Seed-Free Synthetic Data Generation Framework for Instruction-Tuning LLMs: A Case Study in Thai

Parinthapat Pengpun[‡], **Can Udomcharoenchaikit**[†],
Weerayut Buaphet[†], **Peerat Limkonchotiwat**[†]
[‡]Bangkok Christian International School, Thailand
[†]School of Information Science and Technology, VISTEC, Thailand
parinzee@protonmail.com
{canu_pro,weerayut.b_s20,peerat.l_s19}@vistec.ac.th

Abstract

We present a synthetic data approach for instruction-tuning large language models (LLMs) for low-resource languages in a data-efficient manner, specifically focusing on Thai. We identify three key properties that contribute to the effectiveness of instruction-tuning datasets: fluency, diversity, and cultural context. We propose a seed-data-free framework for generating synthetic instruction-tuning data that incorporates these essential properties. Our framework employs an LLM to generate diverse topics, retrieve relevant contexts from Wikipedia, and create instructions for various tasks, such as question answering, summarization, and conversation. The experimental results show that our best-performing synthetic dataset, which incorporates all three key properties, achieves competitive performance using only 5,000 instructions when compared to state-of-the-art Thai LLMs trained on hundreds of thousands of instructions. Our code and dataset are publicly available at https://github.com/parinzee/seed-free-synthetic-instruct.

1 Introduction

Large Language Models (LLMs) have achieved a near human-level of performance across multitudes of tasks and domains (OpenAI et al., 2024; Team et al., 2024; Ma et al., 2024; Antaki et al., 2023). However, many evaluation results have shown that this level of performance is often limited to high-resource languages only, with inconsistent levels of performance for lower-resource languages, i.e., Thai (Xue et al., 2024; Zhang et al., 2023; Krause et al., 2023; Huang et al., 2023; Ahuja et al., 2023). The development of LLMs for low-resource languages is crucial for enabling impactful applications for millions of people worldwide. Some applications of these LLMs include medical chatbots (Sanna et al., 2024), intelligent tutoring systems (Sonkar et al., 2023; Afzal et al., 2019), and content moderation tools that could help combat misformation and hate speech (Kumar et al., 2024). These potential applications have motivated researchers to explore methods for improving LLM performance in low-resource languages.

Recently, researchers developed fine-tuning techniques to improve the performance of LLMs in Thai as well. SambaLingo (Csaki et al., 2024) investigated how performing continual pretraining and instruction-tuning on machine-translated English datasets results in good performance in multiple low-resource target languages including Thai. WangchanX (Phatthiyaphaibun et al., 2024) explored using pre-existing Thai datasets to perform instruction-tuning and adapt the SEA-LION model (Singapore, 2023) to the Thai language. Typhoon-Instruct (Pipatanakul et al., 2023) adapted LLaMa-3 (AI@Meta, 2024) to the Thai language through continual pretraining on a filtered web corpus, and instruction-tuning on a combination of machine-translated datasets and Thai synthetic datasets generated with Self-Instruct (Wang et al., 2023). These methods typically use over 50k and sometimes over 100k examples for their instruction-tuning process, making it very costly. Additionally, some of these approaches involve continual pretraining, which further increases the cost and complexity of the model development process.

It remains unclear whether such large datasets are truly necessary for achieving high performance in low-resource languages, as the aforementioned works in Thai LLMs do not address this. However, other works have also shown that LLM alignment in English does not require extensively large datasets (Zhou et al., 2023; Du et al., 2023).

Thus, by carefully designing a high-quality synthetic dataset tailored to the target language, we hypothesize that it may be possible to achieve similar performance improvements in Thai while significantly reducing the data requirements and associated costs.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 438–457
August 11-16, 2024 ©2024 Association for Computational Linguistics

To formulate a high-quality synthetic dataset, we identify three key properties that the datasets used to finetune the current Thai LLMs have:

(1) **Fluency**: the data is grammatically correct and natural-sounding, enabling the model to learn the proper structure and flow of the language.

(2) **Diversity**: the data consists of a wide range of topics and domains, allowing it to generalize better to various downstream tasks.

(3) **Cultural Context**: the data contains instructions and information relating to the culture and beliefs appropriate for the average person from the country of the target language.

These properties are commonly present within the dataset used to train these models. Fluency is inherently present in the way that humans write and thus is within the human-annotated Han Instruct Dataset (Phatthiyaphaibun, 2024) used to train WangchanX. Diversity comes from the fact that existing datasets commonly cover multiple domains. For example, SambaLingo uses a translated version of UltraChat, which covers a wide range of topics. Cultural context is also present in these datasets. For example, Iapp Wiki QA (Viriyayudhakorn and Polpanumas, 2021)— a subset of OpenThaiGPT's training dataset— includes questions on Thai Wikipedia data. We hypothesize that combining all three properties in a dataset will yield a reasonably performant Thai LLM, even if the dataset is synthetic.

To verify our hypothesis, in this paper, we develop a framework to generate synthetic instruction-tuning datasets with controllable parameters for each of these properties. We use our framework to create five datasets with varying combinations of the properties as detailed in Section 4.1. We then perform instruction-tuning with the base model of LLaMa-3 8B (AI@Meta, 2024) on each dataset and evaluate their performance on two benchmarks: culture-specific and non-culture-specific datasets.

Our findings suggest that incorporating all three properties in the training data improves the performance of LLMs in low-resource languages, and using only just 5k rows of our dataset for instruction-tuning allows for similar performance against other methods that used 10-100x larger datasets.

We summarize the contribution of our work as follows:

- We verify our hypothesis that comparable results to current SOTA Thai LLMs can be achieved by carefully constructing a synthetic dataset that is a fraction of the size of the ones used to train these models.
- We propose a seed-data-free framework for synthetically generating finetuning data that is fluent, diverse, and culturally aligned for low-resource languages.
- We conduct a large-scale study on data efficiency using 8 LLMs, 5 synthetic datasets, 2 benchmarks, and 7 tasks.

2 Related Works

2.1 Thai LLMs

The development of Thai LLMs and other low-resource language LLMs (Csaki et al., 2024; Singapore, 2023; Nguyen et al., 2023) have gained attention in the recent year with models such as LLaMa3-8b-WangchanX-sft-Demo (Phatthiyaphaibun et al., 2024), Typhoon-Instruct (Pipatanakul et al., 2023), and OpenThaiGPT (OpenThaiGPT, 2023) being released. LLaMa3-8b-WangchanX-sft-Demo leverages a combination of English Datasets: Dolly-15 (Conover et al., 2023), Math-14k (Hu et al., 2023); Human-written Thai datasets (6k); and a Google Gemini (Team et al., 2024) translated versions of Dolly-15k and Math-14k for instruction-tuning, which results in a total of 64k examples. Typhoon-Instruct uses both continual pretraining on a filtered subset of Oscar (Ortiz Suárez et al., 2020) and finetuned on multiple translated datasets. However, they do not mention the exhaustive list nor the exact number used.

