to uneven cultural coverage in training data, models often struggle to reflect diverse cultural perspectives (Durmus et al., 2024; Ryan et al., 2024), sometimes reinforcing stereotypes (Jha et al., 2023) or producing culturally insensitive responses (Cao et al., 2022). As these systems become more embedded in global and multicultural contexts, ensuring their culturally appropriate behavior has become a core challenge in language technology development.

Growing efforts have focused on curating cultural knowledge resources, with most evaluations of LLM cultural competence relying on single-turn question answering (Koto et al., 2024; Li et al., 2024c; Shi et al., 2024). However, this approach does not capture the complexities of real-life human-AI interactions (Pawar et al., 2024). In practice, cultural understanding unfolds grad-

[*]Work done at Amazon.
[1]CultureTalk Dataset

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 2048–2065
October 24-29, 2026 ©2026 Association for Computational Linguistics

ually through sustained conversation (Figure 1), where users reveal contextual cues across multiple turns and models must retain earlier cultural signals while remaining robust to topic shifts. This highlights the need for benchmarks that evaluate *multi-turn cultural reasoning* under realistic conversational dynamics rather than isolated question–answer scenarios.

In this work, we design a multi-turn data generation pipeline that transforms validated cultural knowledge into dynamic conversational exchanges via turn-by-turn simulation. To better reflect real-world dialogue complexity, we incorporate strategic topic shifts and apply rigorous post-generation filtering to ensure quality. We evaluate cultural response quality using cultural knowledge entailment against an external human-verified cultural knowledge base. Our contributions are as follows:

1. We develop *CultureTalk*, **a benchmark for evaluating multi-turn cultural reasoning** in conversational AI. The dataset contains 1,741 conversations grounded in human-validated cultural knowledge and captures conversational dynamics such as context accumulation, topic shifts, and conversational grounding of cultural knowledge.

2. Through zero-shot evaluation across eight language models of varying sizes and architectures, we reveal a **consistent degradation in cultural reasoning performance in multi-turn settings** compared to single-turn interactions, and identify common failure patterns such as context forgetting and over-generalized responses.

3. We further explore supervised fine-tuning with different conversation data formats and find that **training with augmented single-turn question answering data yields the strongest gains**. Inference-time scaling that elicits reasoning before responding further improves the model performance. Collectively, these approaches help Qwen3-8B achieve up to 60.6% performance improvement compared to the base model.

2 Related Work

Cultural Datasets. Researchers have developed various datasets in recent years to advance cultural understanding in language models. The Cultural NLI dataset (Huang and Yang, 2023) examines how cultural context influences natural language inference between US and Indian annotators. GlobalOpinion QA (Durmus et al., 2024) establishes a quantitative framework for assessing LLMs' representation of diverse global perspectives through cross-national survey responses, highlighting inherent biases towards Western perspectives. CultureLLM (Li et al., 2024a) leverages the World Value Survey with semantic data augmentation to generate culturally diverse training data across nine cultures, achieving substantial improvement over leading models. *CultureBank* (Shi et al., 2024) presents a novel methodology by extracting cultural information from social media platforms and curates descriptive cultural knowledge. The Blend dataset (Myung et al., 2025) assesses everyday cultural knowledge across 16 countries in 13 languages through single-turn MCQs.

However, most of these datasets are restricted to single-turn formats, limiting their ability to capture the contextual dependencies and evolving pragmatics of real-world interactions. Emerging multi-turn cultural benchmarks remain scarce and narrow in scope. For multi-turn data development, ReNoVi (Zhan et al., 2024) constructs multi-turn dialogues focused on understanding and remediating norm violations, combining both human-authored and synthetic conversations. However, it focuses solely on the Chinese culture. Moreover, CulturePark (Li et al., 2024b) introduces an LLM-powered multi-agent framework that simulates cross-cultural human-human communication through agent role-playing. In contrast, our work targets the human-AI interaction. Compared to prior cultural datasets, we emphasize the evaluation of AI response in culturally grounded conversations across distinct cultures and dynamic conversational contexts beyond remediation scenarios.

Multi-turn Data Curation. Developing multi-turn dialogue datasets has become an essential strategy for mitigating the scarcity of naturally occurring conversational data (Li et al., 2025). One line of work relies on *one-shot expansion*, where a single-turn conversation is directly expanded into a multi-turn dialogue through a single prompting step. For example, the SMILE framework (Qiu et al., 2024) prompts ChatGPT to elaborate on single-turn mental health dialogues, producing the SMILECHAT dataset. However, the generated conversations often exhibit limited diversity and strong dependence on the specific prompt template. In

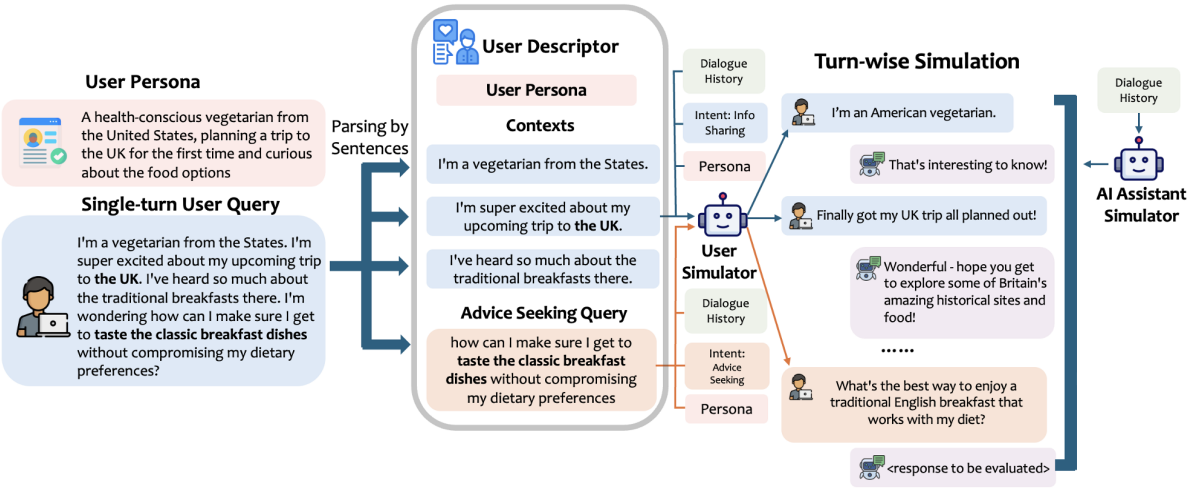

Figure 2: Multi-turn conversation data generation pipeline. We start from the user persona and single-turn user query description, parse it to information sharing queries and advice seeking query. For each turn, we generate natural user utterance based on user persona, user intent, information in the current turn and dialogue history. We generate AI response based on previous conversation context.

contrast, *iterative role-play generation* frameworks simulate interactions between conversational roles turn by turn, better capturing dialogue dynamics. For example, the teaching simulation work by (Hu et al., 2025) employs an iterative process of plan generation, simulated interaction, and reflective refinement to produce high-quality teacher–student dialogues. UltraChat (Ding et al., 2023) leverages self-chat to synthesize large-scale instructional conversations without human queries. Building on these advances, we construct *CultureTalk* through **user–AI role-play**, generating natural and diverse multi-turn dialogues grounded in cultural knowledge. Unlike prior methods, our framework designs user turns to carry culturally relevant contexts and elicit responses that demonstrate the model's awareness of the underlying cultural knowledge.

3 CultureTalk

In this section, we discuss the simulation of multi-turn conversations grounded in cultural knowledge (3.1), the introduction of topic-shift turns in dialogues (3.2), and subsequent filtering and validation (3.3) to ensure data quality.

3.1 Multi-Turn Data Simulation

To develop a multi-turn conversation that is grounded in cultural knowledge, we start with a user persona and a single-turn query in existing cultural dataset, perform context parsing, followed by turn-wise simulation (see Figure 2).

Seed Data Selection. We select *CultureBank* (Shi et al., 2024) as our primary seed dataset, given its comprehensive collection of 23k human-validated cultural norms and descriptive knowledge. Unlike MCQs, CultureBank entries capture richer nuance and contextual detail, and naturally support consulting-style scenarios in which cultural knowledge is applied, making them well-suited for multi-turn conversation generation. The dataset further provides broad coverage across 36 topics spanning approximately 1,850 cultural groups. Given the computational cost of applying the full CultureTalk pipeline to all CultureBank cultures, we select three English-speaking cultures (American, Australian, and British) alongside Japanese culture, which provides a distinct non-Western perspective.

