Cendol: Open Instruction-tuned Generative Large Language Models for Indonesian Languages

Samuel Cahyawijaya[1,4*], **Holy Lovenia**[2,4*], **Fajri Koto**[3,4*], **Rifki Afina Putri**[4,5*], **Emmanuel Dave**[4†], **Jhonson Lee**[4†], **Nuur Shadieq**[4†], **Wawan Cenggoro**[4†], **Salsabil Maulana Akbar**[4†], **Muhammad Ihza Mahendra**[4†], **Dea Annisayanti Putri**[4†] **Bryan Wilie**[1,4], **Genta Indra Winata**[6,4], **Alham Fikri Aji**[3,4*], **Ayu Purwarianti**[4,7,8], **Pascale Fung**[1]

[1]HKUST [2]AI Singapore [3]MBZUAI [4]IndoNLP [5]KAIST
[6]Bloomberg [7]Institut Teknologi Bandung [8]Prosa.ai

Abstract

Large language models (LLMs) show remarkable human-like capability in various domains and languages. However, a notable quality gap arises in low-resource languages, e.g., Indonesian indigenous languages, rendering them ineffective and inefficient in such linguistic contexts. To bridge this quality gap, we introduce Cendol, a collection of Indonesian LLMs encompassing both decoder-only and encoder-decoder architectures across a range of model sizes. We highlight Cendol's effectiveness across a diverse array of tasks, attaining ∼20% improvement, and demonstrate its capability to generalize to unseen tasks and indigenous languages of Indonesia. Furthermore, Cendol models showcase improved human favorability despite their limitations in capturing indigenous knowledge and cultural values in Indonesia. In addition, we discuss the shortcomings of parameter-efficient tunings, such as LoRA, for language adaptation. Alternatively, we propose the usage of vocabulary adaptation to enhance efficiency. Lastly, we evaluate the safety of Cendol and showcase that safety in pre-training in one language such as English is transferable to low-resource languages, such as Indonesian, even without RLHF and safety fine-tuning.[1]

1 Introduction

Indonesia is the fourth most populous country in the world, with around 280 million people spread across more than 17,000 islands within a humongous area of ∼2 million square kilometers. With such a large archipelago surrounding the country, digital services become immensely crucial, making Indonesia the fourth largest internet user in the world, with ∼220 million users. Despite the huge demand, the technology supporting Indonesian digital businesses still lags compared to other much smaller countries. One aspect that it still left behind is the access to state-of-the-art large language model (LLM) technology, such as ChatGPT (OpenAI, 2023a) and GPT4 (OpenAI, 2023b). Although these LLMs support Indonesian and its local languages, these LLMs often have much weaker language representation for such low-resource and underrepresented languages (Cahyawijaya et al., 2023b,a; Asai et al., 2023).

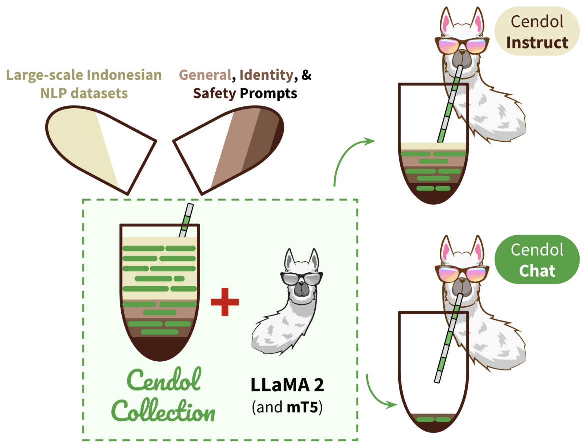

Figure 1: Overview of Cendol Collection and LLM adaptation into Cendol^{inst} and Cendol^{chat} models.

The weak language representation in existing LLMs hurts their ability to generate responses in Indonesian and other underrepresented languages. This also leads to inefficiency during inference due to the vocabulary mismatch, hence texts in these languages are tokenized into much longer tokens (Ahia et al., 2023). Additionally, these LLMs are more prone to safety issues, e.g., giving unsafe responses (Wang et al., 2023b), hallucinations (Guerreiro et al., 2023; Bang et al., 2023), and jailbreaking (Yong et al., 2023; Deng et al., 2023).

To overcome the challenge of weak language representation in Indonesian languages, we introduce Cendol[2], a series of large-scale instruction-

[1]Cendol models are released under Apache 2.0 license and will be made publicly available upon acceptance.

[2]Cendol is an iced sweet dessert that contains droplets of pandan-flavored green rice flour jelly and coconut milk, served with palm sugar syrup. Cendol is popular across Southeast

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14899–14914
August 11-16, 2024 ©2024 Association for Computational Linguistics

tuned LLMs specifically tailored for handling Indonesian indigenous languages. Cendol covers both decoder-only and encoder-decoder LLMs that spread across various scales from 300M up to 13B parameters. Various strategies are incorporated to enable instruction tuning across various scales. We assess the effectiveness of Cendol on a comprehensive evaluation suite, covering various general NLP tasks (e.g., sentiment analysis, topic modeling, machine translation, summarization, etc.), local knowledge, and cultural values evaluations.

Our work highlights the following contributions:

- We introduce Cendol, a collection of state-of-the-art Indonesian LLMs, which outperforms all existing multilingual, Southeast Asian (SEA), and Indonesian LLMs.
- We curate the Cendol Collection, a rigorous instruction-tuned corpus for Indonesian and local languages, covering 23 tasks and 10 languages, with a total of ∼50M instructions.
- We highlight the generalization of Cendol through a comprehensive evaluation suite, showcasing its adaptability towards various Indonesian NLP tasks and languages.
- We demonstrate the ineffectiveness of parameter-efficient tuning approaches, exemplified by LoRA (Hu et al., 2022), in achieving high-quality regional LLMs. This prompts a consideration of the significance of parameter-efficient methods for language adaptation.
- We evaluate the safety of Cendol and showcase that safety in pre-training in one language such as English is transferable to low-resource languages, such as Indonesian.

2 Related Work

2.1 Indonesian Language Models

Various pre-trained Indonesian language models (LMs) have emerged in the past years, including IndoBERT (Wilie et al., 2020; Koto et al., 2020, 2021), IndoBART (Cahyawijaya et al., 2021), and IndoGPT (Cahyawijaya et al., 2021). These models have smaller parameter sizes compared to recent LLMs and have primarily been evaluated only on standard NLP benchmarks. Concurrently, advancements in LLMs have led to the development of multilingual LLMs like BLOOM (Scao et al., 2022) and mT5 (Xue et al., 2021), which include Indonesian, Javanese, and Sundanese. Yet, they fall short of covering other underrepresented Indonesian local languages. LLaMA-2 (Touvron et al., 2023b) also incorporates Indonesian, although it comprises a small portion (0.03%), diminishing its usability in the Indonesian context. Additionally, multilingual LLMs focusing on Southeast Asian languages, such as SEA-LION (Singapore, 2023) and SeaLLM (Nguyen et al., 2023), are beginning to rise, indicating an increasing demand for refining LLMs for underrepresented languages.

2.2 Instruction Tuning

Instruction tuning is a technique to fine-tune LLMs using instruction-and-response pairs. Instruction-tuning allows zero-shot task generalization of LLMs (Sanh et al., 2022; Wei et al., 2022a; Ouyang et al., 2022). Various instruction-tuned LLMs have been developed, both monolingual and multilingual instruction-tuned LLMs using various backbone LLMs such as T5 (Raffel et al., 2020), mT5 (Xue et al., 2021), GPT-3 (Brown et al., 2020), BLOOM (Scao et al., 2022), LLaMA (Touvron et al., 2023a), LLaMA2 (Touvron et al., 2023b), etc. Various efforts have created large-scale instruction-tuned datasets covering different types of instructions, including NLP task-specific instructions (Sanh et al., 2022; Wei et al., 2022a; Muennighoff et al., 2022; Longpre et al., 2023; Cahyawijaya et al., 2023c), multi-turn conversation (Wang et al., 2023a; Chiang et al., 2023), safety prompts (Bai et al., 2022; Touvron et al., 2023b), chain-of-thought instructions (Wei et al., 2022b; Kojima et al., 2022; Liu et al., 2023b), etc.

