

CP Capstone Project Proposal
รายงานโครงการวิศวกรรมคอมพิวเตอร์

เรื่อง

WASAN: A Thai Vision-Text Large Language Model

วสันต์: โมเดลภาษาขนาดใหญ่รูปแบบข้อความ และรูปภาพสำหรับภาษาไทย

โดย

กณวรรณ	วิลาศรี	รหัสนิสิต 6430001121
ณัฐดนัย	ตฤณธวัช	รหัสนิสิต 6431316021
พิเชษฐ	พ่วงรอด	รหัสนิสิต 6432114821
นิพัทธ์	เชนธนากิจ	รหัสนิสิต 6430215121
ญาณภัทร	พัชรวิวัฒน์พงษ์	รหัสนิสิต 6432090321

อาจารย์ที่ปรึกษา

ผศ.ดร.เอกพล ช่างสุวนิช

รายงานนี้เป็นส่วนหนึ่งของวิชา 2110489 โครงการรวบยอดวิศวกรรมคอมพิวเตอร์ 2
ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ประจำปีการศึกษา 2567

บทคัดย่อ

งานวิจัยนี้นำเสนอความพยายามสองแนวทางที่เสริมกันเพื่อพัฒนาโมเดลประสาทร่วมระหว่างภาพและภาษา (Vision-and-Language Models: VLMs) สำหรับภาษาไทย โดยเน้นการปรับปรุงการประเมินคุณภาพการแปลภาษาและการรู้จำข้อความจากภาพ (OCR) สำหรับใบเสร็จรับเงินเป็นหลัก ในส่วนแรก เรารวบรวมข้อความภาษาอังกฤษจากหลากหลายโดเมนและทำการแปลเป็นภาษาไทย จากนั้นสร้างชุดข้อมูลที่มีการใส่คำอธิบายโดยมนุษย์ตามกรอบการประเมินข้อผิดพลาด MQM (Multidimensional Quality Metrics) โดยใช้ระบบแปลภาษาอัตโนมัติ 10 ระบบในการสร้างผลลัพธ์การแปล ซึ่งถูกนำมาใช้ฝึก COMETH ซึ่งเป็นโมเดลประเมินคุณภาพการแปลที่ต่อยอดจาก COMET เราทำการเปรียบเทียบค่าความสัมพันธ์ของลำดับ (Spearman correlation) ระหว่างผลลัพธ์จาก COMETH, COMET ดั้งเดิม และโมเดลภาษาขนาดใหญ่ (LLMs) อีก 3 ตัว ได้แก่ Gemini 2.0 Flash, GPT-4o Mini และ Claude 3.5 Sonnet กับผลการประเมินของมนุษย์ พบว่าในกลุ่ม LLMs นั้น Claude 3.5 Sonnet มีความสอดคล้องกับมนุษย์มากที่สุด อย่างไรก็ตาม COMETH เวอร์ชันที่ถูกฝึกด้วยข้อมูลทั้งจากมนุษย์และ Claude 3.5 Sonnet ให้ผลลัพธ์ที่ใกล้เคียงการตัดสินของมนุษย์มากที่สุดโดยรวม

ในส่วนที่สอง เราศึกษาการใช้เทคนิคปัญญาประดิษฐ์แบบกำเนิด (Generative AI) และการประมวลผลภาพเพื่อสร้างชุดข้อมูลใบเสร็จอิเล็กทรอนิกส์และใบค่าใช้จ่ายแบบสังเคราะห์สำหรับการปรับแต่งโมเดล OCR เราใช้ชุดข้อมูลสังเคราะห์นี้ในการปรับจูนโมเดล Qwen2.5-VL-Instruct แม้ว่าความแม่นยำในการรู้จำข้อความโดยตรงจะเพิ่มขึ้นไม่มากนัก แต่โมเดลสามารถเข้าใจโครงสร้างและรูปแบบของเอกสารได้ดีขึ้นอย่างมีนัยสำคัญ ความพยายามทั้งสองแนวทางนี้มีจุดมุ่งหมายร่วมกันในการผลักดันการพัฒนาโมเดล VLM ที่รองรับภาษาไทย โดยเฉพาะในด้านการประเมินการแปลที่มีคุณภาพสูงและระบบ OCR สำหรับใบเสร็จที่แข็งแกร่งและแม่นยำยิ่งขึ้น

CP Capstone Final Project

WASAN: Thai Vision Language Model

May 2025

Department of Computer Engineering Faculty of Engineering
Chulalongkorn University

Submitted by: WASAN Research Team

Under the supervision of: Ekapol Chuangsuwanich, Ph.D.
Department of Computer Engineering

This page intentionally left blank.

WASAN: Thai Vision Language Model

Kanawat Vilasri, Natdanai Trintawat, Phichet Phuangrot, Nipat Chenthanakij,
Yanapat Patcharawiwatpong, and Ekapol Chuangsuwanich

Department of Computer Engineering, Faculty of Engineering, Chulalongkorn
University

May 2025

Abstract

This work presents two complementary efforts aimed at advancing Thai-language vision-and-language models (VLMs), focusing on improving machine translation evaluation and optical character recognition (OCR) for receipts. In the first part, we collect English text from diverse domains and translate it into Thai, creating a human-annotated dataset using the Multidimensional Quality Metrics (MQM) framework. Ten machine translation (MT) systems are used for translations, and the resulting data is used to train COMETH, an MT evaluation model based on COMET. We compare the system-level Spearman correlation of COMETH, baseline COMET, and three large language models (LLMs)—Gemini 2.0 Flash, GPT-4o Mini, and Claude 3.5 Sonnet—against human MQM annotations. Among the LLMs, Claude 3.5 Sonnet achieves the highest correlation, but our augmented COMETH model, trained on both human and LLM-generated annotations, outperforms all models in alignment with human judgments.

In the second part, we explore generative AI and image processing techniques to synthesize a dataset of e-receipts and expense bills for OCR fine-tuning. Using this artificial dataset, we fine-tune the Qwen2.5-VL-Instruct model. While raw OCR accuracy shows limited improvement, the model gains enhanced understanding of document structure and formatting. Together, these two lines of work aim to support the development of a Thai-capable VLM, with a particular emphasis on high-quality translation evaluation and robust receipt OCR.

1 Introduction

1.1 Background

In recent years, Vision-Language Models (VLMs) represent the forefront of AI technology, showcasing impressive capabilities across a wide range of tasks, such as Visual Question Answering (VQA) [1] and document understanding. Despite these advancements, most VLMs are designed for high-resource languages, with limited attention given to Thai. Existing Thai VLMs, such as Llama-3 Typhoon Instruct Vision Preview [12] and Pathumma-llm-vision-1.0.0 [13], represent early efforts in this domain. However, these models primarily focus on general language tasks and are trained on datasets that lack coverage for Thai-specific vision tasks and localized applications, such as interpreting e-receipts or receipts. These scenarios, which involve unique formats and language structures, underscore the critical gap in adapting state-of-the-art VLM technology to support Thai, a low-resource language with specialized needs.

To bridge this gap, we plan to translate English-based datasets, such as Bunny [4], VQA [1], and GQA [6], into Thai and generate data specific to Thai contexts, such as e-slips and receipts. However, this translation process requires a better Thai translation evaluation system. The current evaluation systems, often based on general-purpose metrics, do not fully capture the nuances of Thai. Therefore, developing an effective evaluation system for Thai translation is a crucial step in improving the performance of Thai VLMs and advancing AI capabilities in this area.

1.2 Objective

This study aims to develop a Thai Vision-Text Large Language Model that can read and understand images, process Thai text, respond in Thai and understand data from Thai bank receipt and expense receipts.

1.3 Scope of work

1. Develop a Machine Translation Evaluation Model that supports the Thai language, aimed at enhancing the accuracy of evaluating Thai translation outputs. This will contribute to the advancement of the Thai Vision-Text Multimodal Model by ensuring better translation quality assessments.
2. Create a tool to generate datasets for both bank receipts and general receipts, which will be utilized to enhance the Vision-Text Multimodal Model’s capability to process and interpret data from these receipts.
3. Fine-tune the Thai Vision-Text Large Language Model and design evaluation metrics to measure its performance, ensuring the model accurately processes vision-text data and generates fluent Thai text.

1.4 Methods

1. Evaluation Model

- (a) Collect domain-specific data for evaluation.
- (b) Select appropriate machine translation systems for comparison.
- (c) Develop a platform for annotators to review and score translations.
- (d) Fine-tune the evaluation model using the annotated data to improve accuracy.

2. Receipt Data Generation

- (a) Generate a dataset of synthetic bank e-receipts.
- (b) Generate a dataset of synthetic expense receipts.
- (c) Create a pipeline for processing and fine-tuning the model using the generated datasets.

3. Thai Vision-Text Large Language Model

- (a) Collect and translate vision-text data into Thai using the evaluation model to assess translation quality, including the receipt datasets.
- (b) Implement a training pipeline to fine-tune the Vision-Text Large Language Model on the processed dataset.

