Komodo: A Linguistic Expedition into Indonesia’s Regional Languages

Anonymous ACL submission

Abstract

The recent breakthroughs in Large Language Models (LLMs) have mostly focused on languages with sufficient resources, such as English. However, there remains a significant gap for languages that lack sufficient linguistic resources. Our work introduces Komodo-7B, a 7-billion-parameter LLM designed to address this gap by operating across Indonesian, English, and 11 regional languages in Indonesia. Komodo-7B consists of Komodo-7B-Base and Komodo-7B-Instruct. Komodo-7B-Instruct achieves state-of-the-art performance in various tasks and languages, outperforming benchmarks set by models such as OpenAI’s GPT-3.5, Mixtral-8x7B-Instruct-v0.1, Gemma-7B-it and many more. This model demonstrates superior performance in both language-specific and overall assessments, highlighting its capability to excel in linguistic diversity. Our commitment aims to bridge the gap for languages with limited resources. Additionally, Komodo-7B-Instruct’s cross-language understanding addresses educational disparities in Indonesia by offering direct translations from English to 11 regional languages. Komodo-7B represents a crucial step towards inclusivity and effectiveness in language models, catering to the linguistic needs of diverse communities.

1 Introduction

Since the introduction of transformers (Vaswani et al., 2017) based language model, there is a significant gap when it comes to models tailored to specific regional languages. While models like GPT-3.5 (Brown et al., 2020) and Llama-2 (Touvron et al., 2023) excel in various tasks, their performance is primarily benchmarked in English. However, these models tend to underperform when dealing with languages other than english, on the other hand, there are multilingual models like Aya-101 (Üstün et al., 2024), Bactrian-X (Li et al., 2023a), Qwen-1.5 (Bai et al., 2024), and Mixtral (Jiang et al., 2024), that perform well in tasks involving multiple languages. Yet, when it comes to individual languages or small regional languages with limited available data, these models lack specialized expertise. Also significant advancements have been achieved in creating datasets for pre-training and instruction tuning, such as Alpaca (Taori et al., 2023), UltraChat (Ding et al., 2023), Dolly (Conover et al., 2023), OpenAssistant (Köpf et al., 2023), and LMSYS-Chat (Zheng et al., 2023). Additionally, there have been efforts to establish evaluation benchmarks like AlpacaEval (Li et al., 2023b) and MT-Bench. However, it’s important to note that the majority of these developments have primarily focused on the English language.

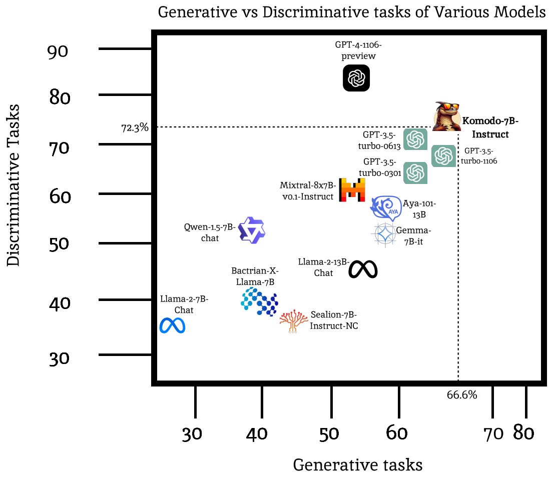


Figure 1: A plot illustrating Komodo-7B-Instruct adeptness in balancing generative and discriminative tasks, showcasing strong performance across diverse language challenges.

Nevertheless, the way data is represented, the efficiency of tokenizers, and the overall performance in tasks related to Indonesian and regional languages lag significantly behind those of English. Even in closed-source models like GPT-3.5, GPT-4 (OpenAI, 2023), and similar ones, the perfor-

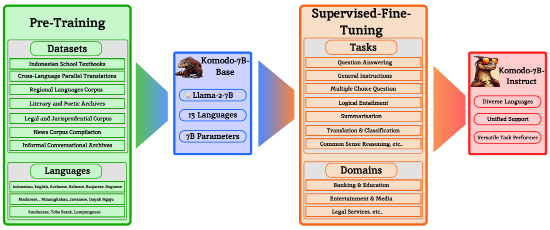

Pre-Training
Datasets
Indonesian School Textbooks
Cross-Language Parallel Translations
Regional Languages Corpus
Literary and Poetic Archives
Legal and Jurisprudential Corpus
News Corpus Compilation
Informal Conversational Archives
Languages
Indonesian, English, Acehnese, Balinese, Banjarese, Buginese
Madurese, , Minangkabau, Javanese, Dayak Ngaju
Sundanese, Toba Batak, Lampungnese
Komodo-7B-Base
Llama-2-7B
13 Languages
7B Parameters
Supervised-Fine-Tuning
Tasks
Question-Answering
General Instructions
Multiple Choice Question
Logical Entailment
Summarisation
Translation & Classification
Common Sense Reasoning, etc..
Domains
Banking & Education
Entertainment & Media
Legal Services. etc..
Komodo-7B-Instruct
Diverse Languages
Unified Support
Versatile Task Performer

Figure 2: The Evolution of Komodo-7B-Instruct Language Model. The diagram illustrates the transformation from the Komodo-7B-Base model, initially trained on diverse datasets encompassing various languages, to the refined Komodo-7B-Instruct model through targeted Supervised Fine-Tuning (SFT) on specific tasks and domains. The journey involves strategic pretraining on comprehensive datasets, followed by fine-tuning for enhanced performance and adaptability across a spectrum of language-related challenges.

mance in Indonesian languages falls short when compared to their performance in English. This performance gap highlights the need for focused attention and improvement in addressing the specific challenges posed by Indonesian and regional languages in the realm of language models. So, overall, currently, there is a notable absence of high performing LLMs specifically designed for Indonesia, trained on Indonesian data and evaluated against benchmarks for Indonesia's regional languages. In response to this gap, we present Komodo-7B-Instruct, a language model with 7 billion parameters, tailored for 13 languages including Indonesian, English and 11 regional languages.

In our comparison with several big models, both those available to everyone and those with restricted access like ChatGPT, Llama-2, and Mixtral, we have achieved top-notch performance in a few tasks. We've surpassed the capabilities of various multilingual models like Cohere's Aya-101, MBZUAI's Bactrian-X-llama-7B, Qwen-1.5, and Mixtral-8x7B-Instruct-v0.1, across several tasks. Notably, our performance also surpasses the existing Indonesian Large Language Model (LLM) SEA-LION (AISingapore, 2023) in almost every task by a significant margin. This demonstrates our model's effectiveness and superiority in handling diverse tasks and outperforming existing benchmarks. In addition, our model shows excellent cross-language understanding, making it a valuable tool to bridge the education gap in Indonesia. With the capability to directly translate from English to 11 regional languages, or even from one regional language to others, our model surpasses the limitations of Google Translate (Wu et al., 2016), which only supports Indonesian, Javanese, and Sundanese. Importantly, our model also serves to a broader range of regional languages, ensuring accessibility for people across various regions in Indonesia.

It's worth noting that Javanese and Sundanese are predominantly spoken on the Java island, where innovation and education quality are relatively higher compared to other regions in Indonesia. Our model's support for a diverse set of regional languages ensures that educational resources and information can be more widely disseminated, contributing to a more inclusive and equitable educational landscape throughout the country.

2 Dataset

The dataset employed in both the pre-training and fine-tuning phases of our language model was created not only from diverse open-source datasets but also from the manually collected data for regional languages of Indonesia. Our approach is informed by the noteworthy outcomes demonstrated by models such as Phi-1 (Gunasekar et al., 2023) and Phi-1.5 (Li et al., 2023c), indicating that achieving commendable results does not necessarily depends on vast amounts of data. Instead, a judicious selection of high-quality data has proven effective, even yielding State-of-the-Art performance under certain circumstances. As data preprocessing is a crucial step in scaling the language models, we took some preprocessing steps inspired by (Rae et al., 2022), as follows:

Repetition Removal: Excessive repetition of words or phrases can lead to uninformative content and is a common issue in language models. To address this, we identify and remove documents with a high proportion of repeated lines or paragraphs. Utilizing multiple approaches, we calculate the fraction of duplicate content for lines and paragraphs separately, considering different n-gram sizes. This ensures a comprehensive removal of redundancy, contributing to the refinement of our dataset.

Quality Filtering: Implementing measures to filter out low-quality or irrelevant data, contributing to a more refined dataset. To enhance data quality, we implement straightforward heuristic filters. We exclude documents with insufficient or excessive word counts, ensuring they fall within a specific range. By applying these filters, we aim to retain high-quality, coherent text relevant to language model training.

Deduplication (Mou et al., 2023): Duplicate text across various documents can introduce redundancy in the dataset. Identifying and removing duplicate entries to prevent redundancy and ensure the uniqueness of the dataset is an important step towards maintaining high-quality data.

These preprocessing steps are integral to our commitment to data quality, aiming to enhance the effectiveness and performance of our language model.