To better align with human preferences, instruction-tuning can also be coupled with reinforcement learning (RL). There are mainly two approaches for such alignment, i.e., reinforcement learning with human feedback (RLHF) (Christiano et al., 2017; OpenAI, 2023a) and reinforcement learning with artificial intelligence feedback (RLAIF) (Lee et al., 2023; Bai et al., 2022). The reward models are trained to reflect human-preferred qualities such that RL enables the generated responses of LLMs to be more human-aligned.

2.3 LLM Evaluation in Indonesian Languages

Due to the disparity of LLM performance across languages (Blasi et al., 2022), significant efforts have focused on evaluating LLMs in Indonesian (Koto et al., 2023; Blasi et al., 2022). Scao et al. (2022) evaluate BLOOM capabilities in Indonesian through slot-filling, intent classification,

Asia, especially in Indonesia.

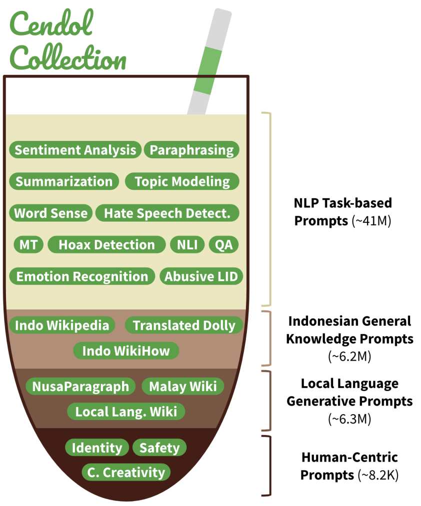

Figure 2: The overview of Cendol Collection. Cendol Collection covers diverse prompts covering various types of instructions with a total of ∼53.5M prompts.

dialogue system, and machine translation tasks. Wei et al. (2023); Ahuja et al. (2023); Asai et al. (2023) compile multilingual NLU and NLG benchmarks and evaluated suites of LLMs in a wide range of NLP tasks.

Further, Koto et al. (2023); Nguyen et al. (2023) provide a more localized perspective by evaluating a suite of LLMs on multi-task language understanding benchmark for Indonesian culture and languages through questions from primary school to university entrance exams in Indonesia and M3Exam (Zhang et al., 2023). ChatGPT model family has also been put to the test by Bang et al. (2023); Leong et al. (2023) for Indonesian NLU, NLG, and reasoning tasks. These evaluations highlight the challenges and opportunities in enhancing LLMs performance for Indonesian and its local languages, particularly addressing issues of dialectal variations and cultural context. We extend this research by evaluating the LLMs we develop across a spectrum of tasks, employing benchmarks specific to Indonesian indigenous languages.

3 Cendol Collection

In total, we create 53.5M prompts, covering a wide range of prompt types including NLP task-based prompts (41M), Indonesian general knowledge prompts (6.2M), local language generative prompts (6.3M), and human-aligned prompts (8.2K). Figure 2 shows the detailed coverage of Cendol Collection across different sources.

3.1 NLP Task-Based Prompt

We collect NLP task-based prompts gathered from 124 dataset subsets covering various tasks, e.g., sentiment analysis, emotion recognition, topic modeling, hate speech detection, natural language inference, machine translation, summarization, question answering, and paraphrasing. The datasets are gathered from NusaCrowd (Cahyawijaya et al., 2023a). We gathered 10-20 prompts for each task type, resulting in a total of ∼41M prompts.

3.2 Indonesian General Knowledge Prompt

To enable better generalization towards general knowledge, we extract general knowledge prompt from Indonesian Wikipedia[3] and Indonesian WikiHow.[4] Additionally, we add an Indonesian machine-translated dataset from Databricks-Dolly-15k.[5] The dataset is translated using a distilled NLLB model with 1.3B parameters.[6] In total, we accumulate ∼6.24M prompts for the Indonesian general knowledge prompt.

3.3 Local Language Generative Prompt

To cover more underrepresented languages spoken in Indonesia, we collect Indonesian local language prompts from two sources, i.e., Wikipedia in local languages and NusaParagraph (Cahyawijaya et al., 2023b). We covered 18 local languages including Sundanese (sun), Javanese (jav), Acehnese (ace), Banjarese (bjn), Buginese (bug), and Gorontalo (gor). Since many local languages in Indonesia are derived from standard Malaysian (msa), we also collect the prompt from Malaysian Wikipedia.[7] For the prompt from Wikipedia, we incorporate the same prompt generation strategy as in §3.2, while for the generative prompt from NusaParagraph, we invert the input and output label of the dataset to make a sentence generation task for the specified local language. In total, we collect ∼6.27M local language generative prompts.

3.4 Human-Centric Prompts

The quality of human-computer interaction is the essence of developing a dialogue agent. To improve the human-computer interaction quality of Cendol, we incorporate three types of human-centric

[3] `https://id.wikipedia.org`
[4] `https://id.wikihow.com/`
[5] `https://huggingface.co/datasets/databricks/databricks-dolly-15k`
[6] `facebook/nllb-200-distilled-1.3B`
[7] `https://ms.wikipedia.org`

prompts, i.e., identity prompt, safety prompt, and computational creativity prompt.

Identity Prompt Identity prompts are incorporated to provide a faithful identity of the Cendol models. These identity prompts include the personal identity of Cendol, the etymology of the word "cendol", the creator information of Cendol, and the neutrality of Cendol on various aspects, e.g., gender, religion, and political stance. In addition, we also include some trivia prompts to increase the engagingness of using Cendol. In total, we cover 125 identity prompts and to increase the representation of these prompts, we upsample the number of identity prompts by 500 in the Cendol Collection.

Safety Prompt We manually construct safety prompts to prevent Cendol from responding to queries that are not appropriate according to cultural norms and values in Indonesia. The safety prompts include prompts for guard-railing illegal activities, e.g., prostitution, gambling, illegal drugs, terrorism, racism, etc. Hate speech, offensive, and biased queries, especially regarding sensitive topics in Indonesia, such as religion and politics, are also guard-railed. In addition, we also prevent Cendol from providing unfaithful answers to queries that require knowledge from an expert, such as legal-related and medical-related queries. In total, we cover 187 safety prompts, and to increase the representation, we upsample the number of safety prompts to 500 in the Cendol Collection.

Computational Creativity Creativity is the essence of humanity (Wilson, 2017). To embed creativity into LLMs, we train Cendol with an open-source poem dataset, i.e., IndoPuisi (Cahyawijaya et al., 2023a), endowing Cendol models the ability to generate Indonesian poems. The dataset covers 7,223 Indonesian poems and we upsample the number of the prompts by 20 in the Cendol Collection.

4 Cendol Recipe

In this section, we describe the configurations for preparing our Cendol models and report the computational resources used in our experiments.

