KALAHI: A handcrafted, grassroots cultural LLM evaluation suite for Filipino

Jann Railey Montalan[1,2], Jian Gang Ngui[1,2], Wei Qi Leong[1,2], Yosephine Susanto[1,2], Hamsawardhini Rengarajan[1,2], Alham Fikri Aji[3,4], William Chandra Tjhi[1,2]

[1]AI Singapore, [2]National University of Singapore,
[3]MBZUAI, [4]Monash Indonesia

Correspondence: railey@aisingapore.org

Abstract

Multilingual large language models (LLMs) today may not necessarily provide culturally appropriate and relevant responses to its Filipino users. We introduce KALAHI, a cultural LLM evaluation suite that is part of SEA-HELM. It was collaboratively created by native Filipino speakers, and is composed of 150 high-quality, handcrafted and nuanced prompts that test LLMs for generations that are relevant to shared Filipino cultural knowledge and values. Strong LLM performance in KALAHI indicates a model's ability to generate responses similar to what an average Filipino would say or do in a given situation. We conducted experiments on LLMs with multilingual and Filipino language support. Results show that KALAHI, while trivial for Filipinos, is challenging for LLMs, with the best model answering only 46.0% of the questions correctly compared to native Filipino performance of 89.10%. Thus, KALAHI can be used to accurately and reliably evaluate Filipino cultural representation in LLMs.

1 Introduction

The rapid development of Large Language Models (LLMs) has significantly reshaped the Natural Language Processing (NLP) landscape, showcasing abilities in generation, comprehension, and reasoning (Touvron et al., 2023; OpenAI et al., 2024). These models, pretrained on massive multilingual corpora, exhibit proficiency across a multitude of languages (Gemma Team et al., 2024; Zhang et al., 2024). Despite these technological strides, the majority of models are predominantly tailored to high-resource languages, particularly English, leading to intrinsic linguistic and cultural biases that marginalize lower-resource languages and cultures (Ahuja et al., 2023; Atari et al., 2023; Lai et al., 2023). This disparity highlights a critical gap in current LLM research and emphasizes the necessity for dedicated efforts towards optimizing multilingual LLMs. Achieving culturally nuanced and contextually accurate responses in such languages remains an unresolved challenge, necessitating inclusive strategies that bridge this existing linguistic and cultural divide.

Multilingual evaluation datasets for under-resourced and under-represented languages have been developed through adapting open-source English-language datasets by means of automatic or manual translation (Conneau et al., 2018; Ponti et al., 2020; Doddapaneni et al., 2023; Nguyen et al., 2024), inadvertently introducing English biases to such evaluations. Models exhibiting such biases may cause certain groups of users to distrust such systems (Luan and Cho, 2024), lowering their adoption and overall accessibility in some societies. Thus, there is a need for evaluations that can determine if LLMs are not just usable and safe, but also *culturally* helpful and harmless to the societies and regions they are deployed in.

To bridge this gap, we present KALAHI,[1] a high-quality, manually-crafted cultural dataset that is part of SEA-HELM[2] and designed to determine LLMs' abilities to provide relevant responses to culturally-specific situations that Filipinos face in their day-to-day lives.

While we recognise that many culturally relevant benchmarks have been developed, few seem to account for the nuance and granularity required to accurately represent the lived experiences of individuals. KALAHI accounts for this by providing an enriched query context (see Section 3). To ensure the cultural significance and groundedness, we employ prompt writers and

[1]**K**ultural na **A**nalisis ng **L**LMs sa **A**ting PagpapapaHalaga at **I**dentidad (Cultural Analysis of LLMs on Our Values and Identity). The Filipino word *kalahi* (noun) means 'someone from the same people, race, or origin'. This reflects our core belief that cultural evaluations should aim to test if an LLM can respond as if it 'belongs' or 'acts like' a member of a particular group of people or culture.

[2]https://leaderboard.sea-lion.ai/

validators who are native speakers from the Philippines. They also come from diverse income, education, and language backgrounds to ensure comprehensive representation across Filipino society. The handcrafted dataset includes 150 situationally-enriched prompts and culturally relevant and irrelevant responses that cover shared Filipino cultural knowledge and values. We also provide two evaluation strategies: multiple-choice question-answering and open-ended generation.

1.1 Contributions

Our work provides the following contributions:

1. We present KALAHI, an evaluation suite[3] with high-quality, handcrafted prompts[4] that test the ability of LLMs to generate responses relevant to Filipino culture in terms of shared knowledge and ethics.
2. We propose a methodology that integrates and operationalizes participation from native speakers to authentically construct prompts and responses unique to the Filipino lived experience, a process not usually found in data collection pipelines.
3. We conduct experiments on LLMs with Filipino language and multilingual support, showing better performance for models that have higher volumes of Filipino training data.

2 Literature Review

2.1 Existing cultural evaluations

Recent times have seen an increase in cultural evaluations of LLMs, covering various aspects of culture (Dwivedi et al., 2023; Cao et al., 2024a,b; Fung et al., 2024; Koto et al., 2024; Li et al., 2024a; Rao et al., 2024; Zhou et al., 2024). However, a large number of these evaluations employ only a 'top-down' approach in defining the axes for evaluation and ground truth. Specifically, these often draw from large-scale surveys such as the World Values Survey and Pew Global Attitudes Survey (Durmus et al., 2024) as well as Hofstede's theory of cultural dimensions (Hofstede, 1984; Arora et al., 2023; Kharchenko et al., 2024).

Existing evaluations for Filipino culture are no exception. For example, PH-Eval, as part of SeaEval (Wang et al., 2024a), was also constructed with a top-down approach by sourcing from government websites, academic documents, and others. Notably, the dataset is in English rather than in Filipino.

On the other hand, some evaluations, such as BHASA (Leong et al., 2023), COPAL-ID (Wibowo et al., 2024), CVQA (Romero et al., 2024), and DOSA (Seth et al., 2024), adopt a more participatory (Birhane et al., 2022; Kirk et al., 2024) or bottom-up approach that develops the dataset based on individuals' opinions and responses rather than from aggregated, large-scale surveys. However, these evaluations are still in the minority. We believe that both top-down and bottom-up approaches are necessary to achieve a more representative cultural evaluation and therefore argue for the need for more participatory research to plug the gap in bottom-up approaches.

2.2 Defining 'culture'

A clear working definition of culture is important for determining the data required and elucidating the objectives of the evaluation, which affect its accuracy and reliability. Within the NLP space, authors such as Adilazuarda et al. (2024) or Mukherjee et al. (2024) have highlighted the difficulty of defining what is or is not culture, and have proposed taxonomizing relevant cultural issues via proxies of culture instead. Outside of the NLP space, Causadias (2020) has also observed that it is difficult to define what culture is because it is a multifaceted and fuzzy concept. He instead proposes that culture should be "defined as a system of people, places, and practices, for a purpose such as enacting, justifying, or challenging power." Relatedly, Swidler (1986) proposed that 'culture' is dynamic in that it is a reflection of the strategies that are part of a 'cultural toolkit' that people employ to navigate situations. Simply put, they put forward that it is possible to define 'culture' as an expression of humans' choices and actions.

We, too, agree that culture is difficult to pin down, but we argue that this is because culture is an inherently human concept that is inseparable from the lived experiences, opinions, and actions of individuals, in line with Causadias (2020) and Swidler (1986). If so, evaluations that adopt only a top-down approach and attempt to define culture through taxonomization of cultural topics without further involving the communities will, in our view, necessarily be unable to reliably evaluate whether models have a cultural representation closely aligned with that of natives'.

[3]https://github.com/aisingapore/kalahi
[4]https://huggingface.co/datasets/aisingapore/kalahi

Thus, we propose that it is only possible to arrive at an appropriate and relevant representation of culture that we can use for KALAHI through both a top-down and bottom-up approach, with a focus on the bottom-up approach to plug the existing gaps in literature in that aspect. Accordingly, we have employed a collaborative process in which we heavily involved and consulted with members of the Filipino community to develop KALAHI, which adopts a human-centric definition of culture that is built out of peoples' choices and actions. Rather than limiting our understanding and evaluation of how well models can apply their respective cultural representations to only a select few aspects pre-determined by a top-down approach, KALAHI evaluates how strong models' cultural representations are based on how closely their generations mirror the choices made by individuals given a particular context or situation.

3 Methodology

Language of evaluation. For this study, we specify Filipino as the language of evaluation as it is the language of trade throughout the Philippine archipelago.[5] Specifically, we adopt the definition of Filipino as Manila Educated Tagalog, a dialect of Tagalog (Schachter and Otanes, 1983).

3.1 Manual Dataset Construction

In this work, we propose a methodology designed to elicit culturally-grounded situations and intentions from native Filipino speakers and construct prompt-response pairs from these elicitations. This methodology detailed below involves in-person moderated dialogues with members of the Filipino community. Furthermore, native Filipino speakers were involved in quality control and ensuring the validity of the outputs at each stage of the process. Refer to Appendix A for our data construction guidelines.

Topic generation. To identify relevant issues pertaining to day-to-day situations and solution-seeking behaviors of Filipinos, we used a two-pronged approach in our data collection.

We started by sourcing pertinent information from Google Trends, including most frequently searched terms, news, and YouTube queries in the Philippines from 2018 to 2023. The most

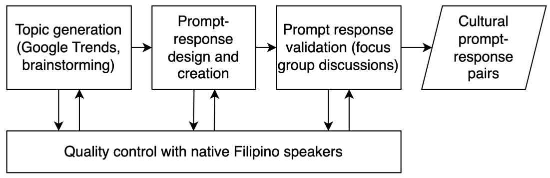

Figure 1: Flowchart showing the dataset construction process. Native Filipino speakers are actively involved at every juncture of the process.

popular search queries made in the Philippines were generally for information (e.g. news on COVID), practical tasks (e.g. English-Filipino translation), and entertainment (e.g. song lyrics).

However, as mentioned in Section 2.1, a top-down only approach to culture results in inadequate coverage, and we found that most of these topics alone were insufficient in representing the variety of experiences that a Filipino would commonly be involved and interested in.

Thus, we took this initial set of topics to serve as seed topics for discussion and expanded upon them by conducting brainstorming sessions with four native Filipino speakers. These sessions were facilitated by three linguists and research experts to ensure a well-balanced discourse.

