2209491 PROJECT IN LANGUAGE TECHNOLOGY II 2568

Title: Assessment of a Large Language Model's Reading and Comprehension Proficiency in the Thai Language

Researcher: Nanthipat Kongborrirak

Advisor: Dr. Jakapun Tachaiya

Abstract

The rapid increase in Large Language Models (LLMs) makes it important to carefully evaluate their abilities in different languages, especially for languages like Thai, which have fewer digital resources. This study tests how well several well-known LLMs can understand written Thai. The models included those that support multiple languages, like GPT-4o, Gemini Flash 2.0, and DeepSeek V3, as well as models specifically adjusted for Thai, such as Typhoon2 8B and 70B. We used a set of questions taken from Thai national exams (A-Level, CU-TEP, O-NET). The models were judged by their exact answer correctness using different ways of asking questions (prompts). These methods included basic naive prompts, Chain of Thought (CoT), few-shot learning method, and prompt optimization. Results indicate that larger models achieve significantly higher accuracy, and while LLMs surpass average human performance on less complex questions (O-NET), a gap remains for advanced reasoning tasks (A-Level). Few-shot prompting emerged as a particularly effective strategy for enhancing the performance of smaller models. This work provides a systematic evaluation of LLM Thai comprehension, offering insights into the factors influencing their performance and identifying areas for future improvement.

Contents

1. Introduction

1.1. Problem Statement

Large Language Models (LLMs) are becoming increasingly common in education, helping with tasks like question answering, summarizing text, and even tutoring. However, most existing benchmarks focus on English or high-resource languages, leaving significant gaps in our understanding of LLM performance in low-resource languages like Thai. Given Thai's unique script, grammar, and relatively limited annotated datasets, it is essential to assess how well LLMs generalize to Thai-language educational tasks.

Despite advances in multilingual training and instruction-tuned models, the comprehension capabilities of LLMs in Thai remain underexplored. Specifically, there is a lack of systematic evaluation of various factors, such as model size, prompting strategies, and prompt language that could affect model performance in Thai reading comprehension tasks.

1.2. Objectives and Contributions

This project aims to evaluate multiple LLMs on Thai reading comprehension using real-world multiple-choice exam datasets. The key contributions are as follows:

1) Benchmarking Thai Comprehension: We evaluate multilingual and Thai fine-tuned LLMs using reading comprehension exams as datasets (A-Level, O-NET M3, O-NET M6).
2) Analyzing Model Scale: We compare models of varying sizes (e.g., 8B vs. 70B) to assess how model capacity affects Thai comprehension performance.
3) Prompting Strategy Evaluation: We assess the impact of different prompting methods—naive (zero-shot), Chain of Thought (CoT), DSPy-based prompt optimization, and few-shot learning—on accuracy.
4) Prompt Language Comparison: We investigate whether using Thai or English in the prompt leads to better comprehension performance.

2. Literature Review

2.1. Prompting Techniques in LLMs

Recent advancements in prompting methods have significantly enhanced the performance and adaptability of large language models (LLMs). According to Liu et al. (2024), prompting has emerged as a lightweight and efficient alternative to fine-tuning, enabling models to adapt to diverse downstream tasks with minimal parameter updates. The authors classify prompting methods into two major categories: efficient computation and efficient design.

Efficient computation techniques aim to reduce resource consumption by minimizing the number of tokens or using compressed representations. Methods such as prompt distillation, prefix-tuning, and adapter-based approaches allow models to generalize with fewer examples and less memory.

On the other hand, efficient design focuses on the quality and structure of the prompt itself. Techniques in this category include chain-of-thought prompting, few-shot prompting, and instruction tuning. These approaches guide the model toward more interpretable and robust reasoning by structuring the prompt to mimic human-like thinking patterns. Chain-of-thought prompting, in particular, has shown to significantly improve model performance in reasoning tasks by encouraging step-by-step deduction before producing an answer.

2.2. DSPy Framework for Declarative LM Programming

The DSPy framework (Declarative Self-improving Python) introduces a modular and declarative approach to large language model (LLM) programming, aiming to overcome the brittleness and inefficiencies of traditional prompt engineering. Instead of manually crafting prompts, DSPy allows developers to define programs using composable modules and signatures (input-output field definitions) that describe the structure of an LLM task.

One of the key innovations of DSPy is its self-improving compiler that automatically tunes and optimizes the LLM's behavior by generating and selecting better instructions or demonstrations using black-box access to the language model. The framework supports prompt optimization, few-shot selection, multi-step reasoning, and even retrieval-augmented generation in a unified, high-level interface. This enables researchers and developers to build more robust and adaptive pipelines with minimal manual intervention.

In evaluations, DSPy has demonstrated improvements in reasoning tasks, few-shot performance, and the ability to generalize across tasks and models with reduced engineering overhead (Khattab et al., 2023).

2.3. Multilingual Models

To evaluate large language models' ability to generalize across languages, this study incorporates state-of-the-art multilingual models that support Thai. GPT-4o (OpenAI, 2024) is the latest flagship model optimized for multimodal and multilingual tasks, demonstrating strong performance across various benchmarks, including low-resource languages. Gemini 2.0 Flash, developed by Google DeepMind, is designed for high efficiency and fast inference while retaining competitive performance, making it suitable for practical educational applications. DeepSeek V3 (DeepSeek-AI, 2024) is another open-weight multilingual model, trained with a large amount of English and Chinese data, and has recently gained attention for its robust cross-lingual generalization capabilities. These models were selected based on their popularity, performance, and ability to support Thai as a target language for evaluation.

