

# ThaiPaliBench: Benchmarking Large Language Models on Thai-Script Pali Understanding

Nobpharat Phetsri<sup>1,2</sup>, Anon Saeoueng<sup>2</sup>, Sarawoot Kongyoung<sup>2</sup>, Waranrach Viriyavit<sup>1</sup>

<sup>1</sup>Faculty of Informatics, Burapha University, Chonburi, Thailand

<sup>2</sup>National Science and Technology Development Agency (NSTDA), Pathum Thani, Thailand

Correspondence: sarawoot.kon@nectec.or.th

## Abstract

The application of large language models (LLMs) to Pali written in the Thai script remains limited in accuracy due to its low-resource setting, complex grammar, and the absence of standardized evaluation benchmarks. In this paper, we introduce **ThaiPaliBench**, a comprehensive benchmark dataset designed to systematically evaluate LLM proficiency in Thai-script Pali. Our proposed dataset is an LLM-assisted benchmark derived from the Royal Pali Examination corpus and spans Parian Tham levels 1-9. ThaiPaliBench, comprises multiple-choice and sentence-pair tasks that evaluate grammatical knowledge, reading comprehension, natural language inference, and translation understanding. To verify the effectiveness of our benchmark, we evaluate over 10 model variants across 3 commercial and open-source LLM families, including ChatGPT, Claude, Kimi, Grok, Typhoon, GPT-OSS-120B, ThaiLLM, and SEA-LION. Our experiments show that even the best-performing model, Kimi K2.5 (1T), achieves only 68.3% overall accuracy, with performance dropping substantially on grammatical reasoning and semantic inference tasks. These results highlight the unique challenges posed by Thai-script Pali and confirm that ThaiPaliBench serves as a reliable, linguistically grounded benchmark for evaluating LLM proficiency in this low-resource language. Our proposed dataset is publicly available at: <https://github.com/ainrg-llm-dev/ThaiPaliBench>

## 1 Introduction

Large language models (LLMs) have demonstrated remarkable multilingual capabilities, enabling cross-lingual understanding and generation across a wide range of languages (Li et al., 2025; Ni et al., 2025). However, low-resource languages — particularly classical languages such as Pali — remain a significant challenge. The scarcity of avail-

able training data in these languages means that current LLMs lack sufficient linguistic knowledge to accurately process and interpret them (Jadhav et al., 2025). This limitation is especially consequential for Pali, a classical Indic language of profound religious and cultural significance (Gurusinghe and Jayatilleke, 2026).

Although Pali serves as the canonical scriptural language of Theravada Buddhism and remains an essential subject of study for Buddhist monks throughout Southeast Asia, including Thailand (Songkhaeo, 2021), comprehensive evaluation frameworks for assessing LLM capabilities in this language remain largely absent, leaving a considerable gap in multilingual benchmarking research (Arçon et al., 2026).

In particular, Thai monastic and academic institutions formally teach Pali — a Middle Indo-Aryan language closely related to Sanskrit — using the Thai script. The national curriculum, known as Parian Tham (Appendix A), comprises nine graded levels (1–9) that progressively assess grammatical analysis, translation proficiency, doctrinal interpretation, and textual comprehension (Pansa, 2020). However, existing resources focus primarily on modern Thai or Romanized Pali (Wanjan, 2021), and no comprehensive, exam-based benchmark currently exists to evaluate LLM proficiency on Thai-script Pali. As a result, the extent to which current LLMs can understand Thai-script Pali remains largely unexplored.

To address the lack of evaluation resources, we introduce **ThaiPaliBench**, the first standardized benchmark for evaluating LLM proficiency on Thai-script Pali. We build the dataset directly from the Royal Pali Examination corpus, following established benchmark design principles (Miranda et al., 2025; Shin et al., 2025), converting real exam materials into structured multiple-choice and sentence-pair formats. The benchmark spans all nine Parian Tham levels and contains 4,000 test

questions, evenly divided into four tasks: grammatical knowledge, reading comprehension, natural language inference, and translation understanding. Questions are sampled across proficiency levels, with grammatical analysis assessed exclusively at the foundational levels (1–3), and model performance is evaluated using accuracy metrics throughout. Question stems and answer keys are drawn directly from these official examinations, while distractor options are LLM-generated under constraints anchored to the official answer; we therefore describe ThaiPaliBench as an LLM-assisted benchmark derived from official examination corpora.

To validate ThaiPaliBench, we evaluate 10 LLM variants spanning proprietary commercial and open-weight architectures. Our commercial selection includes GPT-5 mini (OpenAI, 2025), Claude Haiku 4.5 (Anthropic, 2025), and Grok 4.1 (xAI, 2025) Fast as representative frontier systems.<sup>1</sup> For open-weight models, we evaluate massive frontier systems (Kimi (Kimi Team, 2026) and GLM (GLM Team et al., 2026)), regionally specialized models with continual pre-training (Typhoon (Pipatanakul et al., 2024), ThaiLLM (Lertpiya et al., 2026), and SEA-LION (Ong and Limkonchotiwat, 2023)), and GPT-OSS-120B (OpenAI et al., 2025) as a general baseline. Our experiments reveal that even the top-performing model achieves only 68.3% overall accuracy, confirming that Thai-script Pali remains a substantially underserved language for current LLMs.

