SeaEval for Multilingual Foundation Models: From Cross-Lingual Alignment to Cultural Reasoning

Bin Wang[1*], **Zhengyuan Liu**[1*], **Xin Huang**[1], **Fangkai Jiao**[2], **Yang Ding**[1], **AiTi Aw**[1], **Nancy F. Chen**[1,3]
[1]Institute for Infocomm Research (I^{2}R), A*STAR, Singapore
[2]Nanyang Technological University, Singapore
[3]Centre for Frontier AI Research (CFAR), A*STAR, Singapore
{wang_bin,liu_zhengyuan,nfychen}@i2r.a-star.edu.sg

Abstract

We present *SeaEval*, a benchmark for multilingual foundation models. In addition to characterizing how these models understand and reason with natural language, we also investigate how well they comprehend cultural practices, nuances, and values. Alongside standard accuracy metrics, we investigate the brittleness of foundation models in the dimensions of semantics and multilinguality. Our analyses span both open-sourced and closed models, leading to empirical results across classic NLP tasks, reasoning, and cultural comprehension. Key findings indicate (1) Many models exhibit varied behavior when given paraphrased instructions. (2) Many models still suffer from exposure bias (e.g., positional bias, majority label bias). (3) For questions rooted in factual, scientific, and commonsense knowledge, consistent responses are expected across multilingual queries that are semantically equivalent. Yet, most models surprisingly demonstrate inconsistent performance on these queries. (4) Multilingually-trained models have not attained "balanced multilingual" capabilities. Our endeavors underscore the need for more generalizable semantic representations and enhanced multilingual contextualization. *SeaEval* can serve as a launchpad for more thorough investigations and evaluations for multilingual and multicultural scenarios.[1]

1 Introduction

Over the past years, there has been rapid development of large language models (LLMs), also known as a type of foundation models (FMs) (Bommasani et al., 2021), demonstrating their generalizability and adaptability across various downstream tasks (Scao et al., 2022; Chowdhery et al., 2022; OpenAI, 2023; Touvron et al., 2023b; Wang et al., 2023a). The proliferation of LLMs has raised urgent requirements for extensively evaluating their performance in various contexts, and understanding their limitations (Wei et al., 2024). Therefore, recent efforts on LLM evaluation are focusing on more challenging and human-centric tasks including complex reasoning (Clark et al., 2018; Zellers et al., 2019; Hendrycks et al., 2021a) and domain-knowledge-intensive problems (Hendrycks et al., 2021a; Lin et al., 2022a; Zhong et al., 2023; bench authors, 2023), math problems (Zhong et al., 2023), human exams (Clark et al., 2018; Zhong et al., 2023; OpenAI, 2023), and using LLMs as judges for open question answering (Zheng et al., 2023).

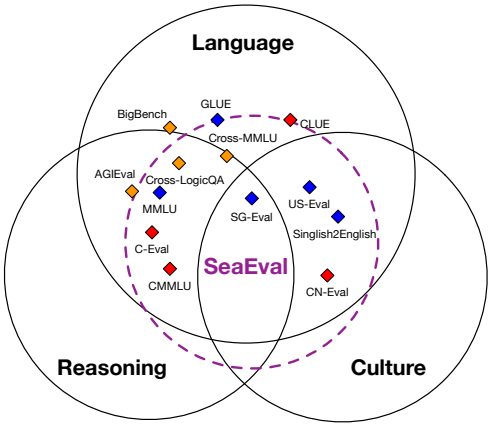

Figure 1: *SeaEval* for multilingual foundation models. English is represented by the color blue, Chinese by red, and a mix of multiple languages by yellow. *SeaEval* includes the datasets within the dotted-line circle.

Compared to other species on earth, humans are adept in using language to symbolize and encode thoughts, represent and document knowledge, and express emotions. Whether it is spoken or written form, language has become a default means for humans to conduct and communicate our reasoning process and logic (Logan, 1986; Li and Gleitman, 2002). Three variables are language, region, and culture. Our cultural behavior, rituals, and values are often embodied and represented through lan-

[1]Datasets, evaluation toolkit, and leaderboard are available at https://github.com/SeaEval/SeaEval.
*: Equal contribution.

guage (Kramsch, 1991; Ji et al., 2004). Therefore, the capabilities of language models go beyond language per se. Evaluating multilingual language model should incorporate deep cultural understanding and reasoning. To this end, we expand the current evaluation criteria to cover more linguistic and cultural contexts (Ahuja et al., 2023; Lai et al., 2023), which are under-explored in prior research.

Another important yet under-explored aspect of evaluating multilingual foundation models is their knowledge transferability in language dimensions. Multilingual foundation models are expected to demonstrate consistent performance across languages in the context of region-invariant common knowledge (Zhu et al., 2023b). Monolingual or bilingual benchmarks cannot adequately capture this aspect. Therefore, we introduce cross-lingual consistency evaluation with tailored datasets and specialized metrics.

SeaEval aims to assess the capabilities of multilingual foundation models in four dimensions: (a) Classic NLP tasks that are centered around language understanding and generation; (b) Complex reasoning; (c) Cultural understanding and reasoning; and (d) Cross-lingual knowledge transfer and contextualization. *SeaEval* encompasses a total of 28 datasets, including 6 new datasets constructed for cultural reasoning and cross-lingual consistency assessments.

Key findings from our investigations and experimental results indicate: (1) Most models show varying responses to rephrased instructions. (2) Exposure bias (e.g., positional bias, majority label bias) of label arrangements still prevails. (3) Most models give inconsistent answers when the same fact-based questions are asked in different languages. This counter-intuitive observation suggests semantic representations are not generalized in the multilingual context. (4) Multilingually-trained models fall short of achieving "balanced multilingual" proficiency. To sum up, our contributions are multifold:

- We offer fresh insights into multilingual foundation models and their evaluations.
- We introduce 7 new datasets for assessing cultural understanding and cross-lingual consistency, along with tailored metrics to fill existing gaps in model evaluation.
- We present a comprehensive evaluation benchmark derived from extensive experiments across models, tasks, datasets and metrics. This framework facilitates in-depth exploration of multilingual and multicultural tasks using foundation models.

2 Essential Properties of Multilingual Foundation Models and Benchmarks

In this section, we delve into the desired properties of multilingual foundation models and explore the ideology for crafting a comprehensive benchmark.

2.1 Multilingual Foundation Model

Multilingual foundation models should possess additional properties beyond monolingual models to effectively handle diverse languages and cultural contexts. Here are the key properties:

Multilinguality The applicability of multilingual LLMs can be elevated when they demonstrate proficiency across diverse languages, including low-resource languages and their dialects (Wu et al., 2016; Kulshreshtha et al., 2020). The advantage of multilingualism is that it allows these models to bridge the linguistic gaps that exist between different cultures and communities. Multilingual capability is not a combination of various monolingual capabilities (Ye et al., 2023); instead, it represents a holistic approach to understanding and processing languages. For example, the model should be adept at bilingual tasks such as machine translation and code-switching scenarios, which offers the advantage of preserving linguistic and cultural information.

Reasoning Capability Reasoning has long been treated as a complex yet essential capability in cognition that goes beyond fundamental language understanding (McCarthy, 2022), which can be demonstrated via extracting critical information from the text environment and drawing correct conclusions (Nilsson, 1991). The advancement of natural language reasoning has evolved from explicit and superficial reading comprehension and natural language inference (Rajpurkar et al., 2016; Wang et al., 2019; Wang and Li, 2023) to encompass implicit, complex, and specific reasoning capabilities such as multi-hop reasoning, numerical reasoning, and logical reasoning (Lai et al., 2017; Huang et al., 2019; Jiao et al., 2022). While reasoning can be challenging due to the various relations and expressions involved, which are difficult to transcribe into symbolic or formal languages, FMs are shown

Language	English	Chinese	Indonesian	Spanish
Question	*Please choose the correct answer for the following question.* *When white light passes through a prism, the light that bends more than green is?* *(A) Red* *(B) Yellow* *(C) Blue* *(D) None of these*	回答下面问题，选择正确答案。 当白光通过棱镜时，比绿光弯曲更多的光是? (A) 红色的 (B) 黄色的 (C) 蓝色的 (D) 都不是	*Silakan pilih jawaban yang benar untuk pertanyaan berikut.* *Ketika cahaya putih melewati sebuat prisma, cahaya manakah yang memiliki sudut deviasi lebih besar daripada cahaya hijau?* *(A) Merah* *(B) Kuning* *(C) Biru* *(D) Tak ada satupun*	*Por favor elija la respuesta correcta para la siguiente pregunta.* *Cuando la luz blanca pasa a través de un prisma, la luz que se desvía más que la verde es* *(A) Rojo* *(B) Amarillo* *(C) Azul* *(D) Ninguna de las anteriores*
Answer	*The correct answer is:* *(C) Blue*	(A) 红色的 ***In English**: (A) Red*	*(D) Tak ada satupun* ***In English**: (D) None of them*	*La respuesta correcta es:* *(A) Rojo* ***In English**: (A) Red*
Correctness	✓	✗	✗	✗

Table 1: An example from Cross-MMLU dataset for evaluating cross-lingual consistency. Outputs are from ChatGPT. The answers are inconsistent for the same question posed in different languages. This inconsistency highlights insufficient alignment across languages, leading to suboptimal multilingual contextualization and representations.

to serve as a proxy to compress abundant knowledge, and solve various tasks following human instructions with less specialization (Wei et al., 2022; Wang et al., 2023b; Jiao et al., 2023).