2.1 Pretraining & Supervised-Fine-Tuning Data

Moreover, our research extends to the incorporation of Indonesian textbooks spanning grades 1 through 12, consisting a number of subjects including arts, civics, mathematics, sports, religion, local cultures, and engineering, among others. This strategic integration serves the purpose of enhancing the model's general knowledge capabilities, covering a broad array of topics including science, daily activities, and more. We've also integrated colloquial data extracted from various sources such as movie subtitles, news, informal conversations, movie reviews, poems, and more.

Additionally, we made use of freely available datasets primarily collected in Indonesian and other regional languages, such as Javanese, Sundanese, Acehnese, and many more. We aim to ensure that our language model is well-versed not only in Indonesian but also in other regional languages. The primary objective is to improve our model's comprehension of regional languages, as currently, no language model understands different regional languages in Indonesia. This approach helps improve the model's overall language skills and adaptability to various cultural contexts.

The intention behind this comprehensive dataset combination is to imbue the language model with a better and deep understanding of the language, encouraging adeptness in handling diverse contextual cues and promoting a more inclusive comprehension of language and its applications.

In our dataset, we've taken inspiration from OpenHathi (SarvamAI, 2023) approach to include English datasets and alternate parallel data, aiming to enhance our model's understanding of code-mixed sentences. Alternate parallel, as inspired by OpenHathi, involves a unique approach to teaching cross-lingual understanding. Instead of following the traditional method of monolingual next-token prediction with translated Indonesian text, we employ a bilingual next-token prediction strategy.

This bilingual approach introduces alternate sentences in English and Indonesian. The significance lies in requiring the model to cross-lingually attend to information during next-token prediction. For instance, predicting an English token in the second sentence would necessitate attending to Indonesian tokens in the preceding sentence. We hypothesize that this approach increases alignment between English and Indonesian. Moreover, it naturally balances the exposure of the model to both languages during training, promoting a more robust understanding.

It's crucial to note that our use of alternate parallelism is not limited to English and Indonesian only; rather, it encompasses all combinations of English, Indonesian, and the 11 regional languages, including Acehnese, Balinese, Banjarese, Buginese, Dayak Ngaju, Javanese, Lampungnese, Madurese, Minangkabau, Sundanese, and Toba Batak. This inclusive approach ensures a comprehensive and diverse training set for our model, contributing to its proficiency in handling code-mixed sentences across multiple languages.

To facilitate translation between English and Indonesian whenever needed, we also leverage translation engines like Google Translate API. This additional step further enriches our dataset and supports the model's ability to comprehend and generate content in diverse linguistic scenarios.

The SFT dataset is derived from open-source data, covering a wide array of tasks. We also do manual label creation from unlabeled data , such as the creation of a Multiple-Choice Question task using the Indonesian dictionary (KBBI). The SFT dataset also incorporates responses from ChatGPT, adding nuanced and diverse perspectives to the labeled data. The meticulous curation process ensures a comprehensive and adaptable dataset, making it valuable for training models across various language-related tasks.

2.2 Benchmarking Datasets

IndoMMLU (Koto et al., 2023): The IndoMMLU dataset is designed to assess language proficiency by focusing across diverse subjects including STEM, humanities, social sciences, and more. Questions within the dataset prompt responses that cover a spectrum of topics, providing a thorough evaluation of language understanding and expression in Indonesian and regional languages.

Indotexbook Bilingual (ID-EN) Entailment: The Indotexbook Bilingual Entailment dataset involves paired sentences in Indonesian and English, intended for entailment analysis. It tasks the model with determining the relationship between two sentences, one in Indonesian and the other in English. The dataset facilitates evaluation for cross-lingual entailment understanding. This is manually created data (held-out set) based on the Indonesian textbooks data that we utilized during pretraining and SFT.

X-Copa (ID) (Ponti et al., 2020): X-Copa, specific to Indonesian (ID), presents contextualized sentence pairs to evaluate the model's commonsense reasoning abilities. The dataset includes scenarios where the model must infer the consequence or outcome of a given situation. For instance, inferring the result of a person turning on a tap. This dataset aids in assessing the model's ability to draw logical inferences.

NusaXSenti (Winata et al., 2023): NusaXSenti focuses on sentiment analysis and classification into positive, negative, or neutral sentiments. The dataset includes sentences in Indonesian, English, and multiple regional languages, broadening the scope of sentiment analysis across diverse linguistic contexts. However, it is worth noting that during benchmarking, we're filtering out the English subset since our goal is to measure the model's performance on Indonesian and regional languages.

Intent Classification: Intent Classification involves classifying the intent behind a given sentence. The dataset provides cases of sentences where the model needs to identify the specific intent, contributing to the development and evaluation of models for intent recognition tasks. The dataset is based on BANKING77-OOS (Zhang et al., 2022) data, where we sample only 5 indomain intents and 1 out-of-domain out-of-scope intent (ood-oos). Original utterances are in English, while we translate them into Indonesian, Javanese, and Sundanese.

Colloquial Detection: Colloquial Detection aims to classify the level of formality in a given sentence. The dataset provides tasks for the model to classify whether a given sentence is a colloquial sentence or not. We created this dataset based on the Twitter data from the emotion classification task released by IndoNLU (Wilie et al., 2020). We mark this set as the colloquial sentences. While for the formal sentences, we sampled several lines from our Indonesian textbooks pre-training data.

ID-HateSpeech (Alfina et al., 2017): As the name suggests, ID-HateSpeech is designed specifically for hate speech detection. This dataset aids in evaluating the model to identify and categorize content containing hate speech, contributing to the development of robust hate speech detection systems.

TydiQA-ID (Cahyawijaya et al., 2021) (Clark et al., 2020): TydiQA-ID contributes to the benchmarking efforts with a focus on Indonesian language question answering. It is an extractive Question Answering dataset. This aids in evaluating the model's comprehension of historical and factual information in Indonesian.

IndoSum (Kurniawan and Louvan, 2018): IndoSum is geared towards summarization tasks, providing the model with text to generate concise and informative summaries. This dataset is valuable for evaluating the model in conceptual summarization techniques.

NusaX-MT (Winata et al., 2023): NusaX-MT involves machine translation tasks, where the model is tasked with translating sentences from one language to another. It instructs the model to translate a given sentence from one regional language to another regional language, showcasing the dataset's multilingual translation challenges. Similar to NusaXSenti, this data also consists of Indonesian, English, and 10 regional languages.

3 Training and Experimental Setup

We have built our language model on top of the Llama-2 model, but with some unique adjustments for our needs.

3.1 Expanding the Vocabulary

Recognizing the importance of linguistic diversity, we focused on enhancing our language model's proficiency in both Indonesian and regional languages. To achieve this, we systematically expanded the tokenizer's vocabulary by identifying and incorporating approximately 2,000 frequently used words specific to Indonesian and 1,000 words for regional languages that were absent in the Llama-2 model.

The standard method for enhancing a vocabulary typically involves developing a new tokenizer and integrating it with the existing one. This technique has shown impressive results in projects like Chinese-LLaMA (Cui et al., 2023) and Open-Hathi (SarvamAI, 2023). The effectiveness of this strategy can be attributed to the significant linguis-

Model Name	**Mean Fertility Score**			**Vocab Size**	**% Improvement**		
	Indonesian	**Regional**	**English**		**Indonesian**	**Regional**	**English**
Llama-2-7B	2.858	2.658	1.666	32000	–	–	–
Komodo-7B	2.031	1.996	1.633	35008	28.90%	24.90%	1.98%

Table 1: Fertility Score Analysis. We compare the mean fertility scores of Llama-2-7B and Komodo-7B, measuring their token-splitting behavior across Indonesian, regional, and English languages.

tic distinctions between languages such as Chinese and Hindi when compared to English. In contrast, the Indonesian language employs the same Latin script as English, which presents a different set of challenges.

We tested the traditional method, as well as a new approach where we included the top n words (not tokens) from the Indonesian vocabulary. We discovered that with the new approach, we could achieve better fertility scores by adding around 3000 new vocabulary words. Adding more than 3000 words did not significantly improve the fertility score further, but it increased the size of the embedding matrix, leading to longer training times.

3.2 Optimizing for Efficiency

In our experiment of optimal performance, we ensured that the total number of tokens in our model's vocabulary is a multiple of 64. Drawing inspiration from the recent advancements in nanoGPT (Karpathy, 2023), we acknowledge the significance of seemingly minor details, such as this, in improving the computational efficiency. This strategic decision enhances the model's execution speed, allowing it to leverage computing power more effectively. The final iteration of our model produces an increased curated vocabulary consisting of 35,008 tokens, attuned to strike a harmonious balance between linguistic richness and computational efficiency.

Since we are expanding the tokenizer vocabulary, we also need to initialize the embeddings for these new vocabs. Building on the ideas presented in (Hewitt, 2021), we initialize all new embeddings by averaging all existing embeddings. This approach limits the KL-divergence between the token-level distributions of language models before and after expansion, leading to improved performance during fine-tuning.