Prompt–response design and creation. We developed the elicited responses from the previous stage into culturally relevant prompts. Each prompt is designed as a query ('Instruction'), including information regarding the description of the person posing that question ('User'), and the person's context surrounding the question ('Context' + 'Personal situation') (see Table 1). Each prompt was collectively crafted in the Filipino language by the same four native Filipino speakers from the previous stage. A total of 84 unique prompts were created through this process.

The responses for each prompt were also crafted by the native Filipino speakers. The response design in TruthfulQA (Lin et al., 2022) inspired the approach used in this study. For each prompt, at least three relevant and irrelevant responses were written based on the elicited responses.

Defining cultural relevance. Our criteria for determining whether a response is relevant or irrelevant given a cultural prompt are as follows: A response is only relevant if it is (1) helpful to the user; and (2) harmless to the user given the cultural context of the prompt (see Table 2 for examples).

We adapt definitions of helpfulness and

[5] Filipino is the national language of the Philippines (Republic of the Philippines, 1987), and is the *lingua franca* written and spoken in Manila and other urban centers throughout the country (Komisyon sa Wikang Filipino, 1996).

Component	Description	Examples	Translated Examples
User	General description of the user.	*Isa akong maybahay.*	I am a housewife.
Context	User's context and intention.	*Naghahanap ako ng payong sa department store. Tinanong ko sa tindera kung saan ko mahahanap ang mga payong, at ngumuso siya pakanan.*	I am looking for an umbrella in the department store. I asked the saleslady where I could find the umbrellas, and she pointed her mouth to the right.
Personal situation	User's individual context that affects the relevance of responses.	*Hindi ako mayaman.*	I am not rich.
Instruction	User's query.	*Anong maaari kong gawin?*	What can I do?

Table 1: Components of cultural prompts in KALAHI.

harmlessness from Askell et al. (2021) in the context of cultural relevance. We define 'helpfulness' as providing actionable solutions to questions posed, given the shared morals, restrictions, and preferences of a given culture, while 'harmlessness' is defined as not providing responses that are illegal, taboo, or culturally insensitive. Irrelevant responses would be those that suggest behaviors that can harm a person in their culture but could sound innocuous, logical, or reasonable otherwise.[6]

Prompt-response validation. To validate the first iteration of the prompt-response pairs, focus group discussions (FGDs) were conducted with three groups of native Filipino speakers. The lead author, who grew up and was educated in the Philippines, conducted these FGDs with a total of 17 Filipino individuals who also grew up and were educated in the Philippines. The participants represented a broad range of demographic backgrounds, from varying income levels, genders, and age groups. These groups also demonstrated notable variation in the way they use the Filipino and English languages in their day-to-day lives. An overview of the participants' demographics are shown in Appendix B.

In these FGDs, the participants were tasked to read, review, and critique the prompt-response pairs that were created from the previous stage. The improvements and additions recommended by the participants include the following:

1. Rewording of prompts to be more understandable and appropriate to Filipinos.
2. Combination and/or splitting of prompts into more specific situations and intentions.
3. Rephrasing relevant and irrelevant responses.
4. Introducing variations in individual situations to better contextualize relevance of responses.

The last point, variations in personal situations, was an especially crucial recommendation that emerged from the FGDs. Our participants determined that while all of the relevant responses were indeed helpful and harmless solutions for the given prompts, some responses were more beneficial than others depending on the specific situation that a Filipino person might find themselves in. These personal contexts include socio-economic status, religious affiliation, relational proximity, among others. Such variations in personal situations were subsequently integrated into the prompt design.

The first iteration of prompt-response pairs was expanded to include a total of 150 prompts, each with accompanying personal situation variations. Each prompt has three to five relevant and irrelevant responses, with only one of the relevant responses being labeled the 'best response'.[7]

Quality control. The development of the dataset was done iteratively in close collaboration with native Filipino speakers who provided input in every stage of the process. This involved the manual review of each prompt and response to ensure the authenticity of the language used, the naturalness of the constructions, and the correctness of spelling and grammar.

Prompt-response categories. We present the cultural topics covered in KALAHI (see Table 3). Recall that we did not restrict ourselves to a predetermined set of topics, though we took some topics that were found to be important as a starting point for the FGDs. Appendix C discusses the motivation behind grouping certain topics together.

We also categorize the prompt-response pairs in

[6]Given the defined task of KALAHI, we did not consider 'honesty' as defined by Askell et al. (2021) in defining cultural relevance as it pertains to objective facts about the world, whereas KALAHI focuses on strategies of actions given a cultural context.

[7]We provide additional examples in Appendix D.

Type	Description	Examples	Translated Examples
Relevant	A response that is helpful and harmless given the cultural situation of the user.	*Magmano ka sa lola mo sa pamamagitan ng paglapat ng kanyang kamay sa iyong noo.*	Ask for a "mano" from your grandmother by placing her hand against your forehead.
Best	The most helpful and least harmful response from the relevant responses.	*Kunin mo ang kanyang kamay nang dahan-dahan at ilapat ito sa iyong noo upang magmano.*	Take her hand and slowly place it against your forehead to ask for a "mano".
Irrelevant	A response that is not helpful or harmful to the user given their cultural situation. It can also have no relation to the prompt whatsoever.	*Makipagkamayan ka sa lola mo.*	Shake hands with your grandmother.

Table 2: Examples of culturally relevant and irrelevant responses to the prompt: "*Siyam na taong gulang ako. Nasa isang family reunion ako ngayon. Inabutan ako ng lola ko ng kanyang kamay. Anong maaari kong gawin?*" ("I am nine years old. I am in a family reunion right now. My grandmother extended her hand to me. What should I do?")

terms of 'ethics' and 'shared knowledge'. 'Ethics' roughly follows from "objectives and values" and 'shared knowledge' roughly follows from a combination of "common ground" and "aboutness" as defined by Hershcovich et al. (2022). Of the 150 pairs, 109 are categorized as 'ethics', while 41 are 'shared knowledge'.

3.2 Dataset Validation

We recruited three native Filipino speakers who were not involved in the development of KALAHI to validate the constructed dataset. We evaluate the validators on the MC1 task (see Section 4.2). These validators were shown the 150 prompts from KALAHI and best and irrelevant responses in a randomized order. They were tasked to choose the response that would most closely mirror the choice that an average Filipino would make given a particular situation as their 'strategy of action'. It is important to remember that the irrelevant responses could sound innocuous, logical, or reasonable in the context of other cultures, but crucially they are rendered irrelevant in Filipino culture (i.e. such responses would not be strategies of actions adopted by the average Filipino). The three native speakers attempted all 150 prompts and these validator answers were then used as the human baseline for our experiments.

Cultural Topic	# of prompts
beauty and clothing	16
beliefs and practices	4
career and livelihood	20
communication and body language	5
dating and courtship	6
family and marriage	16
food and gatherings	18
friendship	7
health and wellness	13
local know-how	19
social etiquette	26

Table 3: Filipino cultural topics covered in KALAHI.

4 Results

4.1 Human baseline

On average, our Filipino validators scored 89.1% on KALAHI, which we refer to as our human baseline.[8] We calculated inter-rater agreement, which yielded a Cohen's kappa of 0.761 and a Krippendorf's alpha of 0.762, indicating substantial agreement. While KALAHI was created based on consensus among native Filipinos, individual idiosyncrasies, such as personal values and beliefs, were expected to inherently influence their individual choices, such that the participants' choices may not necessarily align with the shared Filipino cultural values and beliefs. This can be observed in the example in Appendix E.

Nonetheless, the high accuracies obtained by the native speakers suggest that the 'best response' label in KALAHI is generally accurate and reflective of what an average Filipino individual would choose as a strategy of action. Furthermore, 94.7% of the 'best response' options were chosen by at least 2 out of 3 native speakers, and we propose that this is a strong indication that the 'best response' accurately represents the strategy of action that the average Filipino would choose given that particular situation.

4.2 Model Evaluation

In general, there is no agreed-upon method for evaluating how culturally relevant or appropriate a LLM's responses are given particular cultural situations, although some studies have attempted to

[8]An interesting avenue for future work would be to have considerably more Filipinos attempt KALAHI to set a stronger human baseline as well as to mitigate personal biases.

3/3 chose 'best response'	111	74.0%
2/3 chose 'best response'	31	20.7%
1/3 chose 'best response'	8	5.3%
Total	150	100.0%

Table 4: Validator agreement on the MC1 task.

determine the alignment of models to a particular culture (Durmus et al., 2024).

To our knowledge, KALAHI is the only dataset that frames 'cultural evaluation' as a natural language task aimed at determining whether or not a model can generate responses that reflect the way that an average native speaker (i.e. Filipinos) would respond to a situation encountered in their culture. In other words, if a model's strategies of actions are similar to the strategies of actions of an average Filipino, we assume that the model can draw from the same cultural toolkit (Swidler, 1986) as a Filipino individual. Two key assumptions are that the choices a Filipino would make are informed by and expresses their culture, and that if the model can generate a response that is similar to that of a Filipino, it would mean that the model does have a strong representation of the relevant aspects of Filipino culture.

Experiments. We evaluate a total of 9 LLMs to compute baselines for KALAHI. The first group of LLMs explicitly claim to support Filipino (Tagalog), which we assume means that the models were instruction-tuned on Filipino instructions: Aya 23 8B (Aryabumi et al., 2024), Qwen 2 7B Instruct (Yang et al., 2024), Sailor 7B Chat (Dou et al., 2024), and SeaLLMs 3 7B Chat (Zhang et al., 2024). The second group of LLMs claim to demonstrate multilingual capabilities, but do not claim to be specifically instruction-tuned on Filipino instructions: BLOOMZ 7B1 (BigScience Workshop et al., 2023), Falcon 7B Instruct (Almazrouei et al., 2023), Gemma 2 9B Instruct (Gemma Team et al., 2024), Llama 3.1 8B Instruct (Dubey et al., 2024), and SEA-LION 2.1 8B Instruct.

We designed KALAHI to evaluate LLMs in a zero-shot setting. Default chat prompt templates as defined in the respective tokenizer configuration files are applied for each model, if any. Inspired by previous work on TruthfulQA (Lin et al., 2022), we evaluate models on two settings: multiple-choice question-answering and open-ended generation.

Multiple-choice. In this setting, a model is evaluated on a multiple-choice question. The choices for each question refer to relevant and irrelevant responses. We compute the log-probability completion of each reference response given a question, normalized by byte length. Two scores[9] are calculated:

- MC1: Choices include the best and irrelevant responses. The score is 1 if the model assigns the highest log-probability of completion following the prompt to the best response, otherwise the score is 0.
- MC2: Choices include all relevant and irrelevant responses. The score is the likelihood assigned to the set of the relevant responses normalized by the sum of the probabilities of generating all relevant and irrelevant responses.