Cultural Understanding Language is deeply tied to culture and local norms (Pennycook, 2006). The meaning of linguistic elements can differ considerably across cultures. Cultural understanding capability can help large language models better interpret content with local communication conventions (Zampieri et al., 2020) and avoid stereotypes and biases. In the context of philosophy, language, reasoning, and culture are considered three important pillars that play a significant role in shaping people's understanding of the world (Ji et al., 2004; Kramsch, 2014) as depicted in Figure 1. They intersect and influence one another in various ways across studies in linguistics, philosophy, and psychology. Logan (1986) presents a provocative proposal that language can be used to account for cultural differences in reasoning styles. Therefore, in pursuit of advancing multilingual foundation models, it is desired that models not only acquire proficiency across languages but also gain a profound comprehension of cultural concepts influencing human behaviors. An illustrative example highlighting the impact of local cultural conventions is shown in Table 2, where a particular model must draw insights from locally sourced content to address this problem properly.

Cross-Lingual Knowledge Transfer An important advantage of encompassing multiple languages is the ability to access information from various language resources simultaneously, a characteristic that is also desired in multilingual foundation models. An effective cross-lingual knowledge transfer method can significantly enhance model capabilities across all languages, as they mutually reinforce each other. Additionally, world knowledge is typically dispersed in various languages and regions and may not be easily accessible in a single source, which demonstrates the need for cross-lingual knowledge transfer. On the other hand, some world knowledge should be kept consistent across different languages, such as factual, scientific, and commonsense knowledge. In Table 1, we see an illustration of the same question posed in 4 languages, revealing inconsistent answers attributed to inadequate cross-lingual alignments of multilingual foundation models. Up to 16 languages are tested to illustrate this phenomenon as depicted in Section F.

Question	*Which drink in Singapore has the highest calories?* *(A) Teh O* (B) *Teh Siew Dai* (C) *Kopi* (D) *Kopi C*
Multicultural Reasoning Steps	**Multilingual Understanding** (Hokkien) Teh = Tea (Cantonese) Siew Dai = Less Sweet/Sugar (Malay) Kopi = Coffee **Cultural/Personal Preferences** Teh = Tea + Condensed Milk + Sugar Teh O = Tea + Sugar Kopi = Coffee + Condensed Milk Kopi C = Coffee + Evaporated Milk + Sugar **Reasoning with Dietary Knowledge** Condensed milk = Sweetened = Sugar was Added Sugar = Calories Pure Tea or Coffee = Almost No Calories
Answer	(C) Kopi

Table 2: An example from SG-Eval dataset. To accomplish the task, one needs to employ reasoning that incorporates multilingual and cultural knowledge.

2.2 Multilingual Benchmarks

Motivated by the preceding discussion regarding the desired model characteristics, we introduce the targeted aspects for benchmarks:

Monolingual and Cross-lingual Capabilities The focus on monolingual tasks ensures the model's proficiency in comprehending and generating text within a single language. The cross-lingual tasks, such as machine translation and code-switch comprehension, can access the communication capabilities across different languages, reflecting a comprehensive understanding of multilingual contexts. In terms of evaluation aspects, both fundamental NLP capabilities and complex reasoning capabilities should be examined under monolingual and cross-lingual settings.

Knowledge Transfer Ability Language-related knowledge can be categorized into: 1) cultural knowledge and local norms tied to language and 2) common (universal) knowledge. Cultural knowledge refers to language-related information that is specific to a particular culture, community, or region. It includes the nuances, customs, and norms associated with language use within a specific cultural context. Common knowledge is widely applicable across languages and communities, encompassing factual, scientific, and real-world knowledge, etc (Hendrycks et al., 2021a). When designing evaluation benchmarks, it is essential to include a diverse set of language-related cultural tasks while also evaluating how effectively the universal knowledge is shared across different languages. Since such datasets are not readily accessible, this evaluation aspect is severely constrained in existing benchmarks.

Robustness and Stability The robust context modeling and stable output generation are important to ensure LLMs work as intended (functionality) when applied to real-world applications (reliability) (Haduong et al., 2023). When built on the auto-regressive framework, language models are originally trained to predict the next token given a sequence of previous ones, and their in-context learning and zero-shot inference performance depends on the prompts they receive (Ouyang et al., 2022). Consequently, minor variations of the input can possibly lead to distinct outputs with unpredictable formats. In particular, since FMs do not attain "balanced multilingual" capabilities, they are more sensitive to input variations such as multilingual and code-switch under real-world scenarios. Therefore, recognizing the models' instruction sensitivity should be a crucial aspect of the evaluation framework.

3 SeaEval

In this section, we present our *SeaEval* benchmark from task selection, data curation, to evaluation protocols. Besides the evaluation of fundamental capabilities and complex reasoning, we also include the evaluation tasks on cultural understanding and cross-lingual alignment. The datasets are summarized in Table 3.

3.1 Task Selection

Fundamental Language Capabilities The fundamental capabilities can be evaluated by a combination of classic NLP tasks of language understanding and generation. To ensure the diversity regarding both task and language, we collected 18 representative datasets from 5 languages. Previous studies (Shi et al., 2022; Ye et al., 2023) show that English-centric LLMs demonstrate certain multilingual transfer ability, where the skills learned from one source language can be readily transferred to other languages. Therefore, for discriminative tasks, we select 8 tasks from the GLUE benchmark (Wang et al., 2019), including SST-2, COLA, QQP, QNLI, MNLI, WNLI, RTE, and MRPC. Furthermore, we incorporate DREAM for English dialogue comprehension, OCNLI and C3 for Chinese comprehension, and Indo-Emotion dataset (Saputri et al., 2018; Wilie et al., 2020) to gauge emotion comprehension in Indonesian. To build a generative task basis, we include translation and summarization datasets from FLoRes, SAMSum, and DialogSum.

Complex Reasoning Classic NLP benchmarks (e.g., GLUE, SQuAD) primarily focus on text understanding rather than complex reasoning abilities aligned with intricate real-world scenarios. As language models continue to grow in size and complexity, it becomes increasingly important to assess their abilities in performing complex reasoning and problem-solving tasks that humans typically excel at (Wong et al., 2023). Therefore, here we add evaluation datasets from recent representative human-centric benchmarks, which are derived from high-standard and professional exams. We include the MMLU dataset to assess knowledge comprehension in English. To assess the reasoning capability in a multilingual setting, we include C-Eval,

Dataset	Task Description	Languages	Metrics	# of Samples
Multicultural and Multilingual Understanding				
SG-Eval▲	Cultural Understanding	Eng	Accuracy	102
US-Eval▲	Cultural Understanding	Eng	Accuracy	102
CN-Eval▲	Cultural Understanding	Zho	Accuracy	105
PH-Eval▲	Cultural Understanding	Eng	Accuracy	100
Singlish2English▲	Multilingual Translation	Eng, Singlish	BLEU	546
Cross-Lingual Consistency				
Cross-MMLU▲	Reasoning	Eng, Zho, Ind, Spa, Vie, Zsm, Pil	AC3	900
Cross-LogiQA▲	Logic Reasoning	Eng, Zho, Ind, Spa, Vie, Zsm, Pil	AC3	1,056
Complex Reasoning				
MMLU (Hendrycks et al., 2021b)	Mixed Knowledge	Eng	Accuracy	857
C-Eval (Sun et al., 2019)	Subject Knowledge	Zho	Accuracy	1,346
CMMLU (Li et al., 2023a)	Subject Knowledge	Zho	Accuracy	280
ZBench (Chen and et al., 2023)	Subject Knowledge	Zho	Accuracy	33
Classic NLP Tasks				
FLoRes-Lang2eng (Guzmán et al., 2019)	Translation, Bilingual	Ind, Vie, Zho, Zsm, Eng	BLEU	3,988
Ind-Emotion (Saputri et al., 2018)	Sentiment Analysis	Ind	Accuracy	300
OCNLI (Hu et al., 2020)	Textual Entailment	Zho	Accuracy	300
C3 (Sun et al., 2020)	Reading Comprehension	Zho	Accuracy	300
SAMSum (Gliwa et al., 2019)	Summarization	Eng	ROUGE	300
DialogSum (Chen et al., 2021)	Summarization	Eng	ROUGE	300
DREAM (Sun et al., 2019)	Dialogue Comprehension	Eng	Accuracy	300
8 GLUE Tasks (Wang et al., 2019)	Fundamental NLP	Eng	Accuracy	2,148
29 Datasets	Mixed	8	Mixed	13,263

Table 3: Datasets from **SeaEval**. Language abbreviations are from ISO 639-3 standard, where Eng, Zho, Ind, Spa, Vie, Zsm, and Fil indicate English, Chinese (Mandarin), Indonesian, Spanish, Vietnamese, Malay (Malaysian), and Filipino, respectively. Examples from our newly collected datasets (▲) are shown in Table 5.