4.1 Backbone Models

We prepare Cendol models of various base models to enable thorough comparison and analysis across different scales. Specifically, we train Cendol from models of different sizes, from 300M up to 13B

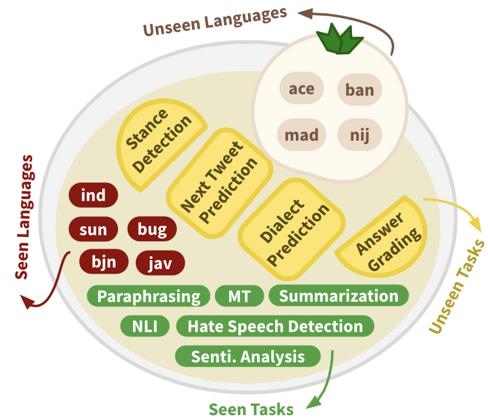

Figure 3: Tasks and languages covered in our Indonesian indigenous benchmark evaluation suite.

parameters, to see the impact of size on performance. We also explore using decoder-only and encoder-only models, and lastly, we also explore using models of different origins to see if the starting base model has any impact. Finally, by producing Cendol in different configurations, users can choose their models based on needs and constraints. Specifically, we train Cendol by continuously fine-tuning decoder-only models, i.e., LLaMA-2 7B and LLaMA-2 13B (Touvron et al., 2023b), as well as encoder-decoder models, i.e., $mT5_{small}$, $mT5_{base}$, $mT5_{large}$, $mT5_{XL}$, $mT5_{XXL}$ (Xue et al., 2021). For all backbone models with <10B parameters, we conduct a full parameter fine-tuning, while for >10B parameter models (i.e., LLaMA-2 13B and $mT5_{XXL}$), we utilize a parameter efficient fine-tuning approach, LoRA (Hu et al., 2022).

4.2 Multi-Phase Tuning

To develop a better instruction-tuned model, we develop the model in two phases of instruction-tuning for each backbone model. The first phase consists only of the NLP task-based prompt data with a total of 18 million instructions over-sampled from the NLP task-based prompt (§3.1) in Cendol Collection. While the second phase consists of other prompt types including general knowledge prompts (§3.2), local language generative prompts (§3.3), and human-centric prompts (§3.4) with a total of 12.8 million instructions. We divide the tuning into two phases to develop both stronger NLP task-specific and more general conversational LLMs. We denote the first phase models as Cendol-Instruct (Cendol^{inst}) and the second phase models as Cendol-Chat (Cendol^{chat}). We report the complete hyperparameters used in Appendix A.

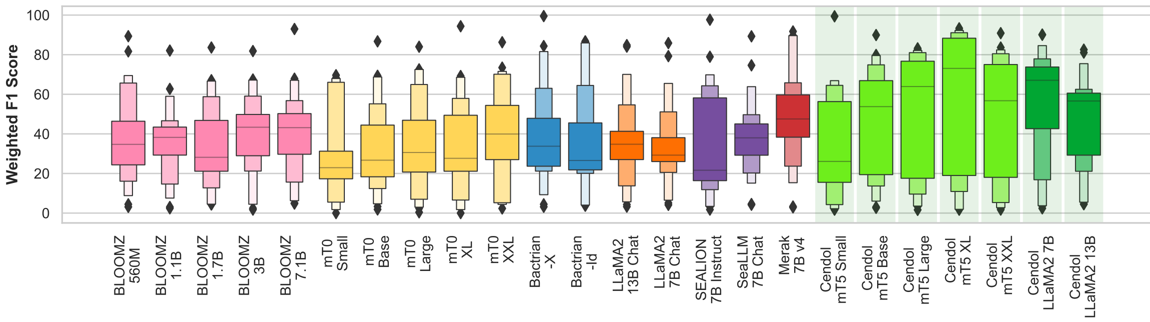

Figure 4: Performance comparison of Cendol models with various multilingual, Southeast Asian, and Indonesian LLMs on NLU tasks. Largest fully fine-tuned Cendol variants, i.e, Cendol mT5$_{XL}$ and Cendol LLaMA2 7B, significantly outperform existing LLMs by ∼20% weighted F1-score.

4.3 Computational Resources

For the instruction tuning, we utilize a 4x40GB A100 GPU server for all models except for the fully fine-tuned LLaMA2-7B model where we use an 8x80GB A100 GPU server. We run the instruction-tuning using DeepSpeed ZeRO-3 (Rajbhandari et al., 2020) to optimize the computation time. The whole instruction tuning takes ∼40 days of training time, with around a 60:40 compute ratio between the first and the second phase instruction tuning. For evaluation, we run the evaluation on a single 40GB A100 GPU server.

5 Evaluation Suite

We evaluate Cendol on various aspects of language proficiency: NLU and NLG (§5.1); generalization capability on unseen tasks and languages (§5.1); as well as local knowledge and cultural commonsense ability (§5.2). In addition, our evaluation includes the first Indonesian safety evaluation for LLMs (Appendix B). Our evaluation suite consists of 10 local languages spoken in Indonesia and spreads across 23 evaluation datasets.

5.1 Indonesian Indigenous Evaluation

To assess the language capability of Cendol models across Indonesian indigenous languages, we design an evaluation benchmark with 15 datasets covering 10 languages including Indonesian and 9 local languages spoken in Indonesia, i.e., Acehnese (ace), Balinese (ban), Banjarese (bjn), Buginese (bug), Javanese (jav), Madurese (mad), Minangkabau (min), Ngaju (nij), and Sundanese (sun).

As shown in Figure 3, this benchmark is split into four subsets: seen tasks, unseen tasks, seen languages, and unseen languages. The seen task subset shows how well the model performs on tasks it has encountered during training, while the unseen task subset assesses the model's ability to generalize to new tasks. The seen language subset and the unseen language subset test the model's performance and generalization to languages that are and are not part of the training data, respectively. For all tasks and datasets, we evaluate the model in a zero-shot prompting setting.

5.2 Local Knowledge and Cultural Commonsense Evaluation

Regional LLMs not only have to understand the local languages but also capture the understanding of local culture and nuances. To demonstrate this, we benchmark Cendol on several datasets. First, we test Cendol on the COPAL-ID (Wibowo et al., 2023) dataset for local-nuanced commonsense reasoning. In COPAL-ID, a scenario is provided and two options are given, one of which is more plausible. All scenarios in COPAL-ID are infused with Indonesian local nuances and context. Next, we also utilize MABL (Kabra et al., 2023), a binary classification dataset where the model is asked to interpret the meaning of a figure of speech in a sentence. We use the Indonesian, Javanese, and Sundanese subsets of MABL. We further benchmark Cendol on IndoStoryCloze (Koto et al., 2022), an Indonesian sentence completion dataset where the model is given two story endings, one of which is more plausible. Lastly, we also use MAPS (Liu et al., 2023a) that benchmarks LLM's understanding of multicultural proverbs and sayings.

6 Impact and Consideration

6.1 Comparison with Existing LLMs

We present the results of the NLU and NLG evaluations of Cendol^{inst} compared to existing LLMs in