OpenThaiGPT also performs both continual pretraining and instruction-tuning. Although they do not explicitly mention the dataset composition nor count for the current version (v1.0.0). Previous versions, however, used an extensively large corpus for finetuning consisting of both human-generated and machine-translated data. For example, `openthaigpt-0.1.0-beta` used a combination of 200k samples, and `openthaigpt-gpt2-instructgpt-poc-0.0.1` used 300k samples [1]. These previous versions used the GPT-2 architecture (Radford et al., 2019) with 1.5B Parameters. For their latest version, they scaled up the model to LLaMa3-8B and LLaMa3-70B [2].

While current works in Thai LLM development have focused on scaling the quantity of the dataset

[1] Information regarding dataset composition is obtained from OpenThaiGPT's Github

[2] Information regarding architecture obtained from model cards in OpenThaiGPT's HuggingFace

and model size, our work aims to improve the performance of Thai LLMs from a data-centric perspective, thereby reducing the costs needed for fine-tuning a model. Furthermore, our method does not rely on continual pretraining, which also reduces the required computation and time required as well.

2.2 Synthetic Data Generation for LLMs

LLMs are dependent on a large number of high-quality datasets in order to achieve good performance (Longpre et al., 2023). Traditionally, instruction-tuning datasets were created by human annotators, which is costly and time-consuming. Synthetic dataset generation has emerged as a promising approach to address these limitations.

Self-Instruct (Wang et al., 2023) used 175 human-generated instructions as a seed, then prompted GPT to generate unique instructions and tasks, resulting in a dataset consisting of 82k samples. WizardLM (Xu et al., 2023) proposed improving LLMs by synthetically generating complex and difficult questions using prompt engineering to increase the difficulty of an instruction or generate a completely new instruction in the same domain as a given instruction. In addition, WizardLM uses Alpaca's training data as the initial data and applies the pipeline, which results in a total of 250k samples of instruction-tuning data. UltraChat (Ding et al., 2023) proposes a method for generating a large-scale multi-turn dialogue dataset for instruction-tuning. UltraChat obtains context data using various techniques, such as utilizing meta-information from Wikidata and search engines, extracting material types from web pages in the C4 corpus, and prompting GPT-3 to generate instructions for different types of writing tasks. This information is then used to perform iterative prompting between two ChatGPT models to simulate user-assistant interactions. Experimental results from these works have demonstrated that synthetic data can improve the performance of LLMs without extensive human effort.

Despite the promising results achieved by existing synthetic data approaches, there remains a gap in the literature regarding the application of these techniques to low-resource languages. Furthermore, these works also utilize high-quality seed instructions, which may be difficult to obtain in low-resource settings. We address this gap by proposing a seed-data-free pipeline for generating instruction-tuning data for low-resource languages.

2.3 Data Efficient Instruction-Tuning for LLMs

Training large language models (LLMs) often requires extensive data, posing challenges for low-resource languages due to dataset scarcity and high computational costs. Researchers have explored techniques for efficient instruction-tuning with limited data.

Zhou et al. (2023) explored constructing a high-quality 1000 sample instruction-tuning data using data from StackExchange, Wikihow, and other online sources. To ensure diversity, the dataset also includes human-annotated instructions as well. In human evaluations, the LIMA model trained on this dataset was found to produce outputs that were strictly preferred to or on par with those from GPT-4 in 43% of cases, Claude in 46% of cases, Bard in 58% of cases, and InstructGPT (DaVinci003) in 65% of cases. Through ablation studies, they also found that data diversity and quality were more important than quantity for improving the model's performance, as doubling the dataset quantity alone did not contribute to performance increases. Models trained on more diverse data from StackExchange outperformed those trained on a larger quantity of homogeneous data from wikiHow, and models trained on quality-filtered data outperformed those trained on unfiltered data.

Du et al. (2023) propose a model-oriented data selection (MoDS) approach for efficiently selecting valuable instruction data to fine-tune an LLM. Their method considers instruction quality, coverage, and necessity based on the abilities of the specific target LLM. First, they use a quality evaluation model to filter the original dataset for high-quality instructions. Then, they apply a k-center greedy algorithm to select a maximally diverse seed dataset from this filtered set. The model is initially fine-tuned on this seed data, then further refined with an augmented dataset addressing performance gaps. The final fine-tuning is done on the combination of the seed and augmented data. An LLM fine-tuned with 4,000 MoDS-selected examples outperformed a model trained on the full 214k dataset.

Although these works have shown success in English, there is a lack of literature regarding data-efficient training in low-resource languages. Our framework addresses this gap by providing empirical evidence that using a small but high-quality synthetic dataset can result in competitive performance for an LLM in the Thai language.

3 Synthetic Dataset Generation Pipeline

3.1 Overview

Based on the literature review, current Thai LLMs use extensively large scale datasets. However, we hypothesize that it may be possible to create an LLM model that is comparable to existing Thai LLMs while only using a fraction of the SFT data for finetuning. To verify our hypothesis, we construct and train on 5 synthetic datasets with varying combinations of the three aforementioned properties: Fluency, Diversity, and Cultural Context. As shown in Figure 1, our synthetic datasets are generated through our framework as follows. The pipeline uses an LLM, in this case, Claude-3 Haiku, to first randomly generate a given number of topics that either are general topics or relate to a specific culture. Using the topics, we search Wikipedia for a related text and then prompt Haiku to generate instructions related to that text. Our pipeline generates instructions in the target language (Thai) directly for 4 tasks: Closed Question Answering (Closed QA), Summarization, Conversation, and Multiple Choice. The data then goes through a diversity control step, where we filter out closely related samples using their semantic embedding vectors to ensure a high-diversity dataset. We perturb the configuration of the pipeline to obtain the 5 synthetic datasets.

3.2 Model Selection

We choose Claude-3 Haiku as the LLM for our synthetic data generation pipeline for several reasons. First, it has demonstrated strong performance across various natural language tasks (Anthropic, 2024), making it well-suited for generating high-quality instruction data. Second, it is relatively low cost for model of its performance when compared to other state-of-the-art models, this allows for a more cost-efficient dataset generation process. Third, from our observation, Claude-3 Haiku produced Thai output that is much more fluent and coherent than other LLMs in a similar price range, such as GPT-3.5-Turbo. This aligns with Enis and Hopkins (2024), which has shown that Claude-3 is a strong translator, indicating a good multilingual understanding. Claude-3's tokenizer is also more efficient in tokenizing Thai characters when compared to GPT-3.5's tokenizer (Claude's tokenizer uses fewer tokens for Thai text).

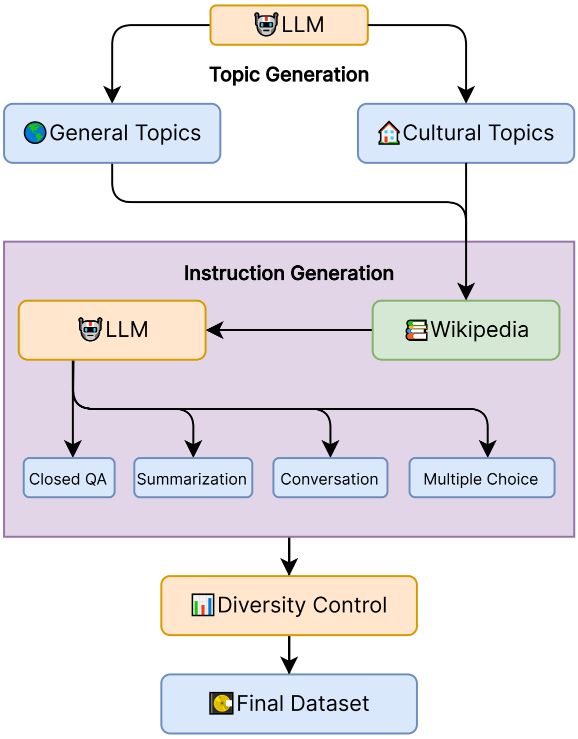

Figure 1: Our proposed framework for generating synthetic instruction-tuning datasets for low-resource languages from scratch with fluency, diversity, and cultural context.

