Unveiling Cultural Blind Spots: Analyzing the Limitations of mLLMs in Procedural Text Comprehension

Amir Hossein Yari[1]   **Fajri Koto**[2]

[1]Sharif University of Technology
[2]Mohamed bin Zayed University of Artificial Intelligence
ahyari2002@gmail.com, fajri.koto@mbzuai.ac.ae

Abstract

Despite the impressive performance of multilingual large language models (mLLMs) in various natural language processing tasks, their ability to understand procedural texts, particularly those with culture-specific content, remains largely unexplored. Texts describing cultural procedures, including rituals, traditional craftsmanship, and social etiquette, require an inherent understanding of cultural context, presenting a significant challenge for mLLMs. In this work, we introduce **CAPTex**, a benchmark designed to evaluate mLLMs' ability to process and reason over culturally diverse procedural texts in multiple languages. Using a range of evaluation methods, we find that (1) mLLMs struggle with culturally contextualized procedural content, particularly in low-resource languages; (2) performance varies across cultural domains, with some proving more difficult than others; and (3) models perform better on multiple-choice tasks presented in conversational formats than on direct questions. These results highlight the current limitations of mLLMs and emphasize the need for culturally informed benchmarks like **CAPTex** to support more accurate and inclusive language understanding.[1]

1 Introduction

Procedural texts encompass a genre of writing that provides systematic instructions or guidance to navigate a sequence of actions or steps, aiming to achieve a specific outcome. These texts are common in various fields, including technical documentation, user manuals, and cookbooks. The core characteristic of procedural texts is their sequential and organized structure, with each instruction building on the previous one to ensure readers can successfully reach the intended outcome. Unlike other writing styles, such as narrative or descriptive, procedural texts emphasize clarity, accuracy, and a straightforward progression of actions to enable effective task completion.

Large Language Models (LLMs) have demonstrated exceptional capabilities across various natural language processing (NLP) tasks, such as text summarization (Jin et al., 2024), multi-modal machine translation (Shen et al., 2024a), solving complex tasks modeled as state machines (Wu et al.), and code generation and understanding (Wong et al., 2023). Unlike traditional models that rely on task-specific training, LLMs can be adapted to a wide range of applications through effective prompting strategies, making them suitable for diverse and dynamic contexts (Ouyang et al., 2022; Dai et al., 2023).

One particularly significant application area for LLMs is their ability to accurately interpret procedural texts. This capability is becoming increasingly vital as these models are employed in tasks like generating automated instructions and facilitating human-computer interactions (Kosch and Feger, 2024). In such scenarios, the demand for clear, contextually appropriate, and executable steps is critical. However, inaccuracies or ambiguities in interpreting procedural instructions can result in miscommunication, errors, and inefficiencies. These issues are particularly concerning in real-world domains such as healthcare, education, and technical support, where precision and clarity are paramount. Consequently, ensuring that LLMs can reliably process procedural texts is essential for their effective and responsible integration into various systems.

Culture plays a fundamental role in both the creation and comprehension of procedures. A step-by-step instruction in one culture may rely on shared knowledge or tacit understanding that is absent in others (Steffensen et al., 1979). For instance, instructions on performing a ritual practice may have

[1] This dataset is publicly released under a Creative Commons license at https://huggingface.co/datasets/AmirHossein2002/CAPTex

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 20151–20170
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

Step	Iran	Indonesia
1	The body is taken to the cemetery for burial.	The traditional Toraja house is prepared in silence, and tools are gathered.
2	The deceased's body is washed according to Islamic rituals.	The body is wrapped, and the coffin is decorated in a ritual performance.
3	The body is wrapped in a plain white shroud.	A cultural parade is held, transporting the body from the house to the burial site.
4	A special prayer is performed for the deceased.	A traditional Toraja dance is performed as part of the ceremony.
5	The body is buried facing Mecca, with a layer of dirt and stones over the grave.	Animal sacrifices, typically buffalo and pigs, are offered as part of the final rites.

Table 1: Comparison of funeral practices in Iran and Indonesia. For Indonesia, it's a tradition from North Sumatra.

meanings, symbols, and steps deeply embedded in the traditions of a particular culture, as demonstrated in Table 1. If language models are unable to recognize and navigate these cultural dimensions, they risk misrepresenting the intent or structure of the procedure, leading to errors or misinterpretations.

Studies have shown that current large language models (LLMs) tend to exhibit biases favoring Western perspectives, mirroring the cultural norms and values of Western, educated, industrialized, rich, and democratic (WEIRD) societies (Durmus et al., 2024; Naous et al., 2024), while often inadequately representing other cultural contexts (Prabhakaran et al., 2022). These biases primarily stem from the nature of training data (Arora et al., 2022; Ganguli et al., 2022; Nadeem et al., 2021) and design decisions, including model architecture, tokenization approaches, evaluation methods, and instruction-tuning techniques. The significance of LLMs accurately comprehending these texts extends beyond the technical understanding of instructions; it also involves ensuring fairness, accessibility, and cultural sensitivity. Evaluating how well LLMs can decode procedural texts, particularly when cultural context plays a pivotal role, is critical for advancing their capability to serve a diverse range of users across different linguistic and cultural backgrounds.

Cultural procedural texts, which are deeply intertwined with societal norms, values, and traditions, pose unique challenges for LLMs. This complexity raises several critical questions about the ability of LLMs to effectively navigate and reason within culturally specific contexts: (1) How do LLMs perform in understanding procedural texts in low-resource languages compared to high-resource languages? (2) How effectively can LLMs recognize, interpret, and preserve the cultural nuances embedded in procedural texts? (3) Do LLMs demonstrate consistent performance across different cultural domains, such as food preparation, religious rituals, and celebration setups? (4) Are there noticeable strengths or weaknesses in LLMs' understanding depending on the cultural context of the procedural text?

To address these questions, we introduce **CAPTex** (**C**ulturally-**A**ware **P**rocedural **Tex**ts), an innovative dataset crafted to evaluate multilingual LLMs' (mLLMs) ability to reason culturally through the lens of procedural text understanding across diverse tasks, including reordering tasks, multiple-choice questions, and conversational frameworks, each of which is elaborated upon in detail in Section 3.2. **CAPTex** is carefully developed with contributions from native speakers representing seven culturally distinct regions—China, India, Indonesia, Iran, Japan, Nigeria, and Pakistan—ensuring authentic and nuanced cultural representation.

2 Related Work

Procedural text analysis has been a focal point of research, addressing a wide array of tasks within this domain. For example, Cao et al. (2024) tackles the cultural adaptation of recipes between Chinese and English-speaking cuisines. Their work aims to automate the translation and cultural adaptation of recipes, ensuring that cultural nuances—including ingredients, cooking techniques, and unit conversions—are appropriately represented. In contrast, our work extends beyond the food domain, encompassing multiple cultural contexts across seven countries, thereby offering a broader perspective on cross-cultural procedural knowledge.

Several studies have also focused on advancing entity tracking methodologies. NCET (Gupta and Durrett, 2019) introduces a mechanism for continuous-space entity tracking, employing a conditional random field (CRF) to ensure sequential consistency in predictions. Similarly, Huang et al.

(2021) utilizes a graph neural network to model semantic relationships among entities, actions, and locations, enhancing the understanding of procedural text.

