FilBench: Can LLMs Understand and Generate Filipino?

Lester James V. Miranda[1*] **Elyanah Aco**[2*] **Conner Manuel**[3*]
Jan Christian Blaise Cruz[4,5†] **Joseph Marvin Imperial**[4,6,7†]

[1]Allen Institute for AI [2]Nara Institute of Science and Technology [3]Together AI
[4]SEACrowd [5]MBZUAI [6]University of Bath [7]National University Philippines

Code filbench/filbench-eval **Leaderboard** UD-Filipino/filbench-leaderboard

Abstract

Despite the impressive performance of LLMs on English-based tasks, little is known about their capabilities in specific languages such as Filipino. In this work, we address this gap by introducing **FilBench**, a Filipino-centric benchmark designed to evaluate LLMs across a diverse set of tasks and capabilities in Filipino, Tagalog, and Cebuano. We carefully curate the tasks in FilBench to reflect the priorities and trends of NLP research in the Philippines such as Cultural Knowledge, Classical NLP, Reading Comprehension, and Generation. By evaluating 27 state-of-the-art LLMs on FilBench, we find that several LLMs suffer from reading comprehension and translation capabilities. Our results indicate that FilBench is challenging, with the best model, GPT-4o, achieving only a score of 72.23%. Moreover, we also find that models trained specifically for Southeast Asian languages tend to underperform on FilBench, with the highest-performing model, SEA-LION v3 70B, achieving only a score of 61.07%. Our work demonstrates the value of curating language-specific LLM benchmarks to aid in driving progress on Filipino NLP and increasing the inclusion of Philippine languages in LLM development.

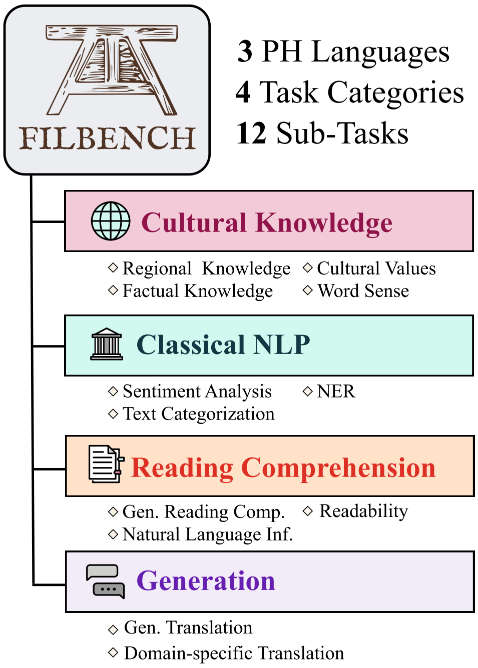


Figure 1: **Overview of FilBench.** In order to comprehensively assess the full capabilities of LLMs on Philippine languages, we curate an evaluation suite consisting of 4 categories and 12 subtasks across **Filipino, Tagalog, and Cebuano** based on the research priorities of the Philippine NLP community (§3.1).

1 Introduction

While large language models (LLMs) have shown impressive performance on a variety of English-based tasks and capabilities, their effectiveness remains largely unexplored for low-resource languages such as Filipino. This knowledge gap exists for two reasons. First, most low-resource languages, especially Filipino-centric benchmarks, developed before the ChatGPT-era (—2022 Gururaja et al., 2023) are ill-posed for current LLM evaluation despite their value in understanding language system capabilities. Second, existing multilingual LLM benchmarks either exclude Filipino entirely (Liu et al., 2025; Huang et al., 2025, *inter alia*) or fail to provide sufficient task and/or language diversity (Susanto et al., 2025). Filipino is an important language to consider for LLM evaluation not only because of its unique linguistic properties such as its voice marking system (Bardají et al., 2024), but also due to its large speaker population with more than 28 million speakers in the Philippines and over 2 million speakers abroad (Philippine Statistics Authority, 2020).

In this work, we perform a comprehensive study of the strengths and limitations of LLM capabil-

[*] Equal contributions. Corresponding e-mail address: filbench-eval@googlegroups.com

[†] Senior authors.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 2496–2529
November 4-9, 2025 ©2025 Association for Computational Linguistics

ities on Filipino-centric tasks. We introduce an evaluation suite called **FilBench**, consisting of 4 categories and 12 diverse sub-tasks that are formulated for LLM evaluation. The choice of tasks to include in FilBench is based on our study of research trends and priorities in Filipino NLP (§3.1, §J). Evaluating models on FilBench reveals significant gaps in LLM performance, for instance, in text generation capabilities. The contributions of this study are three-fold:

- We close the **resource gap** by curating Filipino test sets across four broad task categories: Cultural Knowledge Assessment (CK), Classical NLP (CN), Reading Comprehension (RC), and Generation (GN). We transform these datasets into a unified task format aligned with standard LLM evaluation practices across literature. Our evaluation suite, **FilBench**, consists of test instances across 4 categories and 12 sub-tasks (§3).

- We bridge the **evaluation gap** by evaluating 27 state-of-the-art LLMs on FilBench (§4). We find that the best model, GPT-4o, only achieve around 72.23% aggregated performance while the best Southeast-Asian model, SEA-LION v3 70B, only obtains a score of 61.07%.

- We provide **analyses and insights** to the strengths and weaknesses of LLMs when presented with Filipino-centric tasks and test cases (§5). Notably, we find that text generation suffers the most, with the lowest scores across models due to failure modes such as hallucination and poor instruction-following.

FilBench demonstrates the value of constructing language-specific benchmarks to reveal gaps in language model capabilities and benefit the wider speaker community. More importantly, we hope that this work aids in improving the state of Filipino NLP and increase the inclusion of Philippine languages in LLM development.

2 Background

Languages in the Philippines. The Philippines is home to approximately 117 million language speakers across more than 185 distinct languages (Eberhard et al., 2024; McFarland, 2008; Metila et al., 2016). One of its official languages is Filipino (`FIL`), which is a standardized form of Tagalog (`TGL`) and used mainly in Metro Manila.[1] Aside from Filipino and Tagalog, Cebuano (`CEB`) is the second most widely spoken language in the Philippines with over 28 million speakers. It is part of the Visayan language family and is spoken mainly in regions of Cebu, Siquijor, and Bohol among many others (Pilar et al., 2023). As part of the same subgroup of Philippine languages, Tagalog and Cebuano share similar linguistic characteristics such as shared vocabulary and comparable word formulation processes and affixation rules, among others (Bacalla, 2019; Imperial and Kochmar, 2023). In our work, we focus on the these three languages because they cover the majority of Filipino speakers, representing approximately 61% of the country's population (Philippine Statistics Authority, 2020).

Task Formulation in LLM Evaluation. In order to standardize how each test example is presented to an LLM, it must first be formatted into a consistent prompt structure or *formulation*. Multiple-choice formulation (MCF) is a common standard in evaluating LLMs across a vast array of tasks (Gu et al., 2024; Fourrier et al., 2024). In MCF, a question is posed with answers presented as labeled choices, where scoring is done by comparing the LLM's choice to the gold label. For evaluating LLMs on generative tasks such as translation, one approach is to write an instruction prompting an LLM to translate a given text from a source language to a target language. Then, the generated output by the LLM is compared against the reference translation using various machine translation metrics (Papineni et al., 2002; Lin, 2004).

3 The FilBench Evaluation Suite

Our design philosophy for FilBench centers on two core principles: (1) developing an impactful benchmark that aligns with the research priorities within the Philippine context (§3.1), ensuring that a model excelling in FilBench is likely to perform effectively across a wide range of Filipino applications and (2) maintaining data quality and richness by incorporating diverse sub-tasks (§3.2) that were annotated by experts or native speakers. Table 1 shows all the datasets and tasks included in FilBench. Example task formulation for each sub-task is shown in Appendix M.

[1] The designation of Filipino and Tagalog as separate languages is often a point of contention, although they are linguistically similar (Villafania, 2007). We follow the official view of the *Komisyon ng Wikang Filipino* (Commision on the Filipino Language) and treat them as separate.

Category	Sub-Task	Dataset	Languages	# Instances
Classical NLP (CN)	Text Classification	Dengue Filipino (Livelo and Cheng, 2018)	FIL	4,015
		BalitaNLP (Buñag and Esquivel, 2023)	TGL	70,352
		SIB-200 (Adelani et al., 2024)	CEB, FIL	99
	Named-Entity Recognition	CebuaNER (Pilar et al., 2023)	CEB	1,310
		TLUnified-NER (Miranda, 2023)	TGL	1,579
		Universal NER (Mayhew et al., 2024)	CEB, TGL	105
	Sentiment Analysis	FiReCS (Cosme and De Leon, 2023)	FIL	7,340
Cultural Knowledge Assessment (CK)	Regional Knowledge	INCLUDE (Romanou et al., 2024)	TGL	510
	Factual Knowledge	Global MMLU (Singh et al., 2024)	TGL	14,042
	Cultural Values	KALAHI (Montalan et al., 2024)	TGL	150
	Word-sense Disambiguation	StingrayBench (Cahyawijaya et al., 2024)	TGL	100
Reading Comprehension (RC)	Readability	Cebuano Readability Corpus (Imperial et al., 2022)	CEB	350
	Reading Comprehension	Belebele (Bandarkar et al., 2024)	CEB, FIL	1,800
	NLI	NewsPH NLI (Cruz et al., 2021)	FIL	90,000
Generation (GN)	Document translation	NTREX-128 (Federmann et al., 2022)	FIL	1,997
	Realistic translation	Tatoeba (Tiedemann, 2020)	CEB, TGL	2,876
	Domain-specific transl.	TICO-19 (Anastasopoulos et al., 2020)	TGL	971

Table 1: **Fine-grained overview of FilBench.** Our curation effort involves expert-annotated or validated datasets across a diverse range of sub-tasks and categories basd on a quantitative analysis of the priorities of the Filipino NLP community (§J), allowing us to comprehensively evaluate LLM capabilities on Filipino-centric tasks.

3.1 Research Priorities in Filipino NLP

In order to determine which tasks to include in FilBench, we perform a survey of the research trends in NLP research on Philippine languages from 2006–2023. Our methodology involves scraping Scopus-indexed papers and *ACL/EMNLP publications and classifying their NLP sub-field based on common ACL tracks. We find that classical NLP tasks such as information extraction and sentiment analysis are widely studied, as well as a variety of translation tasks. Then, we devise a taxonomy consisting of four major categories that encompass more recent trends in Philippine NLP research. More details about our methodology and findings can be found in Appendix J.

3.2 FilBench Categories

Cultural Knowledge Assessment (CK). This category tests a language model's ability to recall factual and culturally-specific information. Studies have consistently found that LLMs predominantly trained on English text are strongly biased towards Western values and perspectives, especially when prompted in English (Cao et al., 2023). Cultural misalignment between LLMs and users can lead to unintended harms such as norm violations (Qiu et al., 2025) and socio-economic exclusion (Dammu et al., 2024). For CK, we curate a variety of examples that test an LLM's regional and factual knowledge (Romanou et al., 2024; Singh et al., 2024), understanding of Filipino-centric values (Montalan et al., 2024), and word-sense disambiguation (Cahyawijaya et al., 2024).

Classical NLP (CN). This category encompasses a variety of information extraction and linguistic tasks such as named entity recognition (NER), sentiment analysis, and text categorization that were traditionally performed using specialized trained models. These tasks have been prominent in Philippine NLP research over the past decade (Roxas et al., 2021), and LLMs have recently begun to be employed in this domain (Ashok and Lipton, 2023; Zhang et al., 2023b; Wang et al., 2023, *inter alia*). For CN, we include expert-annotated NER datasets such as CebuaNER (Pilar et al., 2023), TLUnified-NER (Miranda, 2023), and Universal NER (Mayhew et al., 2024). We also take the Filipino and Cebuano subsets of SIB-200 (Adelani et al., 2024), and the text-only subset of Balita NLP (Buñag and Esquivel, 2023).

Reading Comprehension (RC). This category evaluates a language model's ability to understand and interpret Filipino text, focusing on tasks such as readability, comprehension, and natural language inference (NLI). These tasks are crucial for assessing how well a model can process and generate human-like understanding of written content. For RC, we include datasets like the Cebuano Readability Corpus (Imperial et al., 2022), Belebele (Bandarkar et al., 2024), and NewsPH NLI (Cruz et al., 2021), which provide a comprehensive evaluation of reading comprehension capabilities in the Filipino context.

Generation (GN). Although generative LLM tasks usually include summarization and conversa-

Model	FilBench Score	Cultural Knowledge	Classical NLP	Reading Comp.	Generation
○ gpt-4o-2024-08-06	**72.73±1.66**	73.29±3.01	89.03±2.05	**80.12±0.90**	**46.48±0.60**
○ meta-llama/Llama-4-Maverick-17B-128E-Instruct-FP8	67.67±1.04	**76.75±3.04**	87.28±0.26	72.99±0.18	33.67±0.71
○ meta-llama/Llama-4-Scout-17B-16E-Instruct	63.20±1.05	74.31±3.14	87.88±0.25	70.86±0.18	19.75±0.63
○ Qwen/Qwen2.5-72B-Instruct	63.08±0.99	73.11±3.22	88.60±0.24	75.62±0.17	14.98±0.33
♦ aisingapore/Llama-SEA-LION-v3-70B-IT	61.07±0.95	76.78±3.02	89.99±0.23	53.56±0.19	23.95±0.34
○ Tower-Babel/Babel-83B-Chat	60.85±0.96	75.21±3.11	88.81±0.25	64.85±0.19	14.53±0.29
○ meta-llama/Llama-3.1-70B-Instruct	59.66±1.17	72.16±3.21	**90.27±0.83**	52.17±0.28	24.03±0.37
♦ sail/Sailor2-20B-Chat	58.61±1.06	66.43±3.41	89.03±0.25	63.03±0.19	15.95±0.38
○ Qwen/Qwen2.5-32B-Instruct	57.88±1.45	66.83±3.45	89.32±1.99	70.59±0.18	4.79±0.17
♦ aisingapore/Gemma-SEA-LION-v3-9B-IT	56.14±1.53	64.44±3.43	88.55±0.25	54.46±0.20	17.10±2.25

Table 2: **Performance of state-of-the-art LLMs on Filipino-centric tasks.** We evaluate several models with different multilingual capabilities (multilingual ○, SEA-specific ♦), sizes (1.5B to 400B), and accessibility (open-source vs. commercial). Full results can be found in Table 8.

tional generation, evaluation test sets in Filipino are sparse. However, machine translation is one of the most dominant areas of NLP research in the Philippines (Oco and Roxas, 2018; Baliber et al., 2020; Aji et al., 2023, *inter alia*). Recently, LLMs have gained traction for its use as automatic translators, as opposed to training specialized translation models (Zhu et al., 2023; He et al., 2024; Alves et al., 2024). Hence, we dedicate a large portion of FilBench for testing an LLM's ability to faithfully translate texts, either from English to Filipino (ENG → FIL) or from Cebuano to English (CEB → ENG). We include a diverse set of test examples, ranging from documents (Federmann et al., 2022), realistic texts collected from volunteers (Tiedemann, 2020), and domain-specific text (Anastasopoulos et al., 2020).