Context Parsing. Each consulting scenario offered in *CultureBank* contains a user persona (*eval_persona*) such as *"a health-conscious vegetarian"* and context-rich user query (*eval_question*) such as *"I'm excited about upcoming trip to the UK ... how can I get to taste the classic breakfast dishes ..."* (see Figure 2). To construct user descriptor D_i for each dialogue, we select the *eval_persona* field as $persona_i$ and parse the *eval_question* field by sentences to get contexts and advice seeking query. We extract the advice seeking $query_i$ by looking for sentence ending with question mark or containing key words such as *"I'd love to know"/"I'm wondering how"*. The rest non-query sentences explaining the contexts will be $context_{i1}$, $context_{i2}$, ..., $context_{iN_i}$ where N_i is the total number of

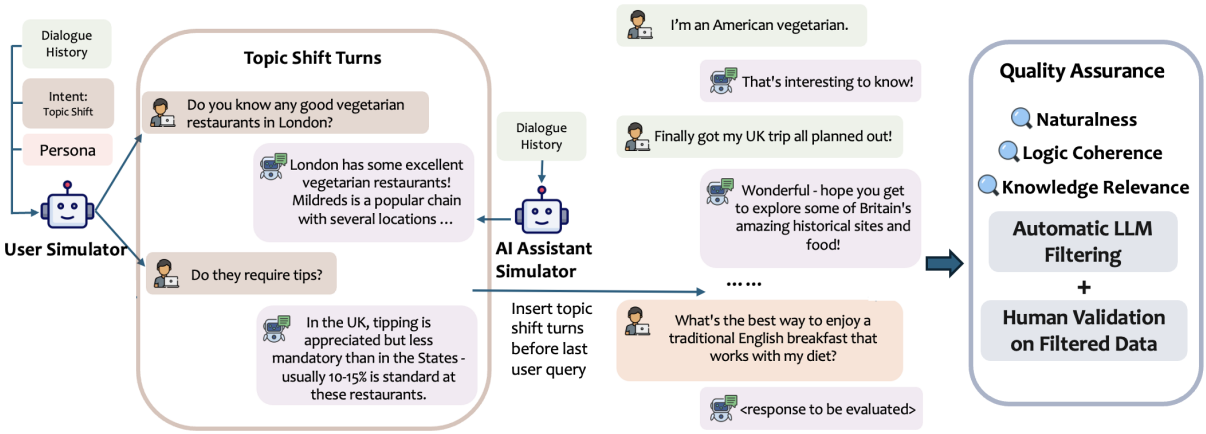

Figure 3: Multi-turn conversation data generation pipeline with topic shift turns. We build upon generated multi-turn conversation and introduce additional two turns that goes along with the dialog history but introduces topic change. We then apply automatic LLM filtering and human validation to ensure quality of conversations developed.

parsed contexts in the single-turn user query.

Algorithm 1 Turn-wise User and AI Assistant Simulation for Cultural Conversations

Require: Conversation descriptor $D_i = \{persona_i, context_{i1}, ..., context_{ik}, query_i\}$

Define the dialogue history up to turn T_j as:
$H_{ij} = \{utterance_user_{i1},\ utterance_ai_{i1}, \ldots, utterance_user_{i,j-1},\ utterance_ai_{i,j-1}\}$
for each turn T_j in conversation **do**
  if $1 \leq j \leq N_i$ **then** ▷ Information sharing turns
    Generate user utterance:
    $P(utterance_user_{ij} \mid intent = info_sharing, persona_i, H_{ij}, context_{ij})$
  else if $j = N_i + 1$ **then** ▷ Advice seeking turn
    Generate user utterance:
    $P(utterance_user_{ij} \mid intent = advice_seeking, persona_i, H_{ij}, query_i)$
  end if
  Generate AI assistant response:
  $P(utterance_ai_{ij} \mid H_{ij}, utterance_user_{ij})$
end for

Turn-wise Simulation. As illustrated in Algorithm 1, we adopt turn-wise user simulation grounded on context descriptor for the specific conversation $D_i = \{persona_i, context_{i1}, context_{i2}, \ldots, context_{iN_i}, query_i\}$. This improves contextual faithfulness of the generated conversation. We consider two utterance types for each turn T_j in a multi-turn conversation.

The first type is an information-sharing turn where $1 \leq j \leq N_i$. In these turns, we prompt Claude-Sonnet-3.5-v2 to generate a user utterance $utterance_user_{ij}$ that focuses on information sharing, conditioned on user's persona, prior conversation history, and the context information in the current turn. For the final turn of the conversation, we simulate an advice-seeking utterance to elicit culturally grounded responses from the model. We prompt the language model to generate a user utterance with an advice-seeking intent, based on the preceding conversation. To simulate AI Assistant responses to all history turns within the conversational context, we input the dialogue history and prompt the language model to produce a response for the current turn.

3.2 Topic Shift Introduction

In real-life conversations, topic shifts naturally occur as speakers change subjects. Such shifts pose challenges to conversational AI systems (Li et al., 2025), which must maintain coherence and relevance across varying subjects while preserving contextual understanding. Building on the multi-turn conversation data generated in Section 3.1, we introduce two strategic topic-shift turns before the final knowledge query from the user. This design helps to stress-test the model's ability to retain focus on prior contexts relevant to the final cultural knowledge query despite the subject changes. Placing topic shift turns immediately before the final cultural knowledge query increases the contextual distance from the relevant information, thereby challenging the model to navigate and retrieve the relevant conversation context.

As shown in Figure 3, we prompt the language

model to generate user utterances that are consistent with the dialogue history and with a topic shift over two turns. The AI simulator then generates corresponding responses to these utterances.

3.3 Quality Filtering and Validation

To ensure cultural authenticity and overall quality, we design a three-tier filtering framework for the generated multi-turn dialogues. The framework consists of three criteria: (1) **Cultural Knowledge Relevance Check**, which evaluates whether the last user knowledge query appropriately relates to the intended cultural knowledge in *CultureBank* (e.g., "no tipping in Japan"), ensuring a response that incorporates the cultural knowledge needed for contextual appropriateness and cultural sensitivity; (2) **Naturalness Check**, which evaluates conversational flow and human-likeness of the dialogue by assessing turn-taking dynamics, use of discourse markers, natural language patterns, contextual coherence, formality level, and realistic dialogue progression; and (3) **Logic Coherence Check**, which verifies the logical consistency and coherence of the conversation by evaluating internal consistency, temporal coherence, cause-effect relationships, information flow, response relevance, and contextual appropriateness of responses. Each criterion produces a binary (yes/no) decision, and only dialogues passing all three criteria are retained in the final dataset. We use Deepseek-R1 as a judge for post-generation annotation based on these criteria and retain dialogues that pass all three checks (see Appendix B.3). Over 86% of conversations across all four cultures passed all three filters, indicating the robustness of our multi-turn data curation pipeline. The final evaluation set comprises 1,741 conversations across Australian, American, British, and Japanese cultures (see Table 3).

We perform stratified random sampling of 200 filtered conversations across cultures. Four authors independently rate each sample along three dimensions: cultural knowledge relevance, naturalness, logic coherence (Likert scale of 1 to 5). The average human ratings for all cultures (naturalness: 4.80, logic coherence: 4.91, relevance: 4.62) suggest strong overall quality of conversations generated using our simulation method.

4 Experiments

In this section, we explain our zero-shot evaluation experiment setup and discuss evaluation results in multi-turn with and without topic shift.

4.1 Experiment Setup

Evaluated LLMs. To evaluate cultural competence across model architectures and scales, we benchmark both proprietary and open-source language models. Proprietary models include Claude-3.5 Haiku and Sonnet, representing different capacity levels. Open-source models include Qwen3 (8B and 32B), Mistral-7B, and Mixtral-8×7B. We also include two reasoning-focused models: GPT-OSS-20B and DeepSeek-r1. This model suite enables analysis of scale, architecture, and reasoning effects on cultural understanding and comparison between proprietary and open-source systems.

Data. We evaluate model performance across four cultures. For each culture, we report results on all cultural knowledge, high- and mid-frequency clusters (> 20 supporting samples on social media), and low-frequency clusters ($5-20$ supporting samples on social media) following *CultureBank*.