3.3 Topic Generation

We separate our categorization of topics into 2 categories: General Topics and Cultural Topics. For both of these categories, we use a temperature of 0.95. We then prompt Haiku to randomly generate these topics; the specific prompts are below. We repeat this process until we obtain the desired amount of topics. Afterward, the topics are filtered for duplicates and removed.

General Topics Prompt:

> *Please generate 20 completely random topics. These can be about absolutely anything from everyday conversation, advice, random thoughts, mathematics, science, history, philosophy, etc. Each topic should be a short phrase or sentence. Ensure your output is in the format of a list of strings, where each string is a topic. Your output should be one line in the aforementioned format without anything else.*

Cultural Topics Prompt:

> *You are a native Thai person with expert knowledge of Thai culture, history,*

> *language, and customs. Ensure that everything you act, do, say, and generate matches with this fact. Please generate 20 completely random topics relating to your culture. These can be about anything related to your culture such traditions, history, food, language, etc. Each topic should be...*

The rest of this prompt is omitted as it is the same as the General Topics Prompt.

3.4 Instruction Generation

Context Selection/Generation Given a topic, the pipeline first randomly chooses whether to select a related context from Wikipedia or not. If a random context from Wikipedia is not chosen, we prompt Haiku to generate a context related to the topic based on a randomly selected style from this list: news article, blog post, text messages, fictional short story, video transcript, song, poem, scientific study, medical report, social media post with replies, email, tweet, or a how-to article. Otherwise, we search Wikipedia through its MediaWiki API using the topic as our query. We find the top 10 most similar articles to the topic and randomly pick one. We split each article based on their sections, and each of those serves as one context for the instruction.

The following describes our goal and hyperparameters in prompting Haiku to generate instructions for each of these tasks. The full prompt for each task is described in Appendix A.

Closed Question Answering. For this task, we take the context from the previous step and prompt Haiku to generate 5 question-answer pairs for each context. We emphasize our prompting to ensure that Haiku generates questions that come from roughly throughout the whole context. Furthermore, we also noticed that sometimes, Haiku would use “common knowledge” that is assumed to be known when generating answers to its questions. To alleviate this, we also emphasize not using any “external information” in our prompt. We use a temperature of 0.35 for this task.

Summarization. We take the context from the previous step and prompt Haiku to generate a summary for context. The summary is generated in one of three styles, which is randomly picked and embedded into the prompt: bullet points, paragraphs, or numbered lists. We use a temperature of 0.35 for this task.

Conversation. This task does not require any context and is designed to mimic how a human might talk to a chatbot— hence the name of the task is called “Conversation.” We prompt Haiku to generate a random conversation between an AI assistant and human that relates to a given topic. We emphasize that the assistant must maintain a friendly and casual conversation. We use a temperature of 0.8 for this task.

Multiple Choice. We use the context from the previous step and prompt Haiku to generate a question regarding the context and possible answer choices (with only 1 correct answer). Please note that we later also shuffled the answer choices as we noticed that Haiku has a tendency to put correct answer choices as the first one. Because we later do this, we also prompt Haiku to not use any ordinal information in the answer choices, i.e., “the first and third choice” or “B and D”. We use a temperature of 0.4 for this task.

3.5 Diversity Control

Although we use relatively high temperatures for these tasks, there may still be cases where we get multiple samples of instructions that are quite similar. To ensure that our dataset is diverse, we filter out any samples that are closely related to each other semantically. We first use BGE-M3 (Chen et al., 2024) to encode all of the samples. BGE-M3 is chosen due to its exceptional Thai performance. The samples are formatted by concatenating the instruction, context, and output. For each sample, we do an approximate nearest neighbor search across the whole dataset. If the cosine similarity of the nearest match of that sample is over 0.95, we remove that sample. This process ensures that our final dataset is sufficiently diverse.

4 Experimental Setting

4.1 Training Datasets

We generate 5 synthetic datasets with varying combinations of fluency, diversity, and cultural contexts using our pipeline to demonstrate that all 3 properties are required for a high-performance model. Each of these datasets has 5,000 samples of instructions. This number is similar to other works in other languages (e.g., LIMA (Zhou et al., 2023) used 1,000 samples, and MoDS (Du et al., 2023) used 4,000 samples).

- **Fluency + Cultural Context + Diversity (F+C+D+):** Constructed by running the pipeline fully with quality control using 750 randomly generated topics in total (400 cultural and 300 general). This dataset is generated in Thai directly by our pipeline and is not translated.

- **Fluency Only:** Constructed by running the pipeline with only 10 randomly generated topics (general topics only) and without any diversity control to reduce overall diversity. Only general topics were used to ensure no cultural context.
- **Diversity Only:** Constructed by running the pipeline with diversity control using 750 general topics. To artificially reduce fluency, we use nllb-200-distilled-600M (Team et al., 2022) to translate all samples to English and back-translate them to Thai again. This effectively simulates using machine translation to translate an English dataset to Thai. This dataset is constructed to demonstrate the impacts of not having fluency or cultural context.
- **Cultural Context Only:** Constructed by using the F+ C+ D+ dataset as a basis. We randomly select 1000 samples; then, we use NLLB to translate them into English. We then use QCPG (Bandel et al., 2022) to paraphrase the dataset. For each sample, we perform 4 paraphrases, resulting in a total count of 5,000 (4,000 paraphrases + 1,000 originals), thereby reducing the overall domain diversity. Then, we translate everything back to Thai again, reducing fluency.
- **No Properties:** We randomly select 1,000 rows from the UltraChat-200k dataset (no Thai cultural context). We use QCPG to perform paraphrasing— generating 4 paraphrases for each sample (reduce diversity), resulting in a total count of 5,000. Then, we translate everything to Thai using NLLB (reduce fluency).

4.2 Models

We perform instruction finetuning of the base version of Llama-3 8B using these datasets with QLoRa on a single RTX 3090. The total amount of GPU hours used is around 80 hours. Hyperparameters are shown in Table 1. In addition to our own models, we also evaluate standard Thai LLMs, such as Typhoon-Instruct-v1.5 8B, OpenThaiGPT-v1.0.0 8B, and LLaMa3-8b-WangchanX-sft-Demo.

4.3 Evaluation

We use WangchanThaiInstruct [3] as our benchmark as it provides both a Thai *culture-specific* version and a *non-culture-specific* version. It consists of 6,287 samples in total (both versions combined) spanning 3 domains: Legal, Medical, and Finance. The dataset is created and quality assured by human annotators during the whole process. Using this dataset allows us to assess the performance of our LLMs on instructions that require an understanding of Thai culture, as well as instructions that are more general in nature.