The main contributions of this work are as follows:

- We introduce **ThaiPaliBench**, the first benchmark for Thai-script Pali. We build the dataset from the Royal Pali Examination corpus across all 9 proficiency levels.
- We use 4 standard evaluation tasks: grammatical knowledge, reading comprehension, natural language inference, and translation understanding. We structure the exam materials into multiple-choice and sentence pairs for objective assessment.

## 2 Related Work

Prior work relevant to ThaiPaliBench spans three areas: general LLM evaluation frameworks that de-

<sup>1</sup>Gemini is excluded from evaluation as it was utilized during dataset construction.

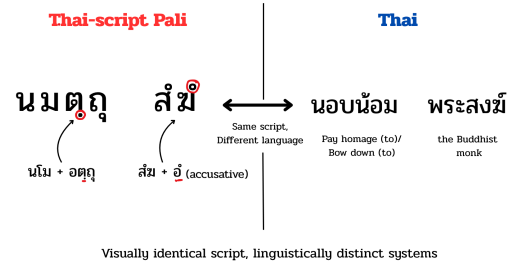


Figure 1: Thai-script Pali and Modern Thai share identical orthography but represent linguistically distinct systems.

fine our benchmarking methodology (Section 2.1), Thai-specific benchmarks that reveal the gap our work addresses (Section 2.3), and computational approaches to classical languages that contextualize the linguistic challenges we evaluate (Section 2.2).

### 2.1 LLM Benchmarks and Cultural Evaluation

The evaluation of Large Language Models (LLMs) recently shifted from general language tests to cultural evaluations. Frameworks like MMLU (Hendrycks et al., 2021) established broad knowledge benchmarks but heavily favor Western concepts. In response, regional benchmarks such as SeaEval (Wang et al., 2024) and SEA-HELM (Sung et al., 2025) introduced cultural reasoning to show the limits of multilingual models in Southeast Asian contexts. The regional frameworks effectively measure model performance on standard reasoning tasks and cultural alignment.

These broad evaluation frameworks measure only modern language fluency and rely entirely on standard sentence structures. The generalized benchmarks fail to capture the specific script-switching features that local writing systems like Thai-script Pali.

### 2.2 LLM Evaluation for Classical Languages

Processing classical languages presents significant computational challenges, particularly due to rich morphology and limited annotated resources. In Pali, a single nominal stem can appear in many inflected surface forms depending on case, number, and gender. Subword tokenization methods such as Byte-Pair Encoding (BPE) often split these forms into multiple fragments that do not correspond to meaningful morphemes, making it harder for models to consistently associate related inflec-

tions with the same lexical root (Lerner and Yvon, 2025). Similar observations have been reported for Sanskrit language models such as San-BERT (Bhatnagar et al., 2023). Pali also exhibits productive phonetic compounding through Sandhi, where adjacent words are merged and phonologically altered (e.g., *Namatthu* from *Namo + Atthu*). Prior computational work on Pali sandhi processing has primarily focused on rule-based splitting and joining systems (Basapur et al., 2019), highlighting the linguistic complexity of recovering original word boundaries in compounded forms. Current computational methods rely heavily on Romanized corpora following Western conventions (von Hinüber, 1996) and focus strictly on developing traditional NLP pipelines. The capability of modern LLMs to understand classical Pali directly within the Thai script remains unmeasured. While our work diverges from traditional tool development by focusing exclusively on benchmark construction.

### 2.3 Thai Benchmarks

Thai language evaluation focuses mostly on modern conversational and domain-specific fluency. Benchmarks such as WangchanThaiInstruct (Limkonchotiawat et al., 2025) and the Thai LLM Leaderboard (SCB 10X et al., 2024) standardize fluency tests for modern secular Thai. These frameworks primarily evaluate mathematical reasoning, reading comprehension, and everyday conversational capabilities for modern Thai LLMs.

However, existing Thai benchmarks do not evaluate the handling of Thai-script Pali morphology and compound formation. When processing Thai-script Pali, modern Thai models frequently map classical tokens into modern semantic spaces, leading to misinterpretation of inflected forms and sandhi compounds (Appendix(B, C)). Our work extends beyond modern Thai fluency assessment by explicitly evaluating syntactic and morphological comprehension in Thai-script Pali contexts.

## 3 Methodology

This section describes how we construct ThaiPaliBench from Royal Pali Examination archives and formulate its four evaluation tasks.

### 3.1 Data Collection & Construction

We based ThaiPaliBench on official Royal Pali Examination archives from 1986 to 2025 (Parian Tham see details in Appendix A). Supervised by

the Sangha Supreme Council, this national education system provides a verified 9-level difficulty.

Processing Thai-script Pali also introduces linguistic challenges that standard models often struggle to resolve. Models must handle segmentation ambiguities in text written without spaces (*scriptio continua*), disambiguate "false friends" where Pali loanwords shift meaning in modern Thai, and manage frequent code-switching between classical verses and Thai commentary.