2.4. Thai Fined-Tune Models

Large language models fine-tuned specifically for the Thai language have emerged to address the limitations of multilingual models on low-resource languages. One notable advancement is Typhoon, a family of instruction-tuned Thai LLMs based on LLaMA architectures, trained with high-quality Thai instruction datasets across various domains such as education, news, and conversation (Thanajiran et al., 2024). The Typhoon models, available in both 8B and 70B parameter sizes, have demonstrated strong performance in Thai-centric tasks, outperforming baseline multilingual models and other open Thai LLMs on several benchmarks. This study incorporates Typhoon-2 models to evaluate whether Thai-specific instruction tuning and increased model capacity improve reading comprehension in Thai.

3. Data Description

3.1. Dataset Overview

This study utilizes Thai-language multiple-choice reading comprehension exams sourced from official national assessment bodies. The datasets include a total of 110 questions, split into evaluation and training sets:

1) O-NET M3 and O-NET M6 (Grade 9 and 12): 25 questions each, totaling 50 items.
2) A-Level + CU-TPT: Combined into a single dataset of 50 questions.
3) Training Set for Prompt Optimization and Few-Shot Learning: 10 representative questions selected from the full dataset for use in DSPy's prompt optimization (MIPROv2) and few-shot prompting experiments.

The datasets represent increasing levels of reading difficulty, following the order: O-NET M3 < O-NET M6 < A-Level+CU-TPT, based on academic level and cognitive demand. Additionally, O-NET M3 and CU-TPT questions contain 4 answer choices, while O-NET M6 and A-Level contain 5 options.

Each question consists of a reading passage (context), a question, a set of multiple-choice options, and a designated correct answer. The dataset is formatted with the following fields:

1) context: A short passage or paragraph.
2) question: A comprehension or inference-based prompt.
3) options: A list of 4–5 answer choices.
4) correct_answer: The exact choice and text matching the correct response.

context	**question**	**options**	**correct_answer**
งิ้วเป็นไม้เนื้ออ่อน เป็นไม้ธรรมดาที่ไม่ค่อยมีอะไรโดดเด่น ในอดีตคนไทยนิยมนำเส้นใยจากขนของเมล็ดหรือผนังด้านในของผลมาทำเป็นนุ่นยัดหมอน ที่นอน เบาะ เป็นต้น ต่อมาชาวบ้านโค่นต้นงิ้วทิ้งเพื่อนำพื้นที่มาใช้ทำการเกษตร ต้นงิ้วจึงเหลือน้อยลง ทำให้คนไทยในปัจจุบันรู้จักไม้ชนิดนี้ไม่มากนัก "The Kapok tree (Ngio) is a softwood, an ordinary tree	ข้อใดเป็นใจความสำคัญ "Which of the following is the main idea?"	1. งิ้วเป็นไม้เนื้ออ่อนและมีลักษณะไม่โดดเด่น 2. คนไทยในอดีตนิยมปลูกต้นงิ้วเพื่อใช้ประโยชน์ 3. เส้นใยจากต้นงิ้วนำมาใช้ยัดหมอนและที่นอน	5. คนไทยในปัจจุบันไม่ค่อยรู้จักต้นงิ้ว "5. Thai people nowadays do not know kapok trees well."

without many outstanding features. In the past, Thai people popularly used the fibers from its seed hairs or the inner lining of its fruit to make kapok for stuffing pillows, mattresses, cushions, etc. Later, villagers cut down kapok trees to use the land for agriculture. Consequently, kapok trees became fewer, causing Thai people nowadays to not know this type of tree very well."		4. ปัจจุบันต้นงิ้วมีจำนวนน้อยมาก 5. คนไทยในปัจจุบันไม่ค่อยรู้จักต้นงิ้ว "1. The Kapok tree is a softwood and has unremarkable characteristics. 2. In the past, Thai people popularly grew kapok trees for their benefits. 3. Fibers from the kapok tree are used to stuff pillows and mattresses. 4. Currently, kapok trees are very few in number. 5. Thai people nowadays do not know kapok trees well."	

Table 1. Example of a Thai Reading Comprehension Question in the Dataset with English Translation

3.2. Preprocessing and Annotation

All datasets were reviewed to ensure that every item included a complete and coherent passage, as required for evaluating LLM reading comprehension. Questions without clear supporting context were excluded.

For the prompt optimization and few-shot prompting experiments, a subset of 10 questions was selected from the full dataset to serve as the training set for DSPy's MIPROv2 optimization and demonstration examples. These training samples were chosen to be representative of the overall question styles and topics across datasets.

Minimal preprocessing was applied beyond format standardization and ensuring consistent label formatting. The dataset was then split into a training subset (for prompt tuning) and an evaluation subset.

4. Methodology

4.1. Workflow Overview

The evaluation framework comprises three main stages: (1) Prompt Construction, (2) Model Inference, and (3) Evaluation. A variety of prompting strategies are systematically applied to different large language models (LLMs), followed by automated inference and scoring using an exact match accuracy metric. The prompting workflow is structured using the DSPy framework. All LLMs are accessed via the OpenRouter AI API, which provides a unified interface for querying diverse models through OpenAI-compatible endpoints. Figure 1 illustrates the overall workflow.