- (c) Develop evaluation metrics to assess the model’s performance based on accuracy, fluency, and relevance of the generated Thai text.

4. Summarize Results

- (a) Analyze and summarize the evaluation and model training results, highlighting improvements and identifying areas for further development.

1.5 Project plan

Figure 1 shows the Gantt chart, which provides a detailed timeline of the project’s tasks and milestones.

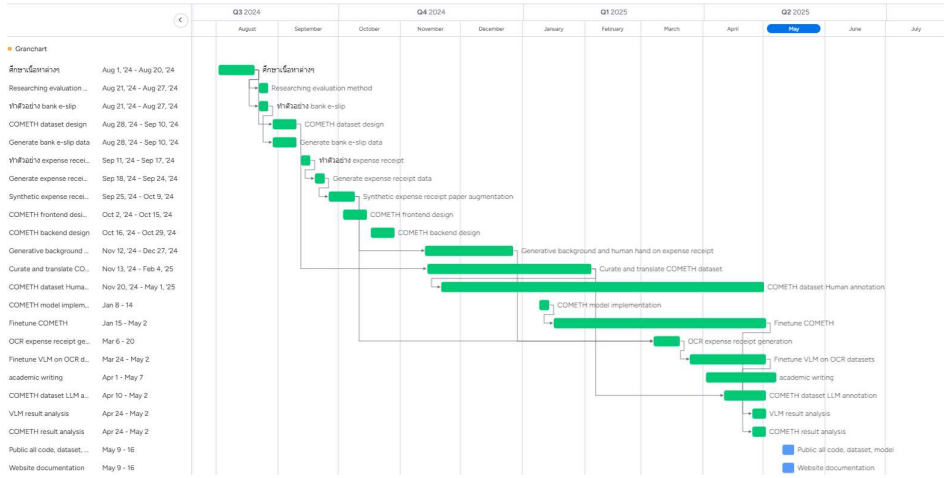


Figure 1: Gantt Chart

1.6 Benefit

The expected benefits of this project include the development of an artificial intelligence system capable of accurately reading and interpreting data from financial document images, such as bank e-receipt and receipts. Additionally, it involves the creation of a Thai Vision-Text Multimodal Model that supports processing both image and text data in the Thai language.

2 Related Work

2.1 Relevant Theories

2.1.1 Automatic Machine Translation Evaluation Metrics

Machine translation evaluation metrics are a critical component for assessing the performance of translation systems. These metrics provide a standardized method for comparing different translation systems and tracking their improvements over time. Additionally, they play a crucial role in guiding development by identifying specific strengths and weaknesses within a translation system.

Automatic machine translation evaluation metrics perform the evaluation task automatically, allowing evaluation in a much larger scale. In addition to a candidate translation, these metrics often require a reference text, usually a human translation, which is compared against to compute a score. The score indicates how close the candidate translation is to the provided reference.

Traditional evaluation metrics, such as BLEU [11], primarily focus on comparing n-gram similarities between a candidate translation and a reference translation. These traditional methods, however, lack the ability to effectively capture semantic nuances and contextual meaning, resulting in a subpar correlation with human judgements.

With advancements in natural language processing, text embedding models such as BERT have been developed to better capture semantic information. Consequently, recent developments in translation evaluation metrics often incorporate these models to go beyond simple word matches. Notable examples of which include BLEURT [17] and COMET [14].

Our work mainly improve upon the COMET model [14], specifically COMETKiwi [15]. COMET leverages cross-lingual pre-trained language models, such as XLM-R, enabling it to handle multiple languages effectively. The model takes three inputs: the source text, the candidate translation, and the reference translation. It is trained on multiple datasets, including Direct Assessments (DA), Multidimensional Quality Metrics (MQM), and Human-mediated Translation Edit Rate (HTER). By incorporating the source text in the evaluation process and utilizing pre-trained language models, COMET achieves remarkable performance and demonstrates a strong correlation with human judgments.

2.1.2 Quality Estimation Model

A typical machine translation evaluation setup requires a reference translation to calculate a quality score. However, an alternative approach exists that evaluates translation quality solely based on the source text and the candidate translation. This approach is known as Quality Estimation.

Quality Estimation is particularly useful for identifying translations that may require additional post-editing effort, enabling efficient quality control without the need for human-provided reference translations. This enhances the overall efficiency of the development process. Other applications include facilitating the selection of translation systems and enabling real-time quality evaluation during translation. Recent advancements in Quality Estimation have been showcased in the WMT Quality Estimation Task, where models such as CometKiwi have been developed.

2.1.3 Relative Ranking Metrics [14]

A variant of COMET metrics computes quality scores using pairwise data consisting of a worse candidate translation and a better candidate translation. The model is trained with a contrastive loss function, which minimizes the distance between the source text and the better candidate (as well as the reference translation) while maximizing the distance to the worse candidate, as illustrated in Figure 2.

This approach has demonstrated a strong correlation with human judgment and offers a simplified annotation process that only requires the selection of better translation from a pair. However, there is a notable gap in the research: no significant work has been done on leveraging ranking data for quality estimation models. Additionally, the potential of listwise comparison, which considers multiple translations simultaneously rather than just pairs, remains unexplored.

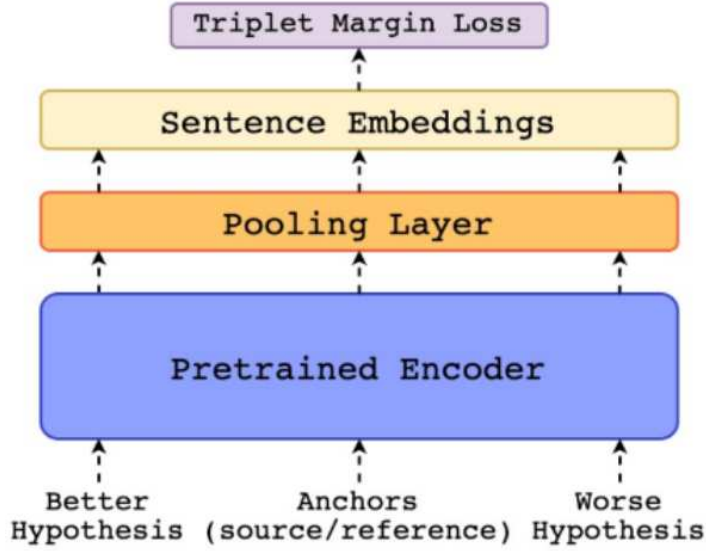


Figure 2: Relative Ranking Model Architecture

2.1.4 Correlation Analysis

To evaluate the accuracy of machine translation metrics and quality estimation models, correlation analysis is employed to determine the strength of the association between the model’s quality scores and the ground truth provided by human annotators. A commonly used method in machine translation research is Spearman’s correlation, which assesses the rank-based relationship between two variables.

2.1.5 Image Processing

Data augmentation is an image processing technique that generates new images from existing ones, improving the training dataset to make expense receipt processing more realistic. Figure 3 shows a comparison of receipts with augmentation and without augmentation.

2.1.6 Augraphy [3]

Augraphy is a Python library designed for creating data augmentation pipelines that simulate common distortions found in real-world document images. The augraphy can create standard office operations like printing, scanning, faxing, ink degradation, and handwritten markings. Augraphy provides 26 unique augmentations, which may be sequenced into pipeline objects which carry out the image manipulation. Figure 4 shows the individual phases of an example pipeline combining to produce a noised document image.

2.1.7 ControlNet [20]

ControlNet is a neural network that adds spatial conditioning—like edges, depth, or pose—to pretrained text-to-image diffusion models such as Stable Diffusion [16]. It works by connecting to the existing model using zero-initialized convolution layers, ensuring safe and stable fine-tuning



Figure 3: augmentation (left) and without augmentation (right)

without disrupting the original model. ControlNet can handle single or multiple conditions, with or without text prompts, and works well on both small and large datasets. This enables more precise and versatile control over image generation.

2.1.8 Character Error Rate (CER)

Character Error Rate (CER) is a metric used to evaluate the performance of text recognition systems, like those in automatic speech recognition (ASR) or optical character recognition (OCR), by measuring the percentage of incorrectly predicted characters. It quantifies the difference between the predicted text and the correct reference text, considering insertions, deletions, and substitutions. A lower CER indicates a better performance. The CER was computed for each translation output using the following formula:

$$\text{CER} = \frac{S + D + I}{N} \quad (1)$$

where:

- S is the number of substitutions,
- D is the number of deletions,
- I is the number of insertions, and
- N is the total number of characters in the reference (ground truth) text.