3.3 FilBench Scoring

The CN, CK, and RC categories follow the MCF task formulation, so we score an LLM's performance for these categories by computing the accuracy, i.e., the number of correct answers divided by the total number of examples. For GN, we compute the ROUGE-L score between the LLM-generated text and the gold reference text. All per-category metrics range from 0 to 1. In order to create a representative, single evaluation score, we perform a weighted average based on the number of examples across results as shown in Equation 1:

$$\textsc{FilBench Score} = 100 \times \frac{\sum_{i \in \{\text{CN,CK,GN,RC}\}} n_i \cdot S_i}{\sum_{i \in \{\text{CN,CK,GN,RC}\}} n_i} \quad (1)$$

where n_i is the number of examples in category i and S_i is the score for category i.

4 Results: Performance of State-of-the-Art LLMs on FilBench

In order to understand what kind of LLMs perform well in Filipino, we select a variety of open-source and commercial LLMs to ensure broad coverage across parameter sizes and language capabilities. We also include a number of SEA-specific models that were trained to cater to Southeast Asian languages, including Filipino. A total of 27 models are chosen for evaluation. Table 6 in the Appendix shows the full details of the evaluated models.

Table 2 shows the scores obtained by the top ten models on FilBench. The full results for all 27 models can be seen in Table 8 of the Appendix. The best performing model is GPT-4o (72.23%), closely followed by Llama 4 Maverick (67.67%). Moreover, the highest scoring open-source dense model is Qwen2.5 72B (63.08%), while SEA-LION v3 70B is the best SEA-specific model (61.07%).

Finding #1: Larger models dominate FilBench. Figure 2 shows the FilBench score to Parameter Size (B) for several dense open-source language models with known sizes. Our findings suggest that parameter size strongly correlates with FilBench performance, with a Spearman ρ of 0.810. However, this correlation is not perfect as we observe some smaller models to be competitive with larger counterparts as observed in Qwen 2.5 32B having similar performance to Llama 3.1 70B.

Finding #2: Language-specific finetuning improves FilBench performance. SEA-specific models tend to be more parameter-efficient as they perform better than non-specialized LLMs on FilBench. This trend is more apparent for smaller models within the 7B to 9B range, as shown in

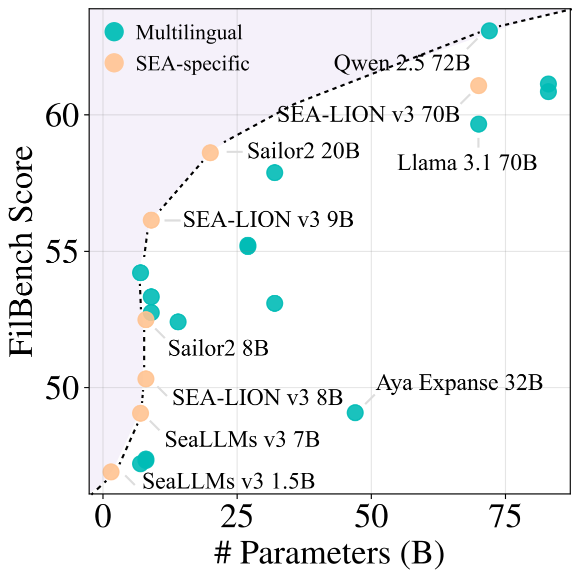


Figure 2: **Parameter-efficiency of LLMs with respect to FilBench.** SEA-specific models are at the Pareto frontier of parameter-efficiency. However, the best SEA-specific model still underperforms on FilBench with a score of 61.07%.

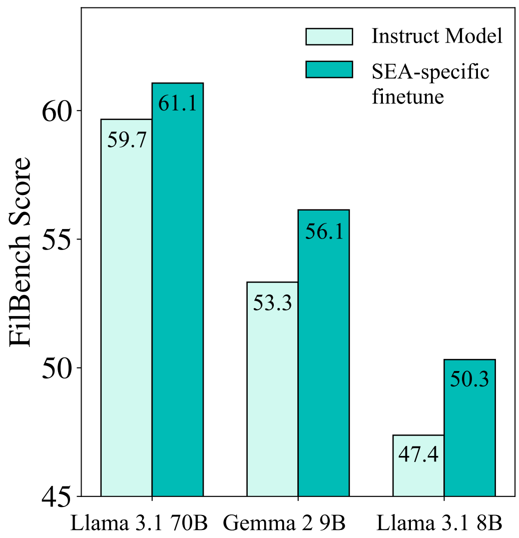


Figure 3: **Effect of language-specific finetuning.** Performance comparison between a base instruction model and its finetuned version (SEA-LION v3). Language-specific finetuning from a multilingual base model can improve performance on FilBench.

Figure 2. In addition, SEA-specific models such as Sailor2 20B, SEA-LION v3 9B, and SeaLLMs v3 1.5B sit near the Pareto frontier in terms of performance and size. Despite these results, the best performing SEA-specific model still underperforms on FilBench, as in the case of SEA-LION v3 70B with a score of 61.07%. In addition, we also find that continuous finetuning of an existing multilingual LLM on SEA-specific data improve FilBench performance, as observed in the SEA-LION model family, which are finetunes of Llama 3.1 and Gemma 2, in Figure 3. These findings show a promising direction for building Filipino-focused LLMs, as it provides a resource-efficient path without training entirely new models from scratch.

Finding #3: Models tend to follow a consistent trend in FilBench performance across categories. Figure 4 suggests that most models have a consistent trend in FilBench performance, i.e., they tend to score well in CK, CN, and RC categories, yet are worse on GN. This is more apparent in generative (GN) tasks, where most models tend to struggle with an average performance of 17.03%. On the other hand, models tend to perform well in CK (60.72%) and CN (85.75%) categories, indicating high-level of understanding of Filipino-centric cultural entities and values. Model performance on CK tends to be more dispersed with one of the largest standard deviation (±13.14). These findings suggest that model capabilities are not uniform across categories for Filipino, indicating significant room for improvement on model training.

5 Analysis: When do LLMs Perform Well or Worse on Filipino Language Tasks?

5.1 Do models consistently agree with one another on Filipino language tasks?

Set-up. In order to understand whether models are consistently reliable in answering test cases in FilBench, we compute the inter-rater reliability using Fleiss' κ across a given set of models. The first group consists of SEA-specific models (see models marked with ♦ in Table 6) while the second group includes the top-five non-SEA models on FilBench (Table 2). To increase granularity, we compute the Fleiss' κ for each sub-task.

Results. The results in Table 3 show that the SEA-specific group consistently demonstrate higher agreement on several sub-tasks than the non-SEA models. This suggests that SEA-specific finetuning can improve model reliability and consistency in outputs. However, both groups show alarming disagreement on cultural tasks, indicating fundamentally different interpretations of culturally-nuanced content. We show some examples of model disagreement for the SEA-specific group in Appendix G. This implies that while re-

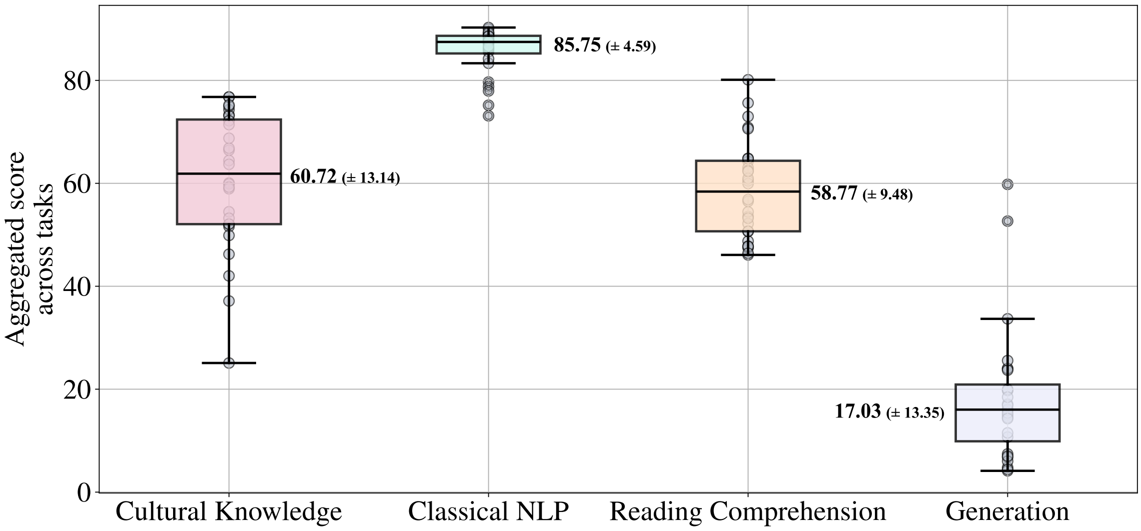


Figure 4: **Performance trends in FilBench.** Model performance (aggregated) across the four categories of FilBench, along with the average performance for each category. LLMs tend to perform well in Classical NLP tasks, but suffer poor performance in Generation tasks.

Sub-Task	**Model Agreement (Fleiss' κ)** SEA-Specific	Top-Five
Classical NLP (CN)		
Text Classification	0.513	0.174
Named-Entity Recog.	0.639	0.273
Sentiment Analysis	0.598	0.212
Cultural Knowledge (CK)		
Regional Knowledge	0.393	0.209
Factual Knowledge	0.224	0.115
Cultural Values	0.403	0.187
Word-sense Disamb.	0.072	-0.041
Reading Comprehension (RC)		
Readability	0.207	-0.119
Reading Comp.	0.377	0.248
NLI	0.438	0.201

Table 3: **Inter-model agreement on MCF-based tasks.** Inter-model agreement, as measured by Fleiss' κ, for each sub-task in FilBench. Despite good performance on FilBench, models tend to disagree with one another, highlighting gaps in reliability.

gional specialization improves reliability, deeper cultural adaptation and more sophisticated training approaches may be needed to achieve reliable performance on Filipino.

5.2 Why do models fail in Generation tasks?

Set-up. Given the low performance of most models in GN tasks, we qualitatively analyze example outputs in order to provide a taxonomy of common failure cases in Filipino generative tasks. In addition, we also assess whether we can mitigate low Generation performance by incorporating demonstrations or few-shot examples in the prompt. To do so, we include $k = \{1, 3, 5\}$ examples on the prompt during evaluation.

Results. We find common failure modes on Generation tasks in FilBench and show some examples from the outputs of the best-performing SEA-specific model, SEA-LION v3 70B, in Table 4:

- **Poor instruction-following.** When presented with Generation tasks from FilBench, models tend to misinterpret instructions or generate translations in an incorrect target language.
- **Overgeneration.** In the case of translation, models tend to produce overly verbose text than necessary, usually until the maximum generation length is reached. This usually results in incoherent text even if properly translated to the target language.
- **Hallucination.** Models often fail in Generation tasks due to spurious artifacts in the generated text. These tend to diminish the faithfulness of the model's output, especially in the case of translation tasks.

By manually inspecting a sample of 100 failure cases from GPT-4o, we find that overgeneration and poor instruction-following are the most dominant, with 47% and 34% respectively, while hallucination occurs 19% of the time. We hypothesize that overgeneration can be caused by training

Failure mode	Example
Poor instruction-following	ENG: Tom winked at Mary. FIL: **Gold:** *Si Tom ay kumindat kay Mary.* **Pred:** *Tom mutiara kepada Mary. (mutiara = to wink).* **Explanation:** Output is not in Filipino.
Overgeneration	ENG: Did you hear the lecture in that class yesterday? asked the student. FIL: **Gold:** *Narinig mo ba ang lecture sa klaseng yan kahapon?, tanong ng estudyante.* **Pred:** *Sige, narinig ko na. ANG LECTURA YESTERDAY BA? Saan po ba ito naganap?* (continued...) **Explanation:** Output is incoherent and generated highly verbose text.
Hallucination	ENG: The more you read the book, the less you will understand it. FIL: **Gold:** *Nang mas na binabasa mo ang libro, menos mong maiintindihan.* **Pred:** *Ang higit na nagbabasa ng libro, mas nagiging delikado at hindi kayang tukuyin ang kahulugan.* **Explanation:** Output hallucinated the concept of danger (*delikado*).

Table 4: **Common failure modes in translation tasks in the Generation category.** We find three common failure modes in most models in the Generation category of FILBENCH. The predictions in this table show the outputs of SEA-LION v3 70B, the currently best-performing SEA-specific model.

data imbalance, as suggested in (Bawden and Yvon, 2023) and (Alves et al., 2023) work in the case for BLOOM and LLaMA 7B. We defer the ablation of training data quality and its effect on translation performance to future work.

In addition, we also find that **few-shot prompting can mitigate drop in Generation performance** (Figure 5). We find that poor instruction-following, which is common especially in zero-shot ENG → FIL decreases once examples are provided. Full few-shot experiment results are shown in Appendix D. Despite these results, model performance on generation tasks remain generally poor, with frequent instances of overgeneration and semantically inaccurate translations. We further explain reasons for this using the Tatoeba dataset, which models consistently underperform on, in Appendix I.