CMMLU, and ZBench, which are specifically tailored for evaluating intricate reasoning in Chinese.

Multilingual and Cultural Understanding An effective multilingual language model is trained with text corpus from diverse sources. It enables the model to acquire cultural knowledge related to languages, which is important when serving users from different cultural backgrounds. In order to assess the model's cultural comprehension abilities, we manually construct 4 datasets containing multiple-choice questions that encompass 4 distinct regions: the United States (English), Singapore (English), China (Chinese), and the Philipines (English). The corresponding datasets are US-Eval, SG-Eval, CN-Eval, PH-Eval. Unlike monolingual models, multilingual models should demonstrate a strong capability for effectively transferring common knowledge. Therefore, we introduce two datasets, Cross-MMLU and Cross-LogiQA, to evaluate this feature across 7 diverse languages: English, Chinese, Indonesian, Spanish, Vietnamese, Malay, and Filipino.

3.2 Data Curation

Considering the size of LLMs, evaluation on the full test set can incur significant computational and economic expenses. Therefore, for existing datasets on evaluating model's fundamental capability and complex reasoning, we randomly sampled a subset. The numbers are listed in Table 3, which results in over 13k samples in total.

The output formats of autoregressive language models (e.g., GPT) cannot be easily controlled for open-ended tasks, making it difficult to assess the accuracy of their predictions. Consequently, in order to quantitatively evaluate their performance, we have transformed all the discriminative datasets (e.g., emotion classification, natural language inference, dialogue comprehension) into multiple-choice questions. While, for generative tasks such as summarization and translation, we have retained the original evaluation process, as it relies on word-matching metrics and human-annotated references are readily applicable.

Cultural Reasoning There are no publicly available datasets for explicitly evaluating cultural knowledge in different regions. To effectively evaluate such knowledge, certain criteria should be met. First, the knowledge should originate directly from respective regions, distinct from widespread content. Second, it should encompass an understanding of the intricate norms of each culture under

examination. Third, certain cultural expressions can be challenging to fully convey in another language, making it preferable to retain the knowledge in its original language.

Therefore, we hired linguistic experts to construct datasets to evaluate the knowledge from three distinct regions, including the United States (US-Eval), Singapore (SG-Eval), China (CN-Eval) and the Philipines (PH-Eval). For each dataset, over 100 questions are sourced from a variety of channels, including local residencies' proposals, government websites, historical textbooks and exams, local cultural heritage materials, and pre-existing academic research datasets. CN-Eval and US-Eval also include questions carefully selected from MMLU, C-Eval and CMMLU datasets. Meanwhile, Singapore serves as an exceptional illustration, blending a harmonious fusion of diverse Southeast Asian cultures, enriched by a wealth of local content (Deterding, 2007; Liu et al., 2022; Wang et al., 2024a). We also introduce a new dataset for Singlish to standard English translation with 546 sentences. Singlish incorporates elements of various languages, including Malay, Chinese dialects, and Tamil, and often includes unique vocabulary, grammar, and pronunciation. It has distinct local characteristics and requires a deep understanding of local practices. The samples from each dataset are illustrated in Table 5.

Cross-Lingual Consistency As shown in Table 1, for existing multilingual LLMs, the same question posed in different languages leads to inconsistent answers, which is undesired for multilingual foundation models. To qualitatively evaluate the model's capability in cross-lingual consistency, we present two datasets: Cross-MMLU and Cross-LogiQA with paralleled questions in 7 languages: English, Chinese, Indonesian, Spanish, Vietnamese, Malay, and Filipino. The selected questions are carefully curated to test common knowledge (e.g. commonsense, scientific), which is universally acceptable and transferrable between languages. Cross-MMLU and Cross-LogiQA are originated from MMLU dataset (Hendrycks et al., 2021a) and LogiQA2.0 dataset (Liu et al., 2023), respectively. To prepare questions that do not have equivalents in the target language, we utilize Google Translate first and enlist native speakers to perform proofreading and editing. This approach helps prevent translation errors and ensures accurate expressions, avoiding any potential misinterpretations.

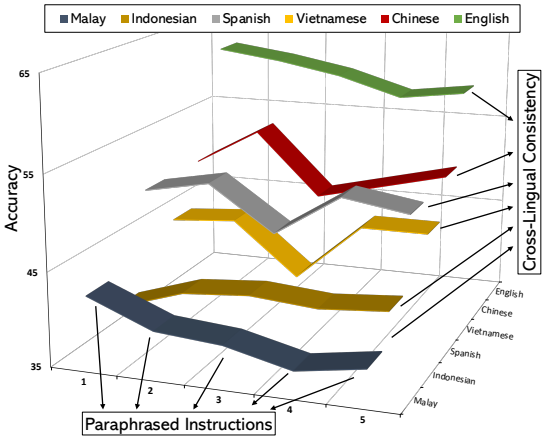

Figure 2: Two new evaluation protocols for multilingual foundation models in *SeaEval*. The performance result is taken from ChatGPT on Cross-LogiQA dataset.

3.3 Evaluation Protocols

Conventional benchmarks typically emphasize a single metric evaluation per dataset. As illustrated in Figure 2, it becomes apparent that they do not provide enough coverage in multilingual FM evaluation. Therefore, in addition to standard metrics, we introduce two new evaluation dimensions called instruction sensitivity and cross-lingual consistency to measure a model's stability across instructions and languages. Regarding standard evaluation metrics, we use accuracy scores for multiple-choice questions. In the case of translation assessments, we report the BLEU-4 score (Papineni et al., 2002), while for summarization tasks, we deploy the average of ROUGE-1/2/L scores (Lin, 2004).

Cross-Lingual Consistency Besides the standard *Accuracy* metric for evaluating multi-choice questions, we compute the cross-lingual *Consistency* score as a measurement of whether the answers are consistent for the same question in 7 different languages without considering the answer's correctness. Specifically, for a question set $Q = \{q^1, q^2, ..., q^N\}$, each question q^i is represented in 7 languages $q^i = \{q^i_{eng}, q^i_{zho}, q^i_{ind}, q^i_{spa}, q^i_{vie}, q^i_{msa}, q^i_{fil}\}$, and a^i_{lang} is model's answer to q^i_{lang}, the *Consistency* score is computed as

$$M_{\{l_1,l_2,..,l_s\}} = \frac{\sum_{i=1}^{N} \mathbb{1}_{\{a^i_{l_1}=a^i_{l_2}=..=a^i_{l_s}\}}}{N}$$

$$Consistency_s = \frac{\sum_{\{l_1,l_2,..,l_s\}\in C(s,q_i)} M_{\{l_1,l_2,..,l_s\}}}{C^s_7}$$

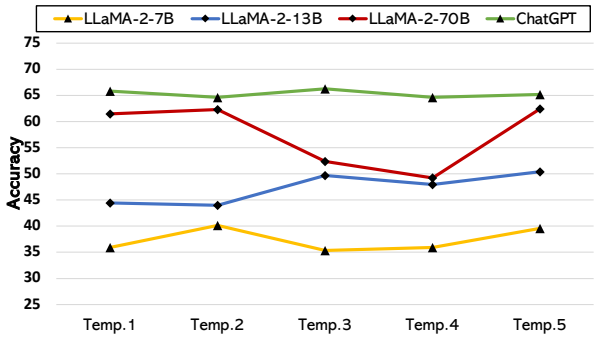

Figure 3: Performance on MMLU dataset with paraphrased instruction. Some models show large performance variances with paraphrased instruction templates.

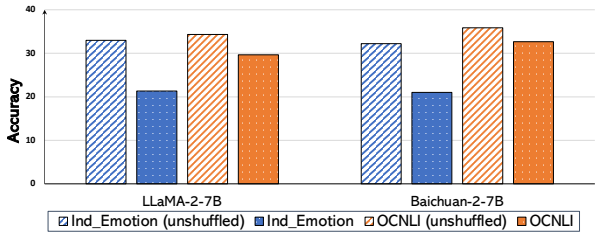

Figure 4: Effect on label order. Performance varies when labels are shuffled, revealing inherent label biases.

where $s \in [2, 7]$. It measures the answer's consistency of any combination of s languages. The model gets rewarded if it generates consistent answers across the sampled languages. The consistency requirement is enhanced to more languages with increased s. Given that both *Accuracy* and *Consistency* alone do not provide a comprehensive assessment of models' performance on cross-lingual datasets, we introduce the *AC3* score as a holistic measure, which is calculated as the harmonic mean of both scores:

$$AC3_s = 2 \cdot \frac{Accuracy \cdot Consistency_s}{Accuracy + Consistency_s}$$

where *AC3* is within range $[0, 1]$. We deploy *AC3* with $s = 3$ as the default value for Cross-MMLU and Cross-LogiQA datasets. Figure 8 illustrates the impact on variable s.