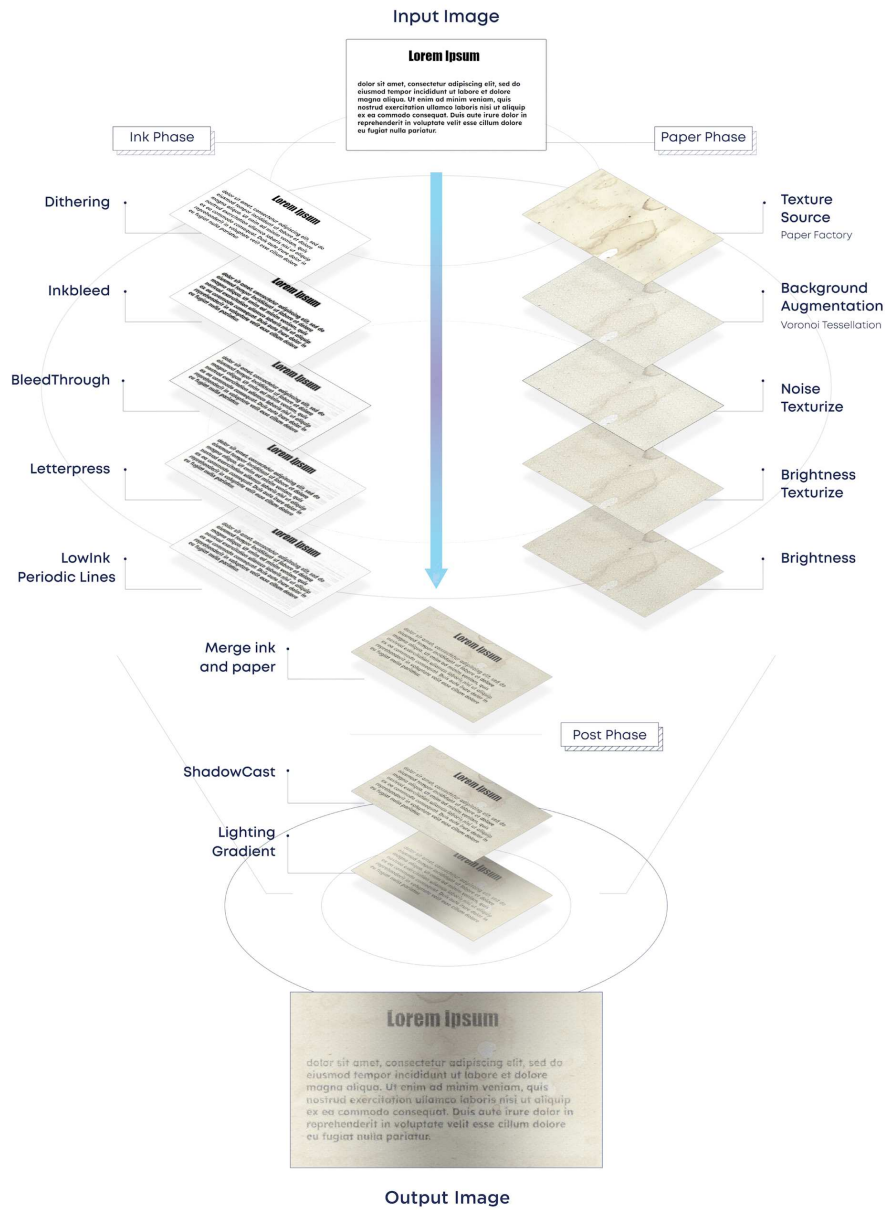


Figure 4: Visualization of an Augraphy pipeline, showing the composition of several image augmentations together with a specific paper background

2.1.9 Word Error Rate (WER)

Word Error Rate (WER) is a metric used to measure the accuracy of speech-to-text systems, often used in automatic speech recognition (ASR) and machine translation. It represents the

percentage of words that are incorrectly transcribed or translated compared to a reference transcript. The WER was computed for each translation output using the following formula:

$$\text{WER} = \frac{S + D + I}{N} \quad (2)$$

where:

- S is the number of substitutions,
- D is the number of deletions,
- I is the number of insertions, and
- N is the total number of words in the reference (ground truth) text.

2.1.10 Low-Rank Adaptation of Large Language Models [5]

Low-Rank Adaptation (LoRA) is a technique designed to make the fine-tuning of large-scale models more efficient. It involves freezing the pretrained model weights while introducing trainable low-rank decomposition matrices into each layer of the Transformer architecture. This allows the model to adapt to downstream tasks without the need to retrain all parameters, significantly reducing the computational cost and memory requirements compared to full fine-tuning.

For large models, such as Qwen-Vision, which consists of billions of parameters, fine-tuning all parameters can be prohibitively expensive. By using LoRA, only a small number of additional parameters are introduced, making it possible to fine-tune the model efficiently without sacrificing performance. This reduces both the cost and time required for training, making it a scalable and cost-effective solution for adapting large vision-language models to specific tasks and domains.

2.1.11 MQM Scores

We evaluated the performance of 10 machine translation systems using two key metrics: MQM scores derived from data annotation. Below, we detail the calculation methods and the comparative results.

The MQM score was computed for each translation output using the following formula:

$$\text{MQM Score} = 100 - \frac{(I_{\text{minor}} + 5 \cdot I_{\text{major}} + 10 \cdot I_{\text{critical}})}{\text{Sentence Length}} \cdot 100$$

where:

- I_{minor} , I_{major} , and I_{critical} represent the number of minor, major, and critical errors identified in the translation.
- Sentence Length is the number of tokens in the translated sentence.
- The weight factors (1 for minor, 5 for major, and 10 for critical errors) reflect the relative severity of each error type.

The MQM scores were averaged across all sentences for each system to calculate the overall MQM performance score.

2.1.12 Relative Ranking Calculation

Relative rankings were determined using annotated listwise rankings provided by human evaluators. For each input sentence, evaluators ranked all 10 systems based on translation quality, considering fluency, adequacy, and overall error severity. The ranks were aggregated across all sentences to compute the average rank for each system, providing a holistic view of their performance.

2.1.13 Spearman’s Rank Correlation

Spearman’s rank correlation coefficient, denoted as ρ or r_s , is a non-parametric measure of statistical dependence between two ranked variables. Unlike Pearson correlation, which captures linear relationships and assumes normally distributed data, Spearman’s correlation assesses the strength and direction of a monotonic relationship—whether linear or not—between two variables by comparing their rank orders. It is particularly well-suited for scenarios where the data may not meet the assumptions of parametric tests or where only ordinal information is available.

Mathematically, Spearman’s correlation is defined as the Pearson correlation coefficient between the ranked variables. For a set of n paired observations (x_i, y_i) , the formula is:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where d_i is the difference between the ranks of each pair of observations. The value of ρ ranges from -1 (perfect negative correlation) to $+1$ (perfect positive correlation), with 0 indicating no correlation.

In our evaluation, we apply Spearman’s rank correlation to compare the MT outputs as scored by various evaluation methods—including LLM-based MQM annotation, baseline COMET and COMETH variants—with those produced by human annotators and LLM annotation. This approach allows us to measure how closely each system’s output aligns with human judgments, providing a robust indicator of the system’s ability to replicate human-like quality assessments.

2.2 Relevant Papers

2.2.1 COMETKiwi [15]

COMETKiwi is an advanced extension of the COMET framework designed to evaluate machine translations with greater semantic and contextual precision. It uses the XLM-R cross-lingual pre-trained language model and incorporates human-annotated datasets, including MQM and Direct Assessments (DA), to train its scoring mechanism. The model excels at identifying semantic errors and contextual mismatches, outperforming traditional n-gram-based metrics like BLEU in capturing translation quality.

In our methodology, we adopt COMETKiwi as the backbone model for developing our own translation evaluation system. COMETKiwi’s pre-trained parameters provide a strong starting point, which we fine-tune using our custom dataset, COMETH, to further align with our specific evaluation needs. This customization allows us to optimize the model for dimensions critical to our tasks, such as semantic adequacy and fluency.

2.2.2 WMT 23 (Workshop on Machine Translation 2023) [8]

WMT 23 is a key resource for machine translation research, providing multilingual corpora with rich linguistic diversity. Among the datasets offered, the English corpus is notable for its

wide-ranging vocabulary and contextual variations. However, these datasets primarily focus on general-purpose evaluation and lack specific annotations for certain language pairs like English-Thai. In our methodology, we extract data from the WMT 23 English corpus, selecting samples with high lexical diversity to ensure a comprehensive representation of linguistic styles. We then translate these English samples into Thai and annotate the resulting English-Thai pairs using the MQM framework. These annotated pairs form a significant portion of our COMETH dataset, which is specifically designed to improve English-Thai translation evaluation. This focus allows us to address the unique challenges posed by Thai, such as tonal variations and complex grammatical structures

2.2.3 Multidimensional Quality Metrics (MQM) [10]

Multidimensional Quality Metrics (MQM) is a comprehensive evaluation framework that categorizes translation errors across dimensions such as fluency, adequacy, grammar, and style [multidimensional]. It provides detailed and human-aligned quality scores, making it a gold standard for evaluating translations.