Hyperparameter	Value
Load in 4-bit	True
Sequence Length	4000
Adapter Type	QLoRA
LoRA Rank	32
LoRA Alpha	16
LoRA Dropout	0.05
LoRA Target Linear	True
Grad. Accum. Steps	8
Micro Batch Size	1
Number of Epochs	3
Optimizer	Paged AdamW 8bit
Learning Rate	0.00015
BF16 Precision	True
Grad. Checkpointing	True
Flash Attention	True
Warmup Ratio	0.5
Evals per Epoch	1
Saves per Epoch	1
Weight Decay	0.0
Seed	42

Table 1: Hyperparameters used to finetune our models.

Both versions of the benchmark consist of seven tasks in total: **Brainstorming**, which evaluates the model's ability to generate creative ideas and solutions based on a given prompt or scenario; **Classification**, which requires the model to assign a given input to one or more predefined categories; **Closed QA**, where the model must locate relevant information within a given text to answer a question; **Creative Writing**, which assesses the model's ability to generate coherent, engaging, and creative pieces of writing based on a prompt or theme; **Open QA**, similar to Closed QA but with more open-ended questions that may not have a single, definitive answer within the provided text; **Multiple Choice**, where the model must select the most appropriate or correct answer from a set of options; and **Summarization**, which involves generating a concise and coherent summary of a given piece of text. By evaluating the Thai LLMs on these diverse tasks and domains, we can gain a comprehensive understanding of their performance across different aspects of language understanding and generation.

Metrics. We follow the WangchanX-10k's suggested metrics for evaluation. We use BLEU, METEOR, ChrF, ROUGE, and BERTScore to mea-

[3] https://huggingface.co/datasets/airesearch/WangchanThaiInstruct_7.24

Table 2: Average evaluation results across all 7 tasks on the Thai Culture and General Test Sets. F, C, and D denote Fluency, Culture, and Diversity, respectively. The plus sign (+) indicates the presence of the corresponding attribute, while the minus sign (-) indicates its absence.

Metric	**F- C- D-**	**F+ C- D-**	**F- C+ D-**	**F- C- D+**	**F+ C+ D+**	**WangchanX**	**Typhoon**	**OpenThai**
BERTScore	45.90	57.30	49.60	48.20	69.50	68.80	**74.10**	64.50
BLEU	0.02	0.01	0.00	0.00	0.10	2.24	**2.32**	0.95
ChrF	4.38	5.18	2.90	2.74	9.47	**17.28**	17.21	14.54
METEOR	2.20	3.70	1.70	1.70	6.70	11.30	**12.70**	8.20
ROUGE-1	1.30	3.60	1.80	1.90	7.70	13.40	**20.70**	12.20
ROUGE-2	0.20	1.00	0.20	0.30	3.30	5.80	**11.80**	5.60
ROUGE-L	1.20	3.60	1.80	1.90	7.50	12.60	**20.00**	11.70
ROUGE-Lsum	1.20	3.50	1.80	1.90	7.60	12.70	**20.00**	11.60
SQuAD F1	0.50	2.80	0.82	0.82	5.32	**8.30**	7.10	3.58

(Thai Culture Test Set)

Metric	**F- C- D-**	**F+ C- D-**	**F- C+ D-**	**F- C- D+**	**F+ C+ D+**	**WangchanX**	**Typhoon**	**OpenThai**
BERTScore	50.90	59.70	52.60	49.50	73.20	72.20	**76.50**	67.50
BLEU	0.04	0.01	0.00	0.00	0.08	2.08	**2.61**	0.88
ChrF	4.91	5.04	2.82	2.56	9.56	17.24	**17.53**	14.84
METEOR	2.50	3.50	1.70	1.60	6.70	11.10	**12.90**	8.40
ROUGE-1	1.50	3.00	1.50	1.60	6.60	14.00	**18.80**	10.40
ROUGE-2	0.40	1.10	0.20	0.30	2.70	7.10	**10.00**	4.40
ROUGE-L	1.50	3.00	1.50	1.60	6.40	13.30	**18.10**	9.90
ROUGE-Lsum	1.50	3.00	1.50	1.60	6.50	13.40	**18.00**	9.90
SQuAD F1	0.86	2.29	0.83	0.73	4.58	**7.43**	7.26	3.31

(General Test Set)

sure the performance in these tasks. However, we do note that the WangchanX-10k mentioned that BERTScore is the most reliable metric as it measures semantic similarity, while other traditional metrics yield inconclusive results.

5 Experimental Results

5.1 Main Results

Results. Table 2 presents the average evaluation results across all tasks on both the Thai Culture Test Set and the General Test Set. Our synthetic datasets are denoted by the presence (+) or absence (-) of three key attributes: Fluency (F), Culture (C), and Diversity (D). The best-performing model for each metric is highlighted in bold, while the second-best model is underlined. On the Thai Culture Test Set, our best-performing synthetic dataset, F+ C+ D+, which incorporates all three key attributes, achieves the second-highest BERTScore of 69.50%, surpassing WangchanX (68.80%) and OpenThaiGPT (64.50%). The Typhoon-Instruct model obtains the highest BERTScore of 74.1%. The results on the General Test Set follow a similar pattern, with F+ C+ D+ maintaining its second-place position in terms of BERTScore 73.20%, outperforming WangchanX 72.2% and OpenThaiGPT 67.50%. The Typhoon-Instruct model achieves the highest BERTScore of 76.5%. The full evaluation results for each task are listed in Appendix B.

Discussion. The experimental results demonstrate the effectiveness of our data-centric approach for improving the performance of Thai LLMs, particularly when considering the BERTScore metric, which is deemed the most reliable by the benchmark authors. F+ C+ D+ achieves the second-highest BERTScore on both the Thai Culture Test Set and the General Test Set, surpassing WangchanX and OpenThaiGPT, suggesting that our data generation pipeline is capable of producing high-quality data that can enhance the model's performance on a wide range of tasks, while being still data-efficient. Namely, we use only 5,000 samples of synthetic data to surpass OpenThaiGPT (200k samples + pretraining) and WangchanX (64k samples), both of which use a mix of human-annotated and machine-translated data.

The consistent top performance of F+ C+ D+ and the lower performances of other synthetic sets, which only consist of one property, demonstrates that all three properties are required to build a strong synthetic dataset. In conclusion, all these results verify our hypothesis that it is indeed possible to construct a small synthetic dataset that performs competitively against much larger datasets.