Metric. We adopt the entailment score (Shi et al., 2024) as our evaluation metric. Cultural correctness is measured by checking whether a model response entails the corresponding *human-verified cultural knowledge* curated in *CultureBank*, rather than relying on the model's inherent knowledge. An LLM-as-a-judge (Claude-Sonnet-3.5-v1) is provided with the model response and the reference knowledge and predicts whether the response entails the knowledge. Following Shi et al. (2024), this entailment-based metric serves as a reliable proxy for cultural appropriateness. As our evaluation is performed only on the final turn response, the evaluation setup is methodologically consistent with the setup used in CultureBank.

We validate the reliability of this metric through a human–model agreement study (Appendix F), where the LLM judge achieves 77% agreement with human annotations (Cohen's $\kappa = 0.53$).

4.2 LLMs' Cultural Competence in Conversational Settings

Multi-turn interaction poses greater challenges. As shown in Table 1, multi-turn conversations consistently reduce models' cultural knowledge application compared to single-turn settings (performance drops across almost all models, knowledge cluster types and cultures in the second and third vertical blocks). We identify two primary errors. The first is *knowledge failure* (71.7%), where the

Culture	Model	CultureTalk-SingleTurn			CultureTalk-MultiTurn			CultureTalk-TopicShift		
		All	High/Mid	Low	All	High/Mid	Low	All	High/Mid	Low
US	Claude-Haiku	62.0	64.5	61.0	50.5 ↓	54.1 ↓	49.1 ↓	48.7 ⇓	51.2 ⇓	47.7 ⇓
	Claude-Sonnet	65.4	66.9	64.8	53.8 ↓	59.3 ↓	51.6 ↓	53.0 ⇓	57.6 ⇓	51.1 ⇓
	Qwen3-8B	48.2	52.3	46.6	36.4 ⇓	40.1 ⇓	34.9 ⇓	38.4 ↓	41.3 ↓	37.2 ↓
	Qwen3-32B	55.6	57.0	55.0	51.8 ↓	56.4 ↓	50.0 ↓	49.3 ⇓	56.4 ↓	46.6 ⇓
	Mistral-7B	54.3	57.0	53.2	41.5 ↓	46.5 ↓	39.5 ↓	41.0 ⇓	46.5 ↓	38.8 ⇓
	Mixtral	55.1	56.4	54.6	40.3 ↓	41.9 ↓	39.7 ↓	39.0 ⇓	39.5 ⇓	38.8 ⇓
	Deepseek-r1	**70.8**	**75.0**	**69.2**	**59.2** ↓	**66.3** ↓	**56.4** ↓	54.3 ⇓	57.0 ⇓	53.2 ⇓
	GPT-OSS-20B	56.6	57.6	56.2	55.4 ⇓	52.9 ⇓	**56.4** ⇓	**57.7** ↓	**58.1** ↓	**57.5** ↓
UK	Claude-Haiku	49.5	55.5	48.1	41.6 ↓	42.7 ↓	41.4 ↓	39.8 ⇓	40.0 ⇓	39.7 ⇓
	Claude-Sonnet	56.1	61.8	54.8	43.5 ⇓	42.7 ⇓	43.7 ⇓	43.9 ↓	43.6 ↓	43.9 ↓
	Qwen3-8B	42.0	47.7	40.7	33.8 ⇓	38.3 ⇓	32.8 ⇓	34.6 ↓	39.3 ↓	33.6 ↓
	Qwen3-32B	50.0	57.9	48.2	44.9 ↓	52.3 ↓	43.2 ⇓	45.1 ⇓	51.4 ⇓	43.6 ↓
	Mistral-7B	46.8	58.9	44.1	36.2 ⇓	43.9 ↓	34.4 ⇓	36.7 ↓	39.2 ⇓	36.1 ↓
	Mixtral	49.3	50.5	49.1	37.0 ↓	38.3 ⇓	36.7 ↓	35.0 ⇓	39.3 ↓	34.0 ⇓
	Deepseek-r1	**60.4**	**71.0**	**58.0**	**53.2** ↓	**60.7** ↓	**51.6** ↓	**51.7** ⇓	**58.9** ⇓	**50.1** ⇓
	GPT-OSS-20B	57.7	63.6	56.4	51.4 ↓	53.3 ⇓	50.9 ↓	50.0⇓	54.2 ↓	49.1 ⇓
AU	Claude-Haiku	56.1	57.6	55.7	44.2 ↓	37.6 ⇓	45.9 ↓	42.7 ⇓	40.0 ↓	43.4 ⇓
	Claude-Sonnet	59.5	57.6	59.9	48.1 ⇓	42.4 ⇓	49.5 ↓	48.3 ↓	48.2 ↓	48.3 ⇓
	Qwen3-8B	38.6	40.0	38.2	38.8	37.6 ↓	39.1	36.4 ⇓	34.1 ⇓	37.0 ⇓
	Qwen3-32B	47.3	49.4	46.8	44.2 ↓	42.4 ↓	44.6 ↓	43.9 ⇓	42.4 ↓	44.3 ⇓
	Mistral-7B	48.3	55.3	46.5	41.5 ↓	41.6 ↓	41.1 ↓	36.4 ⇓	32.9 ⇓	37.3 ⇓
	Mixtral	46.6	51.8	45.3	35.4 ⇓	32.9 ⇓	36.1 ↓	35.7 ↓	37.6 ↓	35.2 ⇓
	Deepseek-r1	**62.4**	**68.2**	**60.9**	**52.4** ↓	50.6 ↓	**52.9** ↓	**51.0** ⇓	50.6 ↓	**51.1** ⇓
	GPT-OSS-20B	53.9	58.8	52.6	49.5 ⇓	**58.8**	47.1 ⇓	49.8 ↓	**55.3** ⇓	48.3 ↓
JP	Claude-Haiku	58.6	60.0	58.2	54.9 ↓	**62.9** ↓	52.0 ↓	54.1 ⇓	62.9 ↓	51.0 ⇓
	Claude-Sonnet	62.4	62.9	62.2	54.1 ↓	**62.9**	51.0 ↓	48.9 ⇓	54.3 ⇓	46.9 ⇓
	Qwen3-8B	47.4	62.9	41.8	42.1 ⇓	45.7 ↓	40.8 ⇓	45.1 ↓	45.7 ↓	44.9 ↓
	Qwen3-32B	48.1	57.1	44.9	50.4	65.7	44.9	50.3	62.9	45.9
	Mistral-7B	54.1	**65.7**	50.0	45.9 ⇓	51.4 ⇓	43.9 ⇓	48.1 ↓	54.3 ↓	45.9 ↓
	Mixtral	55.6	62.9	53.1	39.8 ↓	48.6 ⇓	36.7 ↓	39.8 ↓	54.3 ↓	34.7 ⇓
	Deepseek-r1	**65.4**	57.1	**68.4**	54.9 ↓	**62.9** ↓	52.0 ↓	52.6 ⇓	57.1 ⇓	51.0 ⇓
	GPT-OSS-20B	60.1	62.9	59.2	**56.4** ↓	60.0 ⇓	**55.1** ↓	**55.6** ⇓	**65.7**	**52.0** ⇓

Table 1: Performance comparison across different cultures, models, and conversational settings. We find that multi-turn and topic shifts consistently and notably reduce LLM accuracy. We mark the change from the single-turn performance as ↓(lower performance compared to single-turn) or ⇓(lowest performance among single-turn, multi-turn and topic shift scenarios. High/mid/low marks culture knowledge cluster size.).

model lacks the required knowledge in both single- and multi-turn settings, indicating a knowledge gap. The second is *specificity failure* (28.3%), where models that produce grounded responses in single-turn settings revert to overly general answers in multi-turn interactions, omitting key details and cultural nuances (see qualitative examples in G).

Model scale and reasoning enhance cultural competence. Model scaling laws also extend to cultural competence in conversational settings: larger models such as Claude-3.5-Sonnet and Qwen3-32B outperform smaller counterparts. Reasoning models (Deepseek-r1 and GPT-OSS-20B) generally demonstrate the highest performance in both single- and multi-turn conversations.

Topic shifts further challenge certain models. Topic shift injections introduce additional challenges for certain models (Claude-3.5-Haiku and Deepseek-r1), with overall performance further declining for all four cultures, compared to multi-turn scenarios without topic shift.

4.3 Enhance LLM's Cultural Competence in Conversations

While retrieval-augmented generation (Lewis et al., 2021) grounds responses in external knowledge, it introduces retrieval latency at inference time. We therefore investigate supervised fine-tuning (SFT) as an alternative, enabling LLMs to internalize and retain new knowledge through direct training.