To transform these scanned examinations into a clean evaluation dataset, we designed a pipeline with 3 core processing modules, as shown in Figure 2.

**Input Data (Raw PDF Exam).** Before running the pipeline, we collected original examination papers in scanned PDF format from official archives. As shown in the *Input Data* component of Figure 2, these documents contain Thai-script Pali text, examination questions, and level-specific sections. They serve as the raw source for the benchmark.

**Module 1: OCR & Digitization.** From figure 2, in the first module, we convert scanned PDF images into machine-readable text. We initially tested conventional OCR systems, including olmOCR-2-7B-1025<sup>2</sup> and Typhoon OCR 1.5 2B<sup>3</sup>, but these systems produced frequent transcription errors and hallucinated content, especially on complex Thai-script Pali characters and diacritical marks. To address these limitations, we adopted Gemini3-Pro-Preview<sup>4</sup> as a Vision-Language Model (VLM) for digitization (See details in Appendix D).

**Module 2: LLM-Based Task Synthesis** We then pass the digitized text to the second module for task synthesis (the *LLM-Based Task Synthesis* component in Figure 2). We use Gemini 3 flash to structure extracted content into task formats. Carefully designed prompts guide the model to generate 4 task types: Grammar, Reading Comprehension (RC), Translation, and Natural Language Inference (NLI). We structure NLI as sentence pairs and the remaining tasks as multiple-choice questions with explicit answer options. This synthesis stage systematically maps source content to task type and proficiency level. To make the division between source and model explicit, Appendix F details which parts of each task are drawn from the official examinations and which are LLM-generated.

<sup>2</sup><https://olmocr.allenai.org/>

<sup>3</sup><https://opentyphoon.ai/blog/th/typhoon-ocr-release>

<sup>4</sup><https://deepmind.google/technologies/gemini/>

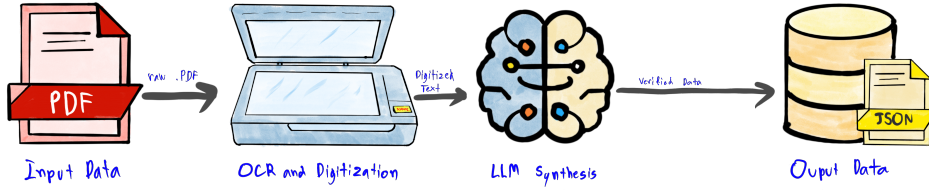


Figure 2: Data construction pipeline: The workflow transforms raw Parian Tham examination papers into structured ThaiPaliBench.

**Module 3: Output Data (ThaiPaliBench JSONL)** In the final module, we standardize and merge generated instances into JSONL files, matching the *Output Data* component in Figure 2. Each record includes structured metadata: source year, proficiency level, task type, question text, answer options, and the correct label.<sup>5</sup>

| Level        | Grammar      | RC           | Translation  | NLI          | Total        |
|--------------|--------------|--------------|--------------|--------------|--------------|
| L1–2         | 500          | 125          | 125          | 125          | 875          |
| L3           | 500          | 125          | 125          | 125          | 875          |
| L4           | –            | 125          | 125          | 125          | 375          |
| L5           | –            | 125          | 125          | 125          | 375          |
| L6           | –            | 125          | 125          | 125          | 375          |
| L7           | –            | 125          | 125          | 125          | 375          |
| L8           | –            | 125          | 125          | 125          | 375          |
| L9           | –            | 125          | 125          | 125          | 375          |
| <b>Total</b> | <b>1,000</b> | <b>1,000</b> | <b>1,000</b> | <b>1,000</b> | <b>4,000</b> |

Table 1: Distribution of questions across task types and proficiency levels in ThaiPaliBench. Grammar questions focus on foundational levels (Levels 1–3), while comprehension, translation, and reasoning tasks span all 9 levels.

The resulting dataset, as shown in Table 1, comprises **4,000** questions evenly distributed across **4** task types, with **1,000** instances per task. Grammar questions concentrate in foundational levels (Levels 1–3), reflecting emphasis on morphological and syntactic capability in earlier stages. In contrast, Reading Comprehension, Translation, and Natural Language Inference span all **9** proficiency levels, with **125** items per level for each task. This design ensures balanced coverage across the difficulty spectrum and enables systematic evaluation of linguistic knowledge and higher-order reasoning.

### 3.2 Pipeline Validation

We validate the two stages of our construction pipeline shown in Figure 2.

**Module 1: OCR & Digitization.** We compared three candidate models on a held-out set of ground-truth Thai-script Pali passages. Gemini 3 Pro Preview achieved the best transcription accuracy and was selected to digitize the exam corpus (Appendix D).

**Module 2: LLM Task Synthesis.** We evaluated candidate generation models by prompting roughly ten real exam questions through web chat interfaces. We chose Gemini Flash because its distractor options and explanations best matched official grading standards while remaining cost-effective. After synthesis, authors spot-checked 10–20 items per task against the original scanned exams, finding 100% answer-key accuracy on all checked items.