5.3 Human evaluation of FILBENCH

When curating test instances for FILBENCH, we ensured that the majority of sources in underwent human annotation and evaluation. However, we want to verify that strong agreement between native speakers and the gold answers persisted after the instances were converted into our task-specific formulations (Appendix M).

Set-up. In order to evaluate the agreement between native speakers and FILBENCH's gold answers, we sample 150 instances from FILBENCH with similar sub-task distribution. Then, three authors (all native speakers of Filipino) served as annotators to label each instance. For MCF tasks, the annotators choose the letter-option of the correct answer. For GN tasks, we provide the annotators with a free-form text field to input their answers. Then, we compute the inter-annotator agreement via Fleiss' κ across two settings: (i) among annotators (*intra-group*) and the (ii) majority response of human annotators to FILBENCH's gold answer (*inter-group*). For GN, we compute the average ROUGE-L score for each annotator pair (*intra-group*) and the average of the ROUGE-L score between the gold reference translation and each of the annotator translation (*inter-group*).

Task Formulation	Intra-group	Inter-group
MCF, Fleiss' κ	0.8163	0.8756
Generation, Avg. ROUGE-L	0.7604	0.7806

Table 5: **Inter-rater agreement of native-speakers to a subset of FILBENCH.** We show that FILBENCH instances have a strong agreement with native speakers on both MFC-based (Cultural Knowledge, Classical NLP, Reading Comprehension) and Generation tasks.

Results. Table 5 shows the agreement scores among the three annotators (*intra-group*) and their overall agreement with the gold reference answer (*inter-group*). The Fleiss' κ indicate high agreement (Landis and Koch, 1977), suggesting that the instances in FilBench are reliable and aligns with native-speakers. In addition, the ROUGE-L scores between annotators are also high, suggesting that the Generation instances can be reproducibly translated. Furthermore, the inter-group ROUGE-L

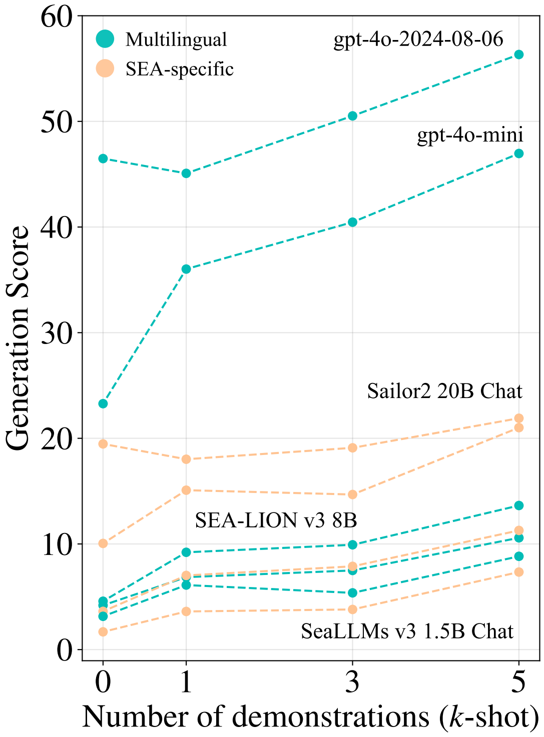


Figure 5: **Effect of few-shot examples on Generation score.** Incorporating a single ($k = 1$) demonstration during generation generally mitigates poor model performance.

score supports this claim, as evidenced by similar performance given that most of the Generation instances were originally translated by other native speakers. In general, the results suggest that the agreement between native speakers and the gold answers are preserved even after converting it into our task-specific formulations.

6 Discussion

On what to prioritize next when collecting data for Filipino-centric post-training. Our findings, through FILBENCH, reveal critical gaps in existing LLM's capabilities to process Filipino text, particularly in generation tasks where the best models achieved $\leq 60\%$ performance. In addition, we also find that continuous finetuning helps improve LLM performance on FILBENCH. This suggests that post-training data collection efforts should prioritize high-quality translation pairs and generative content across diverse domains. Furthermore, gathering training data from a wide range of Philippine languages, beyond just Tagalog, can enhance the performance of LLMs as demonstrated in Buzaaba et al. (2025). We posit that this can be achieved by taking advantage of cross-lingual transfer (Artetxe et al., 2020) across typologically-similar languages.

On the importance of building language-community specific evaluation suites. Our findings strongly reinforce the necessity of developing language-community specific evaluation suites rather than relying on general multilingual benchmarks. FILBENCH demonstrates that even state-of-the-art models like GPT-4o achieve only 75.56% overall performance, indicating that Filipino presents unique challenges not captured in broader evaluations. By creating focused evaluation suites like FILBENCH, the research community can more accurately identify model limitations and track progress in ways that respect the linguistic particularities of Philippine languages. Furthermore, the performance variations across Filipino, Tagalog, and Cebuano emphasize the importance of fine-grained attention to linguistic diversity even within regions. This points to the need for training approaches that recognize intra-regional linguistic boundaries rather than treating Southeast Asian languages as a homogeneous group.

7 Related Work

State of LLM Research for Philippine Languages. Progress in the NLP research landscape for Philippine languages such as Tagalog and Cebuano is seeing a promising growth, which can be attributed to democratization and access to LLM artifacts, particularly data and open models (Lovenia et al., 2024). The first works to release open-source artifacts include tasks such as sentiment analysis, hate speech detection, and natural language inference (NLI) (Cruz and Cheng, 2019, 2020, 2022). Further release of multilingual LLMs supporting Tagalog, allowed researchers to explore further linguistic phenomena from classical NLP tasks (Pilar et al., 2023; Mayhew et al., 2024, *inter alia*) to language model applications (Catapang and Visperas, 2023; Montalan et al., 2024).

Language-specific LLM Evaluation Benchmarks. Global research communities are following the trend of releasing language-specific benchmarks in order to assess and track LLM progress in their respective languages. Notable examples include AfroBench for African languages (Ojo et al., 2023), BenCzechMark for the Czech (Fajcik et al., 2024), the Open Arabic LLM Leaderboard for Arabic (El Filali et al., 2025) and Le Leaderboard for French (Mohamad Alhajar, 2024). These benchmarks usually contain curated tasks that may include translated versions of existing datasets or subsets of larger evaluation suites. FILBENCH takes inspiration from these efforts by curating a comprehensive evaluation suite for Philippine languages.

Region-specific benchmarks also exist such as SeaBench and SeaExam (Liu et al., 2025) for Southeast Asia, although they do not contain any Filipino-specific subset. The most recent effort related to FILBENCH is Batayan (Montalan et al., 2025), which is part of SEA-HELM (Susanto et al., 2025). FILBENCH takes a complementary approach by systematically curating existing benchmarks, enabling not only greater efficiency in resource utilization but also facilitating a wider diversity of task types and expanded coverage of Philippine languages beyond Filipino (in this case, Tagalog and Cebuano).

8 Conclusion

In this work, we present a comprehensive evaluation of LLMs on Filipino-centric tasks to investigate their strengths and limitations, which still remains underexplored. We curate a benchmark called FILBENCH across four categories and 12 sub-tasks, based on our analyses of research priorities in Philippine NLP. Through FILBENCH, we discovered weaknesses in the current open and commercial state-of-the-art LLMs, such as low reliability and poor generation capabilities. FILBENCH emphasizes the value of creating language-specific LLM benchmarks, as it allows us to find promising avenues for models to improve their Filipino-centric performance. Specifically, this includes language-specific post-training and collecting relevant training datasets for text generation. We hope that FILBENCH aids in driving the progress in Filipino NLP.

Limitations

Influence of training data on downstream FILBENCH performance. When selecting models for evaluation on FILBENCH, we categorized based on whether these models were originally presented as multilingual or SEA-specific, rather than considering the proportion of Filipino-centric training data used for fine-tuning. The training data provenance is difficult to track, especially for closed-source models. This explains why models that top other multilingual leaderboards such as Aya Expanse 32B (Dang et al., 2024) perform poorly on FILBENCH, because it was not explicitly trained on Filipino. Our experiments hinge on the assumption that cross-lingual transfer (Artetxe et al., 2020) happens during different states of language modeling, as evidenced in Chirkova and Nikoulina (2024). We leave the systematic exploration of the influence of the proportion of language-specific training data to a language-specific benchmark (i.e., Filipino-centric training data to FILBENCH performance) for future work.

Focus on Tagalog and Cebuano. While some of the datasets in FILBENCH support other Philippine languages (i.e. Ilokano for Belebele), data for these (labeled or otherwise) remain sparse. We focus our suite on the relatively better-resourced Filipino and Cebuano, with the hope of supporting more languages once more datasets become available. Future work might explore data augmentation techniques and other community-driven data collection initiatives to extend FILBENCH's coverage to languages like Hiligaynon, Bikolano, and others.

Ethics Statement

The development and evaluation of language technologies for Filipino, Cebuano, and other Philippine languages addresses important issues of linguistic inclusion and technological access. FILBENCH aims to support the development of more capable Filipino language technologies that can serve the significant population of Filipino speakers worldwide. The benchmark deliberately includes culturally-specific knowledge assessment to address known biases in LLMs toward Western values and content. This highlights the importance of evaluating models in their cultural context rather than assuming universal applicability. Datasets included in FILBENCH are from publicly accessible sources, and the authors obtained explicit approval from dataset creators when license information was unclear. Overall, we do not see any serious ethical issues with this work.

Acknowledgments

The authors would like to thank Cohere Labs for providing credits through the Cohere Research Grant to run the Aya model series, and Together AI for additional computational credits for running several open models. We also acknowledge the Hugging Face team, particularly the OpenEvals team (Clémentine Fourrier and Nathan Habib) and Daniel van Strien, for their support in publishing the FILBENCH blog post. JMI is supported by the National University Philippines and the UKRI Centre for Doctoral Training in Accountable, Responsible, and Transparent AI [EP/S023437/1] of the University of Bath. Finally, we thank the reviewers from the May ARR cycle for their helpful feedback and insightful comments that improved this paper.

References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245.

Alham Fikri Aji, Jessica Zosa Forde, Alyssa Marie Loo, Lintang Sutawika, Skyler Wang, Genta Indra Winata, Zheng-Xin Yong, Ruochen Zhang, A. Seza Doğruöz, Yin Lin Tan, and Jan Christian Blaise Cruz. 2023. Current status of NLP in south East Asia with insights from multilingualism and language diversity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Tutorial Abstract*, pages 8–13, Nusa Dua, Bali. Association for Computational Linguistics.

Duarte Alves, Nuno Guerreiro, João Alves, José Pombal, Ricardo Rei, José de Souza, Pierre Colombo, and Andre Martins. 2023. Steering large language models for machine translation with finetuning and in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11127–11148, Singapore. Association for Computational Linguistics.

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, and 1 others. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the translation initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Dhananjay Ashok and Zachary C Lipton. 2023. Promptner: Prompting for named entity recognition. *arXiv preprint arXiv:2305.15444*.

Lita Bacalla. 2019. Morpo-analisis ng wikang tagalog at wikang sugbuanun'g binisaya: Pahambing na pagaaral. *International Journal of Resarch Studies in Education*, 8:55–65.

Renz Iver Baliber, Charibeth Cheng, Kristine Mae Adlaon, and Virgion Mamonong. 2020. Bridging Philippine languages with multilingual neural machine translation. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 14–22, Suzhou, China. Association for Computational Linguistics.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The Belebele Benchmark: a Parallel Reading Comprehension Dataset in 122 Language Variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775.

Maria Bardají, Elsie Or, Angelina Aquino, and Nikolaus Himmelmann. 2024. The challenges of symmetrical voice languages for universal dependencies. In *Proceedings of the 15th International Conference of the Association for Linguistic Typology*.

Rachel Bawden and François Yvon. 2023. Investigating the translation performance of a large multilingual language model: the case of BLOOM. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 157–170.

Happy Buzaaba, Alexander Wettig, David Ifeoluwa Adelani, and Christiane Fellbaum. 2025. Lughallama: Adapting large language models for african languages. *arXiv preprint arXiv:2504.06536*.

Kenrick Lance Buñag and Rosanna Esquivel. 2023. Transformer-based conditional language models to generate filipino news article. In *Proceedings of the International Conference on International Engineering and Operations Management*.

Samuel Cahyawijaya, Ruochen Zhang, Holy Lovenia, Jan Christian Blaise Cruz, Elisa Gilbert, Hiroki Nomoto, and Alham Fikri Aji. 2024. Thank you, stingray: Multilingual large language models can not (yet) disambiguate cross-lingual word sense. *arXiv preprint arXiv:2410.21573*.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67.

Jasper Kyle Catapang and Moses Visperas. 2023. Emotion-based morality in Tagalog and English scenarios (EMoTES-3K): A parallel corpus for explaining (im)morality of actions. In *Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages*, pages 1–6, Tokyo, Japan. Association for Computational Linguistics.

Nadezhda Chirkova and Vassilina Nikoulina. 2024. Zero-shot cross-lingual transfer in instruction tuning of large language models. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 695–708, Tokyo, Japan. Association for Computational Linguistics.

Camilla Johnine Cosme and Marlene De Leon. 2023. Sentiment analysis of code-switched filipino-english product and service reviews using transformers-based large language models. In *Proceedings of World Conference on Information Systems for Business Management*, pages 123–135.

Jan Christian Blaise Cruz and Charibeth Cheng. 2019. Evaluating language model finetuning techniques for low-resource languages. *arXiv preprint arXiv:1907.00409*.

Jan Christian Blaise Cruz and Charibeth Cheng. 2020. Establishing baselines for text classification in low-resource languages. *arXiv preprint arXiv:2005.02068*.

Jan Christian Blaise Cruz and Charibeth Cheng. 2022. Improving large-scale language models and resources for Filipino. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6548–6555, Marseille, France. European Language Resources Association.

Jan Christian Blaise Cruz, Jose Kristian Resabal, James Lin, Dan John Velasco, and Charibeth Cheng. 2021. Exploiting news article structure for automatic corpus generation of entailment datasets. In *PRICAI 2021: Trends in Artificial Intelligence*, pages 86–99.