5.2 Error Analysis

In this study, we demonstrate error analysis across different tasks to decipher why our model performs

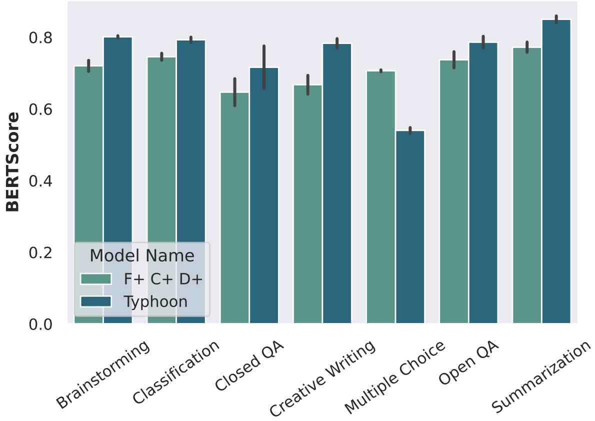

Figure 2: Comparison of BERTScores of our best synthetic model and Typhoon-Instruct on the average scores from both test sets. We also performed Wilcoxon rank-sum tests (Wilcoxon, 1945) comparing F+ C+ D+ against Typhoon-Instruct for each task on both the Thai culture-specific and general test sets, and found that the differences were statistically significant ($p < 0.05$) for all tasks, with an average Wilcoxon statistic of -6.512 and an average p-value of 0.00073 across all comparisons.

worse than current Thai LLMs in certain tasks. As evidenced in Figure 2, our model performs slightly lower than Typhoon-Instruct in some tasks. When we examine these tasks, it is evident that the tasks with the largest gaps are Brainstorming, Creative Writing, and Summarization.

After investigation, we discovered that our model has a tendency to produce shorter and more concise responses on average. This is shown in Figure 3. This could lead to it omitting some information that the reference includes. Hence this leads to a lower score on these open-ended tasks. Since this does not impact tasks that require short responses (i.e., Classification and Open QA), we can see that the difference between Typhoon-Instruct and F+ C+ D+ is much smaller. Furthermore, our model even beats Typhoon-Instruct in Multiple Choice by a large margin. The fact that our model produces shorter responses on average also explains why our model has lower scores when using evaluated n-gram based metrics, which effectively measure text overlap. We conjecture that our model's tendency for brevity stems from the fact that our synthetic data pipeline currently only generates short single-turn dialogues. However, other Thai LLMs, such as Typhoon-Instruct, are trained on machine-translated versions of long multi-turn dialogue datasets like UltraChat.

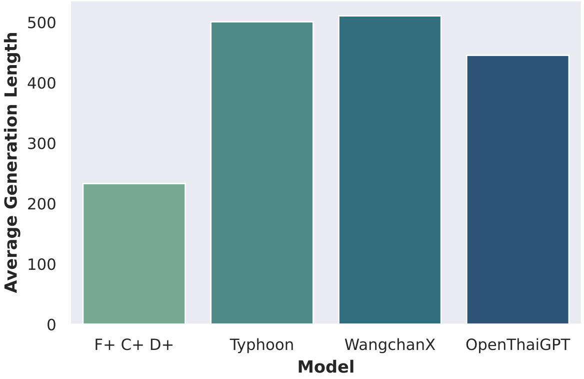

Figure 3: Comparison of average generation lengths across all tasks and both benchmarks. A Wilcoxon rank-sum test was conducted to compare the generation lengths of our best model (F+ C+ D+) and Typhoon-Instruct. The results showed a statistically significant difference ($W = -54.233$, $p < 0.00001$), indicating that our model generates significantly shorter outputs compared to Typhoon-Instruct.

6 Conclusion and Future Work

In conclusion, this study demonstrates the effectiveness of a data-centric approach for improving the performance of large language models in Thai, a low-resource language. We identified three key properties that lead to well-performing Thai LLMs: fluency, diversity, and cultural context. We proposed a seed-data-free framework for generating high-quality instruction-tuning data that incorporates these properties. Experiments conducted across multiple models, synthetic datasets, benchmarks, and tasks provide empirical evidence that it is possible to achieve competitive results compared to state-of-the-art Thai LLMs trained on 10-100x larger datasets. While our model tends to generate more concise responses compared to the top-performing Typhoon-Instruct model, impacting its performance on open-ended generative tasks, it still achieves impressive results overall, beating other models such as OpenThaiGPT-v1.0.0 and achieving comparable results to WangChanX LlaMa3-8B SFT Demo.

For future work, there are several promising directions to explore. One important avenue is to extend our pipeline to generate multi-turn dialogue datasets, which can help alleviate the length issues observed in the current study and enhance the model's ability to handle more realistic, conversation-based scenarios. Additionally, conducting experiments with a stronger base model, such as upgrading from Claude-3 Haiku to a more advanced model, could potentially yield even better performance without requiring significant modifications to the data generation process. To assess the generalizability of our approach, it would be valuable to expand our experiments to other low-resource languages, adapting the framework to handle different linguistic properties and cultural contexts.

Acknowledgements

We extend our sincere gratitude to Potsawee Manakul for his invaluable assistance during the early stages of this project.

This research has received funding support from the NSRF via the Program Management Unit for Human Resources & Institutional Development, Research and Innovation Grant Number B46G670083.

Limitations

Our limitation in this paper is we did not investigate the optimal combination of synthetic and human-generated data that could provide insights into the most effective data composition strategies. This could involve comparing the performance of models trained solely on synthetic data with those trained on a combination of synthetic and carefully filtered human-generated data. In addition, conducting extensive human evaluations would be crucial for assessing the practical usability and perceived quality of the generative models.

References

Shazia Afzal, Tejas Dhamecha, Nirmal Mukhi, Renuka Sindhgatta, Smit Marvaniya, Matthew Ventura, and Jessica Yarbro. 2019. Development and deployment of a large-scale dialog-based intelligent tutoring system. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 114–121, Minneapolis, Minnesota. 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 *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

AI@Meta. 2024. Llama 3 model card.

Fares Antaki, Daniel Milad, Mark A Chia, Charles-Édouard Giguère, Samir Touma, Jonathan El-Khoury, Pearse A Keane, and Renaud Duval. 2023. Capabilities of gpt-4 in ophthalmology: an analysis of model entropy and progress towards human-level medical question answering. *British Journal of Ophthalmology*.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku.

Elron Bandel, Ranit Aharonov, Michal Shmueli-Scheuer, Ilya Shnayderman, Noam Slonim, and Liat Ein-Dor. 2022. Quality controlled paraphrase generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 596–609, Dublin, Ireland. Association for Computational Linguistics.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

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.

Zoltan Csaki, Bo Li, Jonathan Li, Qiantong Xu, Pian Pawakapan, Leon Zhang, Yun Du, Hengyu Zhao, Changran Hu, and Urmish Thakker. 2024. Sambalingo: Teaching large language models new languages. *Preprint*, arXiv:2404.05829.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051, Singapore. Association for Computational Linguistics.

Qianlong Du, Chengqing Zong, and Jiajun Zhang. 2023. Mods: Model-oriented data selection for instruction tuning. *Preprint*, arXiv:2311.15653.

Maxim Enis and Mark Hopkins. 2024. From llm to nmt: Advancing low-resource machine translation with claude. *Preprint*, arXiv:2404.13813.

Zhiqiang Hu, Nancy Chen, and Roy Lee. 2023. Adapter-TST: A parameter efficient method for multiple-attribute text style transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 693–703, Singapore. Association for Computational Linguistics.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.

Lea Krause, Wondimagegnhue Tufa, Selene Baez Santamaria, Angel Daza, Urja Khurana, and Piek Vossen. 2023. Confidently wrong: Exploring the calibration and expression of (un)certainty of large language models in a multilingual setting. In *Proceedings of the Workshop on Multimodal, Multilingual Natural Language Generation and Multilingual WebNLG Challenge (MM-NLG 2023)*, pages 1–9, Prague, Czech Republic. Association for Computational Linguistics.

