ThaiInstruct: An instruction-following Dataset for Culturally-Aware, Multitask, and Multi-domain Evaluation in Thai

Anonymous ACL submission

Abstract

Large language models excel at instruction-following in English, but their performance in low-resource languages like Thai remains underexplored. Existing benchmarks often rely on translations, missing cultural and domain-specific nuances needed for real-world use. We present ThaiInstruct, a human-authored Thai dataset for evaluation and instruction tuning, covering four professional domains and seven task types. Created through a multi-stage quality control process with annotators, domain experts, and AI researchers, ThaiInstruct supports two studies: (1) a zero-shot evaluation showing performance gaps on culturally and professionally specific tasks, and (2) an instruction tuning study with ablations isolating the effect of native supervision. Models fine-tuned on ThaiInstruct outperform those using translated data in both in-domain and out-of-domain benchmarks. These findings underscore the need for culturally and professionally grounded instruction data to improve LLM alignment in low-resource, linguistically diverse settings.

1 Introduction

Large language models (LLMs) are crucial in NLP applications due to their instruction following capabilities, enabling zero-shot/few-shot learning for diverse tasks (e.g., summarization, machine translation) and eliminating task-specific model training. However, evaluation of LLMs remains centered on English, such as Open LLM Leaderboard (Zheng et al., 2023) and MT-Bench (Myrzakhan et al., 2024), leaving performance in underrepresented languages like Thai under-investigated. This gap is especially critical in domain-specific applications—*such as legal, financial, medical, and retail use cases*—where professional knowledge and cultural grounding are required.

Recent efforts have extended LLM evaluations to Southeast Asian languages—including Indonesian, Vietnamese, and Thai—through benchmarks like SEA-Bench (Liu et al., 2025), SEA-HELM (Susanto et al., 2025), and SEA-Crowd (Lovenia et al., 2024). However, these benchmarks predominantly rely on translated English data and lack domain specificity. This gap risk inflated performance estimates that do not reflect actual usage in native, domain-sensitive contexts. In the specific case of Thai, the problem is further exacerbated by the lack of accessible, high-quality native-language benchmarks. While authentic resources such as ThaiH6 and ThaiCLI (Kim et al., 2024) have been developed, they remain closed-source and unavailable to the broader research community, hindering transparent and reproducible evaluation of LLM performance on native Thai instructions.

Moreover, developing LLMs for Thai remains challenging due to the lack of high-quality supervised datasets. Although recent advances such as Llama (Grattafiori et al., 2024), Gemma (Team et al., 2024), SEA-LION (Ng et al., 2025), and Qwen (Qwen et al., 2025) have demonstrated promising support for Thai, their understanding of culturally and professionally grounded instructions remains underexplored. As shown in Figure 1, state-of-the-art models like ChatGPT-4o struggle to respond accurately to instructions rooted in Thai cultural and domain-specific contexts, despite their general fluency. Motivated by these challenges, we aim to address two research questions: **RQ1**: *"How effectively do current large language models (LLMs) handle Thai culturally and professionally specific instructions?"* **RQ2**: *"To what extent does fine-tuning large language models on native Thai data improve their culturally and professionally specific understanding and response accuracy?"*

To address these RQs, we present ThaiInstruct, a Thai instruction-following dataset for evaluating and improving LLM performance on Thai instructions. ThaiInstruct comprises 28,098 training and 6,916 test samples across four domains (*Medical, Law, Finance, and Retail*) and seven task types

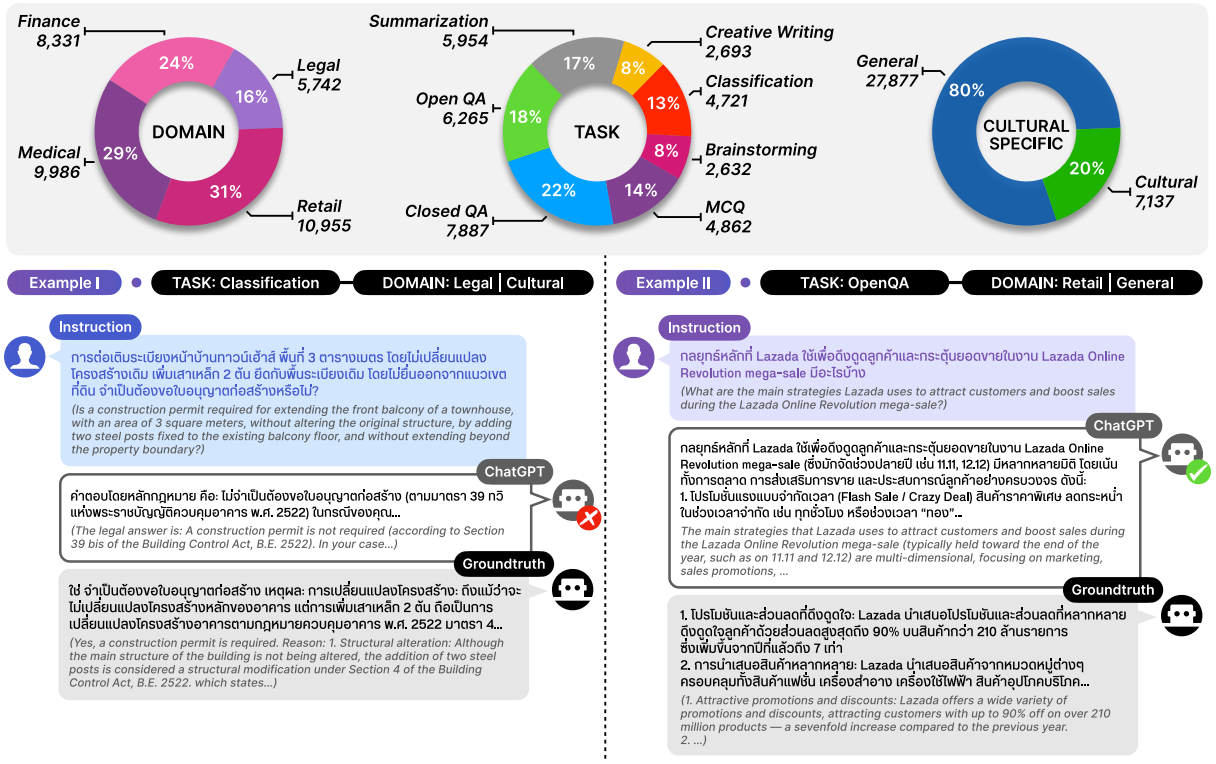

Figure 1: **(Top)** Sample distribution by domain and task type, including culture-specific vs. general data. **(Bottom)** Examples from ThaiInstruct with a sample ChatGPT-4o responses. **(Left)** A culturally specific question on Thai construction law, where ChatGPT contradicts the ground truth about permit requirements for balcony extensions. **(Right)** A general question on Lazada's marketing strategies, correctly answered by ChatGPT.

(*Brainstorming, Classification, Closed QA, Creative writing, Multiple choice, Open QA, and Summarization*). All samples are human-authored, with no reliance on LLM-generated content. We also introduce an evaluation protocol aligned with human preferences, using an LLM-as-a-judge approach to reflect Thai cultural and domain-specific contexts.

Using ThaiInstruct, we conducted two studies. For **RQ1**, we evaluated state-of-the-art LLMs in a zero-shot setting with the test split, and an LLM-as-a-judge protocol to assess factual accuracy and reasoning. We analyzed model performance across culturally and professionally grounded instructions. For **RQ2**, we fine-tuned LLMs on the training split and compared them to models trained on translated datasets like Alpaca and Dolly. Ablation studies isolate the effect of native supervision under matched data size and format, revealing how domain-specific data improves alignment.

Experimental results show that zero-shot LLMs struggle with culturally and domain-specific Thai instructions, especially in legal and multiple-choice tasks. Reasoning evaluations reveal poor rationale quality, with models often failing our judgment-based metric despite fluent outputs. In contrast, ThaiInstruct-tuned models show clear gains in both in-domain and out-of-domain settings. These findings highlight the need for deeper alignment with linguistic, cultural, and professional norms in Thai.

The contributions of our paper are as follows:

- We introduce ThaiInstruct, a human-authored Thai instruction dataset spanning multiple domains and task types, designed to support both cultural evaluation and instruction tuning in real-world, context-sensitive applications.
- We conduct two structured studies: a zero-shot evaluation revealing performance gaps on culturally and professionally specific Thai instructions, and an instruction tuning study with ablation experiments offering practical guidance on how to use ThaiInstruct effectively.
- We present a reproducible dataset development process, covering sourcing, task design, cultural annotation, and multi-stage quality control, generalizable to other underrepresented languages. We release all artifacts, including the dataset, evaluation splits, training scripts, and all fine-tuned models, including those with SOTA results on Thai LLM benchmarks, to provide strong baselines, promote reproducibility, and enable future research on applications that require Thai-specific knowledge.

Dataset	Tasks	Domains	Cultural Data?	Training&Test Data?	Human QC?	Expert QC?	Human Craft Data?	License
Multilingual Benchmarks								
SEACrowd	2	2	×	×	×/✓	×	×/✓	Mixed
SEA-Bench	1	1	×	×	✓	×	✓	Apache-2.0
SEA-HELM	4	2	×	×	×/✓	×	×/✓	CC BY-SA
Thai Benchmarks								
ThaiLLM Leaderboard	3	2	×	×	×/✓	×	×/✓	MIT
Thai-H6	1	1	×	×	✓	×	×	Closed-source
ThaiCLI	1	1	✓	×	✓	✓	✓	Closed-source
ThaiInstruct (Ours)	7	4	✓	✓	✓	✓	✓	CC BY-SA/NC

Table 1: Benchmark comparison between previous works and ours. For multilingual benchmarks, the number is displayed only for Thai data in each benchmark. Note that when we put both ×and ✓, meaning that the benchmark combines both choices, depending on the samples.

2 Related Works

2.1 Thai Language Models

The development of Thai language models has gained increasing momentum in recent years. WangchanBERTa (Lowphansirikul et al., 2021) represented one of the first pre-trained Thai language models based on BERT (Devlin et al., 2019). Followed by PhayaThaiBERT (Sriwirote et al., 2023), which expanded on this approach through expanding its tokenizer with foreign vocabularies to better handle loaned words. Following the release of GPT (Brown et al., 2020) style models, multiple works adapted these architectures for Thai. Recently, works like OpenThaiGPT (Yuenyong et al., 2025) made significant progress by finetuning base foundation LLMs like Qwen or Llama on large synthetic Thai instruction tuning datasets. Meanwhile, Typhoon (Pipatanakul et al., 2024) represents the state-of-the-art in Thai LLMs, extending its support to multimodal domains like speech and vision. While progress has been made, many works heavily rely on machine-translated instruction datasets often lacking Thai nuances, idiomatic expressions, and cultural context. This underscores the need for high-quality, culturally grounded, human-curated datasets reflecting real-world usage.