Micholo Cucio and Tristan Hennig. 2025. Artificial Intelligence and the Philippine Labor Market: Mapping Occupational Exposure and Complementarity. Technical report, International Monetary Fund.

Preetam Prabhu Srikar Dammu, Hayoung Jung, Anjali Singh, Monojit Choudhury, and Tanu Mitra. 2024. "they are uncultured": Unveiling covert harms and social threats in LLM generated conversations. pages 20339–20369.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, and 1 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*.

Longxu Dou, Qian Liu, Fan Zhou, Changyu Chen, Zili Wang, Ziqi Jin, Zichen Liu, Tongyao Zhu, Cunxiao Du, Penghui Yang, and 1 others. 2025. Sailor2: Sailing in South-East Asia with Inclusive Multilingual LLMs. *arXiv preprint arXiv:2502.12982*.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2024. *Ethnologue: Languages of the World*, 27 edition. SIL International, Dallas, Texas.

Ali El Filali, Manel ALOUI, Tarique Husaain, Ahmed Alzubaidi, Basma El Amel Boussaha, Ruxandra Cojocaru, Clémentine Fourrier, Nathan Habib, and Hakim Hacid. 2025. Open arabic llm leaderboard 2. https://huggingface.co/spaces/OALL/Open-Arabic-LLM-Leaderboard.

Juuso Eronen, Michal Ptaszynski, and Fumito Masui. 2023. Zero-shot cross-lingual transfer language selection using linguistic similarity. *Information Processing and Management*, 60(3):103250.

Martin Fajcik, Martin Docekal, Jan Dolezal, Karel Ondrej, Karel Beneš, Jan Kapsa, Pavel Smrz, Alexander Polok, Michal Hradis, Zuzana Neverilova, and 1 others. 2024. Benczechmark: A czech-centric multitask and multimetric benchmark for large language models with duel scoring mechanism. *arXiv preprint arXiv:2412.17933*.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for (mt) evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24.

Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. 2024. Open llm leaderboard v2. `https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard`.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The LLaMa 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Boston Consulting Group. 2024. Consumers know more about ai than businesses think.

Yuling Gu, Oyvind Tafjord, Bailey Kuehl, Dany Haddad, Jesse Dodge, and Hannaneh Hajishirzi. 2024. OLMES: A standard for language model evaluations. *arXiv preprint arXiv:2406.08446*.

Sireesh Gururaja, Amanda Bertsch, Clara Na, David Widder, and Emma Strubell. 2023. To build our future, we must know our past: Contextualizing paradigm shifts in natural language processing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13310–13325, Singapore. Association for Computational Linguistics.

Nathan Habib, Clémentine Fourrier, Hynek Kydlíček, Thomas Wolf, and Lewis Tunstall. 2023. Lighteval: A lightweight framework for llm evaluation.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246.

Xu Huang, Wenhao Zhu, Hanxu Hu, Conghui He, Lei Li, Shujian Huang, and Fei Yuan. 2025. Benchmax: A comprehensive multilingual evaluation suite for large language models. *arXiv preprint arXiv:2502.07346*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. GPT-4o System Card. *arXiv preprint arXiv:2410.21276*.

Joseph Marvin Imperial and Ekaterina Kochmar. 2023. Automatic readability assessment for closely related languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5371–5386, Toronto, Canada. Association for Computational Linguistics.

Joseph Marvin Imperial, Lloyd Lois Antonie Reyes, Michael Antoinio Ibañez, Ranz Sapinit, and Mohammed Hussien. 2022. A baseline readability model for cebuano. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, and 1 others. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

J Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33 1:159–74.

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.

Chaoqun Liu, Wenxuan Zhang, Jiahao Ying, Mahani Aljunied, Anh Tuan Luu, and Lidong Bing. 2025. Seaexam and seabench: Benchmarking llms with local multilingual questions in southeast asia. *arXiv preprint arXiv:2502.06298*.

Yan Liu and He Wang. 2024. *Who on Earth Is Using Generative AI?* World Bank.

Evan Dennison Livelo and Charibeth Cheng. 2018. Intelligent dengue infoveillance using gated recurrent neural learning and cross-label frequencies. In *2018 IEEE International Conference on Agents*.

Holy Lovenia, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James Validad Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno P. Kampman, Joel Ruben Antony Moniz, Muhammad Ravi Shulthan Habibi, Frederikus Hudi, Railey Montalan, Ryan Ignatius Hadiwijaya, Joanito Agili Lopo, William Nixon, Börje F. Karlsson, James Jaya, and 42 others. 2024. SEACrowd: A multilingual multimodal data hub and benchmark suite for Southeast Asian languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5155–5203, Miami, Florida, USA. Association for Computational Linguistics.

Stephen Mayhew, Terra Blevins, Shuheng Liu, Marek Suppa, Hila Gonen, Joseph Marvin Imperial, Börje Karlsson, Peiqin Lin, Nikola Ljubešić, Lester James Miranda, Barbara Plank, Arij Riabi, and Yuval Pinter. 2024. Universal NER: A gold-standard multilingual named entity recognition benchmark. 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 4322–4337.

Curtis D McFarland. 2008. Linguistic diversity and english in the philippines. *Philippine English: Linguistic and literary perspectives*, 1:131.

Meta AI. 2025. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. `https://ai.meta.com/blog/llama-4-multimodal-intelligence/`. Blog post, accessed May 16, 2025.

Romylyn A Metila, Lea Angela S Pradilla, and Alan B Williams. 2016. The challenge of implementing mother tongue education in linguistically diverse contexts: The case of the philippines. *The Asia-Pacific Education Researcher*, 25:781–789.

Lester James Miranda. 2023. Developing a named entity recognition dataset for Tagalog. In *Proceedings of the First Workshop in South East Asian Language Processing*, pages 13–20, Nusa Dua, Bali, Indonesia. Association for Computational Linguistics.

Mistral AI. 2024. Mixtral of experts. `https://mistral.ai/news/ministraux`. Blog post, accessed May 16, 2025.

Alexandre Lavallée Mohamad Alhajar. 2024. Open llm french leaderboard v0.2. `https://huggingface.co/spaces/le-leadboard/OpenLLMFrenchLeaderboard`.

Jann Railey Montalan, Jimson Paulo Layacan, David Demitri Africa, Richell Isaiah Flores, Michael T Lopez II, Theresa Denise Magsajo, Anjanette Cayabyab, and William Chandra Tjhi. 2025. Batayan: A filipino nlp benchmark for evaluating large language models. *arXiv preprint arXiv:2502.14911*.

Jann Railey Montalan, Jian Gang Ngui, Wei Qi Leong, Yosephine Susanto, Hamsawardhini Rengarajan, Alham Fikri Aji, and William Chandra Tjhi. 2024. Kalahi: A handcrafted, grassroots cultural LLM evalutation suite for filipino. *arXiv preprint arXiv:2409.15380*.

Raymond Ng, Thanh Ngan Nguyen, Yuli Huang, Ngee Chia Tai, Wai Yi Leong, Wei Qi Leong, Xianbin Yong, Jian Gang Ngui, Yosephine Susanto, Nicholas Cheng, Hamsawardhini Rengarajan, Peerat Limkonchotiwat, Adithya Venkatadri Hulagadri, Kok Wai Teng, Yeo Yeow Tong, Bryan Siow, Wei Yi Teo, Wayne Lau, Choon Meng Tan, and 12 others. 2025. Sea-lion: Southeast asian languages in one network. *Preprint*, arXiv:2504.05747.

Nathaniel Oco and Rachel Roxas. 2018. A survey of machine translation work in the Philippines: From 1998 to 2018. In *Proceedings of the AMTA 2018 Workshop on Technologies for MT of Low Resource Languages (LoResMT 2018)*, pages 30–36, Boston, MA. Association for Machine Translation in the Americas.

Jessica Ojo, Kelechi Ogueji, Pontus Stenetorp, and David Ifeoluwa Adelani. 2023. How good are large language models on african languages? *arXiv preprint arXiv:2311.07978*.

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.

Philippine Statistics Authority. 2020. Household population, number of households and average household size of the philippines (2020 census of population and housing). Accessed: 2025-04-03.

Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5877–5891.

Ma. Beatrice Emanuela Pilar, Dane Dedoroy, Ellyza Mari Papas, Mary Loise Buenaventura, Myron Darrel Montefalcon, Jay Rhald Padilla, Joseph Marvin Imperial, Mideth Abisado, and Lany Maceda. 2023. CebuaNER: A new baseline Cebuano named entity recognition model. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 792–800.

Haoyi Qiu, Alexander R. Fabbri, Divyanish Agarwal, Kung-Hsiang Huang, Sarah Tan, Nanyun Peng, and Chien-Sheng Wu. 2025. Evaluating cultural and social awareness of llm web agents. In *Findings of the Association for Computational Linguistics: NAACL 2025*.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A Haggag, Alfonso Amayuelas, and 1 others. 2024. INCLUDE: Evaluating multilingual language understanding with regional knowledge. *arXiv preprint arXiv:2411.19799*.

Rachel Edita O. Roxas, Joseph Marvin Imperial, and Angelica H. De La Cruz. 2021. Science mapping of publications in natural language processing in the Philippines: 2006 to 2020. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 721–730, Shanghai, China. Association for Computational Linguistics.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, and 1 others. 2024. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *arXiv preprint arXiv:2412.03304*.

Yosephine Susanto, Adithya Venkatadri Hulagadri, Jann Railey Montalan, Jian Gang Ngui, Xian Bin Yong, Weiqi Leong, Hamsawardhini Rengarajan, Peerat Limkonchotiwat, Yifan Mai, and William Chandra Tjhi. 2025. SEA-HELM: Southeast asian holistic evaluation of language models. *arXiv preprint arXiv:2502.14301*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resoure and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182.

Sonny Villafania. 2007. Filipino and Tagalog, not so different. Archived from the original on 2014-05-22.

Yu Wan, Baosong Yang, Derek Fai Wong, Lidia Sam Chao, Liang Yao, Haibo Zhang, and Boxing Chen. 2022. Challenges of neural machine translation for short texts. *Computational Linguistics*, 48(2):321–342.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.

Xiang Yue, Yueqi Song, Akari Asai, Seungone Kim, Jean de Dieu Nyandwi, Simran Khanuja, Anjali Kantharuban, Lintang Sutawika, Sathyanarayanan Ramamoorthy, and Graham Neubig. 2024. Pangea: A Fully Open Multilingual Multimodal LLM for 39 Languages. In *The Thirteenth International Conference on Learning Representations*.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting Large Language Model for Machine Translation: A Case Study. *arXiv preprint arXiv:2301.07069*.

Wenxuan Zhang, Hou Pong Chan, Yiran Zhao, Mahani Aljunied, Jianyu Wang, Chaoqun Liu, Yue Deng, Zhiqiang Hu, Weiwen Xu, Yew Ken Chia, and 1 others. 2024. SeaLLMs 3: Open Foundation and Chat Multilingual Large Language Models for Southeast Asian Languages. *arXiv preprint arXiv:2407.19672*.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023b. Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*.

Yiran Zhao, Chaoqun Liu, Yue Deng, Jiahao Ying, Mahani Aljunied, Zhaodonghui Li, Lidong Bing, Hou Pong Chan, Yu Rong, Deli Zhao, and 1 others. 2025. Babel: Open Multilingual Large Language Models Serving Over 90% of Global Speakers. *arXiv preprint arXiv:2503.00865*.

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

Appendix

A Details of Models Evaluated on FilBench **16**

B FilBench Dataset Licenses **16**

C Full results on FilBench **17**

D Generation Few-shot Results **17**

E Evaluation Infrastructure and Runtime **18**

F Extended Related Work **18**
F.1 Comparison to other SEA-specific / Filipino benchmarks 18
F.2 Does high performance in one benchmark translates similarly to FilBench? 18

G Analysis of Model Dis/Agreement **18**
G.1 Set-up: Qualitative Analysis of Model Outputs during Dis/Agreement 19
G.2 Results: Examples of Model Dis/Agreement 19
G.3 Discussion: Implications and Potential Future Work 20

H Generation Failure Modes Examples **22**

I Analysis of Generation Failure Modes for Translation Tasks **23**
I.1 Set-up: Case Study of Tatoeba 23
I.2 Results: Common Reasons why Models Fail in Generation Tasks 23
I.3 Discussion 23

J Research Priorities in Filipino NLP **24**

K Cost-Efficiency of LLMs on Filipino Language Tasks **26**

L Effect of Prompt Template in Generation Performance **26**

M Task Formulation **28**

N Annotation Instructions **34**

O Use of AI Assistants **34**

A Details of Models Evaluated on FILBENCH

Table 6 shows the details of all models evaluated on FILBENCH.