3.3 Training & Finetuning

For pretraining, our tokenizer processed about 8.79 billion tokens. Incremental pretraining, building upon Llama-2-7B-Base, was conducted over 3 epochs using LORA (Hu et al., 2021). This approach safeguards against catastrophic forgetting and optimizes hardware and cost requirements. The training utilized 8 x A100 40GB GPUs, taking approximately 300 hours. Supervised Fine-Tuning (SFT) further refined our model on diverse tasks for 5 epochs using LORA. Employing the same GPU configuration, the SFT process took about 36 hours.

4 Evaluation & Results

4.1 Tokenizer Fertility Analysis

In comparing the tokenizer performance between Llama-2-7B, our baseline model, and Komodo-7B, the enhanced version, notable distinctions emerge as shown in Table 1. Llama-2-7B showcases mean fertility scores of 2.858 for Indonesian, 2.658 for regional languages, and 1.666 for English, with a vocabulary size of 32,000. On the other hand, Komodo-7B exhibits substantial improvements with mean fertility scores of 2.031 for Indonesian, 1.996 for regional languages, and 1.633 for English, coupled with an expanded vocabulary size of 35,008. This expansion results in significant percentage improvements—28.90% for Indonesian, 24.90% for regional languages, and 2% for English—highlighting the improved tokenization capabilities of Komodo-7B compared to the Llama-2-7B baseline. This means during inference, Komodo-7B will provide better latency compared to Llama-2. These results underscore the tangible advancements achieved through our model's refined tokenizer, offering improved word tokenization across diverse languages.

4.2 Embedding Position Analysis

We aimed to examine the effectiveness of our model in refining embeddings over the course of

Organization	Model Name	Discriminative Tasks							Generative Tasks		Average
		Indo MMLU	ID-EN	XCOPA-ID	Intent Classification	Colloquial Detection	NusaX-Senti	ID-Hate Speech	TydiQA-ID	Indosum	
OpenAI	GPT-3.5-turbo-0301	51.3	64.5	70.0	82.0	64.1	47.2	68.0	85.3	41.0	63.7
	GPT-3.5-turbo-0613	52.7	66.8	88.2	84.0	75.1	63.3	63.7	86.4	40.0	68.9
	GPT-3.5-turbo-1106	53.3	69.7	89.3	84.0	64.2	59.8	56.6	88.0	42.0	67.4
	GPT-4-preview-1106	**69.8**	78.0	**98.3**	**89.0**	**92.7**	66.1	**73.4**	72.0	33.0	**74.7**
Meta	Llama-2-7B-Chat	30.4	45.6	41.5	57.0	31.4	2.9	41.3	11.7	34.0	32.9
	Llama-2-13B-Chat	32.0	61.7	38.0	59.0	31.1	58.7	57.2	71.9	40.0	50.0
Google	Gemma-7B-it	37.4	73.6	57.7	77.1	18.8	44.2	54.8	73.3	44.0	53.4
Mistral	Mixtral-8x7B-v0.1-Instruct	45.2	57.8	88.7	86.0	41.1	52.8	68.8	90.3	14.0	60.5
AISingapore	Sealion-7B-Instruct-NC	23.9	26.9	41.3	37.0	41.8	30.7	57.3	65.3	26.0	38.9
Cohere	Aya-101-13B	47.7	47.3	84.0	64.0	18.9	74.6	72.7	81.3	39.0	58.8
MBZUAI	Bactrian-X-Llama-7B	23.6	43.2	45.3	42.0	50.3	44.5	42.4	65.0	15.0	41.3
Alibaba	Qwen-1.5-7B-chat	40.0	56.0	29.5	85.0	41.8	58.7	63.9	51.22	29.0	50.6
-	Komodo-7B-Instruct	43.2	**90.5**	79.6	84.0	73.6	**79.3**	56.2	**90.3**	**43.0**	**71.1**

Table 2: This table breaks down how well Komodo-7B-Instruct tackles various language tasks compared to other models. **Notes:** (1) For Sealion & Mixtral, we have used the prompts provided by the authors. (2) Performance of GPT4 in the TydiQA-ID data is low because the model refuse to answer the query most of the time due to hallucination prevention. (3) All evaluation functions are attached in the Appendix section.

pretraining. To do this, we selected 8-10 complete words from various word categories, beginning with pronouns, verbs, and adjectives, and progressing to include specific regional language words like Sundanese and Javanese. Initially, we plotted the embeddings' starting positions. Subsequently, we conducted three epochs of training and generated another plot to visualize the changes in the embeddings.

Figure 4 in the appendix section illustrates the initial and final positions of words in two plots by employing Principal Component Analysis to improve the interpretability of these plots. The plot on the left represents the initial positions, while the plot on the right shows their positions after three epochs of training. An important observation is that words belonging to the same group tend to cluster together after training. The results indicate an improvement in the model's ability to position similar words closer to each other in its memory as training progresses.

4.3 Downstream Tasks

This section evaluates Komodo-7B-Instruct performance across both discriminative and generative tasks. A comparative analysis with other models, including colloquial examples and English proficiency regression, enhances the comprehensive exploration of Komodo-7B-Instruct capabilities. There were situations during evaluation where the models produced results which couldn't be directly be evaluated using heuristics, so we used the help of GPT-4-0613 in that case, we have mentioned the prompts used for each task in the Appendix-section. As shown in Table 2, Komodo-7B-Instruct outperforms various models across different tasks.

4.3.1 Baselines

Our evaluation encompasses a range of both open-source and closed-source massively multilingual models to ensure a thorough assessment. These models include GPT-3.5-turbo-0301, GPT-3.5-turbo-0613, GPT-3.5-turbo-1106, GPT-4-1106-preview, Llama-2-7B-Chat, Llama-2-13B-Chat, Mixtral-8x7B-v0.1-Instruct, Gemma-7B-it (DeepMind, 2024), Sealion-7B-Instruct-NC, Aya-101 (13B), Bactrian-X-Llama-7B, and Qwen-1.5-7B-Chat.

4.3.2 Discriminative Tasks

Let's take a closer look at how Komodo-7B-Instruct is doing in different tasks. Table 2 provides insights that tell us how well Komodo-7B-Instruct understands and discriminates between various types of language challenges. For tasks like IndoMMLU, ID-EN (Indonesian-English) sentences , and X-Copa-ID, Komodo-7B-Instruct consistently scores well, showing it's good at understanding different kinds of language tasks. Special focus is its high score of 90.5 in ID-EN, which is better than many other models. This shows that Komodo-7B-Instruct excels in cross-language understanding, especially between Indonesian and English.

In tasks like Intent Classification, where it figures out what users are trying to do, and Colloquial Detection, where it understands informal language, Komodo-7B-Instruct does really well. Scoring above 80 in Intent Classification and excelling in Colloquial Detection shows that Komodo-7B-Instruct is versatile—it gets what people mean and can handle casual language.

When it comes to understanding sentiment in different languages, Komodo-7B-Instruct has shown it's really good. We can see that Komodo-7B-Instruct achieves SOTA on the NusaX-Senti dataset, with a spectacular score of 79.3.

Figure 6 in the appendix section further analyze the performance breakdown of all benchmarked models on each of the 11 languages within the NusaX-Senti dataset. We can see that Komodo-7B-Instruct excels especially in languages like Acehnese, Buginese, Dayak Ngaju, Madurese, and Toba Batak. In these languages, Komodo-7B-Instruct is the top-performer, being the best at figuring out if the words express positive or negative sentiment.

In other languages like Banjarese, Minangkabau, Sundanese, and Balinese, Komodo-7B-Instruct stands second, just behind models including GPT-4-1106-preview and Aya-101-13B.

4.3.3 Generative Tasks

4.3.4 Generative vs Discriminative

In Figure 1, the plot provides a visual comparison of Komodo-7B-Instruct performance in generative and discriminative tasks. Each point on the plot corresponds to a specific model, showcasing how well Komodo-7B-Instruct balances the creation of meaningful content and accurate discrimination across diverse language challenges. Komodo-7B-Instruct excels in both Generative and Discriminative tasks, scoring 66.6 and 72.3, respectively. It demonstrates strong proficiency, trailing only slightly behind the GPT-4-1106-preview model, which excels notably in Discriminative tasks with a score of 81. This shows Komodo-7B-Instruct robust performance and versatility in handling various language tasks.

4.3.5 Translation Performance Comparison with Google Translate

Figure 5 in the appendix section serves as a valuable analysis for evaluating the translation capabilities of Komodo-7B-Instruct in comparison to Google Translate. The visual representation allows us to discern the languages each platform supports. On the right side, the heatmap illustrates Google Translate's proficiency, primarily in Javanese, English, Indonesian, and Sundanese. However, this leaves numerous language spaces unoccupied.