Deepak Kumar, Yousef AbuHashem, and Zakir Durumeric. 2024. Watch your language: Investigating content moderation with large language models. *Preprint*, arXiv:2309.14517.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. 2023. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, toxicity. *Preprint*, arXiv:2305.13169.

Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2024. Eureka: Human-level reward design via coding large language models. *Preprint*, arXiv:2310.12931.

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

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenThaiGPT. 2023. Openthaigpt 7b 1.0.0-beta. https://openthaigpt.aieat.or.th/. Released.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Wannaphong Phatthiyaphaibun. 2024. Han instruct dataset.

Wannaphong Phatthiyaphaibun, Surapon Nonesung, Patomporn Payoungkhamdee, Peerat Limkonchotiwat, Can Udomcharoenchaikit, Jitkapat Sawatphol, Chompakorn Chaksangchaichot, Ekapol Chuangsuwanich, and Sarana Nutanong. 2024. Wangchanlion and wangchanx mrc eval. *Preprint*, arXiv:2403.16127.

Kunat Pipatanakul, Phatrasek Jirabovonvisut, Potsawee Manakul, Sittipong Sripaisarnmongkol, Ruangsak Patomwong, Pathomporn Chokchainant, and Kasima Tharnpipitchai. 2023. Typhoon: Thai large language models. *Preprint*, arXiv:2312.13951.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Leonardo Sanna, Patrizio Bellan, Simone Magnolini, Marina Segala, Saba Ghanbari Haez, Monica Consolandi, and Mauro Dragoni. 2024. Building certified medical chatbots: Overcoming unstructured data limitations with modular RAG. In *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024*, pages 124–130, Torino, Italia. ELRA and ICCL.

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.

Shashank Sonkar, Naiming Liu, Debshila Mallick, and Richard Baraniuk. 2023. CLASS: A design framework for building intelligent tutoring systems based on learning science principles. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1941–1961, Singapore. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen

Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck,

Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Kobkrit Viriyayudhakorn and Charin Polpanumas. 2021. iapp_wiki_qa_squad.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. *Preprint*, arXiv:2212.10560.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *Preprint*, arXiv:2304.12244.

Boyang Xue, Hongru Wang, Weichao Wang, Rui Wang, Sheng Wang, Zeming Liu, and Kam-Fai Wong. 2024. A comprehensive study of multilingual confidence estimation on large language models. *Preprint*, arXiv:2402.13606.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927, Singapore. Association for Computational Linguistics.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment. *Preprint*, arXiv:2305.11206.

A Prompts for Each Task in Instruction Generation

Closed Question Answering:

> *Generate 5 questions focusing on different aspects / parts of this given context. Use only the given context to create your questions. Do not use external information. <context>[context]</context> Ensure your output is in the format of a list of dictionaries, where each dictionary contains a 'question' key and an 'answer' key. Your output should be one line in the aforementioned format without anything else.*

Summarization:

> *Generate a concise summary in [summary style] format of the following context related to [topic]: <context> [context] </context> Ensure your output is in the format of a dictionary with a 'summary' and 'instruction' key, where 'summary' is your summary in the specified format and 'instruction' is a sentence you would instruct someone to get this summary (for example: 'Please summarize in [summary style] format the following text passage'). Your output should be one line in the aforementioned format, and in the correct language without anything else.*

Conversation:

> *Generate a conversation between a user and an AI assistant on the topic of [topic]. The user's message should be a question or a statement related to [topic], and the AI assistant should provide a relevant, engaging response to maintain a friendly and casual conversation. The output should be in the following format: <format>Input: User's message Output: AI assistant's response</format> Ensure your output contains ONLY ONE input-output pair exactly in the specified format without any additional text.*

Multiple Choice:

> *Generate a multiple-choice question focusing on the given context. The question should only have one correct choice. Use only the given context to create your question and answer choices. Do not use external information. <context>[context]</context> DO NOT USE any ordinal information (DO NOT USE eg: first answer is correct, all of the above is correct, etc) of the choices to answer your question as the choices will be shuffled later. Ensure your output is in the following format:<format> Question: Your question Choices: - [Choice 1] - [Choice 2] - [Choice 3] - [Choice 4] Answer: [Explaination + Reasoning + Correct Answer (in this order exactly)] </format> Your output should contain ONLY ONE multiple-choice question exactly in the specified format without any additional text.*

B Full Evaluation Results For Every Task

- **Brainstorming**: Table 3
- **Classification**: Table 4
- **Closed Question Answering**: Table 5
- **Creative Writing**: Table 6
- **Multiple Choice**: Table 7
- **Open Question Answering**: Table 8
- **Summarization**: Table 9

Table 3: Average evaluation results for the Brainstorming task on the Thai Culture and General Test Sets. F, C, and D denote Fluency, Culture, and Diversity, respectively. The plus sign (+) indicates the presence of the corresponding attribute, while the minus sign (-) indicates its absence.

Metric	**F- C- D-**	**F+ C- D-**	**F- C+ D-**	**F- C- D+**	**F+ C+ D+**	**WangchanX**	**Typhoon**	**OpenThai**
BERTScore	54.63	55.38	51.58	53.41	70.48	68.99	**80.43**	71.76
BLEU	0.00	0.00	0.00	0.00	0.04	0.39	1.20	**2.13**
ChrF	5.03	3.92	3.04	3.52	9.54	14.51	18.41	**22.42**
METEOR	2.59	2.04	1.54	1.80	5.68	7.71	10.98	**11.43**
ROUGE-1	2.73	1.84	1.55	0.95	6.20	12.03	21.96	**22.01**
ROUGE-2	0.48	0.11	0.00	0.00	2.77	4.45	**11.42**	11.13
ROUGE-L	2.84	1.80	1.56	0.96	6.04	11.41	20.78	**21.45**
ROUGE-Lsum	2.82	1.87	1.52	1.01	6.00	11.44	20.58	**21.59**

(Thai Culture Test Set)

Metric	**F- C- D-**	**F+ C- D-**	**F- C+ D-**	**F- C- D+**	**F+ C+ D+**	**WangchanX**	**Typhoon**	**OpenThai**
BERTScore	55.80	58.46	55.55	58.54	73.55	71.76	**79.79**	71.64
BLEU	0.00	0.00	0.00	0.00	0.02	0.50	1.11	**1.33**
ChrF	5.14	4.27	3.15	3.94	9.29	16.38	17.89	**19.85**
METEOR	2.65	2.45	1.75	2.28	5.60	8.56	**11.26**	10.74
ROUGE-1	1.86	1.91	0.75	2.06	4.66	15.37	**21.08**	17.72
ROUGE-2	0.21	0.30	0.10	0.52	1.59	7.19	**11.35**	8.42
ROUGE-L	1.82	1.91	0.75	1.99	4.54	14.80	**19.93**	16.67
ROUGE-Lsum	1.81	1.92	0.74	2.00	4.52	14.84	**20.02**	16.72

(General Test Set)

Table 4: Average evaluation results for the Classification task on the Thai Culture and General Test Sets. F, C, and D denote Fluency, Culture, and Diversity, respectively. The plus sign (+) indicates the presence of the corresponding attribute, while the minus sign (-) indicates its absence.