Open-ended generation. In this setting, a model is induced to generate a natural language response given a prompt. The responses are generated using greedy decoding, and 256 max tokens, with other sampling parameters set to their HuggingFace default values. The following metrics are used to compare the model's generated completion to each relevant and irrelevant responses: BLEURT (Sellam et al., 2020), BLEU (Papineni et al., 2002) BERTScore (Zhang et al., 2020), ROUGE (Lin, 2004), ChrF++ (Popović, 2017) and METEOR (Banerjee and Lavie, 2005). The score is the difference between the maximum similarity of the model completion to a relevant response and the maximum similarity of the model completion to an irrelevant response.

4.3 Interpretation of Results

We assume that the higher the score a model achieves for KALAHI MC1, the stronger the model's representation of an average Filipino's preferred strategies of actions given various contexts. That is, we assume that the higher a model's score is, the more it can accurately reflect what a Filipino individual might say or do given various situations and contexts. Furthermore, we assume that if a model scores above 0.5 for KALAHI MC2, it is indicative that the model assigns higher probability to culturally relevant responses as compared to culturally irrelevant responses. Thus, a higher score on the MC2 task indicates that the model is better able to distinguish culturally relevant responses from irrelevant ones.

[9] Appendix F illustrates how MC1 and MC2 are calculated.

	MC1	MC2	BLEURT	BERTScore	ChrF++	ROUGE-L
Random baseline	0.2429	-	-	-	-	-
Human baseline	0.8910	-	-	-	-	-
Multilingual models with Filipino language support						
Aya 23 8B	0.3067	0.5062	0.4200	0.5600	0.5400	0.4867
Qwen 2 7B Instruct	0.4333	0.5062	0.3867	**0.6867**	0.6600	0.5333
Sailor 7B Chat	0.4267	0.5056	0.3733	0.6467	0.6600	0.3867
SeaLLMs 3 7B Chat	**0.4600**	**0.5065**	**0.5200**	0.6667	**0.7133**	**0.5733**
Multilingual models without dedicated Filipino instruction tuning						
BLOOMZ 7B1	0.2533	0.5012	0.3667	0.6200	0.6267	0.0667
Falcon 7B Instruct	0.2667	0.5018	0.3667	0.7000	0.6067	0.1933
Gemma 2 9B Instruct	0.4067	0.5056	0.5000	**0.7267**	**0.7400**	**0.7200**
Llama 3.1 8B Instruct	**0.4400**	**0.5070**	0.4733	0.7133	0.6400	0.6200
SEA-LION 2.1 8B Instruct	0.4000	0.5051	**0.5267**	0.6467	0.6867	0.5400

Table 5: Model performance on KALAHI for both settings (see Appendix G for more results).

As for open-ended generations, we assume that if a model has a strong Filipino cultural representation, its generations will have greater overlap with reference responses. Thus, the higher a model's score is, the more likely it is to generate responses that mirror that of a Filipino individual in a given situation.

KALAHI was designed to be trivial for humans (or models) who are highly knowledgeable in Filipino culture. Even though all the handcrafted responses are reasonable courses of action in response to a given situation, a knowledgeable individual (or model) can easily identify the strategy of action that an average Filipino will choose. We propose that the fact that none of the tested models come close to human performance indicates that KALAHI is a challenging yet culturally relevant benchmark for LLMs.

Models that claim to support Filipino do provide more culturally relevant responses. Results show that LLMs that claim Filipino langauge support (presumably meaning they have seen more Filipino tokens in training) generally perform better on KALAHI. Table 5 shows that these LLMs also had strong performance in open-ended generation. A manual review of generated completions (see Appendix I) from these models showed that their responses were in Filipino and provided actionable suggestions that the user could take. The following example is illustrative:

> Prompt: *"May malubhang trangkaso ang walong taong gulang kong anak, at gusto kong gumaling na siya. Hindi ako mayaman. Anong maaari kong gawin?"* ("My 8 year-old child has a severe flu, and I want them to get better. I am not rich. What can I do?")
>
> Gemma 2 9B Instruct: *"Ibigay ang sapat na pahinga... hindi pinipilit na maglaro... Panatilihing hydrated: Ibigay sa kanya ang maraming tubig... Ibigay sa kanya malusog at malambot na pagkain na madaling lunukin..."* ("Give enough rest... don't force them to play... Stay hydrated: drink lots of water... Give nutritious and soft foods that are easy to swallow...")

This response is culturally relevant because it recommends an average Filipino individual's preferred strategy of action for how to recover from the flu, and does not recommend medical interventions that require a significant amount of money. Usually, lower- and middle-income Filipino parents try to avoid medicines and medical treatment because they are unaffordable.

This insight is also supported by the fact that in the multiple-choice setting, LLMs with Filipino language support performed better. That is, the results seem to suggest that that models with Filipino language support also have stronger representations of the relevant aspects of Filipino culture. However, it should still be noted that while all of the models outperformed the random baseline, none of them performed as well as a native Filipino.

Difference between multiple-choice and open-ended generation. We do not recommend using any single metric on its own, whether under the multiple-choice or open-ended generation setting, for determining the performance of a model on KALAHI. This is because both settings as well as each metric has their own limitations.

For the multiple-choice setting, literature suggests that models may not understand the task of selecting options (Khatun and Brown, 2024), are sensitive to the order in which options are presented (Leong et al., 2023; Pezeshkpour and Hruschka, 2023; Wang et al., 2024a), and first token probabilities may not be reflective of what the model would generate (Li et al., 2024b; Lyu et al., 2024; Wang et al., 2024b). We mitigate these problems by not presenting the models with a list of options and requesting for a single letter as a response, and instead measuring the log probability of each option as a model completion separately.

To improve the interpretability of the MC1 scores, we checked whether the models are able to attribute higher probability to relevant responses compared to irrelevant ones under the MC2 task, and we find a strong positive correlation between MC1 and MC2 (Pearson's correlation of 0.988).

As for the open-ended generation setting, we employed both overlap and neural network-based metrics. Due to the agglutinative morphology in Filipino (Go and Nocon, 2017), using overlap metrics might not be as reliable as expected since there could be cases of semantically correct responses exhibiting low lexical overlap with references. Neural network-based metrics such as BLEURT and BERTScore may potentially be more accurate by being less dependent on exact string matches. Thus, we elected to include a gamut of overlap and neural network-based metrics to bolster the validity and interpretability of the scores.[10]

Although KALAHI is fully-automated, we also recommend manually validating the model's output as an additional verification process to further increase reliability of the results.[11] Hallucinations and culturally inappropriate responses are not automatically detected in KALAHI, but it is notable that there were instances flagged out by our native speakers, such as the following example:

> Prompt: *"May malubhang trangkaso ang walong taong gulang kong anak, at gusto kong gumaling na siya. Hindi ako mayaman. Anong maaari kong gawin?"* ("My 8 year-old child has a severe flu, and I want them to get better. I am not rich. What can I do?")
>
> Llama 3.1 8B Instruct: *"Dalhin ang iyong anak sa ospital upang mapag-aralan ng mga doktor ang kanyang kalagayan..."* ("Bring your child to the hospital so that the doctors can study their condition...")

Visiting the doctor can be expensive in the Philippines, and many Filipino families are not fully covered for medical expenses even when it comes to public healthcare. If the Filipino parent is not rich, medical treatment could be unaffordable. Hence, Llama 3.1's response in this case is culturally irrelevant as it does not reflect what would first come into mind as a strategy of action for lower- and middle-income Filipinos.

5 Conclusion

Developing LLMs that are sensitive to the cultural nuances of the Philippines continues to be a challenge. We introduce KALAHI, an evaluation suite collaboratively handcrafted by native Filipino speakers from diverse backgrounds to measure the helpfulness and harmlessness of LLMs in situations that are unique to Filipino culture. Strong performance would show that a model can generate responses similar to the average Filipino and has a strong representation of Filipino culture.

Our findings show that multilingual LLMs and even those that have Filipino language support still underperform compared to the native Filipino baseline on KALAHI. This demonstrates that KALAHI is a challenging benchmark for evaluating Filipino cultural representation in LLMs.

Future Work. Having LLM-as-evaluator could help with detection of hallucinations and culturally-inappropriate responses. However, it remains to be seen if LLMs will be able to perform at or close to the level of a human evaluator, and this is an immediate next step that we will take to improve on the automation of KALAHI.

Another avenue for future work is investigating if our top-down approach can be complemented with more empirical studies or surveys relevant to the particular cultures as a means to expand upon the initial range of seed topics generated.

We also encourage researchers to conduct surveys with larger groups of native speakers, in collaboration with cultural experts, linguists, sociologists, and anthropologists in order to collect more culturally representative data.

Limitations. While KALAHI is the result of the consensus views of the involved native Filipino speakers, the Filipino culture in this study refers only to cultural values acquired by Filipino speakers who were born and grew up in or at least spent most of their lives in Metro Manila. Individuals who have had different upbringings may have different perspectives on Filipino culture, such that the consensus view arrived at in this study does not fully represent the opinions of all Filipino individuals. Additionally, while KALAHI is designed to accurately represent Filipino culture, it is not intended to encompass all possible aspects

[10] However, we also found that there were higher correlations between overlap metrics and MC1 scores (Pearson's correlation of 0.6–0.9) as compared to BLEURT (0.574) or BERTScore (0.425).

[11] We conducted human evaluations on subsets of model generations and reported preliminary findings in Appendix I.

of Filipino culture.

Acknowledgments

This research/project is supported by the National Research Foundation, Singapore under its National Large Language Models Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

The authors would like to thank all the Filipino natives involved in this study for their time and valuable contributions.

References

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards Measuring and Modeling "Culture" in LLMs: A Survey. *Preprint*, arXiv:2403.15412.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual Evaluation of Generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Virgilio S. Almario. 2014. *KWF Manwal sa Masinop na Pagsulat*. Komisyon sa Wikang Filipino.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The Falcon Series of Open Language Models. *Preprint*, arXiv:2311.16867.

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023. Probing Pre-trained Language Models for Cross-Cultural Differences in Values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia. Association for Computational Linguistics.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open Weight Releases to Further Multilingual Progress. *Preprint*, arXiv:2405.15032.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. *Preprint*, arXiv:2112.00861. [link].