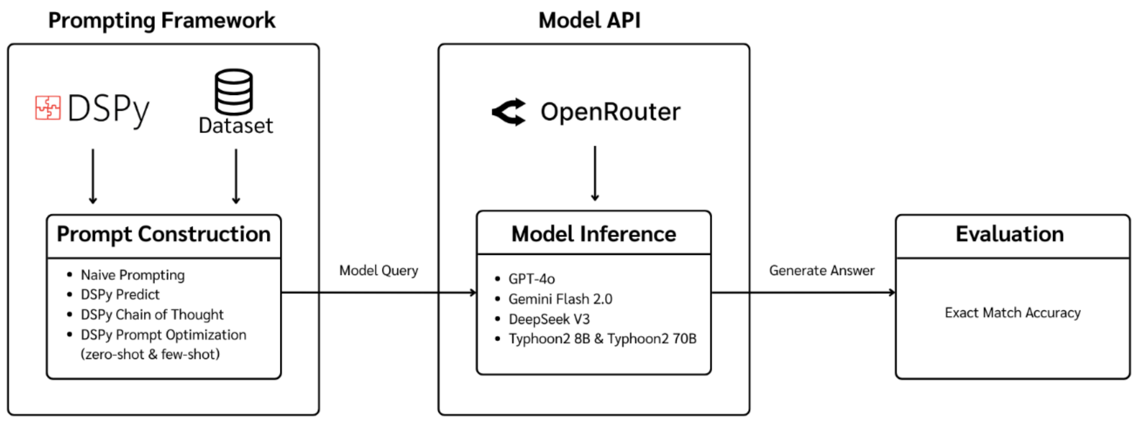

Figure 1. System architecture of the evaluation pipeline, integrating DSPy for prompt construction, OpenRouter for model inference, and exact match accuracy for evaluation.

4.2. Component Description

This study investigates the performance of large language models (LLMs) using a range of prompting strategies and model types. All models are accessed via the OpenRouter AI API, which provides a unified interface for interacting with multiple LLM providers. Two categories of LLMs are evaluated: multilingual models, including GPT-4o, Gemini Flash 2.0, and DeepSeek V3, and Thai fine-tuned models, specifically Typhoon2 8B and Typhoon2 70B. The multilingual models are general-purpose and trained on diverse languages, while the Typhoon2 models are instruction-tuned specifically for the Thai language.

To evaluate the impact of prompting strategies, four main prompting methods are employed. The **Naive Prompting** method serves as the baseline and consists of a straightforward instruction in either Thai or English. This method asks the model to read a passage and select the most accurate answer from multiple choices. The same prompt instruction is reused across all methods to ensure consistency in input structure.

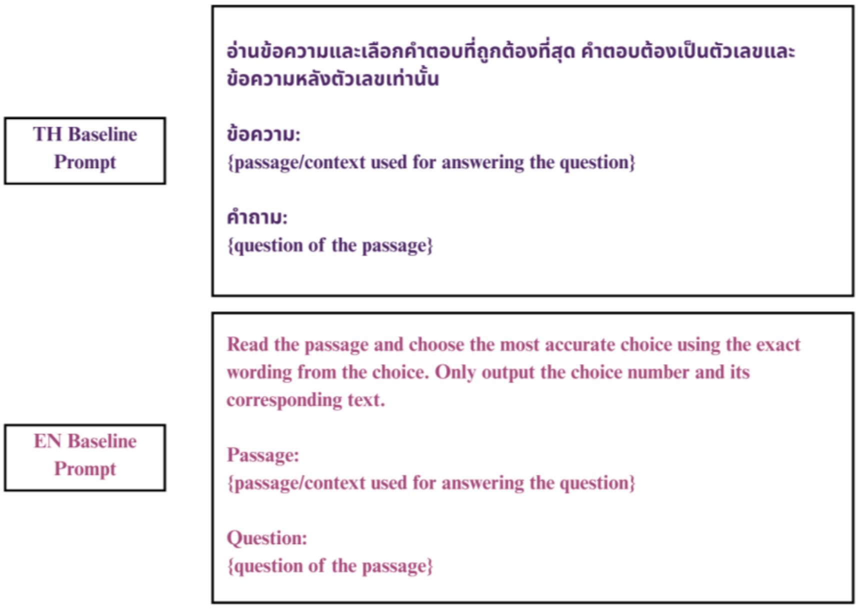

Figure 2. Thai and English Baseline Prompts Used in Naive Prompting.

The **DSPy Predic**t method leverages DSPy's Signature abstraction to provide a structured prompt format, clearly defining input and output fields (e.g., context, question, and answer). This structured approach helps standardize the model's behavior and improve consistency in responses.

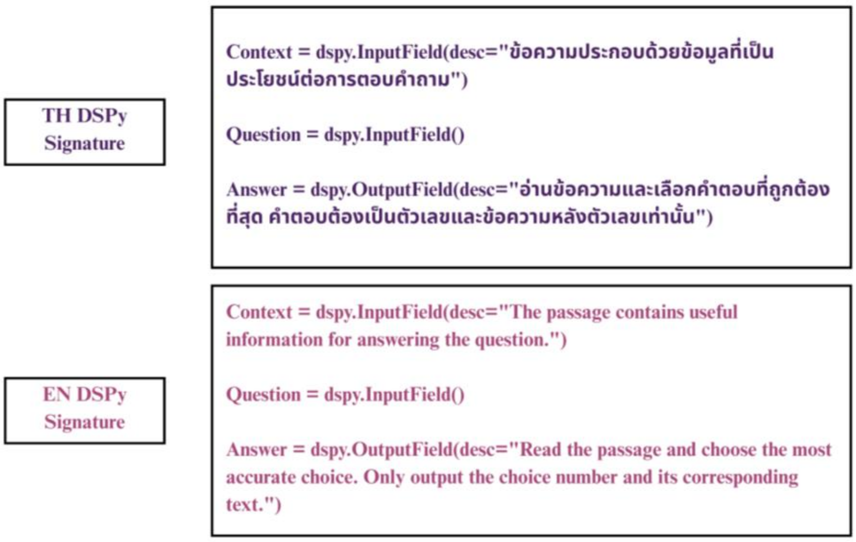

Figure 3. Structured DSPy Signatures in Thai and English for the DSPy Predict Method.