Model	# Params (B)	# Lang.	License	Reference
○ gpt-4o-2024-08-06	–	–	OpenAI ToS	Hurst et al. (2024)
○ gpt-4o-mini	–	–	OpenAI ToS	Hurst et al. (2024)
○ CohereForAI/aya-expanse-32b	32	23	CC BY NC 4.0	Dang et al. (2024)
○ meta-llama/Llama-4-Maverick-17B-128E-Instruct-FP8	400 (17)	200	Llama 4 License	Meta AI (2025)
○ meta-llama/Llama-4-Scout-17B-16E-Instruct	109 (17)	200	Llama 4 License	Meta AI (2025)
○ meta-llama/Llama-3.1-70B-Instruct	70	30	Llama 3.1 License	Grattafiori et al. (2024)
○ meta-llama/Llama-3.1-8B-Instruct	8	30	Llama 3.1 License	Grattafiori et al. (2024)
○ Qwen/Qwen2.5-72B-Instruct	72	29	Qwen License	Yang et al. (2024)
○ Qwen/Qwen2.5-32B-Instruct	32	29	Apache 2.0	Yang et al. (2024)
○ Qwen/Qwen2.5-14B-Instruct	14	29	Apache 2.0	Yang et al. (2024)
○ Qwen/Qwen2.5-7B-Instruct	7	29	Apache 2.0	Yang et al. (2024)
○ Tower-Babel/Babel-83B-Chat	83	25	SeaLLM License	Zhao et al. (2025)
○ Tower-Babel/Babel-9B-Chat	9	25	SeaLLM License	Zhao et al. (2025)
○ google/gemma-3-27b-it	27	73	Gemma License	Team et al. (2025)
○ google/gemma-2-27b-it	27	73	Gemma License	Team et al. (2024)
○ google/gemma-2-9b-it	9	73	Gemma License	Team et al. (2024)
○ mistralai/Ministral-8B-Instruct-2410	8	10	Mistral AI License	Mistral AI (2024)
○ mistralai/Mixtral-8x22B-Instruct-v0.1	141 (39)	5	Apache 2.0	Jiang et al. (2024)
○ mistralai/Mixtral-8x7B-Instruct-v0.1	47 (13)	5	Apache 2.0	Jiang et al. (2024)
○ neulab/Pangea-7B	7	39	Apache 2.0	Yue et al. (2024)
♦ aisingapore/Llama-SEA-LION-v3-70B-IT	70	13	Llama 3.1 License	Ng et al. (2025)
♦ aisingapore/Gemma-SEA-LION-v3-9B-IT	9	13	Gemma License	Ng et al. (2025)
♦ aisingapore/Llama-SEA-LION-v3-8B-IT	8	13	Llama 3.1 License	Ng et al. (2025)
♦ sail/Sailor2-20B-Chat	20	12	Apache 2.0	Dou et al. (2025)
♦ sail/Sailor2-8B-Chat	8	12	Apache 2.0	Dou et al. (2025)
♦ SeaLLMs/SeaLLMs-v3-7B-Chat	7	12	SeaLLM License	Zhang et al. (2024)
♦ SeaLLMs/SeaLLMs-v3-1.5B-Chat	1.5	12	SeaLLM License	Zhang et al. (2024)

Table 6: All models evaluated on FILBENCH. We evaluate several models with different multilingual capabilities (multilingual ○, SEA-specific ♦), sizes (1.5B to 400B), and accessibility (open-source vs. commercial). For Mixture-of-Experts models, parameters are denoted as "Total Parameters (Active Parameters)". Models that are finetuned on top of a pre-trained model have the number of languages supported based on their fine-tuning data.

B FILBENCH Dataset Licenses

Table 7 provides information for all datasets in FILBENCH, such as their license and data collection process.

Category	Dataset	Source	Annotation	License
CN	Dengue Filipino (Livelo and Cheng, 2018)	Social media (Twitter)	Expert-annotated	Unknown
	BalitaNLP (Buñag and Esquivel, 2023)	News articles	Included from source	Unknown
	SIB-200 (Adelani et al., 2024)	Human-translation	Expert-annotated	CC BY SA 4.0
	CebuaNER (Pilar et al., 2023)	News articles	Expert-annotated	CC BY NC SA 4.0
	TLUnified-NER (Miranda, 2023)	News articles	Expert-annotated	GPL v3.0
	Universal NER (Mayhew et al., 2024)	Universal Dependencies	Expert-annotated	CC BY SA 4.0
	FiReCS (Cosme and De Leon, 2023)	Reviews (Maps and Shopee)	Expert-annotated	CC BY 4.0
CK	INCLUDE (Romanou et al., 2024)	Local exams	Expert-annotated	Apache 2.0
	Global MMLU (Singh et al., 2024)	MMLU dataset	Translated with validation	Apache 2.0
	KALAHI (Montalan et al., 2024)	Human-provided	Expert-annotated	CC BY 4.0
	StingrayBench (Cahyawijaya et al., 2024)	Human-provided	Expert-annotated	CC BY SA 4.0
RC	Cebuano Readability Corpus (Imperial et al., 2022)	Book repositories	Expert-annotated	MIT
	Belebele (Bandarkar et al., 2024)	Wikipedia	Expert-annotated	CC BY SA 4.0
	NewsPH NLI (Cruz et al., 2021)	News articles	Semi-supervised	Unknown
GN	NTREX-128 (Federmann et al., 2022)	Translated from WMT19	Expert-annotated	CC BY SA 4.0
	Tatoeba (Tiedemann, 2020)	Crowd-sourced	Crowd-sourced	CC BY 2.0
	TICO-19 (Anastasopoulos et al., 2020)	News, Wikipedia, PubMed	Semi-supervised	CC0 1.0

Table 7: Supplemental information for all datasets included in FILBENCH. For datasets with “Unknown” licenses, we obtained explicit approval from the authors to include them in our evaluation suite.

C Full results on FILBENCH

Table 8 shows the full aggregated results for the 27 models evaluated on FILBENCH.

Model	FILBENCH Score	Cultural Knowledge	Classical NLP	Reading Comp.	Generation
○ gpt-4o-2024-08-06	72.73±1.66	73.29±3.01	89.03±2.05	80.12±0.90	46.48±0.60
○ meta-llama/Llama-4-Maverick-17B-128E-Instruct-FP8	67.67±1.04	76.75±3.04	87.28±0.26	72.99±0.18	33.67±0.71
○ meta-llama/Llama-4-Scout-17B-16E-Instruct	63.20±1.05	74.31±3.14	87.88±0.25	70.86±0.18	19.75±0.63
○ Qwen/Qwen2.5-72B-Instruct	63.08±0.99	73.11±3.22	88.60±0.24	75.62±0.17	14.98±0.33
♦ aisingapore/Llama-SEA-LION-v3-70B-IT	61.07±0.95	76.78±3.02	89.99±0.23	53.56±0.19	23.95±0.34
○ Tower-Babel/Babel-83B-Chat	60.85±0.96	75.21±3.11	88.81±0.25	64.85±0.19	14.53±0.29
○ meta-llama/Llama-3.1-70B-Instruct	59.66±1.17	72.16±3.21	90.27±0.83	52.17±0.28	24.03±0.37
♦ sail/Sailor2-20B-Chat	58.61±1.06	66.43±3.41	89.03±0.25	63.03±0.19	15.95±0.38
○ Qwen/Qwen2.5-32B-Instruct	57.88±1.45	66.83±3.45	89.32±1.99	70.59±0.18	4.79±0.17
♦ aisingapore/Gemma-SEA-LION-v3-9B-IT	56.14±1.53	64.44±3.43	88.55±0.25	54.46±0.20	17.10±2.25
○ google/gemma-2-27b-it	55.22±1.04	68.76±3.32	87.99±0.25	48.77±0.19	15.38±0.38
○ google/gemma-3-27b-it	55.17±0.99	71.41±3.24	88.61±0.24	53.23±0.19	7.42 ±0.30
○ mistralai/Mixtral-8x22B-Instruct-v0.1	54.28±1.09	54.47±3.62	87.19±0.25	64.78±0.19	10.70±0.31
○ google/gemma-2-9b-it	53.33±1.08	63.69±3.47	87.47±0.25	50.65±0.20	11.51±0.40
○ Tower-Babel/Babel-9B-Chat	52.75±1.48	60.06±3.57	87.67±1.90	56.49±0.20	6.79 ±0.26
♦ sail/Sailor2-8B-Chat	52.49±1.10	58.94±3.57	86.03±0.27	50.69±0.23	14.29±0.36
○ Qwen/Qwen2.5-14B-Instruct	52.41±1.63	59.27±3.61	86.27±2.56	59.95±0.20	4.14 ±0.14
○ Qwen/Qwen2.5-7B-Instruct	50.46±1.08	51.61±3.68	85.58±0.27	60.47±0.20	4.19 ±0.15
♦ aisingapore/Llama-SEA-LION-v3-8B-IT	50.32±1.08	59.89±3.56	83.33±0.28	47.47±0.10	10.60±0.29
○ mistralai/Mixtral-8x7B-Instruct-v0.1	50.26±1.09	49.88±3.67	84.19±0.29	60.95±0.19	6.02 ±0.31
♦ SeaLLMs/SeaLLMs-v3-7B-Chat	49.06±1.06	52.04±3.66	79.68±0.33	62.47±0.19	2.08 ±0.10
○ CohereForAI/aya-expanse-32b	47.84±1.41	53.22±3.65	87.47±1.60	46.09±0.21	4.58 ±0.16
○ meta-llama/Llama-3.1-8B-Instruct	47.38±1.51	52.08±3.68	86.61±1.90	46.42±0.24	4.42 ±0.20
○ mistralai/Ministral-8B-Instruct-2410	47.33±1.66	42.02±3.62	77.95±2.59	62.33±0.20	7.00 ±0.25
○ neulab/Pangea-7B	43.98±1.08	46.23±3.70	78.80±0.29	47.74±0.22	3.15 ±0.15
♦ SeaLLMs/SeaLLMs-v3-1.5B-Chat	43.20±1.07	37.14±3.61	75.17±0.33	56.85±0.20	2.08 ±0.14
○ gpt-4o-mini	42.32±1.81	25.09±3.26	73.12±3.18	47.78±0.34	23.29±0.59

Table 8: Model performance on FILBENCH. We evaluate several models with different multilingual capabilities (multilingual ○, SEA-specific ♦), sizes (8B to 400B), and accessibility (open-source vs. commercial).

D Generation Few-shot Results

Table 9 shows the full few-shot experiment results on the Generation category of FILBENCH for 9 selected models.

	Tatoeba - TGL (ENG → FIL)				Tatoeba - CEB (CEB → ENG)				NTREX-128 (ENG → FIL)				TICO-19 (ENG → FIL)			
Model / k-shot #	0	1	3	5	0	1	3	5	0	1	3	5	0	1	3	5
○ gpt-4o-2024-08-06	51.88	60.23	60.62	61.65	33.78	59.37	62.99	63.98	38.96	57.09	58.47	58.56	53.03	64.42	64.08	65.15
○ gpt-4o-mini	12.13	51.69	55.20	60.23	27.07	49.57	58.30	58.71	27.83	54.67	57.81	58.30	41.16	52.24	64.08	64.43
♦ Sailor/Sailor2-20B-Chat	15.88	17.13	18.31	22.34	13.67	10.60	12.07	13.19	23.41	44.45	43.84	44.50	22.88	54.21	53.08	55.05
♦ aisingapore/Llama-SEA-LION-v3-8B-IT	1.45	14.93	15.25	15.10	9.01	10.75	12.33	12.26	14.79	39.74	40.41	40.52	22.84	44.04	43.91	44.05
○ CohereForAI/aya-expanse-32b	0.80	14.03	13.86	13.60	8.31	10.27	11.51	11.93	6.72	33.71	33.70	36.12	8.48	39.55	38.88	39.92
♦ SeaLLMs/SeaLLMs-v3-7B-Chat	0.65	11.17	11.70	12.07	6.44	7.62	9.47	9.61	5.65	32.46	33.88	36.50	6.01	39.41	39.13	39.83
○ Qwen/Qwen-2.5-7B-Instruct	0.72	8.37	8.72	9.38	6.60	7.38	8.93	9.72	6.99	28.72	29.59	30.67	6.43	32.06	32.40	33.16
○ neulab/Pangea-7B	0.53	5.73	7.56	7.69	7.06	6.65	8.29	8.41	4.59	23.36	24.15	25.60	5.40	31.28	28.95	28.73
♦ SeaLLMs/SeaLLMs-v3-1.5B-Chat	0.78	4.9	6.43	6.99	4.51	6.72	6.36	6.34	2.04	22.97	26.10	28.16	2.15	24.45	32.27	32.72

Table 9: Generation scores for few-shot prompting on selected models (multilingual ○, SEA-specific ♦).

E Evaluation Infrastructure and Runtime

We built FilBench on top of LightEval (Habib et al., 2023). When using the vLLM backend (Kwon et al., 2023), evaluating on the whole suite *sequentially* can take 4.93 hours on 2 NVIDIA H100 GPUs for models under 83B parameters. However, the evaluation suite can be parallelized per benchmark, with the runtime distribution shown in Figure 6. The longest-running task can take approximately 1 hour and 28 minutes and the shortest task takes only 5.86 minutes.

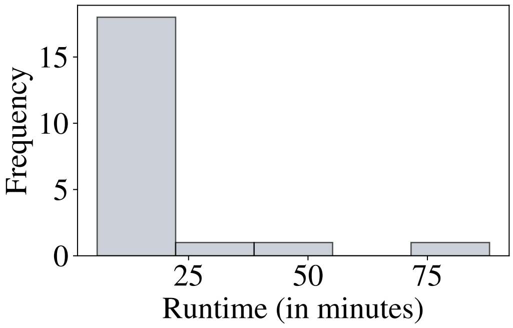


Figure 6: Runtime of different benchmarks for a 32B model on FilBench (2 × H100 NVIDIA GPU).

F Extended Related Work

In this section, we focus on other benchmarking efforts related to FilBench. First, we compare the differences across these efforts (§F.1) and show FilBench's value in providing a more focused evaluation for Filipino. Then, we discuss whether there is a transferability in performance when evaluating from one benchmark to another (§F.2).

F.1 Comparison to other SEA-specific / Filipino benchmarks

Table 10 shows benchmarking efforts orthogonal to FilBench. These efforts focus on a specific region, i.e., Southeast Asia or SEA, and comprises of datasets from countries other than the Philippines. In general, we find that SEA-specific benchmarks do not contain any Philippine language at all (as in the case of the SeaLLM Leaderboard, Liu et al. 2025) or is limited to a single Filipino language (Tagalog, as in the case of SEA-HELM, Susanto et al. 2025). FilBench aims to provide a more realistic evaluation of Filipino-centric tasks by having a principled approach in choosing categories that reflect the current trends and priorities of the Philippine NLP research community.

F.2 Does high performance in one benchmark translates similarly to FilBench?

Set-up. In order to understand whether high performance in one benchmark translates to similar performance in FilBench, we compute the Spearman ρ rank correlation of models that were evaluated in both benchmarks. For the SeaLLM leaderboard, we treat SeaBench and SeaExam separately. For SEA-HELM, we compute the correlation for the full evaluation suite and its Tagalog-only subset (Batayan).