Metric	**F- C- D-**	**F+ C- D-**	**F- C+ D-**	**F- C- D+**	**F+ C+ D+**	**WangchanX**	**Typhoon**	**OpenThai**
BERTScore	53.22	62.66	52.55	52.28	73.55	73.36	**78.52**	66.44
BLEU	0.00	0.00	0.00	0.00	0.01	**0.85**	0.16	0.31
ChrF	4.62	3.85	2.41	2.74	7.68	**14.88**	11.76	12.07
METEOR	2.33	2.96	1.33	1.70	5.32	**8.72**	8.25	6.60
ROUGE-1	1.05	2.46	1.34	1.03	4.58	5.70	**7.57**	3.78
ROUGE-2	0.07	0.59	0.04	0.19	1.59	1.14	**2.37**	1.00
ROUGE-L	1.01	2.33	1.33	1.01	4.45	5.16	**7.35**	3.50
ROUGE-Lsum	0.99	2.32	1.33	1.00	4.43	5.21	**7.36**	3.51
SQuAD F1	0.52	2.49	0.64	0.95	3.47	**3.96**	3.51	2.04

(Thai Culture Test Set)

Metric	**F- C- D-**	**F+ C- D-**	**F- C+ D-**	**F- C- D+**	**F+ C+ D+**	**WangchanX**	**Typhoon**	**OpenThai**
BERTScore	60.33	64.61	56.39	54.93	75.55	75.70	**80.03**	71.11
BLEU	0.01	0.00	0.00	0.00	0.01	0.33	0.16	**0.61**
ChrF	5.19	3.40	2.75	2.60	8.03	12.52	11.55	**13.78**
METEOR	2.78	2.78	1.65	1.70	5.55	7.52	**8.18**	8.11
ROUGE-1	1.29	2.35	0.61	1.27	4.88	6.19	**8.12**	6.48
ROUGE-2	0.20	0.72	0.07	0.31	1.38	2.30	**3.19**	2.21
ROUGE-L	1.29	2.33	0.60	1.24	4.88	5.96	**7.96**	6.40
ROUGE-Lsum	1.26	2.32	0.61	1.27	4.90	5.93	**7.98**	6.35
SQuAD F1	1.01	2.46	0.96	1.11	3.03	3.67	**3.81**	3.30

(General Test Set)

Table 5: Average evaluation results for the Closed QA task on the Thai Culture and General Test Sets. F, C, and D denote Fluency, Culture, and Diversity, respectively. The plus sign (+) indicates the presence of the corresponding attribute, while the minus sign (-) indicates its absence.

Metric	**F- C- D-**	**F+ C- D-**	**F- C+ D-**	**F- C- D+**	**F+ C+ D+**	**WangchanX**	**Typhoon**	**OpenThai**
BERTScore	18.18	48.22	42.27	36.98	60.91	61.49	**65.67**	44.32
BLEU	0.01	0.00	0.00	0.00	0.00	**1.20**	1.09	0.00
ChrF	1.97	3.67	2.08	1.25	7.51	14.62	**15.12**	4.40
METEOR	0.98	2.44	1.35	0.98	6.91	12.52	**13.89**	2.92
ROUGE-1	0.42	2.87	1.39	0.63	10.94	16.86	**19.74**	5.10
ROUGE-2	0.29	0.68	0.17	0.00	5.99	10.45	**12.50**	2.02
ROUGE-L	0.42	2.80	1.36	0.62	10.76	16.07	**19.14**	4.87
ROUGE-Lsum	0.42	2.82	1.36	0.64	10.86	16.07	**19.07**	4.87
SQuAD F1	0.29	2.23	0.83	0.23	9.01	**14.46**	13.31	2.00

(Thai Culture Test Set)

Metric	**F- C- D-**	**F+ C- D-**	**F- C+ D-**	**F- C- D+**	**F+ C+ D+**	**WangchanX**	**Typhoon**	**OpenThai**
BERTScore	32.35	56.25	50.62	40.48	68.44	70.43	**77.57**	56.49
BLEU	0.11	0.00	0.00	0.00	0.00	**4.20**	3.68	0.00
ChrF	4.16	5.20	2.70	0.95	7.98	19.93	**20.49**	5.45
METEOR	1.91	3.43	1.50	0.61	5.94	15.15	**15.92**	3.13
ROUGE-1	1.12	3.62	1.34	0.32	8.25	16.99	**21.40**	4.86
ROUGE-2	0.37	1.89	0.33	0.05	5.04	11.03	**13.40**	2.24
ROUGE-L	1.15	3.57	1.32	0.24	8.21	16.30	**20.44**	4.75
ROUGE-Lsum	1.12	3.59	1.31	0.24	8.14	16.33	**20.47**	4.70
SQuAD F1	0.65	2.85	0.58	0.25	7.88	**16.06**	14.00	1.75

(General Test Set)

Table 6: Average evaluation results for the Creative Writing task on the Thai Culture and General Test Sets. F, C, and D denote Fluency, Culture, and Diversity, respectively. The plus sign (+) indicates the presence of the corresponding attribute, while the minus sign (-) indicates its absence.

Metric	**F- C- D-**	**F+ C- D-**	**F- C+ D-**	**F- C- D+**	**F+ C+ D+**	**WangchanX**	**Typhoon**	**OpenThai**
BERTScore	52.10	52.74	47.08	48.56	64.13	64.18	**79.61**	69.70
BLEU	0.00	0.00	0.00	0.01	0.01	0.49	**1.38**	0.32
ChrF	3.79	4.25	2.14	3.74	8.18	18.98	**20.35**	15.52
METEOR	1.94	2.05	1.23	2.13	3.68	9.06	**12.81**	8.28
ROUGE-1	1.02	2.42	0.31	1.51	4.93	9.71	**34.92**	17.03
ROUGE-2	0.00	1.05	0.00	0.31	2.50	2.78	**26.55**	10.03
ROUGE-L	0.99	2.42	0.31	1.42	4.60	8.85	**35.16**	16.82
ROUGE-Lsum	0.99	2.42	0.31	1.46	4.65	9.05	**34.47**	16.81

(Thai Culture Test Set)

Metric	**F- C- D-**	**F+ C- D-**	**F- C+ D-**	**F- C- D+**	**F+ C+ D+**	**WangchanX**	**Typhoon**	**OpenThai**
BERTScore	52.34	54.56	47.36	50.82	69.37	69.33	**77.02**	69.20
BLEU	0.00	0.00	0.00	0.00	0.02	1.15	**1.46**	1.08
ChrF	4.30	4.16	2.07	4.52	9.03	**21.93**	20.77	21.87
METEOR	2.40	2.40	1.14	2.43	5.06	11.41	**12.47**	11.03
ROUGE-1	2.11	2.61	0.34	1.50	8.12	15.68	**22.92**	13.56
ROUGE-2	1.00	0.83	0.00	0.35	3.37	8.76	**14.07**	6.55
ROUGE-L	1.79	2.66	0.34	1.47	8.00	15.18	**22.17**	12.92
ROUGE-Lsum	1.80	2.68	0.34	1.48	8.05	15.37	**22.07**	12.99

(General Test Set)

Table 7: Average evaluation results for the Multiple Choice task on the Thai Culture and General Test Sets. F, C, and D denote Fluency, Culture, and Diversity, respectively. The plus sign (+) indicates the presence of the corresponding attribute, while the minus sign (-) indicates its absence.