The **DSPy Chain of Thought (CoT)** method introduces a reasoning prefix—"การให้เหตุผล:" in Thai and "Reasoning:" in English—encouraging the model to articulate its thought process before arriving at an answer. This intermediate reasoning step aims to enhance performance by mimicking human-like problem-solving.

TH CoT Prompt

```
{TH DSPy Signature}

rationale_type = dspy.OutputField(
prefix="การให้เหตุผล: ", desc="Let's think step by step in Thai...")
```

EN CoT Prompt

```
{EN DSPy Signature}

rationale_type = dspy.OutputField(
prefix="Reasoning: ", desc="Let's think step by step in English...")
```

Figure 4. Chain of Thought (CoT) Signature in Thai and English for DSPy Program.

Finally, the **DSPy Prompt Optimization** method uses MIPROv2 to automatically generate and select effective instructions, both in zero-shot and few-shot settings. In the few-shot setting, 1–3 example questions are included to demonstrate the task. This method explores prompt-and-example combinations that optimize performance based on a target evaluation metric.

DSPy teleprompter

```
optimized_predictor = teleprompter.compile(
        base_predictor.deepcopy(), # DSPy CoT Module
        trainset=trainset, # 10 questions
        max_bootstrapped_demos=num_fewshot_samples # 0-3 shots
        max_labeled_demos=num_fewshot_samples # 0-3 shots
```

Figure 5: Compilation of the DSPy Teleprompter for Automated Generation and Selection of Optimized Prompts (Zero-Shot and Few-Shot).

Each of these prompting strategies is applied across all models and both Thai and English prompt languages, allowing for controlled comparisons of model behavior and performance under varying prompting conditions.

4.3. Experiment Setup

All experiments were conducted using a framework that integrates multiple prompting methods and large language models (LLMs) via API calls. The LLMs were accessed through OpenRouter.ai, which supports a unified interface for querying various commercial and open-source models. This enabled seamless evaluation across multilingual general-purpose models and Thai fine-tuned models.

The primary models evaluated include GPT-4o (OpenAI), Gemini Flash 2.0 (Google), and DeepSeek V3 for multilingual capabilities, as well as Typhoon2 8B and Typhoon2 70B, which are fine-tuned for Thai language comprehension. All models were prompted using a consistent structure in both Thai and English, depending on the language setting of the input.

Prompting techniques tested include:

1) Naive Prompting using static baseline instructions in Thai or English.
2) DSPy Predict, a method that standardizes model input/output using a structured signature template.
3) Chain-of-Thought Prompting (CoT), which instructs the model to reason step by step before answering.
4) DSPy Prompt Optimization, which automatically searches for optimal instructions and examples using MIPROv2, under both zero-shot and few-shot conditions (1–3 examples).

The evaluation dataset consists of 110 Thai-language reading comprehension questions compiled from three national-level exams: O-NET M3, O-NET M6, and A-Level + CU-TPT. The dataset was split into 100 evaluation questions and 10 training samples used for prompt optimization and few-shot prompting.

Model outputs were assessed using Exact Match Accuracy, requiring both the choice number and the corresponding answer text to exactly match the gold standard. All inference results were collected and stored for comparison across different models and prompting strategies.

5. Evaluation & Results

5.1. Metrics Used

The evaluation criterion employed in this study was Exact Match Accuracy. For an answer to be considered correct, it was required that both the selected multiple-choice option number and the corresponding answer text precisely matched the ground truth. This metric ensures a clear and unambiguous measure of the models' ability to not only identify but also correctly articulate the answer.

5.2. Results

The performance of different LLMs, including multilingual models (GPT-4o, Gemini Flash 2.0, DeepSeek V3) and Thai fine-tuned models (typhoon2 8B, typhoon2 70B), was evaluated across several Thai reading comprehension datasets (A-Level + CU-TEP, O-NET M3-M6)

1) Impact of Model Size on Accuracy

As illustrated in Figure 6 ("Model Size vs. Accuracy"), there is a clear correlation between model size and comprehension performance. Larger models, such as GPT-4o and typhoon2 70B, demonstrated higher and more consistent accuracy. Conversely, smaller models like typhoon2 8B exhibited lower and more variable accuracy scores. This suggests that model scale is a significant factor in achieving robust Thai reading comprehension.