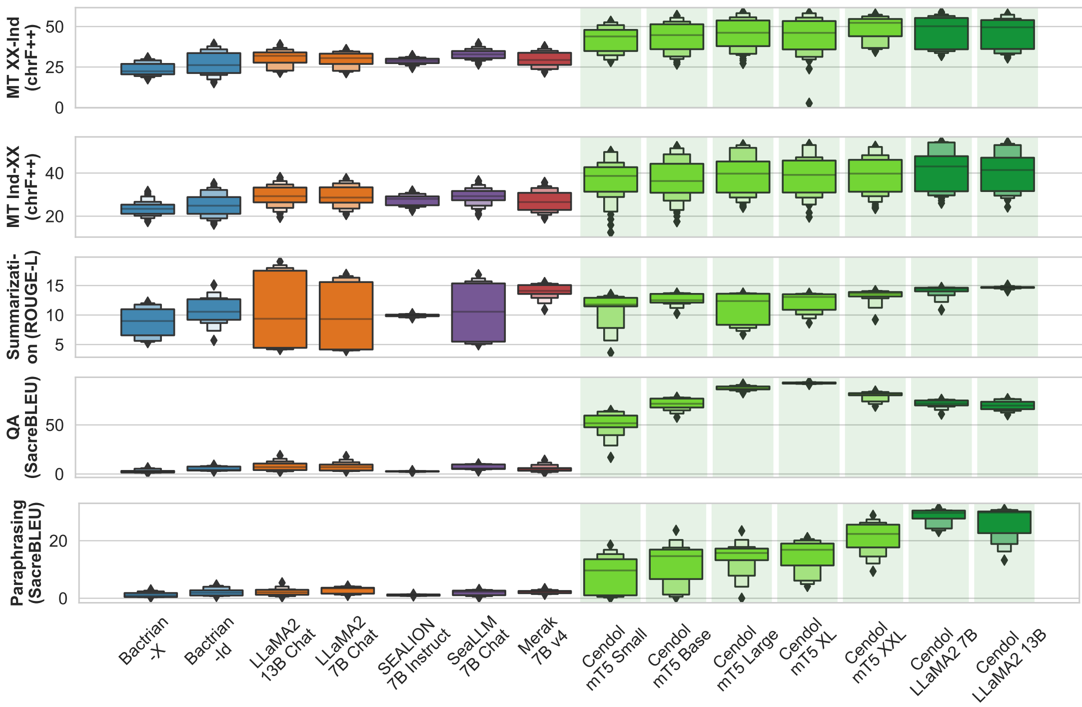

Figure 5: Performance comparison of Cendol^{inst} models with multilingual, SEA, and Indonesian LLMs on NLG tasks: **(1)** machine translation from local languages to Indonesian, **(2)** machine translation from Indonesian to local languages, **(3)** Indonesian language summarization, **(4)** Indonesian language question answering, and **(5)** Indonesian language paraphrasing. BLOOMZ and mT0 are not included since the evaluation datasets are exposed in xP3.

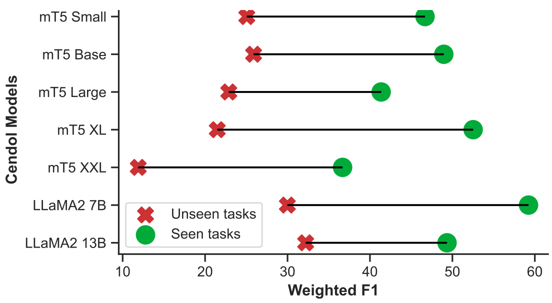

Figure 6: Seen and unseen tasks performance of different Cendol models. All models consistently produce much lower performance for unseen tasks.

Model Type	Cendol^{inst}	Cendol^{chat}	ΔPerf.
Cendol mT5$_{small}$	30.02	29.84	-0.18
Cendol mT5$_{base}$	45.08	35.87	-9.21
Cendol mT5$_{large}$	48.82	40.13	-8.69
Cendol mT5$_{XL}$	58.84	55.79	-3.05
Cendol mT5$_{XXL}$	46.95	37.16	-9.79
Cendol LLaMA2 7B	56.80	50.34	-6.46
Cendol LLaMA2 13B	48.16	45.29	-2.87

Table 1: Comparison of NLU performance between Cendol^{inst} and Cendol^{chat} models.

Figure 4 and Figure 5, respectively. In terms of language understanding capability, the best Cendol^{inst} model (i.e., Cendol mT5$_{XL}$) outperforms all existing LLMs both multilingual, SEA languages, and Indonesian LLMs on the comparable size by ~20% weighted F1-score. Even smaller Cendol^{inst} models (i.e., Cendol mT5$_{base}$ and Cendol mT5$_{large}$ with 600M and 1.2B parameters, respectively), outperform larger LLMs with 7B and 13B parameters. Similarly for language generation, we observe huge improvements in MT, QA, and paraphrasing tasks with at least ~20 increase in chrF++ (Popović, 2015) and SacreBLEU (Post, 2018), respectively. For summarization tasks, Cendol^{inst} models perform similarly to Merak 7B V4 and outperform other baseline LLMs by ~5% ROUGE-L. Our results signify the importance of large-scale instruction tuning to improve the zero-shot NLP capability for underrepresented regional languages.

6.2 Generalization Towards Unseen Data

Unseen Tasks Figure 6 showcases the performance of various Cendol models when evaluated on seen and unseen tasks. There is a huge performance drop (20%-30% weighted F1 score) in the unseen tasks, which can be attributed to two underlying reasons, i.e., 1) the difficulty of the unseen tasks themselves and 2) the generalization of the models towards the unseen tasks. When we

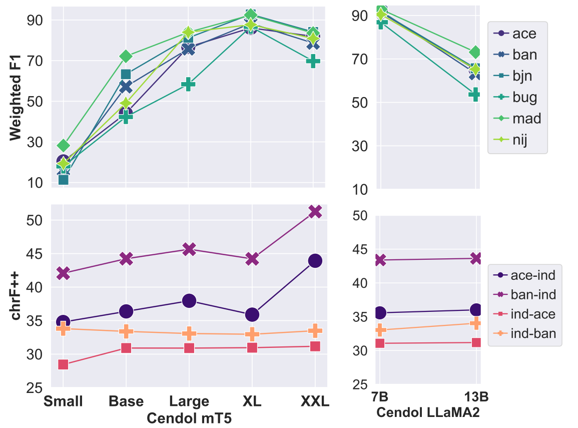

Figure 7: Unseen language performance on **(top)** NLU and **(bottom)** NLG. Cendol shows considerable improvement as the model scales. The drops on LLaMA2-13B and mT5$_{XXL}$ are due to the use of LoRA.

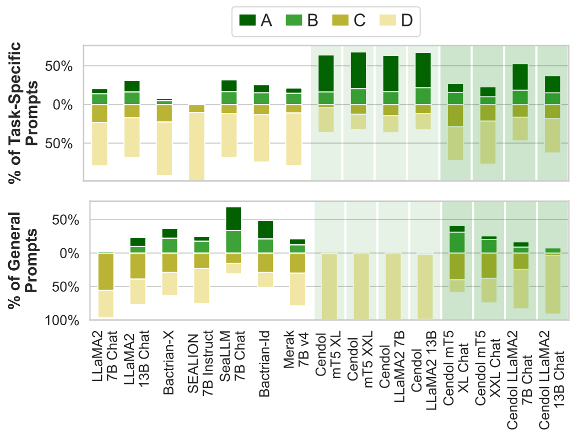

Figure 8: Human evaluation results of the baselines, Cendol^{inst} models, and Cendol^{chat} models on natural **(top)** task-specific and **(bottom)** general prompts prompts. A is the best and D is the worst.

fine-tune a smaller IndoBERT model (Wilie et al., 2020) into the seen and unseen tasks, there is a ∼10% weighted F1 score difference between the unseen and seen tasks. Hence, we attribute the rest ∼10%-20% weighted F1 score as the generalization bottleneck of the Cendol models.

Unseen Languages As shown in Figure 7, the NLU performance to unseen languages follows the scaling law of LLMs. The performance improvements on NLG tasks are less apparent, moreover, no effect of scaling is observed on the translation from Indonesian to the unseen language direction. This showcases that, despite being able to better understand the unseen languages, the LLMs still have difficulty generating sentences in these unseen languages. Interestingly, we observe degradation in terms of NLU performance from the LoRA-tuned models, i.e., LLaMA2-13B and mT5$_{XXL}$, despite their increase in NLG performance.