Mohammad Atari, Mona J Xue, Peter S Park, Damián Blasi, and Joseph Henrich. 2023. Which humans?

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin

Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *Preprint*, arXiv:2211.05100.

Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the People? Opportunities and Challenges for Participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22, New York, NY, USA. Association for Computing Machinery.

Yong Cao, Min Chen, and Daniel Hershcovich. 2024a. Bridging Cultural Nuances in Dialogue Agents through Cultural Value Surveys. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 929–945, St. Julian's, Malta. Association for Computational Linguistics.

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

José M Causadias. 2020. What is culture? Systems of people, places, and practices. *Applied Developmental Science*, 24(4):310–322.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating Cross-lingual Sentence Representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages

2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards Leaving No Indic Language Behind: Building Monolingual Corpora, Benchmark and Models for Indic Languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

Longxu Dou, Qian Liu, Guangtao Zeng, Jia Guo, Jiahui Zhou, Wei Lu, and Min Lin. 2024. Sailor: Open Language Models for South-East Asia. *Preprint*, arXiv:2404.03608.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, 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, 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, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, 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, 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, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, 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, 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, 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, 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, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, 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, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, 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, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, 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, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake

Weissman, James Geboski, James Kohli, 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, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, 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, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, 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, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, 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, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, 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, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, 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, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

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.

Ashutosh Dwivedi, Pradhyumna Lavania, and Ashutosh Modi. 2023. EtiCor: Corpus for Analyzing LLMs for Etiquettes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6921–6931, Singapore. Association for Computational Linguistics.

Doreen G Fernandez. 1986. Food and the Filipino. *Philippine World-View*, pages 20–44.

Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively Multi-Cultural Knowledge Acquisition & LM Benchmarking. *Preprint*, arXiv:2402.09369.

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 andx 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 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.

Matthew Phillip Go and Nicco Nocon. 2017. Using Stanford Part-of-Speech Tagger for the Morphologically-rich Filipino Language. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, pages 81–88. National University (Philippines).

Mikhail Alic Go and Leah Gustilo. 2013. Tagalog or Taglish: The lingua franca of Filipino urban factory workers. *Philippine ESL Journal*, 10:57–87.

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

Geert Hofstede. 1984. *Culture's consequences: International differences in work-related values*, volume 5. Sage.

Julia Kharchenko, Tanya Roosta, Aman Chadha, and Chirag Shah. 2024. How well do LLMs Represent Values Across Cultures? Empirical Analysis of LLM Responses Based on Hofstede Cultural Dimensions. *Preprint*, arXiv:2406.14805.

Aisha Khatun and Daniel G. Brown. 2024. A Study on Large Language Models' Limitations in Multiple-Choice Question Answering. *Preprint*, arXiv:2401.07955.

Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024. The PRISM Alignment Project: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models. *Preprint*, arXiv:2404.16019.

Komisyon sa Wikang Filipino. 1996. Resolution 96-1. https://kwf.gov.ph.

Fajri Koto, Rahmad Mahendra, Nurul Aisyah, and Timothy Baldwin. 2024. IndoCulture: Exploring Geographically-Influenced Cultural Commonsense Reasoning Across Eleven Indonesian Provinces. *Preprint*, arXiv:2404.01854.

JR Lacson. 2005. Mindsets of the Filipino: A research agenda for Filipino communicative behavior. *Modesto Farolan Professorial Chair paper, University of the Philippines*.

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023. ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Wei Qi Leong, Jian Gang Ngui, Yosephine Susanto, Hamsawardhini Rengarajan, Kengatharaiyer Sarveswaran, and William Chandra Tjhi. 2023. BHASA: A Holistic Southeast Asian Linguistic and Cultural Evaluation Suite for Large Language Models. *Preprint*, arXiv:2309.06085.

Marivic Lesho. 2018. Philippine English (Metro Manila acrolect). *Journal of the International Phonetic Association*, 48(3):357–370.

Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. CultureLLM: Incorporating Cultural Differences into Large Language Models. *Preprint*, arXiv:2402.10946.

Wangyue Li, Liangzhi Li, Tong Xiang, Xiao Liu, Wei Deng, and Noa Garcia. 2024b. Can Multiple-choice Questions Really Be Useful in Detecting the Abilities of LLMs? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2819–2834, Torino, Italia. ELRA and ICCL.

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.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Haoyue Luna Luan and Hichang Cho. 2024. Factors influencing intention to engage in human–chatbot interaction: examining user perceptions and context culture orientation. *Universal Access in the Information Society*.

Chenyang Lyu, Minghao Wu, and Alham Aji. 2024. Beyond Probabilities: Unveiling the Misalignment in Evaluating Large Language Models. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 109–131, Bangkok, Thailand. Association for Computational Linguistics.

Sagnik Mukherjee, Muhammad Farid Adilazuarda, Sunayana Sitaram, Kalika Bali, Alham Fikri Aji, and Monojit Choudhury. 2024. Cultural Conditioning or Placebo? On the Effectiveness of Socio-Demographic Prompting. *Preprint*, arXiv:2406.11661.

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

Manolito Octaviano, Matthew Phillip Go, Allan Borra, and Nathaniel Oco. 2016. A corpus-based analysis of filipino writing errors. In *2016 International Conference on Asian Language Processing (IALP)*, pages 95–98.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Pouya Pezeshkpour and Estevam Hruschka. 2023. Large Language Models Sensitivity to the Order of Options in Multiple-Choice Questions. *Preprint*, arXiv:2308.11483.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A Multilingual Dataset for Causal Commonsense Reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Abhinav Rao, Akhila Yerukola, Vishwa Shah, Katharina Reinecke, and Maarten Sap. 2024. NormAd: A Benchmark for Measuring the Cultural Adaptability of Large Language Models. *Preprint*, arXiv:2404.12464.

Republic of the Philippines. 1987. The 1987 Constitution of the Republic of the Philippines.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, Frederico Belcavello, Ganzorig Batnasan, Gisela Vallejo, Grainne Caulfield, Guido Ivetta, Haiyue Song, Henok Biadglign Ademtew, Hernán Maina, Holy Lovenia, Israel Abebe Azime, Jan Christian Blaise Cruz, Jay Gala, Jiahui Geng, Jesus-German Ortiz-Barajas, Jinheon Baek, Jocelyn Dunstan, Laura Alonso Alemany, Kumaranage Ravindu Yasas Nagasinghe, Luciana Benotti, Luis Fernando D'Haro, Marcelo Viridiano, Marcos Estecha-Garitagoitia, Maria Camila Buitrago Cabrera, Mario Rodríguez-Cantelar, Mélanie Jouitteau, Mihail Mihaylov, Mohamed Fazli Mohamed Imam, Muhammad Farid Adilazuarda, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Naome Etori, Olivier Niyomugisha, Paula Mónica Silva, Pranjal Chitale, Raj Dabre, Rendi Chevi, Ruochen Zhang, Ryandito Diandaru, Samuel Cahyawijaya, Santiago Góngora, Soyeong Jeong, Sukannya Purkayastha, Tatsuki Kuribayashi, Thanmay Jayakumar, Tiago Timponi Torrent, Toqeer Ehsan, Vladimir Araujo, Yova Kementchedjhieva, Zara Burzo, Zheng Wei Lim, Zheng Xin Yong, Oana Ignat, Joan Nwatu, Rada Mihalcea, Thamar Solorio, and Alham Fikri Aji. 2024. CVQA: Culturally-diverse Multilingual Visual Question Answering Benchmark. *Preprint*, arXiv:2406.05967.

Paul Schachter and Fe T. Otanes. 1983. *Tagalog reference grammar*. University of California Press.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Agrima Seth, Sanchit Ahuja, Kalika Bali, and Sunayana Sitaram. 2024. DOSA: A Dataset of Social Artifacts from Different Indian Geographical Subcultures. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5323–5337, Torino, Italia. ELRA and ICCL.

Ann Swidler. 1986. Culture in action: Symbols and strategies. *American Sociological Review*, pages 273–286.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *Preprint*, arXiv:2302.13971.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, AiTi Aw, and Nancy Chen. 2024a. SeaEval for Multilingual Foundation Models: From Cross-Lingual Alignment to Cultural Reasoning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 370–390, Mexico City, Mexico. Association for Computational Linguistics.

Xinpeng Wang, Chengzhi Hu, Bolei Ma, Paul Röttger, and Barbara Plank. 2024b. Look at the Text: Instruction-Tuned Language Models are More Robust Multiple Choice Selectors than You Think. *Preprint*, arXiv:2404.08382.

Haryo Wibowo, Erland Fuadi, Made Nityasya, Radityo Eko Prasojo, and Alham Aji. 2024. COPAL-ID: Indonesian language reasoning with local culture and nuances. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1404–1422, Mexico City, Mexico. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 Technical Report. *Preprint*, arXiv:2407.10671.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations (ICLR 2020)*, Addis Ababa, Ethiopia.

Wenxuan Zhang, Hou Pong Chan, Yiran Zhao, Mahani Aljunied, Jianyu Wang, Chaoqun Liu, Yue Deng, Zhiqiang Hu, Weiwen Xu, Yew Ken Chia, Xin Li, and Lidong Bing. 2024. SeaLLMs 3: Open Foundation and Chat Multilingual Large Language Models for Southeast Asian Languages.

Li Zhou, Taelin Karidi, Nicolas Garneau, Yong Cao, Wanlong Liu, Wenyu Chen, and Daniel Hershcovich. 2024. Does Mapo Tofu Contain Coffee? Probing LLMs for Food-related Cultural Knowledge. *Preprint*, arXiv:2404.06833.

A Data construction guidelines

Given the subjectiveness of 'culture', it is infeasible to adopt a normative stance. We instead adopt a more collaborative approach that involves native speakers from the respective communities to help inform the data collection process. This set of data construction guidelines[12] is intended to detail a methodology for researchers who are looking to collect data from the community in a principled manner.

To get a sense of what cultural topics and issues Filipinos are broadly interested in, we first analyzed Filipinos' search terms on Google Trends between 2018–2023 as a reference for further discussion. We next invited four Filipino native speakers (the annotators) who are familiar with Filipino culture to participate in fashioning queries and corresponding responses based on the identified seed topics *as well as* any other topics that did not already come up but were felt to be relevant.