In our work, MQM serves as the annotation framework for the English-Thai translation pairs in the COMETH dataset. By applying MQM to these pairs, we ensure that our model is trained to evaluate translations in a way that aligns with human judgments, focusing on dimensions critical to the English-Thai language pair. This process allows our fine-tuned COMETKiwi model [15] to deliver precise and reliable assessments of English-Thai translations, particularly in capturing cultural and contextual nuances.

2.2.4 Listwise Ranking [18]

The listwise methodology provides a robust framework for ranking tasks. The approach transitions from evaluating pairwise comparisons, which assess relative quality between two items, to listwise comparisons, which optimize rankings over entire lists [18]. This transition ensures higher consistency and improved performance in ranking-based applications.

In our work, we integrate the listwise approach with MQM annotations to rank translation systems effectively. By leveraging this ranking method, we generate a hierarchy of translation outputs based on their MQM scores, ensuring an evaluation framework that prioritizes holistic quality over isolated pair comparisons. This technique is particularly impactful in assessing English-Thai translations, where subtle nuances like tonal accuracy and contextual appropriateness play a critical role in overall ranking.

2.2.5 Typhoon-Vision [12]

Typhoon-Vision is a multimodal AI model designed specifically for processing both text and vision tasks in Thai. The model is built upon Typhoon 1.5 8B Instruct and the SigLIP vision encoder, with a total of 8.5 billion parameters to support tasks involving both image and language understanding in the Thai context. The architecture of Typhoon-Vision uses a 2-layer GELU MLP that connects the vision and language components, inspired by the LLaVA architecture.

The training data used to develop the Vision-Language Model (VLM) for Thai includes a variety of datasets that have been translated into Thai using ChatGPT-4. Notable datasets include:

Bunny Dataset [4]: Used for both pre-training and fine-tuning, divided into two main subsets:

Bunny-pretrain-LAION-2M: A filtered version of the larger LAION-2B dataset for more efficient learning.

Bunny-695K: A dataset sourced from various other datasets based on SVIT-mix665K.

WizardLM-evol-instruct-70K [19]: A dataset replacing ShareGPT-40K within the SVIT-mix665 to improve performance.

Typhoon Self-Instruct: Approximately 10,000 examples from this dataset were used to maintain text-only performance.

Wang Handwritten OCR: A dataset related to handwritten text (OCR).

Synthetic OCR Dataset: A custom dataset created using Synhtiger, with about 50,000 examples from OCR data.

This model is specifically designed to handle Thai language and vision tasks, with a strong focus on image and text understanding within the Thai context.

2.2.6 Qwen2.5-VL [2]

Qwen 2.5-VL is a multimodal large language model developed by Alibaba Cloud’s Qwen team. Building on the foundation of the Qwen 2.5 series, it integrates advanced vision-language processing, allowing it to interpret and generate content across text, images, and videos. The model is available in multiple configurations, including 3B, 7B, 32B, and 72B parameters.

Key Features of Qwen 2.5-VL:

- **Multimodal Understanding:** Capable of processing and interpreting text, images, and videos, facilitating tasks such as visual question answering, document OCR, and object detection.
- **Multilingual Capabilities:** Supports a wide range of languages, including English, Chinese, and Thai, allowing it to understand and generate text in various languages, making it versatile for global applications.
- **Enhanced Vision Capabilities:** Demonstrates significant advancements in general image recognition, expanding the categories of images it can identify, including plants, animals, landmarks, and various products.
- **Structured Data Extraction:** Supports structured outputs from documents like invoices, forms, and tables, making it beneficial for applications in finance and commerce.
- **Dynamic Resolution Processing:** Introduces dynamic resolution processing and absolute time encoding, enabling it to handle images of varying sizes and videos of extended durations, with second-level event localization.

2.2.7 GEMBA MQM [7]

The paper “GEMBA-MQM: Detecting Translation Quality Error Spans with GPT-4” by Tom Kocmi and Christian Federmann introduces GEMBA-MQM, a GPT-4-based evaluation metric designed to identify translation quality errors without relying on human reference translations. Utilizing a fixed three-shot prompting approach, GEMBA-MQM instructs GPT-4 to mark error spans according to the Multidimensional Quality Metrics (MQM) framework. A notable feature of this method is its language-agnostic prompt design, eliminating the need for manual prompt customization across different languages. Experimental results demonstrate that

GEMBA-MQM achieves state-of-the-art accuracy in system-level rankings, outperforming traditional metrics such as COMET and BLEURT. However, the authors advise caution in academic applications due to the proprietary nature of GPT-4, which poses challenges related to transparency, reproducibility, and potential fluctuations in model performance over time.

3 Methodology

This data generation process focuses on creating e-slip and expense receipt due to the lack of available Thai bank e-slip and expense receipt datasets.

3.1 Receipt Data generation

This data generation process focuses on creating e-slip and expense receipt due to the lack of available Thai bank e-slip and expense receipt datasets.

3.1.1 Bank e-receipt datasets

Create fake bank e-receipt by using photoshop to remove text from real bank e-receipt obtained online, creating editable templates for each bank. Additionally, build a generator for each bank, enabling automated receipt generation in the specified formats. Figure 5 shows examples of fake e-receipt from SCB, Krungthai, and KBank. The generator supports receipt generation for the following banks and formats.



Figure 5: SCB, Krungthai ,Kbank fake bank e-slip.

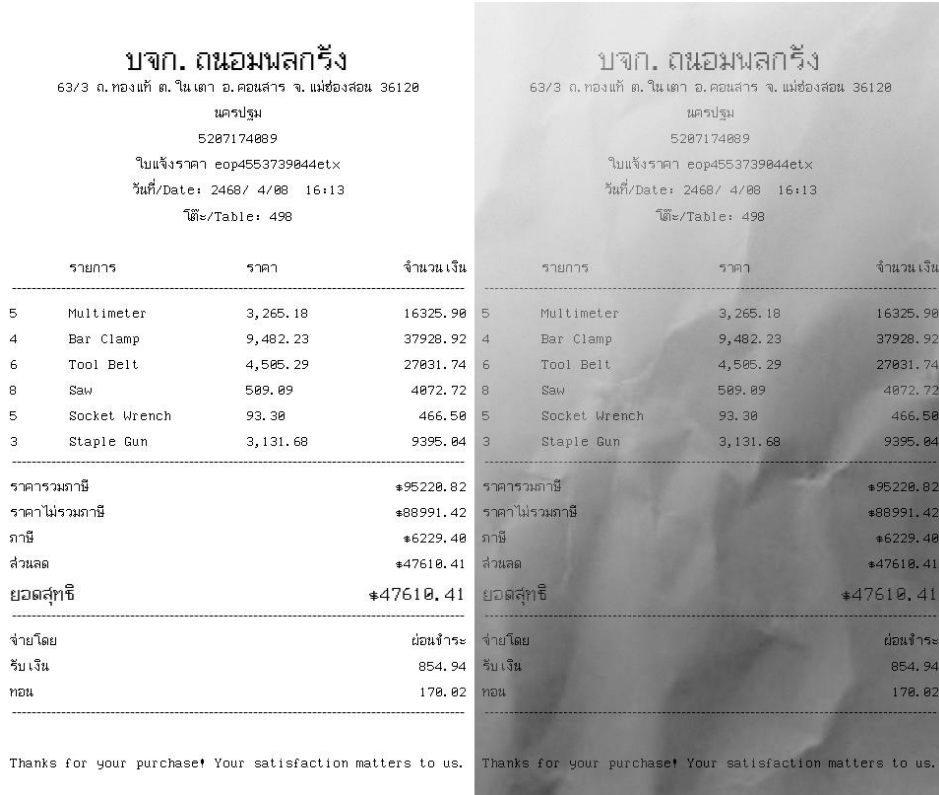
Bank	Formats	Types
SCB (Siam Commercial Bank)	plain	biller, note, info, default
KBank (Kasikorn Bank)	transfer, payment, top-up	default
Krungthai Bank	compact, full, plain	default, note

3.1.2 Expenses receipt datasets

To generate synthetic expense receipt datasets supporting both Thai and English, we implement a multi-stage pipeline. First, we use the OpenCV library in Python to create artificial receipt

images with a structured layout, incorporating typical elements such as the store name, date, itemized purchases, prices, VAT, and total amount. These synthetic receipts are then enhanced to improve realism through various preprocessing techniques. This includes simulating common receipt characteristics such as blurring, scaling, color shifts, random noise, geometric distortions, and shadow effects—mimicking artifacts commonly found in real-world printed receipts. To further enhance visual authenticity, the receipts are blended with paper textures, producing a realistic textured appearance, as illustrated in Figure 6.

In the final stage, we use generative AI based on Stable Diffusion prompts to generate synthetic backgrounds and human hands interacting with the receipts. This generative augmentation creates scenes that resemble real-life scenarios, as shown in Figure 7.