Incorporating temporal aspects into procedural comprehension, Rajaby Faghihi and Kordjamshidi (2021) propose the Time-Stamped Language Model (TSLM), which augments pre-trained language models with timestamp embeddings. This approach has significantly improved performance on datasets such as Propara (Dalvi et al., 2018) and NPN-Cooking. Additionally, Tang et al. (2020) introduces the Interactive Entity Network (IEN), a recurrent network with memory designed to capture diverse entity interactions for state tracking. Meanwhile, Amini et al. (2020) develops an algorithm for procedural reading comprehension, translating texts into a formalism that represents processes as sequences of transitions over entity attributes.

Efforts to integrate multimodal data have also advanced procedural text analysis. For instance, Liu et al. (2020) introduces a transformer-based model that combines textual and visual information for processing multimodal recipe datasets effectively. Building on this, Wu et al. (2022) conducts benchmarking on reasoning and sequencing unordered multimodal instructions, highlighting that state-of-the-art models still fall short of human-level performance. While their work primarily focuses on step reordering, our evaluation framework is more comprehensive, introducing three additional tasks to assess LLMs' capabilities. Furthermore, rather than being restricted to English, our research incorporates the native languages of the targeted countries, ensuring that the procedures analyzed are culturally unique rather than globally common.

Despite these advancements, a holistic benchmark for procedural text comprehension remains elusive. Our work sets a new standard by extending beyond food-related tasks to encompass multiple domains, incorporating a diverse range of languages beyond English, and evaluating the capabilities and limitations of mLLMs through a multifaceted assessment framework. In the following section, we will elaborate on these methodologies in detail, highlighting how our benchmark surpasses prior efforts.

3 CAPTex

Language	**Class**
Chinese (Mandarin)	5 - The Winners
Japanese	5 - The Winners
Persian	4 - The Underdogs
Hindi	4 - The Underdogs
Indonesian	3 - The Rising Stars
Urdu	3 - The Rising Stars
Hausa	2 - The Hopefuls

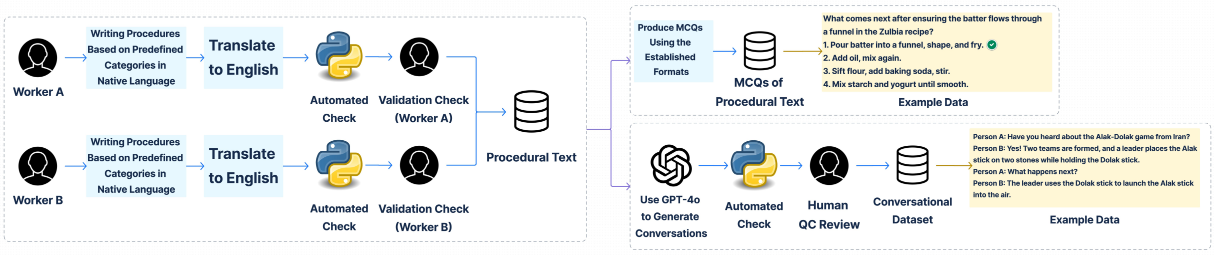
Table 2: Resource availability of languages in **CAPTex**

To address our research objectives, we introduce **CAPTex**, a dataset that incorporates cultural procedural texts across English and seven additional languages from diverse linguistic and geographical backgrounds. **CAPTex** is built upon three foundational components: (1) a curated collection of procedures spanning ten unique categories, (2) a series of thoughtfully crafted multiple-choice questions designed to evaluate the comprehension of each procedure, and (3) a rich corpus of conversational exchanges offering clarifications on the corresponding procedures.

3.1 Data Construction

Language Coverage Contemporary large language models demonstrate impressive performance in languages with abundant training data; however, their capabilities diminish when applied to low-resource languages and intricate cultural contexts, thereby constraining their global applicability (Rasheed et al., 2025). To promote linguistic diversity, we selected languages—Chinese (Mandarin), Japanese, Persian, Hindi, Indonesian, Urdu, and Hausa—representing a spectrum of resource availability, as measured by the criteria established by Joshi et al. (2020). Table 2 presents a detailed summary of the resource availability for each language featured in **CAPTex**.

Topic Taxonomies The procedures in **CAPTex** span ten culturally significant domains: (1) food and cuisine, (2) celebrations and festivals, (3) social etiquette and hospitality, (4) craftsmanship and artisan skills, (5) traditional attire and dress, (6) agricultural and seasonal practices, (7) religious and spiritual practices, (8) life milestones and family rites, (9) sports, games, and competitions, and (10) environmental and nature-based practices. These categories were selected by adapting and expanding upon the taxonomy from `IndoCulture` (Koto et al., 2024b) to ensure comprehensive coverage of cultural traditions and everyday practices. Please refer


Worker A
Writing Procedures Based on Predefined Categories in Native Language
Translate to English
Automated Check
Validation Check (Worker A)
Worker B
Writing Procedures Based on Predefined Categories in Native Language
Translate to English
Automated Check
Validation Check (Worker B)
Procedural Text
Produce MCQs Using the Established Formats
MCQs of Procedural Text
Example Data
Use GPT-4o to Generate Conversations
Automated Check
Human QC Review
Conversational Dataset
Example Data

What comes next after ensuring the batter flows through a funnel in the Zulbia recipe?
1. Pour batter into a funnel, shape, and fry. ✓
2. Add oil, mix again.
3. Sift flour, add baking soda, stir.
4. Mix starch and yogurt until smooth.

Person A: Have you heard about the Alak-Dolak game from Iran?
Person B: Yes! Two teams are formed, and a leader places the Alak stick on two stones while holding the Dolak stick.
Person A: What happens next?
Person B: The leader uses the Dolak stick to launch the Alak stick into the air.

Figure 1: End-to-End process of dataset creation

to Appendix for more detailed topic taxonomies.

Writing Procedural Text As shown in Figure 1, for each language, we employed two native speakers from each target nation to manually write procedural texts in their native language along with their English translations. These workers have deep cultural ties, having spent their entire lives in their native lands, ensuring strong familiarity with local traditions. The only exceptions are two individuals who lived in their home countries for the first 25 years before moving abroad, but have been residing outside their homeland for fewer than five years.[2]

Given a specific topic or category, each worker manually wrote procedural texts in their native language and translated them into English. They were strictly prohibited from using AI-based text generation tools but were allowed to reference reliable literature to verify cultural details and improve the accuracy of their writing. In total, we produce 1,400 human-written procedural texts (100 texts per language × 7 native languages × 2, including English translations).

Quality Control We ensure the high quality of `CAPTex` through two quality checks. First, we conduct an automated check to verify that each procedure consists of 5–10 steps and that the step count aligns between the native language and its English translation. Second, we perform a manual review by having workers cross-check procedural texts written by their peers. This evaluation follows a detailed checklist assessing conceptual accuracy, cultural relevance, logical progression, step order, grammatical correctness, and the accuracy and consistency of English translations. Any issues identified during the manual review are addressed by having the original worker revise their text accordingly. A common challenge arises from the lack of logical progression, which requires consolidating certain steps or eliminating those that do not impact the sequence. This restructuring makes the order of the remaining steps critical, as they become inherently non-interchangeable. For a more detailed description of the data collection process and quality control measures, refer to Appendix A.2.

3.2 Task Formulation

Using `CAPTex`, we developed four tasks, each designed with specific objectives. The following sections provide a detailed explanation of these tasks and their intended goals.