6.3 General vs. Task-Specific LLMs

We compare the task-specific Cendol^{inst} models with the general Cendol^{chat} models. As shown in Table 1, the task-specific performance of Cendol^{chat} models decreases significantly by up to ∼10% weighted F1-score. We further evaluate the models through human evaluation for both task-specific and general prompts (Figure 8). For the task-specific human evaluation, we sample 60 generation results from all the evaluated NLG tasks. For the general prompts human evaluation, we generate responses from 100 prompts that require some local knowledge about Indonesia. The responses are then rated by 3 annotators with a moderate inter-annotator agreement (κ=0.59). The annotation guideline is described in Appendix C.

Task-specific prompts The result of human evaluation on task-specific prompts is shown in Figure 8. Cendol^{inst} models significantly outperform all other models scoring a large portion of A rating compared to others. Cendol^{chat} models achieve lower ratings, with LLaMA2-based Cendol^{chat} performing slightly lower scores compared to the Cendol^{inst} models, while mT5-based Cendol^{chat} models perform on a par with other multilingual and regional LLMs.

General prompts As shown in Figure 8, Cendol^{inst} models fail to answer general prompts in almost all cases, while Cendol^{chat} models show a trend similar to the NLP task-specific performance of Cendol^{inst} where larger models perform better than a smaller model with the exception on the LoRA-tuned model, i.e., mT5$_{XXL}$. Despite the huge quality shift after the second phase of instruction-tuning, there are only several responses that are accurate and comparable to human standards (Rate A). This shows that supervised fine-tuning alone is not enough and further human-alignment tuning strategy, such as RLHF (Christiano et al., 2017) or RLAIF (Bai et al., 2022), is necessary to generate human-aligned responses.

6.4 Capturing Local Knowledge

We evaluate local knowledge using 7 cultural and local knowledge tasks. The results are presented in Table 2. Our Cendol models are out-competed

Model	MABL id	MABL jv	MABL su	MAPS	COPAL	Indo MMLU	IndoStory Cloze
Multilingual LLM							
BLOOMZ 7.1B	63.83	52.50	50.96	67.14	60.26	28.66	65.12
mT0 XXL	64.79	55.34	54.59	**86.79**	**64.30**	39.85	58.92
Bactrian-X	61.57	52.21	50.67	52.62	52.62	18.83	65.70
LLaMA2 13B*	56.94	51.01	48.97	43.15	49.61	34.98	65.05
Southeast Asian LLM							
SEALION 7B*	59.71	51.68	48.65	34.00	55.21	21.92	63.79
SeaLLM 7B*	64.24	53.46	49.72	58.54	55.11	33.60	68.55
Indonesian LLM							
Bactrian-Id	**65.34**	51.79	48.48	45.84	56.27	22.95	**69.14**
Merak 7B v4	62.30	52.46	50.65	77.78	55.82	**46.27**	66.78
Cendol							
mT5 Small	53.42	53.95	50.50	50.75	48.67	13.07	50.31
mT5 Base	54.58	53.67	51.23	48.99	48.86	14.62	52.50
mT5 Large	56.49	54.93	52.49	46.45	49.21	14.80	55.60
mT5 XL	57.31	54.26	53.80	35.35	50.07	16.36	55.64
mT5 XXL	62.30	**55.80**	52.71	43.02	53.95	14.46	56.87
LLaMA2 7B	58.19	52.74	**55.46**	40.82	50.33	23.54	57.41
LLaMA2 13B	56.82	52.21	54.08	37.32	52.52	21.87	59.09

Table 2: Comparison of Cendol against various LLMs on local knowledge and cultural commonsense tasks. *We use the instruction-tuned versions.

by some existing LLMs on all Indonesian language tasks, especially on IndoMMLU and IndoStoryCloze, where Cendol models perform the worst among all LLMs. Nonetheless, the best Cendol models achieve state-of-the-art performance on two local language tasks, i.e., MABL-jv and MABL-su. This highlights the existing multilingual, Southeast Asian, and Indonesian LLMs' limited understanding of Indonesian local languages. Furthermore, we observed a huge variance over different LLMs in some tasks, such as MAPS and IndoMMLU, which raises the question of whether some LLMs have seen the corresponding evaluation datasets.

6.5 Parameter Efficient Tuning Is Ineffective

We compare the effectiveness and efficiency of the parameter-efficient tuning method with LoRA (Hu et al., 2022) with fully fine-tuned models with a similar training throughput. Specifically, we compare the LoRA-based Cendol mT5$_{XXL}$ model with two other models, i.e., Cendol mT5$_{large}$ and Cendol mT5$_{XL}$. As shown in Table 3, the training throughput of Cendol mT5$_{XXL}$ is ~1.5x higher than Cendol mT5$_{XL}$ and is ~0.34x lower than Cendol mT5$_{large}$. Nonetheless, in terms of other efficiency aspects, such as inference throughput and storage size, Cendol mT5$_{XXL}$ is less efficient than other models. In terms of quality, the training and evaluation losses of Cendol mT5$_{XXL}$ are much higher which leads to a worse performance downstream task performance for both NLU and NLG. These results demonstrate that despite reducing the computational resources compared to fully fine-

Aspect	Cendol mT5$_{large}$	Cendol mT5$_{XL}$	Cendol mT5$_{XXL}$ (LoRA r=128)
Train Throughput (↑)	**120**	28	41
Eval Throughput (↑)	**299**	85	75
Parameter Size (↓)	**1.2B**	3.7B	13B
Train Loss (↓)	0.5819	**0.2898**	0.8015
Eval Loss (↓)	0.5938	**0.2991**	0.7715
Storage Size (↓)	**4.8GB**	14.8GB	52GB
NLU Perf. (↑)	52.07	**59.77**	47.47
NLG Perf. (↑)	35.21	**42.76**	32.09

Table 3: Performance efficiency comparison of smaller fully fine-tuned and parameter-efficient tuning LLMs.

Factor	**Vocab^{ind}**	**Vocab^{orig}**	**ΔPerf.**
Model Efficiency			
Token Efficiency (Ind)	46.34 tokens	58.87 tokens	↑21.28%
Token Efficiency (Oth)	52.61 tokens	61.74 tokens	↑14.79%
Training (per batch)	3.14s	3.55s	↑11.50%
Inference (per 100 steps)	6.63s	8.15s	↑18.71%
Downstream Performance			
NLU Performance	58.51	55.4	↑5.61%
NLG Performance	45.27	45.79	↓1.14%

Table 4: Efficiency and downstream tasks performance comparison between LLaMA2-7b with Indonesian-adapted vocabulary (Vocab^{ind}) and the LLaMA2-7b with original vocabulary (Vocab^{orig}).

tuned models of the same size, parameter-efficient tuning methods are less effective and less efficient compared to the smaller fully fine-tuned models in the case of language adaptation. We provide the Pareto-efficiency curve of Cendol mT5$_{XXL}$ compared to the other fully fine-tuned Cendol mT5 models in Appendix D. As an alternative solution for language adaptation, we introduce an alternative approach for improving the modeling efficiency through vocabulary adaptation (Chau et al., 2020; Poerner et al., 2020; Tai et al., 2020; Koto et al., 2021).

Subword tokenization in LLMs generally produces longer sequences for low-resource language (Ahia et al., 2023) which makes it less efficient. By performing vocabulary adaptation (Chau et al., 2020; Tai et al., 2020; Poerner et al., 2020), we can improve the efficiency during the instruction tuning phase. Specifically, we use a subword vocabulary developed from Indonesian corpora, with the embedding initialized through averaging (Koto et al., 2021), and perform the first phase instruction tuning. As shown in Table 4, the vocabulary adapted model steadily improve the token efficiency by 21.28% for Indonesian text and 14.79% for other local language text. This token efficiency results in improvements in both training and inference with around $\sim$11.50% and $\sim$18.71% efficiency improvement, respectively. Additionally, in terms of downstream performance, the vocabulary-adapted model yields a similar performance compared to the model with the original vocabulary, with 5% improvement on the NLU tasks and 1.14% reduction on the NLG tasks. Our result suggests that vocabulary adaptation with subword averaging provides an adequately representative initialization resulting in a significantly better efficiency and similar downstream task performance after the instruction tuning phase.