Instruction Sensitivity Early methods for training LLMs to follow instructions primarily use task instruction sets, which are compiled by combining task instruction templates with instances from standard NLP tasks (Chung et al., 2022). However, such approaches often fall short of capturing the intricacies of practical user instructions, as these instructions tend to originate from artificial NLP tasks designed to test specific aspects of machine capabilities. Real-world user instructions, on the other hand, are significantly more diverse and complex (Ouyang et al., 2022; Wang et al., 2024b), and it is necessary to evaluate the performance under varied instructions. Therefore, we build 5 human paraphrased instructions with NLP experts for each dataset. We show in Figure 3 about the LLaMA-2 and ChatGPT models on their performance with five instructions and witnessed that ChatGPT models are more robust to instruction paraphrases. Some instructions possess the ability to unlock the model's full potential, potentially surpassing its more efficient counterparts, which may lead to biased evaluation (***Our Finding 1***). Hence, it becomes crucial to utilize multiple instructions to obtain a more comprehensive assessment of model capabilities. Evaluating the model's resilience to paraphrased instructions is also a significant aspect. To report performance, we employ the median value derived from five instructions as the ultimate result.

Exposure Bias on Label Arrangements Recent work demonstrates that LLMs have inherently exhibited exposure biases from many factors (Fei et al., 2023) including majority label bias, recency bias, and common token bias (Zhao et al., 2021). In our study, we found the positional arrangement of labels is a potential source of exposure bias, especially for smaller-sized models. Figure 4 shows the results of LLaMA-2 and Baichuan-2 models on two datasets. We observe that some models are prone to rely on intrinsic biases of label arrangements when making predictions which lead to higher evaluation results. Ignoring such patterns could raise unanticipated advantages on specific models (***Our Finding 2***). Therefore, in *SeaEval*, we shuffle all labels whenever possible to avoid exposure biases on label arrangements. Note that for position-sensitive labels such as 'all above', we manually keep their order unchanged.

4 Evaluation Results and Discussion

We show the evaluation results on five datasets for cross-lingual consistency and cultural reasoning in Figure 5 and our key findings are as follows.

Firstly, GPT-4 demonstrates outstanding performance on most datasets, surpassing others by a substantial margin across cultures and languages, demonstrating its superior capability in handling multilingual tasks.

Second, becoming an expert in cultural knowledge necessitates extensive pre-training with a di-

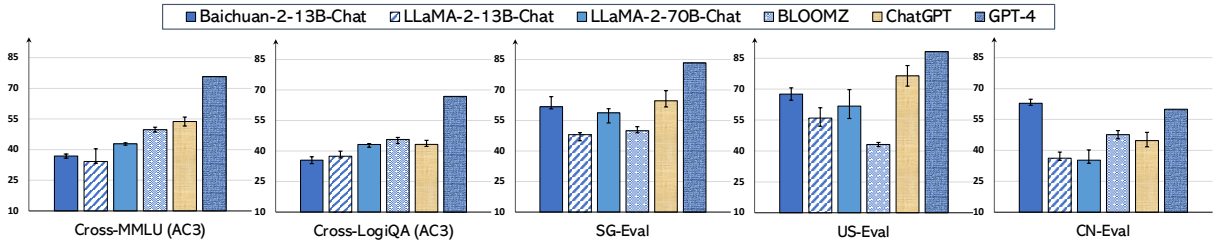

Figure 5: Evaluation results of six representative LLMs on a subset of *SeaEval*. AC3 and Accuracy scores are reported. The error bar covers performances from five different instruction templates.

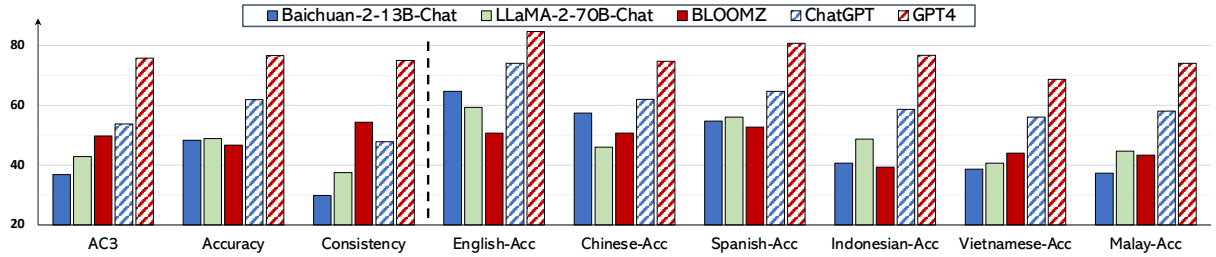

Figure 6: Evaluation results on Cross-MMLU. Both overall and language-specific scores are shown.

verse and extensive collection of multilingual textual data such as books, articles, websites, historical documents, and cultural artifacts. Baichuan-2 model has shown remarkable performance in understanding Chinese culture (CN-Eval), even outperforming GPT4. In contrast, LLaMA models are primarily focused on English, with approximately 90% of English pre-training data. This specialization makes them less proficient in handling multilingual and multicultural scenarios.

Detailed results regarding cross-lingual consistency are presented in Figure 6. The report includes comprehensive evaluation metrics: *AC3*, *Accuracy*, *Consistency*, and *Accuracy* for each language. The consistency score clearly demonstrates that BLOOMZ stands out for its better performance in aligning knowledge across languages, solidifying its position as a leading open-source multilingual foundational model. Even being the worst in overall accuracy, BLOOMZ surpasses ChatGPT in cross-lingual consistency, achieving a score of 54% compared to 47%. However, they are still showing unsatisfactory consistency scores, highlighting the inconsistency in the sharing of common knowledge across various languages (***Our Finding 3***). While GPT-4 achieves a 75% consistency score, it drops to 64% when $s = 6$ as shown in Figure 8, which suggests ample opportunity to further enhance cross-lingual knowledge alignment, aiming for optimal multilingual models.

Last, when assessing models' accuracy in individual languages, it is evident that their problem-solving capability in English usually surpasses that in other selected languages. This illustrates that the proficiency of models varies unevenly across different languages (***Our Finding 4***). Compared to high-resource languages, the performance of low-resource languages is inferior. For example, English, Chinese and Spanish rank within the top 5 out of 46 languages present in BLOOMZ corpus. Therefore, the multilingual foundation model's capability in low-resource languages needs to be further improved to match the more centralized languages. The disparity in cross-lingual consistency highlights the need for more robust alignment efforts. Particularly under low-resource constraints, enhancements in this aspect have the potential to elevate the overall performance across all languages through effective knowledge transfer, facilitating further development of multilingual language models (Kulshreshtha et al., 2020; Huang et al., 2023a; Zhu et al., 2023b; Muennighoff et al., 2023).

5 Conclusions

We introduced *SeaEval* benchmark for multilingual foundation model evaluation, grounded in comprehensive experimentation across languages, models, tasks, and datasets. *SeaEval* encompasses 29 datasets, including 7 new ones for cultural understanding and cross-lingual consistency. Our empirical analysis demonstrates four key findings on the capabilities of multilingual foundation models:

1) Sensitivity to paraphrased instructions; 2) Exposure bias of label arrangements, 3) Inconsistent performance across multilingual questions that are semantically equivalent, and 4) Imbalanced multilingual proficiency. These findings accentuate the importance of more generalizable semantic representations and enhanced multilingual contextualization. We hope that our endeavors in *SeaEval* can pave the way for more in-depth investigations into multilingual and multicultural tasks using foundation models.

Limitations

In this study, our primary focus is multilingual foundation models' language capabilities. Nonetheless, there remain several evaluation aspects to be included in order to provide a complete reflection of the capability of multilingual foundation models in practical applications.

First, there is a need for the inclusion of more languages and cultural reasoning datasets. Expanding the linguistic and cultural diversity within the benchmark is a resource-intensive endeavor, as the acquisition of suitable datasets for various languages and culture-related contexts can be challenging. Nevertheless, as we aspire for this benchmark to comprehensively cover a wide range of languages, there is a pressing need to explore automated methods for data collection. Such an approach can help ensure the acquisition of high-quality datasets while mitigating the resource-intensive nature of manual data curation, thereby enhancing scalability.

Second, *SeaEval* ensures a robust quantitative evaluation benchmark, incorporating datasets that facilitate more straightforward performance quantization. In real-world usage cases, foundation models are also used for information-seeking purposes, where users may pose subjective questions and engage in dialogues. This poses challenges in evaluating the faithfulness, expertise and engagement during interactions. Existing approaches adopt powerful FMs as the evaluation criteria which may not necessarily replicate the judgments from humans (Zheng et al., 2023), underscoring the necessity of practical automatic assessment approaches for open-ended questions.