Task 1: Step Reordering In this task, procedural steps are initially shuffled and labeled with sequential letters (such as A, B, C, D). The model is then tasked with reconstructing the correct sequence, outputting a comma-separated list of these letters without any additional explanation.

To gauge performance, we utilize three established metrics. The first, Spearman's rank correlation (Spearman, 1904), examines the monotonic relationship between the predicted and actual rankings. The second, Levenshtein distance (Levenshtein, 1965), measures the minimum number of edit operations needed to transform the predicted sequence into the correct one. Lastly, Kendall's Tau rank correlation (Kendall, 1938) assesses the ordinal agreement between the predicted and true sequence by counting pairwise swaps. This assessment is performed in both English and a native language, ensuring a thorough evaluation of the model's ability to generalize across different linguistic environments.

Task 2: Procedure-Based Multiple-Choice Questions (PB-MCQ) We design a comprehensive multiple-choice question (MCQ) framework for each procedure to evaluate mLLMs' comprehension and reasoning abilities in identifying both sub-

[2] Each worker is compensated fairly based on a five-day workload, with payments aligned with the minimum wage in their respective country.

sequent and preceding steps. We created affirmative and negative versions for each question type, constructing them in the original language as well as in English. This approach ensured consistency across all question types (Subsequent Affirmative, Subsequent Negative, Antecedent Affirmative, Antecedent Negative) in both languages. To maintain linguistic parity, we construct all questions in both the original language and English. Each question consists of four answer choices, with one correct option. As a result, we generate eight MCQs per procedure, leading to a total of 5,600 questions (4 question types × 100 procedures × 14 languages, including native languages and their English translations).

In our question formulation, we ensured that for queries about upcoming steps, the correct answer was the next step, while three incorrect options were randomly chosen from earlier steps. Likewise, for questions regarding previous steps, the correct response was the preceding step, with three incorrect choices randomly selected from later steps. We confirmed that each question had one correct answer and three incorrect options. An illustrative example of MCQs is provided in Table 7. For additional information about MCQs, please refer to appendix A.3

The language model is prompted to generate only the correct choice option (A, B, C, or D) as its output. The primary evaluation metric for assessing model performance in this task is accuracy, measured by the proportion of correctly selected answers. Beyond evaluating the cultural understanding and procedural reasoning of language models, this task also enables an analysis of how well mLLMs comprehend affirmative and negative questions. Additionally, it assesses the model's ability to predict both the subsequent and the antecedent procedural steps, further refining our understanding of its reasoning and contextual awareness.

Task 3: Conversation-Based Multiple-Choice Questions (CB-MCQ) To evaluate the reasoning capabilities of large language models (LLMs) using procedural text, we created procedurally grounded conversations in English. We utilized GPT-4o along with a specially designed prompt (detailed in Appendix A.4) to generate a series of four-utterance dialogues, conditioned on `CAPTex`. These dialogues simulate a natural conversation between two individuals, referred to as Person A and Person B. In the conversation, Person A starts by asking about a particular procedure. Person B, acting as a knowledgeable respondent, introduces the procedure and describes the initial steps involved. Person A then asks a follow-up question to clarify the next step in the process, and Person B provides a detailed explanation in response. Table 8 presents an example of the generated conversations.

Based on the conversation structure, the third utterance presents a question regarding the next step of the process. However, the fourth utterance, which contains the explanation of that step, is intentionally omitted. The language model must then select the correct answer from four options corresponding to the missing explanation, using the same answer choices as PB-MCQ. Comparing PB-MCQ and CB-MCQ allows us to assess whether models perform better in structured question-answering or conversational reasoning. PB-MCQ evaluates direct procedural knowledge, while CB-MCQ tests inference within dialogue, providing insights into model adaptability across different task formats.

For quality control, we employ a native speaker to review each dialogue assessing conceptual accuracy, grammatical correctness, and the consistency of Person B's explanations with the actual steps outlined in the procedure. If any errors are identified, the worker manually corrects them to maintain accuracy and coherence. Similar to PB-MCQ, we use accuracy as the evaluation metric for CB-MCQ.

Task 4: Conversation-Based Question Answering (CB-QA) This task is similar to CB-MCQ but differs in that it requires the model to generate a response as if it were the conversational participant, rather than selecting from predefined choices. This task is chosen over CB-MCQ to evaluate the model's ability to produce natural, contextually appropriate responses, reflecting a deeper understanding of procedural knowledge in dialogue. The performance of mLLMs is evaluated using three metrics: ROUGE-L score (Lin, 2004), BERTScore (Zhang et al., 2020), and additional semantic similarity score (Corley and Mihalcea, 2005), which quantifies meaning alignment between generated and reference responses.

3.3 Data Statistics

For each country and category, `CAPTex` incorporates 10 procedures, totaling 100 distinct procedures per country. Figure 2 represents the distribution of procedural steps across countries, with num-

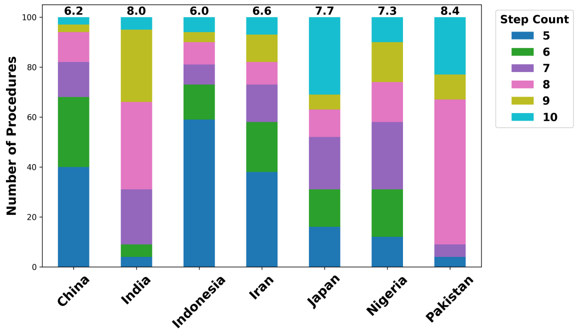


Figure 2: Procedures step counts by country

bers above each bar indicating the average number of steps.

The methodology guarantees an equitable distribution of multiple-choice questions (MCQs) across various countries, categories, and formats. For each designated format[3] a single question is crafted. This approach results in the creation of 2,800 unique questions in the original language and an additional 2,800 in English, culminating in a total of 5,600 MCQs. Tables 3 and 6 present the mean lexical density of **CAPTex**, analyzed across countries and categories, respectively.

The conversations dataset is thoroughly balanced, with one conversation constructed for each procedure. This approach yielded a total of 700 English conversations.

4 Experiments

4.1 Setup

We assessed 31 multilingual language models of different sizes, including DeepSeek (DeepSeek-AI et al., 2025), Gemma-2 (Team et al., 2024), Llama-3 (Grattafiori et al., 2024), Mamba (Gu and Dao, 2024), Mistral (Jiang et al., 2023), Qwen2.5 (Qwen et al., 2025), BLOOMZ (Muennighoff et al., 2023), Aya-Expanse (Dang et al., 2024), mT0 (Muennighoff et al., 2023), GPT-4 (OpenAI et al., 2024), and O3-mini (OpenAI, 2025).

We conducted zero-shot evaluations using prompt templates exclusively in English. Prior research has shown that prompting in different languages can lead to variations in responses to similar queries (Lin et al., 2022; Shen et al., 2024b). Moreover, studies on multilingual LLMs have consistently found that these models tend to perform better when prompted in English rather than in other languages (Muennighoff et al., 2023; Ozsoy, 2024; Koto et al., 2024a).

4.2 Results and Analysis

Table 4 provides a comprehensive overview of model performance across all four tasks, consistently highlighting GPT-4o as the top-performing model. Among open-weight models, Gemma-2-9b-it outperforms others in the reordering task, while Qwen2.5-14B-Instruct achieves the highest accuracy in MCQ tasks, and Qwen2.5-7B demonstrates the strongest performance in the CB-QA task.