Culture	Setting	Base Model	SFT - Knowledge	SFT - One SQA	SFT - Multi SQA	SFT - One Culture	SFT - Multi MQA	SFT - Sequential	Inference-Scaling
AU	single	38.6	52.2	56.1	62.1	56.3	52.4	56.1	**63.6**
	multi	38.8	42.2	52.4	57.5	52.7	48.3	51.9	**58.5**
	shift	36.4	42.5	51.9	**57.0**	50.5	46.1	50.5	54.6
JP	single	47.4	51.1	57.1	64.7	**71.4**	59.4	57.1	60.9
	multi	42.1	50.4	60.2	55.6	**59.4**	51.9	50.4	**59.4**
	shift	45.1	47.4	57.1	60.9	**64.7**	47.4	50.4	58.6
UK	single	42.0	50.7	59.9	**66.6**	63.0	55.8	58.7	66.4
	multi	33.8	41.0	52.6	**57.8**	55.8	47.8	51.2	57.5
	shift	34.6	38.7	51.2	56.1	**57.7**	47.6	47.6	57.2
US	single	48.2	58.4	65.4	69.8	**70.5**	60.3	62.5	69.8
	multi	36.4	40.0	59.3	62.8	61.6	53.6	53.0	**63.8**
	shift	38.4	37.9	56.7	61.0	60.3	51.8	49.8	**62.1**
Average	single	43.8	53.8	60.7	**66.5**	64.7	56.8	59.3	**66.5**
	multi	36.5	41.7	55.5	59.3	57.4	50.3	51.9	**60.1**
	shift	37.2	40.0	53.7	**58.4**	57.4	48.7	49.3	**58.4**

Table 2: Cultural entailment scores for base and fine-tuned Qwen3-8B with different methods. We achieve performance improvement with fine-tuning on conversational data. Fine-tuning on augmented single-turn QA (*SFT-Multi SQA*) is the most effective method and inference-time reasoning further enhances its performance.

4.3.1 Training Setup

Within a unified SFT framework, we study how training data format, scale, and cultural scope affect cultural competence using three training formats: (a) cultural knowledge statements (*SFT–Knowledge*), (b) single-turn question–answer (QA) pairs (*SFT–One/Multi SQA*), and (c) multi-turn QA conversations (*SFT–Multi MQA*).

Multi-turn training data are constructed using the turn-wise simulation (Section 3.1), expanding single-turn conversations into multi-turn dialogues with no overlap with evaluation data. To assess data scale, we contrast fine-tuning on a single QA per knowledge item (*SFT–One SQA*) with fine-tuning on ten independently generated QAs (*SFT–Multi SQA*). We also investigate cultural specificity by comparing fine-tuning on mixed-culture data versus culture-specific subsets (*SFT–One Culture*). Additionally, we test a sequential fine-tuning strategy that first trains on augmented single-turn QAs and then on multi-turn QAs (*SFT–Sequential*). Finally, we evaluate inference-time scaling, where models finetuned on ten augmented single-turn QAs are explicitly instructed to reason before responding (*Inference Scaling*).

We hold out 200 randomly selected samples from train data, each corresponding to a unique piece of cultural knowledge, as the validation set. We use a balanced validation split (50 instances per culture) and select the epoch with the highest average entailment score to avoid overfitting.

4.3.2 Results and Discussion

Finetuning on QA is better than on knowledge. Finetuning on conversation data grounded in cultural knowledge (*SFT–One SQA*) injects knowledge to the base model and significantly improves its ability to apply the knowledge in its response (single-turn: 43.8 to 60.7; multi-turn: 36.5 to 55.5; with topic shift: 37.2 to 53.7). Notably, compared to the baseline method of fine-tuning on cultural knowledge (*SFT–Knowledge*) which only results in slight improvement, fine-tuning on single-turn QA (*SFT–One SQA*) achieves performance improvement on both single-turn (53.8 to 60.7), multi-turn without (41.7 to 55.5), and with topic shift (40.0 to 53.7) scenarios. This suggests that conversation-based fine-tuning helps the model learn to contextualize cultural knowledge during interaction, rather than merely memorizing facts (Appendix G.1).

Knowledge-based data augmentation helps with knowledge learning. By augmenting each single-turn QA (*SFT–One SQA*) to multiple single-turn QAs for every knowledge (*SFT–Multi SQA*), it results in more effective cultural learning in both single-turn and multi-turn scenarios (single-turn: 60.7 to 66.5; multi-turn: 55.5 to 59.3; multi-turn with topic shift: 53.7 to 58.4). This aligns with prior findings that fact-based data augmentation improves knowledge injection through SFT (Mecklenburg et al., 2024).

Finetuning on single-turn QA is more effective than on multi-turn QA. Performance gain by

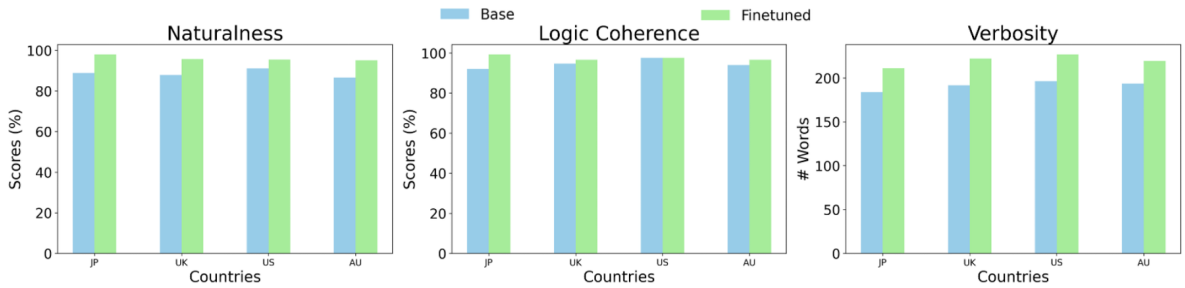

Figure 4: Naturalness, logic coherence, and verbosity of the base and SQAs-finetuned Qwen3-8B models.

fine-tuning with multi-turn data (*SFT–Multi MQA*) (single-turn: 43.8 to 56.8, multi-turn: 36.5 to 50.3) is not as effective as fine-tuning with single-turn conversational data (*SFT–Multi SQA*) (single-turn: 43.8 to 66.5, multi-turn: 36.5 to 59.3). Also, fine-tuning further on both single-turn and multi-turn data (*SFT–Sequential*) can lead to performance drop compared to finetuning only on single-turn data (*SFT–Multi SQA*). This may stem from the increased noise in multi-turn conversations, which dilutes the knowledge signals and complicates the model's learning dynamics to find connections between knowledge and relevant conversation context (see Appendix E.1).

Finetuning on mixed culture achieves comparable performance to finetuning on specific culture. As shown in Table 2, models fine-tuned on culture-specific (*SFT–One Culture*) and mixed-culture data (*SFT–Multi SQA*) achieve comparable performance, with almost the same performance level for all three settings (single-turn: 64.7 vs. 66.5, without topic shift: 57.4 vs. 59.3, with topic shift: 57.4 vs. 58.4). This indicates minimal intercultural knowledge conflict within CultureBank data which we develop from.

Limited and inconsistent gains from inference-time reasoning. *Inference-Scaling* (Kojima et al., 2023) on augmented single-turn QAs (*SFT–Multi SQA*) yields inconsistent improvements: while it benefits certain cultures and settings (particularly in multi-turn and topic-shift scenarios), it provides little to no gain on average (single-turn: 66.5 vs. 66.5; multi-turn: 59.3 vs. 60.1; topic-shift: 58.4 vs. 58.4). This underscores the intrinsic difficulty of multi-turn conversational cultural competence, which cannot be reliably acquired through naive inference-time prompting alone and instead requires training-time adaptation.

Improved cultural knowledge entailment does not degrade response quality. To assess whether gains in cultural entailment come at the expense of conversational quality, we evaluate model responses along three dimensions: naturalness, logical coherence, and verbosity. Naturalness and logical coherence are assessed using Claude-3.5-Sonnet as an automated judge, while verbosity is measured quantitatively by average response length (word count). We compared the base model with the model finetuned on multiple SQAs (*SFT - Multi SQA*), which achieves the strongest entailment performance. We observe no regression in naturalness or logical coherence and instead, both dimensions improve slightly (see Figure 4). In contrast, verbosity increases modestly, indicating a trade-off whereby higher cultural knowledge entailment is associated with longer responses.