Third, but certainly not the least, safety and efficiency are two important dimensions of FMs. Ensuring the safety of models in real-time and dynamic contexts is critical, especially to avoid generating harmful or biased content. Meantime, striking a balance between the effectiveness and efficiency of FMs is challenging and requires more ongoing research efforts. Therefore, our pursuit of a comprehensive benchmark should extend to these vital dimensions of model performances.

Acknowledgement

This work is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-GC-2022-005). The computational work for this article was partially performed on resources of the National Supercomputing Centre (NSCC), Singapore. It is also partially supported by Cloud TPUs from Google's TPU Research Cloud (TRC). We thank Xunlong Zou and Geyu Lin for participating in research discussions, and Siti Umairah Md Salleh, Siti Maryam Binte Ahmad Subaidi, Nabilah Binte Md Johan, Wiwik Karlina, Xuan Long Do, Fabian Ritter Gutierrez and Ayrton San Joaquin for their contribution to cross-lingual resource construction and verification.

References

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.

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, et al. 2023. Benchmarking foundation models with language-model-as-an-examiner. *arXiv preprint arXiv:2306.04181*.

BIG bench authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

Fangzhou Chen and et al. 2023. Z-bench. https://github.com/zhenbench/z-bench.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. PaLM: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

David Deterding. 2007. *Singapore English*. Edinburgh University Press.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland. Association for Computational Linguistics.

Yu Fei, Yifan Hou, Zeming Chen, and Antoine Bosselut. 2023. Mitigating label biases for in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14014–14031, Toronto, Canada. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Zhouhong Gu, Xiaoxuan Zhu, Haoning Ye, Lin Zhang, Jianchen Wang, Sihang Jiang, Zhuozhi Xiong, Zihan Li, Qianyu He, Rui Xu, et al. 2023. Xiezhi: An ever-updating benchmark for holistic domain knowledge evaluation. *arXiv preprint arXiv:2306.05783*.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the EMNLP-IJCNLP 2019*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Nikita Haduong, Alice Gao, and Noah A. Smith. 2023. Risks and NLP design: A case study on procedural document QA. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1248–1269, Toronto, Canada. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kübler, and Lawrence Moss. 2020. OCNLI: Original Chinese Natural Language Inference. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3512–3526, Online. Association for Computational Linguistics.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023a. Not all languages are created equal in llms: Improving multilingual capability by cross-lingual-thought prompting. *arXiv preprint arXiv:2305.07004*.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2391–2401. Association for Computational Linguistics.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023b. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems*.

Li-Jun Ji, Zhiyong Zhang, and Richard E Nisbett. 2004. Is it culture or is it language? Examination of language effects in cross-cultural research on categorization. *Journal of personality and social psychology*, 87(1):57.

Fangkai Jiao, Yangyang Guo, Xuemeng Song, and Liqiang Nie. 2022. MERIt: Meta-Path Guided Contrastive Learning for Logical Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3496–3509, Dublin, Ireland. Association for Computational Linguistics.

Fangkai Jiao, Zhiyang Teng, Shafiq R. Joty, Bosheng Ding, Aixin Sun, Zhengyuan Liu, and Nancy F. Chen. 2023. Logicllm: Exploring self-supervised logic-enhanced training for large language models. *CoRR*, abs/2305.13718.

Claire Kramsch. 1991. Culture in language learning: A view from the united states. *Foreign language research in cross-cultural perspective*, 2:217–240.

Claire Kramsch. 2014. Language and culture. *AILA review*, 27(1):30–55.

Saurabh Kulshreshtha, Jose Luis Redondo Garcia, and Ching-Yun Chang. 2020. Cross-lingual alignment methods for multilingual BERT: A comparative study. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 933–942, Online. Association for Computational Linguistics.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard H. Hovy. 2017. RACE: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 785–794. Association for Computational Linguistics.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023a. CMMLU: Measuring massive multitask language understanding in chinese.

Peggy Li and Lila Gleitman. 2002. Turning the tables: Language and spatial reasoning. *Cognition*, 83(3):265–294.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023b. AlpacaEval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

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. 2022a. 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.

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

Hanmeng Liu, Jian Liu, Leyang Cui, Zhiyang Teng, Nan Duan, Ming Zhou, and Yue Zhang. 2023. Logiqa 2.0—an improved dataset for logical reasoning in natural language understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2947–2962.

Zhengyuan Liu, Shikang Ni, Aiti Aw, and Nancy Chen. 2022. Singlish message paraphrasing: A joint task of creole translation and text normalization. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3924–3936.

Robert K Logan. 1986. *The alphabet effect*. New York: Morrow.

John McCarthy. 2022. Artificial intelligence, logic, and formalising common sense. *Machine Learning and the City: Applications in Architecture and Urban Design*, pages 69–90.

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

Nils J. Nilsson. 1991. Logic and artificial intelligence. *Artif. Intell.*, 47:31–56.

R OpenAI. 2023. GPT-4 technical report. *arXiv*, pages 2303–08774.

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

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.

Alastair Pennycook. 2006. *Global Englishes and transcultural flows*. routledge.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Mei Silviana Saputri, Rahmad Mahendra, and Mirna Adriani. 2018. Emotion classification on indonesian twitter dataset. In *2018 International Conference on Asian Language Processing (IALP)*, pages 90–95. IEEE.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. DREAM: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2020. Investigating prior knowledge for challenging Chinese machine reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:141–155.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An instruction-following llama model.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Bin Wang and Haizhou Li. 2023. Relational sentence embedding for flexible semantic matching. In *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, pages 238–252, Toronto, Canada. Association for Computational Linguistics.

Bin Wang, Geyu Lin, Zhengyuan Liu, Chengwei Wei, and Nancy F Chen. 2024a. Craft: Extracting and tuning cultural instructions from the wild. *arXiv preprint arXiv:2405.03138*.

Bin Wang, Zhengyuan Liu, and Nancy Chen. 2023a. Instructive dialogue summarization with query aggregations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7630–7653, Singapore. Association for Computational Linguistics.

Bin Wang, Chengwei Wei, Zhengyuan Liu, Geyu Lin, and Nancy F Chen. 2024b. Resilience of large language models for noisy instructions. *arXiv preprint arXiv:2404.09754*.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023b. Voyager: An open-ended embodied agent with large language models. *CoRR*, abs/2305.16291.

Chengwei Wei, Yun-Cheng Wang, Bin Wang, and C.-C. Jay Kuo. 2024. An overview of language models: Recent developments and outlook. *APSIPA Transactions on Signal and Information Processing*, 13(2).

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, and Ayu Purwarianti. 2020. IndoNLU: Benchmark and resources for evaluating Indonesian

natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 843–857, Suzhou, China. Association for Computational Linguistics.

Lionel Wong, Gabriel Grand, Alexander K. Lew, Noah D. Goodman, Vikash K. Mansinghka, Jacob Andreas, and Joshua B. Tenenbaum. 2023. From word models to world models: Translating from natural language to the probabilistic language of thought. *CoRR*, abs/2306.12672.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Jiacheng Ye, Xijia Tao, and Lingpeng Kong. 2023. Language versatilists vs. specialists: An empirical revisiting on multilingual transfer ability. *arXiv preprint arXiv:2306.06688*.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *arXiv preprint arXiv:2306.05179*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023a. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.

Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023b. Extrapolating large language models to non-english by aligning languages. *arXiv preprint arXiv:2308.04948*.

A Related Work

A.1 Existing Benchmarks

The field of LLM evaluation is expanding quickly owing to the rapid development of model capabilities. Chang et al. (2023) provides a comprehensive review of different evaluation methods. Even though there are thousands of languages around the world, the vast majority of LLM evaluation benchmarks concentrate on English or Chinese (Liang et al., 2022; Li et al., 2023b; Zhong et al., 2023), which has a solid foundation of well-annotated resources. They are mainly focused on complex reasoning datasets which are normally collected from human examinations including SAT, math tests or Chinese college examinations (Zhong et al., 2023; Hendrycks et al., 2021b; Huang et al., 2023b; Li et al., 2023a). Gu et al. (2023) gathered questions from various disciplines like economics, jurisprudence and literature. Besides the subjective test, Li et al. (2023b) and Bai et al. (2023) propose to use LLMs as the judger to provide objective evaluations on generated content for objective scores and implement pairwise model ranking.

There are pioneering efforts on multilingual large language model evaluation. Lai et al. (2023) propose to evaluate large language models for their multilingual capability with a series of classic NLP tasks. Zhang et al. (2023) expands multilingual evaluation to 9 languages associated with a toolkit. Zhu et al. (2023a) evaluate LLM with machine translation test sets which are multilingual inherently. In this work, we expand multilingual foundation model evaluation benchmarks beyond combinations of monolingual tasks.