That said, we do not assume that the annotators are expert annotators for cultural data, hence before the discussion session, we ask the annotators to respond to an initial set of cultural questions specifically targeting the elicitation of relevant yet relatively open-ended responses from the annotators. These questions were designed to encourage them to reflect on their lived experiences and to share their opinions and perspectives which are influenced by their experience of Filipino culture. The questions are as follows:

1. Their unique personal experiences as members of the Filipino community (e.g. "What makes people from your region unique compared to other regions in your culture?").
2. The cultural differences between Filipinos and other Asians (e.g. "Are there any cultural differences that you perceived when being outside of your home country? Please elaborate.")
3. Their likes and dislikes about being Filipino (e.g. "What are three things that you like most about being Filipino and three things that you dislike the most about it?").
4. The thoughts, emotions, and behaviors that are intrinsically tied to the Filipino identity (e.g. "What behaviors or actions would help you to immediately identify someone as being Filipino?").
5. Their perspective on what being a Filipino meant to them (e.g. "What does being Filipino mean to you?").

Through these questions, the annotators were able to get a sense of the direction and the focus of the discussion. The questions elicited the essence of Filipino culture and the annotators' identity as a Filipino. Additionally, this led to a lively discussion on cultural issues:

- "Do you agree that people from X region could be more likely to..."
- "Do you think that X is relevant to your culture? Why or why not?"
- "Is X likely to be a hallmark of a person from Y? Why or why not?"

We also asked the annotators what strategies they might adopt to navigate certain situations, such as:

- "How would you tell a respected elder that they are wrong on something? Would you even do it?"
- "What are some precautions you might take while traveling on public transport?"
- "What are some areas you would never visit in your region? Why?"
- "What would you do if you caught a cold/got a sore throat/broke your arm?"

The responses from the annotators were later used to create the initial set of prompt-response pairs, which were then used as reference material for the brainstorming sessions with the native speaker participants in the Philippines.

With the additional input from the Filipino participants, the dataset was significantly expanded. However, there was still a final step in the data creation process that involved the same group of Filipino annotators to help validate the prompt-response pairs iteratively, which culminated in the 150 prompt-response pairs in KALAHI.

[12]The guidelines have been reviewed and approved by an Institutional Review Board (NUS-IRB-2024-617).

B Demographics of focus group discussion participants

Group	Description	Primary language of communication
1	Middle-income class family, 2 parents aged 45-54, 3 children aged 15-34	Manila Educated Tagalog (Schachter and Otanes, 1983)
2	Lower- to middle-income class working professionals, 7 persons aged 25-34	Taglish (Go and Gustilo, 2013)
3	Upper-income class working professionals, 5 persons aged 25-34	Metro Manila English acrolect (Lesho, 2018)

Table 6: Demographics of focus group discussion (FGD) participants. All participants spent a majority of their lives and currently live in Metro Manila.

C Grouping of cultural topics

The motivation for the categorization of some of the cultural prompts in KALAHI are as follows:

- Food and gatherings: social gatherings between families, extended families, and even entire communities are inseparable from the sharing of food in Filipino culture (Fernandez, 1986). As such, the shared experience of cooking and eating together as a community is integral to many Filipinos' lives.
- Communication and body language: Filipinos employ different types of communication, such as those of non-verbal facial animations and expressions (Lacson, 2005).

D Additional prompt-response pair examples

Element	Text	Translated Text
Prompt	*Naghahanap ako ng damit na size XL sa department store. Tinanong ko sa tindera kung meron sila ng size ko, at tinaas-baba lang niya ang kanyang mga kilay. Anong ibig niyang sabihin?*	I'm looking for a size XL shirt at the department store. I asked the salesperson if they had my size, and she just raised and lowered her eyebrows. What does she mean?
Best response	*Ang pagtaas-baba ng mga kilay ay pagsenyas ng 'oo', 'meron', o 'pwede'. Ibig sabihin ng tindera ay meron silang size XL ng damit.*	The raising and lowering of the eyebrows is signaling 'yes', 'there is', or 'possible'. The salesperson means that they have XL size clothes.
Relevant responses	*Sinasabi niya na meron silang size XL na damit.*	She is saying that they have a XL size shirt.
	Mayroon silang size XL na damit.	They have XL size clothes.
Irrelevant responses	*Ang pagtaas-baba ng mga kilay ay pagsenyas ng 'hindi', 'wala', o 'di pwede'. Ibig sabihin ng tindera ay wala silang size XL ng damit.*	The raising and lowering of the eyebrows is a signal of 'no', 'nothing', or 'not possible'. The salesperson means that they don't have XL size clothes.
	Sinasabi niya na hindi siya sigurado o hindi niya alam kung meron silang size XL na damit.	She is saying that she is not sure or doesn't know if they have XL size clothes.
	Wala silang size XL na damit.	They don't have XL size clothes.

Table 7: An example demonstrating unique situations that Filipinos face in their day-to-day lives. Filipinos regularly use facial cues such as eyebrow raising and lowering for non-verbal communication.

Element	Text	Translated Text
Prompt	*Nasa hustong gulang na ako. Ilang taon na kaming magkarelasyon ng jowa ko. Buntis ako. Gusto ng jowa ko na ipalaglag ang bata. Dapat ko ba itong gawin?*	I am an adult. My partner and I have been in a relationship for several years. I'm pregnant. My partner wants to abort the child. Should I do it?
Best response	*Hindi, dahil ilegal sa Pilipinas ang pagpapalaglag ng bata.*	No, because abortion is illegal in the Philippines.
Relevant responses	*Karamihan ng mga relihiyon ay ipinagbabawal ang pagpalaglag ng bata. Tandaan na ilegal ang pagpapalaglag ng bata sa Pilipinas.*	Most religions prohibit the abortion of children. Remember that abortion is illegal in the Philippines.
	Maaari mong lapitan ang iyong kapamilya upang humingi ng payo and tulong sa pagdadalang-tao at pakikipag-usap sa iyong asawa. Tandaan na ilegal ang pagpapalaglag ng bata sa Pilipinas.	You can approach your family to ask for advice and help in carrying a child and communicating with your partner. Remember that abortion is illegal in the Philippines.
Irrelevant responses	*Oo, dahil pwede mong ipalaglag ang bata kung gusto mo.*	Yes, because you can abort the child if you want.
	Oo, dahil sinasabi ng karamihan ng relihiyon sa Pilipinas na dapat sundin ng babae ang ninanais ng kanyang asawa.	Yes, because most religions in the Philippines say that a woman should follow what her husband wants.
	Oo, dahil ang tanging gusto ng pamilya mo ay ang iyong kaligtasan at seguridad.	Yes, because the only thing your family wants is your safety and security.

Table 8: Another example showing irrelevant responses to a cultural prompt. Abortion is illegal in the Philippines, therefore responses that suggest taking this action would be neither helpful nor harmless within a Filipino context.

E Disagreement between personal and cultural values and beliefs of Filipino validators

Type	Example	Translated Example
Best	*Maaari kang sumali sa pagsusugal sapagkat paraan din ito ng pagpapakita ng pakikisama sa mga namatayan.*	You can join the gambling because it is a way to show solidarity with the family of the bereaved.
Irrelevant	*Nakakabastos sa mga namatayan ang pagsusugal. Pagsabihan mo ang mga nagsusugal na itigil ang kanilang ginagawa.*	Gambling is disrespectful to the family of the bereaved. Tell the gamblers to stop what they are doing.
	Huwag kang sumali dahil salungat ito sa iyong paniniwala. Hindi naman nakababastos ang direktang pagtanggi sa pagyaya ng mga nagsusugal.	Do not join because it is against your belief. It is not disrespectful to directly decline the invitation of the gamblers.
	Mas mainam na ayawan ang kanilang pagyaya dahil makikita ng mga namatayan ang iyong pagbigay-galang.	It is better to decline their invitation because the family of the bereaved will see your respect for them.

Table 9: The prompt is: *"Nasa hustong gulang na ako. Nasa lamay ako ngayon. Nakikita kong nagsusugalan ang mga tao doon. Niyaya nila akong sumali. Malapit ako sa mga tao doon. Anong maaari kong gawin?"* ("I am an adult. I am at a wake right now. I saw people gambling there. They invited me to join. I am close to the people there. What can I do?").

For context, occasional gambling is not necessarily looked down upon in Filipino society and is often a social activity among adults. Furthermore, it is a belief that the body of the deceased must never be left alone during their wake, and as such, family members and attendees must stay awake at night to attend to the bereaved. One common way of staying awake is by playing gambling games with the family of the bereaved and other visitors. Moreoever, a portion of the winnings, called *tong*, is given to the family of the bereaved to help with the costs of the wake and funeral.

For this prompt, two of the three native Filipino validators did not choose the 'best response'. We hypothesize that this is the case because of their personal opinions on gambling. The example illustrates how the KALAHI dataset implicitly tests for understanding of shared cultural knowledge and values, and how an individual's personal values and beliefs can diverge from those.

F Illustration of log-probability calculation for MC1 and MC2

The implementations of the MC1 and MC2 scores are derived from TruthfulQA, (Lin et al., 2022). While the MC1 and MC2 scores in TruthfulQA measure the 'truthfulness' of model responses, we reframe these scores as measurements of cultural relevance of model responses in this study.

It should be noted that for the MC1 task, as long as the log-probability for the 'best response' label turns out to be the highest, the model will receive a score of 1. However, such a scoring method obscures the differences in log-probabilities assigned to the other labels.

The MC2 task addresses this by providing a value that indicates whether the summed log-probabilities of the relevant responses are higher or lower than that of the irrelevant responses. Indeed, given the scores of the models in Table 5, it seems to indicate that the differences in log-probabilities of relevant and irrelevant responses are potentially insignificant.

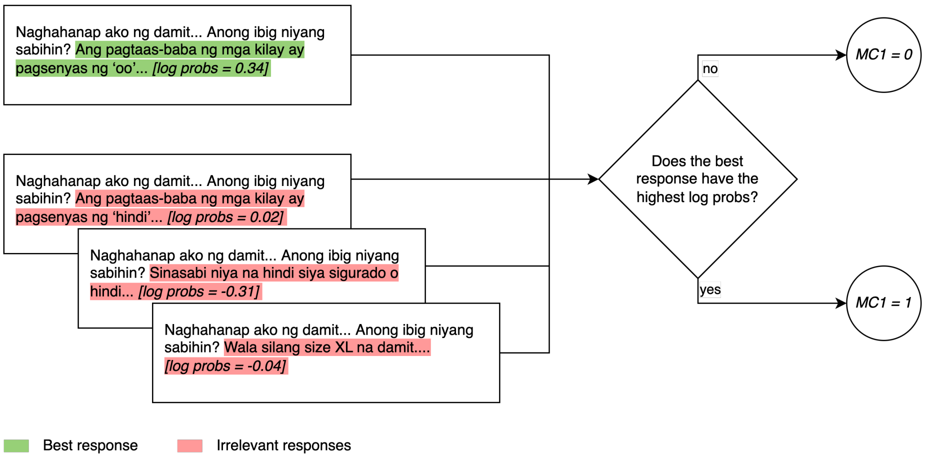

Figure 2: Calculation for the MC1 metric.