We further evaluated the final dataset using two validation checks:

- **Programmatic TRUTH ANCHOR Check:** We ran a verification script across all 3,000 multiple-choice items (Grammar, Reading Comprehension, and Translation). The script confirms that every correct option exactly matches the official answer key and that all five choices are non-empty and mutually distinct. At a strict similarity threshold ( $\geq 0.95$ ), the script flagged 2 exact and 21 near-duplicates in Grammar, 1 near-duplicate in Reading Comprehension, and 3 exact and 4 near-duplicates in Translation. After manual review, we found that several flagged cases are simple morphological variants (for example, *rajjam* vs. *raññam*) or standard grammar question formats, leaving 13 cases in Grammar, 1 in Reading Comprehension, and 4 in Translation as relevant duplicates. For NLI, we found 226 items where Pali premises overlap with Translation or Reading Comprehension source sentences because these tasks share the same canonical texts. We leave the frozen

<sup>5</sup>ThaiPaliBench is under CC BY-NC 4.0.

dataset unchanged and report these residual artifacts; their estimated effect on overall benchmark scores is under 0.15 percentage points per task.

- **Manual Validation Against Source Papers:** We manually validated a stratified random sample of 389 items (100 Grammar, 100 Reading Comprehension, 100 Translation, 89 NLI) against the original scanned examination papers, checking whether each question stem originates from the source exam and whether the marked key matches the official answer. Stem provenance was 99.7% (388/389; 95% Wilson CI [98.6, 100.0]) and key accuracy 95.1% (370/389; CI [92.5, 96.9]). Per-task key accuracy was 95.0% (Grammar), 92.0% (RC), 94.0% (Translation), and 100% (NLI). Errors are not uniformly distributed: in Reading Comprehension, four of eight errors occur on items asking for morphosyntactic analysis rather than passage comprehension, and in Translation, errors concentrate on segmentation misalignment between the Pali source and the Thai target span. We report these rates on the frozen dataset and leave the affected items unchanged. NLI labels were checked by semantic judgment rather than against an official answer key. This validation was performed by the authors; no certified Pali scholar reviewed the benchmark.

### 3.3 Task Formulation & Metrics

We formulate all benchmark tasks as discriminative evaluations rather than open-ended generation. We create tests of Grammar, Translation, and Reading Comprehension as **5-option** multiple choice questions (MCQs), and Natural Language Inference (NLI) that use sentence-pair classification. This structure aligns with evaluation benchmarks such as MMLU, which prioritize objective, verifiable answers over generation quality (see Appendix F for detailed task provenance). The benchmark measures **4** core abilities:

- **Grammatical Knowledge:** We tested the model for understanding Pali grammar, including verb conjugation and noun declension, following a CoLA-style format (Warstadt et al., 2019).
- **Reading Comprehension:** We measure passage understanding by presenting passages

written in Thai-script Pali and asking the correct Thai answer, following SQuAD-style design (Rajpurkar et al., 2018).

- **Natural Language Inference (NLI):** Each item pairs a Pali premise with a Thai hypothesis, labeled entailment, neutral, or contradiction, following the sentence-pair classification format used in XNLI (Conneau et al., 2018). Pali premises are extracted directly from official examinations. Entailment hypotheses are synthesized via LLM as semantically faithful Thai paraphrases anchored strictly to official reference translations, while neutral and contradiction hypotheses introduce unverifiable facts or logical negations, respectively. Because all hypothesis variants are LLM-generated, this task assesses cross-lingual inference closely tied to translation semantics rather than fully independent expert-verified reasoning.
- **Translation Understanding:** We test models can select the most accurate Thai translation of a Thai-script Pali sentence or passage, given several similar-looking options.

## 4 Results

This section reports the systematic evaluation of **10** commercial and open-weights models on ThaiPaliBench.

### 4.1 Overall Model Performance

Table 2 reports the average accuracy across all ThaiPaliBench tasks. Kimi k2.5 achieves the highest accuracy (68.33% [66.90, 69.77]), while GPT-OSS records the lowest (40.65%), reflecting a 27.68% gap among evaluated models.

Large open-weight models (Kimi k2.5 and GLM-5) occupy the top tier, outperforming all commercial models in this evaluation. Item-level paired bootstrap tests show that adjacent models within groups share indistinguishable performance ( $p > 0.05$ ), clustering systems into five distinct performance tiers rather than a strict linear ranking.

### 4.2 Performance Across Parian Tham Levels

Figure 4 displays model accuracy across eight level buckets covering all nine Parian Tham levels, ordered from easiest to most advanced. However, higher levels do not consistently yield lower accuracy.