A.2 Foundation Language Models

The foundation language models, as general task solvers, include both pre-trained language models and their instruction-tuned variants. ChatGPT (Ouyang et al., 2022) and GPT4 (OpenAI, 2023) are showing superior capabilities across various applications. A series of foundation models

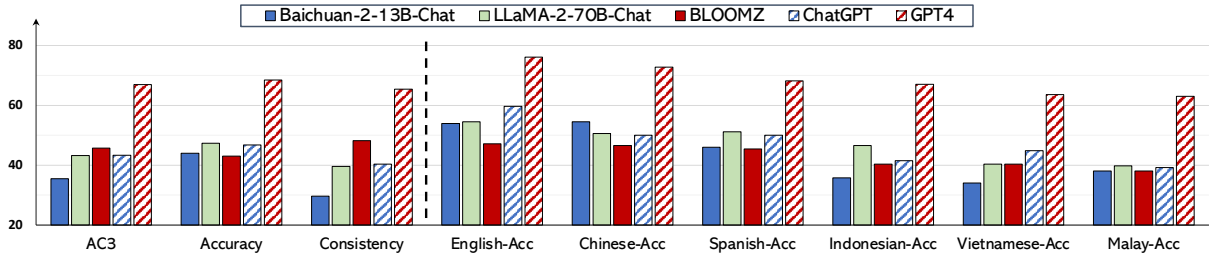

Figure 7: Evaluation results on Cross-LogiQA. Both overall and language-specific scores are shown.

Question	*Which of the following items would be considered the least suitable gift to bring to a Singaporean family during Lunar New Year?* *(A) An pineapple* *(B) Cash money* *(C) Two organges* *(D) A red packet*
Multicultural Reasoning Steps	**Multilingual Understanding** · Pineapple = Ong lai (Hokkien), sounds like 'good fortune to come' · Orange sounds like 'good luck' in Cantonese **Cultural Preferences** · Chinese like 'double' as a representation of unity, completeness and harmony. · Red packet is preferred as associated with luck, prosperity and happiness.
Answer	(B) Cash Money

Table 4: The 2nd example of cultural reasoning from SG-Eval.

are released afterwards including Flan-T5 (Chung et al., 2022), Alpaca (Taori et al., 2023), Vicuna (Chiang et al., 2023) and LLaMA-2 (Touvron et al., 2023b). Their multilingual capability is inferior to English due to the unbalanced training corpus and vocabulary settings. For applicable to multilingual scenarios, a series of bilingual or multilingual models are proposed by pertaining from scratch (XGLM (Lin et al., 2022b), BLOOM (Scao et al., 2022), ChatGLM (Du et al., 2022)), expansion of vocabulary sizes (Cui et al., 2023) or aligning multilingual instructions (Zhu et al., 2023b). In the foreseeable future, we anticipate a surge of multilingual language models, underscoring the need for effective multilingual LLM evaluation benchmarks.

B Selected Models

In this work, we evaluate the performance of various large language models on our benchmark datasets. They show disparate capabilities in various tasks. The included models are Flan-T5 (Flan-T5-Small, Flan-T5-Base, Flan-T5-Large, Flan-T5-XL, FLAN-T5-XXL, FLAN-UL2) (Chung et al., 2022), LLaMA-1 (LLaMA-7B, LLaMA-13B, LLaMA-30B, LLaMA-65B) (Touvron et al., 2023a), LLaMA-2 (LLaMA-2-7B, LLaMA-2-7B-Chat, LLaMA-2-13B, LLaMA-2-13B-Chat, LLaMA-2-70B, LLaMA-2-70B-Chat) (Touvron et al., 2023b), Baichuan (Baichuan-7B, Baichuan-13B, Baichuan-13B-Chat, Baichuan-2-7B, Baichuan-2-7B-Chat, Baichuan-2-13B, Baichuan-2-13B-Chat) (Yang et al., 2023), Alpaca-7B (Taori et al., 2023), Vicuna (Vicuna-7B-v1.3, Vicuna-13B-v1.3, Vicuna-7B-v1.5, Vicuna-13B-v1.5, Vicuna-33B-v1.3) (Chiang et al., 2023), ChatGLM (ChatGLM-6B, ChatGLM2-6B) (Du et al., 2022), BLOOM (BLOOMZ-7B1, MT0-XXL) (Scao et al., 2022), Colossal-LLaMA-2-7B-Base, ChatGPT (gpt-3.5-turbo-0613, gpt-4-0613) (OpenAI, 2023). In this paper, we report the result of the following models as a representative set considering their overall performance and multilingual support.

- **Baichuan-2**: is an open-source multilingual language model with emphasis on English and Chinese. It shows competitive performance compared to models of the same size and generally outperforms LLaMA-2 model through better pre-training and human alignment techniques. Baichuan-2-13B-Chat is selected in our experiments.
- **LLaMA-2**: is an open-source language model released by Meta. Even though it supports multilingual, LLaMA is trained with most data (close to 90%) in English which makes it an English-centric model. It performs the best for English use cases than other languages. The Chat variant is further tuned for improved helpfulness and safety. LLaMA-2-13B-Chat and LLaMA-2-70B-Chat are selected in our experiments.
- **BLOOMZ**: is the leading open-source multilingual large language model further tuned

with diverse instructions. It supports over 40 languages with a more balanced pertaining and fine-tuning corpus. Note that BLOOMZ is instruction-tuned with supervised datasets which may cause supervision leakage on certain datasets (e.g. SAMSum, DREAM) and unjustified comparison. BLOOMZ-7B1 is selected in our experiments.

- **ChatGPT**: is closed-source model developed by OpenAI. It has good multilingual support and demonstrates more robust performance compared to open-source models. The model is updating over time and we select GPT3.5 (referred to as ChatGPT) and GPT4 on version 0613 in all our experiments.

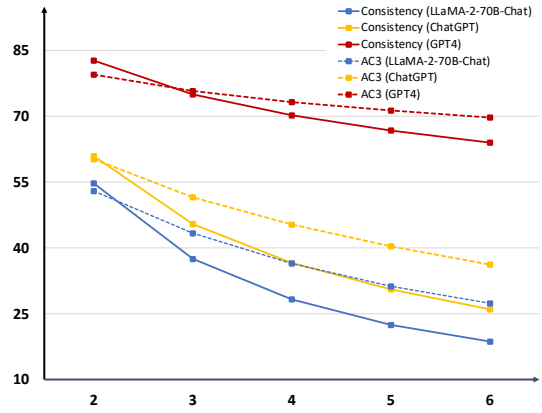

Figure 8: Consistency and AC3 Scores with $s \in [2, 6]$ on Cross-MMLU dataset. (Filipino excluded)

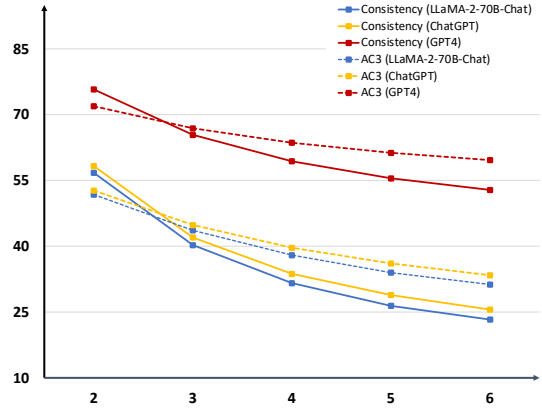

Figure 9: Consistency and AC3 Scores with $s \in [2, 6]$ on Cross-LogiQA dataset. (Filipino excluded)

C Cross-Lingual Consistency

In this paper, we spot the cross-lingual inconsistency problem for multilingual foundation models. To better evaluate this aspect, we collect and propose two new datasets with respective metrics. In this section, we provide more analysis on the cross-lingual consistency study and the effectiveness of s in the proposed *Consitency* and *AC3* metrics.

The evaluation results for Cross-LogiQA dataset is depicted in Figure 7. As the leading open-source multilingual model, BLOOMZ outperforms other open-source models and ChatGPT in terms of consistency but falls short in achieving high performance in specific languages. This suggests that BLOOMZ provides more consistent answers across languages, possibly due to its training on a more balanced multilingual corpus and fine-tuning with multilingual instruction data to improve cross-language alignment. As observed in Figure 6, the performance is better for higher-resource languages such as English, Chinese, and Spanish compared to lower-resource languages like Indonesian, Vietnamese, and Malay. GPT4 surpasses other models in both "Accuracy" and "Consistency," highlighting significant potential for further enhancement of all other models.

In Section 3.3, AC3 score is presented, taking into account both accuracy and consistency scores. We have one tolerance hyperparameter s which requires the answers to be consistent across s languages to be rewarded in consistency score. We deploy $s = 3$ as the default hyperparameter in above experiments. Here, we conduct more systematic study of s with its effect on the final scores.