รายการ	ราคา	จำนวนเงิน
5 Multimeter	3,265.18	16325.98
4 Bar Clamp	9,482.23	37928.92
6 Tool Belt	4,585.29	27831.74
8 Saw	589.89	4872.72
5 Socket Wrench	93.38	466.58
3 Staple Gun	3,131.68	9395.84
รวมภาษี		95228.82
รวมไม่รวมภาษี		88991.42
ภาษี		6229.48
ส่วนลด		47610.41
ยอดสุทธิ		47610.41
จ่ายโดย		ผ่อนชำระ
รับเงิน		854.94
ทอน		178.82

Thanks for your purchase! Your satisfaction matters to us.

Figure 6: Synthetic expense receipts generated with a structured layout and realistic visual effects

3.2 COMETH Datasets

The data selection process focuses on ensuring lexical diversity and eliminating redundant or low-quality samples. Techniques that solve the Maximum Coverage Problem are employed to achieve this goal. The number of segments for each domain is given in Table 1.

3.2.1 Data Selection Process

To construct the dataset, we first identify the domains we want to cover based on the target application or research focus. Once the domains are selected, we locate suitable data sources

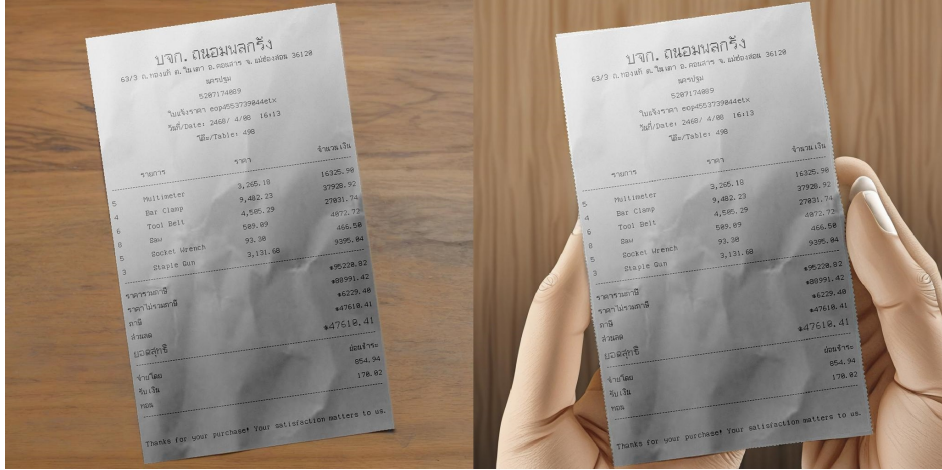


Figure 7: Generative ai generated an expense receipts with background and realistic human hands

Source	Domain	#segments
wmt-mqm-train-22	Conversation	289
	Social	280
	News	247
	E-commerce	184
bunny	E-commerce	125
vqa	Smart city	125
gqa	Smart city	125
ted	Education	125
khan	Education	230
treaty	Officials	250
WHO	Medical	250
manuals	Manuals	250
Total		2480

Table 1: Dataset Overview by Source and Domain

that align with these domains, ensuring that the data is diverse and relevant as measured by lexical diversity. Finally, we specify the desired number of segments to extract from each source to ensure a balanced representation across domains.

Regarding lexical diversity, we select a subset of sentences that maximizes the number of unique words. This is an instance of Maximum Coverage Problem, a known NP-hard problem. We utilized an approximated greedy algorithm, selecting the set that results in largest union cardinality at each step. The word present in each sentence is represented as Bitmap implemented with Roaring [9] library.

3.2.2 Word Count Distribution

The distribution of word counts across datasets is analyzed to maintain consistency and address potential biases in sentence lengths. Figure 11 shows the comparison of word count distributions

between the selected datasets and the WMT22 dataset.

3.2.3 System Selection

The COMETH platform incorporates diverse translation systems, carefully selected to represent different approaches to machine translation with emphasis on Thai language processing. Systems are categorized into three tiers - high, mid, and low performance - enabling comprehensive evaluation across different capability levels. The results are summarized in Table 2.

Model	Tier	Cost	Characteristics
GPT-4o-mini	High	Paid	High performance, requires API
Claude 3.5 Sonnet	High	Paid	High performance, requires API
LLaMa3-8b-WangchanX-sft-Full	High	Free	Open-source, fine-tuned for Thai tasks
Typhoon-v1.5x-70b-instruct	Mid	Free	Designed for Thai language
Qwen2.5-72B-Instruct	Mid	Free	Open-source, strong multilingual support
Grok-beta	Mid	Free	Grok 2 with better performance
Gemma2-9b-cpt-sea-lionv3-instruct	Mid	Free	Specialized in SEA languages
ggt-sheet	Low	Free	Lightweight, limited features
nllb-200-1.3B	Low	Free	Less-common languages
Openthaigpt1.5-72b-instruct	Low	Free	Thai LLM based on Qwen 2.5

Table 2: Translation Model Comparison

3.2.4 Model Selection Strategy

In the high-performance tier, GPT-4o-mini and Claude 3.5 Sonnet were selected for their exceptional multilingual capabilities and proven track record with Thai language tasks. While these models require paid API access, they provide a robust baseline for optimal translation performance. LLaMa3-8b-WangchanX-sft-Full, an open-source model specifically fine-tuned for Thai language tasks, represents high-performance capabilities in the open-source domain.

The mid-tier category features models like Typhoon-v1.5x-70b-instruct and Qwen2.5-72B-Instruct, offering strong performance while maintaining free accessibility. These models demonstrate particular strength in Thai language processing, with Typhoon being specifically designed for Thai language tasks. Grok-beta and Gemma2-9b-cpt-sea-lionv3-instruct provide additional architectural diversity, with the latter specializing in Southeast Asian languages.

For the low-tier category, ggt-sheet, nllb-200-1.3B, and Openthaigpt1.5-72b-instruct were selected. While having more limited capabilities, these models offer important insights into minimum viable performance levels for Thai translation tasks and serve as benchmarks for measuring progress.

3.3 COMETH Platform

The COMETH platform implements a modern web architecture prioritizing user experience while maintaining robust evaluation capabilities. The design philosophy centers on creating an intuitive interface for translation evaluation while ensuring high-quality data collection. The platform employs a progressive enhancement approach, delivering core functionality to all users while providing enhanced features where supported.

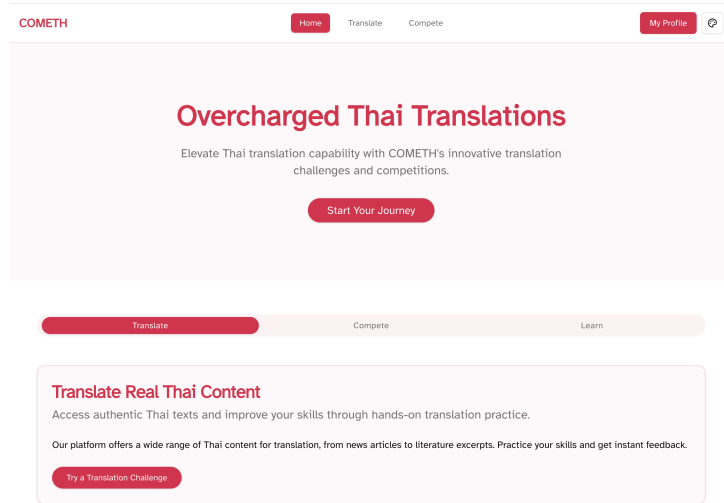


Figure 8: Homepage of COMETH website

3.3.1 Translation Evaluation Workflow

The platform implements a four-phase evaluation process that guides users through translation, evaluation, ranking, and results analysis. Each phase has been carefully designed to maintain consistent quality assessment while ensuring user engagement throughout the process.

Translation Phase The translation phase presents users with an English source text requiring Thai translation. The interface prominently displays the source text alongside a dedicated translation area. Users receive guidance through contextual translation tips and character count indicators. The system automatically captures timing data from initial display through submission. A skip functionality allows users to bypass challenging content, maintaining translation quality by enabling focus on confidently manageable texts.

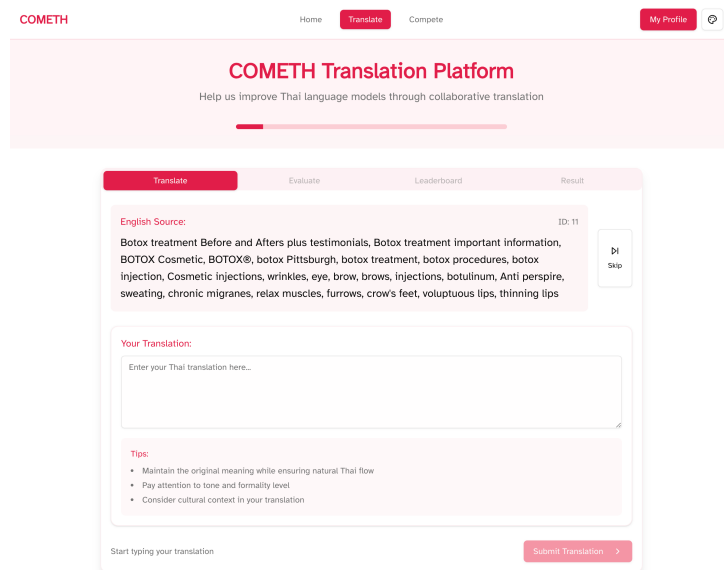


Figure 9: Translation Phrase (1) in Cometh Platform Translation

Evaluation Phase During the evaluation phase, users employ the Multidimensional Quality Metrics (MQM) framework to assess multiple candidate translations. The interface facilitates error categorization across minor, major, and critical classifications, with each category clearly defined through examples. The system provides real-time feedback on error marking and maintains running totals, supporting consistent evaluation standards throughout the assessment process.