Conversely, the left side of the heatmap showcases the comprehensive linguistic capabilities of Komodo-7B-Instruct, encompassing a total of 11 regional languages. This inclusive approach extends the reach of education in Indonesia by enabling direct translation from English to a diverse set of regional languages including languages which are not supported by many models and translation systems like Acehnese, Balinese, Banjarese, Buginese, Madurese, Minangkabau, and Toba Batak. Our benchmarking data doesn't contain Lampungnese but it is worth noting that Komodo-7B-Instruct also supports Lampungnese. This is a significant advancement, considering that Google Translate's support is limited to Javanese and Sundanese, primarily spoken in Java. Additionaly, Komodo-7B-Instruct can translate between the regional languages without the need of any intermediate language such as English or Indonesian.

The broader coverage of Komodo-7B-Instruct ensures that individuals across various regions in Indonesia, beyond Java, can benefit from education in their native languages. This not only enhances accessibility but also addresses the challenge of language diversity in educational settings. Therefore, Komodo-7B-Instruct stands as a promising solution for bridging educational gaps and encouraging inclusivity in language learning.

4.3.6 Comparison with Other Fine-Tuned Models

Figure 3 shows that Komodo-7B-Instruct demonstrates robust performance across both generative and discriminative tasks, achieving an average score of 72.63%. In comparison, Gemma-7B-finetuned scores slightly lower at 70.1%, and Llama-2-7B-finetuned stands at 68.9%.

It is essential to highlight that we manually finetuned Llama-2-7B-Base and Gemma-7B-base using the same SFT data employed to train Komodo-7B-Instruct. This approach aims to demonstrate that the pretraining phase conducted on Komodo-7B-Instruct contributes significantly to the enhancement of language understanding capabilities for Indonesian and regional languages. The incremental pretraining performed on the Llama-2-7B-Base model proves beneficial, enabling Komodo-7B-Instruct to achieve superior performance compared to Llama-2-7B-Finetuned.It is worth noting that Gemma shows promising results as a base model compared to Llama-2.

4.3.7 Qualitative Analysis

In addition to quantitative benchmarking, we conducted qualitative testing by providing various general instructions to the model. Figure 7 in the appendix displays a sample of these instructions along with the responses from Llama-2-7B-Finetuned, Aya-101, Mixtral-8x7B-Instruct, and Komodo-7B-Instruct. Notably, the response from Mixtral-8x7B-Instruct is in English, despite the input being in Indonesian. While Llama-2-7b-Finetuned responds in Indonesian, the reply lacks empathy and helpfulness. Aya-101's response is more empathetic but falls short in terms of helpfulness. Conversely, Komodo-7B-Instruct response is highly empathetic and helpful, offering both consolation and solutions.

4.3.8 English Capability Regression Analysis

Perplexity serves as a significant metric for evaluating pretrained language models. Table 3 presents the perplexity benchmarking results for Llama-2-7B-Base and Komodo-7B-Base, focusing on their performance in both English and Indonesian contexts. Llama-2-7B-Base demonstrates a perplexity of 7.7173 in English and 6.2608 in Indonesian, while Komodo-7B-Base exhibits a slightly higher perplexity of 7.8648 in English and a lower perplexity of 5.6477 in Indonesians. These figures provide valuable insights into the language comprehension capabilities of the models, with lower perplexity values indicating enhanced predictive performance. The analysis underscores Komodo-7B-Base's competitive performance, maintaining favorable perplexity levels in both English and Indonesian language domains.

In a detailed examination of Komodo-7B-Base and Llama-2-7B-Base performances across various English language downstream tasks, Komodo-7B-Base consistently demonstrates proficiency comparable to Llama-2-7B-Base in several instances, affirming its competence in handling diverse linguistic challenges. We utilized the LM Evaluation Harness (Sutawika et al., 2023) repository to perform the benchmarking on these datasets.

As shown in the Figure 8 in appendix, Komodo-7B-Base is able to maintain the performance of Llama-2-7B-Base across all tasks, except GSM8k, which consists of mathematical task. This probably happens because our pre-training data consists of very less mathematical data. Komodo-7B-Base excels in Indonesian and regional languages, prioritizing linguistic variations unique to the Indonesian context. While it may show minor disparities in English tasks, its exceptional performance in Indonesian and regional languages demonstrates its effectiveness in its designated domain.

Model	English PPL	Indonesian PPL
Llama-2-7B-Base	7.7173	6.2608
Komodo-7B-Base	7.8648	5.6477

Table 3: This illustrates perplexity benchmarking results for Llama-2-7B & Komodo-7B, emphasizing Komodo-7B's competitive performance in both English and Indonesian contexts.

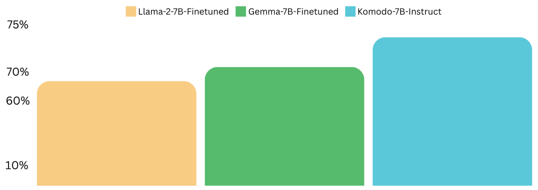


Figure 3: Performance Comparison: Komodo-7B-Instruct versus Llama-2-7B-Finetuned versus Gemma-7B-Finetuned across diverse language tasks

5 Conclusion

In this paper, we introduced Komodo-7B, a robust language model tailored for Indonesian and regional languages. Our extensive evaluation demonstrated its impressive performance across various tasks, showcasing versatility in both generative and discriminative challenges. The model's balanced proficiency makes it well-suited for addressing diverse language nuances. The collaborative efforts and advancements in architecture and training techniques have resulted in a valuable resource for linguistically rich regions. Furthermore, it's noteworthy that Komodo-7B-Instruct supports 11 Indonesia's regional languages not covered by Google Translate. Models accommodating these languages often exhibit subpar performance, emphasizing Komodo-7B-Instruct importance in bridging the language gap for these linguistic nuances.

6 Acknowledgement

We'd like to express our gratitude to the teams behind vLLM (Kwon et al., 2023) and Text Generation Inference (Face, 2023) (by Hugging Face) for their invaluable tools, which boosted our model's data generation capabilities. The ability to distribute parallel requests across multiple GPUs using tensor parallelism has made our work much more efficient. This acknowledgment is a shout-out to the collaborative spirit and the wonderful possibilities that open-source communities bring.

7 Limitations

The current performance of the system is limited by the base model, which is Llama-2. While Llama-2 has proven to be an effective base model for this application, there are potential limitations to its capabilities. As technology advances and new base models are developed, such as Llama-3 (Meta-AI, 2024) or Gemma, it would be beneficial to explore these alternatives. These newer models may offer improved performance, enhanced features, or better compatibility with the specific requirements of the system. Conducting a comparative analysis of different base models could provide valuable insights and guide the decision-making process for future upgrades.

Another limitation lies in the coverage of regional languages. The system currently supports 12 regional languages, which is a commendable achievement considering the linguistic diversity in Indonesia. However, it is important to note that there are more regional languages in Indonesia that are not currently supported by the system. The reason for this limitation is primarily due to data availability. Gathering comprehensive and accurate data for all regional languages can be a challenging task, requiring significant time and resources. Despite these challenges, it is a goal worth pursuing, as expanding the language coverage would make the system more inclusive and beneficial to a larger population. Future efforts should focus on addressing.

References

AISingapore. 2023. Sea-lion (southeast asian languages in one network): A family of large language models for southeast asia. `https://github.com/aisingapore/sealion`.

Ika Alfina, Rio Mulia, Mohamad Ivan Fanany, and Yudo Ekanata. 2017. Hate speech detection in the indonesian language: A dataset and preliminary study. In *2017 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2017*, pages 233–237. Institute of Electrical and Electronics Engineers Inc.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Fei Huang, Binyuan Hui, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Tianyu Liu, Keming Lu, Jianxin Ma, Rui Men, Na Ni, Xingzhang Ren, Xuancheng Ren, Zhou San, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Jin Xu, An Yang, Jian Yang, Kexin Yang, Shusheng Yang, Yang Yao, Bowen Yu, Jianwei Zhang, Yichang Zhang, Zhenru Zhang, Bo Zheng, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2024. Qwen1.5. Technical report, Alibaba.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Samuel Cahyawijaya, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kuncoro, Sebastian Ruder, Zhi Yuan Lim, Syafri Bahar, Masayu Leylia Khodra, Ayu Purwarianti, and Pascale Fung. 2021. Indonlg: Benchmark and resources for evaluating indonesian natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8875–8898.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydiqa a benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

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

Google DeepMind. 2024. Gemma: Introducing new state-of-the-art open models. Technical report, Google DeepMind.

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

Hugging Face. 2023. Text generation inference.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

John Hewitt. 2021. Initializing new word embeddings for pretrained language models.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

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, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Andrej Karpathy. 2023. Tweet on nanogpt optimization.