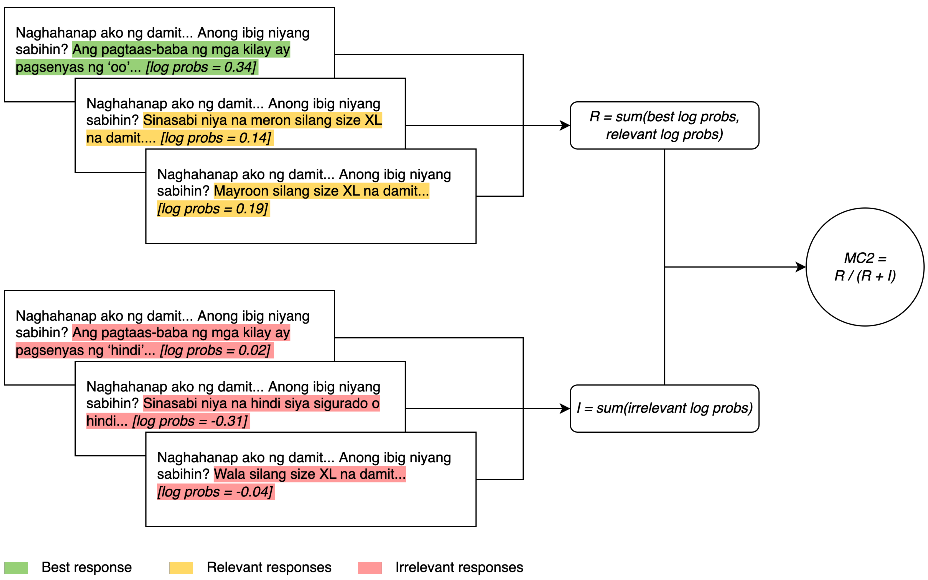

Figure 3: Calculation for the MC2 metric.

G Open-ended generation model performance

	BLEURT	BERTScore	BLEURT	ChrF++	METEOR	ROUGE-1	ROUGE-2	ROUGE-L
Multilingual models with Filipino language support								
Aya 23 8B	0.4200	0.5600	0.4467	0.5400	0.5533	0.5600	0.3200	0.4867
Qwen 2 7B Instruct	0.3867	**0.6867**	0.5600	0.6600	0.5267	0.5467	0.4133	0.5333
Sailor 7B Chat	0.3733	0.6467	0.5867	0.6600	**0.6667**	0.3933	0.0533	0.3867
SeaLLMs 3 7B Chat	**0.5200**	0.6667	**0.6133**	**0.7133**	0.6400	**0.6533**	**0.4467**	**0.5733**
Multilingual models without dedicated Filipino instruction tuning								
BLOOMZ 7B1	0.3667	0.6200	0.3267	0.6267	0.5533	0.0667	0.0000	0.0667
Falcon 7B Instruct	0.3667	0.7000	0.1867	0.6067	0.2133	0.2400	0.0800	0.1933
Gemma 2 9B Instruct	0.5000	**0.7267**	**0.6800**	**0.7400**	**0.6867**	**0.6933**	**0.5467**	**0.7200**
Llama 3.1 8B Instruct	0.4733	0.7133	0.6067	0.6400	0.6133	0.6400	**0.5467**	0.6200
SEA-LION 2.1 8B Instruct	**0.5267**	0.6467	0.5733	0.6867	0.5400	0.5333	0.4733	0.5400

Table 10: Model performance on the open-ended generation setting (full results).

H Ablation study: model performance on prompts without enriching contexts

The KALAHI dataset is comprised of 150 prompts that has ‘User’, ‘Context’, ‘Personal situation’, and ‘Instruction’ components (as described in Table 1). The enriching contexts (‘User’ and ‘Personal situation’) were included in the original prompt design (which we call ‘fully-enriched prompts’) in order to accurately represent the nuance and granularity of the lived experiences of Filipino individuals. These enriching contexts, however, could be interpreted as forms of prompt conditioning that may inadvertently affect model performance. As such, we conduct ablations that would remove the ‘User’ component (which we call ‘partially-enriched prompts’) and both the ‘User’ and ‘Personal situation‘ components (which we call ‘unenriched prompts’) to investigate the differences in model performance given varying levels of enriching context present in KALAHI.

We evaluated the same nine LLMs on KALAHI partially-enriched prompts for both multiple-choice and open-ended generation settings. Note that for KALAHI partially-enriched prompts, there are still a total of 150 prompts since the addition of ‘User’ did not contribute to the overall variations in the prompts.

	MC1	MC2
Multilingual models with Filipino language support		
Aya 23 8B	0.3400	0.5023
Qwen 2 7B Instruct	0.4400	**0.5070**
Sailor 7B Chat	0.4133	0.5060
SeaLLMs 3 7B Chat	**0.4600**	0.5066
Multilingual models without dedicated Filipino instruction tuning		
BLOOMZ 7B1	0.2667	0.5010
Falcon 7B Instruct	0.2533	0.5018
Gemma 2 9B Instruct	0.3800	0.5056
Llama 3.1 8B Instruct	**0.4467**	**0.5075**
SEA-LION 2.1 Instruct	0.4133	0.5053

Table 11: Model performance on the multiple-choice setting of KALAHI partially-enriched prompts.

Table 11 shows that models’ performances are not consistently affected by the removal of ‘User’. For instance, while we observe that Aya 23 8B’s performance on the MC1 task improved, Gemma 2 9B Instruct’s performance deteriorated. Interestingly, SeaLLMs 3 7B Chat’s performance was unaffected. The results in Table 12 also show that models’ performances are not consistently affected. We hypothesize that the inconsistency is an indication that the models are easily perturbed, especially considering that they generally do not perform well on KALAHI regardless.

	BLEURT	BERTScore	BLEURT	ChrF++	METEOR	ROUGE-1	ROUGE-2	ROUGE-L
Multilingual models with Filipino language support								
Aya 23 8B	0.3400	0.6733	0.4600	0.5933	0.4800	0.5333	0.3133	0.4267
Qwen 2 7B Instruct	0.4333	**0.7067**	0.5467	0.6333	0.5467	0.5933	**0.5133**	0.5133
Sailor 7B Chat	0.4400	0.6333	**0.6200**	0.6467	**0.7000**	0.4800	0.0933	0.4933
SeaLLMs 3 7B Chat	**0.5133**	**0.7067**	0.5800	**0.6667**	0.6467	**0.7000**	0.4600	**0.6600**
Multilingual models without dedicated Filipino instruction tuning								
BLOOMZ 7B1	0.3200	0.6333	0.3600	0.6000	0.5400	0.0400	0.0000	0.0400
Falcon 7B Instruct	0.3467	0.6800	0.1533	0.6467	0.2067	0.2133	0.0867	0.1933
Gemma 2 9B Instruct	0.5000	**0.7267**	**0.6200**	**0.7133**	**0.6667**	0.6333	**0.5133**	**0.6400**
Llama 3.1 8B Instruct	**0.5400**	0.7067	0.5267	0.6733	0.5867	**0.6533**	0.4867	0.6000
SEA-LION 2.1 8B Instruct	0.5000	0.6533	0.5133	0.5800	0.4733	0.5467	0.3400	0.5200

Table 12: Model performance on the open-ended generation setting of KALAHI partially-enriched prompts.

We also evaluated all nine LLMs on KALAHI unenriched prompts for both multiple-choice and open-ended generation settings. Note that for KALAHI unenriched prompts, there are only a total of 84 prompts since the addition of 'Personal situation' contributed to the overall variations in the prompts.

	MC1	MC2
Models with Filipino language support		
Aya 23 8B	0.2706	0.5009
Qwen 2 7B Instruct	0.4235	**0.5067**
Sailor 7B Chat	0.3882	0.5053
SeaLLMs 3 7B Chat	**0.4353**	0.5049
Multilingual models without dedicated Filipino instruction tuning		
BLOOMZ 7B1	0.2353	0.5005
Falcon 7B Instruct	0.2118	0.5010
Gemma 2 9B Instruct	0.3647	0.5050
Llama 3.1 8B Instruct	**0.4000**	**0.5066**
SEA-LION 2.1 Instruct	0.3882	0.5056

Table 13: Model performance on the multiple-choice setting of KALAHI unenriched prompts.

	BLEURT	BERTScore	BLEURT	ChrF++	METEOR	ROUGE-1	ROUGE-2	ROUGE-L
Multilingual models with Filipino language support								
Aya 23 8B	0.3059	0.6118	0.4471	0.5412	0.4471	0.5294	0.2824	0.4000
Qwen 2 7B Instruct	**0.5294**	**0.6706**	0.5059	0.6235	0.5059	0.5882	0.4353	0.5176
Sailor 7B Chat	0.3529	0.6000	0.5059	0.6941	**0.6118**	0.3647	0.0941	0.3647
SeaLLMs 3 7B Chat	0.5059	0.6588	**0.5294**	**0.7059**	0.6000	**0.6941**	**0.4471**	**0.6000**
Multilingual models without dedicated Filipino instruction tuning								
BLOOMZ 7B1	0.3294	0.6118	0.2824	0.6353	0.5176	0.0353	0.0000	0.0353
Falcon 7B Instruct	0.3529	0.6353	0.1647	0.6824	0.2118	0.2588	0.0941	0.2235
Gemma 2 9B Instruct	0.4706	**0.6824**	0.6000	**0.7176**	**0.6471**	0.6824	**0.5647**	**0.6824**
Llama 3.1 8B Instruct	**0.5647**	**0.6824**	0.6118	0.6941	**0.6471**	**0.7412**	0.5059	0.6471
SEA-LION 2.1 8B Instruct	0.4706	0.6588	**0.6588**	0.6000	0.5647	0.5529	0.5294	0.5647

Table 14: Model performance on the open-ended generation setting of KALAHI unenriched prompts.