Among models of comparable size, Qwen2.5 demonstrates superior performance relative to its counterparts. Notably, Mamba exhibits significantly weaker performance in procedural text comprehension compared to transformer-based models. Our findings indicate that increasing the number of parameters within the same model family generally enhances performance. Additionally, for language models with an available Instruct variant, the Instruct versions consistently achieve higher performance—except in the cases of Gemma-2-2B and Qwen2.5-1.5B.

We calculate Kendall rank correlation scores across four tasks to assess task sensitivity when comparing LLMs. Our analysis shows strong correlations (0.8–0.9) between reordering, PB-MCQ, and CB-MCQ, indicating that these tasks rank models similarly. However, CB-QA (the generation task) has a lower correlation (0.4–0.5) with the others, suggesting that text generation captures different aspects of procedural reasoning and is a valuable addition to the evaluation.

Analysis of PB-MCQ Subtypes Table 5 showcases the performance of the top models across PB-MCQ task for four distinct question types: The results reveal that antecedent affirmative (AA) questions are the easiest for language models, while subsequent affirmative (SA) questions are the most challenging. This suggests that models find it easier to reason about preceding steps than following ones. Interestingly, while prior studies (Truong et al., 2023; She et al., 2023; Kassner and Schütze, 2020) indicate that negation generally weakens model performance in NLP tasks, this pattern does not hold in the procedural context. Subsequent negative (SN) leads to better performance, whereas affirmative negative (AN) questions result in a notable decline in accuracy.

Language Effects on Performance We examine the influence of language on the performance of Qwen2.5-14B-Instruct, the top-performing open-

[3] Subsequent Affirmative, Subsequent Negative, Antecedent Affirmative, and Antecedent Negative

Country	Procedures	MCQ (English/Native)			Conversations (Utterances)			
	(English/Native)	Question	Correct Ans.	Incorrect Ans.	First	Second	Third	Fourth
China	157.8 / 138.8	38.0 / 39.8	24.5 / 21.2	26.0 / 22.8	9.4	46.2	13.1	34.3
India	37.5 / 50.4	16.5 / 21.8	4.8 / 6.3	4.7 / 6.3	9.3	37.2	12.7	23.6
Indonesia	121.9 / 98.9	39.1 / 31.9	19.5 / 15.6	20.1 / 16.4	14.8	41.1	12.4	30.1
Iran	144.5 / 166.4	34.3 / 40.3	21.6 / 24.7	22.4 / 25.8	10.0	44.7	13.2	30.2
Japan	120.6 / 147.1	30.6 / 45.5	16.2 / 19.6	15.8 / 19.2	12.3	43.8	13.1	29.5
Nigeria	162.3 / 173.3	35.8 / 38.1	21.1 / 22.8	21.8 / 23.5	11.0	45.7	13.2	29.6
Pakistan	76.6 / 99.7	23.9 / 36.4	9.2 / 12.0	9.4 / 12.2	12.0	42.0	12.9	27.0

Table 3: Average word counts for **CAPTex** components (Procedures, MCQs, and Conversations) by country.

Model	Reordering			PB-MCQ	CB-MCQ	CB-QA		
	$\rho \uparrow$	LD$\downarrow$	$\tau \uparrow$			R-L	BS	SS
Random	0.00	5.56	0.00	0.25	0.25	0.00	0.00	0.00
DeepSeek-R1(Distill-Llama-8B)	0.30	5.48	0.14	0.27	0.38	0.20	0.53	0.48
DeepSeek-R1(Distill-Qwen-14B)	0.43	4.47	0.33	0.42	0.54	0.22	0.55	0.50
Gemma-2-2b	0.30	5.49	0.14	0.27	0.30	0.23	0.58	0.55
Gemma-2-2b-it	0.21	5.48	0.10	0.08	0.26	**0.25**	0.55	0.48
Gemma-2-9b	0.48	4.94	0.32	0.37	0.57	0.21	0.53	0.47
Gemma-2-9b-it	**0.75**	**3.09**	**0.66**	0.43	0.46	0.15	0.50	0.42
Llama-3.1-8B	0.30	5.49	0.14	0.28	0.32	0.24	**0.59**	0.55
Llama-3.1-8B-Instruct	0.43	4.69	0.33	0.37	0.48	0.24	**0.59**	0.55
Llama-3.2-1B	0.29	5.56	0.12	0.25	0.27	0.22	0.57	0.52
Llama-3.2-1B-Instruct	0.31	5.44	0.15	0.26	0.34	0.23	0.58	0.54
Llama-3.2-3B	0.29	5.56	0.12	0.26	0.28	0.23	0.58	0.54
Llama-3.2-3B-Instruct	0.25	5.35	0.14	0.33	0.36	0.24	**0.59**	0.54
Mamba-1.4b-hf	0.30	5.59	0.14	0.00	0.00	0.13	0.39	0.25
Mamba-2.8b-hf	0.25	5.51	0.14	0.00	0.00	0.12	0.39	0.26
Mistral-7B-Instruct-v0.2	0.51	4.60	0.40	0.34	0.48	0.23	0.58	0.54
Mistral-7B-v0.3	0.19	5.45	0.11	0.32	0.31	0.24	0.58	0.54
Mistral-7B-Instruct-v0.3	0.38	4.75	0.30	0.37	0.41	0.24	**0.59**	0.55
Mistral-Nemo-Base-2407	0.33	5.34	0.18	0.34	0.50	0.20	0.53	0.43
Mistral-Nemo-Instruct-2407	0.43	4.59	0.34	0.39	0.57	0.21	0.53	0.43
Qwen2.5-1.5B	0.38	5.15	0.26	0.31	0.34	0.23	0.58	0.55
Qwen2.5-1.5B-Instruct	0.42	4.82	0.32	0.33	0.21	0.23	0.58	0.54
Qwen2.5-7B	0.63	3.94	0.52	0.45	0.54	**0.25**	**0.59**	**0.57**
Qwen2.5-7B-Instruct	0.69	3.65	0.60	0.50	0.60	0.24	**0.59**	0.56
Qwen2.5-14B	0.70	3.33	0.62	0.48	0.64	0.26	**0.59**	0.55
Qwen2.5-14B-Instruct	0.72	3.21	0.64	**0.56**	**0.70**	0.24	**0.59**	0.56
Aya-Expanse-8b	0.47	4.71	0.37	0.39	0.43	0.24	0.58	0.54
Bloomz-560m	0.37	5.39	0.22	0.15	0.00	0.14	0.45	0.32
Bloomz-7b1	0.29	5.48	0.13	0.09	0.20	0.17	0.50	0.38
mT0-xxl	0.29	5.54	0.13	0.35	0.27	0.14	0.50	0.43
GPT-4o	**0.81**	**2.38**	**0.75**	0.58	**0.74**	**0.29**	**0.62**	**0.60**
O3-mini	0.78	2.54	0.72	**0.66**	0.65	0.27	0.61	0.59

Table 4: Models' performance across tasks. Metrics include Spearman's Rank Correlation (ρ) [-1,1], Levenshtein Distance (LD) [0,∞], and Kendall's Tau Rank Correlation (τ) [-1,1] for Reordering; accuracy [0,1] for PB-MCQ and CB-MCQ; and ROUGE-F1 [0,1], BERT-F1 [0,1], and Semantic Similarity (SS) [0,1] for CB-QA. Higher values indicate better performance for all metrics except LD, where lower is better.

weight model, in the reordering and PB-MCQ tasks. To quantify this effect, we first normalize the evaluation metrics for these tasks and then compute an aggregated score using a weighted sum approach.[4] As depicted in Figure 3, with the exception of China, Qwen2.5-14B-Instruct generally

[4] The weighted sum assigns 60% of the total score to the reordering task (20% for each metric), with Levenshtein Distance (LD) normalized as $1-(L/L_{\max})$ to ensure higher values indicate better performance. The remaining 40% is assigned to PB-MCQ accuracy.