Fajri Koto, Nurul Aisyah, Haonan Li, and Timothy Baldwin. 2023. Large language models only pass primary school exams in indonesia: A comprehensive test on indommlu". In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Kemal Kurniawan and Samuel Louvan. 2018. Indosum: A new benchmark dataset for indonesian text summarization. In *2018 International Conference on Asian Language Processing (IALP)*, pages 215–220.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang1, Ying Sheng1, Cody Hao Yu Lianmin Zheng, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Attention is all you need. In *Proceedings of the 29th Symposium on Operating Systems Principles*, page 611–626.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. 2023. Openassistant conversations - democratizing large language model alignment. *arXiv preprint arXiv:2304.07327*.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023a. Bactrian-x: Multilingual replicable instruction-following models with low-rank adaptation. *arXiv preprint arXiv:2305.15011*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023b. Alpacaeval: An automatic evaluator of instruction-following models.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023c. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.

Meta-AI. 2024. [link].

Chenghao Mou, Chris Ha, Kenneth Enevoldsen, and Peiyuan Liu. 2023. Chenghaomou/text-dedup: Reference snapshot.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Edoardo Maria Ponti, Goran Glava, Olga Majewska, Qianchu Liu, Ivan Vuli, and Anna Korhonen. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, and Richard Powell et al. 2022. Scaling language models: Methods, analysis and insights from training gopher.

SarvamAI. 2023. Openhathi series: An approach to build bilingual llms frugally. Technical report, SarvamAI.

Lintang Sutawika, Leo Gao, Hailey Schoelkopf, Stella Biderman, Jonathan Tow, Baber Abbasi, Ben Fattori, Charles Lovering, Farzaneh Nakhaee, Jason Phang, Anish Thite, Fazz, Aflah, Niklas Muennighoff, Thomas Wang, Sdtblck, Nopperl, Gakada, Tttyuntian, Researcher2, Chris, Julen Etxaniz, Zdeněk Kasner, Khalid, Jeffrey Hsu, AndyZwei, Pawan Sasanka Ammanamanchi, Dirk Groeneveld, Ethan Smith, and Eric Tang. 2023. A framework for few-shot language model evaluation.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, and et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, X. Li, Zhi Yuan Lim, S. Soleman, R. Mahendra, Pascale Fung, Syafri Bahar, and

A. Purwarianti. 2020. Indonlu: Benchmark and resources for evaluating indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*.

Genta Indra Winata, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2023. Nusax: Multilingual parallel sentiment dataset for 10 indonesian local languages.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google’s neural machine translation system: Bridging the gap between human and machine translation.

Jianguo Zhang, Kazuma Hashimoto, Yao Wan, Zhiwei Liu, Ye Liu, Caiming Xiong, and Philip S. Yu. 2022. Are pretrained transformers robust in intent classification? a missing ingredient in evaluation of out-of-scope intent detection.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric. P Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2023. Lmsys-chat-1m: A large-scale real-world llm conversation dataset.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D’souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

A Functions used for Evaluation of Generated Output by Models

A.1 Example of Evaluation function used for IndoMMLU

```
def evaluate_correctness(row):
    output = str(row[’Output’]).lower()
    answer = str(row[’answer’]).lower()
    Input = str(row[’Input’])

    if len(output) == 1:
        return output[0] == answer
    elif len(output) > 1 and (output[1]
        == ’.’):
        return output[0] == answer
    else:
        return GPT_4_evaluator(Input,
            output, answer)
```

A.2 Example of Evaluation function used for ID-EN

```
def evaluate_correctness(row):
    answer = str(row[’answer’]).strip()
    output = str(row[’Output_Mapped’]).
        strip()

    if output in [’1’, ’0’]:
        return output[0] == answer
    else:
        return GPT_4_evaluator(output,
            answer)
```

A.3 Example of Evaluation function used for X-Copa-ID

```
def evaluate_correctness(row):
    answer = str(row[’answer’])
    output = str(row[’Output’])
    if answer.lower() in output.lower():
        return True
    elif ’Saya tidak dapat menemukan
        jawaban atas pertanyaan
    yang diajukan.’.lower() in output.
        lower():
        return False
    else:
        return GPT_4_evaluator(output,
            answer)
```

A.4 Example of Evaluation function used for Intent-Classification

```
def check_occurrence(sentence, words_set
    ):
    count = sum(1 for word in words_set
        if word.lower() in sentence.
        lower())
    return count >= 2

def return_final_output_intent(output,
    negative_intent="tidak ada"):
    if isinstance(output, float):
        output = str(output)
    intent_list = [
        ’automatic top up’, ’balance not
             updated after cheque or
        cash deposit’,
        ’declined card payment’, ’
            declined transfer’,
        ’edit personal details’,
    ]

    if check_occurrence(output,
        intent_list):
        return negative_intent

    for expected_intent in intent_list:
        if expected_intent.lower() in
            output.lower():
            return expected_intent.lower
                ()

    return negative_intent
```

A.5 Example of Evaluation function used for Colloquial Detection

```
def check_occurrence(sentence, words_set
    ):
    count = sum(1 for word in words_set
        if word.lower() in sentence.
        lower())
    return count >= 2

def return_in_format(response):
    if response is None or isinstance(
        response, (int, float)):
        return -1

    words_set = (”ceremonial”, ”polished
        ”, ”everyday”, ”conversational”,
    ”colloquial”)

    if check_occurrence(response,
        words_set):
        return -1
    elif any(word in response.lower()
        for word in (”ceremonial”, ”
        polished”,
    ”everyday”)):
        return 0
    elif any(word in response.lower()
        for word in (”conversational”,
    ”colloquial”)):
        return 1
    else:
        return response.lower()
```

A.6 Example of Evaluation function used for NusaX-Senti

```
dictionary = {
    ’positive’: ’positif’,
    ’negative’: ’negatif’,
    ’neutral’: ’netral’
}

def evaluate(output, answer):
    output = output.replace(’.’, ’’)
    if ’ ’ not in output:
        output_lower = output.lower()
        answer_lower = answer.lower()
        if output_lower == answer_lower:
            return True
        elif output_lower in dictionary:
            return dictionary[
                output_lower].lower() ==
                 answer_lower
        else:
            return False
    else:
        result = GPT_4_evaluator(output,
             answer)
        return result
```

A.7 Example of Evaluation function used for ID-HateSpeech

```
def evaluate_correctness(row):
    answer = str(row[’answer’]).strip()
    output = str(row[’Output’]).strip()
    output = output.replace(’.’,’’)
    if len(output) == 1:
        return output == answer
    elif output[0] == ’1’ or output[0]
        == ’0’:
        return output[0] == answer
    else:
        return GPT_4_evaluator(output,
            answer)
```

A.8 Example of Evaluation function used for TydiQA-ID

```
def evaluate_correctness(row):
    answer = str(row[’answer’])
    output = str(row[’Output’])
    if answer.lower() in output.lower():
        return True
    elif ’Saya tidak dapat menemukan
        jawaban atas pertanyaan
    yang diajukan.’.lower() in output.
        lower():
        return False
    else:
        return GPT_4_evaluator(output,
            answer)
```

A.9 Example of Evaluation function used for IndoSum

```
from rouge import Rouge
rouge = Rouge()
def calculate_rouge_l(row):
    scores = rouge.get_scores(str(row[’
        Output’]),
    str(row[’answer’]))[0][’rouge-l’]
    return scores[’r’], scores[’p’],
        scores[’f’]
```

B Prompts Used for Evaluation in GPT-4-0613

B.1 Evaluation Prompt for IndoMMLU

Given the following options:{Options}.The model’s generated response is:{output_text}.The correct *answer is: {answer}.Your task is to check if the model's response is correct or not? Provide a response with Yes or No only.*

B.2 Evaluation Prompt for X-Copa-ID

Your task is to check if the Actual Answer is present in the Generated Answer.Generated Answer:{generated_answer},Actual Answer:{actual_answer}.Provide a response with Yes or No only.

B.3 Evaluation Prompt for TydiQA-ID

Your task is to check if the Actual Answer is present in the Generated Answer.Generated Answer:{generated_answer},Actual Answer:{actual_answer}.Provide a response with Yes or No only.

B.4 Evaluation Prompt for ID-HateSpeech

Your task is to Verify if the given output is same as expected answer. Output: {output_text}, Expected Answer: {expected_answer}.Provide a response with Yes or No only.

B.5 Evaluation Prompt for NusaX-Senti

Your task is to Verify if the given output is same as expected answer. Output: {output_text}, Expected Answer: {expected_answer}.Provide a response with Yes or No only.

C Qualitative Samples for Downstream Tasks

C.1 Caption Generation

Input:
"Di bawah ini adalah instruksi yang menjelaskan tugas. Tulis respons yang secara tepat melengkapi permintaan.
###Instruksi:
Buatkan caption untuk post Instagram dengan tema ""bunga putih""
###Respons:"

Input(Translated):
"Below are instructions that explain the task. Write a response that precisely completes the request.
### Instruction:
Make a caption for Instagram post with the theme ""White Flower""
### Response:"

Llama-2-7B-Finetuned Output:
Sensasi keindahan bunga putih yang memikat hati dan menenangkan pikiran. Nikmati keindahan alam dalam setiap detik.

Aya-101:
Sebuah buket bunga putih yang indah, sempurna untuk hari yang cerah!