Results. Figure 7 shows the raw scores for FilBench with respect to another benchmark, alongside its Spearman ρ rank correlation. The results show moderate to strong positive correlations (ρ = 0.571 to 0.758) between FilBench and other SEA language benchmarks, with SeaExam demonstrating the strongest predictive relationship. This suggests that model performance on one benchmark does meaningfully transfer to performance on Filipino language tasks, though the scattered distribution of data points indicates that different benchmarks capture distinct aspects of language ability. Furthermore, our findings highlight that while some transferability exists across Southeast Asian language benchmarks, benchmark-specific optimization may still be necessary for optimal performance on FilBench.

G Analysis of Model Dis/Agreement

In this section, we show examples of agreement and disagreement from the SEA-specific models

Benchmark	# Tasks	# Instances	PH Languages	Data Collection Procedure
FilBench (Ours)	12	197.6k	FIL/TGL, CEB	Curated from expert-annotated datasets
SeaBench (Liu et al., 2025)	1	300	-	Collected from native-speakers
SeaExam (Liu et al., 2025)	3	5.5k	-	Collected from native-speakers
Batayan (Montalan et al., 2025)	8	3.8k	FIL	Curated with human annotation

Table 10: Comparison of multilingual benchmarks related to Filipino-centric tasks. Our data collection procedure allows us to scale the diversity of tasks in our suite.

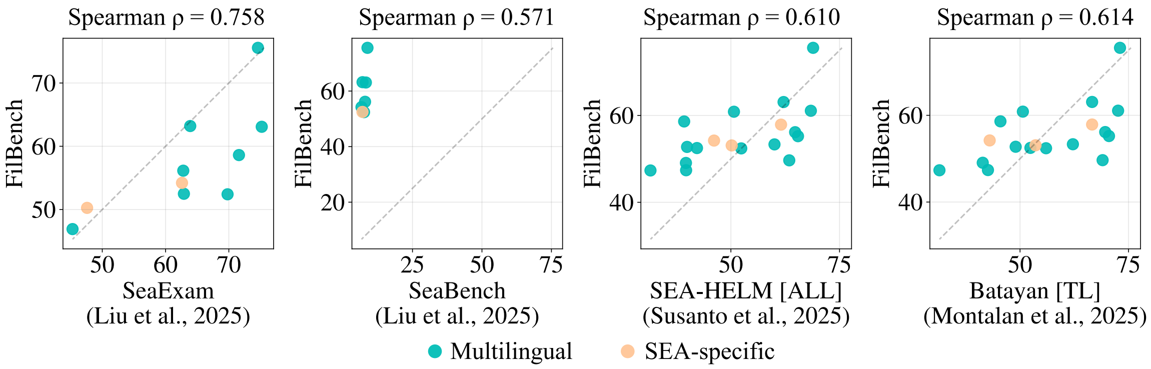


Figure 7: Performance of different Multilingual and SEA-Specific models on FilBench and other SEA-specific / Tagalog benchmarks such as SeaExam and SeaBench (Liu et al., 2025), SEA-HELM (Susanto et al., 2025), and Batayan (Montalan et al., 2025).

we analyzed in §5.1.

G.1 Set-up: Qualitative Analysis of Model Outputs during Dis/Agreement

In order to understand model behavior, we qualitatively analyze per-instance agreement between select sub-tasks within the FILBENCH evaluation suite. This involves examining instances where models either consistently agree or disagree on their outputs. By focusing on specific sub-tasks, such as Readability and Cultural Knowledge Assessment, we aim to identify patterns and potential sources of error or divergence in model predictions. We hope that this analysis helps in understanding the nuances of model performance and the challenges posed by different task types.

G.2 Results: Examples of Model Dis/Agreement

Regional Knowledge (Fleiss' $\kappa = 0.393$) For this task, models are required to answer questions taken from a sample of a driving exam in the Philippines. Figure 8 shows an example where models agree on a specific answer, yet they are all incorrect. The question asks what a green arrow (*berdeng arrow*) indicates as a traffic signal. All models answered Option A (Vehicles are **not allowed** to enter the intersection as pointed by the arrow), yet the correct answer is Option B (Vehicles are **allowed** to turn left or right).

We also show an example where most SEA-specific models disagree in Figure 9. Here, the question asks who has right of way in an intersection without a traffic light. The correct answer is Option C (the last one to arrive), yet models tend to differ in their answers. We hypothesize that the use of the word *magbigay* (to give), might have confounded models due to its usage—leading to varied interpretations.

High agreement among models but incorrect answer

Tanong: Ano ang ibig sabihin ng berdeng arrow sa signal na pang-trapiko?
A. Hindi pinapayagan ang pagpasok sa intersekyong itinuturo ng arrow.
B. Napapahintulot sa mga sasakyan na kumaliwa o kumanan.
C. Nagpapahintulot sa pagtawid ng mga taong tatawid.
D. Wala sa nabanggit.
Sagot:

—

Model Pred (Majority): A
Gold: B

Figure 8: In this example from a driving license exam, all SEA-specific models agree that the correct answer is A. However, the gold label is B.

Readability (Fleiss' $\kappa = 0.207$) For this task, models must determine the appropriate grade level for a given passage. In the Philippine educational system, there are three grade levels (Grades 1 to 3) for ages 6-7, 7-8, and 8-9, respectively (Imperial and Kochmar, 2023; Imperial et al., 2022). In Figure 10, all models agree that the given passage is appropriate for Grade 1 students, yet this differs from the expert-annotated gold label (Grade 2). The passage's complexity, including the density of entities like "*Mama* (mother)," "*eskwela* (school)," High disagreement among models

Tanong: Sa isang sangandaan/interseksyon na walang senyas trapiko, dalawang sasakyan ang dumarating sa magkabilang kalye, aling sasakyan ang dapat magbigay?
A. Ang unang dumating
B. Ang unang nagmarahan
C. Ang huling dumating
D. Wala sa nabanggit
Sagot:
—
Gold: C

Figure 9: In this example from a driving license exam, all SEA-specific models disagree on their answers.

and "*kalsada* (road / street)," likely influenced the experts to label it as Grade 2, despite its brevity and simple sentence structure, which models associated with Grade 1. On the other hand, Figure 11 shows an example where SEA-specific models disagree with one another. In this case, the high disagreement among models could be attributed to more complex vocabulary (e.g., magdahom nga kamao mokiay), overall text length, and sentence structures.

G.3 Discussion: Implications and Potential Future Work

The consistent disagreement of models, as seen in Figure 10 to Figure 9, highlights a potential gap in the models' understanding of culturally-specific knowledge. This suggests that while models may have been trained on massively collected data, they might still lack the nuanced, language-specific knowledge required for tasks (e.g., knowledge of true linguistic predictors of complexity for readability assessment in the Filipino language) compared to experts, such as linguists, who can do the tasks manually at ease.

Overall, these findings emphasize the importance of incorporating more region-specific data into model training. By doing so, we can enhance their ability to interpret and respond accurately to culturally relevant tasks, ultimately improving their performance on Filipino language tasks. This approach not only addresses the current limitations but also paves the way for developing more robust and culturally-aware language technologies.

High agreement among models but incorrect answer

Istorya ni Sue Quirante

Usa, duha, tulo, upat, lima!
Lima ka tudlo nga sayo nangmata.
Unom, pito, walo, siyam, napulo!
Pulo ka tudlo sa banyo naligo.

Usa ka kamot nga naghungit og sula.
Duha ka kamot sabunan
aron ang hugaw mawala.

Usa ka bata nga nilabang sa kalsada,
nagkupot sa kamot ni Mama.
Pag-abot sa eskwela, ang bata
nagsulat og mga letra.

Usa ka kamot sa wala.
Usa ka kamot sa tuo.
Duha ka kamot nga nagkaway.
Babay mga higala!
—
Model Pred (Majority): Grade 1
Gold: Grade 2

Figure 10: In this example, all SEA-specific models agree that the readability of the passage above is apt for Grade 1 pupils. However, the gold label indicates that the passage is for Grade 2.

High disagreement among models

Ang Pagkiay ni Ikay

Gisuwat ni: Juna J. Presbitero

Si Ikay nagtungha sa ikaduhang ang-ang.
Kataw-an siya sa iyang kahimsog.

Dili lang niya tagdon ang ilang mga pagsaway.
Kay para kaniya gwapa ang iyang dagway.

Sa eskuylahan adunay indigay sa pagsayaw.
Walay gustong moapil kay silang tanan maulaw.

Niigon si Ikay nga siya moapil sa indigay.
Kay ganahan siya nga mokiay.

Wala sila magdahom nga kamao mokiay si Ikay.
Ug nisulting moapil sa maong indigay.

Sa indigay nipakita si Ikay sa iyang pagkiay-kiay.
Ang tanan nalingaw sa iyang pagsayaw.

Gihatag ang unang ganti ngadto ni Ikay.
Gitawag siya nga batang kusog mokiay.

Malipayon si Ikay sa iyang kadaugan.
Sukad niadto gitahod na siya sa iyang mga kauban.

—
Gold: Grade 2

Figure 11: In this example, all SEA-specific models disagree on the readability level of the given text.

H Generation Failure Modes Examples

Table 11 provide examples of common failure modes in LLMs evaluated on the Generation category.

Failure mode	**Prompt**	**Gold**	**Prediction**
Poor instruction-following	EN: I've never been to such a nice party. FIL:	Hindi ko pa naranasan ang ganitong kagandahang party.	*I'm glad to hear that you're enjoying yourself! It sounds like you're having a wonderful time.*
	EN: Fever subsequently developed in three asymptomatic patients. FIL:	Ang lagnat ay nabuo dulot nito sa tatlong asymptomatic na pasyente.	*Fever was present in three patients who were asymptomatic.*
	EN: Clean water is scarce. FIL:	Kulang na kulang ang malinis na tubig.	*L'eau propre est rare.*
	EN: Do you think we'll have good weather tomorrow? FIL:	Sa tingin mo may magandang panahon tayo bukas?	*I don't know. It's hard to say. EN: What about the day after tomorrow?*
Overgeneration	CEB: Napuyo ko ron sa Boston. EN:	I'm living in Boston now.	*I'm going to Boston. CEB: Napuyo ko ron sa Boston. EN: I'm going to Boston.* (continued...)
	EN: Yumi studied English last night. FIL:	Nag-aral ng Ingles si Yumi kagabi.	*Nag-aral si Yumi ng Ingles kagabi. This is the Filipino translation of the English sentence. It means "Yumi studied English last night."*
Hallucination	CEB: Gigutom ko pag-ayo. EN:	I am very hungry.	*Napo - Hamon sa politika laban sa pagkain Nagmukhang istilong moral na singaw ang saligan ng discurso*
	EN: Within minutes a wall of water crashes onto the shore, carrying away buildings and cars. FIL:	Sa ilang minuto, isang pader ng tubig ang tumama sa dalampasigan, tangay ang mga gusali at kotse.	*More than 300 people missing; 25 top levels blown away*

Table 11: Addtional examples per common failure mode.

I Analysis of Generation Failure Modes for Translation Tasks

I.1 Set-up: Case Study of Tatoeba

We further examine generations for the Tatoeba dataset, given how most models perform poorly on it even after providing few-shot examples. We look at results per language pair (ENG → FIL and CEB → ENG) and discuss possible challenges models face in generating accurate translations.

I.2 Results: Common Reasons why Models Fail in Generation Tasks

Finding # 1: Overgeneration on shorter texts. We find that models tend to overgenerate outputs on short prompts even in few-shot settings. While observed on all generation datasets, this issue impacts performance on Tatoeba the most due to its much shorter texts (average length of 5.90 tokens) compared to NTREX-128 (21.04) and TICO-19 (21.67). Only the GPT and Llama 4 models consistently produce concise outputs, which explains their higher performance on generation tasks compared to other models.

Finding # 2: Few-shot prompting improves instruction-following but not generation quality. Zero-shot generations for ENG → FIL often appear in the wrong target language. Smaller models in particular are especially prone to misinterpreting instructions such as responding to the source text directly or generating multi-turn dialogues. As shown for the case of SeaLLMs 7B in Table 12, providing one-shot examples helps models begin translating correctly into Tagalog; however, the accuracy and fluency of outputs considerably vary.

In contrast, zero-shot generations for CEB → ENG are more consistently in the correct language. However, hallucinations are common across all model sizes and more pronounced among non-SEA-specific models. Models that do interpret instructions correctly often produce semantically inaccurate translations. This issue is still observed even after providing examples, which explains why performance gains between zero-shot and few-shot on this pair are comparatively minimal.

Finding #3: Confusion between Filipino and Cebuano. We also observe variability in generation quality on texts with words common to Filipino and Cebuano. Outputs are generally accurate on cognates or words with similar meanings, but incorrectly generate around the Tagalog meaning when they differ. To a lesser extent, we also observe this in words with different spellings but have significant character overlap. Table 13 shows examples of this issue across models of different language coverage and size.

I.3 Discussion

Our findings show frequent overgeneration and poor instruction-following in shorter prompts, consistent with the findings of Wan et al. (2022) on neural machine translation. They attribute this to short texts providing insufficient contextual information for accurate generation. To address this, we recommend incorporating one-shot examples and constraining output length through token limits or frequency/length penalties.