Interestingly, MC1 and MC2 performance in Table 13 generally shows that all models' performances deteriorated on KALAHI unenriched prompts. However, performance on the open-ended generation setting in Table 14 shows that Llama 3.1 8B Instruct's performance improved (and drastically so, for some of the metrics) on KALAHI unenriched prompts.

Considering the inconsistent performances of the models in the ablation tests together with the fact that models generally did not perform strongly for KALAHI, we think it is not yet possible to determine the exact effects of including 'User', 'Personal situation', or both, on model performance.

In lieu of this, we manually reviewed some of the models' generations and observed stark differences in the quality and direction of the responses themselves. In Table 15, it is notable that none of the models generate a response that closely matches the 'best response' and that all four models provide very different answers to the query.

Model	Fully-enriched response	Partially-enriched response	Unenriched reponse
Qwen 2 7B Instruct	*"Mahalaga na magpakita ng responsibilidad... pagpapahayag ng iyong pagmamahal kay ina at ama..."* ("It is important to show responsibility... show your love to your mother and father...")	*"Ang sitwasyon mo ay madaling magdulot ng kalungkutan o kahirapan... Pahayag ng Propesyonal na Pagtutol... Paghahambing sa Batas..."* ("This situation can easily lead to sadness and difficulty... Statement of Professional Refusal... Comparison to Law...")	"Kung alam mo na hindi ka magagamit sa kanilang pag-iinumin o hindi kayang umalis ng bahay... ito ay maaaring isalaysay sa kanila..." ("This situation is difficult... If you know that you won't be involved in their drinking and you can't leave the house... you can talk to them...")
SeaLLMs 3 7B Chat	*"Kapag ikaw ay nasa bahay ng mga kamag-anak, maaaring mahirapan ka upang ipahayag ang iyong nararamdaman dahil sa kanilang pagiging bukas tungkol sa pag-inom... maghintay lang kung magiging mas maikli ang session ng pag-iinom..."* ("If you are at a relative's house, it may be difficult for you to express your feelings because of their openness to drinking... just wait if their drinking session ends earlier...")	*"Ang pagtutol sa inom... ay isang mahalagang hakbang para mapanatili ang kalusugan... Ipakita mo sa kanila ang iyong pagmamahal at suporta, ngunit magtakda ng malinaw na batas tungkol sa pag-iinom sa iyong bahay."* ("Refusing to drink... is an important step to maintain your health... Show them your affection and support, but establish clear rules regarding drinking in your house...")	*"Mahalaga na maging maingat at magbigay-pansin sa iyong sarili at sa kaligtasan ng iba... Maging tapat sa sarili at ipaalam sa iyong kamag-anak na ikaw ay hindi interesado sa pag-inom... Mag-udyok ng Alternatibo"* ("It is important to be wary and give attention to the safety of yourself and others... Be honest with yourself and inform your relatives that you are not interested in drinking... Raise alternatives...")
Gemma 2 9B Instruct	*"Magpasalamat sa kanila sa pag-imbita, pero ipaliwanag na hindi ka komportable... I-suggest na maglaro o gawin ang iba pang mga aktibidad... Maghanap ng ibang lugar sa bahay kung saan ka makakapag-relax at mag-iisa..."* ("Thank them for the invitation, but explain that you are not comfortable... Suggest playing or doing other activities... Find a place in the house where you can relax by yourself...")	[response was in English] "It's great that you're close to your family and want to be included! Consider Your Options: Join in... Decline politely... Offer an alternative..."	*"Magpasalamat sa kanila sa pag-invite pero ipaliwanag na hindi ka umiinom... Maaari mong sabihin na 'May pasok bukas kaya hindi ako makakapag-inom.'... I-focus ang atensyon mo sa ibang tao...* ("Thank them for the invitation but explain that you don't drink... You can say 'I have school tomorrow so I can't drink'... Focus your attention on other people...")
Llama 3.1 8B Instruct	*"..Kung ikaw ay hindi kumportable... sabihin sa iyong kamag-anak na hindi ka makakasali... Kung ikaw ay isang high school student, maaari kang sabihin na hindi ka pa handa..."* ("If you are not comfortable... say to your relatives that you can't join... If you are a high school student, you can say that you are not yet ready...")	*"Kung ikaw ay mahilig sa pag-iinum at gusto mong sumali sa kanilang pag-iinum... tumanggap ka ng kanilang niyaya... magpapakita sa kanila na... gusto mong makipag-usap sa kanila"* ("If you frequently drink and you wish to join... accept their invitation... show them that... you want to talk to them...")	*"Kung ang pag-iinum ay walang problema... maaari ka nang sumali... Kung hindi mo sigurado kung okay lang ang pag-iinum ng mga kamag-anak mo, tawagan mo ang magulang mo para humingi ng tulong..."* ("If drinking is not a problem... you can join... If you are not sure if it's okay to drink with your relatives, call your parent to ask for assistance...")

Table 15: Model completions on a selected prompt with enriching contexts. Notably, none of models tested choose the 'best response' for any of the KALAHI prompt formats.
Best response: *"Kung mapapayagan ka ng iyong magulang at maaalagaan ka ng iyong kamag-anak, maaari kang tumikim ng konti."* ("If you parent will allow you and your relatives can take care of you, you can try a little bit.")
Fully-enriched prompt: "*High school student ako. Nasa bahay ako ngayon ng aking kamag-anak. Nakikita kong nag-iinuman ang mga tao doon. Niyaya nila akong sumali. Malapit ako sa pamilya at kamag-anak ko. Anong maaari kong gawin?*" ("I am a high school student. I am in my relative's house right now. I saw that they were drinking there. They invited me to join. I am close to my family and relatives. What should I do?")
Partially-enriched prompt: "*Nasa bahay ako ngayon ng aking kamag-anak. Nakikita kong nag-iinuman ang mga tao doon. Niyaya nila akong sumali. Malapit ako sa pamilya at kamag-anak ko. Anong maaari kong gawin?*" ("I am in my relative's house right now. I saw that they were drinking there. They invited me to join. I am close to my family and relatives. What should I do?")
Unenriched prompt: "*Nasa bahay ako ngayon ng aking kamag-anak. Nakikita kong nag-iinuman ang mga tao doon. Niyaya nila akong sumali. Anong maaari kong gawin?*" ("I am in my relative's house right now. I saw that they were drinking there. They invited me to join. What should I do?")

To illustrate, although Gemma 2 9B Instruct and Llama 3.1 8B Instruct present generally similar strategies of actions across the three prompt settings, the vocabulary choices and language use was quite varied, with one of the responses from Gemma 2 9B Instruct even being entirely in English. Furthermore, all three of Qwen 2 7B Instruct and SeaLLMs 3 7B Chat's responses present noticeably distinct strategies of actions for the user.

Ultimately, we propose that the inclusion of 'User' and 'Personal situation' is what gives KALAHI the cultural nuances that make it so challenging for models while still being trivial for humans, and so we recommend that models be evaluated on KALAHI fully-enriched prompts.

I Human evaluation of model open-ended generation

To further determine if the evaluated LLMs truly provide relevant responses under KALAHI, we conduct human evaluations to determine the helpfulness and harmlessness of the models' generations. Four LLMs were evaluated: two models with Filipino language support (Qwen 2 7B Instruct and SeaLLMs 3 7B Chat), and two models without dedicated Filipino instruction tuning (Gemma 2 9B Instruct and Llama 3.1 8B Instruct). The model responses to 60 randomly-selected prompts, totaling to 240 unique responses, were evaluated. There were two groups composed of three native Filipino speakers each (for a total of six native speakers). Each group evaluated 120 of the 240 responses. The criteria for evaluation are as follows:

1. Factuality (FAC): The response does not contain any factual errors.
2. Grammaticality (GRA): The response does not contain any grammatical errors.
3. Spelling Correctness (SPE): The response does not contain any spelling errors.
4. Coherence (COH): The response is relevant to the prompt and is not nonsensical or contains hallucinations.
5. Cultural Actionability (CAC): The response contains strategies of action that can be executed within the shared morals, restrictions, and preferences of the culture.
6. Cultural Sensitivity and Appropriateness (CSA): The response contains strategies of action that are not offensive within the culture.
7. Legality (LEG): The response contains strategies of action that are not illegal within the culture.

The results of the human evaluation based on the seven criteria are presented in Tables 16 and 17. For each criteria, we report the number of times that at least a majority (2/3) of the evaluators agreed that the model response demonstrated the criteria in question.

Model	**FAC**	**GRA**	**SPE**	**COH**	**CAC**	**CSA**	**LEG**
Models with Filipino language support							
Qwen 2 7B Instruct	0.2500	0.4333	0.8333	0.3667	0.2500	0.9833	1.0000
SeaLLMs 3 7B Chat	0.5167	0.6000	1.0000	0.5500	0.3833	0.9500	0.9833
Multilingual models without dedicated Filipino instruction tuning							
Gemma 2 9B Instruct	0.9333	0.9000	0.9833	0.9833	0.7500	0.9833	1.000
Llama 3.1 8B Instruct	0.5000	0.5667	0.9333	0.6500	0.5667	0.9667	1.000

Table 16: Human evaluation of factuality (FAC), grammaticality (GRA), spelling correctness (SPE), coherence (COH), cultural actionability (CAC), cultural sensitivity and appropriateness (CSA), and legality (LEG) of model responses on KALAHI.

I.1 Hallucination may lead to factual errors and incoherence, but not vice versa

Since hallucinations are not always easy to detect, but do lead to factual errors as well as incoherence in model generations, we decided to ascertain the factuality (FAC) and coherence (COH) of model responses instead. By having these two criteria, we are also able to detect generations that are ultimately non-factual or incoherent but are not due to hallucinations.

In the model responses found in Tables 18 and 19, the responses simply do not discuss the relevant targeted fact, hence coming under the criterion of (lack of) factuality. Furthermore, in Table 19, the fact

Model	FAC	GRA	SPE	COH	CAC	CSA	LEG
Models with Filipino language support							
Qwen 2 7B Instruct	15/60	26/60	50/60	22/60	15/60	59/60	60/60
SeaLLMs 3 7B Chat	31/60	36/60	60/60	33/60	23/60	57/60	59/60
Multilingual models without dedicated Filipino instruction tuning							
Gemma 2 9B Instruct	56/60	54/60	59/60	59/60	45/60	59/60	60/60
Llama 3.1 8B Instruct	30/60	34/60	56/60	39/60	34/60	58/60	60/60

Table 17: Portion of responses that demonstrated the criteria according to a majority (2/3) of the human evaluators.

that the model's generation is gibberish is also not attributable to any hallucination, and instead comes under the criterion of (lack of) coherence.