Model	PB-MCQ			
	SA	SN	AA	AN
Gemma-2-9b-it	0.31	0.51	0.56	0.36
Llama-3.1-8B-Instruct	0.24	0.38	0.55	0.33
Mistral-Nemo-Instruct-2407	0.29	0.34	0.60	0.34
Qwen2.5-14B-Instruct	0.45	0.53	0.69	0.55
Aya-Expanse-8b	0.33	0.38	0.61	0.24
GPT-4o	0.46	0.57	**0.73**	0.54
O3-mini	**0.51**	**0.70**	0.72	**0.72**

Table 5: Model performance on PB-MCQ across question types. "SA", "SN", "AA", and "AN" denote Subsequent Affirmative, Subsequent Negative, Antecedent Affirmative, and Antecedent Negative, respectively.

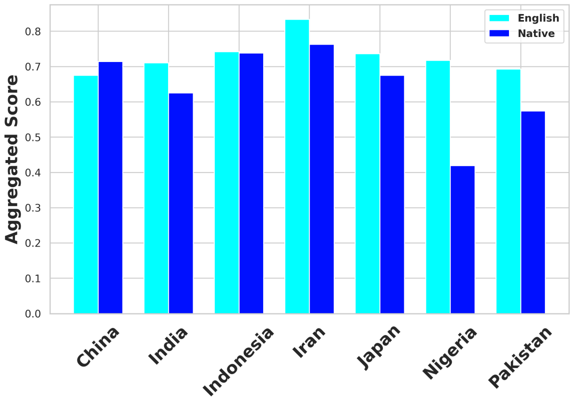


Figure 3: Language impact on Qwen2.5-14B-Instruct performance

outperforms in English across other countries, especially in low-resource languages like Hausa and Urdu. This divergence may stem from the model's extensive proficiency in English, whereas the linguistic nuances, idiomatic expressions, and procedural reasoning structures inherent to Chinese contexts might be underrepresented in the training data.

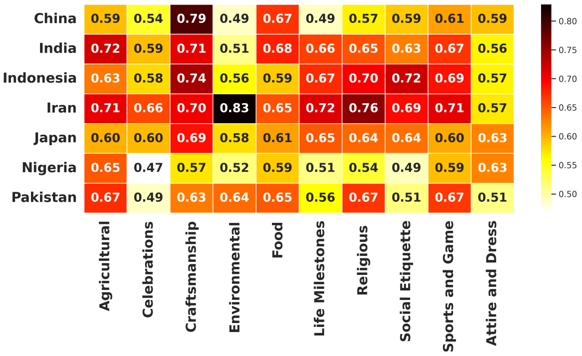


Figure 4: Cultural dimension performance by country

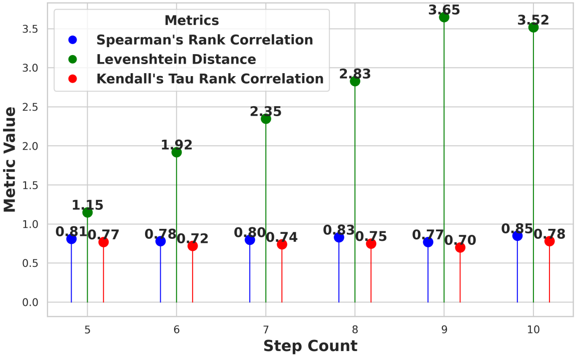


Figure 5: Impact of step count on GPT-4o reordering

Performance Across Cultural Dimensions Figure 4 presents the aggregated performance scores of the top-performing open-weight model, Qwen2.5-14B-Instruct, across all evaluation tasks. Scores are normalized and combined using a weighted sum approach, with task-specific weights designed to better reflect the varying cognitive demands and real-world relevance of each evaluation. Specifically, reordering is given a higher weight (30%) due to its requirement for holistic sequence reconstruction, encompassing temporal logic, global coherence, and dependency tracking across steps. In contrast, PB-MCQ and CB-MCQ (each weighted at 20%) test narrower, one-step reasoning skills and are thus weighted slightly lower. CB-QA is assigned a higher weight (30%) as it evaluates open-ended generative capabilities, probing the model's contextual understanding, fluency, and ability to engage in realistic procedural dialogue. Notably, its lower correlation with the other tasks highlights its role in capturing distinct aspects of procedural reasoning. This weighted aggregation offers a more nuanced and representative measure of procedural competence than a simple average.

Our analysis reveals that Qwen2.5-14B-Instruct's knowledge of procedural texts varies across cultural domains. For instance, it demonstrates strong familiarity with Indian agriculture but performs less effectively on Chinese agricultural topics. Conversely, for craftsmanship and artisan skills, Qwen2.5-14B-Instruct encodes Chinese cultural knowledge better than other countries. In the food category, Indian cuisine is better represented, while Iranian religious practices appear more prominently. Additionally, Indonesian social etiquette are well captured by Qwen2.5-14B-Instruct, suggesting variation in how different cultural aspects are reflected in the model.

Impact of Procedure Length on Ordering Figure 5 shows that as procedure length increases, Levenshtein distance rises, indicating greater reordering difficulty. However, Spearman's rank and Kendall's Tau correlations remain high, suggesting that GPT-4o generally preserves step order despite complexity. Shorter procedures introduce more ambiguity, making errors more impactful, while longer sequences benefit from stronger local dependencies, aiding order retention. Notably, procedures with ten steps achieve the highest rank correlations, indicating that structural cues in longer sequences enhance model performance. Most errors involve minor swaps rather than complete misordering.

5 Conclusion

We introduce **CAPTex**, a benchmark for evaluating mLLMs' ability to process culturally diverse procedural texts across seven languages. Our findings show that model performance varies across cultural domains, with greater challenges in tasks requiring implicit cultural knowledge, such as environmental practices, while structured domains like craftsmanship are better handled. Multiple-choice tasks in conversational contexts improve reasoning, while generation evaluation highlights gaps in procedural text comprehension. As an important benchmarking tool, **CAPTex** helps identify specific areas where mLLMs need improvement in handling culturally diverse procedural text. It thereby establishes a crucial foundation for guiding future advancements towards developing more robust, culturally aware, and effective mLLMs capable of complex procedural reasoning across a wide range of languages and cultural contexts.

Limitations

Our study provides valuable insights into the performance of mLLMs on procedural texts, but there are a few limitations to consider. Firstly, our research is limited to textual data and does not include multimodal inputs, such as procedural texts with images. Incorporating images would enhance model understanding, but due to the added complexity, this is reserved for future work.

Additionally, our dataset focuses on seven countries, primarily due to budget constraints. While this may seem limited, the selected countries offer diverse language categories and varying resource levels, ensuring a meaningful analysis of cultural gaps in mLLMs. These findings can be generalized to a broader context, given the representativeness of the samples.