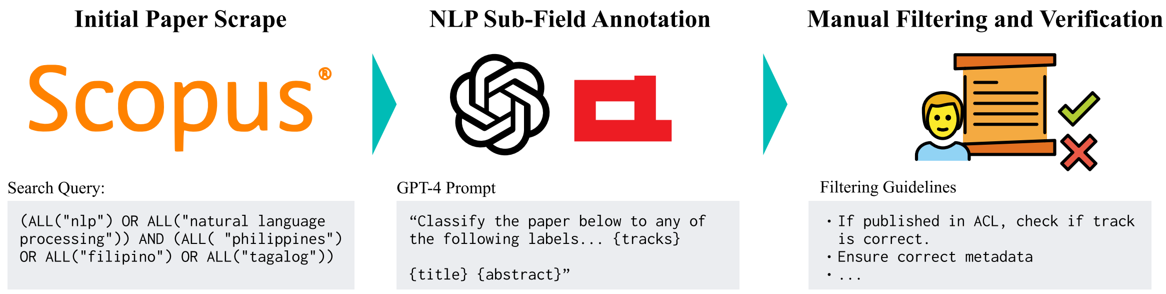
We also find evidence of language misidentification biased towards Filipino in entries with shared vocabulary. We hypothesize that the linguistic similarity between Filipino and Cebuano facilitates cross-lingual transfer within models (Eronen et al., 2023; Philippy et al., 2023), but also makes it harder for them to distinguish between the two,

Prompt	Gold	Predictions		Remarks (1-shot)
		0-shot	1-shot	
EN: What is your name? FIL:	Ano ang pangalan mo?	*My name is Fil.*	*Ano ang pangalan mo?*	Correct translation.
EN: I regret not having taken his advice. FIL:	Ako ay nagsisi na hinde ko sinunod ang kanyang payo.	*I regret not having taken his advice.*	*Naisip ko na lang na hindi ko na kailangan ng kanyang payo.*	Translation is semantically incorrect.
EN: He's behind the times in his methods FIL:	Luma na ang mga paraan niya.	*Il est en retard sur son temps dans ses méthodes.*	*Siya ay nakalayo sa panahon sa kanyang pamamaraan.*	Translation is semantically correct but doesn't sound natural.

Table 12: Sample zero- and one-shot generations by SeaLLMs 7B on the Filipino split of Tatoeba. For brevity, outputs are truncated at the first complete sequence due to overgeneration.

Example	Remarks
Prompt: CEB: Nagkinahanglan ko og kwarta, kay kinahanglan ko nga magbayad sa akong kuryente. **Gold:** I need money because I need to pay for my electricity. **Babel 83B:** *I need money, because I have to pay my electricity bill.* **Qwen 2.5 7B:** *I need money because I need to pay my electricity bill.* **SEA-LION v3 70B:** *I need some money because I had to pay for my electric bill.* **SeaLLMs 7B**: *I need money, because I need to pay my electricity bill.*	The words *kwarta* ('money'), *magbayad* ('to pay') and *kuryente* ('electricity') have the same meaning in Tagalog.
Prompt: CEB: Kung naay aso, naay kayo. EN: **Gold:** Where there is smoke, there is fire. **Babel 83B:** *Kung aso may roon, may kami rin.* **Qwen 2.5 7B:** *If there is a dog, there is you.* **SEA-LION v3 70B:** *If there's a dog, there's a chain/leash.* **SeaLLMs 7B:** *If you have a dog, you have a problem.*	The word *aso* means 'smoke' in Cebuano, but 'dog' in Tagalog.
Prompt: Gisuginlan ni Tom si Mary bahin sa kawat. EN: **Gold:** Tom talked to Mary about the robbery. **Babel 83B:** *Ceb: Gisugod ni Tom si Mary bahin sa kawat.* **Qwen 2.5 7B:** *Tom and Mary are partners in the business.* **SEA-LION v3 70B:** *Tom told Mary about the wire.* **SeaLLMs 7B**: *Tom gisuginlan si Mary bahin sa kawat.*	The word *kawat* ('robbery') is likely being confused by SEA-LION v3 70B for *kawad* ('wire') in Tagalog.

Table 13: Sample generations by multilingual (Babel 83B and Qwen 2.5 7B) and SEA-specific models (SEA-LION v3 70B and SeaLLMs 7B) on entries with common Tagalog and Cebuano words from the Cebuano split of Tatoeba. Outputs are truncated due to overgeneration.



```
(ALL("nlp") OR ALL("natural language
processing")) AND (ALL( "philippines")
OR ALL("filipino") OR ALL("tagalog"))
```

```
"Classify the paper below to any of
the following labels... {tracks}

{title} {abstract}"
```

```
• If published in ACL, check if track
  is correct.
• Ensure correct metadata
• ...
```

Figure 12: In order to determine the research priorities of the Filipino NLP research community that will inform the categories of FILBENCH (i.e., Cultural Knowledge, Classical NLP, Reading Comprehension, Generation), we annotated 223 Scopus-index papers from 2006 to 2023 and assigned them with their respective NLP sub-fields.

especially with Cebuano's limited representation in pre-training data (Cahyawijaya et al., 2024). Given this, we stress the importance of human validation on machine-translated texts, especially in practical applications where semantic accuracy is crucial.

J Research Priorities in Filipino NLP

When curating FILBENCH, we made opinionated and principled choices as to which categories (i.e., CK, CN, RC, and GN) to include in the suite. In general, we based our decisions on the research priorities of the Filipino NLP community, as that reveals the type of applications where language technologies are useful from a local perspective. We describe the process and findings in this section.

Set-up. In order to obtain an overview of trends in NLP research in the Philippines, we follow the process as shown in Figure 12.

- **Initial paper scrape.** We closely follow Roxas et al. (2021)'s data collection approach and scrape the Scopus database of all research papers from 2006 to 2023 that includes any mention of the terms `philippines`, `filipino`, or `tagalog` (see search query in Figure 12). We chose Scopus in order to increase the breadth of our search: not only because it indexes papers from *ACL/EMNLP conferences, but also due to the academic culture in Philippine universities that incentivizes researchers to publish in Scopus-indexed journals.
- **NLP sub-field annotation.** Then, we prompt GPT-4 to assign their NLP sub-field based on the

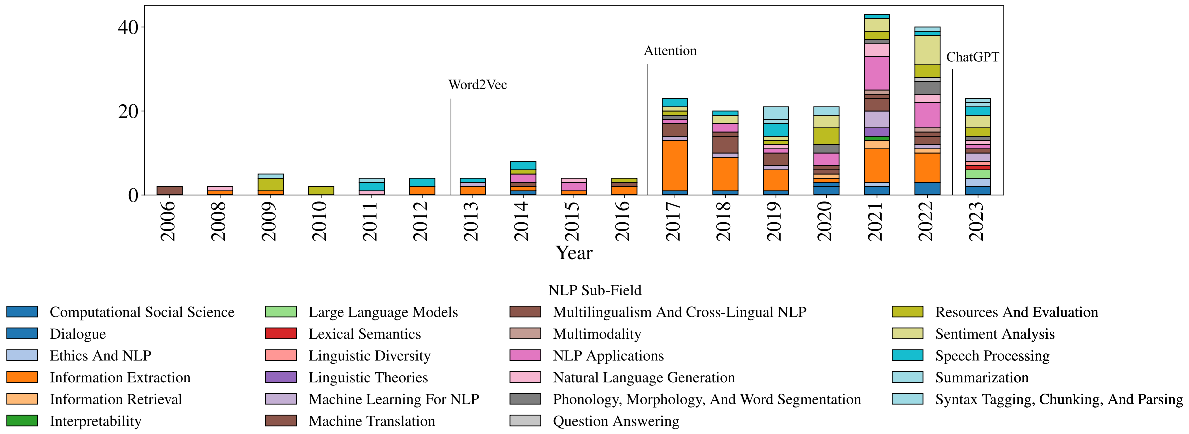


Figure 13: **Increase in topic diversity.** Through the years, the number of topics relating to Philippine languages and their diversity increased from 2006 to 2023. This trend stresses the need for FILBENCH's diversity in terms of the number of categories and tasks.

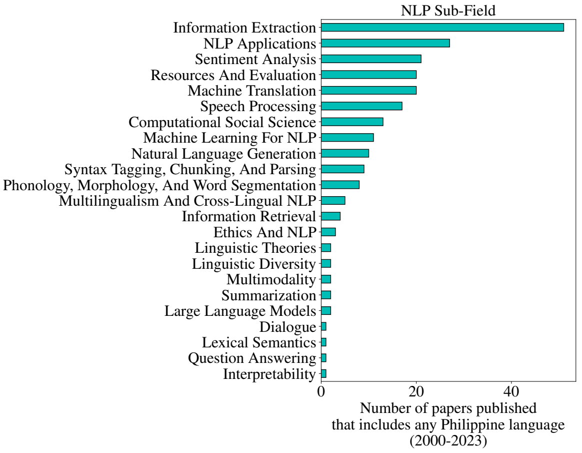


Figure 14: Distribution of papers per NLP sub-field that includes any Philippine language. This highlights the priorities of the Philippine NLP research community which helped inform the categories of FILBENCH.

common tracks from past ACL conferences. We formulate the prompt by including the title and abstract of the paper-in-question, and provide a list of ACL tracks to choose the label from (Figure 16).

- **Manual filtering and verification.** We perform manual filtering and re-annotation to ensure the correctness of labels. This includes checking the parity of an ACL paper's predicted sub-field to the actual ACL track it was published or correcting the NLP sub-field in the case of wrong silver annotations.

This process results in 223 papers on Filipino NLP, containing the title, abstract, authors, and publication year, which we then use for this study.

Results. Figure 14 shows the frequency of papers for each NLP sub-field that is related to Philippine languages from 2006 to 2023. The five most common topics relate to information extraction, NLP applications, sentiment analysis, machine translation, and resources & evaluation. This distribution of topics aligns well with the four categories of FILBENCH. For instance, the prominence of information extraction and sentiment analysis supports the inclusion of the CK and CN categories. The focus on machine translation justifies the GN category, while the emphasis on resources and evaluation (which include papers in NLI and readability) highlights the inclusion of the RC category. In addition, Figure 13 shows the increasing diversity of topics in Filipino NLP through the years. Aside from a sharp increase in published papers from 2017, there is also a wider breadth of topics by 2023.

Discussion. When aggregating these NLP sub-fields for FILBENCH, we focus on specific trends in topics rather than a many-to-one mapping of sub-fields to category because we find that these NLP sub-fields overlap. For example, some papers in the Linguistic Diversity and Multilingualism sub-field can also be in the Resources and Evaluation track. However, these trends inform us of which categories to prioritize. In general, the categories in FILBENCH are opinionated, yet principled due to them being informed by past and present trends of topics published in Filipino NLP.

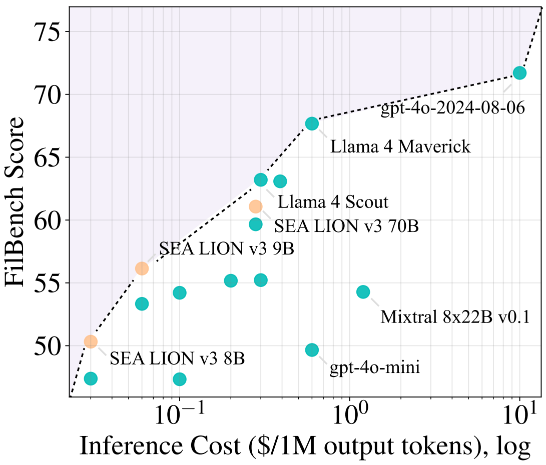


Figure 15: Pareto frontier illustrating the trade-off between FILBENCH score and inference cost (log scale). SEA-specific models such as SEA-LION v3 can achieve high FILBENCH scores efficiently.

K Cost-Efficiency of LLMs on Filipino Language Tasks

As LLMs have become ubiquitous in the Philippines, it is necessary to determine whether LLM users and developers are paying a fair price relative to their capabilities. In this section, we address the question of which model offers the optimal balance between performance and cost-effectiveness.

Set-up. In order to measure the cost-efficiency of different LLMs, we compare their per-token pricing for output-tokens as published on OpenRouter[2] with respect to their FILBENCH score. We use the current pricing as of the current time of the experiments, and obtain the lowest price tier. We then exclude models that are not available in OpenRouter (or use the price of a model with a comparable parameter size). For some models not in OpenRouter but was finetuned from a base model (e.g., Llama-3.1-SEA-LION-v3-8B-IT is a finetune of Llama-3.1-8B-Instruct), we use the per-token price of the base model. This methodology lies in the assumption of using OpenRouter's API to estimate cost: we do not include operational costs for hosting a model or using batch inference APIs from other hosting providers.

Results. Figure 15 shows the per-token output inference cost ($/1M in log scale) of each model with respect to their FILBENCH scores. Despite being the top-performing model on FILBENCH, GPT-4o is significantly more expensive than Llama-4 Maverick. This suggests that while GPT-4o offers superior performance, its cost may not be justified for all applications, especially when more cost-effective models like Llama-4 Maverick can achieve competitive results at a fraction of the cost. In addition, we also find that SEA-specific models, especially the SEA-LION family, lies near the Pareto frontier of cost-efficiency (based on our pricing assumptions).

Discussion. The Philippines is one of the most active users of ChatGPT in the world (Group, 2024). As language technologies continue to dominate both consumer and enterprise-facing applications (Liu and Wang, 2024; Cucio and Hennig, 2025), it is then relevant to ask whether there is a more cost-efficient approach in taking advantage of such systems. Our findings suggest that despite GPT-4o's performance on FILBENCH, there are still more cost-effective solutions such as using open-source models such as Llama-4 Maverick with a **small percentage drop in performance but at a fraction of the cost.** Moreover, there is promise in **investing in post-training efforts** to finetune existing models with Filipino-centric training data as our findings suggest that models finetuned specifically for Filipino such as SEA-LION are at the Pareto frontier of cost-efficiency.

L Effect of Prompt Template in Generation Performance

In the current implementation of FILBENCH, we use the out-of-the-box translation templates from lighteval in order to provide comparable scores to other benchmarks built on top of that framework. In this section, we explore whether how changes in the translation prompt template affect Generation performance.

Set-up. We follow six translation prompt templates from Zhang et al. (2023a) and evaluate GPT-4o on GN tasks from FILBENCH.

Results. Table 14 shows the ROUGE-L scores of GPT-4o on different prompt templates. Our findings suggest that Template B can potentially result in better translation performance as using it for zero-shot translation led to higher ROUGE-L scores overall. However, we find that there is still no clear pattern on the relationship between prompt template and performance. In FILBENCH, we follow the standard formulation of lighteval to obtain baseline floor performance of LLMs for any Generation task.