6.6 Safety Transferability

We conduct safety evaluations on truthfulness and harmful responses. The experiment and results are detailed in Appendix B. It shows that Cendol is on par with other existing multilingual and regional LLMs in terms of safety. Interestingly, LLaMA2-based Cendol yields a much better safety score than mT5-based Cendol, suggesting the transferability of LLaMA2's safety pre-training to Cendol.

7 Conclusion

We introduce Cendol, a collection of Indonesian LLMs covering both decoder-only and encoder-decoder architecture over various model sizes, and Cendol Collection, a large-scale instruction-tuning dataset for Indonesian and its local languages. We highlight the effectiveness of Cendol on a wide range of tasks, achieving $\sim$20% improvement for both NLU and NLG tasks. Cendol also generalizes to the local languages in Indonesia. Furthermore, we demonstrate our effort for human alignment through a supervised fine-tuning approach, which yields a significant improvement in terms of human favorability. We also discuss two limitations of Cendol: 1) human-preferred response needs to be further enhanced with a better human alignment approach, and 2) despite their amazing performance on NLP tasks, Cendol models still fall behind on capturing local knowledge and cultural values in Indonesia. Moreover, we analyze the generalization of Cendol to unseen tasks and languages, the ineffectiveness and inefficiency of LoRA for language adaptation, vocabulary adaptation as an efficient tuning alternative, and the safety of Cendol models.

Acknowledgements

We thank Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi (Ditjen DIKTI), Indonesia for partially supporting the computing resources used in this project. This work is also supported by the National Research Foundation, Singapore under its AI Singapore Programme; PhD Fellowship Award, the Hong Kong University of Science and Technology; and PF20-43679 Hong Kong PhD Fellowship Scheme, Research Grant Council, Hong Kong.

Limitations

Better Human Alignment through Reinforcement Learning One limitation of our work is the limited exploration of human value alignment. While our model can generate human-like text, it cannot generate responses that are aligned with human values, goals, and preferences. This lack of explicit human value alignment may result in the model producing outputs that are not only irrelevant but also potentially harmful or offensive to humans. Furthermore, without proper alignment, the model may not be able to understand and respond appropriately to the nuances and complexities of human communication, leading to misunderstandings and misinterpretations. Therefore, future work should focus on exploring and integrating human alignment techniques to ensure that the LLM can be safely and effectively used in real-world applications that involve human interaction.

Capturing Local Knowledge Another notable limitation that has emerged is the insufficient capability of our Cendol models to capture and reflect local cultural values accurately. This shortfall is partly due to the underrepresentation of diverse cultural contexts within the datasets used to train these LLMs. The majority of data feeding into the development of LLMs tends to be sourced from dominant languages and cultures, often overlooking the rich and nuanced expressions found in less-represented communities. Consequently, LLMs may exhibit biases that favor certain cultural norms and idioms, leading to misinterpretations or inappropriate responses when dealing with languages or dialects that are embedded with local cultural significance. The lack of cultural sensitivity in LLMs not only hinders effective communication but can also perpetuate stereotypes and misunderstandings.

Safety Evaluation Despite being the first to evaluate safety in the Indonesian, our current safety evaluation is done through translating the existing English safety corpora, i.e., TruthfulQA, ToxiGen, and ImplicitHate. Although most of the sentences remain valid, some of them are not natural and less culturally relevant to the regional context of Indonesia. These translated corpora are likely to miss important features such as local and cultural nuances, as well as contextual language which can hinder the effectiveness of the safety evaluation. To make the safety evaluation more culturally relevant to the Indonesian context, the evaluation should utilize locally sourced Indonesian safety corpora. We expect future work to explore this direction to help ensure the safety evaluation is sensitive to the local cultural and social environment and provide more accurate insights into potential safety risks specific to the regional values.

Single-Turn Human-Computer Interaction Although our instruction-tuning data and user-oriented evaluation primarily focus on building general-purpose LLMs which are commonly expected to be able to respond interactively in a multi-turn manner. It is essential to acknowledge that our Cendol^{inst} and Cendol^{chat} models are not currently optimized for handling multi-turn dialogues. In other words, they are not expected to be able to engage in a continuous human-computer interaction. This can result in less coherent and less effective responses when compared to models specifically designed for a continuous human-computer interaction. Therefore, future work should focus more on developing a multi-turn dialogue system, by preserving the context and the interactions between the user and the model in previous turns and carry them to future turns.

Ethics Statement

Our research underscores the imperative of democratizing access to NLP technology for underrepresented languages, with a particular emphasis on Indonesian and its local languages. We recognize and embrace the ethical responsibilities inherent in language research, acutely aware of its potential impact on diverse linguistic communities. Our commitment to inclusivity, cultural relevance, and fairness is the cornerstone of our study. Transparent and equitable collaboration is the lifeblood of our work, and we uphold a fair and transparent scoring guideline that aligns with our core principles.

Throughout our study, we have made conscious efforts to engage with language communities, involve local experts, and respect their linguistic and cultural nuances. This effort is not merely a component of our research - it is an ongoing dialogue, fostering mutual respect and understanding. Our ultimate goal is to contribute to a more inclusive NLP landscape, one that celebrates linguistic diversity and mitigates biases. By encouraging further collaboration and ensuring that the voices of underrepresented language communities are heard, we aim to address their specific needs in the development of language technology.

References

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? tokenization in the era of commercial language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9904–9923, Singapore. Association for Computational Linguistics.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Akari Asai, Sneha Kudugunta, Xinyan Velocity Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2023. Buffet: Benchmarking large language models for few-shot cross-lingual transfer. *arXiv preprint arXiv:2305.14857*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei

Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.

Damián Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505.

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.

Samuel Cahyawijaya, Holy Lovenia, Alham Fikri Aji, Genta Indra Winata, Bryan Wilie, Rahmad Mahendra, Christian Wibisono, Ade Romadhony, Karissa Vincentio, Fajri Koto, Jennifer Santoso, David Moeljadi, Cahya Wirawan, Frederikus Hudi, Ivan Halim Parmonangan, Ika Alfina, Muhammad Satrio Wicaksono, Ilham Firdausi Putra, Samsul Rahmadani, Yulianti Oenang, Ali Akbar Septiandri, James Jaya, Kaustubh D. Dhole, Arie Ardiyanti Suryani, Rifki Afina Putri, Dan Su, Keith Stevens, Made Nindyatama Nityasya, Muhammad Farid Adilazuarda, Ryan Ignatius, Ryandito Diandaru, Tiezheng Yu, Vito Ghifari, Wenliang Dai, Yan Xu, Dyah Damapuspita, Cuk Tho, Ichwanul Muslim Karo Karo, Tirana Noor Fatyanosa, Ziwei Ji, Pascale Fung, Graham Neubig, Timothy Baldwin, Sebastian Ruder, Herry Sujaini, Sakriani Sakti, and Ayu Purwarianti. 2023a. Nusacrowd: Open source initiative for indonesian nlp resources.

Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Dea Adhista, Emmanuel Dave, Sarah Oktavianti, Salsabil Maulana Akbar, Jhonson Lee, Nuur Shadieq, Tjeng Wawan Cenggoro, Hanung Wahyuning Linuwih, Bryan Wilie, Galih Pradipta Muridan, Genta Indra Winata, David Moeljadi, Alham Fikri Aji, Ayu Purwarianti, and Pascale Fung. 2023b. Nusawrites: Constructing high-quality corpora for underrepresented and extremely low-resource languages.