Finally, the conversation dataset consists of exchanges with exactly four utterances. While real-world dialogues are typically longer and more dynamic, this limitation was made for practical reasons. Despite the brevity, we found that mLLMs still struggle with maintaining coherence and understanding conversational flow, underscoring the challenges these models face, even in simpler dialogue settings.

Acknowledgments

We would like to thank the workers and reviewers whose contributions were essential in creating **CAPTex**. Their dedication in annotating data and providing valuable feedback greatly enhanced the quality of this work.

References

Aida Amini, Antoine Bosselut, Bhavana Dalvi Mishra, Yejin Choi, and Hannaneh Hajishirzi. 2020. Procedural reading comprehension with attribute-aware context flow. *Preprint*, arXiv:2003.13878.

Kushal Arora, Layla El Asri, Hareesh Bahuleyan, and Jackie Cheung. 2022. Why exposure bias matters: An imitation learning perspective of error accumulation in language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 700–710, Dublin, Ireland. Association for Computational Linguistics.

Yong Cao, Yova Kementchedjhieva, Ruixiang Cui, Antonia Karamolegkou, Li Zhou, Megan Dare, Lucia Donatelli, and Daniel Hershcovich. 2024. Cultural adaptation of recipes. *Transactions of the Association for Computational Linguistics*, 12:80–99.

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.

Wei Dai, Jionghao Lin, Hua Jin, Tongguang Li, Yi-Shan Tsai, Dragan Gašević, and Guanliang Chen. 2023. Can large language models provide feedback to students? a case study on chatgpt. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, pages 323–325. IEEE.

Bhavana Dalvi, Lifu Huang, Niket Tandon, Wen-tau Yih, and Peter Clark. 2018. Tracking state changes in procedural text: a challenge dataset and models for process paragraph comprehension. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1595–1604, New Orleans, Louisiana. Association for Computational Linguistics.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

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*.

Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, Sheer El Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Scott Johnston, Andy Jones, Nicholas Joseph, Jackson Kernian, Shauna Kravec, Ben Mann, Neel Nanda, Kamal Ndousse, Catherine Olsson, Daniela Amodei, Tom Brown, Jared Kaplan, Sam McCandlish, Christopher Olah, Dario Amodei, and Jack Clark. 2022. Predictability and surprise in large generative models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1747–1764, New York, NY, USA. Association for Computing Machinery.

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, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth

Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun

Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Albert Gu and Tri Dao. 2024. Mamba: Linear-time sequence modeling with selective state spaces. *Preprint*, arXiv:2312.00752.

Aditya Gupta and Greg Durrett. 2019. Tracking discrete and continuous entity state for process understanding. In *Proceedings of the Third Workshop on Structured Prediction for NLP*, pages 7–12, Minneapolis, Minnesota. Association for Computational Linguistics.

Hao Huang, Xiubo Geng, Jian Pei, Guodong Long, and Daxin Jiang. 2021. Reasoning over entity-action-location graph for procedural text understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5100–5109, Online. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Hanlei Jin, Yang Zhang, Dan Meng, Jun Wang, and Jinghua Tan. 2024. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. *arXiv preprint arXiv:2403.02901*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

M. G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Thomas Kosch and Sebastian Feger. 2024. Risk or chance? large language models and reproducibility in hci research. *Interactions*, 31(6):44–49.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024a. ArabicMMLU: Assessing massive multitask language understanding in Arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5622–5640, Bangkok, Thailand. Association for Computational Linguistics.

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

Vladimir I. Levenshtein. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10:707–710.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ao Liu, Shuai Yuan, Chenbin Zhang, Congjian Luo, Yaqing Liao, Kun Bai, and Zenglin Xu. 2020. Multi-level multimodal transformer network for multimodal recipe comprehension. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1781–1784, New York, NY, USA. Association for Computing Machinery.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

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.

OpenAI. 2025. Openai o3-mini system card.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Makbule Gulcin Ozsoy. 2024. Multilingual prompts in llm-based recommenders: Performance across languages. *Preprint*, arXiv:2409.07604.

Vinodkumar Prabhakaran, Rida Qadri, and Ben Hutchinson. 2022. Cultural incongruencies in artificial intelligence. *Preprint*, arXiv:2211.13069.

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

Hossein Rajaby Faghihi and Parisa Kordjamshidi. 2021. Time-stamped language model: Teaching language models to understand the flow of events. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4560–4570, Online. Association for Computational Linguistics.

Hanoona Rasheed, Muhammad Maaz, Abdelrahman Shaker, Salman Khan, Hisham Cholakal, Rao M Anwer, Tim Baldwin, Michael Felsberg, and Fahad S Khan. 2025. Palo: A polyglot large multimodal model for 5b people. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1745–1754. IEEE.

Jingyuan S. She, Christopher Potts, Samuel R. Bowman, and Atticus Geiger. 2023. ScoNe: Benchmarking negation reasoning in language models with fine-tuning and in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1803–1821, Toronto, Canada. Association for Computational Linguistics.

Huangjun Shen, Liangying Shao, Wenbo Li, Zhibin Lan, Zhanyu Liu, and Jinsong Su. 2024a. A survey on multi-modal machine translation: Tasks, methods and challenges. *Preprint*, arXiv:2405.12669.

Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. 2024b. The language barrier: Dissecting safety challenges of LLMs in multilingual contexts. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2668–2680, Bangkok, Thailand. Association for Computational Linguistics.

C. Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Margaret S. Steffensen, Chitra Joag-Dev, and Richard C. Anderson. 1979. A cross-cultural perspective on reading comprehension. *Reading Research Quarterly*, 15(1):10–29.

Jizhi Tang, Yansong Feng, and Dongyan Zhao. 2020. Understanding procedural text using interactive entity networks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7281–7290, Online. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. 2023. Language models are not naysayers: an analysis of language models on negation benchmarks. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 101–114, Toronto, Canada. Association for Computational Linguistics.

Man-Fai Wong, Shangxin Guo, Ching-Nam Hang, Siu-Wai Ho, and Chee-Wei Tan. 2023. Natural language generation and understanding of big code for ai-assisted programming: A review. *Entropy*, 25(6).

Te-Lin Wu, Alex Spangher, Pegah Alipoormolabashi, Marjorie Freedman, Ralph Weischedel, and Nanyun Peng. 2022. Understanding multimodal procedural knowledge by sequencing multimodal instructional manuals. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4525–4542, Dublin, Ireland. Association for Computational Linguistics.

Yiran Wu, Tianwei Yue, Shaokun Zhang, Chi Wang, and Qingyun Wu. Stateflow: Enhancing llm task-solving through state-driven workflows. In *NeurIPS 2024 Workshop on Open-World Agents*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Dataset

A.1 Detailed Categories

1. **Food and Cuisine**: Captures cultural practices related to the preparation and consumption of food.

2. **Celebrations and Festivals**: Represents rituals and activities associated with cultural events and festivities.

3. **Social Etiquette and Hospitality**: Reflects societal norms and traditions in interpersonal interactions.

4. **Craftsmanship and Artisan Skills**: Showcases traditional methods of creating culturally significant artifacts.

5. **Traditional Attire and Dress**: Highlights practices involving culturally significant clothing and adornments.

6. **Agricultural and Seasonal Practices**: Documents cultural procedures tied to agricultural cycles and seasonal changes.

7. **Religious and Spiritual Practices**: Encapsulates practices integral to religious and spiritual traditions.

8. **Life Milestones and Family Rites**: Represents cultural customs marking significant life events.