2.2 Benchmark for Thai Language Models

Recently, there has been growing interest in evaluating LLMs on Southeast Asian (SEA) data. As shown in Table 1, efforts such as SEACrowd (Lovenia et al., 2024) and SEA-Helm (Susanto et al., 2025) extend classical LLM benchmarks to Southeast Asian languages by collecting and verifying machine-translated samples through large-scale crowd sourcing. However, these resources often rely on generic questions that do not adequately probe deeper cultural nuances of the target language. Subsequent work like SEA-Bench (Liu et al., 2025) incorporates regional exams to introduce greater contextual relevance, yet still falls short of reflecting the diverse, real-world scenarios encountered by native speakers. Within Thai specifically, efforts like the ThaiLLM Leaderboard[1] curated Thai-specific evaluation datasets that still heavily rely on machine-translated data and a small-scale set of manually written MT-Bench[2] style evaluation, limiting their effectiveness for comprehensive evaluation. Later efforts, such as Thai-H6 and ThaiCLI (Kim et al., 2024), introduced benchmarks aimed at assessing cultural understanding in Thai. Nonetheless, these remain closed-source and lack accompanying training datasets, limiting their accessibility to the broader research community.

3 Dataset and Study Design

As mentioned in the introduction section, we aim to address two core research questions. To answer these questions, we design two studies using ThaiInstruct: (1) a zero-shot evaluation to assess model performance on culturally specific Thai instructions (RQ1), and (2) an instruction-tuning study to measure the effect of native supervision compared to translated datasets (RQ2), including ablation and generalization tests.

These studies require a dataset that supports both evaluation and instruction tuning. Specifically, we need a culturally grounded test set that enables fine-grained, zero-shot evaluation across task types and domains for RQ1, and a high-quality training set that enables controlled fine-tuning experiments for RQ2. In response to this need, we introduce ThaiInstruct, a human-authored dataset of Thai instructions covering diverse real-world use cases, explicitly designed to support both types of evaluation. We now describe how ThaiInstruct was constructed to support these two studies.

[1] https://huggingface.co/spaces/ThaiLLM-Leaderboard/leaderboard

[2] https://huggingface.co/datasets/ThaiLLM-Leaderboard/mt-bench-thai

3.1 ThaiInstruct Dataset Overview

To support the proposed studies, ThaiInstruct is constructed as a dual-purpose dataset for both zero-shot evaluation and instruction fine-tuning. As illustrated in Figure 1, it comprises 35,014 human-authored instruction-response pairs, split into 28,098 for training and 6,916 for testing. The dataset is designed to probe model behavior across both general and culturally specific Thai instructions (RQ1) and to provide high-quality supervision for instruction tuning (RQ2).

ThaiInstruct spans four domains—Finance, Legal, Medical, and Retail—each sourced from curated Thai websites to ensure topical relevance and language authenticity. It includes seven task types, such as Open QA, Closed QA, Summarization, and Creative writing, selected to reflect a range of real-world applications. Each domain covers diverse subtopics (e.g., 18 in Finance, 133 in Legal, 28 in Medical), offering a rich and balanced distribution for both evaluation and generalization analysis. We provide the full list of subtopics in Figure 7.

As shown in Figure 2, to ensure the dataset's reliability, we employed a three-stage annotation pipeline with dedicated groups responsible for (1) initial data creation, (2) content verification, and (3) formatting consistency. The cultural specificity is annotated by domain experts to enable subgroup analysis in RQ1. A detailed account of the collection and annotation process is provided in the following sections.

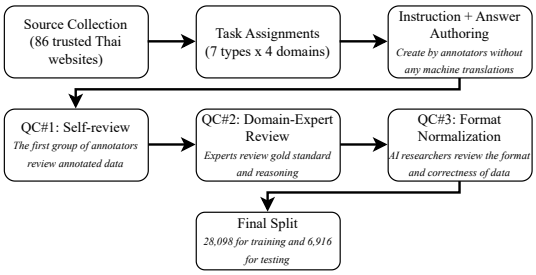

Figure 2: The overview of our data collection and annotation pipeline.

3.2 Data Collection and Annotation

Data Collection. We collected documents and articles from a diverse set of Thai websites, primarily operated by government agencies and publicly listed companies. We utilize these documents for annotators to formulate data according to the given document, instead of allowing annotators to come up with a random topic. In total, we collected 27,351 documents from 86 websites. During preprocessing, the raw HTML is cleaned by removing unnecessary elements such as HTML tags and images. We retained only content-rich documents that fell within the top 80% in terms of token count. Duplicate or highly similar documents were eliminated using `mUSE` (Yang et al., 2019), with a cosine similarity threshold of 0.8. A complete list of the source websites is provided in Appendix D.

Annotation. Annotators were selected based on two criteria: (i) native fluency in Thai and (ii) successful completion of an annotation exam covering all seven task types. Only annotators who passed this evaluation were involved in the project. For each annotation round, we provided annotators with a sample from a Thai website and a specific task type, where one website could be annotated for multiple tasks across different rounds. Annotators were instructed to create questions and answers according to the assigned task type. The annotation guideline and rules are discussed in Appendix B.1.

To ensure the quality of the gold standards, annotators were required to provide the reasoning behind each answer (except for creative writing, brainstorming, and summarization tasks), which experts in the field would later verify. This approach encouraged annotators to ensure they understood the questions and could answer them correctly. When a website was unsuitable for creating questions or answers (e.g., the provided sample was low quality or the topic was sensitive), annotators were permitted to skip that sample.

3.3 Quality Control (QC) by Human

To formulate high-quality data, we implemented three layers of quality control. In particular, we ask reviewers to check the correctness of the question, gold standard, and formatting for all data in both training and test sets. We developed an annotation platform, shown in Figure 4 of Appendix F, for annotators to evaluate the instructions. In addition, we discuss the annotation guidelines and criteria for all annotator groups in Appendix B. We describe each aspect as follows:

Question QC by Annotators. For the initial quality control, we sampled 10% of the examples and asked the same annotators to check and edit the questions, instructions, answers, and reasoning for correctness.

Gold Standard QC by Experts. In this step, we ask domain specialists who work in the target fields (i.e., medical, legal, and finance) to verify the questions and answers from the previous annotation

stage. Each example is reviewed by an expert, and if it is found to be of poor quality in either the question or the answer, or incorrect, those samples are returned to the first step. When experts reject samples, they are required to provide references and reasons why the samples are incorrect. We repeated this process until all the data were accepted by experts. Moreover, we asked experts to add a cultural tag when the sample is Thai-specific content or not, e.g., a question about Thai stocks, Thai medicine, or Thai laws.

Format QC by AI Researchers. To ensure that ThaiInstruct could serve as high-quality supervision for instruction tuning, we included a dedicated formatting stage led by AI researchers. While the first two QC layers focused on content accuracy and domain correctness, this final step emphasized structural consistency with instruction-tuning best practices. AI researchers with hands-on experience in training LLMs curated formatting guidelines tailored to each task type (e.g., proper multiple-choice structure, full-sentence closed-form QA, and inclusion of rationale). A separate group of annotators then implemented these refinements under close supervision. This step ensured that the dataset conformed to task-specific conventions expected by modern LLMs, enabling cleaner training and more reliable evaluation.

Data Leakage. Since our data has training and test data, and to prevent data leakage, we split the samples by the original document. Specifically, samples created from the same document, regardless of task type, were categorized in the same data split (either training or test sets).

4 Experiment Setup

4.1 Evaluation Setup

Since we provide a train-test split, we evaluate the model under zero-shot settings (the model was never trained on our data) and in-domain settings (the model was trained using our training data). Moreover, we evaluate those models using out-of-domain benchmarks, such as the Thai LLM leaderboard and the Thai MT-Bench benchmarks, to assess the robustness of models that were trained with our dataset. We describe each setting as follows.

Test Data Evaluation: We evaluate existing instruction LLMs without any further training using the same prompt for all models. We selected well-known instruction LLMs that supported Thai, representing a range of sizes and architectures. The models evaluated include: Gemini 2.0, Qwen2.5 7B and 72B (Qwen et al., 2025), Llama-3.1 8B and 70B (Grattafiori et al., 2024), and Sailor2 8B and 20B (Dou et al., 2024). We excluded the GPT family from this evaluation set, as GPT 4.1 was used as the judge model to assess response quality on our benchmark tasks.

Training Data Evaluation: In this setting, we test the robustness of our training data by training base LLMs using our data and testing on our test data and out-of-domain benchmarks. Given the long-form nature of the outputs in our dataset, we found it more effective to integrate with existing instruction datasets rather than using them as a standalone resource. Thus, we chose to combine our training dataset with two widely used Thai instruction datasets: `alpaca-cleaned-52k-th`[3] and `databricks-dolly-15k-th`[4]. To ensure fair comparisons, we conduct sample combinations such that the total number of examples remains consistent across setups. For instance, `Alpaca 10k` is directly compared with `Alpaca 5k + ThaInstruct 5k`, maintaining the same total size while varying the composition, where the range is from 2.5k to 30k. Additionally, we included full-size mixtures—such as `Alpaca 52k + Thai Instruct 28k` and `Dolly 15k + Thai Instruct 28k`. Our experiments utilized three base models with proven capabilities for Thai and multilingual: Gemma-2-9B (Team et al., 2024), Llama-3.1-8B (Grattafiori et al., 2024), and SEA-LIONv2-8B (Ng et al., 2025). For each model, we maintained consistent hyperparameters, shown in Table 6, across training runs for fair comparison.

For *out-of-domain benchmarks*, we use out-of-domain datasets, namely the Thai LLM Leaderboard. Although this benchmark is translation (human verified) and non-cultural data, we want to compare the generalization of models for a fair comparison with previous Thai LLM works. We use the default inference setup and codes from the original Thai LLM Leaderboard.