Samuel Cahyawijaya, Holy Lovenia, Tiezheng Yu, Willy Chung, and Pascale Fung. 2023c. InstructAlign: High-and-low resource language alignment via continual crosslingual instruction tuning. In *Proceedings of the First Workshop in South East Asian Language Processing*, pages 55–78, Nusa Dua, Bali, Indonesia. Association for Computational Linguistics.

Samuel Cahyawijaya, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kuncoro, Sebastian Ruder, Zhi Yuan Lim, Syafri Bahar, Masayu 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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ethan C. Chau, Lucy H. Lin, and Noah A. Smith. 2020. Parsing with multilingual BERT, a small corpus, and a small treebank. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1324–1334, Online. Association for Computational Linguistics.

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

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Yue Deng, Wenxuan Zhang, Sinno Jialin Pan, and Lidong Bing. 2023. Multilingual jailbreak challenges in large language models.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nuno M. Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Saghar Hosseini, Hamid Palangi, and Ahmed Hassan Awadallah. 2023. An empirical study of metrics to measure representational harms in pre-trained language models. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 121–134, Toronto, Canada. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Fikri Aji, Genta Winata, Samuel Cahyawijaya, Anuoluwapo Aremu, Perez Ogayo, and Graham Neubig. 2023. Multi-lingual and multi-cultural figurative language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8269–8284, Toronto, Canada. Association for Computational Linguistics.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

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*, pages 12359–12374, Singapore. Association for Computational Linguistics.

Fajri Koto, Timothy Baldwin, and Jey Han Lau. 2022. Cloze evaluation for deeper understanding of commonsense stories in Indonesian. In *Proceedings of the First Workshop on Commonsense Representation and Reasoning (CSRR 2022)*, pages 8–16, Dublin, Ireland. Association for Computational Linguistics.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. IndoBERTweet: A pretrained language model for Indonesian Twitter with effective domain-specific vocabulary initialization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10660–10668, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fajri Koto, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. 2020. IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 757–770, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback.

Wei Qi Leong, Jian Gang Ngui, Yosephine Susanto, Hamsawardhini Rengarajan, Kengatharaiyer Sarveswaran, and William Chandra Tjhi. 2023. Bhasa: A holistic southeast asian linguistic and cultural evaluation suite for large language models. *arXiv preprint arXiv:2309.06085*.

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

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2023a. Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings.

Hanmeng Liu, Zhiyang Teng, Leyang Cui, Chaoli Zhang, Qiji Zhou, and Yue Zhang. 2023b. LogiCoT: Logical chain-of-thought instruction tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2908–2921, Singapore. Association for Computational Linguistics.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, et al. 2023. Seallms–large language models for southeast asia. *arXiv preprint arXiv:2312.00738*.

OpenAI. 2023a. Chatgpt.

OpenAI. 2023b. Gpt-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2020. Inexpensive domain adaptation of pretrained language models: Case studies on biomedical NER and covid-19 QA. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1482–1490, Online. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

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

Wen Tai, H. T. Kung, Xin Dong, Marcus Comiter, and Chang-Fu Kuo. 2020. exBERT: Extending pre-trained models with domain-specific vocabulary under constrained training resources. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1433–1439, Online. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

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, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023a. Openchat: Advancing open-source language models with mixed-quality data.

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael R Lyu. 2023b. All languages matter: On the multilingual safety of large language models. *arXiv preprint arXiv:2310.00905*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, et al. 2023. Polylm: An open source polyglot large language model. *arXiv preprint arXiv:2307.06018*.

Haryo Akbarianto Wibowo, Erland Hilman Fuadi, Made Nindyatama Nityasya, Radityo Eko Prasojo, and Alham Fikri Aji. 2023. Copal-id: Indonesian language reasoning with local culture and nuances. *arXiv preprint arXiv:2311.01012*.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, and Ayu Purwarianti. 2020. IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 843–857, Suzhou, China. Association for Computational Linguistics.

Edward O Wilson. 2017. *The origins of creativity*. Liveright Publishing.

Minghao Wu, Abdul Waheed, Chiyu Zhang, Muhammad Abdul-Mageed, and Alham Fikri Aji. 2023. Lamini-lm: A diverse herd of distilled models from large-scale instructions.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Zheng-Xin Yong, Cristina Menghini, and Stephen H. Bach. 2023. Low-resource languages jailbreak gpt-4.

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

A Tuning Hyperparameters

We provide detailed hyperparameter tuning for developing Cendol^{inst} and Chat for each different model architecture in Table 5.

Hyperparameter	mT5 small..xl	mT5 xxl	LLaMA-2 7B	LLaMA-2 13B
		Cendol^{inst}		
max_input_length	512	512	512	512
max_output_length	256	256	768	768
batch_size	128	128	128	128
bfp16	True	True	True	True
zero_config	zero-3	zero-3	zero-3	zero-3
lr	3e-4	2e-4	2e-5	2e-4
lora_r	-	128	-	128
lora_alpha	-	128	-	128
lora_dropout	-	0.05	-	0.05
		Cendol^{chat}		
max_input_length	512	512	512	512
max_output_length	256	256	768	768
batch_size	128	128	128	128
bfp16	True	True	True	True
zero_config	zero-3	zero-3	zero-3	zero-3
lr	3e-5	1e-4	1e-5	1e-4
lora_r	-	128	-	128
lora_alpha	-	128	-	128
lora_dropout	-	0.05	-	0.05

Table 5: List of hyperparameter settings used during the instruction-tuning of Cendol^{inst} and Cendol^{chat}.

B Safety Evaluation

model	**lang**	**accuracy**
Bactrian-X	ind	67.26%
Bactrian-Id	ind	**98.41%**
LLaMA2 7B Chat	ind	95.72%
LLaMA2 13B Chat	ind	97.28%
SEALION 7B Instruct-nc	ind	53.46%
SeaLLM 7B Chat	ind	76.90%
Merak 7B v4	ind	**88.98%**
Cendol mT5 Small	ind	47.25%
Cendol mT5 Base	ind	87.76%
Cendol mT5 Large	ind	98.56%
Cendol mT5 XL	ind	95.32%
Cendol mT5 XXL	ind	<u>**98.75%**</u>
Cendol LLaMA2 7B	ind	31.21%
Cendol LLaMA2 13B	ind	71.63%

Table 6: Evaluation of Cendol and benchmark LLMs on the automatic truthful benchmark (Higher is better). The overall most truthful model is denoted by <u>underline</u>, while within the group, they are denoted by **bold**.