9. **Sports, Games, and Competitions**: Covers traditional recreational and competitive activities.

10. **Environmental and Nature-Based Practices**: Focuses on cultural interactions with and stewardship of the natural environment.

A.2 Procedures Collection

In the initial phase, we conducted a pilot study using Prolific [5] to recruit workers and gather some procedures. However, upon evaluating the quality of the collected procedures, we determined that Prolific was unsuitable for this specific task. Consequently, we opted to directly engage reliable workers who met our stringent requirements. Native speakers of the target languages were responsible for crafting each procedure in their native language. These were then precisely translated into English, ensuring both versions accurately conveyed the same content. Each procedure consisted of five to ten sequential steps, where the order was crucial for the proper understanding and execution of the tasks. To preserve the dataset's authenticity, workers were strictly prohibited from using AI-based text generation tools, which could introduce inaccuracies or fabrications, thereby compromising the reliability of the procedures. We implemented a two-step quality control process, comprising automated checks and peer evaluations through cross-verification. An all-inclusive checklist was employed to assess conceptual precision, cultural authenticity, logical flow, the necessity of maintaining step sequence, grammatical accuracy, and adherence to the required number of steps. This rigorous process ensured that both the native language and English versions met our high standards.

A.3 MCQ Design and Structure

As previously outlined, four distinct categories of questions are systematically generated using Python code. These questions adhere to the following templates for inquiries in English:

- Subsequent Affirmative: *In the procedure “Procedure Name”, what is the next step after: “Reference Step”?*

[5] http://prolific.com/

- Subsequent Negative: *In the procedure "Procedure Name", which one is not the step before: "Reference Step"?*

- Antecedent Affirmative: *In the procedure "Procedure Name", which step must be completed before: "Reference Step"?*

- Antecedent Negative: *In the procedure "Procedure Name", which step does not come after: "Reference Step"?*

These templates ensure clarity and consistency in question generation, aligning with the procedural framework.

A.4 Conversation Generation

The following prompt is used to generate conversations using GPT-4o: *Create a short conversation between two people, Person A and Person B, based on the following procedure. The conversation should begin with (Have you heard about the "Procedure Name" from "Country"?) and progress naturally, with each message reflecting a clear and logical flow of ideas related to the steps of the procedure. The conversation should consist of four messages. In the third message, Person A asks a question about the next step in the procedure. In the last message, Person B should respond according to the procedure's step, providing a clear answer or action related to the next step, and explicitly mention the step number.*

Example Structure:
Person A: Have you heard about the "Procedure Name" from "Country"?
Person B: (Response introducing the procedure and discussing some of the first steps)
Person A: (Asks a follow-up question to clarify the next step)
Person B: (Explains the first next step in the procedure according to the procedure's step)
Next step: (number of next step)

IMPORTANT: Ensure the generated conversation adheres to the above structure. The last message should always be "Next step: (number)", where (number) is the next step in the procedure.

The procedure is as follows:
"Procedure Steps"

Given the suboptimal performance of GPT-4o in processing lower-resource languages, this section is exclusively dedicated to generating conversations in English.

The generated conversations are subjected to a two-phase verification process. The initial phase involves automated verification using a Python script, ensuring that each conversation comprises exactly four utterances. The subsequent phase entails thorough evaluation by qualified human annotators. These annotators assess each dialogue against a detailed checklist, which evaluates conceptual accuracy, grammatical precision, and alignment of Person B's responses with the procedural steps. During the review process, two types of errors were identified: 76 conversations had an incorrect next step number, and 45 conversations included explanations of multiple steps in the final utterance.

B Additional Results

Figure 8 presents the aggregated performance[6] of the models across all four tasks.

Language Effects on Performance As shown in Figure 6, the results for GPT-4o, the top-performing model, are presented in a manner similar to those for Qwen2.5-14B-Instruct in Figure 3. GPT-4o generally performs better in English across most countries, with the exceptions of China and Indonesia. A comparison of the two figures reveals that GPT-4o maintains a more balanced performance between English and native languages in most countries. While GPT-4o tends to perform better in English, reflecting its extensive training on high-resource languages, it displays a smaller performance gap in low-resource languages such as Hausa (Nigeria) and Urdu (Pakistan). In contrast, Qwen2.5-14B-Instruct shows a more pronounced decline in native languages, indicating that GPT-4o may possess superior cross-lingual capabilities and stronger support for languages with limited resources. This gives GPT-4o a distinct advantage in multilingual contexts, while Qwen2.5-14B-Instruct shows a stronger preference for English, particularly in regions with fewer linguistic resources.

Performance Across Cultural Dimensions The aggregated performance scores of GPT-4o, the top-performing model, are displayed in Figure 7 for all evaluation tasks. Our analysis reveals that GPT-4o's understanding of procedural texts differs

[6] Scores are normalized and computed using a weighted sum approach, with the following weight distribution: 0.3 for reordering, 0.1 for each metric, 0.2 for PB-MCQ, 0.2 for CB-MCQ, and 0.3 for CB-QA.

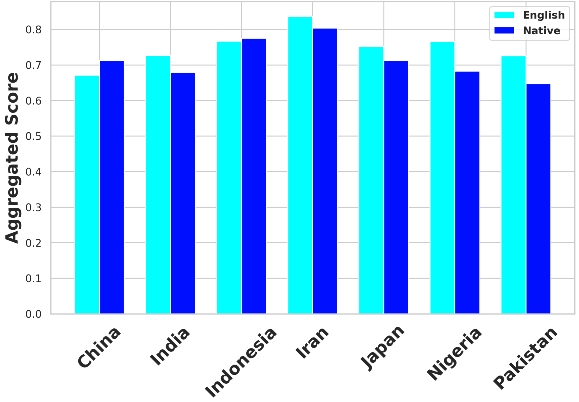


Figure 6: Language impact on GPT-4o performance

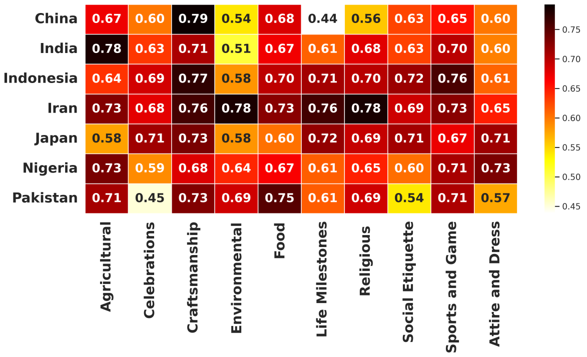


Figure 7: Cultural dimension performance by country

across cultural contexts. Specifically, the model demonstrates robust familiarity with Indian agricultural practices, while its performance on Japanese agricultural topics is comparatively weaker. In contrast, GPT-4o excels in encoding Japanese cultural knowledge related to celebrations and festivals, surpassing its representations of other cultures. In the culinary domain, Pakistani cuisine is more accurately captured, whereas Iranian religious practices are more prominently reflected. Furthermore, the model effectively encapsulates Indonesian social etiquette, highlighting the diversity in how various cultural elements are represented within the model.