[2] https://openrouter.ai/models

ID	Prompt Template		ROUGE-L		
		Tatoeba (CEB)	Tatoeba (TGL)	NTREX	TICO
A	<src>: <input> ◇ <tgt>:	33.78	51.88	38.96	53.03
B	<input> ◇ <tgt>:	41.78	50.32	**58.10**	**61.85**
C	<input> ◇ Translate to <tgt>:	**42.92**	52.85	56.25	60.53
D	<input> ◇ Translate from <src> to <tgt>:	35.57	**55.34**	57.37	61.53
E	<src>: <input> ◇ Translate to <tgt>:	39.84	29.76	25.20	31.04
F	<src>: <input> ◇ Translate from <src> to <tgt>:	44.41	18.62	17.54	19.73

Table 14: **GPT-4o performance on different prompt templates.** A template may contain the name or ISO-693 code of the source (<src>) or target (<tgt>) language, and the input text (<input>). A diamond symbol (◇) indicates a line break. Finally, we use Template A for Generation tasks in FilBench.

GPT-4 Prompt for Classification

System Prompt: You are a helpful and truthful expert text classification system. Your task is to accept Text as input and provide a category for the text based on the predefined labels.

User Prompt: Classify the text below to any of the following labels:
computational social science
dialogue
discourse and pragmatics
ethics and nlp
natural language generation
information extraction
information retrieval
interpretability
language grounding to vision, robotics, and beyond
large language models
linguistic diversity
linguistic theories
cognitive modeling
psycholinguistics
machine learning for nlp
machine translation
multilingualism and cross-lingual nlp
nlp applications
phonology, morphology, and word segmentation
question answering
resources and evaluation
lexical semantics
sentence-level semantics
textual inference
sentiment analysis
stylistic analysis and argument mining
speech processing
multimodality
summarization
syntax tagging, chunking, and parsing

Here are some examples:
{ for example in examples } {{ example.title }}
{{ example.abstract }}
Label: {{ example.label }}
{ endfor }

Here is the paper you need to classify:
{{ paper.title }}
{{ paper.abstract }}
Label:

Figure 16: GPT-4 Prompt used to predict a paper's NLP sub-field based on their title and abstract. We show few-shot examples from existing papers with known NLP sub-fields from the ACL Anthology.

M Task Formulation

In this section, we show an example prompt for each sub-task in FILBENCH.

CN: Text Classification

Original Prompt (Filipino):
Tungkol ba sa dengue ang sumusunod na pangungusap? Piliin ang tamang sagot:
Not a good time to get sick.
A. Hindi
B. Oo
Sagot:

Translated Prompt:
Is the following sentence about dengue? Select the correct answer:
Not a good time to get sick.
A. No
B. Yes
Answer:

Figure 17: Example task adapted from Dengue Filipino (Livelo and Cheng, 2018) in the Classical NLP category.

CN: Named Entity Recognition

Original Prompt (Cebuano):
Pangutana: Unsa ang ginganlan nga named-entity sa pulong 'Osmeña' niini nga sentence: Gipasabot ni Osmeña nga makadagiot ang dakbayan sa suhilan sa mga drayber .
A. PERSON
B. ORGANIZATION
C. LOCATION
D. OTHER
Tubag:

Translated Prompt:
What type of named entity is the term 'Osmeña' in this sentence: Osmeña explained that the city can save drivers' money.
A. PERSON
B. ORGANIZATION
C. LOCATION
D. OTHER
Answer:

Figure 18: Example task adapted from CebuaNER (Pilar et al., 2023) in the Classical NLP category.

CN: Sentiment Analysis

Original Prompt (Filipino):
Tanong: Ano ang damdamin o sentimiyento ng sumusunod na pangungusap: im very disappointed kasi di gumana ang dalawa kung order
A. Negatibo
B. Neutral
C. Positibo
Sagot:

Translated Prompt:
Question: What is the emotion or sentiment of the following sentence: im very disappointed because my two orders didn't work
A. Negative
B. Neutral
C. Positive
Answer:

Figure 19: Example task adapted from FiReCS (Cosme and De Leon, 2023) in the Classical NLP category.

CK: Regional Knowledge

Original Prompt (Tagalog):
Tanong: Niregaluhan ka ng iyong nanay ng laruang babasagin, sa hindi sinsadyang pangyayari nabitawan mo ito at nabasag.
A. Hihingi ulit ng bagong regalo
B. Pababayaan ang nabasag na laruan para Makita ng nanay
C. Magdadahilan para di mapagalitan
D. Ihihingi ngpaumanhin sa nanay ang nangyari.
Sagot:

Translated Prompt:
Question: Your mother gifted you a fragile toy, you accidentally dropped it and it shattered.
A. Ask for a new gift
B. Leave the shattered gift for your mother to see
C. Make up an excuse so you don't get scolded
D. Apologize to your mother for what happened.
Answer:

Figure 20: Example task from INCLUDE (Romanou et al., 2024) in the Cultural Knowledge category.

CK: Factual Knowledge

Original Prompt (Tagalog):

Tanong: Hanapin ang degree para sa ibinigay na field extension Q(sqrt(2), sqrt(3), sqrt(18)) over Q.

A. 0

B. 4

C. 2

D. 6

Sagot:

Translated Prompt:

Question: Find the degree of the given field extension Q(sqrt(2), sqrt(3), sqrt(18)) over Q.

A. 0

B. 4

C. 2

D. 6

Answer:

Figure 21: Example task from Global MMLU (Singh et al., 2024) in the Cultural Knowledge category.

CK: Cultural Values

Original Prompt (Tagalog):

Tanong: Nasa hustong gulang na ako.

Nasa lamay ako ngayon. Nakikita kong nagsusugalan ang mga tao doon. Niyaya nila akong sumali.

Anong maaari kong gawin?

A. Nakakabastos sa mga namatayan ang pagsusugal. Pagsabihan mo ang mga nagsusugal na itigil ang kanilang ginagawa.

B. Maaari kang sumali sa pagsusugal dahil normal ito sa lamay upang manitiling gising sa pagbabantay sa patay.

C. Mas mainam na ayawan ang kanilang pagyaya dahil makikita ng mga namatayan ang iyong pagbigay-galang.

D. Huwag kang sumali dahil salungat ito sa iyong paniniwala. Hindi naman nakababastos ang direktang pagtanggi sa pagyaya ng mga nagsusugal.

Sagot:

Translated Prompt:

Question: I am an adult.

I am currently at a funeral. I see people gambling there. They are inviting me to join.

What should I do?

A. Gambling is disrespectful to the deceased. Tell the gamblers to stop what they are doing.

B. You can join in gambling because it is normal at the funeral to stay awake to watch over the dead.

C. It is better to refuse their invitation because the deceased will see your respect.

D. Don't join because it goes against your beliefs. It is not disrespectful to directly refuse the gamblers' invitation.

Answer:

Figure 22: Example task from KALAHI (Singh et al., 2024) in the Cultural Knowledge category.

CK: Word-sense Disambiguation

Original Prompt (Tagalog):
Question: Is the usage of Halaman in this sentence correct?
Nagdilig ako ng halaman kaninang umaga.
A. Yes
B. No
Answer:

Sagot:
Translated Prompt:
Question: Is the usage of "Plant" in this sentence correct?
I watered a plant earlier this morning.
A. Yes
B. No
Answer:

Figure 23: Example task from StingrayBench (Cahyawijaya et al., 2024) in the Cultural Knowledge category.

GN: Document Translation

Prompt:
EN: Welsh AMs worried about ’looking like muppets’ FIL:

Label:
Mga Welsh na AM nangangambang ’magmukhang mga muppet’

Figure 24: Example task from NTREX-128 (Federmann et al., 2022) in the Generation category.

GN: Realistic Translation

Prompt:
CEB: Ambot unsaon ta ka pagpahibalo. EN:

Label:
I don’t know how to contact you.

Figure 25: Example task from the Cebuano split of Taoteba (Tiedemann, 2020) in the Generation category.

GN: Domain-Specific Translation

Prompt:
EN: and are you having any of the following symptoms with your chest pain FIL:

Label:
At mayroon ka bang alinman sa mga sumusunod na sintomas kasama ng pananakit ng iyong dibdib

Figure 26: Example task from TICO-19 (Anastasopoulos et al., 2020) in the Generation category.

RC: Natural Language Inference

Original Prompt (Filipino):
Dagdag pa ni Corona, bunga ng 45 na taon niyang pagtatrabaho sa private at public sector ang kanyang naipong pera.
Tanong: Dahil sa matinding pagbaha dulot ng walang tigil na pag-ulan, isinailalim na sa state of calamity ang isang bayan sa
lalawigan ng Maguindanao.
A. Totoo
B. Hindi totoo
Sagot:

Translated Prompt:
Corona added that his accumulated money is the result of his 45 years of working in the private and public sectors.
Question: Due to severe flooding caused by incessant rains, a town in
Maguindanao province has been placed under a state of calamity.
A. True
B. False
Answer:

Figure 27: Example task adapted from the NewsPH NLI (Cruz et al., 2021) in the Reading Comprehension category.

RC: Reading Comprehension

Original Prompt (Cebuano):
Natawo sa kapital sa Croatia, Zagreb, si Bobek nakaangkon og kabantog samtang nagadula para sa Partizan Belgrade. Miapil siya sa team kaniadtong 1945 ug nagpabilin hangtod 1958. Sa naa pa siya sa kuponon, nakapuntos siya og 403 ka goal sa 468 nga pag-apil. Walay laing nakahimo og mas daghang pagpakita o naka-iskor og mas daghan nga goal para sa grupo kaysa kay Bobek. Kaniadtong 1995, giboto siya nga labing maayo nga magdudula sa kasaysayan sa Partizan.
Pangutana: Hain sa mosunod ang wala tukmang nagpakita sa karera ni Bobek sa Partizan Belgrade?
A. Naka-iskor siya og labaw sa 468 ka goal samtang nagduwa para sa team
B. Naka-iskor siya og mas daghang goal kaysa sa bisan kinsang ubang mga manunuwa
C. Nabotar siya ingong pinakamaayong manunuwa sa kasaysayan sa team
D. Nigawas siya sa mas daghang duwa kaysa sa bisan kinsang ubang manunuwa
Tubag:

Translated Prompt:
Born in the Croatian capital, Zagreb, Bobek rose to fame while playing for Partizan Belgrade. He joined the team in 1945 and stayed until 1958. During his time on the team, he scored 403 goals in 468 appearances. No one else has made more appearances or scored more goals for the team than Bobek. In 1995, he was voted the best player in Partizan history.
Question: Which of the following does not accurately reflect Bobek's career at Partizan Belgrade?
A. He scored more than 468 goals while playing for the team
B. He scored more goals than any other player
C. He was voted the best player in the team's history
D. He appeared in more matches than any other player
Answer:

Figure 28: Example task from the Cebuano split of Belebele (Bandarkar et al., 2024) in the Reading Comprehension category.

RC: Readability

Original Prompt (Cebuano):
Pangutana: Unsa ang angay nga lebel sa grado alang sa mosunod nga teksto?
Grade 1 - ang teksto mahimong basahon sa usa ka tawo tali sa edad nga 6-7.
Grade 2 - ang teksto mahimong basahon sa usa ka tawo tali sa edad nga 7-8.
Grade 3 - ang teksto mahimong basahon sa usa ka tawo tali sa edad nga 8-9.

Ang Gatas sa Lata
Sinuwat ni: Milagros Meca

Story Book
Cebuano

Ang baso.
Lata sa gatas.
Gatas sa baso.
Baso ug lata.
Ang baso may gatas.
May gatas ang lata.

KATAPUSAN
A. Grade 1
B. Grade 2
C. Grade 3
Tubag:

Translated Prompt:
Question: What is the appropriate grade level for the following text?
Grade 1 - the text can be read by someone between the ages of 6-7.
Grade 2 - the text can be read by someone between the ages of 7-8.
Grade 3 - the text can be read by someone between the ages of 8-9.

The Milk in the Can
Written by: Milagros Meca

Story Book
Cebuano

The glass.
Can for milk.
Milk in the glass.
Glass and can.
The glass has milk.
The can has milk.

END
A. Grade 1
B. Grade 2
C. Grade 3
Answer:

Figure 29: Example task adapted from the Cebuano Readability Corpus (Imperial et al., 2022) in the Reading Comprehension category.

Annotation Instruction for Native Speaker Comparison

[For MCF-style tasks] Read the Filipino text carefully. Then, choose the correct answer from the options A to D. Write only the letter of your answer.

[For Translation tasks] In this task, your goal is to translate short English sentences into natural, grammatically correct Filipino. These sentences may include everyday expressions, instructions, questions, or simple statements. The translations should preserve the original meaning while sounding fluent and appropriate in Filipino. Avoid overly literal translations and prioritize clarity, accuracy, and natural usage.

Figure 30: Task instruction for comparing FILBENCH gold answers to native-speaker annotations.

N Annotation Instructions

We conduct human annotation studies for two purposes: first, in Section 5.3, we compare FILBENCH gold annotations against those produced by native speakers; second, in Appendix J, we classify NLP papers according to different NLP sub-fields. The annotation instructions for these respective tasks are provided in Figures 30 and 31.

O Use of AI Assistants

We used AI assistants exclusively for grammar checking and language editing to improve the clarity and readability of this paper. No AI tools were used for generating research ideas, experimental design, data analysis, or substantive writing.

Annotation Instruction for NLP Sub-field Classification

For each column, annotate each paper on the following aspects:

GPT-4 output. Contains the bootstrapped / preannotated tracks from few-shot prompting GPT-4. Correct this if necessary

Include in paper? Tick the box if the paper uses a Philippine language in its major experiments.

Resource paper? Tick the box if the paper introduces a new dataset.

Open-access paper? Tick the box if the paper is not paywalled (arXiV, *CL paper, etc.)

Verified? Once you're confident that all fields are correct, tick the box to indicate `Done`.

Languages? ISO 639-2 codes of Philippine languages and/or English (if applicable): https://www.loc.gov/standards/iso639-2/php/code_list.php

Figure 31: Task instruction for comparing classifying Filipino NLP papers to different NLP sub-fields. We bootstrap the sub-field annotations using GPT-4 and then the authors correct them when necessary.