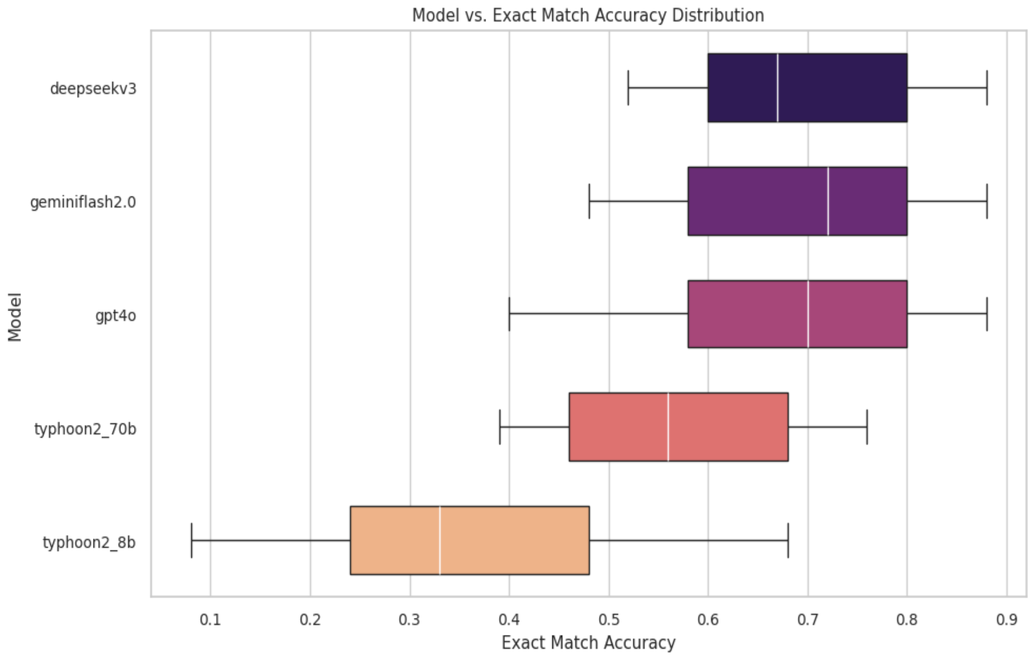

Figure 6. Model Size vs. Accuracy.

2) Comparison with Human Performance:

Figure 7 ("Average Model vs. Average Human Performance by Dataset") provides a comparative view of model performance against average human benchmarks.

When comparing average model performance with human scores across datasets, we observe that LLMs perform best on the O-NET-M3 (≈0.83) and O-NET-M6 (≈0.68) datasets, significantly outperforming average human scores—by up to 31% on O-NET-M3 (human: 0.52) and 25% on O-NET-M6 (human: 0.43). This performance gap suggests that O-NET questions are relatively less complex and more aligned with the capabilities of current LLMs, enabling them to excel well beyond typical human performance.

In contrast, the A-Level dataset presents a more challenging evaluation setting, with an average human score of 0.56 and model performance closely clustered around that mark. This narrower gap indicates that A-Level questions demand more advanced comprehension and reasoning, revealing the limitations of even the most powerful models when faced with nuanced or abstract reading tasks.

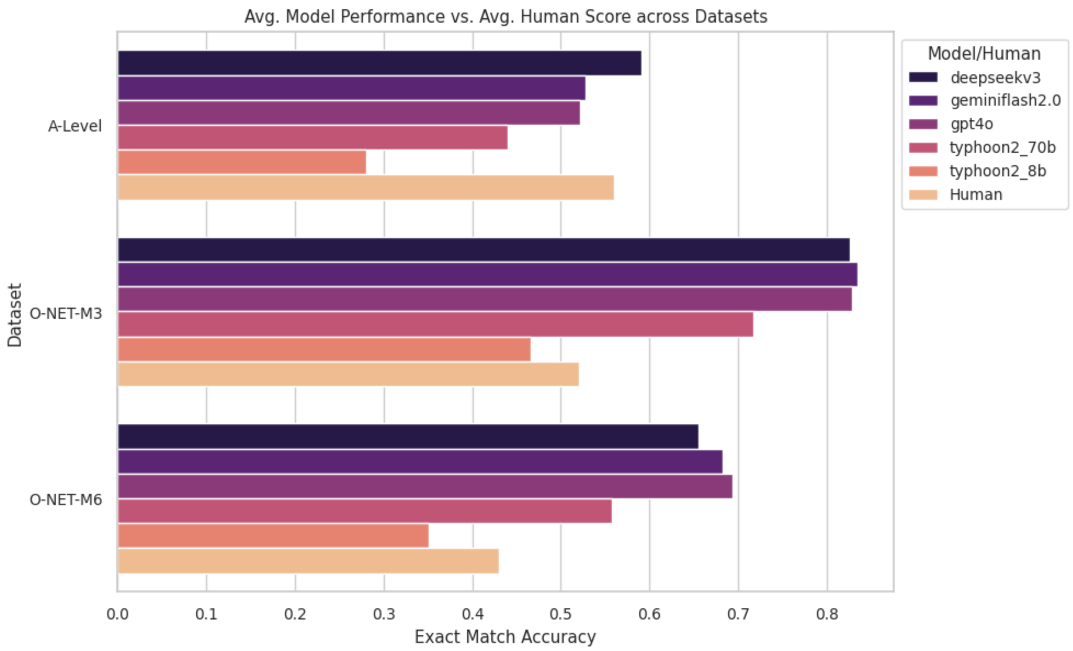

Figure 7. Average Model vs. Average Human Performance by Dataset.

3) Efficacy of Few-Shot Prompting

The introduction of few-shot examples via DSPy Prompt Optimization generally enhanced model accuracy, as shown in Figure 8. ("Few-Shot Examples vs. Accuracy").

Introducing few-shot examples generally improves model accuracy, especially for smaller models. The most notable improvement is observed in typhoon2_8b, which jumps from ~0.16 (zero-shot) to nearly 0.50 with just one example—an increase of over 34%. This underscores the impact of in-context learning for smaller models. In contrast, larger models like gpt4o and geminiflash2.0 show more consistent but modest gains, suggesting that while few-shot prompting helps across the board, its relative benefit is most pronounced for less capable models.