4.2 Metric

A key aspect of our benchmark is evaluating the justification and reasoning behind each answer (see reasoning examples in Figure 6). Traditional metrics such as BLEU or ROUGE-L fail to capture this

[3] https://huggingface.co/datasets/Thaweewat/alpaca-cleaned-52k-th

[4] https://huggingface.co/datasets/Thaweewat/databricks-dolly-15k-th

Model	Brainstorming	Classification		Closed QA		Creative Writing	Multiple Choice		Open QA		Summarization
	Fluency	Accuracy (%)	Rating	Accuracy (%)	Rating	Fluency	Accuracy (%)	Rating	Accuracy (%)	Rating	Fluency
Cultural											
Gemini 2.0	**7.68**	**71.50**	**7.47**	**96.18**	**9.20**	**8.17**	**62.70**	**7.02**	**62.13**	**6.82**	8.35
Qwen2.5-72B	7.05	63.77	6.75	93.75	9.02	7.63	54.37	6.44	45.59	5.64	**8.60**
Llama-3.1-70B	5.98	52.90	4.33	90.66	7.09	6.03	55.17	4.01	34.05	4.19	6.13
Sailor2-20B	6.41	62.32	6.55	92.01	9.07	6.07	58.73	6.15	44.85	5.53	7.87
Llama-3.1-8B	4.70	34.30	4.02	79.86	6.87	5.15	40.87	4.06	15.81	3.23	6.56
Sailor2-8B	6.50	60.87	6.38	90.97	8.91	7.00	59.92	6.59	39.34	5.24	7.99
Qwen2.5-7B	5.95	55.07	5.84	87.85	8.32	6.71	45.63	5.59	27.94	4.24	8.22
General											
Gemini 2.0	**7.68**	**84.38**	**8.23**	**97.51**	9.00	**7.95**	**69.79**	**7.46**	**78.53**	**7.77**	8.32
Qwen2.5-72B	6.91	83.43	7.96	94.43	9.03	7.14	66.58	7.33	69.08	7.05	**8.83**
Llama-3.1-70B	5.92	78.43	5.86	94.40	7.38	6.07	66.02	4.71	54.65	5.38	6.14
Sailor-2-20B	6.25	83.79	7.94	95.74	**9.13**	5.78	66.71	6.79	70.14	7.12	7.93
Llama-3.1-8B	4.90	57.99	5.38	79.70	6.92	4.83	50.13	4.82	31.24	4.10	7.05
Sailor2-8B	6.60	79.53	7.79	93.53	8.99	6.56	66.07	7.15	64.08	6.81	8.08
Qwen2.5-7B	6.20	72.43	6.83	89.25	8.44	6.31	52.44	6.15	53.67	5.68	8.46

Table 2: Comparison of zero-shot performance across cultural and general samples.

dimension, since two correct explanations may differ in wording or format yet still arrive at the same valid conclusion. Building on the foundation of MTBench (Zheng et al., 2023), we utilize LLMs to assess both correctness for tasks with objective answers (e.g., QA tasks) and to assign a 1–10 **rating** based on the quality of reasoning or justification. For creative tasks, including brainstorming, creative writing, and summarization, response quality is measured through assessments of **fluency** and cohesiveness. For tasks requiring definitive answers, such as question answering and classification, our evaluation is twofold: we verify the factual **accuracy** and evaluate the accompanying reasoning against human-labeled annotations that justify correct or incorrect responses. The complete prompts for both categories are shown in Fig. 5. Moreover, we study the robustness of existing metrics (i.e., BLEU and ROUGE-L) and our metrics compared to human preferences in Appendix A.

5 Experiment Results

5.1 Test Data Evaluation

To answer **RQ1**, we evaluate existing LLMs using our test data without any fine-tuning.

Results. As shown in Table 2, while Gemini 2.0 outperforms open-source models across several tasks, Qwen2.5-72B demonstrates comparable performance in various tasks. For example, in the cultural test set, Qwen2.5-72B outperforms Gemini 2.0 on Summarization, and also outperforms Gemini 2.0 on the Closed QA in the general set. We also observed that Sailor2-20B and Llama-3.1-70B perform comparably to Gemini 2.0 on several metrics. These findings suggest promising potential for open-source models to match or exceed the performance of proprietary models like Gemini.

Discussion. From our experimental results, we found that the most challenging tasks are multiple choice and Open QA. Although previous works demonstrate a high performance for LLMs in Multiple choice (Li et al., 2024; Balepur et al., 2024), the accuracy of this task is low compared to Closed QA or Classification. This is because our Multiple choice samples require correct reasoning for the answer to be correct. This emphasizes the challenge of our datasets; although many previous works demonstrated that Multiple choice datasets were solved by LLMs, our dataset demonstrates a contradiction with the previous works.

Reasoning Evaluation. As shown in Table 2, we evaluate the LLMs' reasoning against the gold standard using a rating metric. To quantify the relationship between reasoning quality and correctness, we compute Spearman's rank correlation between rating and accuracy, yielding a strong correlation of 0.78. This result supports our hypothesis that reasoning quality is closely tied to answer accuracy, where models that provide better reasoning are likely to produce correct answers.

Cultural vs. General. Moreover, we also observe that the performance drops when comparing cultural and general sets. When we evaluate the accuracy metric, we found that performance dropped in all models and tasks, except for Closed QA. This emphasizes that the cultural set needs cultural understanding, not only world knowledge like other benchmarks. In contrast, when evaluating the free-form tasks, i.e., Brainstorming, Creative writing, and Summarization, the performance of these tasks is almost the same for both cultural and general sets. This is because these tasks do not have an accurate answer; the texts can be answered in a similar form or text.

Model	Finance			Medical			Retail			Legal		
	Fluency	Accuracy (%)	Rating	Fluency	Accuracy (%)	Rating	Fluency	Accuracy (%)	Rating	Fluency	Accuracy (%)	Rating
Gemini 2.0	**8.18**	**79.89**	**7.82**	8.19	**88.71**	**8.70**	**7.88**	**80.78**	**7.89**	8.42	**72.35**	**7.60**
Qwen2.5-72B	7.86	76.89	7.71	**8.69**	83.84	8.18	7.53	77.10	7.77	**8.63**	61.64	6.85
Llama-3.1-70B	6.04	73.87	6.00	6.47	73.62	5.83	5.99	77.22	6.05	6.16	56.54	4.70
Sailor2-20B	6.43	76.48	7.57	8.47	85.75	8.21	6.62	77.45	7.65	8.02	62.07	6.65
Llama-3.1-8B	5.68	57.81	5.50	7.16	51.37	5.14	5.52	59.01	5.54	6.49	40.55	4.43
Sailor2-8B	6.96	71.15	7.41	8.55	84.07	8.19	6.99	73.64	7.54	8.17	60.20	6.64
Qwen2.5-7B	7.12	67.26	6.80	8.41	69.15	6.90	6.93	69.58	6.94	8.25	49.73	5.77

Table 3: Comparison of zero-shot performance across each domain.

5.2 Domain Performance

Since our dataset spans four distinct domains, LLM performance may vary across them. To address the dataset gap and challenges (**RQ1**), we analyze each model's domain-specific performance using both general and cultural sets from Table 2.

As shown in Table 3, we found that the Legal domain was the most challenging. No model achieves over 73% accuracy in this domain, whereas Gemini surpasses 88% accuracy in other domains. This may be due to limited legal data in LLM training, compounded by the fact that laws vary significantly across countries and legal systems, making it harder for LLMs to provide accurate answers. In contrast, the Medical domain, despite its real-world complexity, yields the highest accuracy, with Gemini achieving 88.71%. However, when examining fluency, all domains perform similarly. This suggests that while generating fluent Thai text is not difficult for these models, producing factually accurate legal responses remains a key challenge.

5.3 Using Our Training Data

To answer **RQ2**, we evaluate the effectiveness of our training data by fine-tuning base LLMs (Llama-3.1, Gemma-2, and SEA-LIONv2) using our data alongside comparison datasets such as Dolly and Alpaca. Additionally, we conduct experiments with both the full dataset and a size-matched subset. To avoid biased in-domain evaluation, we also assess performance on the Thai LLM Benchmark using MT-Bench, NLU, and NLG datasets. The full experimental setup is detailed in Section 4.1.

5.3.1 Full Data Comparison

Table 4 demonstrates the results of our training data compared with other training data. We found that, when we use the full dataset for both Dolly/Alpaca and our data, we outperform using only Dolly/Alpaca in 31 out of 42 cases (73% of cases that we improved the performance). We observe that we clearly obtain improvements in all cases on in-domain data. This shows that using our dataset with other data can improve the performance of LLMs, emphasizing the robustness of our training data, which not only improves the in-domain performance but also improves the out-of-domain performance.

Moreover, we also observe consistent improvement for Llama-3.1-8b on the MT Bench and NLU benchmarks. However, for other base models, we found that there were mixed results in NLG tasks. This is because NLG tasks are measured using exact match metrics like BLEU and ROUGE-L, which is often ineffective for measuring the performance of LLMs, as we see minor differences in performance of Alpaca/Dolly and our dataset. In contrast, we found that when we use models that are specific to Southeast Asian (SEA) like SEA-LION, the performance difference gap is larger because the model is designed specifically for Thai and other languages in SEA, resulting in a greater improvement when trained in SEA languages.

5.3.2 Balanced Dataset Study

To ensure that the observed performance improvement was not merely due to increased dataset size, we matched the sizes of our dataset and Alpaca/Dolly, and conducted experiments using the same benchmark. For example, we compare Alpaca 30k with a mixed dataset of Alpaca 15k and ThaiInstruct 15k, and also experimented with varying total training sizes of 10k, 20k, and 30k to assess the effect of dataset size.

As shown in Table 8, when we split the data equally, we outperform Dolly and Alpaca by 31 out of 42 cases using Llama-3.1-8b, 41 out of 42 cases using Gemma-2-9b, and 28 out of 42 cases using SEA-LIONv2-8B. Although there were mixed results in NLG datasets, we still outperform competitive datasets in NLU datasets for both Llama-3.1 and Gemma-2. In addition, the in-domain performance also increases greatly in all cases. This emphasizes the effectiveness of our training data for in-domain and out-of-domain evaluation regardless of the size of the training data.