5 Conclusion

We construct a high-quality multi-turn conversation dataset grounded in cultural knowledge, providing a benchmark that moves evaluation closer to real-life cultural interactions by capturing how context unfolds across turns. Our zero-shot evaluation reveals that multi-turn conversations pose greater challenges to existing models: as context spreads across turns, model responses become overly general and lose cultural nuance. This suggests that existing cultural evaluation practices may underestimate the complexity of real-world conversations. Fine-tuning on augmented single-turn conversations proves to be most effective in enhancing cultural competence during conversations, implying that cultural knowledge is better internalized when embedded in dialogue-like contexts. We also find that inference-time reasoning is only marginally effective in further improving performance, suggesting the challenging nature of conversational cultural competence.

Limitations

Since our dataset, *CultureTalk*, is built upon *CultureBank*, an existing cultural knowledge base developed by prior work, its knowledge coverage and accuracy are inherently constrained by the scope and quality of the underlying knowledge base. However, our goal in this work is not to exhaustively validate cultural knowledge itself, but to study how models retain and apply cultural knowledge in multi-turn conversational settings. Our data generation method is generalizable and can be readily applied to other cultural knowledge bases of higher quality as they become available. Furthermore, as social and cultural norms evolve over time, the same methodology can be employed with updated cultural knowledge to ensure continued relevance.

As an initial exploration, we develop *CultureTalk* entirely in English but multilingual culture-grounded conversations can better reflect real-world conversations. With multilingual data, more cultural nuances could be expressed and captured. Given the computational cost of applying the full CultureTalk pipeline to all CultureBank cultures, we select four cultures as an initial controlled testbed. Thus, broader coverage across more cultures is an important direction for future work. Moreover, our evaluation focuses on knowledge entailment, which captures an important aspect of cultural competence. Future work may consider additional dimensions of culturally competent dialogue, such as tone calibration, turn-level adaptation, and discourse management.

Ethical Statement

Our work builds upon the *CultureBank* dataset, which was constructed from large-scale online community data. As acknowledged by the dataset creators, despite rigorous filtering, the dataset may still contain stereotypes, controversial statements, or negative content. We encourage researchers to engage with our data thoughtfully, taking into account the potential risk of perpetuating biases.

References

Keqin Bao, Jizhi Zhang, Wenjie Wang, Yang Zhang, Zhengyi Yang, Yancheng Luo, Chong Chen, Fuli Feng, and Qi Tian. 2023. A bi-step grounding paradigm for large language models in recommendation systems. *Preprint*, arXiv:2308.08434.

Yang Trista Cao, Anna Sotnikova, Hal Daumé III, Rachel Rudinger, and Linda Zou. 2022. Theory-grounded measurement of U.S. social stereotypes in English language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1276–1295, Seattle, United States. Association for Computational Linguistics.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. *Preprint*, arXiv:2305.14233.

Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards measuring the representation of subjective global opinions in language models. *Preprint*, arXiv:2306.16388.

Bihao Hu, Jiayi Zhu, Yiying Pei, and Xiaoqing Gu. 2025. Exploring the potential of llm to enhance teaching plans through teaching simulation. *npj Science of Learning*, 10(1):7.

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.

Akshita Jha, Aida Davani, Chandan K. Reddy, Shachi Dave, Vinodkumar Prabhakaran, and Sunipa Dev. 2023. Seegull: A stereotype benchmark with broad geo-cultural coverage leveraging generative models. *Preprint*, arXiv:2305.11840.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. *Preprint*, arXiv:2205.11916.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Boda Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024. Arabicmmlu: Assessing massive multitask language understanding in arabic. *Preprint*, arXiv:2402.12840.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. Culturellm: Incorporating cultural differences into large language models. *Preprint*, arXiv:2402.10946.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024b. Culturepark: Boosting cross-cultural understanding in large language models. *Preprint*, arXiv:2405.15145.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024c. Cmmlu: Measuring massive multitask language understanding in chinese. *Preprint*, arXiv:2306.09212.

Yubo Li, Xiaobin Shen, Xinyu Yao, Xueying Ding, Yidi Miao, Ramayya Krishnan, and Rema Padman. 2025. Beyond single-turn: A survey on multi-turn interactions with large language models. *Preprint*, arXiv:2504.04717.

Nick Mecklenburg, Yiyou Lin, Xiaoxiao Li, Daniel Holstein, Leonardo Nunes, Sara Malvar, Bruno Silva, Ranveer Chandra, Vijay Aski, Pavan Kumar Reddy Yannam, Tolga Aktas, and Todd Hendry. 2024. Injecting new knowledge into large language models via supervised fine-tuning. *Preprint*, arXiv:2404.00213.

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, and 3 others. 2025. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Preprint*, arXiv:2406.09948.

Keivalya Pandya and Mehfuza Holia. 2023. Automating customer service using langchain: Building custom open-source gpt chatbot for organizations. *Preprint*, arXiv:2310.05421.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2024. Survey of cultural awareness in language models: Text and beyond. *Preprint*, arXiv:2411.00860.

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2024. SMILE: Single-turn to multi-turn inclusive language expansion via ChatGPT for mental health support. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 615–636, Miami, Florida, USA. Association for Computational Linguistics.

Michael J. Ryan, William Held, and Diyi Yang. 2024. Unintended impacts of llm alignment on global representation. *Preprint*, arXiv:2402.15018.

Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Chunhua yu, Raya Horesh, Rogério Abreu de Paula, and Diyi Yang. 2024. Culturebank: An online community-driven knowledge base towards culturally aware language technologies. *Preprint*, arXiv:2404.15238.

Haoyi Xiong, Jiang Bian, Yuchen Li, Xuhong Li, Mengnan Du, Shuaiqiang Wang, Dawei Yin, and Sumi Helal. 2024. When search engine services meet large language models: Visions and challenges. *Preprint*, arXiv:2407.00128.

Haolan Zhan, Zhuang Li, Xiaoxi Kang, Tao Feng, Yuncheng Hua, Lizhen Qu, Yi Ying, Mei Rianto Chandra, Kelly Rosalin, Jureynolds Jureynolds, Suraj Sharma, Shilin Qu, Linhao Luo, Ingrid Zukerman, Lay-Ki Soon, Zhaleh Semnani Azad, and Reza Haf. 2024. RENOVI: A benchmark towards remediating norm violations in socio-cultural conversations. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3104–3117, Mexico City, Mexico. Association for Computational Linguistics.

A Dataset Statistics

Culture	Validated	Generated	% Pass
Australian	412	470	87.7
American	610	700	87.1
British	586	675	86.8
Japanese	133	152	87.5
Total	**1741**	**1997**	**87.2**

Table 3: Statistics of generated and validated evaluation data in *CultureTalk*.

After filtering supported by Deepseek-R1 as a judge, 87.2% of conversations passed all three filters, which suggests the robustness of our multi-turn data curation pipeline. The final evaluation set comprises 1,741 conversations across Australian (412), American (610), British (586), and Japanese (133) cultures (see Table 3).

B Instruction Prompts

B.1 Instruction Prompts for Turn-wise Utterance Simulation

B.1.1 User Simulation (information sharing utterance)

Conversation History: {dialog_history}
You are {persona} who is having a conversation with an AI assistant.
Please generate an utterance by the user in the next turn that goes along with the dialog history and implies only the information in {context}.
This utterance is about information sharing instead of seeking advice.
Please ensure the given utterance is less elaborative and more like what will be said in real world.
Please just output the utterance only without any other explanations.

B.1.2 User Simulation (advice seeking utterance)

Conversation History: {dialog_history}
You are {persona} who is having a conversation with an AI assistant.
Please generate an utterance by the user in the next turn that goes along with the dialog history and implies only the information in {query}.
Please just output the utterance only without any other explanations.

B.1.3 AI Assistant Simulation

Conversation History: {dialog_history}
You are a helpful, respectful and intelligent assistant trained to identify and extract cultural information. Your role is to follow the given instructions precisely and format your responses as required. Keep your responses succinct and limited to the requested information. If you don't know the answer to a question, please don't share false information.
Please generate an utterance by the user in the next turn that goes along with the dialog history and implies only the information in {query}.
Please just output the utterance only without any other explanations.