Instead, models share a common pattern:

| Model                       | Type         | Model Category | Accuracy (%) | Tier |
|-----------------------------|--------------|----------------|--------------|------|
| Kimi k2.5                   | Open-Weights | Standard       | <b>68.33</b> | 1    |
| GLM-5                       | Open-Weights | Standard       | 66.98        | 1    |
| Claude Haiku 4.5            | Commercial   | Standard       | 57.75        | 2    |
| Grok 4.1 Fast               | Commercial   | Standard       | 55.43        | 3    |
| GPT-5 mini                  | Commercial   | Reasoning      | 53.63        | 3    |
| Llama-SEA-LION-v3-70B-IT    | Open-Weights | Standard       | 53.60        | 3    |
| Qwen-SEA-LION-v4-32B-IT     | Open-Weights | Standard       | 52.50        | 3    |
| ThaiLLM-Qwen3-8B-Think v3.0 | Open-Weights | Reasoning      | 52.12        | 3    |
| Typhoon v2.5                | Open-Weights | Standard       | 45.40        | 4    |
| GPT-OSS                     | Open-Weights | Reasoning      | 40.65        | 5    |

Table 2: Overall benchmark performance grouped into tiers. Tier boundaries are placed where adjacent models differ significantly ( $p < 0.05$ , item-level paired bootstrap, 10,000 resamples); see Appendix H.5 for all nine adjacent-pair tests. Because only adjacent pairs are tested, models at opposite ends of the same tier are not necessarily indistinguishable. “Reasoning” denotes explicit chain-of-thought models.

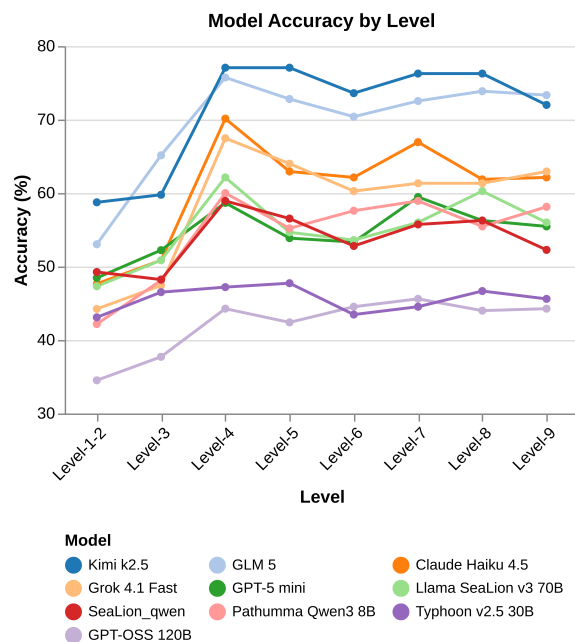


Figure 4: Model accuracy across Parian Tham proficiency levels (L1–L9).

- **Foundational Levels (L1–L3):** All models struggle at Levels 1 and 2. Surprisingly, their scores go up at Level 3. This breaks the assumption that higher levels are always harder.
- **Intermediate Peak (L4):** Scores hit their highest point at Level 4. This big jump suggests that Level 4 introduces easier question types, like reading comprehension, which need less grammar analysis.
- **Advanced Plateau (L5–L9):** Scores level off across the highest tiers. However, the gap between models grows larger. Top models (like Kimi k2.5) keep high scores, while weaker models swing up and down. This widening gap shows that advanced exams separate strong reasoning models from weak ones better than the early exams do.

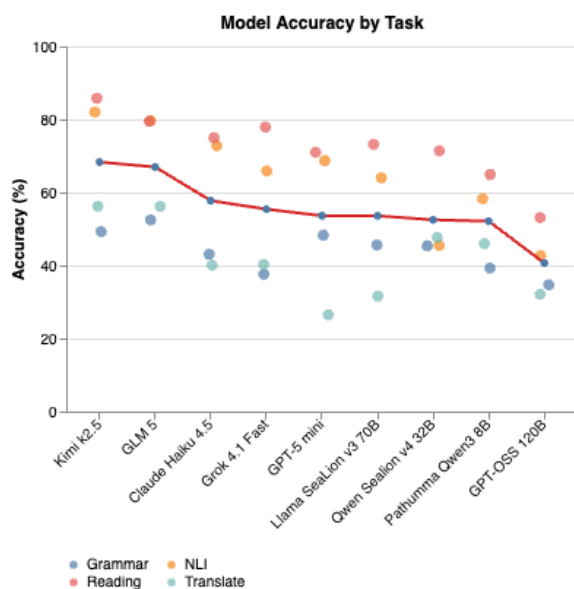


Figure 3: Model accuracy by evaluation category. Reading comprehension achieves higher accuracy across models, while grammar and translation tasks remain comparatively lower.

We suggest 2 main reasons for this unexpected trend in the foundational levels. First, there is a significant imbalance in the dataset size: Levels 1–3 contain 875 questions per level, whereas Levels 4–9 contain only 375 questions per level. This larger sample size in the early levels exposes models to a wider variety of difficult edge cases. Second, the composition of the exam subjects shifts. Levels 4–9 do not include the notoriously difficult grammar questions. Furthermore, within the grammar category itself, questions in Levels 1–2 cover a much

wider diversity of complex rules. In contrast, Level 3 grammar questions are more predictably focused on word morphology, primarily asking models to identify word types, suffixes (*Paccaya dhamma*), Sandhi rules, and basic syntactic functions.