Model	MT Bench Average	NLU Accuracy (%)	NLG Translation (BLEU)	NLG Generation (RougeL)	ThaiInstruct Fluency	ThaiInstruct Accuracy (%)	ThaiInstruct Rating
Llama-3.1-8B							
Alpaca 52k	3.04	48.48	**2.21**	**9.51**	3.10	26.28	3.19
Alpaca 52k + ThaiInstruct 28k	**3.43**	**50.15**	2.17	8.54	**4.52**	**43.38**	**4.58**
Dolly 15k	2.64	42.47	**1.60**	8.10	1.85	40.34	2.63
Dolly 15k + ThaiInstruct 28k	**2.88**	**45.56**	1.28	**8.60**	**4.41**	**43.51**	**4.54**
Gemma-2-9B							
Alpaca 52k	3.43	53.23	1.11	7.04	2.61	14.35	2.26
Alpaca 52k + ThaiInstruct 28k	**4.61**	**53.95**	**1.86**	**8.05**	**4.87**	**55.94**	**5.20**
Dolly 15k	**4.10**	51.43	**1.48**	7.76	1.85	40.34	2.63
Dolly 15k + ThaiInstruct 28k	3.86	**53.88**	1.47	**8.06**	**4.87**	**54.44**	**5.16**
SEA-LIONv2-8B							
Alpaca 52k	**4.80**	**43.94**	14.59	**25.73**	4.75	44.41	4.44
Alpaca 52k + ThaiInstruct 28k	4.76	43.87	**16.40**	16.51	**5.34**	**51.47**	**5.33**
Dolly 15k	3.57	**46.14**	**14.31**	**35.37**	3.24	48.13	4.15
Dolly 15k + ThaiInstruct 28k	**4.13**	43.93	13.38	16.09	**5.17**	**50.52**	**5.11**

Table 4: Train LLMs using our training data with other training data to improve the robustness of LLMs in out-of-domain and in-domain evaluations. We also compare the balance data size in Section 5.3.2.

5.4 Context Length Analysis

Since our test set contains samples with long contexts, ranging from 6 to 26,405 tokens for the test set, we study the gaps and challenges in those samples that contain long contexts. To facilitate this analysis, we divide the samples into three groups based on token length: head (shortest 20%), body (middle 40%), tail (longest 40%), as illustrated in Figure 3. We expect robust models to perform consistently regardless of token lengths.

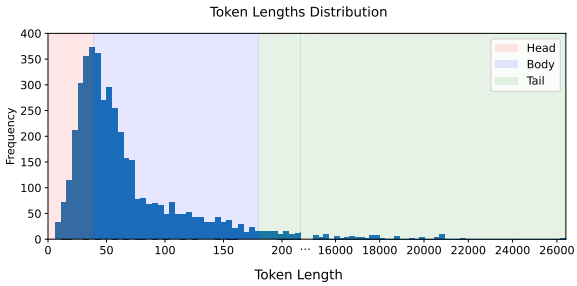

Figure 3: Token length distribution in our dataset.

Table 7 presents the performance of each model across the head, body, and tail distributions. When analyzing the samples with long context, we observe a significant performance drop in the Brainstorming, Multiple choice, Open QA, and Summarization tasks across all models when comparing head and tail groups, with the sole exception of Gemini in Summarization. Notably, models constrained by shorter context length (i.e., Llama-3.1) or those trained on limited context length data (i.e., Sailor2) exhibit a significant performance drop across both the body and tail distributions. For example, Sailor2-20B's performance on Summarization dropped from 8.97 to 6.46 points (∼27% decrease). These findings underscore the difficulty of our dataset when handling long inputs, where even state-of-the-art models like Gemini struggle to maintain robust performance.

6 Conclusion

We introduced ThaiInstruct, a human-authored Thai instruction dataset designed to support evaluation and fine-tuning. Our experiments revealed a consistent performance gap in state-of-the-art LLMs when handling Thai context-sensitive inputs, and demonstrated that fine-tuning with ThaiInstruct leads to measurable improvements, supported by ablation studies under size- and format-controlled conditions. These findings establish ThaiInstruct as both a benchmark and a development tool for culturally and professionally aligned Thai LLMs.

Beyond the dataset, we contribute a transparent and reproducible process for constructing instruction data that reflects both cultural and domain-specific grounding. While ThaiInstruct is language-specific, the underlying methodology is generalizable and can guide similar efforts in other underrepresented languages and application domains. We make all resources publicly available — including the dataset, evaluation splits, training scripts, and all fine-tuned models, even those that achieve state-of-the-art performance on Thai LLM benchmarks — to establish solid baselines, ensure reproducibility, and support future research focused on culturally and professionally relevant Thai applications.

Limitations

While our dataset and benchmark are a step forward toward better representations of Thai in the LLM era, we acknowledge that significant progress is still to be made. We outline several limitations of our study below.

Evaluate Metric. Since the main metric of our benchmark relies on LLM-as-a-judge, reproducing this number in the future might not be straightforward since the API model might update. However, using traditional metrics like BLEU or ROUGE-L also yields the worst evaluation metric, as we discussed in Appendix A.

Judge Models. As we discussed in Appendix A, the limitation of our evaluation metric is the judge model. We found that there are no dominant models to judge Thai knowledge and culture, where the Kendall between annotators and LLM is only 0.1259 points.

Non-Commercial Use Limitation. A significant portion of our released dataset (30,000 samples) is distributed under the CC BY-NC license. This restricts the use of these samples to non-commercial purposes only. While this allows for academic and research use, it can be limiting for potential users who require the data for commercial applications. The remaining samples (5,014 samples) under the more permissive CC BY-SA 4.0 license can be used commercially under the terms of that license.

Ethical Statement

We hire all annotators from an annotation company in Thailand. However, for experts, we contract their working company and hire them directly. The pay rate is higher than the minimum wage in Thailand. For the AI Researcher, we also use the same payment rate similar to the domain experts. For the license of data, we will release them publicly using the license of CC-BY-SA and NC, since some of the data, although they allow us to collect, still want the license to be NC.

References

Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. 2024. Artifacts or abduction: How do LLMs answer multiple-choice questions without the question? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10308–10330, Bangkok, Thailand. Association for Computational Linguistics.

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, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Longxu Dou, Qian Liu, Guangtao Zeng, Jia Guo, Jiahui Zhou, Wei Lu, and Min Lin. 2024. Sailor: Open language models for south-east asia. *Preprint*, arXiv:2404.03608.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Dahyun Kim, Sukyung Lee, Yungi Kim, Attapol Rutherford, and Chanjun Park. 2024. Representing the under-represented: Cultural and core capability benchmarks for developing thai large language models. *arXiv preprint arXiv:2410.04795*.

Wangyue Li, Liangzhi Li, Tong Xiang, Xiao Liu, Wei Deng, and Noa Garcia. 2024. Can multiple-choice questions really be useful in detecting the abilities of LLMs? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2819–2834, Torino, Italia. ELRA and ICCL.

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

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

Lalita Lowphansirikul, Charin Polpanumas, Nawat Jantrakulchai, and Sarana Nutanong. 2021. Wangchanberta: Pretraining transformer-based thai language models. *Preprint*, arXiv:2101.09635.

Rajiv Movva, Pang Wei Koh, and Emma Pierson. 2024. Annotation alignment: Comparing LLM and human annotations of conversational safety. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9048–9062, Miami, Florida, USA. Association for Computational Linguistics.

Aidar Myrzakhan, Sondos Mahmoud Bsharat, and Zhiqiang Shen. 2024. Open-llm-leaderboard: From multi-choice to open-style questions for llms evaluation, benchmark, and arena. *Preprint*, arXiv:2406.07545.

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

Maja Pavlovic and Massimo Poesio. 2024. The effectiveness of LLMs as annotators: A comparative overview and empirical analysis of direct representation. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 100–110, Torino, Italia. ELRA and ICCL.

Kunat Pipatanakul, Potsawee Manakul, Natapong Nitarach, Warit Sirichotedumrong, Surapon Nonesung, Teetouch Jaknamon, Parinthapat Pengpun, Pittawat Taveekitworachai, Adisai Na-Thalang, Sittipong Sripaisarnmongkol, Krisanapong Jirayoot, and Kasima Tharnpipitchai. 2024. Typhoon 2: A family of open text and multimodal thai large language models. *Preprint*, arXiv:2412.13702.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Panyut Sriwirote, Jalinee Thapiang, Vasan Timtong, and Attapol T. Rutherford. 2023. Phayathaibert: Enhancing a pretrained thai language model with unassimilated loanwords. *Preprint*, arXiv:2311.12475.

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

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Multilingual universal sentence encoder for semantic retrieval. *Preprint*, arXiv:1907.04307.

Sumeth Yuenyong, Kobkrit Viriyayudhakorn, Apivadee Piyatumrong, and Jillaphat Jaroenkantasima. 2025. Openthaigpt 1.5: A thai-centric open source large language model. *Preprint*, arXiv:2411.07238.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.

A Measuring Alignment Between Metrics and Human Preferences

We evaluated three state-of-the-art LLMs (Gemini 2.5, GPT 4.1, and Claude Sonnet 3.7) to see how closely they aligned with human when ranking model outputs. To ensure broad coverage, we sampled 1,200 examples uniformly across all domains and task types. Following previous works (Pavlovic and Poesio, 2024; Movva et al., 2024), we had annotators rank the outputs from worst to best, while LLMs produced numerical ratings. We measured the agreement between human rankings and LLM scores using Kendall τ, which explicitly accounts for ties on either side, which often appeared due to the LLMs producing ratings. Surprisingly, all three LLMs showed only weak correlation with human preferences as shown in Table 5, indicating that even top-performing models struggle to capture the nuances of Thai cultural and contextual understanding. In the end, since none of the LLMs strongly aligned with human preferences, we opted for `GPT 4.1` as the judge, as we were limited by the rate limit for both `Claude Sonnet 3.7` and `Gemini 2.5`, resulting in a much longer judge time.

Model	Annotator 1	Annotator 2	Annotator 3	Annotator Avg	Judge Time (hr)
Gemini 2.5	**0.1322**	**0.0964**	0.0156	0.1255	24
GPT 4.1	0.1273	0.0933	0.0107	**0.1259**	1
Sonnet 3.7	0.1044	0.0915	0.0048	0.0962	8
BLEU	-0.0010	0.0427	**0.0198**	0.0319	0.1
ROUGE-L	-0.0115	0.0166	0.0083	0.0053	0.5

Table 5: Kendall τ between annotators and LLM and the judge time per model

B Annotator Guideline

B.1 First Group of Annotators

We have provided information in each domain as a JSON file in the following structure: Title, Texts (Article), URL (Source link). Then, we let the annotator write the instruction and gold standard, where each article will have an instruction task type, such as: Open QA, Closed QA, Multiple choice, Summarization, Brainstorming, Classification, and Creative Writing. We also randomly assign a task type for users to perform, according to the following rules:

- Medical domain has the most data: randomly assign 1 task type per article.
- Finance domain: 2 task types per article.
- Retail domain: 2 task types per article.
- Legal domain: 2 task types per article.

We also control each domain's distribution to have the most balanced distribution of task types. We also asked annotators to check the length of each article first. If the content is not long enough or too short, assign only 1 task type to that article, except for the Medical domain, which is already 1 task type per article.