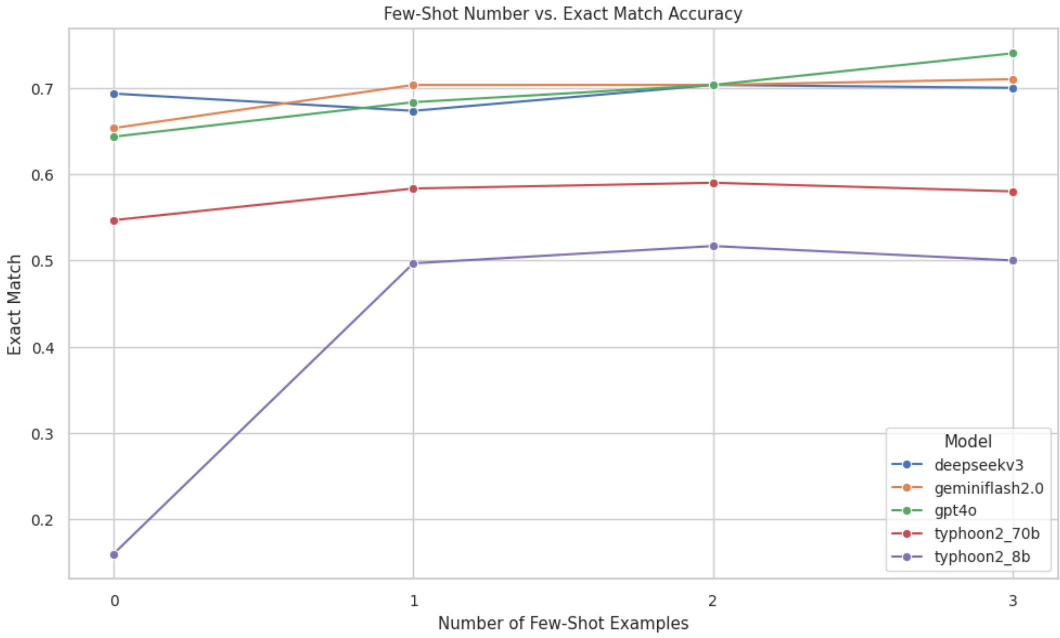

Figure 8. Few-Shot Examples vs. Accuracy.

4) Cross-Lingual Prompting and Method Performance

Figure 9. ("Model & Method Performance Across Languages") shows how different models and prompting methods perform when using English versus Thai prompts.

Model performance varies across languages and prompting methods, with some models showing notable differences. Gemini Flash 2.0 stands out for its consistency, maintaining nearly identical accuracy in both English and Thai, indicating strong multilingual generalization. GPT-4o, on the other hand, performs slightly better in Thai, particularly when using advanced prompting strategies like Chain of Thought, few-shot examples, and prompt optimization, suggesting enhanced reasoning capabilities in Thai under guided conditions. DeepSeekV3 displays the opposite trend, showing slightly stronger performance in English across most methods, which may reflect its training bias or better alignment with English-language instruction.

The Typhoon series reveals mixed patterns: typhoon2_70b maintains relatively consistent performance across both languages, with a small edge for English in some methods. Meanwhile, typhoon2_8b, the smallest of the models compared, shows a clear preference for English, especially in zero-shot and non-optimized methods, highlighting the challenges smaller models face in generalizing to lower-resource languages without additional context or examples.

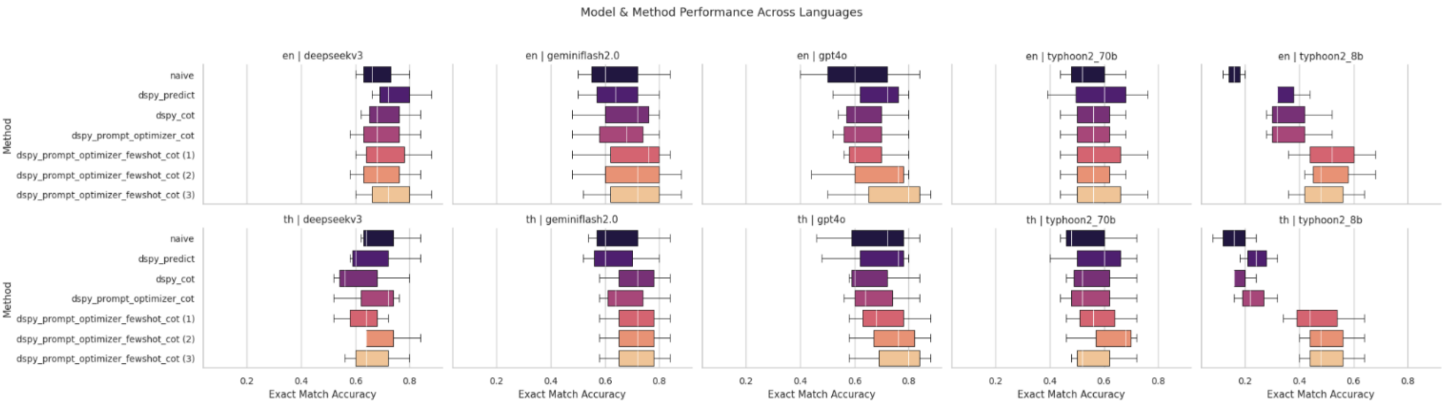

Figure 9. Model & Method Performance Across Languages.