Mixtral-8x7B-Instruct:
Delicate beauty in monochrome, a serene touch of pure white flowers. #bungaputih #floralperfection

Komodo-7B-Instruct:
Keindahan bunga putih yang mempesona, mengingatkanku pada kesucian dan ketulusan hati. #bungaputih #keindahanalam #kesucian

C.2 Email Generation

Input:
"Di bawah ini adalah instruksi yang menjelaskan tugas. Tulis respons yang secara tepat melengkapi permintaan.
### Instruksi:
Tuliskan balasan email yang sopan untuk menolak tawaran pekerjaan yang diberikan karena renumerasi yang ditawarkan terlalu kecil, namun dengan bahasa yang halus.
### Respons:"

Input(Translated):
"Below are instructions that explain the task. Write a response that precisely completes the request.
### Instruction:
Write a polite email reply to refuse the job offer given because the renumeration offered is too small, but with a subtle language.
### Response:"

Llama-2-7B-Finetuned Output:
Kepada [Nama Pemberi Tawaran Pekerjaan],

Saya sangat menghargai kesempatan yang diberikan untuk melamar pekerjaan di perusahaan Anda. Namun, saya ingin menyampaikan bahwa saya harus menolak tawaran pekerjaan ini karena renumerasi yang ditawarkan terlalu kecil untuk memenuhi kebutuhan hidup saya dan keluarga.

Saya sangat mengapresiasi kesabaran dan kerja keras tim Anda dalam mencari kandidat yang tepat untuk posisi ini. Namun, saya harus mempertimbangkan faktor keuangan dan kesejahteraan saya sendiri dan keluarga saya. Oleh karena itu, saya harus menolak tawaran pekerjaan ini dan mengucapkan terima kasih atas waktu dan usaha yang telah diberikan oleh tim Anda.

Saya berharap dapat menemukan kesempatan *pekerjaan yang lebih sesuai dengan kemampuan dan kebutuhan saya di masa depan. Terima kasih atas perhatian Anda.*

Salam hormat

Nama Anda

Aya-101:
Halo, terima kasih atas tawaran Anda. Saya senang mendengar bahwa Anda tertarik pada pekerjaan ini. Namun, saya pikir renumerasi yang Anda tawarkan terlalu kecil untuk saya. Saya pikir saya lebih cocok untuk pekerjaan yang lebih besar dengan gaji yang lebih tinggi. Jika Anda memiliki kesempatan untuk mempertimbangkan saya, saya akan senang untuk menghubungi Anda kembali. Terima kasih atas waktu dan perhatiannya. Sejujurnya, [Nama Anda]

Mixtral-8x7B-Instruct:
Subject: Regarding Your Job Offer

Dear [Hiring Manager's Name],

I hope this email finds you well. I am writing in response to the job offer that I received from your esteemed organization. I am truly honored to have been considered for this position and appreciate the time and effort invested in the recruitment process.

However, after careful consideration, I regret to inform you that I am unable to accept the position at this time. Although I am excited about the opportunity to contribute to [Company Name] and work alongside a talented team of professionals, I have come to the conclusion that the proposed compensation package does not align with my financial needs and career goals.

I want to emphasize that this decision was not made lightly, and I am grateful for the consideration. I am genuinely impressed with [Company Name]'s mission, vision, and culture, and I am confident that the organization will continue to succeed and thrive in the industry.

I would like to express my appreciation to you and the entire team for your time and support throughout the recruitment process. I hope that we can keep the lines of communication open for future opportunities that may be a better fit for both parties.

Thank you again for the opportunity, and I wish you and the entire team all the best.
Sincerely,

Your Name

Komodo-7B-Instruct:
Kepada Tim Rekrutmen,

Terima kasih atas kesempatan yang diberikan untuk mempertimbangkan tawaran pekerjaan yang diberikan. Saya sangat mengapresiasi kesempatan ini dan sangat tertarik dengan posisi yang ditawarkan.

Namun, setelah mempertimbangkan dengan seksama tawaran ini, saya ingin memberitahukan bahwa saya dengan sangat menyesal harus menolak tawaran ini. Sayangnya, renumerasi yang ditawarkan terlalu kecil untuk memenuhi ekspektasi saya dalam hal kompensasi dan manfaat yang diberikan.

Saya ingin mengucapkan terima kasih atas waktu dan usaha yang telah diberikan oleh tim rekrutmen dalam proses rekrutmen ini. Saya berharap kesuksesan dan keberhasilan untuk perusahaan di masa depan.

Salam hormat,

Your Name

C.3 Suggestion Generation

Input:
"Di bawah ini adalah instruksi yang menjelaskan tugas. Tulis respons yang secara tepat melengkapi permintaan.
### Instruksi:
Gue mau nembak gebetan gue di pantai, kasih tips dong supaya bisa romantis pisan
### Respons:"

Input(Translated):
"Below are instructions that explain the task. Write a response that precisely completes the request.
### Instruction:
I want to confess to my crush at a beach, give me some tips to make it romantic
### Response:"

Llama-2-7B-Finetuned Output:
Anjeun kudu nyiapkeun tempat anu alus pikeun némbak, biasana tempat anu cerah sareng sepi.

Anjeun oge kudu mastikeun yén anjeun boga bahan pikeun némbak, saperti bal, papan némbak, jeung papan pikeun némbak. Sanggeus anjeun manggihan tempat anu alus pikeun némbak, anjeun kudu ngabahas jeung babaturan anjeun ngeunaan naon anu anjeun hoyong ngalakukeun. Contona, anjeun bisa ngomong, "Kuring bade némbak bal kuring kahiji". Lamun babaturan anjeun teu boga ideu, anjeun bisa ngomong, "Kuring bade némbak bal kuring kahiji". Sanggeus anjeun geus nyieun rencana, anjeun kudu ngabahas jeung babaturan anjeun ngeunaan cara ngalakukeunana. Contona, anjeun bisa ngomong,....

Aya-101:

Pastikan cuaca nyaman dan bebas gangguan.

Mixtral-8x7B-Instruct:

Selamat malam! Sungguh menyenangkan dengan tujuan liburan romancemu ke pantai. Berikut adalah beberapa tips untuk membuat gebetanmu menjadi romantis:

1. Pilih waktu: Pilih waktu sore hari atau malam hari ketika cuaca sedang sejuk dan angin menyapu pantai. Warnanya akan menjadi lebih terang, dan keindahan pasir putih akan menjadi lebih menonjol.

2. Temukan tempat yang romantis: Cari sebuah tempat yang tenang dan tenang, seperti area yang jauh dari keramaian. Temukan area pasir yang lembut dan bersih, dan jika mungkin, dekat dengan air laut.

3. Bring the essentials: Jangan lupa untuk membawa bahan-bahan untuk membuat gebetan yang romantis, seperti lampu tengah malam, selimut lembut, dan bantal. Kamu juga mungkin ingin membawa beberapa makanan ringan dan minuman untuk menikmati bersama.

4. Dress up: Pilih pakaian yang nyaman dan romantis. Pilih warna yang cerah dan berkilau, seperti biru muda atau hijau muda, yang akan membuatmu terlihat cantik di bawah sinar matahari terbenam.

5. Take your time: Jangan terlalu cepat membuat gebetan. Nikmati saat itu, bersantai, dan main-main di pasir. Ingat, liburan romance ini adalah tentang kalian dua, bukan hanya tentang gebetan.

6. Berbagi cinta: Sebelum memulai gebetan, berbagi cinta dengan pasanganmu. B"

Komodo-7B-Instruct:

1. Pilih waktu yang tepat. Pilihlah waktu yang tepat untuk mengajak gebetanmu ke pantai. Pilihlah waktu yang tenang dan sunyi, sehingga kamu bisa berdua lebih nyaman dan santai.

2. Siapkan makanan ringan. Siapkan makanan ringan seperti buah-buahan segar, kue-kue kecil, atau camilan lainnya yang bisa dinikmati saat berada di pantai. Ini akan membuat suasana semakin romantis.

3. Siapkan lampu lilin. Siapkan lampu lilin yang bisa dipasang di sekitar pantai. Ini akan memberikan suasana yang lebih romantis dan intim.

4. Siapkan musik. Siapkan musik yang romantis dan sesuai dengan suasana pantai. Ini akan membuat suasana semakin intim dan romantis.

5. Siapkan hadiah kecil. Siapkan hadiah kecil untuk diberikan kepada gebetanmu. Ini akan membuat suasana semakin romantis dan intim.

6. Jangan lupakan tatapan mata. Jangan lupa untuk tatap mata dan berinteraksi dengan gebetanmu. Ini akan membuat suasana semakin romantis dan intim.

7. Nikmati momen bersama. Nikmati momen bersama dengan gebetanmu di pantai. Ini akan membuat suasana semakin romantis dan intim.