B.2 Domain Experts

Criteria are grouped into three categories:

C1: Format and Scope. The output must strictly follow the required structure: "Answer + Elaboration." The inclusion of "Comparison" and "Conclusion" components is optional. Responses must adhere closely to the given instructions and input, avoiding irrelevant content or exceeding the specified scope. For example, in summarization tasks, the response should accurately reflect all key points from the input without introducing information not present in the original content.

C2: Factual Accuracy and Completeness. The output must be factually accurate, logically sound, and fully address all parts of the instruction. For example, if asked to recommend practice (e.g., the duration to avoid heavy meals before exercise), but the output fails to provide the underlying rationale for this recommendation, it is considered incomplete and consequently does not meet C2.

C3: Instruction/Question Relevance. This criterion evaluates whether the instruction or question is appropriate, specific, and well-aligned with the expected task type and input. It also considers the question's relevance to the domain and its suitability for eliciting a meaningful response. Additionally, experts added relevant tags to each example, with no limit on the number of tags as long as they related to the content. This process ensured that the gold standards in our dataset were 100% accurate.

B.3 AI Researcher Annotators

As we discussed in Section 3.3, for this group, we only let them check the format of the instruction and the gold standard. If they found the error or incorrect format, the sample will be sent back to the first group of annotators, similar to the domain experts' pipeline.

C Inference Setup

All models were evaluated using a temperature of 0.2 and a `max new tokens` of 1028, with inference performed via the vLLM [5] engine, with the only exception being Gemini, which was accessed through API calls.

D Data Sources

The data were collected in the form of web page documents from the following websites:

- **Finance**: bam.co.th, finrwealthbuilder.com, kasikornbank.com, longtunman.com, phillip.co.th, setinvestnow.com, finnomena.com, thestandard.co, brandage.com, brandbuffet.in.th, brandinside.asia, ceochannels.com, marketingoops.com, wealthsolution.co.th.

- **Legal**: cpao.go.th, ddproperty.com, dharmniti.co.th, dplawandservice.com, elt-corp.com, justicechannel.org, khemmapat.org, kobkiat.com, lawsiam.com, lawyerthailand.biz, mkclegal.com, moj.go.th, pdpathailand.com, promsaklawyer.com, saranlaw.com, slawconsult.com

- **Medical**: ambu.or.th, bangkokhospital-chiangrai.com, bumrungrad.com, he02.tci-thaijo.org, w1.med.cmu.ac.th, chulalongkornhospital.go.th, dmh.go.th, dst.or.th, thaiepilepsysociety.com, rama.mahidol.ac.th, gastrothai.net, goodbyeitch.com, haamor.com, idthai.org, manarom.com, medparkhospital.com

- **Retail**: ceochannels.com, marketingoops.com, pnstoretailer.com, techsauce.co, storehub.com, brandage.com, brandbuffet.in.th, brandinside.asia, marketthink.co, readthecloud.co.

E Training Setup

All models were trained with LlamaFactory (Zheng et al., 2024) with the hyperparameters specified in Table 6.

Hyperparameter	
Learning Rate	2×10^{-4}
Learning Rate Schedule	Cosine
Batch Size (effective)	128
Max Token Length	2048
Warm up Ratio	0.1
Epochs	3

Table 6: Training Hyperparameters.

[5] https://github.com/vllm-project/vllm

F Annotation Platform

As shown in Figure 4, we developed an annotation platform for the domain experts and QC members to assess the quality of the created instruction stances.

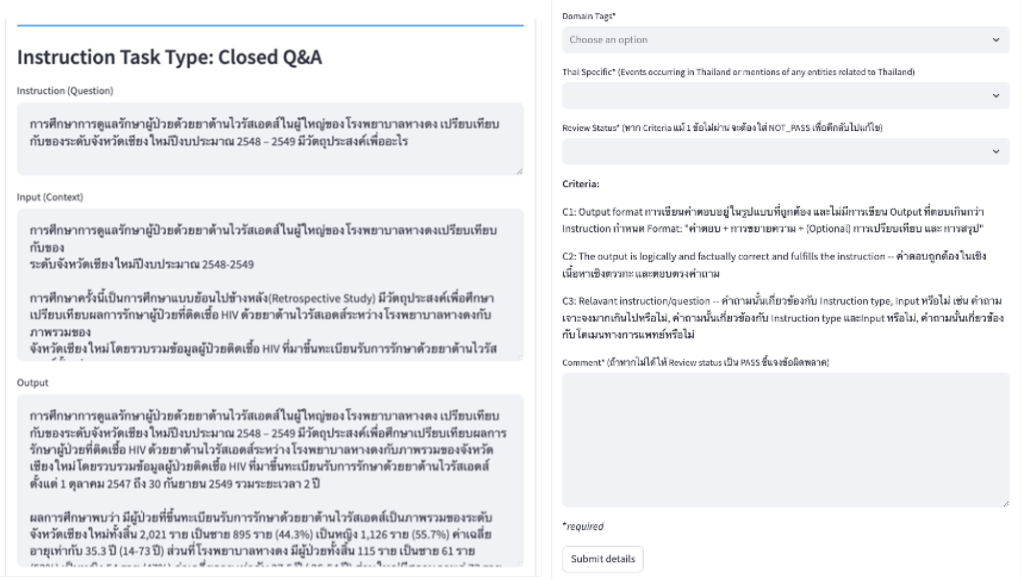

Figure 4: A screenshot of the annotation platform where domain experts can evaluate the quality of instructions based on specific criteria and tag the relevant topics associated with each instruction.

	Brainstorming	Classification		Closed QA		Creative Writing	Multiple Choice		Open QA		Summarization
Model	**Fluency**	**Accuracy (%)**	**Rating**	**Accuracy (%)**	**Rating**	**Fluency**	**Accuracy (%)**	**Rating**	**Accuracy (%)**	**Rating**	**Fluency**
Gemini 2.0											
Head	7.45	80.37	8.10	95.04	8.81	6.61	70.95	7.29	**77.69**	**7.69**	**8.53**
Body	7.61	**84.16**	**8.27**	97.67	9.09	8.00	**71.78**	**7.57**	73.28	7.47	8.22
Tail	**7.86**	80.24	7.89	**97.87**	**9.13**	**8.69**	62.84	7.16	75.00	7.58	8.32
Qwen2.5-72B											
Head	6.79	78.08	7.63	92.37	8.98	5.93	66.67	7.26	63.75	6.64	**9.27**
Body	6.90	**80.85**	**7.82**	**94.94**	9.00	7.30	**67.88**	**7.36**	**65.79**	**6.78**	8.91
Tail	**7.00**	79.02	7.67	94.58	**9.07**	**7.75**	57.70	6.79	61.75	6.74	8.39
Llama-3.1-70B											
Head	5.97	70.48	5.61	88.37	6.82	5.04	**67.44**	**5.07**	39.50	4.53	**6.89**
Body	**6.01**	**72.60**	**5.86**	93.42	7.27	6.04	64.89	4.47	47.58	4.91	6.14
Tail	5.82	71.69	5.01	**95.79**	**7.55**	**6.51**	57.95	4.20	**53.63**	**5.37**	5.80
Sailor2-20B											
Head	6.31	78.54	7.67	93.89	9.08	5.53	64.29	6.29	63.35	6.71	**8.97**
Body	6.13	78.49	**7.68**	**96.89**	**9.25**	**5.89**	**65.21**	6.65	63.16	6.70	8.84
Tail	**6.36**	**81.22**	7.64	89.94	8.53	5.86	64.55	**6.80**	**66.45**	**6.86**	6.46
Llama-3.1-8B											
Head	4.70	49.32	4.96	77.48	6.76	3.72	45.71	4.17	20.72	3.34	**7.21**
Body	4.85	53.43	**5.33**	**83.07**	**7.13**	4.89	**49.39**	4.63	24.70	3.73	7.07
Tail	**5.02**	**55.37**	4.97	77.56	6.76	**5.44**	47.43	**4.88**	**34.83**	**4.39**	6.68
Sailor2-8B											
Head	6.50	70.78	7.22	91.98	8.96	5.88	63.81	6.81	**64.14**	**6.60**	**9.20**
Body	6.55	76.60	**7.62**	**93.77**	**9.05**	6.62	**65.94**	**7.12**	55.47	6.35	9.03
Tail	**6.67**	**77.80**	7.55	87.04	8.34	**6.95**	63.57	7.00	58.76	6.50	6.52
Qwen2.5-7B											
Head	5.96	**69.41**	6.62	87.40	8.33	5.09	50.00	5.90	39.44	4.75	**8.62**
Body	6.00	69.03	**6.65**	**89.49**	8.41	6.46	**55.47**	**6.30**	47.98	5.35	8.43
Tail	**6.48**	68.78	6.63	89.17	**8.46**	**6.92**	46.45	5.78	**52.35**	**5.71**	8.29

Table 7: Token length analysis with bolded scores representing best scores in each group

Tasks with definitive answers prompt template
[Instruction]
You are to act as an impartial judge tasked with evaluating the quality of a response provided by an AI assistant. The evaluation must focus on two primary criteria: **correctness** and **helpfulness**.

You will be given:
- A **user question**.
- A **reference answer**, which is considered the accurate and authoritative response.
- An **assistant's answer**, which you must evaluate.

Your evaluation must begin by determining whether the assistant's answer is **factually correct** and aligns with the information in the reference answer.
If the assistant's answer contains significant factual inaccuracies, clear contradictions with the reference, or fails to address the core question, mark it as incorrect.
If the answer is largely accurate and consistent with the reference, mark it as correct.

Indicate your judgment in the following format:
Correctness: [[1]] (if correct) or Correctness: [[0]] (if incorrect)

Next, provide a **reasoning rating** on a scale from 1 to 10. This rating should reflect how well the assistant's reasoning aligns with the logic, depth, and completeness of the reference answer:
- A score of 10 means the reasoning is entirely sound, complete, and mirrors the clarity and correctness of the reference.
- A score of 5 means the reasoning is partially correct or incomplete, but not misleading.
- A score below 5 indicates flawed or poor reasoning, misunderstanding of the topic, or failure to properly support the answer.
- A score of 1 means the reasoning is severely flawed or entirely off-topic.