5.3. Analysis / Discussion

The results indicate a clear hierarchy in model performance, strongly influenced by model scale. Larger models (GPT-4o, typhoon2 70B) consistently outperform smaller ones, likely due to their increased capacity to learn complex patterns and relationships within the Thai language, leading to better comprehension and reasoning.

The comparison with human performance offers valuable insights. Models excel or perform comparably on datasets perceived as less complex (O-NET, CU-TEP). This suggests that current LLMs are adept at tasks requiring straightforward information retrieval and basic inference. However, their underperformance on the A-Level datasets highlights a current limitation in tackling questions that demand more advanced reasoning, deeper contextual understanding, or nuanced interpretation, areas where human cognition still holds an advantage. The O-NET questions being "relatively less complex" as noted in the image aligns with this interpretation.

The significant positive impact of few-shot prompting, especially for smaller models like typhoon2 8B, underscores the value of in-context learning. Providing a few relevant examples allows models to better understand the task and desired output format, effectively "guiding" them towards more accurate responses. This is a crucial finding for optimizing performance when using models with fewer parameters. Larger models, while also benefiting, may already possess a more generalized understanding that makes the incremental gain from few-shot examples less pronounced.

The exploration of cross-lingual prompting strategies reveals interesting nuances. While models like Gemini Flash 2.0 show robustness across English and Thai CoT, the general trend of English CoT slightly outperforming Thai CoT (except for GPT-4o in some cases) might suggest that the models, even multilingual ones, could have a residual bias or slightly better internal representation for reasoning processes when guided in English. GPT-4o's stronger performance with Thai CoT could be attributed to its advanced multilingual capabilities and potentially better handling of complex instructions in Thai. The mixed results from Typhoon models suggest that fine-tuning and model architecture interact with prompting language in complex ways. The observation that Thai CoT performs slightly worse across most methods warrants further investigation into optimizing CoT for the Thai language specifically.

In summary, model size, task complexity, and prompting strategy are key determinants of LLM performance in Thai reading comprehension. While current models show promise, particularly on less complex tasks and when guided by effective prompting, further advancements are needed to match human proficiency in scenarios requiring deep reasoning. The effectiveness of few-shot learning provides a practical avenue for performance enhancement, especially for smaller models.

6. Conclusion & Future Work

6.1. Summary of Findings

This project aimed to assess the reading comprehension proficiency of various Large Language Models (LLMs) in the Thai language using standardized examination questions. The evaluation demonstrated a clear correlation between model scale and performance, with larger models like GPT-4o and Typhoon2 70B achieving higher and more consistent accuracy on Thai comprehension tasks compared to smaller models.

Our findings indicate that current LLMs perform well, sometimes exceeding average human performance, on datasets representing less complex reasoning tasks (e.g., O-NET). However, a performance gap persists relative to humans on more challenging datasets requiring advanced reasoning and nuanced understanding (e.g., A-Level).

Furthermore, the study highlighted the significant impact of prompting strategies. Few-shot prompting, particularly using the DSPy Prompt Optimization with few-shot examples, proved highly effective in boosting accuracy, especially for smaller models. Comparing English and Thai Chain of Thought (CoT) prompts revealed nuances, with English CoT often yielding slightly better results, although advanced models like GPT-4o showed strong performance with Thai prompts as well. This suggests that while multilingual models handle Thai reasonably well, optimizing prompting strategies specifically for the language remains an important consideration.

6.2. Limitations

1) Dataset Size: The evaluation was conducted using a relatively small dataset comprising 110 questions sourced from Thai examinations (A-Level, CU-TEP, O-NET). While sufficient for initial assessment, with 100 questions used for evaluation and 10 reserved for few-shot examples during prompt optimization, this limited size may not fully capture the breadth and depth of Thai reading comprehension challenges.
2) Model Scope: The selection of LLMs was constrained by availability through the OpenRouter AI platform, chosen for time and resource efficiency. This necessarily excluded other potentially relevant proprietary or open-source models that might exhibit different performance characteristics on Thai language tasks.

6.3. Future Work

1) Dataset Expansion and Diversification: A primary next step should be the collection and curation of a significantly larger and more diverse dataset of

Thai reading comprehension questions. This would enable more robust evaluation and potentially reveal model strengths and weaknesses across a wider range of text types and complexities.

2) Question Categorization and Sub-Skill Analysis: With a larger dataset, questions could be categorized based on the specific reading comprehension sub-skills they target. This would allow for a more granular analysis to determine which specific aspects of Thai reading proficiency LLMs excel at or struggle with. Potential categories include identifying the main idea ("จับใจความ"), interpretation ("ตีความ"), text sequencing ("เรียงลำดับข้อความ"), analysis ("วิเคราะห์"), and detail identification ("จับรายละเอียด"), among others. Such analysis would provide deeper insights into the cognitive abilities mimicked by the models.
3) Identifying Linguistic Skill from World Knowledge: Future investigations could aim to differentiate whether LLM success in Thai comprehension tasks stems primarily from genuine Thai linguistic understanding and processing skills or from leveraging vast amounts of general world knowledge accessible within their parameters. Designing specific test sets or methodologies to probe this distinction would be highly valuable for understanding the true nature of LLM capabilities in specific languages like Thai.
4) Exploring Broader Model Architectures and Fine-Tuning: Extending the evaluation to include a wider array of LLMs, including those not available on OpenRouter, and exploring domain-specific or task-specific fine-tuning techniques for Thai comprehension could yield further performance improvements.