B.2 Instruction Prompts for Topic Shift Injection

B.2.1 User Simulation

Conversation History: {dialog_history}
You are {persona} who is having a conversation with an AI assistant.
Please generate an utterance by the user in the next turn that goes along with the dialog history but introduces a topic change.
Please ensure the given utterance is less elaborative and more like what will be said in real world.
Please just output the utterance only without any other explanations.

B.2.2 AI Assistant Simulation

Conversation History:{dialog_history}
You are a helpful, respectful and intelligent assistant trained to identify and extract cultural information. Your role is to follow the given instructions precisely and format your responses as required. Keep your responses succinct and limited to the requested information. If you don't know the answer to a question, please don't share false information.
Please generate an utterance by the user in the next turn that goes along with the dialog history and implies only the information in {query}.
Please just output the utterance only without any other explanations.

B.3 Instruction Prompts for Quality Checking

B.3.1 Criterion 1: Cultural Knowledge Incorporation Check

Cultural principle: {knowledge}
Conversation: {conversation}
Is the last user utterance in the given conversation relevant to the cultural principle? Please give your answer in 'yes' or 'no' only without other words.

B.3.2 Criterion 2: Conversation Quality Check - Naturalness

Conversation: {conversation}
Does the conversation read like a natural human-AI conversation? Please give your answer in 'yes' or 'no' only without other words.

B.3.3 Criterion 3: Conversation Quality Check - Logic Coherence

Conversation: {conversation}
Is the given conversation logically coherent? Please give your answer in 'yes' or 'no' only without other words.

B.4 Instruction Prompts for Single-Turn QA Generation

We reused the prompt template in CultureBank in curating single-turn QA train data:

You are an AI assistant tasked with mentoring the training of therapists and consultants . Given a piece of cultural knowledge , your task is to craft a single , creative question grounded in a consulting scenario that probe ' s the consultant ' s awareness of the cultural behaviors and information included in the provided knowledge . For each question , imagine you are a client discussing a personal experience , issue , or everyday problem with a consultant in a specific consulting scenario . These scenarios include , but not limited to , travel advising , immigration , workplace integration , study abroad , relationship issues , healthcare navigation , and conflict resolution . In particular , when generating each question , you should :

1. Choose a consulting scenario
2. Generate a persona for the client
3. Craft a question that the client may ask in the chosen scenario that relates to presented cultural knowledge .

Keep the following guidelines in mind :

1. Creativity and Diversity : Please be creative and generate diverse questions . Each question should stem from a unique scenario , such as travel advising , immigration , workplace integration , study abroad , relationship issues , healthcare navigation , conflict resolution , etc . Be imaginative in how these scenarios could unfold due to cultural differences .

2. Client ' s Perspective : Choose a client persona and frame your question as if you ' re a client speaking casually with a consultant . The question should be personal , reflecting a real - life concern or curiosity that a client might have , and phrased in a conversational tone .

3. Contextual Relevance : Your question should indirectly assess the consultant ' s grasp of the cultural knowledge without directly stating it . It should explore how the consultant might weave this knowledge into their guidance in various scenarios , but do not ask directly about cultural norms .

4. Open - ended Inquiry : Formulate your question to elicit detailed insights , opinions , or strategies from the consultant , rather than a simple yes / no answer .

Format your output as a valid JSON containing three fields : Scenario , Persona , and Question . Do not include any additional words .

Knowledge : {cultural_knowledge_description}

Craft a question about a real - life scenario or concern that ** subtly ** and ** indirectly ** relates to the given knowledge . Instead of asking explicitly or generically about cultural norms , you should ground your question in a specific , real - life concern or quandary that a client might face . Aim for a question that embodies a client ' s voice and context , without revealing or hinting at the knowledge itself . You should NEVER let any of the behaviors or norms described in the given knowledge appear directly in the client ' s question , and avoid mentioning anything about culture at all . Instead , use anecdotes from the client ' s perspective . Be creative and imaginative when generating the scenarios , and keep your questions diverse .

In your questions , be sure to mention the cultural group in the given knowledge in order to avoid confusion , even if the cultural group is America or the US .

Format your output as a valid json .

C Topic

CultureTalk contains a diverse coverage of 35 different topics, as shown in Table 4.

Topic	Count	Percentage (%)
Social Norms and Etiquette	255	13.4
Food and Dining	180	9.5
Communication and Language	142	7.5
Miscellaneous	125	6.6
Consumer Behavior	107	5.6
Cultural Exchange	103	5.4
Health and Hygiene	92	4.8
Cultural Traditions and Festivals	91	4.8
Finance and Economy	78	4.1
Environmental Adaptation and Sustainability	77	4.0
Community and Identity	68	3.6
Education and Technology	52	2.7
Social Interactions	48	2.5
Cultural and Environmental Appreciation	47	2.5
Household and Daily Life	43	2.3
Drinking and Alcohol	40	2.1
Entertainment and Leisure	39	2.0
Safety and Security	35	1.8
Migration and Cultural Adaptation	35	1.8
Workplace	33	1.7
Lifestyles	30	1.6
Family Traditions and Heritage	28	1.5
Work-Life Balance	23	1.2
Beauty and Fashion	20	1.1
Family Dynamics	20	1.1
Relationships and Marriage	17	0.9
Transportation	15	0.8
Humor and Storytelling	12	0.6
Time Management and Punctuality	10	0.5
Sports and Recreation	10	0.5
Travelling	9	0.5
Social Infrastructure	7	0.4
Pet and Animal Care	5	0.3
Dress Codes	4	0.2
Religious Practices	3	0.2

Table 4: Distribution of Topics in CultureTalk

D Training Experiment Setup

D.1 Details of Train Data

SFT-Knowledge: cultural knowledge directly sourced from CultureBank for all four cultures.

SFT-One SQA / SFT-Multi SQA: augmented using the CultureBank construction pipeline where single-turn QA pairs are generated via a manually designed prompt grounded in human-verified cultural knowledge for all four cultures.

SFT-One Culture: augmented using the CultureBank construction pipeline where single-turn QA pairs are generated via a manually designed prompt

grounded in human-verified cultural knowledge for each of the four cultures.

SFT-MultiMQA: augmented using the CultureBank and CultureTalk construction pipeline where single-turn QA pairs are first generated via a manually designed prompt grounded in human-verified cultural knowledge and multi-turn QA pairs are then curated using our turn-wise simulation (Figure 2) for all four cultures.

SFT-Sequential: data used in *SFT-MultiSQA* and *SFT-MultiMQA*

The train data developed in all settings is fully disjoint from the evaluation set containing 1741 instances.

D.2 Training Hyperparameters

For all finetuning experiments, we use the hyperparameter $lr_scheduler_type$ = "$cosine$", $warmup_ratio$ = 0.05, $weight_decay$ = 0.01, max_grad_norm = 1.0, $optim$ = "$adamw_torch$". We adjust $learning_rate$ based on the data quantity for different settings: $5e-6$ for *SFT-Knowledge* and *SFT-One SQA*, $1e-5$ for *SFT-Multi SQA*, $2e-6$ for *SFT-Multi MQA*, $1e-6$ for the second round finetuning on MQA in *SFT-Sequential*. We set number of epochs to 10 for every setting and select the epoch with highest entailment score on the hold-out validation set.

E Human Validation Guidelines

Four authors independently rate the generated multi-turn data on a Likert scale. These ratings are intended as a quality check on a random stratified subset rather than a full-scale annotation study. Below is the detailed validation guideline:

Naturalness (1 to 5)

5 (Very Natural): The conversation flows smoothly like a real human interaction, with appropriate responses and natural transitions.

4 (Somewhat Natural): The dialogue maintains a generally natural flow despite occasional minor awkwardness.

3 (Neutral): The conversation shows a mix of natural and artificial elements, with some responses feeling human-like while others seem mechanical.

2 (Somewhat Unnatural): The exchange feels noticeably stilted with awkward responses and unnatural transitions.

1 (Very Unnatural): The conversation is highly artificial and robotic, lacking the flow and coherence of human dialogue.

Logic coherence (1 to 5)

5 (Very Coherent): The conversation maintains clear logical flow with well-connected ideas and appropriate responses that build on previous context.

4 (Somewhat Coherent): The dialogue follows a mostly logical progression with minor inconsistencies or gaps in reasoning.

3 (Neutral): The conversation shows mixed coherence with some logical connections but also contains noticeable jumps or disconnects in reasoning.

2 (Somewhat Incoherent): The exchange often lacks logical connection between turns with several contradictions or non-sequiturs.