Use the following format:
Rating: [[<score>]]
newline
[Question]
question

[The Start of Reference Answer]
`{ref_answer}`
[The End of Reference Answer]

[The Start of Assistant's Answer]
`{answer}`

Creative tasks prompt template
[Instruction]
Please act as an impartial judge and evaluate the quality of the response provided by an AI assistant to the user question displayed below. Your evaluation should consider clarity, sentence fluency, and cohesion of the writing. You will be given a reference answer and the assistant's answer. Begin your evaluation by comparing the assistant's answer with the reference answer. Identify and correct any mistakes. Be as objective as possible. After providing your explanation, you must rate the response on a scale of 1 to 10 by strictly following this format: [[rating]], for example: Rating: [[5]].

[Question]
question

[The Start of Reference Answer]
`{ref_answer}`
[The End of Reference Answer]

[The Start of Assistant's Answer]
`{answer}`
[The End of Assistant's Answer]

Figure 5: Adapted MT-Bench prompt templates to fit our benchmark. For tasks with definitive answers we assess both reasoning and correctness of the answers. For creative tasks we assess the clarity, fluency and cohesion of the response.

Instruction (Original):

จากเหตุการณ์ข้อมูลรั่วไหลของ CP Freshmart เหตุใดจึงส่งผลกระทบต่อลูกค้าในแง่มุมของ Retail มากที่สุด?
(A) ลูกค้าอาจสูญเสียเงินออมในบัญชีธนาคาร
(B) ลูกค้าอาจถูกมิจฉาชีพแอบอ้างนำข้อมูลไปใช้
(C) ลูกค้าอาจถูกติดตามพฤติกรรมการซื้อของ
(D) ลูกค้าอาจไม่สามารถใช้บริการของ CP Freshmart ได้

Groundtruth (Original):

คำตอบคือ (B) ลูกค้าอาจถูกมิจฉาชีพแอบอ้างนำข้อมูลไปใช้ เนื่องจาก เหตุผล: (A) ข้อมูลที่รั่วไหลนั้นไม่มีข้อมูลเกี่ยวกับบัญชีธนาคาร (B) ข้อมูลที่รั่วไหลนั้นมีข้อมูลส่วนตัวของลูกค้า เช่น ชื่อ เบอร์โทรศัพท์ อีเมล ที่อยู่ ซึ่งมิจฉาชีพสามารถนำไปใช้แอบอ้างเป็นลูกค้าเพื่อทำธุรกรรมต่างๆ (C) ข้อมูลที่รั่วไหลนั้นไม่มีข้อมูลเกี่ยวกับพฤติกรรมการซื้อของ (D) ข้อมูลที่รั่วไหลนั้นไม่ส่งผลต่อการใช้งานบริการของ CP Freshmart คำตอบที่ถูกต้อง: (B) ลูกค้าอาจถูกมิจฉาชีพแอบอ้างนำข้อมูลไปใช้ เหตุผลเพิ่มเติม: ข้อมูลส่วนตัว เช่น ชื่อ เบอร์โทรศัพท์ อีเมล ที่อยู่ สามารถนำไปใช้แอบอ้างเพื่อทำธุรกรรมต่างๆ เช่น สมัครบัตรเครดิต กู้เงิน ซื้อสินค้าออนไลน์ มิจฉาชีพสามารถโทรหาลูกค้าเพื่อหลอกลวงให้โอนเงินหรือเปิดเผยข้อมูลส่วนตัวเพิ่มเติม การถูกมิจฉาชีพแอบอ้างนำข้อมูลไปใช้อาจสร้างความเสียหายทางการเงินและความเสียหายต่อชื่อเสียงของลูกค้า ข้อควรระวัง: ลูกค้าควรเปลี่ยนรหัสผ่านของบัญชีต่างๆ ที่ใช้ข้อมูลส่วนตัวที่รั่วไหล ลูกค้าควรระวังการติดต่อจากบุคคลแปลกหน้าที่อ้างว่ามาจาก CP Freshmart ลูกค้าควรติดตามข่าวสารจาก CP Freshmart เกี่ยวกับการแก้ไขปัญหาข้อมูลรั่วไหล

Instruction (Translated):

From the data breach incident of CP Freshmart, which of the following reasons negatively impacted the customers the most?
(A) Customers may lose their savings in bank accounts
(B) Customers may be impersonated by fraudsters using their personal information
(C) Customers may be tracked for their purchasing behavior
(D) Customers may be unable to use CP Freshmart's services

Groundtruth (Translated):

The correct answer is (B): Customers may be impersonated by fraudsters using their personal information because: (A) The leaked data does not include bank account information, (B) The leaked data contains customers' personal information, such as names, phone numbers, email addresses, and home address, which fraudsters can use to impersonate customers and carry out various transactions, (C) The leaked data does not contain information related to purchasing behavior, (D) The leaked data does not affect the availability of CP Freshmart's services. The correct answer is B: Customers may be impersonated by fraudsters using their personal information. Additional reasoning: Personal information such as names, phone numbers, email addresses, and addresses can be used by fraudsters to impersonate individuals for various fraudulent transactions, such as applying for credit cards, taking out loans, or shopping online. Fraudsters may also call customers to deceive them into transferring money or disclosing further personal information. Being impersonated in this way can cause financial loss and damage to the customer's reputation. Precautions: Customers should change passwords for any accounts that may use the leaked personal information. Customers should be cautious of any contact from strangers claiming to be from CP Freshmart. Customers should stay informed by following updates from CP Freshmart regarding how they are addressing the data breach.

Figure 6: An MCQ example from the dataset showing that each correct answer is followed by the reasoning that arrived towards that answer.

Finance Domain Subtopics:
Financial Law, Investment Strategies, Personal Financial Management, Asset Management, Financial Analysis, Financial Analysis and Financial Economics, Digital Finance, Economic and Financial News, Company Financial Information, Financial Literacy, Financial Market, Financial Market and Financial Products and Services, Financial Products and Services, Financial Institutions, Financial Instruments, Financial Technology (FinTech), FinTech and Digital Finance, Financial Economics,

Legal Domain Subtopics:
Ministerial Regulation, Maritime Transport and Navigation Law – Merchant Marine and Marine Rescue, Fiscal Law, International Trade Law, Aviation Law, Gambling Law, Education Law, Education Law – Teachers and Personnel, Informal Lending Law (Share Lending), Trade Competition Law – Price Control, Loan Law, Loan Law – Usufruct and Consumable Loans, Immigration Law, Immigration Law – Employment of Foreigners, Family Law, Alcohol Control Law, Cybersecurity and Privacy Law, Consumer Protection Law, Consumer Protection Law – Product and Service Pricing, Labor Protection Law, Labor Protection Law – Labor Relations and Compensation, Suretyship Law, Traffic Law, Traffic and Land Transport Law, Mortgage Law, Pledge Law, Hire of Work Law, Employment Law, Public Assembly Law, Sales Law, Agency Law, Bill and Cheque Law, Property Law – Ownership, Property Law – Ownership and Limited Real Rights, Intellectual Property Law, Intellectual Property Law (Copyrights, Patents, Trademarks), Military Law, Brokerage Law, Juristic Acts and Contracts Law, Juristic Acts and Contracts Law – Unfair Contract Terms, Juristic Person Law, Current Account and Warehouse Receipt Law, Entertainment Law, Administrative Law, Administrative Procedure and Administrative Court Establishment Law, Administrative Tort Liability Law, Insurance Law, Social Security Law, Compromise and Settlement Law, Anti-Money Laundering Law, Deposit Law, Judiciary Act, Judiciary Act – Royal Decree, Judiciary Act – Jurisdiction Determination between Courts, Energy Law, Merchant Marine Law (Marine Transport and Rescue) – Royal Act, Merchant Marine Law – Multimodal Transport Royal Act, Tax Law, Inheritance Law, Inheritance Law – Wills, Narcotics Law, International Law, International Law – Treaties, Constitutional Law, Transport of Goods and Passengers Law, Tort Law, Law of Evidence, Law of Evidence (Civil and Criminal), Obligations Law, Obligations Law – Royal Act, Obligations Law – Debt Collection Royal Act, Bankruptcy and Business Reorganization Law, Building Control, Condominium, Land Allotment and Land Reclamation Law, Building Control Law (draft), Building Control Law (Building Control Royal Act), Building Control Law – Excavation and Land Fill, Criminal Case Law for Political Office Holders, Computer Crime Law (draft), Computer Crime Law (Computer Act), Tourism Business and Guide Law, Animal Law, Public Health, Medical and Epidemic Law, Human Rights Law, Environmental Law, Environmental, Forestry and National Park Law, Health Law, Securities Law, Partnership and Company Law, Space Law, Real Estate Law, Criminal Law – Petty Offenses, Criminal Law – Public Endangerment (Arson, Flood, etc.), Criminal Law – Public Endangerment, Criminal Law – Commercial Offenses, Criminal Law – Offenses against the Administration (Officials and Government Positions), Criminal Law – Forgery and Falsification, Criminal Law – Justice Administration Offenses, Criminal Law – National Security and Terrorism, Criminal Law – Public Peace, Criminal Law – Secret Societies and Criminal Associations, Criminal Law – Life and Bodily Harm, Criminal Law – Property Offenses, Criminal Law – Theft, Embezzlement, Fraud, Criminal Law – Offenses Related to Corpses, Criminal Law – Sexual Offenses, Criminal Law – Liberty and Reputation, Weapons, Ammunition, Explosives, Fireworks and Imitation Weapons Law, Hire Purchase Law, Lease Law, Juvenile and Family Law, Juvenile and Family Law – Royal Act, Juvenile and Family Law – Domestic Violence Protection Royal Act, Election Law, Labor Law, Exchange and Gift Law, Local Administration Regulation Law, Industrial Factory and Machinery – Mining Law, Cybersecurity and Personal Data Protection Law (PDPA), General News and Statistics, Local Administration Ordinance, Basic Legal Knowledge, Supreme Court Judgments, Administrative Court Judgments, Land Code, Criminal Procedure Code, Civil Procedure Code, Criminal Code, Civil and Commercial Code, Royal Decree, Royal Decree – Supreme Court President Regulations, Emergency Decree, Organic Act, Legal Profession and Lawyer Ethics – Complaints, Petitions, Legal Etiquette,

Medical Domain Subtopics:
Anatomy, Physiology, Alternative Medicine, Pediatrics, Nursing, Ophthalmology, Psychiatry, Dentistry, Gynecology, Forensic Medicine, Dermatology, Pathology, Epidemiology, Radiology, Anesthesiology, Surgery, Veterinary Medicine, Public Health, Obstetrics, Orthopedics, Internal Medicine, Others, Pharmacology, Transfusion Medicine, Emergency Medicine, Rehabilitation Medicine, Nutrition, Otolaryngology / ENT