C.4 Suggestion Generation

Input:

"Di bawah ini adalah instruksi yang menjelaskan tugas. Tulis respons yang secara tepat melengkapi permintaan.
### Instruksi:
Tolong berikan beberapa ide untuk games yang bisa dilakukan keluarga saat chinese new year
### Respons:"

Input(Translated):

"Below are instructions that explain the task. Write a response that precisely completes the request.
### Instruction:
Please give some ideas for games that family can do during Chinese New Year
### Response:"

Llama-2-7B-Finetuned Output:

1. Game kue tradisional: Keluarga dapat membuat kue tradisional seperti kue tahu, kue pisang, atau kue ubi jalar dan bermain bersama-sama.

2. Game kartu: Keluarga dapat bermain game kartu tradisional seperti mahjong atau weiqi.

3. Game tebak-tebakan: Keluarga dapat bermain game tebak-tebakan dengan tema Cina seperti game Mahjong atau Go.

4. Game karaoke: Keluarga dapat menyanyikan lagu-lagu Cina tradisional dan menonton orang lain menyanyikan lagu-lagu tersebut.

5. Game memasak: Keluarga dapat memasak makanan tradisional Cina bersama-sama dan bermain game memasak.
6. Game tebak-tebakan: Keluarga dapat bermain game tebak-tebakan dengan tema Cina seperti game Mahjong atau Go.
7. Game papan: Keluarga dapat bermain game papan tradisional seperti Go atau Shogi dengan tema Cina.
8. Game permainan papan: Keluarga dapat bermain game permainan papan tradisional seperti Go atau Shogi dengan tema Cina.
9. Game tebak-tebakan: Keluarga dapat bermain game tebak-tebakan dengan tema Cina seperti game Mahjong atau Go.
10. Game karaoke: Keluarga dapat menyanyikan lagu-lagu Cina tradisional dan menonton orang lain menyanyikan lagu-lagu tersebut.

Aya-101:
Berikut adalah beberapa ide untuk permainan yang bisa dilakukan keluarga saat chinese new year: - Keluarga bisa bermain catur. - Keluarga bisa bermain catur. - Keluarga bisa bermain catur. - Keluarga bisa bermain catur. - Keluarga bisa bermain catur. - Keluarga bisa bermain catur. - Keluarga bisa bermain catur. - Keluarga bisa bermain catur. - Keluarga bisa bermain catur. - Keluarga bisa bermain catur. - Keluarga bisa bermain catur. - Keluarga bisa bermain ....

Mixtral-8x7B-Instruct:
1. ""Lucky Red Envelope Hunt"": This game involves hiding red envelopes (ang pao) around the house, and having family members search for them. The envelopes can contain chocolate coins or small toys, adding to the excitement.
2. ""Chinese New Year Trivia"": Create a trivia game with questions about Chinese New Year traditions, history, and culture. This can be a fun and educational way to engage the whole family.
3. ""Chinese Character Practice"": Set up a station where family members can practice writing Chinese characters using traditional calligraphy brushes and ink. This can be a relaxing and meditative activity.
4. ""Lantern Making"": Provide materials for family members to create their own lanterns, which can be hung up around the house as decorations. This can be a fun and creative activity for all ages.
5. ""Chinese Zodiac Scavenger Hunt"": Create a scavenger hunt where family members have to find objects or pictures that represent each of the 12 animals in the Chinese zodiac.
6. ""Chinese Cooking Challenge"": Have a cooking competition where family members have to create a dish using traditional Chinese New Year ingredients, such as dumplings, fish, or noodles.
7. ""Chinese Tea Ceremony"": Hold a traditional Chinese tea ceremony, where family members can learn about the history and significance of tea in Chinese culture.
8. ""Chinese Knot Tying"": Teach family members how to tie traditional Chinese knots using ribbons or cords. This can be a fun and challenging activity that results in beautiful decorations.
9. ""Chinese Chess or Mahjong Tournament"": Organize a tournament where family members can play traditional Chinese board games like chess or Mahjong.
10. ""Lion Dance Performance"": Encourage family members to create their own lion dance costumes and put on a performance. This can be a fun and energetic way to celebrate Chinese New Year traditions.

Komodo-7B-Instruct:
1. Permainan tebak-tebakan
2. Permainan kartu seperti Uno, Blackjack, atau Poker
3. Permainan papan seperti Monopoli atau Risk
4. Permainan video seperti Mario Kart atau Mario Party
5. Membuat kue bersama-sama
6. Membuat hiasan atau dekorasi rumah bersama-sama
7. Mengadakan pesta kostum dan foto bersama
8. Mengadakan permainan bola atau olahraga bersama
9. Mengadakan acara memasak bersama-sama
10. Mengadakan acara menonton film bersama-sama.

C.5 Translation

Input:
"Di bawah ini adalah instruksi yang menjelaskan tugas. Tulis respons yang secara tepat melengkapi permintaan.
### Instruksi:
apa arti kata ""dahar"" dalam bahasa sunda?
### Respons:"

Input(Translated): *"Below are instructions that explain the task. Write a response that precisely completes the request.*

### Instruction:
What does the word ""dahar"" mean in Sundanese?
### Response:"

Llama-2-7B-Finetuned Output:

Dahar nyaéta kecap anu dipaké pikeun ngagambarkeun kagiatan ngadahar.

Aya-101:

Dahar hartina dahar

Mixtral-8x7B-Instruct:

Hallo! Kata "dahar" dalam Bahasa Sunda memiliki arti 'perut atau bagian bawah badan'. Semoga dapat membantu Anda!

Komodo-7B:

Dahar artinya makan.

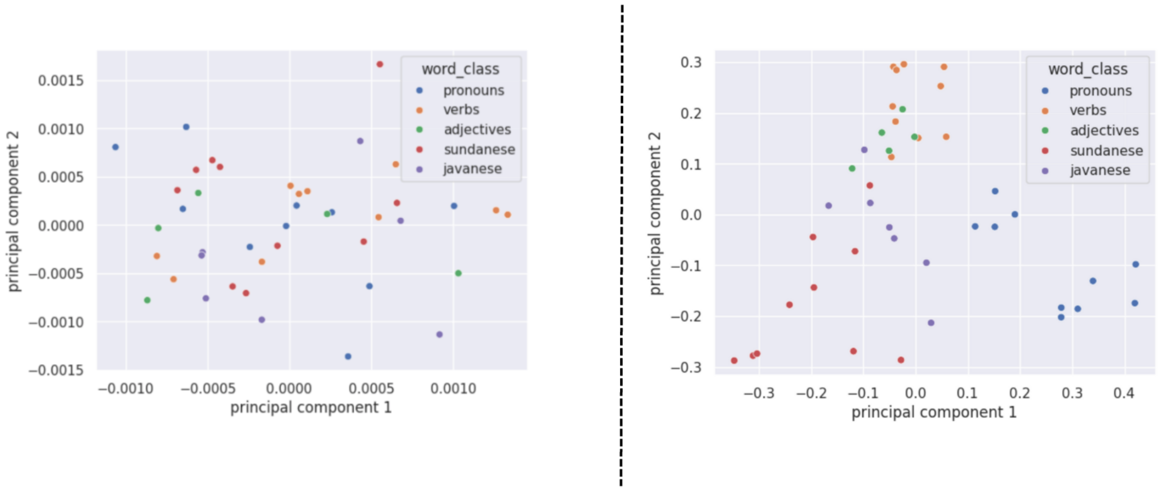


Figure 4: The left plot represents the initial embedding position of words when they are first randomly initialized, while the right plot shows their updated positions after 3 epochs of pre-training. The noticeable grouping of words from the same class in the right plot indicates effective learning and organization of word relationships during pre-training. These plots are created by utilizing PCA with 2 principal components.

Task	**Dataset**	**Metric**	**Unseen Tasks**	**Languages**
Discriminative Tasks				
MCQs	IndoMMLU	Acc.	✗	10
Entailment	ID-EN	Acc.	✗	2
Common Sense Reasoning	X-Copa-ID	Acc.	✗	1
Intent-Classification	Intent-Classification	F1-w-avg	✔	3
Colloquial-Detection	Colloquial-Detection	Acc.	✔	1
Sentiment-Analysis	NusaXSenti	Acc.	✔	11
Hatespeech Detection	ID-Hatespeech	Acc.	✔	1
Generative Tasks				
Translation	NusaX-MT	CHRF++	✗	13
Question-Answering	Tydiqa-ID	Acc.	✗	1
Summarisation	IndoSum	Rouge-L-F1	✗	1

Table 4: Datasets considered for evaluation. Unseen Task refers to tasks entirely excluded from training, which includes the 3 discriminative tasks. The seen tasks refer to the tasks where supervised fine tuning is performed and instances are held-out for evaluation.