### 4.3 Natural Language Understanding Across Difficulty Gradient

Figure 5 shows category-specific performance trends across Parian Tham levels. This breakdown helps us analyze how each skill scales with exam difficulty.

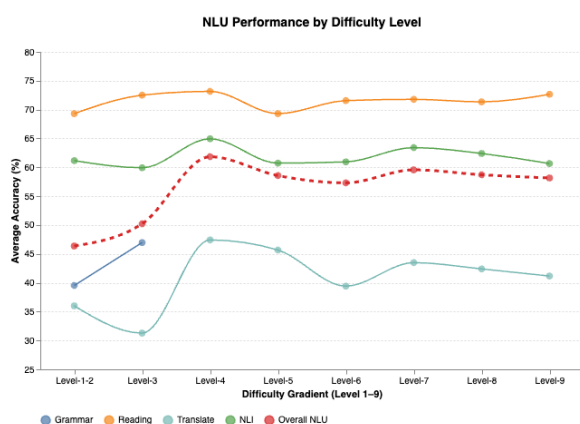


Figure 5: Category-specific performance across proficiency levels (Grammar, RC, NLI, Translation, and overall NLU).

The plotted lines reveal distinct patterns across the evaluation categories:

- **Reading Comprehension Stability and Probing:** Reading comprehension is present across all levels, and its scores remain remarkably stable (between 69.5% and 73.8%). This flat trend contradicts the expected difficulty increase. To understand this behavior, we ran a controlled ablation probe on the full RC subset ( $N = 1,000$ , Levels 1–9) across 9 models<sup>6</sup> under 3 settings: (1) full passage + question, (2) question only, and (3) mismatched passage + question (full results in Table 8).
  - *Passage Reliance:* Removing the passage costs 9.2–18.2 percentage points across all models ( $p < 0.0001$ ; Kimi drops from 86.1% to 67.9%, Claude Haiku from 75.6% to 57.5%, and GPT-OSS from 54.8% to 45.6%). This con-

<sup>6</sup>The probe was conducted across 9 of our 10 evaluated models; Grok 4.1 Fast was excluded as it was deprecated by the provider prior to this ablation run.

firms that models genuinely read the context rather than relying on simple keyword overlap.

- *Scriptural Memorization:* Without any passage, models still score 45.6%–67.9% ( $2.3\text{--}3.4\times$  the 20% chance baseline). Because RC questions come from canonical Buddhist texts available online, models leverage pre-training knowledge. Thus, reported RC scores combine real passage comprehension with memorized scriptural text.
- *Mismatched Passage:* Substituting an unrelated passage (S3) is not equivalent to removing it, and the direction is model-dependent: relative to the question-only setting, 4 of 9 models exhibit significant shifts (Kimi gains 3.0 points,  $p = .030$ ; Claude Haiku gains 2.9 points,  $p = .031$ ; GLM-5 loses 4.9 points,  $p < .001$ ; Llama-SEA-LION loses 2.6 points,  $p = .041$ ), while the remaining 5 models show no significant difference ( $p > .05$ ). These effects are small and we report them as an observation rather than a conclusion. Accuracy under S3 remains 9.0–22.6 points below the full-passage setting across all models ( $p < .0001$ ). On a 300-item pilot these contrasts were non-significant; the full subset was required to detect them.

- **Natural Language Inference Variation:** NLI follows the curriculum difficulty much better than reading comprehension. Scores drop at Level 3 and bounce back strongly at Level 4 (rising from 60.0% to 65.6%). This pattern suggests that intermediate-level logic introduces abstract relationships that briefly confuse the models.
- **Translation Instability:** Translation proves the hardest and most unpredictable task. Models consistently achieve lower accuracy here compared to Reading Comprehension. Rather than allowing for simple keyword matching, Translation requires a genuine cross-lingual process. Although the source text is written entirely in the Thai script, the underlying language is Pali. Models must translate Thai-script Pali into modern Thai, a task that demands deep lexical and morphological knowl-

edge. This deceptive “Thai-to-Thai” visual similarity often misleads models into simple pattern-matching rather than actual translation. Accuracy drops heavily at Level 3 and jumps back up at Level 4 (increasing from 30.9% to 47.0%). These wild shifts suggest that translation difficulty depends on specific passage vocabulary rather than the general curriculum level.

- **Overall NLU Trajectory:** The overall trend combines all skills. Accuracy climbs from the early levels, hits a sharp peak at Level 4 (jumping from 50.7% to 62.2%), and then levels off across advanced tiers (staying around 57% to 60%). The Level 4 peak implies that this tier introduces easier question types. Furthermore, the flat scores show that models hit a performance ceiling; harder exams no longer lower their multiple-choice scores.

This comparison shows that scores do not scale automatically with model size. ThaiLLM-Qwen3-8B reaches 52.12% and outperforms the 120B generalist GPT-OSS (40.65%) by 11.47 points ( $p < 0.001$ ), suggesting that domain-specific pre-training on regional data contributes more than parameter count alone.