Figure 7: Full list of subtopics for Finance, Legal, and Medical domains

Model	MT Bench Average	NLU Accuracy (%)	NLG Translation (BLEU)	NLG Generation (RougeL)	ThaiInstruct Fluency	ThaiInstruct Accuracy (%)	ThaiInstruct Rating
Llama-3.1-8B							
Alpaca 5k + ThaiInstruct 5k	3.00	**47.22**	3.12	8.59	**4.08**	**39.84**	**4.16**
Alpaca 10k	**3.05**	46.54	**4.08**	**11.05**	3.36	28.39	3.33
Alpaca 10k + ThaiInstruct 10k	**3.07**	46.47	2.43	8.54	**4.21**	**42.31**	**4.39**
Alpaca 20k	2.75	**47.31**	**2.79**	**9.14**	2.77	22.32	2.94
Alpaca 15k + ThaiInstruct 15k	**3.26**	44.65	1.86	8.58	**4.35**	**42.16**	**4.46**
Alpaca 30k	2.88	**47.67**	**3.47**	**9.65**	2.83	21.83	2.95
Dolly 2.5k + ThaiInstruct 2.5k	**2.40**	**46.43**	**3.75**	8.72	**3.57**	**35.93**	**3.72**
Dolly 5k	1.88	42.87	1.98	**9.55**	1.75	22.70	2.19
Dolly 5k + ThaiInstruct 5k	**2.28**	**46.43**	**1.36**	8.55	**3.85**	**37.89**	**3.98**
Dolly 10k	1.99	42.74	1.35	**8.96**	1.69	22.35	2.14
Dolly 7.5k + ThaiInstruct 7.5k	2.31	**46.37**	1.48	**8.59**	**3.96**	**39.63**	**4.11**
Dolly 15k	**2.64**	42.47	**1.60**	8.10	1.69	22.21	2.16
Gemma-2-9B							
Alpaca 5k + ThaiInstruct 5k	**4.25**	**53.70**	**2.25**	**8.14**	**4.85**	**54.24**	**5.17**
Alpaca 10k	3.98	51.71	1.39	6.84	4.00	46.62	4.26
Alpaca 10k + ThaiInstruct 10k	4.02	**53.81**	**2.02**	**8.09**	**4.97**	**55.33**	**5.30**
Alpaca 20k	**4.14**	52.40	1.45	6.95	3.53	38.07	3.90
Alpaca 15k + ThaiInstruct 15k	**4.20**	**53.49**	**1.98**	**8.02**	**5.14**	**56.67**	**5.49**
Alpaca 30k	3.79	52.41	1.1	5.73	3.25	32.71	3.43
Dolly 2.5k + ThaiInstruct 2.5k	**3.66**	**54.62**	**1.75**	**8.07**	**4.30**	**51.86**	**4.84**
Dolly 5k	2.59	53.34	1.39	7.58	1.71	42.35	2.45
Dolly 5k + ThaiInstruct 5k	**3.99**	**53.50**	**1.54**	**8.12**	**4.59**	**54.31**	**5.08**
Dolly 10k	2.70	51.98	1.52	7.58	1.81	43.68	2.74
Dolly 7.5k + ThaiInstruct 7.5k	**4.13**	**53.34**	**1.63**	**8.12**	**4.72**	**55.09**	**5.24**
Dolly 15k	4.10	51.43	1.48	7.76	1.85	40.34	2.63
SEA-LIONv2-8B							
Alpaca 5k + ThaiInstruct 5k	4.52	**43.76**	**34.47**	19.39	**5.62**	**52.84**	**5.57**
Alpaca 10k	**4.54**	43.31	28.01	**25.35**	4.61	48.88	4.73
Alpaca 10k + ThaiInstruct 10k	4.55	**44.66**	24.00	17.55	**5.72**	**53.93**	**5.70**
Alpaca 20k	**4.74**	43.98	**24.22**	**25.82**	4.73	49.32	4.53
Alpaca 15k + ThaiInstruct 15k	4.44	**44.51**	**20.58**	16.31	**5.54**	**53.94**	**5.61**
Alpaca 30k	**4.60**	42.96	15.58	**25.68**	5.11	49.66	4.78
Dolly 2.5k + ThaiInstruct 2.5k	**4.25**	44.89	**36.60**	26.82	**5.10**	**50.25**	**5.28**
Dolly 5k	3.69	**45.88**	25.53	**35.66**	3.46	48.04	4.11
Dolly 5k + ThaiInstruct 5k	**4.21**	44.30	**15.64**	23.72	**5.31**	**51.25**	**5.42**
Dolly 10k	3.83	**46.57**	14.40	**37.35**	3.09	48.61	4.04
Dolly 7.5k + ThaiInstruct 7.5k	**4.31**	45.31	13.54	22.00	**5.54**	**53.81**	**5.57**
Dolly 15k	3.57	**46.14**	**14.31**	**35.37**	3.24	48.13	4.15

Table 8: Dataset ablation: in-domain distribution with data size ablations.

Model	Social Science	Extraction	Math	Reasoning	STEM	Writing	Coding	Roleplay	Average
Llama 3.1 8B									
Alpaca 5k + ThaiInstruct 5k	2.9	2.85	2.2	2.3	3.7	3	3.68	3.35	3.00
Alpaca 10k	3.65	2.35	2	1.9	3.25	3.45	3.86	3.9	3.05
Alpaca 10k + ThaiInstruct 10k	4.25	2.15	2.15	2.45	3.5	3.35	3.14	3.55	3.07
Alpaca 20k	3.4	1.8	1.35	1.85	3.4	2.95	3.23	4.05	2.75
Alpaca 15k + ThaiInstruct 15k	3.65	3	2.05	2.9	3.5	3	3.95	4	3.26
Alpaca 30k	3.3	1.8	1.9	1.95	3	3.15	3.77	4.15	2.88
Dolly 2.5k + ThaiInstruct 2.5k	2.65	3	1.4	1.8	2.6	2.75	2.5	2.5	2.40
Dolly 5k	2.45	1.85	1.3	1.8	2	1.65	2	1.95	1.88
Dolly 5k + ThaiInstruct 5k	3.25	2.95	1.6	1.85	1.7	2.7	2.36	1.8	2.28
Dolly 10k	2.15	2.05	1.25	1.7	1.85	2.35	3	1.6	1.99
Dolly 7.5k + ThaiInstruct 7.5k	3.9	2.5	1.1	2.1	2.45	2.5	1.86	2.1	2.31
Dolly 15k	3.35	3.4	1.5	2.9	2.5	1.7	3	2.8	2.64
Gemma 2 9B									
Alpaca 5k + ThaiInstruct 5k	6	3.9	3.05	3.75	4.1	4.05	3.64	5.5	4.25
Alpaca 10k	5.4	2.75	2.95	2.85	5.05	4.2	3.32	5.35	3.98
Alpaca 10k + ThaiInstruct 10k	6	2.95	3.3	4.3	4.1	3.6	3.18	4.75	4.02
Alpaca 20k	6	2.3	4.15	3.7	4.65	4.3	3.59	4.4	4.14
Alpaca 15k + ThaiInstruct 15k	6.05	2.8	3.3	3.95	4.3	4	4.18	5	4.20
Alpaca 30k	4.65	2.45	3.85	2.55	4.75	3.25	3.77	5.05	3.79
Dolly 2.5k + ThaiInstruct 2.5k	5.2	2.75	3.15	3.75	3.65	3.95	2.85	4	3.66
Dolly 5k	3.2	2.4	1.6	2.8	3.4	2.15	2.05	3.15	2.59
Dolly 5k + ThaiInstruct 5k	5.85	3	3	4.2	4.2	4	3.32	4.35	3.99
Dolly 10k	3.4	1.75	1.65	3.05	3.35	3.1	2.23	3.1	2.70
Dolly 7.5k + ThaiInstruct 7.5k	6.3	3.9	3.2	3.95	4.2	4	3.18	4.3	4.13
Dolly 15k	5.85	4.55	3.5	3.75	4.55	3.6	2.77	4.25	4.10
SEA-LIONv2-8B									
Alpaca 5k + ThaiInstruct 5k	6.05	3.95	3.3	3.65	5.8	5.15	3.68	4.6	4.52
Alpaca 10k	6.25	3.9	2.4	3.35	4.95	5.55	4.27	5.65	4.54
Alpaca 10k + ThaiInstruct 10k	7.25	3.25	3.35	3.95	5.2	4.65	3.18	5.6	4.55
Alpaca 20k	6.8	3.35	3.4	4.45	4.95	5.2	4.18	5.55	4.74
Alpaca 15k + ThaiInstruct 15k	5.65	3.65	3	3.75	5.45	4.75	3.59	5.55	4.44
Alpaca 30k	7	3.5	3.45	4.1	5.5	4.7	3.36	5.15	4.60
Dolly 2.5k + ThaiInstruct 2.5k	5.35	4.25	3.1	3.5	4.75	4.75	3.68	4.6	4.25
Dolly 5k	5.1	4.15	1.75	3.8	4.45	3.3	3.14	3.8	3.69
Dolly 5k + ThaiInstruct 5k	5.9	4.3	2.25	4.05	4.45	4.65	3.41	4.65	4.21
Dolly 10k	5.15	2.75	2.5	3.9	4.95	3.6	3.18	4.6	3.83
Dolly 7.5k + ThaiInstruct 7.5k	5.75	4.45	2.2	3.45	5	4.95	3.64	5.05	4.31
Dolly 15k	5	3.55	2.1	3.7	4.2	3.2	3.18	3.6	3.57

Table 9: Dataset ablation: Thai MT Bench

Model	Social Science	Extraction	Math	Reasoning	STEM	Writing	Coding	Roleplay	Average
Llama-3.1-8B									
Alpaca 52k + ThaiInstruct 28k	4.1	2.15	1.85	2.8	3.55	2.95	3.82	4.15	3.43
Alpaca 52k	3.4	2.05	1.65	1.95	4.1	3.45	3.55	4.2	3.04
Dolly 15k + ThaiInstruct 28k	3.75	3.5	1.65	2.3	2.75	2.55	2.64	2.85	2.88
Dolly 15k	3.35	3.4	1.5	2.9	2.5	1.7	3	2.8	2.64
Gemma-2-9B									
Alpaca 52k + ThaiInstruct 28k	6.6	4.0	3.5	3.4	4.5	4.9	4.10	4.95	4.61
Alapaca 52k	3.6	2.35	3.8	2.3	4.7	3.4	3.5	3.8	3.43
Dolly 15k + ThaiInstruct 28k	5.8	3.25	3.4	3.75	4.75	3.75	2.95	3.25	3.86
Dolly 15k	5.85	4.55	3.5	3.75	4.55	3.6	2.77	4.25	4.10
SEA-LIONv2-8B									
Alpaca 52k + ThaiInstruct 28k	6.35	4.05	2.7	3.7	6.2	5.25	4.64	5.15	4.76
Alpaca 52k	6.75	3.65	3.25	4.3	5.7	4.7	4.41	5.6	4.80
Dolly 15k + ThaiInstruct 28k	6.15	3.7	2.2	4.0	4.85	4.25	4.14	3.75	4.13
Dolly 15k	5	3.55	2.1	3.7	4.2	3.2	3.18	3.6	3.57