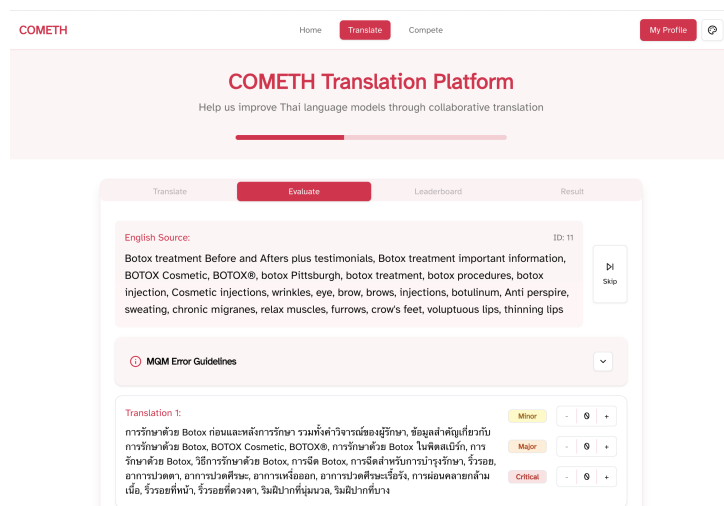


Figure 10: Evaluation Phrase (2) in Cometh Platform Translation

Ranking Phase The ranking phase implements a sophisticated drag-and-drop interface for organizing translations by quality. Users can establish quality tiers, allowing for equal-ranking

groups or strict ordinal arrangements. The interface displays previously marked errors alongside translations to support informed quality judgments. The system captures not only final rankings but also tracks decision-making patterns through interaction timing and movement data.

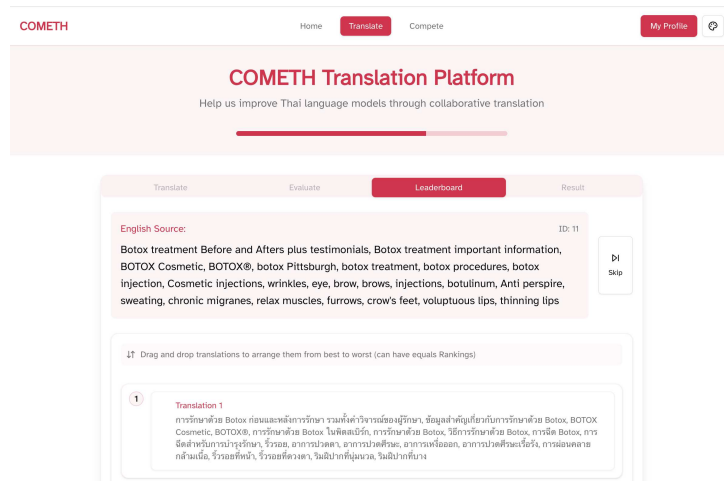


Figure 11: Ranking Phrase (3) in Cometh Platform Translation

Results Phase In the results phase, the system presents a comprehensive analysis of the evaluation session. This includes detailed error counts across categories, correlation analysis between error markings and rankings, and temporal metrics for each evaluation phase. The interface also facilitates comparison with other evaluators' assessments when available, providing valuable context for quality bench-marking.

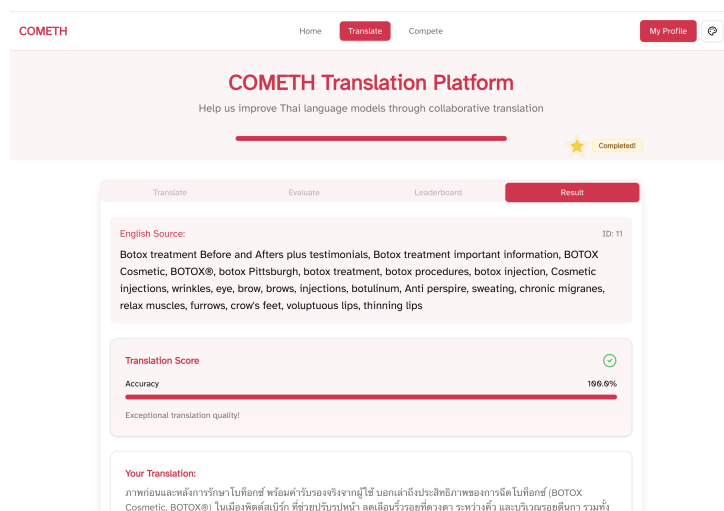


Figure 12: Result Phrase (4) in Cometh Platform Translation

3.3.2 Administrative Workflow

The administrative interface implements privileged functionality for dataset management, analytical processing. This interface restricts access to authorized personnel through user-based authentication protocols, ensuring data integrity and operational security.

Dataset management functionality facilitates corpus augmentation through CSV importation with automated validation protocols. The system parses structured data files containing source text and multiple translation candidates from disparate machine translation systems, automatically validating conformity to required schema specifications before integration into the production database. This methodology enables efficient corpus expansion while maintaining data consistency.

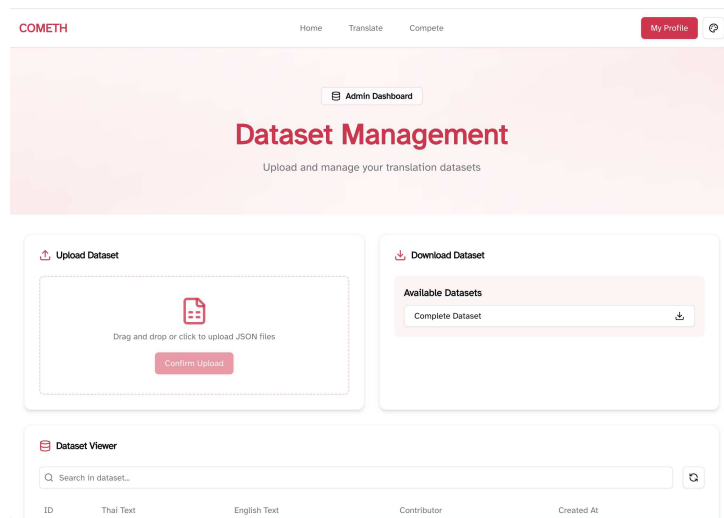


Figure 13: Administrative Page in Cometh Platform Translation

3.3.3 Community Leaderboard Implementation

The platform implements a community leaderboard system for performance visualization and engagement enhancement. The leaderboard establishes a tiered achievement framework categorizing contributors as Champion, Elite, or Expert Translator based on completion thresholds and quality metrics. Statistical indicators display translation counts alongside normalized performance percentages for comprehensive comparison.

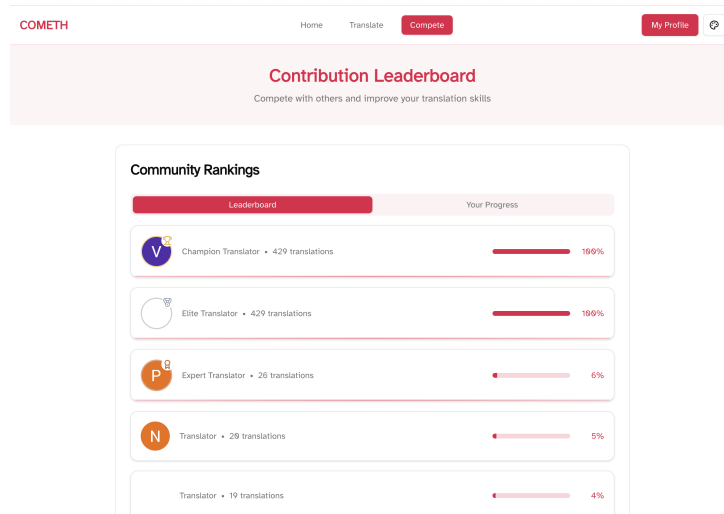


Figure 14: Community Leader board showing Individual Contributions in Cometh Platform Translation

The interface employs a dual-view paradigm presenting both community rankings and personalized progress visualization. The community view arranges contributors in descending performance order, while the individual view contextualizes personal achievements within the broader participant ecosystem. This approach supports both competitive and self-referential improvement motivations.