References

Chang, K., Xu, S., Wang, C., Luo, Y., Liu, X., Xiao, T., & Zhu, J. (2024). *Efficient prompting methods for large language models: A survey*. arXiv. https://doi.org/10.48550/ARXIV.2404.01077

DeepSeek-AI. (2024). *DeepSeek-V3 technical report.* arXiv. https://doi.org/10.48550/ARXIV.2412.19437

DSPy. (n.d.). *DSPy: A framework for programming with foundation models.* https://dspy.ai/

Google DeepMind. (2024, December 6). *Gemini: Our largest and most capable AI model.* Google. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/#gemini-2-0-flash

Khattab, O., Singhvi, A., Maheshwari, P., Zhang, Z., Santhanam, K., Vardhamanan, S., Haq, S., Sharma, A., Joshi, T. T., Moazam, H., Miller, H., Zaharia, M., & Potts, C. (2023). *DSPy: Compiling declarative language model calls into self-improving pipelines.* arXiv. 1 https://doi.org/10.48550/ARXIV.2310.03714

OpenAI. (2024). *GPT-4o system card.* arXiv. https://doi.org/10.48550/ARXIV.2410.21276

OpenRouter. (n.d.). *OpenRouter: Unified API for accessing multiple LLMs.* https://openrouter.ai/

Pipatanakul, K., Manakul, P., Nitarach, N., Sirichotedumrong, W., Nonesung, S., Jaknamon, T., Pengpun, P., Taveekitworachai, P., Na-Thalang, A., Sripaisarnmongkol, S., Jirayoot, K., & Tharnpipitchai, K. (2024). *Typhoon 2: A family of open text and multimodal Thai large language models.* arXiv. https://doi.org/10.48550/ARXIV.2412.13702

A. An example of Optimized prompt

"You are a skilled Thai language comprehension expert. Your task is to read a given context and answer a related question by identifying the most accurate multiple-choice answer. Start by analyzing the context and question carefully. Use a step-by-step reasoning process to derive the answer, ensuring your choice is well-supported by logical conclusions drawn from the context. Output your reasoning and the final answer, selecting the choice number and its corresponding text that best fits the given context."

Figure 10. Optimized Instruction Prompt Generated by DSPy's MIPROv2 for Thai Reading Comprehension.

B. An example of model's response using DSPy CoT

context	**question**	**options**	**response**	**reasoning**
แม้จะมีอายุไม่ถึง ๑๐ ขวบดี เด็กหญิงพรชนกก็ช่วยเหลือตนเองได้เป็นอย่างดี เธอทำอาหารเช้าเอง และเตรียมใส่ปิ่นโตไปกินที่โรงเรียนตอนกลางวันด้วย ที่โรงเรียนเธอก็ทำกิจกรรมหลายอย่างและเป็นประธานชมรมหนูน้อยนักประดิษฐ์ของใช้จากเศษวัสดุ "Even though she is not yet 10 years old, girl Pornchanok can take care of herself very well. She makes her own breakfast and prepares it in a tiffin carrier (food container) to eat at school for lunch. At school, she also participates in	ตามข้อความต่อไปนี้ "เด็กหญิงพรชนก" มีลักษณะนิสัยหลายประการยกเว้นข้อใด "According to the following text, "Girl Pornchanok" has several character traits, except for which one?"	1. เอื้อเฟื้อ 2. ประหยัด 3. รับผิดชอบ 4. ใฝ่ใจรักโลก 5. มีมนุษยสัมพันธ์ "1. Generous 2. Thrifty 3. Responsible 4. Environmentally conscious	1. เอื้อเฟื้อ "1. Generous"	จากข้อความที่ให้มา เด็กหญิงพรชนกมีลักษณะนิสัยที่แสดงถึงความรับผิดชอบ เพราะเธอสามารถช่วยเหลือตนเองได้ดี ทำอาหารเช้าเอง และเตรียมอาหารไปโรงเรียน ซึ่งแสดงถึงความรับผิดชอบ (ข้อ 3) นอกจากนี้ การที่เธอเป็นประธานชมรมหนูน้อยนักประดิษฐ์ ของใช้จากเศษวัสดุ แสดงถึงความใฝ่ใจรักโลก (ข้อ 4) และความประหยัด (ข้อ 2) เนื่องจากเธอใช้เศษวัสดุในการประดิษฐ์สิ่งของ อย่างไรก็ตาม ข้อความไม่ได้กล่าวถึงการเอื้อเฟื้อ (ข้อ 1) หรือการมีมนุษยสัมพันธ์ (ข้อ 5) โดยตรง แต่การเป็นประธานชมรมอาจบ่งบอกถึงมนุษยสัมพันธ์ได้บ้าง ดังนั้น

many activities and is the president of the Little Inventors Club: Useful Items from Waste Materials."		5. Good interpersonal skills"		ข้อที่ไม่สามารถสรุปได้จากข้อความคือข้อ 1. เอื้อเฟื้อ "From the given text, Girl Pornchanok has character traits that demonstrate responsibility because she can take good care of herself, makes her own breakfast, and prepares food for school, which shows responsibility (Option 3). Additionally, her being the president of the "Little Inventors Club: Useful Items from Waste Materials" indicates environmental consciousness (Option 4) and thriftiness (Option 2), as she uses waste materials to create items. However, the text does not directly mention generosity (Option 1) or having good interpersonal skills (Option 5). While being a club president might suggest some interpersonal skills, the trait that cannot be definitively concluded from the text is Option 1: Generous."

Table 2. Example of Model Output with Reasoning with English Translation