1 (Very Incoherent): The conversation is highly illogical with frequent contradictions, irrelevant responses, and broken chains of reasoning.

Relevance (1 to 5)

5 (Very Relevant): The conversation directly addresses and accurately illustrates the given cultural principle throughout the dialogue.

4 (Somewhat Relevant): The dialogue mostly relates to the cultural principle with some minor deviations or indirect connections.

3 (Neutral): The conversation partially addresses the cultural principle but includes content that isn't clearly connected to it.

2 (Somewhat Irrelevant): The exchange has only weak or superficial connections to the cultural principle, often straying from the main topic.

1 (Very Irrelevant): The conversation fails to meaningfully address or connect with the given cultural principle at all.

E.1 Comparison between Training on Single-Turn and Multi-Turn QA

As shown in Table 5, the model finetuned directly on the cultural knowledge focuses on recommending places to go related to Japanese anime and does not discuss cultural meaning and shared engagement among creators and viewers. The prior user-AI conversation turns in multi-turn dialogues may introduce additional noises, making it harder for the model to learn the connection of responses acknowledging the norm (strong engagement with anime) with relevant conversational contexts. On the other hand, finetuning on single-turn QA can better help establish connection between

the knowledge (*deep engagement and relatability with anime*) and the relevant context (*connect with Japanese coworkers on TV interests*) and the model is able to provide a response that acknowledges the prevalent cultural norm about Japanese anime in the related conversational context.

Knowledge
In the context of anime, both creators and viewers engage in the creation, consumption, and expression of relatability towards anime characters, with the goal of enjoying and connecting with these characters on a personal level ...
Train Data Formats
Single-turn QA:
Q: I've noticed my Japanese colleagues often discuss their favorite TV shows during lunch breaks, and they seem really passionate about certain characters ... Any advice on how I can connect with my coworkers over their TV interests without seeming out of place?
A: Your observation about American coworkers discussing TV shows and fictional characters is quite insightful. This behavior is indeed common ... as anime viewers in Japan often relate deeply to characters and discuss them in their daily lives ...
Multi-turn QA:
Q: My Japanese coworkers always get super excited talking about their TV shows during lunch. They really get into those characters.,
A: Yes, Japanese dramas and TV shows are hugely popular! They call them "dorama" and people often follow their favorite actors from show to show. It's really part of the daily culture there.
Q: I want to get into their TV show discussions but I have no idea where to even start.
A: I can help you get started! Some popular recent Japanese dramas include "Alice in Borderland" and "Midnight Diner" which are both on Netflix with English subtitles. They're great entry points since they're easily accessible and have gotten international attention. Which genres do you typically enjoy watching?
Q: Any tips for joining their conversations without coming across as awkward or like I'm trying too hard?
A: A natural way to join in would be to start by asking them about their favorite characters and what makes those characters relatable to them ...
Responses
Response from model finetuned on single-turn QA:
... Your interest in anime and Japanese pop culture aligns perfectly with a significant aspect of modern Japanese society ... anime and manga play a crucial role in Japan's cultural exports and global influence ... Your interest in anime is shared by many Japanese people as well ...
Response from model finetuned on multi-turn QA:
... Japan has several major anime studios you can visit, like Sunrise and Toei Animation, where you can see the production process firsthand. There are also anime fan events and conventions throughout the year where you can meet both industry professionals and fellow fans ...

Table 5: Examples of model responses demonstrating difference in training on single-turn and multi-turn QA.

F Human Evaluation of Entailment Judgments

To evaluate the reliability of the LLM-based entailment metric, we conduct a human-model agreement study. We randomly sample 100 model responses generated by the Qwen3-8B base model and Qwen3-8B finetuned on augmented single-turn QAs for 50 different cultural knowledge items. Three authors are provided the conversation context, model response, and reference cultural knowledge, and label whether the response entails the corresponding knowledge.

The LLM judge achieves 77.0% agreement with human annotations, with Cohen's $\kappa = 0.53$, indicating moderate agreement. Moreover, both human annotations and the LLM judge show that the finetuned Qwen3-8B model attains substantially higher entailment scores than the base model (human: 0.60 vs. 0.22; LLM judge: 0.58 vs. 0.30), demonstrating consistent performance trends across evaluation methods. Overall, these findings suggest that the LLM judge provides a reasonably reliable proxy for cultural knowledge entailment, capturing both individual-label agreement and the correct relative performance trend across models.

G Qualitative Examples

G.1 Effectiveness of Training on QA

Train Data Formats
Knowledge: In Japanese parking lots, ... careful parking ... patiently waiting for parking spots ... ensure efficient and courteous parking practices
Single QA: Q: I'm about to start my new job in Tokyo next week... the company has a small parking lot, and I've heard it can get pretty busy. Do you have any tips on how I can avoid causing any issues or offending anyone when I park my car there? A: ... park carefully ... wait patiently ... ensure efficient and courteous parking
Responses
Response from model finetuned on knowledge: Proof of a designated parking spot... many parking lots in cities utilize automated mechanical systems.
Response from model finetuned on single-turn QA: ... designated parking spot ... cultural emphasis on order and responsibility ... plan ahead and secure a spot ... respect for order ... value of harmony and consideration of others.

Table 6: Examples of model responses demonstrating difference in training on knowledge statement and single-turn QA.

As shown in Table 6, the model finetuned directly on the cultural knowledge focuses on logistical aspects only and does not capture cultural attitudes and behavioral norms related to parking in Japan like carefulness, patience and harmony. Instead, if we finetune the model on a social scenario (*about to start new job in Tokyo and wonder about some tips about parking*) and the response that embeds knowledge, the model is able to contain social-behavioral elements and key information about the cultural attitudes (*harbony, order, consideration*).

G.2 Failures in Multi-turn Scenarios

We report concrete qualitative examples of model outputs in single-turn and multi-turn scenarios for each error type (knowledge error and specificity error) in Table 7. Within knowledge failure, we observe two distinct patterns: **topic drift**, where the model shifts to a related but non-target cultural subtopic, and **wrong concept**, in which the model provides an incorrect cultural detail. For specificity error, we identify two subcategories: **missing normative detail** when too vague to capture the specific social meaning and prevalence in the gold knowledge, and high-level generalization where the model misses the required context specificity. As the knowledge description and model responses are lengthy, we summarize the key points of model responses in the table.

G.3 Failures in Topic-Shift Scenarios

We performed qualitative analysis for instances where model is able to generate response entailing the correct knowledge but fails in topic-shift scenarios. We observe that with topic shift turn introduction, models may tend to be distracted by the topic shift turns and provide a response failing to recognize important nuanaces in cultural differences compared to the multi-turn scenario without topic shifts. Here is one concrete example where the model fails to recognize that the imperial system is predominantly favored in the US with introduction of topic shift turns:

Cultural Knowledge: In the United States and globally, it is common for people and the US to import a variety of products from countries such as Mexico, Canada, and Europe, often labeling them as 'European Imports'. This practice extends to both food produce and other goods, with the imported produce often originating from the same countries as the imported products. This behavior is widely regarded as a standard practice within the sampled population, reflecting a high level of normativity.

Conversation: **User:** "I'm heading to the States for three months to work with our American team on a big project.
Assistant: "That's exciting! Which state will you be based in? I can share some helpful tips about American work culture."
User: "I've heard they run things quite differently in American offices."
Assistant: "Yes, American workplace culture has some distinct differences! For example, Americans tend to be more direct in communication and meetings are usually quite structured. What aspects of US work culture are you most curious about?"
User: "I need to be clear about how I walk through technical specs with them - back home we do things a bit differently."
Assistant: "Let me share some tips about reviewing technical specs in American meetings. They typically prefer detailed documentation upfront, clear action items, and regular status updates. Would you like specific guidance on structuring those technical discussions?"