Table 10: Full dataset: Thai MT Bench results

Model	Belebele	M3 Exam	Thai Exam	Wisesight	xCopa	xNLI	Average
Llama-3.1-8B							
Alpaca 5k + ThaiInstruct 5k	65.89	39.02	36.64	39.12	69.60	33.07	47.22
Alpaca 10k	63.33	40.54	33.27	45.53	66.60	29.98	46.54
Alpaca 10k + ThaiInstruct 10k	63.78	38.70	35.04	37.55	70.40	33.37	46.47
Alpaca 20k	63.11	40.36	33.45	50.51	66.20	30.26	47.31
Alpaca 15k + ThaiInstruct 15k	61.56	38.28	33.10	35.75	66.20	33.01	44.65
Alpaca 30k	62.33	39.39	32.39	50.17	69.20	32.51	47.67
Dolly 2.5k + ThaiInstruct 2.5k	65.11	40.68	35.58	36.20	68.00	33.03	46.43
Dolly 5k	59.44	35.01	29.73	37.18	63.00	32.85	42.87
Dolly 5k + ThaiInstruct 5k	65.11	40.68	35.58	36.20	68.00	33.03	46.43
Dolly 10k	60.78	35.33	28.85	35.19	63.80	32.51	42.74
Dolly 7.5k + ThaiInstruct 7.5k	65.11	39.44	36.99	35.04	68.60	33.03	46.37
Dolly 15k	60.67	35.93	31.50	34.71	60.00	32.00	42.47
Gemma-2-9B							
Alpaca 5k + ThaiInstruct 5k	77.11	50.05	46.55	37.36	77.60	33.53	53.70
Alpaca 10k	76.89	48.11	46.73	33.17	71.00	34.37	51.71
Alpaca 10k + ThaiInstruct 10k	77.56	49.86	44.78	36.80	80.20	33.65	53.81
Alpaca 20k	78.56	47.92	47.43	37.18	68.00	35.33	52.40
Alpaca 15k + ThaiInstruct 15k	78.56	49.82	46.73	34.07	77.20	34.59	53.49
Alpaca 30k	77.44	48.25	46.73	40.51	66.80	34.71	52.41
Dolly 2.5k + ThaiInstruct 2.5k	80.22	49.58	46.90	33.88	82.60	34.53	54.62
Dolly 5k	78.33	49.08	46.37	33.28	78.60	34.37	53.34
Dolly 5k + ThaiInstruct 5k	79.22	49.63	46.37	33.02	79.40	33.37	53.50
Dolly 10k	78.00	48.66	43.89	32.31	75.60	33.41	51.98
Dolly 7.5k + ThaiInstruct 7.5k	80.44	50.14	46.73	31.30	81.60	33.37	53.34
Dolly 15k	75.11	47.32	43.36	33.10	71.60	38.10	51.43
SEA-LIONv2-8B							
Alpaca 5k + ThaiInstruct 5k	57.56	40.45	36.81	32.98	61.40	33.35	43.76
Alpaca 10k	56.11	42.02	38.23	29.88	59.60	34.01	43.31
Alpaca 10k + ThaiInstruct 10k	57.22	39.02	34.69	33.43	70.20	33.37	44.66
Alpaca 20k	58.33	42.62	38.76	27.59	62.60	33.99	43.98
Alpaca 15k + ThaiInstruct 15k	57.44	39.81	36.64	35.38	64.40	33.39	44.51
Alpaca 30k	57.44	42.39	36.99	27.07	59.60	34.25	42.96
Dolly 2.5k + ThaiInstruct 2.5k	57.11	42.62	38.76	36.05	61.40	33.41	44.89
Dolly 5k	62.11	43.54	42.30	33.73	59.00	34.57	45.88
Dolly 5k + ThaiInstruct 5k	58.22	38.84	36.64	35.75	63.00	33.35	44.30
Dolly 10k	63.78	44.42	40.35	34.22	64.00	32.67	46.57
Dolly 7.5k + ThaiInstruct 7.5k	61.11	40.45	38.41	32.38	66.20	33.33	45.31
Dolly 15k	63.22	43.68	38.58	34.89	62.60	33.83	46.14

Table 11: NLU (Accuracy) benchmark results

Model	Belebele	M3 Exam	Thai Exam	Wisesight	xCopa	xNLI	Average
Llama-3.1-8B							
Alpaca 52k + ThaiInstruct 28k	65.56	39.53	36.99	56.31	69.20	33.33	50.15
Alpaca 52k	65.67	39.21	33.81	49.42	70.00	32.18	48.48
Dolly 15k + ThaiInstruct 28k	62.89	39.99	35.40	34.97	64.80	33.31	45.56
Dolly 15k	60.67	35.93	31.50	34.71	60.00	32.00	42.47
Gemma-2-9B							
Alpaca 52k + ThaiInstruct 28k	78.33	50.78	46.73	33.96	74.40	35.49	53.95
Alpaca 52k	79.33	49.49	47.08	36.17	72.00	35.31	53.23
Dolly 15k + ThaiInstruct 28k	79.00	50.78	47.96	28.94	78.60	34.01	53.88
Dolly 15k	75.11	47.32	43.36	33.10	71.60	38.10	51.43
SEA-LIONv2-8B							
Alpaca 52k + ThaiInstruct 28k	59.44	40.18	34.51	34.03	60.00	33.07	43.87
Alpaca 52k	57.00	40.77	34.87	39.20	58.60	33.21	43.94
Dolly 15k + ThaiInstruct 28k	57.22	39.16	36.11	36.54	61.20	33.33	43.93
Dolly 15k	63.22	43.68	38.58	34.89	62.60	33.83	46.14

Table 12: Full datasets: NLU results

Model	Flores200 Eng-Th (BLEU)	Flores200 Th-Eng (BLEU)	XLsum (RougeL)	iApp QA (RougeL)
Llama-3.1-8B				
Alpaca 5k + ThaiInstruct 5k	2.00	4.25	13.29	3.90
Alpaca 10k	2.59	5.58	14.21	7.89
Alpaca 10k + ThaiInstruct 10k	1.86	3.00	13.27	3.82
Alpaca 20k	1.82	3.75	13.34	4.94
Alpaca 15k + ThaiInstruct 15k	1.35	2.37	13.37	3.80
Alpaca 30k	2.95	3.98	13.49	5.81
Dolly 2.5k + ThaiInstruct 2.5k	2.20	5.30	13.30	4.13
Dolly 5k	2.83	1.13	12.49	6.60
Dolly 5k + ThaiInstruct 5k	2.08	0.63	13.24	3.86
Dolly 10k	1.99	0.71	12.31	5.61
Dolly 7.5k + ThaiInstruct 7.5k	1.94	1.01	13.31	3.87
Dolly 15k	2.23	0.96	12.48	3.71
Gemma-2-9B				
Alpaca 5k + ThaiInstruct 5k	2.16	2.33	12.63	3.65
Alpaca 10k	1.64	1.13	10.45	3.23
Alpaca 10k + ThaiInstruct 10k	2.13	1.90	12.50	3.67
Alpaca 20k	1.05	1.84	10.84	3.05
Alpaca 15k + ThaiInstruct 15k	2.12	1.83	12.38	3.65
Alpaca 30k	0.96	1.24	8.55	2.90
Dolly 2.5k + ThaiInstruct 2.5k	2.38	1.12	12.51	3.63
Dolly 5k	2.30	0.48	11.90	3.25
Dolly 5k + ThaiInstruct 5k	2.41	0.67	12.57	3.66
Dolly 10k	2.46	0.57	11.97	3.19
Dolly 7.5k + ThaiInstruct 7.5k	2.53	0.73	12.54	3.70
Dolly 15k	2.40	0.56	12.28	3.23
SEA-LIONv2-8B				
Alpaca 5k + ThaiInstruct 5k	25.91	43.03	14.73	24.05
Alpaca 10k	21.46	34.56	18.02	32.68
Alpaca 10k + ThaiInstruct 10k	24.72	23.28	15.06	20.03
Alpaca 20k	20.83	27.61	18.77	32.87
Alpaca 15k + ThaiInstruct 15k	25.74	15.41	15.31	17.31
Alpaca 30k	20.57	10.58	18.87	32.48
Dolly 2.5k + ThaiInstruct 2.5k	26.20	46.99	15.99	37.65
Dolly 5k	25.90	25.15	19.58	51.73
Dolly 5k + ThaiInstruct 5k	25.19	6.09	15.05	32.38
Dolly 10k	25.24	3.56	18.83	55.86
Dolly 7.5k + ThaiInstruct 7.5k	24.60	2.48	14.94	30.00
Dolly 15k	25.31	3.30	18.65	52.08

Table 13: Dataset ablation: NLG results

Model	Flores200 Eng-Th (BLEU)	Flores200 Th-Eng (BLEU)	XLsum (RougeL)	iApp QA (RougeL)
Llama-3.1-8B				
Alpaca 52k + ThaiInstruct 28k	2.01	2.32	13.19	3.89
Alpaca 52k	1.91	2.51	13.18	5.83
Dolly 15k + ThaiInstruct 28k	1.72	0.83	13.31	3.89
Dolly 15k	2.23	0.96	12.48	3.71
Gemma-2-9B				
Alpaca 52k + ThaiInstruct 28k	2.15	1.56	12.45	3.64
Alpaca 52k	1.36	0.86	11.56	2.51
Dolly 15k + ThaiInstruct 28k	2.32	0.61	12.46	3.65
Dolly 15k	2.40	0.56	12.28	3.23
SEA-LIONv2-8B				
Alpaca 52k + ThaiInstruct 28k	25.79	7.00	15.69	17.32
Alpaca 52k	25.38	3.79	18.75	32r.70
Dolly 15k + ThaiInstruct 28k	24.65	2.10	14.67	17.50
Dolly 15k	25.31	3.30	18.65	52.08

Table 14: Full datasets: NLG results