Metric	**F- C- D-**	**F+ C- D-**	**F- C+ D-**	**F- C- D+**	**F+ C+ D+**	**WangchanX**	**Typhoon**	**OpenThai**
BERTScore	53.65	66.32	52.86	47.64	70.36	69.91	53.20	**71.94**
BLEU	0.00	0.04	0.00	0.00	0.04	**2.65**	0.00	2.30
ChrF	5.00	8.76	3.29	1.81	9.89	15.95	5.54	**18.79**
METEOR	2.76	7.00	1.91	1.51	7.74	11.61	4.62	**11.98**
ROUGE-1	2.10	8.13	4.51	3.99	10.51	**17.70**	15.57	17.56
ROUGE-2	0.22	1.64	0.29	0.32	2.82	**4.65**	3.02	4.31
ROUGE-L	2.13	8.03	4.30	3.87	9.77	**16.71**	14.79	16.02
ROUGE-Lsum	2.08	8.12	4.29	3.95	9.89	**16.79**	14.89	16.05
SQuAD F1	0.73	5.12	1.38	1.41	6.39	**10.97**	7.18	7.21

(Thai Culture Test Set)

Metric	**F- C- D-**	**F+ C- D-**	**F- C+ D-**	**F- C- D+**	**F+ C+ D+**	**WangchanX**	**Typhoon**	**OpenThai**
BERTScore	56.41	65.37	54.17	42.19	**70.90**	66.83	54.80	69.97
BLEU	0.01	0.03	0.00	0.00	0.06	0.37	0.00	**1.70**
ChrF	5.03	7.86	3.23	0.95	9.60	11.00	4.44	**15.02**
METEOR	2.65	5.78	1.99	0.93	7.40	7.99	4.45	**9.40**
ROUGE-1	1.46	3.82	3.93	3.34	6.69	12.34	**13.34**	10.10
ROUGE-2	0.23	0.93	0.16	0.44	1.76	**3.39**	2.23	1.66
ROUGE-L	1.39	3.70	3.96	3.28	6.29	11.44	**13.40**	9.39
ROUGE-Lsum	1.41	3.75	3.97	3.33	6.32	11.43	**13.38**	9.39
SQuAD F1	0.84	2.98	1.09	0.86	5.05	**6.63**	6.60	4.22

(General Test Set)

Table 8: Average evaluation results for the Open QA task on the Thai Culture and General Test Sets. F, C, and D denote Fluency, Culture, and Diversity, respectively. The plus sign (+) indicates the presence of the corresponding attribute, while the minus sign (-) indicates its absence.

Metric	**F- C- D-**	**F+ C- D-**	**F- C+ D-**	**F- C- D+**	**F+ C+ D+**	**WangchanX**	**Typhoon**	**OpenThai**
BERTScore	52.40	58.74	49.79	50.44	71.47	68.23	**76.95**	68.09
BLEU	0.00	0.00	0.00	0.00	0.02	**0.82**	0.70	0.75
ChrF	4.17	4.17	2.26	2.97	8.27	14.64	15.30	**15.77**
METEOR	2.08	2.65	1.22	1.73	5.22	7.56	**9.30**	7.98
ROUGE-1	1.03	2.57	1.04	2.22	6.26	10.30	**12.64**	11.19
ROUGE-2	0.04	0.69	0.15	0.45	2.47	4.60	**6.74**	5.73
ROUGE-L	0.91	2.53	1.03	2.22	6.12	9.81	**12.29**	10.77
ROUGE-Lsum	0.91	2.52	1.02	2.24	6.14	9.81	**12.29**	10.77
SQuAD F1	0.46	1.29	0.44	0.68	2.41	3.80	**4.37**	3.07

(Thai Culture Test Set)

Metric	**F- C- D-**	**F+ C- D-**	**F- C+ D-**	**F- C- D+**	**F+ C+ D+**	**WangchanX**	**Typhoon**	**OpenThai**
BERTScore	58.49	60.27	51.89	53.20	75.97	72.78	**80.28**	74.20
BLEU	0.00	0.00	0.00	0.00	0.04	0.46	0.71	**1.22**
ChrF	4.69	3.93	2.33	3.08	9.18	14.12	15.18	**18.10**
METEOR	2.39	2.39	1.26	1.72	5.63	7.56	9.52	**9.58**
ROUGE-1	1.52	1.63	1.19	1.46	3.89	9.22	12.38	**13.74**
ROUGE-2	0.27	0.42	0.07	0.35	1.37	3.85	6.18	**7.09**
ROUGE-L	1.46	1.62	1.20	1.45	3.86	8.75	11.83	**13.16**
ROUGE-Lsum	1.45	1.63	1.21	1.44	3.85	8.73	11.87	**13.21**
SQuAD F1	0.93	0.87	0.68	0.70	2.37	3.35	**4.61**	3.95

(General Test Set)

Table 9: Average evaluation results for the Summarization task on the Thai Culture and General Test Sets. F, C, and D denote Fluency, Culture, and Diversity, respectively. The plus sign (+) indicates the presence of the corresponding attribute, while the minus sign (-) indicates its absence.

Metric	F- C- D-	F+ C- D-	F- C+ D-	F- C- D+	F+ C+ D+	WangchanX	Typhoon	OpenThai
BERTScore	37.11	57.15	51.25	47.76	75.77	75.74	**84.06**	59.31
BLEU	0.11	0.03	0.00	0.00	0.57	9.25	**11.71**	0.82
ChrF	6.08	7.60	5.08	3.12	15.19	27.36	**33.98**	12.82
METEOR	2.88	6.47	3.20	2.33	12.67	21.91	**28.89**	8.07
ROUGE-1	0.42	5.00	2.52	2.93	10.97	21.10	**32.22**	8.25
ROUGE-2	0.09	2.06	0.52	0.95	5.19	12.48	**19.73**	4.06
ROUGE-L	0.42	4.80	2.47	2.88	10.80	20.27	**30.71**	7.93
ROUGE-Lsum	0.41	4.83	2.48	2.89	10.82	20.33	**30.72**	7.95

(Thai Culture Test Set)

Metric	F- C- D-	F+ C- D-	F- C+ D-	F- C- D+	F+ C+ D+	WangchanX	Typhoon	OpenThai
BERTScore	40.50	58.27	51.96	46.54	78.67	78.37	**85.97**	59.96
BLEU	0.16	0.01	0.00	0.00	0.37	7.55	**11.14**	0.18
ChrF	5.89	6.48	3.53	1.88	13.82	24.81	**32.39**	9.78
METEOR	3.02	5.19	2.43	1.50	11.72	19.82	**28.37**	6.43
ROUGE-1	1.32	5.23	2.10	1.34	9.55	21.80	**32.27**	6.35
ROUGE-2	0.24	2.23	0.31	0.36	4.35	13.03	**19.76**	2.93
ROUGE-L	1.27	5.09	2.06	1.29	9.37	20.72	**30.52**	6.04
ROUGE-Lsum	1.27	5.09	2.06	1.29	9.34	20.74	**30.53**	6.05

(General Test Set)