## 5 Discussion

The results show a persistent performance gap in how LLMs understand Thai-script Pali. Although massive open-weight models achieve performance parity with proprietary commercial systems (Section 4.1), all models consistently score lower on grammar and translation tasks. This trend aligns with broader findings where multilingual benchmarks show systematic degradation on low-resource languages. ThaiPaliBench highlights that models demonstrate weak performance even at foundational grammatical levels. This weakness confirms that current LLM training pipelines lack sufficient coverage of classical language structures, causing models to fail when strict grammatical accuracy is required. When comparing the evaluated systems (Appendix H.1), we find that open-weight models handle grammar better, while commercial APIs maintain strong performance in comprehension tasks.

Beyond model types, ThaiPaliBench reveals a surprising contradiction regarding exam difficulty. Although Level 1 and 2 are designed to be the easiest, our data indicates that models actually struggle

more at these foundational stages than at Level 3 or 4. The analysis in Section 4.2 shows that accuracy does not follow a straight line from easiest (Level 1) to hardest (Level 9). Instead, we see an unexpected accuracy peak at Level 4 and a flat trend across higher tiers. As discussed in Section 4.3, our probing analysis shows that reported comprehension scores combine genuine passage reading with memorized scriptural text. Meanwhile, the lower scores at Levels 1–2 indicate that models cannot yet handle the abstract grammar rules required at the start of the curriculum.

Consequently, these performance anomalies raise questions about model scale. Our scaling analysis indicates that simply adding more parameters does not guarantee better results. The finding in Appendix H.2 that an 8B Thai-specialized model outperforms a 120B generalist by 11.47 points (52.12% vs. 40.65%,  $p < 0.001$ ) indicates that parameter size alone cannot ensure competence in classical languages. This observation matches broader findings from holistic LLM evaluations (Liang et al., 2023). Overall, ThaiPaliBench surfaces limitations that modern-language tests fail to reveal, providing a clear way to diagnose cross-lingual weaknesses.

## 6 Conclusion

This paper introduced ThaiPaliBench, the first comprehensive benchmark for evaluating Large Language Models on Thai-script Pali. By adapting the Royal Pali Examination, we created a rigorous framework to test grammatical knowledge, reading comprehension, natural language inference, and translation understanding across different difficulty levels. Our evaluation of leading models reveals that classical languages remain a significant challenge for modern AI.

Most notably, our findings reveal a persistent performance gap in current model capabilities. Models perform worse on foundational grammar tests (Levels 1 and 2) than on advanced reading tasks.

This anomaly suggests that models combine active reading with memorized scriptural knowledge rather than relying on abstract grammatical reasoning, as confirmed by our probing analysis.

Furthermore, our scaling analysis demonstrates that parameter count alone does not guarantee competence in Thai-script Pali. ThaiLLM-Qwen3-8B, a Thai-specific model with only 8B parameters, outperforms the 120B generalist GPT-OSS (52.12%

vs. 40.65%), suggesting that domain-specific pre-training on regional language data contributes more effectively than raw scale for classical language understanding.

Ultimately, ThaiPaliBench provides a diagnostic resource for the AI research community, surfacing systematic gaps in classical language understanding and offering a foundation for future low-resource model development.

## Limitations

ThaiPaliBench involves methodological constraints related to data construction and preprocessing (Section 3.1). Distractors (i.e., incorrect answer choices in multiple-choice questions) were systematically generated to approximate Royal Pali Examination grading principles across all MCQ tasks, including Grammar, Reading Comprehension, and Translation; however, automated generation may not fully reflect authentic human error distributions. Source texts were digitized using OCR and vision-language tools, and despite correction and review procedures, minor transcription inaccuracies and limited expert validation may remain. Regarding experimental reproducibility, evaluating proprietary APIs carries replication risks due to unannounced backend updates and non-determinism, while synthesizing items via Gemini may introduce stylistic bias. Although we release frozen dataset splits and exact inference parameters, closed API outputs may exhibit minor temporal drift. Our evaluation is discriminative only; generative competence in Thai-script Pali remains untested, although Levels 8–9 of the examination contain open-ended composition tasks that provide a direct path to such an extension. NLI labels were assigned by construction against the official reference translations and were not verified by Pali scholars.

## Acknowledgments

We thank Kun Kerdthaisong and Chaianun Damrongrat for their guidance and helpful feedback throughout this research.

## References

David Alfter. 2015. [Analyzer and generator for pali](#). *Preprint*, arXiv:1510.01570.

Anthropic. 2025. Introducing Claude Haiku 4.5. <https://www.anthropic.com/news/claude-haiku-4-5>. Model version claude-haiku-4-5-20251001. Accessed 2026-05-17.

Tjaša Arčon, Matej Klemen, Marko Robnik-Šikonja, and Kaja Dobrovoljc. 2026. [Evaluating metalinguistic knowledge in large language models across the world’s languages](#). *Preprint*, arXiv:2602.02182.

Swati Basapur, Shivani V, and Sivaja Nair. 2019. [Pāli sandhi – a computational approach](#). In *Proceedings of the 6th International Sanskrit Computational Linguistics Symposium*, pages 181–192, IIT Kharagpur, India. Association for Computational Linguistics.