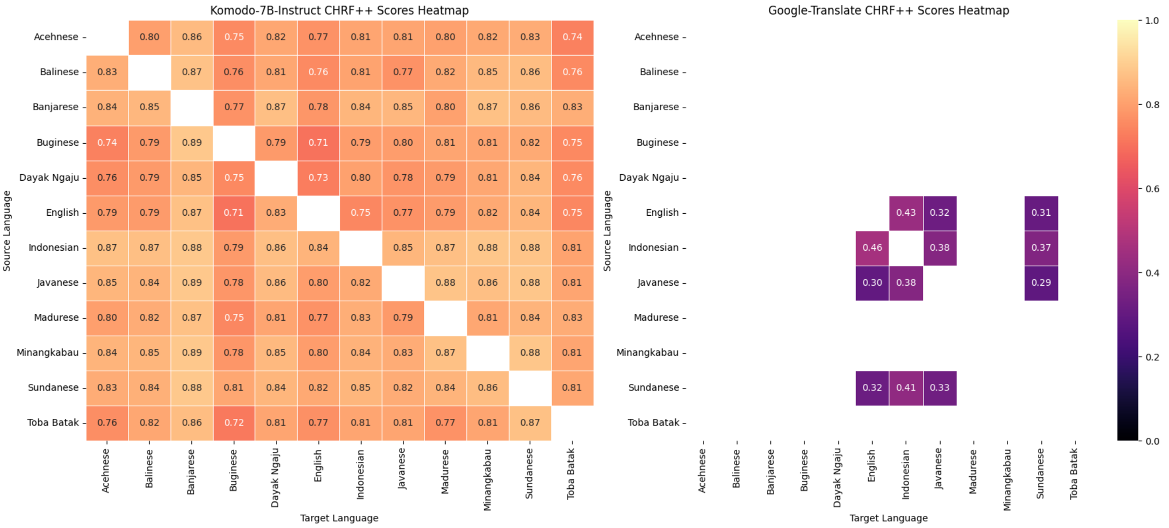


Figure 5: A comparison between the Google-translate & Komodo-7B-Instruct

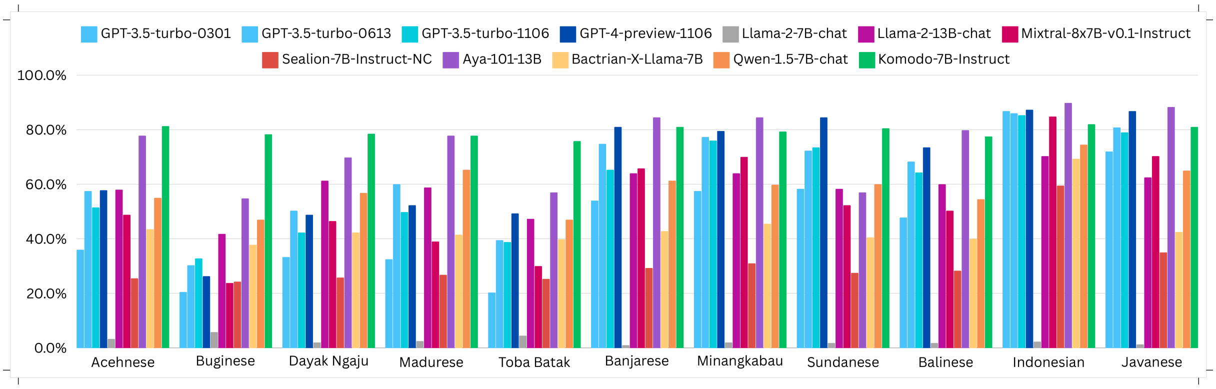
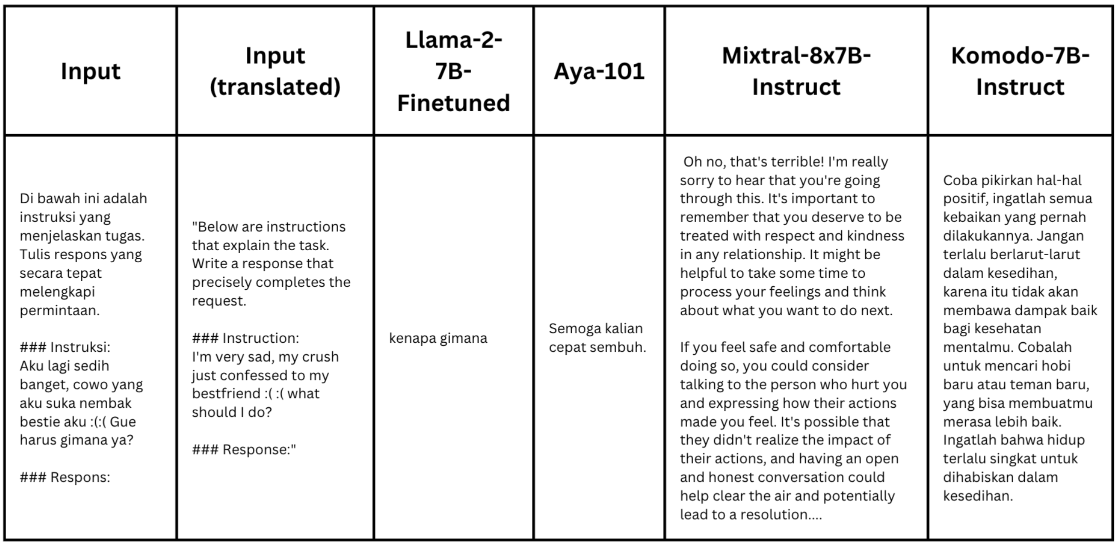


Figure 6: Performance breakdown of all models on NusaX-Senti dataset.



Input	Input (translated)	Llama-2-7B-Finetuned	Aya-101	Mixtral-8x7B-Instruct	Komodo-7B-Instruct
Di bawah ini adalah instruksi yang menjelaskan tugas. Tulis respons yang secara tepat melengkapi permintaan. ### Instruksi: Aku lagi sedih banget, cowo yang aku suka nembak bestie aku :(:(Gue harus gimana ya? ### Respons:	"Below are instructions that explain the task. Write a response that precisely completes the request. ### Instruction: I'm very sad, my crush just confessed to my bestfriend :(:(what should I do? ### Response:"	kenapa gimana	Semoga kalian cepat sembuh.	Oh no, that's terrible! I'm really sorry to hear that you're going through this. It's important to remember that you deserve to be treated with respect and kindness in any relationship. It might be helpful to take some time to process your feelings and think about what you want to do next. If you feel safe and comfortable doing so, you could consider talking to the person who hurt you and expressing how their actions made you feel. It's possible that they didn't realize the impact of their actions, and having an open and honest conversation could help clear the air and potentially lead to a resolution....	Coba pikirkan hal-hal positif, ingatlah semua kebaikan yang pernah dilakukannya. Jangan terlalu berlarut-larut dalam kesedihan, karena itu tidak akan membawa dampak baik bagi kesehatan mentalmu. Cobalah untuk mencari hobi baru atau teman baru, yang bisa membuatmu merasa lebih baik. Ingatlah bahwa hidup terlalu singkat untuk dihabiskan dalam kesedihan.

Figure 7: A sample of general instruction and responses from each model. "Input (translated)" is the translated version of the original "Input" that is passed to the model, this is only for visualization and understanding purpose for non-Indonesian readers.Additional samples of general instructions can be found in the Appendix.

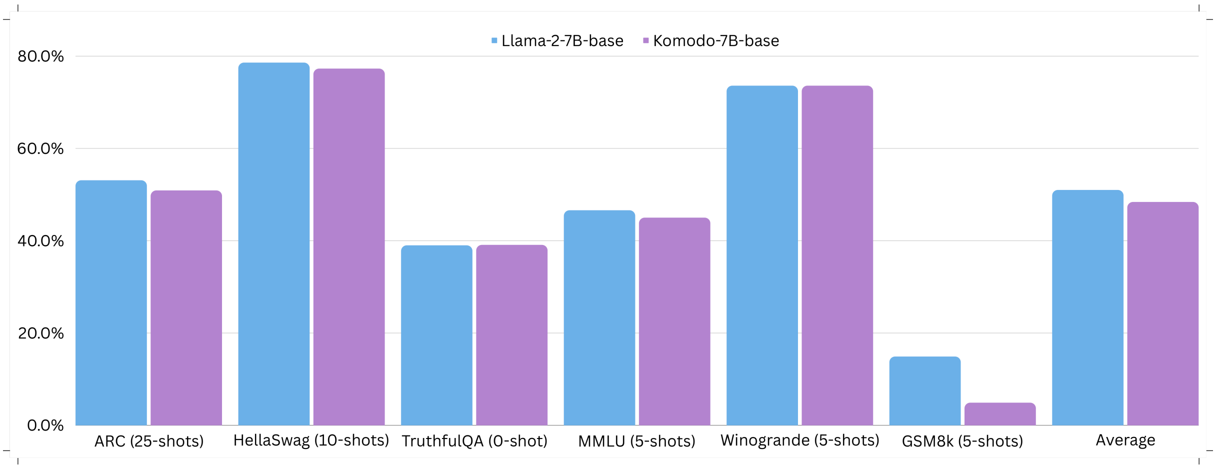


Figure 8: Comparative Analysis of Komodo-7B-Base and Llama-2-7B-Base on English Language Tasks