The results on Cross-MMLU and Cross-LogiQA are shown in Figure 8 and 9. As s increases, the consistency score has dropped dramatically for all models. Among all three models, the performance of GPT4 drops the least, indicating a robust consistency alignment across languages. Even for ChatGPT, when $s = 6$, the consistency score downgrades to around 26% on both Cross-MMLU and Cross-LogiQA datasets. It indicates that only 26% of the cases that ChatGPT is selecting the same answer for the same question across six languages. Hence, we opt for $s = 3$ as the default value for two primary reasons: 1) to facilitate evaluation across multilingual language models, even when not all 6 languages are supported, and 2) to allow for a certain level of tolerance regarding language consistency, without imposing the strict requirement of complete consistency among all responses. This approach can be seen as a more lenient method of assessing language consistency. In the future, as

the model's multilingual capabilities continue to advance, we can increase s accordingly.

D Evaluation Protocols

When assessing foundation models, two distinct settings come into play: zero-shot and few-shot. In this study, we predominantly rely on zero-shot evaluation as the chosen method for all models, primarily for two compelling reasons. Firstly, zero-shot evaluation aligns more closely with real-world application scenarios, where users interact directly with deployed models without undergoing explicit training. Secondly, it's worth noting that even without the process of fine-tuning using human instructions, these models display a noteworthy ability to comprehend and adhere to emerging instructions. Besides, it brings additional benefits to avoid uncertainty caused by the in-context few-shot samples and potential exposure biases.

Another challenge in evaluation is the unstructured form of outputs from large language models. Unlike previous discriminative models, language generation models produce the answer represented by free text. Therefore, there is a gap between the generated content and the ground-truth answers, especially for multi-choice questions. Therefore, we develop a heuristic algorithm to decide the mapping. In general, we first split the answer into sentences. For each sentence, we detect whether the choice symbols (e.g. (A), (B), (C), (D)) exist. If none of them exists, we detect whether the choice description exists. If exists, we count the sentence as the symbol of such label and N.A. otherwise. Finally, we perform majority voting from all sentences as the final answer. From the experiments, we found the algorithm is robust enough to link the generated content with labels among diverse models and languages. Therefore, we chose this algorithm after careful comparison with a few other variants.

E Full Experimental Results and Analysis

Besides the evaluation on 5 datasets shown in Figure 5, the full evaluation results on the other 23 datasets are shown in Figure 10. From the results, we spot several key findings.

First, Baichuan-2-13B-Chat surpasses LLaMA-2-13B-Chat not only in Chinese tasks but also in tasks in English and other languages. Despite being pre-trained with 2T tokens mostly in English, LLaMA-2 falls short in terms of its English proficiency. This highlights the critical role of data quality and diversity. Effective data collection and post-processing play a pivotal role in the development of large language models.

Secondly, when it comes to a multilingual context, BLOOMZ exhibits lower competitiveness in comparison to other models. Despite being trained on a dataset that incorporates more than 40 languages and pre-trained using a more evenly balanced corpus from these languages, it fails to showcase superior performance in various multilingual tasks. This could be attributed to ineffective pre-training and the limitations imposed by the model's size, which consists of 7 billion parameters. It is worth noting that BLOOMZ does display exceptional performance in specific datasets, such as SAMSum and Flores. This can be attributed to the direct fine-tuning of the model with supervision, making it inappropriate to draw direct comparisons with datasets as outlined in (Muennighoff et al., 2023). Nevertheless, the cross-lingual consistency of the BLOOMZ model is good due to its cross-lingual generalization through multitask finetuning (Muennighoff et al., 2023).

Lastly, GPT4 surpass ChatGPT in various aspects including multilingual capability. However, due to its commercial nature, it is generally hard to conduct transparent research with such models.

F More Examples

To have a direct interpretation of the newly proposed six datasets, we further illustrate samples and instructions with English translations in Table 5. Examples of SG culture questions are shown in Table 2 and 4.

For evaluating cross-lingual consistency, we introduce two datasets: Cross-MMLU and Cross-LogiQA, featuring parallel questions in 6 different languages. In this section, we delve deeper into cross-lingual inconsistency phenomena across a broad range of languages. To achieve this, we expand our sample to include questions in 16 languages and prompt ChatGPT for answers. The outcomes are detailed in Table 1, 6 and Figure 11, 12, 13, 14, 15. The languages encompassed in this study are English, Chinese, Indonesian, Spanish, Thai, French, Korean, Malay, Turkish, German, Romanian, Filipino, Tamil, Portuguese, Vietnamese and Arabic. The results substantiate the existence of cross-lingual consistency issues across different languages, underscoring the need

for increased attention to this matter.

G Annotators

The annotators consist of both full-time employees and PhD students. Full-time employees did not receive additional compensation for their annotation work but considered it as part of their regular working hours. Conversely, PhD students had their annotation time recorded and were compensated with fixed-hour claim rates. On average, one round of correction for each language on each dataset took approximately 3-5 hours, varying depending on the languages involved.

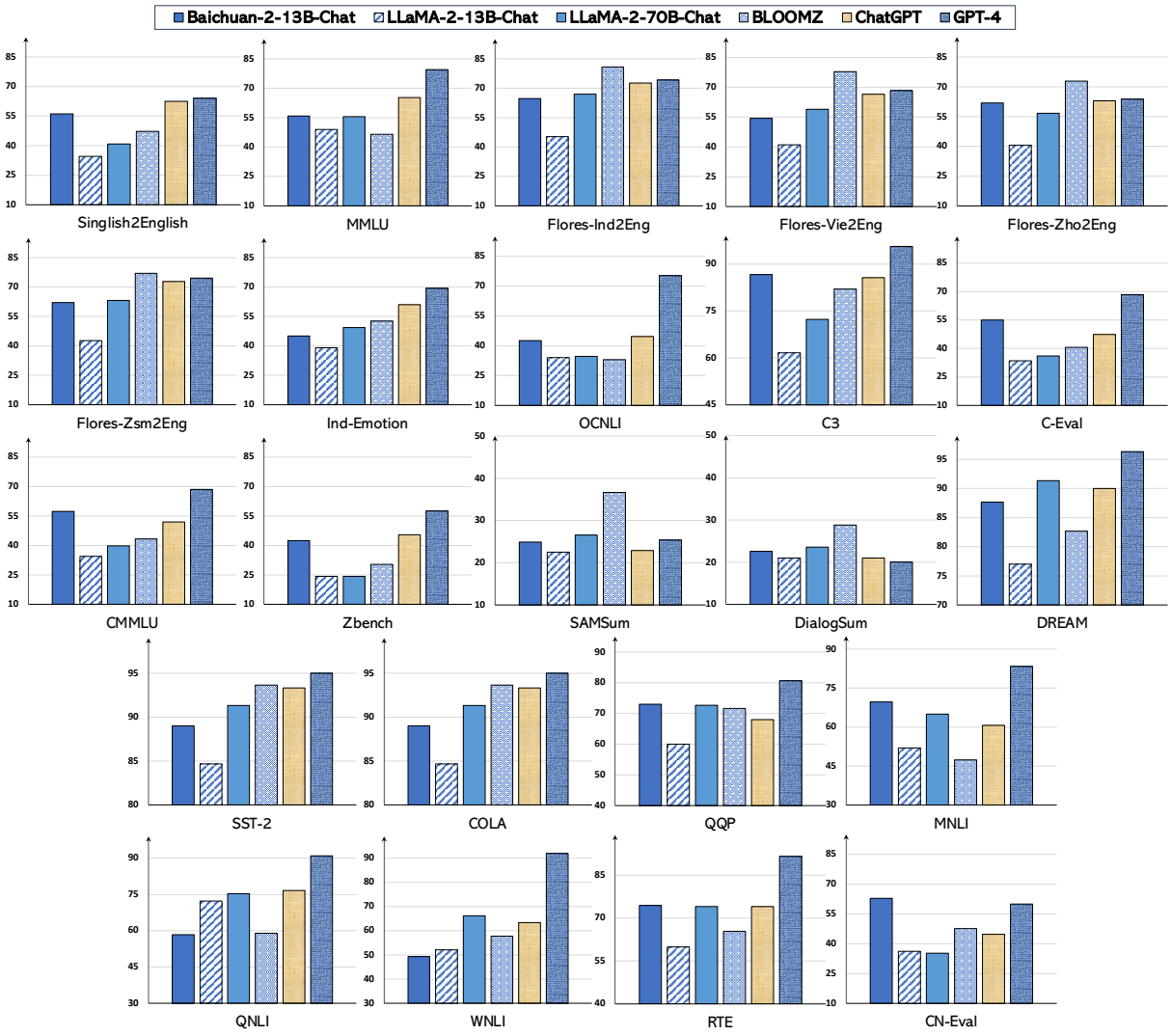

Figure 10: Evaluation results of LLMs on the rest of SeaEval tasks as supplement of Figure 5.