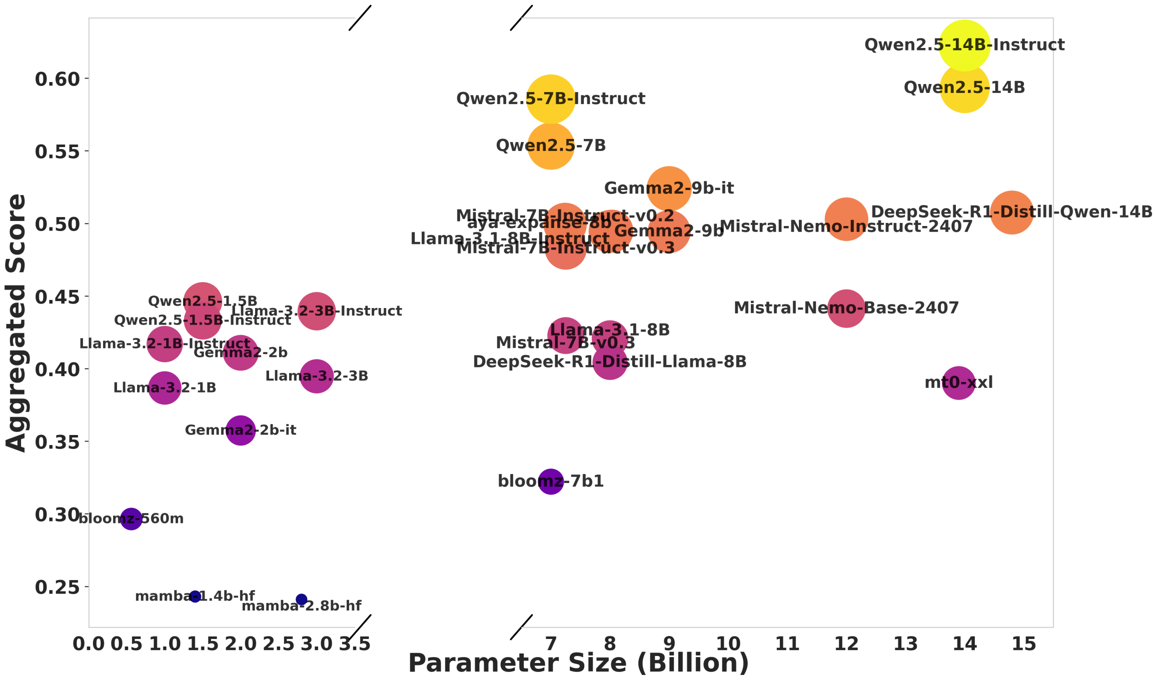


Figure 8: Performance of LLMs by model size

Category	Procedures	MCQs (English/Native)			Conversations (Utternaces)			
	(English/Native)	Question	Correct Answer	Wrong Answer	First	Second	Third	Fourth
Agricultural and Seasonal Procedures	94.6 / 104.1	26.8 / 32.7	13.6 / 14.6	13.4 / 14.3	10.4	41.1	12.7	28.5
Celebrations and Festivals	118.2 / 124.2	31.7 / 35.8	17.4 / 17.9	17.8 / 18.2	11.0	43.3	12.5	29.0
Craftsmanship and Artisan Skills	132.4 / 146.5	32.6 / 39.0	18.6 / 20.5	18.7 / 20.3	10.3	45.0	13.2	30.6
Environmental and Nature-Based Procedures	93.3 / 108.8	29.1 / 36.0	14.2 / 16.3	15.3 / 17.4	11.3	42.5	12.9	28.8
Food and Cuisine	137.8 / 139.9	32.2 / 36.3	18.2 / 18.6	18.5 / 18.5	10.7	43.2	13.1	28.6
Life Milestones and Family Rites	132.3 / 137.0	32.8 / 38.7	18.0 / 18.0	18.7 / 19.3	11.7	43.3	12.7	29.5
Religious and Spiritual Practices	105.1 / 109.7	30.4 / 34.4	15.2 / 15.3	15.7 / 16.0	12.1	41.4	12.5	28.0
Social Etiquette and Hospitality	115.0 / 125.2	32.8 / 38.7	17.2 / 18.6	18.5 / 19.7	13.0	43.5	13.3	29.6
Sports, Games, and Competitions	112.2 / 122.1	30.5 / 36.0	16.1 / 16.9	16.9 / 18.3	10.5	42.6	13.6	28.2
Traditional Attire and Dress	132.5 / 131.8	32.7 / 36.2	18.0 / 17.9	18.4 / 18.1	11.6	43.6	12.9	31.1

Table 6: Average word counts for `CAPTex` components (Procedures, MCQs, and Conversations) by category

Country	Category	Language	Type	Question	Choices
Japan	Food and Cuisine	Japanese	SA	手順 ’昆布巻きの作り方（4人前）’ では、’鍋に6と5の戻し汁を入れ、中火で沸騰させる。’ の後の次のステップは何ですか？	A. 1を昆布と戻し汁に分ける。昆布を端からしっかり巻き付けて爪楊枝で留める。 B. 4を巻き付け、きつく結んだら爪楊枝を抜き、余分なかんぴょうを切る。 C. 醤油大さじ1.5、顆粒和風だし大さじ1、料理酒大さじ1、みりん大さじ2、砂糖大さじ2、塩小さじ1/2を加える。 D. 塩を洗い流し、水気を切ったら20cm幅に切り、12本にする。
Japan	Food and Cuisine	English	SA	In the procedure ’How to Make Kelp Rolls (for 4 Persons)’, what is the next step after: ’Place the rolls into a saucepan and add the stock. Bring to a boil over medium heat.’?	A. Separate the kelp and stock. Wrap the kelp tightly around the filling and secure it with a toothpick. B. Wrap a piece of kanpyo around the roll and tie it tightly. Remove the toothpick and trim any excess kanpyo. C. Add 1.5 tbsp soy sauce, 1 tbsp granulated Japanese soup stock, 1 tbsp cooking sake, 2 tbsp mirin, 2 tbsp sugar, and 1/2 tsp salt. D. Rinse off the salt, drain, and cut the kanpyo into 20 cm lengths, then divide into 12 pieces.

Table 7: Example of MCQs

Country	Category	Conversation
Iran	Sports, Games, and Competitions	Person A: Have you heard about the Alak-Dolak Game from Iran? Person B: Yes, it's quite an interesting traditional game! It starts with forming two teams, and they draw lots to see which team gets to go first. Each team also has a leader who plays a crucial role at the beginning. For instance, the leader places two stones and then positions the "Alak" stick on top of them while holding the "Dolak" stick in their hand. Person A: That's fascinating! What does the leader do next after placing the Alak stick on the stones? Person B: After placing the Alak stick, the leader uses the Dolak stick by positioning its tip under the Alak stick. Then they move the Dolak stick left and right to launch the Alak stick into the air. Next step: 3.
Nigeria	Celebrations and Festivals	Person A: Have you heard about the Osun-Osogbo Festival from Nigeria? Person B: Yes, I have! It's a fascinating cultural event that spans two weeks, starting with cleansing rituals called Iwopopo. The festival begins by preparing the Arugba, a virgin maiden who carries a sacred calabash. She plays a vital role as she represents purity and connects the people with the goddess Osun. It's all part of cleansing the community of evil spirits. Person A: That sounds amazing. What happens after the cleansing rituals? Person B: After the cleansing, traditional Yoruba music, drumming, and dances are performed daily, highlighting the community's rich cultural heritage. Craftsmen and vendors set up in the marketplace near the sacred Osun Grove to display arts, crafts, and local foods. Next step: 2

Table 8: Examples of a procedurally grounded conversation