3.4 Technical Implementation

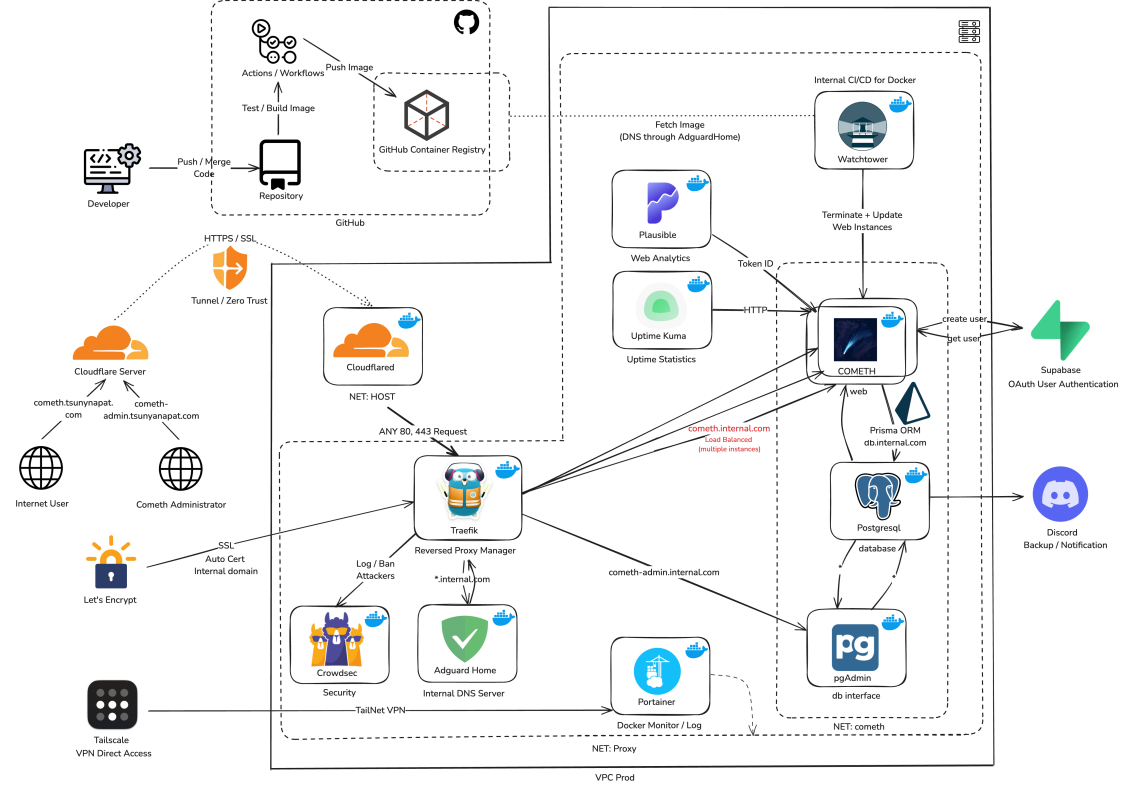


Figure 15: Architecture of Cometh Website

3.4.1 Frontend Architecture

The implementation architecture employs Next.js with React to facilitate server-side rendering and static generation capabilities. TypeScript provides compile-time type verification while functional programming paradigms establish state management protocols. The architecture implements separation between server components for data acquisition and client components for interface interactions, thereby optimizing rendering efficiency.

State management utilizes React hooks implementation with Immer for immutable state transformation, ensuring deterministic state transitions. The architecture instantiates a tab-based navigation system that encapsulates workflow phases in discrete component modules, enhancing maintainability through structural isolation. The user interface implementation utilizes Shadcn components built upon Radix UI primitives, ensuring accessibility compliance and interface consistency. The ranking interface incorporates dnd-kit for interaction management with keyboard navigation alternatives, while Tailwind CSS with class variance authority patterns facilitates responsive design implementation across device categories.

3.4.2 Database Infrastructure

The database architecture implements PostgreSQL with Prisma as the object-relational mapping layer. The schema design establishes four principal entities (Sentence, Candidate, UserProfile, authentication tables) with normalization principles applied to minimize redundancy while strategic denormalization addresses performance requirements.

The Sentence model constitutes the primary workflow foundation, incorporating source text fields, translation data, and phase-specific temporal markers for behavioral analysis. Candidate models maintain machine translation data with comprehensive evaluation metrics and comparative ranking indicators.

Authentication services utilize Supabase for OAuth protocol implementation, establishing appropriate authorization boundaries for user-specific translation assignments. Database schema evolution employs Prisma's migration system with comprehensive versioning to ensure environmental consistency across development contexts.

3.4.3 Continuous Integration and Deployment

The continuous integration and deployment methodology employs GitHub Actions for automated verification and deployment procedures. The workflow architecture establishes discrete validation phases for formatting standards, linting protocols, type verification, and build integrity, with parallel execution paths maximizing efficiency.

Deployment automation initiates upon repository modifications, constructing and publishing containerized images to GitHub Container Registry with semantic versioning implementation. The container build process employs a multi-stage methodology optimizing security and resource utilization: dependency acquisition, compilation, and runtime artifact generation with minimal footprint.

The production infrastructure implements continuous deployment through an automated container orchestration mechanism. WatchTower service monitors the GitHub Container Registry at five-minute intervals, initiating container refreshment upon detecting image updates via SIGTERM signaling. This approach facilitates zero-downtime deployments through Traefik's load balancing capabilities, distributing incoming requests across container instances while seamlessly transitioning traffic to updated containers. The architecture supports horizontal scaling with consistent configuration management through environment variables with high availability.

3.4.4 Deployment Architecture

The deployment architecture implements Docker containerization for environmental consistency. The system architecture comprises three principal services (frontend application, PostgreSQL database, administration interface) orchestrated through Docker Compose, ensuring environmental parity and simplified developer integration processes.

Production deployment utilizes Traefik as a reverse proxy implementing TLS termination, certificate management, and request routing capabilities. The configuration facilitates zero-downtime deployment through container-based service discovery while implementing network isolation protocols that restrict port exposure. External connectivity and analytics implementation utilize CloudFlare Tunnel technology for secure communications channels.

3.5 Finetuning with OCR expense receipt dataset

3.5.1 Training

We use Qwen2.5-VL-7B-Instruct [2] as the base model and perform finetuning using LoRA [5] with a rank of 8. The training data consists of a combination of our custom synthetic receipt dataset and TyphoonVision dataset [12]. Our experiments compare the performance of models finetuned on non-augmented receipts versus those trained on augmented receipts, in order to evaluate the impact of data augmentation on model performance.

3.5.2 Evaluation

Model evaluation is conducted using a real-world expense receipt dataset. This approach, commonly used in OCR research, involves comparing model outputs against ground truth labels. We report three main metrics: word error rate (WER), character error rate (CER), and key mismatch rate.

WER and CER quantify discrepancies between the predicted text and the ground truth.

Key mismatch measures the number of mislabeled or missing structured fields (e.g., total amount, date, item names).

Figure 16 illustrates how WER, CER, and key mismatch are calculated.

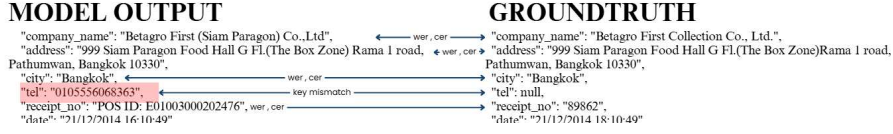


Figure 16: Calculation of word error rate (WER), character error rate (CER), and key mismatch in OCR evaluation.

3.5.3 Results and finding

Our results, as shown in Table 3, demonstrate that using augmented receipts in combination with the TyphoonVision dataset [12] improves the model’s ability to preserve receipt formatting and structural consistency. This performance surpasses both the baseline model and the model trained on non-augmented data. However, the augmentation strategy does not lead to a significant improvement in the model’s OCR text recognition accuracy.

From our experiments and evaluations, we observed that key mismatches most frequently occur in the product section, particularly when the model fails to correctly identify the quantity and price associated with each item.

MODEL	WER	CER	KEY MISMATCH
Qwen2.5-VL-7B-Instruct	0.233	0.183	15.524
Qwen2.5-VL-7B-Instruct-ocr4k-typ40k	0.219	0.154	6.238
Qwen2.5-VL-7B-Instruct-typ44k	0.245	0.181	16.095

Table 3: Qwen2.5-VL-7B-Instruct performance when adding our dataset

3.6 COMETH evaluation

System-Level Correlation with Human Annotations. To assess how well our proposed LLM-based MQM annotators approximate human judgments, we compute the Spearman rank correlation between various systems and human annotations, as presented in Table 4. We evaluate three general-purpose large language models—Gemini 2.0 Flash, GPT-4o Mini, and Claude 3.5 Sonnet—alongside several COMET-based baselines. These include: the baseline COMET model, a COMETH model fine-tuned on human MQM annotations, and an augmented COMETH model trained on both human annotations and those generated by Claude 3.5 Sonnet. Among the LLMs, Claude 3.5 Sonnet achieves the highest correlation with human annotations, indicating that state-of-the-art foundation models can provide high-quality MQM-style error annotations. Additionally, the COMETH models augmented with LLM-generated data show stronger alignment with human judgments, suggesting that synthetic annotations from LLMs can enhance metric training.