Language	**Thai**	**French**
Question	โปรดเลือกคำตอบที่ถูกต้องสำหรับคำถามต่อไปนี้ เมื่อแสงสีขาวส่องผ่านทะลุปริซึม แสงที่หักเหมา -กกว่าสีเขียวคืองอมากกว่าสีเขียวจะเป็น (A) สีแดง (B) สีเหลือง (C) สีน้ำเงินน้ำ (D) ไม่มีข้อใด	*Veuillez choisir la bonne réponse à la question suivante.* *Lorsque la lumière blanche traverse un prisme, la lumière qui se courbe plus que la verte est* *(A) Rouge* *(B) Jaune* *(C) Bleue* *(D) Aucune d'eux*
Answer	(D) ไม่มีข้อใด ***In English:*** *(D) None of these*	*(C) Bleue* ***In English:*** *(C) Blue*
Correctness	✗	✓

Figure 11: An example from our Cross-MMLU dataset on Thai and French

Multicultural and Multilingual Understanding		
SG-Eval	Instruction	*Please carefully read the following question and select the most appropriate answer from the choices.*
	Sample	*Which drink in Singapore has the highest calories?* *(A) Teh O* *(B) Teh Siew Dai* *(C) Kopi* *(D) Kopi C* *Answer: (C) Kopi*
US-Eval	Instruction	*Read the following question carefully and select the correct answer from the choices.*
	Sample	*When daylight-saving time arrives in the spring how do most Americans turn their clocks?* *(A) one hour forward* *(B) one hour backward* *(C) two hours forward* *(D) two hours backward* *Answer: (A) one hour forward*
CN-Eval	Instruction	请仔细阅读以下问题，并从选项中选择最合适的答案。
	Sample	清代官场饮茶有着特殊的程序和含义，有别于一般的茶道，主人若端茶，对客人说“请喝茶”，这表明 (A)对客人不满 (B)请客人品茶 (C)对客人的尊敬 (D)会谈结束送客 答案: (D)会谈结束送客 **Translation:** *Tea drinking in officialdom in the Qing Dynasty had special procedures and meanings, which were different from ordinary tea ceremonies. If the host served tea and said "Please drink tea" to the guests, this meant* *(A) Dissatisfied with the guest* *(B) Invite guest to taste tea* *(C) Show Respect for guest* *(D) End the meeting and seeing off the guest* *Answer: (D) End the meeting and see off the guest*
Singlish2English	Instruction	*Translate the following sentence from Singlish to English. Please only output the translated sentence.*
	Sample	**Source in Singlish:** *Wah this one damn shiok and underrated. The maggi goreng also damn sedap. Bro you got refined taste* **Target in Standard English:** *Wow, this is super enjoyable and underrated. The Maggi Goreng is damn delicious. Brother, you have got a refined taste.*
Cross-Lingual Consistency		
Cross-MMLU	Instruction	*Respond to the question by selecting the most appropriate answer.*
	Sample	Shown in Table 1.
Cross-LogiQA	Instruction	*Kindly choose the correct answer from the options provided for the multiple-choice question.*
	Sample	**English Version:** *Content: At a gathering at which bankers, athletes, and lawyers are present, all of the bankers are athletes and none of the lawyers are bankers.* *Question: If the statements above are true, which one of the following statements must also be true?* *(A) Some of the lawyers are not athletes.* *(B) Some of the athletes are not lawyers.* *(C) None of the lawyers are athletes.* *(D) All of the athletes are bankers.* *Answer: (B) Some of the athletes are not lawyers.* **Chinese Version:** 在银行家，运动员和律师的聚会上，所有银行家都是运动员，没有一个律师是银行家。 如果陈述以上为真，下列哪一项也一定为真? (A)有些律师不是运动员。 (B)有些运动员不是律师。 (C)没有律师是运动员。 (D)所有运动员都是银行家。 答案: (B)有些运动员不是律师。 **Indonesian Version:** *Isi: Pada pertemuan yang dihadiri oleh para bankir, atlet, dan pengacara, semua bankir adalah atlet dan tidak ada satupun pengacara yang merupakan bankir.* *Pertanyaan: Jika pernyataan di atas benar, manakah pernyataan berikut yang juga benar?* *(A) Beberapa pengacara bukanlah atlet.* *(B) Beberapa atlet bukan pengacara.* *(C) Tidak ada pengacara yang merupakan atlet.* *(D) Semua atletnya adalah bankir.* *Jawaban: (B) Beberapa atlet bukan pengacara.* **Same sample in Spain, Vietnamese and Malay languages ...**

Table 5: An example of the instructions and samples from our newly proposed datasets. One instruction is sampled for each dataset.

Language	Korean	Malay	Turkish	German
Question	다음 문제에서 올바른 답을 선택해 주세요. 흰 빛이 프리즘을 통과할 때, 초록보다 더 많이 굴절되는 빛은 무엇입니까? (A) 빨강 (B) 노랑 (C) 파랑 (D) 이 중 어느 것도 아님	*Sila pilih jawapan yang betul untuk soalan berikut.* *Apabila cahaya putih melalui prisma, cahaya yang membengkok lebih daripada hijau ialah* *(A) Merah* *(B) Kuning* *(C) Biru* *(D) Tiada satu pun daripada ini*	*Lütfen aşağıdaki soru için doğru cevabı seçin.* *Beyaz ışık bir prizmadan geçtiğinde yeşilden daha fazla kırılan ışık* *(A) kırmızıdır* *(B) sarıdır* *(C) mavidir* *(D) Bunlardan hiçbiri*	*Bitte wählen Sie die richtige Antwort für die folgende Frage. Welches Licht wird stärker gebeugt als grünes Licht wenn weißes Licht durch ein Prisma fällt?* *(A) Rot* *(B) Gelb* *(C) Blau* *(D) Keine von diesen*
Answer	(C) 파랑 ***In English**: (C) Blue*	*(A) Merah* ***In English**: (A) Red*	*(C) mavidir* ***In English**: (C) Blue*	*(A) Rot* ***In English**: (A) Red*
Correctness	✓	✗	✓	✗

Table 6: An example from our Cross-MMLU dataset on Korean, Malay, Turkish and German

Language	Romanian	Filipino
Question	*Vă rugăm să selectați răspunsul corect la următoarea întrebare:* *când lumina albă trece printr-o prismă, care dintre culorile următoare se îndoaie mai mult decât culoarea verde?* *(A) Roșu* *(B) Galben* *(C) Albastru* *(D) Niciuna dintre acestea*	*Mangyaring piliin ang tamang sagot para sa sumusunod na tanong.* *Kapag ang puting liwanag ay dumaan sa isang prisma, ang liwanag na mas baluktot kaysa berde ay* *(A) Pula* *(B) Dilaw* *(C) Asul* *(D) Wala rito*
Answer	*(C) Albastru* ***In English**: (C) Blue*	*(A) Pula* ***In English**: (A) Red*
Correctness	✓	✗

Figure 12: An example from our Cross-MMLU dataset on Romanian and Filipino

Language	Tamil	Portuguese
Question	பின்வரும் கேள்விக்கான சரியான பதிலைத் தேர்ந்தெடுக்கவும். வெள்ளை ஒளி ஒரு ப்ரிஸம் வழியாக செல்லும் போது, பச்சை நிறத்தை விட அதிகமாக வளைக்கும் ஒளி *(A)* சிவப்பு *(B)* மஞ்சள் *(C)* நீலம் *(D)* இவற்றில் ஏதுமில்லை	*Por favor, escolha a resposta correta para a seguinte pergunta: Quando a luz branca passa por um prisma, qual das seguintes cores se dobra mais do que a cor verde?* *(A) Vermelho* *(B) Amarelo* *(C) Azul* *(D) Nenhuma destas*
Answer	*(B)* மஞ்சள் ***In English**: (B) Yellow*	*(C) Azul* ***In English**: (C) Blue*
Correctness	✗	✓

Figure 13: An example from our Cross-MMLU dataset on Tamil and Portuguese

Language	Vietnamese
Question	Hãy chọn đáp án đúng cho câu hỏi sau. Khi ánh sáng trắng đi qua lăng kính thì ánh sáng lệch nhiều hơn ánh sáng xanh là (A) ánh sáng đỏ (B) ánh sáng vàng (C) ánh sáng xanh da trời (D) Không có cái nào trong số này
Answer	(C) ánh sáng xanh da trời ***In English**: (C) Blue*
Correctness	✓

Figure 14: An example from our Cross-MMLU dataset on Vietnamese

Language	Arabic
Question	اختر الإجابة الصحيحة للسؤال التالي. عندما يمر الضوء الأبيض عبر المنشور، الضوء الذي ينحني أكثر من الأخضر يكون *(A)* أحمر *(B)* أصفر *(C)* أزرق *(D)* لا شيء مما سبق
Answer	*(C)* أزرق ***In English**: (C) Blue*
Correctness	✓

Figure 15: An example from our Cross-MMLU dataset on Arabic