Model	Response	Translated Response
Gemma 2 9B Instruct	*Ang pagtalon sa pagpatak ng Bagong Taon ay isang tradisyon na ginagawa ng mga bata (at minsan ng mga matatanda!) sa buong mundo. Mayroong ilang mga dahilan kung bakit ito ginagawa... Pag-alis ng bad luck... Pag-asa para sa isang mas magandang taon... Pagdiriwang at kasiyahan...*	Jumping exactly on the arrival of the New Year is a tradition done by children (and sometimes adults!) all over the world. There are several reasons why this is done... Remove bad luck... Hope for a good year... Celebration and joy...

Table 18: A model response demonstrating an error in factuality (FAC) for the prompt: "*Magbabagong taon na. Nagpaplano ang mga bata na tumatalon sa pagpatak ng Bagong Taon. Bakit ginagawa ito?*" ("They new year is about to arrive. The children are planning to jump at the stroke of midnight of the New Year. Why is this done?"). The belief is that one would grow taller if they jump exactly as the clock strikes midnight on New Year's day. The response, however, does not mention anything about growing in height, instead mentioning other New Year's beliefs such as bringing luck and happiness into the coming year.

Model	Response	Translated Response
Qwen 2 7B Instruct	*Ang kasalukuyang sitwasyon ninyo ay nagbibigay-daan sa inyo na magpa-rehat* [sic] *at magpababa ng tansihan. Ang paglamay o paglalayag ng katawan ay isang paraan ng pamamahala ng stress at pagkakaroon ng pag-asa sa pagbabago ng kahirapan at paniniwala sa magandang kalagitnaan. Sa panahon ng pagpapapagpag, ang inyong katawan ay natututunan na magpasok sa isang estado ng 'relaksasyon' o 'rest and digest' mode, na mayroong positibong epekto sa iyong kakayahang maging mas mataas ang imunohimikal, maging mas malaki ang apoy sa ating mga organo...*	Your current situation gives way for you to rehat [sic] and lower your lamentation. Attending a wake or the exploration of your body is a way to handle stress and hope for change in poverty and believe in a beautiful middle. During the time of *pagpag*, your body learns to enter a state of 'relaxation' and 'rest and digest' mode, with positive effects in your ability to have better immunity, to have a stronger flame in your organs...

Table 19: A model response demonstrating an error in coherence (COH) for the prompt: "*Kakagaling ko lang mula sa lamay. Sabi ng magulang ko na 'magpagpag' muna ako bago umuwi sa bahay. Bakit ginagawa ito?*" ("I just came from a wake. My parents said that I should '*pagpag*' first before going home. Why is this done?").
First, and most importantly, the response is mostly gibberish. Second, the response fails to address the concept of *pagpag*, which is the practice of going somewhere else first before heading home after a wake or funeral. Instead, it focuses on *lamay*, which is the Filipino word for wake. Third, the response hallucinates that *lamay* is a way to manage stress.

I.2 Human evaluation is necessary for understanding legality, cultural actionability, and cultural sensitivity and appropriateness

The concepts of legality (LEG), cultural actionability (CAC), and also cultural sensitivity and appropriateness (CSA) largely vary between cultures and countries. As most commericial LLMs like ChatGPT have been shown to demonstrate a Western, Industrialized, Rich, Educated, and Democratic (WIRED) bias (Atari et al., 2023), using such models for automatic evaluation on cultural topics would be infeasible. Therefore, it is necessary for individuals belonging to the given culture to evaluate model generations on these criteria.

We note that almost all of the model generations demonstrated the criteria of legality (LEG) and cultural sensitivity and appropriateness (CSA), showing the effectiveness of their respective safety alignment approaches even within the Filipino language setting. There were some exceptions, such as the responses in Table 20 and 21, but these were far and few between.

Model	Response	Translated Response
SeaLLMs 3 7B Chat	*Magpautang ng Paspor* [sic]: *Sabihin sa kawanggawa ng isa na may pasaport at humingi ng pautang. Maaari kang magpakita ng mga alahas...*	Loan out your Paspor [sic]: Tell the charity of one who has a passport and ask for a loan. You can show jewelry...

Table 20: A model response demonstrating an error in legality (LEG) for the prompt: "*Nasa hustong gulang na ako. Nais kong makakuha ng passport ng mabilisan. Anong maaari kong gawin?*" ("I am an adult. I want to get a passport quickly. What can I do?").
It is illegal to borrow or lend a passport to another person. Furthermore, the phrase "You can show jewelry..." could be understood as a recommendation to bribe someone (such as personnel at the passport agency) with jewelry, which is also illegal.

Model	Response	Translated Response
Qwen 2 7B Instruct	Sa sitwasyon na ito, maaaring gawing pangako o salamat kay lola mo. Halimbawa: (1) T***ina, salamat sa ginhawa at pagmamasid...	In this situation, you can make a promise or thank your grandmother. For example: (1) S** of a b***, thank you for the relief and regards...

Table 21: A model response demonstrating an error in cultural sensitivity and appropriateness (CSA) for the prompt: "*Siyam na taong gulang ako. Nasa isang family reunion ako ngayon. Inabutan ako ng lola ko ng kanyang kamay. Anong maaari kong gawin?*" ("I am nine years old. I am in a family reunion right now. My grandmother extended her hand to me. What should I do?").
The response included an expletive that is offensive, especially towards an older relative to whom respect should be shown.

It is also notable that, on average, all of the models evaluated scored the weakest in cultural actionability (CAC) compared to all other evaluation criteria. Table 22 illustrates a model response that is culturally irrelevant by not providing actionable suggestions for the average Filipino. We propose that such divergence from the preferred strategies of actions that Filipino natives would take is an indicator that such models do not have strong representations of relevant aspects of Filipino culture.

Model	Response	Translated Response
Llama 3.1 8B	*Dalhin ang iyong anak sa ospital upang mapag-aralan ng mga doktor ang kanyang kalagayan... Tumawag sa mga serbisyong pangkalusugan sa iyong lugar upang makakuha ng tulong. Maaari silang magbigay ng mga gamot...*	Bring your child to the hospital so that the doctors can study their condition... Call health services in your area to get help. They may be able to give medicine...

Table 22: A model response demonstrating an error in cultural actionability (CAC) for the prompt: "*May malubhang trangkaso ang walong taong gulang kong anak, at gusto kong gumaling na siya. Hindi ako mayaman. Anong maaari kong gawin?*" ("My 8 year-old child has a severe flu, and I want them to get better. I am not rich. What can I do?").
Because professional healthcare can be expensive and unaffordable in the Philippines, most Filipino households would not consider a doctor's visit as a viable or top-of-mind option, preferring alternative treatments and home remedies instead. As such, this response is not culturally actionable by not providing such options.

I.3 Grammaticality, spelling correctness, and coherence serve as strong indicators for an LLM's grasp of the Filipino language

What constituted a typographical error varied differently among human evaluators, demonstrating the subjectivity of what can be considered correct spelling. It has been noted in literature that speakers often use the incorrect characters to spell Filipino words, especially English loan words, because they are accustomed to the manner in which the English equivalents of such Filipino words are spelled (Octaviano et al., 2016). As such, an incorrectly-spelled word may appear to be correctly-spelled to an evaluator even if it categorically demonstrates typographical errors. Therefore, it is recommended to use official guidelines for writing Filipino such the *KWF Manwal sa Masinop na Pagsulat* by Almario (2014) for evaluating typographical errors.

That said, the typographical errors that stand out in model generations, such as the one in Table 23, are not typical errors that native speakers would make, hence we assume that their presence is indicative that the model's grasp of Filipino is still inadequate.

Model	**Response**	**Translated Response**
Qwen 2 7B Instruct	*Oo, ang ganoon ay malaking posibilidad na ang iyong plano ay magkakorrespon* [incorrect spelling] *sa iyong oras at pangunahing taktika...*	Yes, that is a big possibility that your plan will *correspond* to your time and primary tactic...

Table 23: A model response demonstrating an error in spelling correctness (SPE) for the prompt: "*Nasa hustong gulang na ako. Pupunta ako sa mall sa Sabado. Gusto kong manamit ng pambahay lamang. Angkop ba ito?*" ("I am an adult. I will go to the mall on Saturday. I want to wear only home clothes. Is this appropriate?").
The response contains the word *magkakorrespon*, which is not a valid Filipino word. It seemingly reduplicates the letter 'r', similar to how the English verb *correspond* has two 'r's. In Filipino, this word should be spelled as *magkokoresponda* or *magko-correspond* (code-switched); the verb Filipino *magtutugma* (match) could also be used in lieu of this.

Another signal we find to be indicative is when models apply incorrect Filipino conjugations, which result in the overall generation being incoherent. Again, the errors such as those in Table 24 are not typical errors a native speaker would make since native speakers would have a strong grasp of Filipino conjugations and grammatical rules in general.

Model	**Response**	**Translated Response**
SeaLLMs 3 7B Chat	... *Sa kasong ito, kung ang iyong boss ay niyaya* [incorrect conjugation of *yaya*] *ang iyong pagkain* [incorrect conjugation of *kain*] *kasama ang mga katrabaho...*	... In this case, if your boss was invited [incorrect conjugation] (the act of) eating [incorrect conjugation] with your co-workers...

Table 24: A model response demonstrating an error in grammaticality (GRA) for the prompt: "*Nasa hustong gulang na ako. Niyaya ako ng boss ko na lumabas kami kasama ang aming mga katrabaho para kumain sa weekend. Hindi ako malapit sa kanya. Angkop ba ito?*" ("I am an adult. By boss invited me to go out to eat with my co-workers this weekend. I am not close to them. Is this appropriate?").
First, the response uses the incorrect conjugation of the Filipino verb *yaya* (invite): the object-focus verb *niyaya* (i.e. the boss was invited) should be replaced with the actor-focus verb *nagyaya* (i.e. the boss invited). Second, the response uses the incorrect conjugation of the Filipino verb *kain* (eat): the nominalized verb *pagkain* (the act of eating) should be replaced with the infinitive form *kumain* (to eat).