In our analysis, we focus on assessing the language model's performance in terms of its **truthfulness** and **toxicity**. Specifically, **truthfulness** pertains to the model's ability to avoid disseminating information that is inaccurate due to misconceptions or erroneous beliefs. Meanwhile, **toxicity**

model	asian	black	chinese	jewish	latino	lgbtq	mental disability	mexican	middle eastern	muslim	native american	physical disability	women	avg
Bactrian-X	37.34%	31.96%	38.47%	36.74%	29.52%	31.49%	24.94%	29.96%	27.29%	30.93%	**24.89%**	23.06%	23.76%	30.03%
Bactrian-Id	<u>**37.54%**</u>	<u>**33.45%**</u>	<u>**38.51%**</u>	<u>**37.11%**</u>	<u>**30.75%**</u>	<u>**31.99%**</u>	**25.07%**	<u>**31.16%**</u>	**27.37%**	**31.25%**	23.99%	**23.31%**	**24.63%**	<u>**30.47%**</u>
LLaMA2 7B Chat	29.25%	26.09%	28.01%	23.73%	16.55%	21.79%	15.80%	18.19%	19.96%	23.76%	21.08%	16.12%	16.19%	21.27%
LLaMA2 13B Chat	27.07%	24.92%	25.40%	23.08%	16.15%	20.96%	14.88%	17.33%	18.46%	23.10%	21.46%	14.85%	15.51%	20.24%
SEALION 7B Instruct	**36.61%**	32.62%	**37.02%**	**30.99%**	**22.91%**	**30.74%**	**25.96%**	**22.46%**	23.63%	31.32%	26.08%	**24.71%**	**23.98%**	**28.39%**
SeaLLM 7B Chat	34.50%	**33.20%**	33.70%	30.49%	20.95%	24.95%	23.15%	22.25%	**23.83%**	<u>**32.20%**</u>	**27.58%**	21.99%	18.09%	26.68%
Merak 7B v4	24.06%	20.91%	20.98%	20.81%	12.50%	17.12%	9.73%	13.62%	12.45%	19.86%	14.44%	9.28%	10.92%	15.90%
Cendol mT5 Small	1.86%	1.10%	2.17%	1.22%	2.42%	1.30%	2.06%	2.36%	2.12%	1.31%	0.16%	2.15%	1.14%	1.64%
Cendol mT5 Base	8.30%	3.88%	10.10%	9.57%	4.50%	7.23%	7.08%	3.36%	12.64%	7.50%	3.12%	3.32%	4.55%	6.55%
Cendol mT5 Large	4.69%	3.39%	4.86%	3.40%	3.83%	3.37%	4.42%	3.33%	4.67%	4.80%	3.59%	5.16%	4.09%	4.12%
Cendol mT5 XL	6.24%	4.31%	6.35%	5.66%	3.58%	4.60%	6.47%	3.83%	6.08%	7.93%	2.91%	8.87%	6.65%	5.65%
Cendol mT5 XXL	3.09%	1.74%	2.64%	1.44%	1.48%	1.55%	2.23%	1.40%	1.95%	1.80%	1.00%	1.96%	1.42%	1.82%
Cendol LLaMA2 7B	**30.83%**	31.64%	**31.55%**	25.00%	**28.82%**	**30.49%**	<u>**28.07%**</u>	**29.62%**	<u>**28.90%**</u>	27.50%	<u>**31.18%**</u>	<u>**25.10%**</u>	<u>**27.86%**</u>	**28.97%**
Cendol LLaMA2 13B	29.23%	**32.62%**	26.51%	**27.72%**	21.19%	24.85%	22.45%	19.41%	26.66%	**31.85%**	23.86%	18.33%	19.01%	24.90%

Table 7: Evaluation of Cendol and benchmark LLMs on ToxiGen (higher means less toxic). The overall least toxic models are denoted by <u>underline</u>, while within the group, they are denoted by **bold**.

refers to the model's propensity to produce content that is toxic, rude, adversarial, or implicitly hateful. We leverage three distinct datasets: TruthfulQA (Lin et al., 2022), which benchmarks the veracity of responses, ImplicitHate (ElSherief et al., 2021), designed to identify underlying hate speech, and ToxiGen, a resource for gauging the generation of toxic text (Hartvigsen et al., 2022).

We first translate the dataset using a distilled NLLB model with 1.3B parameters [8] and evaluate all the datasets on a set of 31 LLMs including Cendol and multilingual, regional, and Indonesian-adapted LLMs. For TruthfulQA, we report the accuracy score on the MC1 subset as the percentage of generations that are both truthful and informative. For ToxiGen and ImplicitHate, we employ an automatic safety score evaluation metric from Hosseini et al. (2023) as the percentage of likeliness of the model producing benign over harmful sentences. We show the evaluation result on TruthfulQA, ToxiGen, and ImplicitHate, in Table 6, Table 8, and Table 7, respectively.

Safety Evaluation Result Cendol mT5 XXL model outperforms mT0 XXL, demonstrating a 16.75% enhancement in both truthfulness and informativeness. Additionally, evaluations of the Cendol mT5 XXL Chat indicate a reduction in implicit hate speech by 6.28% points and a decrease in toxicity by 2.45% points. A trend is discernible within the Cendol mT5 model series, revealing that an increase in model size correlates with improvements in truthfulness and informativeness. Although the Cendol mT5 series excels in these areas, the Cendol LLaMA2 series demonstrates a superior capability in generating significantly lower levels of toxic and hateful outputs. Interestingly, LLaMA2 (Touvron et al., 2023b) incorporates safety measures during the pretraining, which is lacking in mT5 (Xue et al., 2021). **This suggests that there is a transfer of safety across languages from English to Indonesian**. In addition, Cendol models are also comparably more truthful and less toxic, when compared to the local, regional, and Indonesian-adapted models. For instance, the Cendol mT5-XL model, with 3.7 billion parameters, is found to be 6.34 percentage points more truthful than the Merak-7B-v4 model. In terms of toxicity, the Cendol LLaMA2 7B model is less toxic by 0.58 percentage points compared to its regional counterpart, the SEALION 7B Instruct model. It's noteworthy, however, that regional models are generally less prone to producing implicit hate speech.

model	score
Bactrian-X	29.62%
Bactrian-Id	<u>**30.46%**</u>
LLaMA2 7B Chat	20.18%
LLaMA2 13B Chat	19.44%
SEALION 7B Instruct	**27.57%**
SeaLLM 7B Chat	26.23%
Merak 7B v4	18.10%
Cendol mT5 Small	1.18%
Cendol mT5 Base	5.45%
Cendol mT5 Large	3.68%
Cendol mT5 XL	4.39%
Cendol mT5 XXL	2.19%
Cendol LLaMA2 7B	**25.30%**
Cendol LLaMA2 13B	22.20%

Table 8: Evaluation of Cendol and benchmark LLMs on ImplicitHate (higher means less hateful). The overall least harmful model is denoted by <u>underline</u>, while within the group, they are denoted by **bold**.

C Human Evaluation Guidelines

We adopt the human evaluation and annotation guidelines from prior works (Wu et al., 2023; Li

[8] facebook/nllb-200-distilled-1.3B

et al., 2023). We provide the detailed human evaluation guideline in Figure 9.

Rating	Description
Rate A	• Valid, acceptable and satisfying (subject to the annotator) response; • Accurate in terms of facts, yet comparable to human standards; • The response meets the required criteria, but it may not be in the expected format.
Rate B	• The response is acceptable but has minor errors that can be improved; • Minor errors include out-of-context content, minimal factual errors, partially responding to the instruction, etc.
Rate C	• The response is relevant and responds to the instruction, but it has significant errors in the content.
Rate D	• Invalid and unacceptable response.

Figure 9: Human annotation guideline that is incorporated in our human evaluation of Cendol.

D Pareto-Efficiency of Cendol Models

We showcase the Pareto-efficiency curve of the Cendol mT5 models in Figure 10.

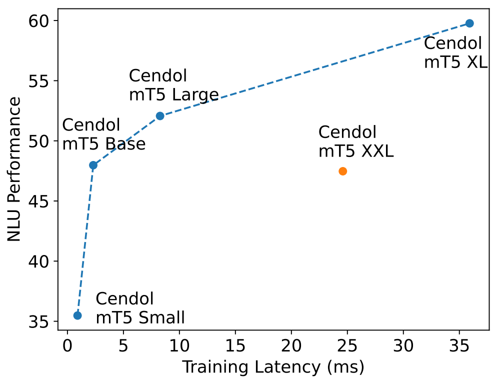

Figure 10: Pareto-efficiency curve of Cendol mT5 models. Parameter efficient method leads to a non-pareto optimal point as shown in the case of Cendol mT5$_{XXL}$.