Cultural Knowledge	User Query / Model Response	Error Analysis in Multi-Turn Settings
In Japanese restaurants, staff and customers exchange the greeting *"irrashaimase"*, creating a hospitable atmosphere. Responding in kind is the cultural norm and is widely acknowledged and practiced.	**Query:** Ask about Japanese dining norm, specifically phrases to use when being greeted. **Response in Single-turn** bowing, using both hands when receiving items, saying *itadakimasu*. **Response in Multi-turn** removing shoes, bowing slightly, saying *itadakimasu* and *gochisousama deshita*.	**Knowledge Failure - Topic Drift** The knowledge emphasizes the greeting *"irrashaimase"*, but responses omit it and focus instead on related concept like table manners and gratitude phrases.
In the United States, particularly in BJ's and Philly, it is common for people to have a preference for hard, salty, and buttery pretzels, which are known for their long shelf life and on-demand availability.	**Query:** Ask what kind of pretzels are iconic on the East Coast. **Response in Single-turn** recommends Philadelphia soft pretzels and related soft-pretzel experiences. **Response in Multi-turn** recommends soft pretzels from Reading Terminal Market, Miller's Twist, Federal Pretzel Bakery.	**Knowledge Failure - Wrong Concept** The cultural knowledge emphasizes *hard* pretzels, while the model confidently answers with *soft* Philly pretzels.
Americans often talk about the healthcare system with a mix of frustration, dissatisfaction, appreciation, and sarcasm.	**Query:** Ask about how locals usually talk about the U.S. healthcare system. **Response in Single-turn** mentions healthcare complexities, high costs, and public frustration. **Response in Multi-turn** mentions may have strong opinions and personal stories, and suggests approaching the topic respectfully.	**Specificity Failure - Missing Normative Detail** The single-turn response captures important cultural nuance, including frustration with the healthcare system. The multi-turn response remains broadly relevant but is too generic, omitting the more specific normative pattern that Americans often discuss healthcare in a frustrated manner sometimes.
During Christmas in Japan, traditions include eating KFC, Christmas cake, and other festive dishes. These practices, influenced by the West, are now widely accepted as distinct Japanese customs.	**Query:** Ask about traditional dishes during Christmas in Japan. **Response in Single-turn** mentions KFC, oden, and Christmas cake. **Response in Multi-turn** lists osechi ryori, hot pot dishes, seasonal sweets.	**Specificity Failure - High-level Generalization** The single-turn response matches the knowledge (KFC, cake). The multi-turn shifts to winter foods, missing Christmas specificity.

Table 7: Examples of cultural knowledge, model responses, and error analyses.

User: "That would help, yeah. Also - what's the deal with tipping over there? I've heard you need to tip for everything." (**topic shift turn 1**)
Assistant: "Tipping is a big part of American culture - typically 15-20% at restaurants, $1-2 per drink at bars, and a few dollars for taxi drivers or hotel staff. For services like haircuts or food delivery, 15% is standard. You generally don't need to tip for retail purchases or professional services though." (**topic shift turn 1**)
User: "Ok, that's helpful about the tipping. And for meetings - should I book them well in advance or is it more casual there?" (**topic shift turn 2**)
Assistant: "For business meetings in the US, it's best to schedule them well in advance - Americans generally prefer structured calendars. Try to book at least a few days ahead when possible, and always send a calendar invite with agenda items. Last-minute meetings are typically seen as disruptive unless there's a genuine urgency." (**topic shift turn 2**)
User: "Being in engineering, I'm wondering how to handle the whole metric versus imperial thing when I'm working with the team there."

Response in Multi-turn without Topic Shift: ... in the United States, you'll need to be prepared to work with imperial measurements ...

Response in Multi-turn with Topic Shift: ... in the US, you'll need to be comfortable with and prepared to work with both metric and imperial measurements ...

Model	Entail in both single- and multi-turn	Entail in single-turn but not multi-turn	Entail in multi-turn but not single-turn	Not entail in either single- or multi-turn
Claude-Haiku	38.5%	15.1%	5.3%	41.0%
Claude-Sonnet	34.7%	12.1%	8.9%	44.3%
Qwen3-8B	46.4%	17.0%	9.7%	26.9%
Qwen3-32B	36.7%	14.5%	10.6%	38.3%
Mistral-7B	42.9%	17.1%	6.8%	33.2%
Mixtral	43.3%	18.7%	5.6%	32.5%
Deepseek-r1	29.1%	15.6%	6.0%	49.3%
GPT-OSS-20B	31.9%	15.4%	11.6%	41.2%

Table 8: Entailment results across models for single- and multi-turn settings.

G.4 Evaluation Data Examples

We provide concrete examples of how knowledge, single-turn QA, and multi-turn QA look like in the CultureTalk evaluation dataset developed (see Table 9).

H Breakdown of Prediction Shifts

In Table 8, we compute the breakdown of prediction shifts for each model. For each example, we compare whether the model is correct in the single-turn setting and whether it remains correct in the corresponding multi-turn setting.

We can see the proportion of casess where models entailing the knowledge in single-turn but not in multi-turn is consistently higher than those where models entailing the knowledge in multi-turn but not in single-turn. . This supports our main finding that multi-turn conversations introduce additional challenges beyond those captured by single-turn evaluation.

Example 1

Knowledge:
In the United States, it is common for people to actively seek and offer help, often relying on community support. This behavior reflects a strong sense of communal responsibility and a willingness to assist others in need. The normativity of this behavior is widely regarded as very high within the sampled population, indicating that seeking and offering help is a deeply ingrained aspect of American culture. This cultural norm emphasizes the value placed on mutual support and cooperation, contributing to a strong sense of community and interconnectedness.

Single QA:
I've recently moved to this new town in the US and I'm trying to settle in, but I'm having a hard time getting used to how open and helpful everyone is here. Back home, I was taught that asking for help is a sign of dependency, but it seems like people here really value offering a hand. How can I adapt to this new way of living without feeling like I'm burdening others with my questions?

Multiturn QA:
User: "Back in my country, we always keep to ourselves and solve problems alone... it's strange seeing people here freely ask others for help."
Assistant: 'I understand that cultural differences around asking for help can feel surprising. Different societies have their own norms about independence versus interdependence. What made you start noticing this difference?
User: "You know, in my culture, we're taught that relying on others shows weakness... but here, I notice people help each other all the time."
Assistant: "That's an interesting observation. While self-reliance is valued in many cultures, seeking help can actually be a sign of strength - it shows wisdom in recognizing when others' perspectives or support could be valuable. What has it been like for you adjusting to this different approach?"
User: "It's still hard... I need to find a good doctor here but I don't know how to even start looking."
Assistant: 'I understand it can feel uncomfortable, but finding a doctor is important for your health. Would you like me to suggest some ways to start your search that might feel more comfortable for you?'
User: "Yes... I think I need some help finding one. It's not easy for me to ask, but I want to stay healthy."
Assistant: "I'm glad you're taking this step. Let me help walk you through some options - we can start by looking for doctors covered by your insurance, or checking reviews on medical websites. What would feel most comfortable for you to try first?
User: "I want to learn how to adjust to this more open way of getting help... but I feel guilty every time I need to ask someone for anything. I keep thinking I'm bothering them...

Example 2

Knowledge:
In various comparisons and preferences, people in Australia commonly express their preference for their own country over others and acknowledge similarities with them. This behavior reflects a strong sense of national pride and a tendency to favor their own culture and lifestyle. The sampled population widely regards this behavior as common and normative within the Australian cultural context.

Single QA:
I've been tasked with a short-term project in Australia and I want to make sure I fit in and do my best to understand the local culture. Can you give me some tips on how to engage with Australians in a way that shows I respect and appreciate their lifestyle, especially when it comes to their preferences and how they view their own country?

Multiturn QA:
User: "I've got a project coming up in Australia - trying to learn about how things work over there so I can fit in better."
Assistant: "What aspects of Australian work culture are you most interested in learning about? I can help explain business etiquette, communication styles, and workplace customs that might be different from what you're used to."
User: "What's the dress code like in Aussie offices? Don't want to stick out like a sore thumb."
Assistant: 'Australian office dress codes tend to be relatively relaxed and business casual in most workplaces, though it can vary by industry. Generally, smart casual attire like collared shirts, slacks or neat skirts, and closed-toe shoes is appropriate. Suits are mainly worn in traditional corporate environments like banking and law firms. Just avoid thongs (flip-flops), shorts, or overly casual wear.'
User: "What's the deal with lunch breaks there? Pretty different from the US?"
Assistant: "Lunch breaks in Australia are typically more relaxed and social compared to the US. Most workers take a proper lunch break, usually 30-60 minutes, and it's common to step away from your desk to eat with colleagues. Many will head out to local cafes or bring food to eat in common areas. The concept of working through lunch isn't as prevalent as it is in the US."
User: "I want to make sure I interact with Australians respectfully - any advice on what they value and how they see their country? I don't want to come across as an ignorant American."

Table 9: Examples of model responses demonstrating difference in training on knowledge statement and single-turn QA.