MQM Scoring of MT Systems. To further evaluate the effectiveness of the MQM annotation pipeline, we compare the performance of ten machine translation systems based on their average MQM scores derived from human annotations. Table 5 presents the average number of MQM errors per sentence for each system, where higher MQM scores correspond to higher translation quality. This evaluation serves as a gold-standard reference for assessing the accuracy and discriminative power of automatic metrics. The results reveal a clear ranking among the MT systems, reinforcing the value of human MQM annotation in distinguishing fine-grained translation errors and establishing reliable benchmarks for system-level comparison.

MODEL	Spearman’s correlation
gemini-2.0-flash	0.3918
claude-sonnet-3.5	0.4383
gpt-4o-mini	0.4352
COMET	0.4570
COMETH (700 sentences)	0.4639
COMETH Augmented (700 sentences + 1500 LLM annotated sentences)	0.4795

Table 4: Spearman’s correlation between human annotation and translation evaluation model on test set

MT systems	average MQM score
claude-sonnet-3.5	0.8696
grok-beta	0.8404
gpt-4o-mini	0.8121
ggt-sheet	0.7933
gemma2-9b-cpt-sea-lionv3-instruct	0.7855
typhoon-v1.5x-70b-instruct	0.7693
LLaMa3-8b-WangchanX-sft-Full	0.7069
Qwen2.5-72B-Instruct	0.6896
openthaigpt1.5-72b-instruct	0.6859
nllb-200-1.3B	0.5316

Table 5: Average MQM score of each MT systems on test set

4 Discussion

This project aims to develop a multimodal vision-text model for the Thai language, focusing on improving the translation and processing of image and text data. The use of OCR will help interpret financial data, such as bank slips and receipts, and synthetic data will be added to enhance the accuracy of the model. The next steps will follow the three planned phases: data preparation, model development, and evaluation. Additionally, results will be reported, and consultations with the supervisor will be conducted in case of any issues or for further guidance.

5 Social impact

The development of the Vision-Text Multimodal Model has significant social impacts in several areas, as follows:

1. **Understanding financial data, bank receipts, and receipts:** Developing a model that can read and interpret information from financial documents, such as bank receipts and general receipts, will enable users to access important data quickly and accurately, especially in the context of complex financial transactions.
2. **Innovation in industry:** Applying the developed model in sectors such as banking and technology companies will help improve operational efficiency, reduce errors in data processing, and foster new innovations that meet consumer needs.
3. **Impact on the development of Thai language technology:** Developing tools that support the Thai language will make it easier for Thai-speaking users to access new technologies and help bridge the technological gap between the Thai language and other languages with more advanced development.

This project will not only impact the technology sector but also have broader effects on society, particularly in areas related to the daily lives of the public.

6 Conclusion

This report introduces COMETH, a novel machine translation evaluation model and platform tailored for the complexities of English-Thai translation. By employing a rigorous three-phase process encompassing human translation, MQM-based error assessment, and list-wise ranking, COMETH addresses the limitations of general-purpose evaluation metrics for Thai. Furthermore, the creation of a specialized dataset of synthetic Thai e-receipts and expense receipts tackles a significant gap in resources for developing Thai Vision Language Models.

In summary, this work represents a crucial step towards enhancing the capabilities of Thai VLMs. The COMETH platform offers a valuable tool for accurately evaluating and improving English-Thai translation quality, while the newly generated receipt datasets provide essential resources for training VLMs to understand and process real-world Thai visual and textual data. These contributions collectively pave the way for more sophisticated and localized Thai AI applications in the future.

References

- [1] Stanislaw Antol et al. “VQA: Visual Question Answering”. In: *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015, pp. 2425–2433. DOI: 10.1109/ICCV.2015.279.
- [2] Shuai Bai et al. *Qwen2.5-VL Technical Report*. 2025. arXiv: 2502.13923 [cs.CV]. URL: <https://arxiv.org/abs/2502.13923>.
- [3] Alexander Groleau et al. *Augraphy: A Data Augmentation Library for Document Images*. 2023. arXiv: 2208.14558 [cs.CV]. URL: <https://arxiv.org/abs/2208.14558>.
- [4] Muyang He et al. *Efficient Multimodal Learning from Data-centric Perspective*. 2024. arXiv: 2402.11530 [cs.CV]. URL: <https://arxiv.org/abs/2402.11530>.
- [5] Edward J. Hu et al. *LoRA: Low-Rank Adaptation of Large Language Models*. 2021. arXiv: 2106.09685 [cs.CL]. URL: <https://arxiv.org/abs/2106.09685>.
- [6] Drew A. Hudson and Christopher D. Manning. *GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering*. 2019. arXiv: 1902.09506 [cs.CL]. URL: <https://arxiv.org/abs/1902.09506>.
- [7] Tom Kocmi and Christian Federmann. *GEMBA-MQM: Detecting Translation Quality Error Spans with GPT-4*. 2023. arXiv: 2310.13988 [cs.CL]. URL: <https://arxiv.org/abs/2310.13988>.
- [8] Tom Kocmi et al. “Findings of the 2023 Conference on Machine Translation (WMT23): LLMs Are Here but Not Quite There Yet”. In: *Proceedings of the Eighth Conference on Machine Translation*. Ed. by Philipp Koehn et al. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 1–42. DOI: 10.18653/v1/2023.wmt-1.1. URL: <https://aclanthology.org/2023.wmt-1.1/>.
- [9] Daniel Lemire et al. “Roaring bitmaps: Implementation of an optimized software library”. In: *Software: Practice and Experience* 48.4 (Jan. 2018), pp. 867–895. ISSN: 1097-024X. DOI: 10.1002/spe.2560. URL: <http://dx.doi.org/10.1002/spe.2560>.
- [10] Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. “Multidimensional quality metrics: a flexible system for assessing translation quality”. In: *Proceedings of Translating and the Computer 35*. London, UK: Aslib, Nov. 2013. URL: <https://aclanthology.org/2013.tc-1.6/>.
- [11] Kishore Papineni et al. “Bleu: a Method for Automatic Evaluation of Machine Translation”. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Ed. by Pierre Isabelle, Eugene Charniak, and Dekang Lin. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, July 2002, pp. 311–318. DOI: 10.3115/1073083.1073135. URL: <https://aclanthology.org/P02-1040/>.
- [12] Parinthat Patil et al. *Llama-3 Typhoon Vision Preview*. <https://huggingface.co/scb10x/llama-3-typhoon-v1.5-8b-vision-preview>. Research preview version, supports text and image input modalities. 2024.
- [13] Thirawarit Pitiphat and NECTEC Team. *nectec/pathumma-llm-vision-1.0.0*. <https://huggingface.co/nectec/Pathumma-llm-vision-1.0.0>. 2024.
- [14] Ricardo Rei et al. “COMET: A Neural Framework for MT Evaluation”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber et al. Online: Association for Computational Linguistics, Nov. 2020, pp. 2685–2702. DOI: 10.18653/v1/2020.emnlp-main.213. URL: <https://aclanthology.org/2020.emnlp-main.213/>.

- [15] Ricardo Rei et al. “CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task”. In: *Proceedings of the Seventh Conference on Machine Translation (WMT)*. Ed. by Philipp Koehn et al. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, Dec. 2022, pp. 634–645. URL: <https://aclanthology.org/2022.wmt-1.60/>.
- [16] Robin Rombach et al. *High-Resolution Image Synthesis with Latent Diffusion Models*. 2022. arXiv: 2112.10752 [cs.CV]. URL: <https://arxiv.org/abs/2112.10752>.
- [17] Thibault Sellam, Dipanjan Das, and Ankur Parikh. “BLEURT: Learning Robust Metrics for Text Generation”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, July 2020, pp. 7881–7892. DOI: 10.18653/v1/2020.acl-main.704. URL: <https://aclanthology.org/2020.acl-main.704/>.
- [18] Fen Xia et al. “Listwise Approach to Learning to Rank: Theory and Algorithm”. In: *Proceedings of the 25th International Conference on Machine Learning (ICML)*. Jan. 2008, pp. 1192–1199. DOI: 10.1145/1390156.1390306.
- [19] Can Xu et al. “WizardLM: Empowering Large Language Models to Follow Complex Instructions”. In: *arXiv preprint arXiv:2304.12244* (Apr. 2023). DOI: 10.48550/arXiv.2304.12244. URL: <https://arxiv.org/abs/2304.12244>.
- [20] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. *Adding Conditional Control to Text-to-Image Diffusion Models*. 2023. arXiv: 2302.05543 [cs.CV]. URL: <https://arxiv.org/abs/2302.05543>.