Kartik Bhatnagar, Sampath Lonka, Jammi Kunal, and Mahabala Rao M G. 2023. [San-bert: Extractive summarization for sanskrit documents using bert and its variants](#). *Preprint*, arXiv:2304.01894.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. [XNLI: Evaluating cross-lingual sentence representations](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

GLM Team, Aohan Zeng, Xin Lv, Zhenyu Hou, and 1 others. 2026. [GLM-5: From vibe coding to agentic engineering](#). *arXiv preprint arXiv:2602.15763*.

Ranidu Gurusinghe and Nevidu Jayatilleke. 2026. [Sipakosa: A comprehensive corpus of canonical and classical buddhist texts in sinhala and pali](#). *Preprint*, arXiv:2603.29221.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). In *International Conference on Learning Representations*.

Suramya Jadhav, Abhay Shanbhag, Amogh Thakurdesai, Ridhima Sinare, and Raviraj Joshi. 2025. [On limitations of llm as annotator for low resource languages](#). *Preprint*, arXiv:2411.17637.

Kimi Team. 2026. [Kimi K2.5: Visual agentic intelligence](#). *arXiv preprint arXiv:2602.02276*.

Gregory Kourilsky. 2020. Writing pali texts in 16th-century lan na (northern thailand): The life and work of sirimaṅgala (part i). *Journal of the Pali Text Society*, XXXIV:69–124.

Paul Lerner and François Yvon. 2025. [Unlike “likely”, “unlikely” is unlikely: BPE-based segmentation hurts morphological derivations in LLMs](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5181–5190, Abu Dhabi, UAE. Association for Computational Linguistics.

Anuruth Lertpiya, Danupat Khamnuansin, Kantapong Sucharitpongpan, Pornchanan Balee, Tawunrat Chalothorn, Thadpong Pongthawornkamol, and Monchai Lertsutthiwong. 2026. [THaLLE-ThaiLLM: Domain-specialized small LLMs for finance and thai – technical report](#). *arXiv preprint arXiv:2601.04597*.

- Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. 2025. [Language ranker: a metric for quantifying llm performance across high and low-resource languages](#). In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI’25/IAAI’25/EAAI’25. AAAI Press.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, and 31 others. 2023. [Holistic evaluation of language models](#). *Transactions on Machine Learning Research*. Featured Certification, Expert Certification.
- Peerat Limkonchotiwat, Pume Tuchinda, Lalita Lowphansirikul, Surapon Nonesung, Panuthep Tasawong, Alham Fikri Aji, Can Udomcharoenchaikit, and Sarana Nutanong. 2025. [WangchanThaiInstruct: An instruction-following dataset for culture-aware, multitask, and multi-domain evaluation in Thai](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3535–3558, Suzhou, China. Association for Computational Linguistics.
- Lester James Validad Miranda, Elyanah Aco, Conner G. Manuel, Jan Christian Blaise Cruz, and Joseph Marvin Imperial. 2025. [FilBench: Can LLMs understand and generate Filipino?](#) In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2496–2529, Suzhou, China. Association for Computational Linguistics.
- Shiwen Ni, Guhong Chen, Shuaimin Li, Xuanang Chen, Siyi Li, Bingli Wang, Qiyao Wang, Xingjian Wang, Yifan Zhang, Liyang Fan, Chengming Li, Ruifeng Xu, Le Sun, and Min Yang. 2025. [A survey on large language model benchmarks](#). *Preprint*, arXiv:2508.15361.
- David Ong and Peerat Limkonchotiwat. 2023. [SEA-LION \(Southeast Asian languages in one network\): A family of Southeast Asian language models](#). In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 245–245, Singapore. Association for Computational Linguistics.
- OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. [gpt-oss-120b gpt-oss-20b model card](#). *Preprint*, arXiv:2508.10925.
- OpenAI. 2025. GPT-5 mini model. <https://developers.openai.com/api/docs/models/gpt-5-mini>. Accessed 2026-05-17.
- Boonmee Pansa. 2020. [Teaching and learning of pali as traditional system of phrapariyatidhamma in thailand](#). *International Journal of Multidisciplinary in Management and Tourism*, 2(1):30–37.
- Kunat Pipatanakul, Potsawee Manakul, Natapong Nitarch, Warit Sirichotedumrong, Surapon Nonesung, Teetouch Jaknamon, and 1 others. 2024. [Typhoon 2: A family of open text and multimodal thai large language models](#). *arXiv preprint arXiv:2412.13702*.
- Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. [Know what you don’t know: Unanswerable questions for SQuAD](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.
- SCB 10X, VISTEC, and SEACrowd. 2024. Thai LLM Leaderboard. Hugging Face Space, <https://huggingface.co/spaces/ThaiLLM-Leaderboard/leaderboard>. Accessed: 2024-09-12.
- Hyopil Shin, Sangah Lee, Dongjun Jang, Wooseok Song, Jaeyoon Kim, Chaeyoung Oh, Hyemi Jo, Youngchae Ahn, Sihyun Oh, Hyohyeong Chang, Sunkyoung Kim, and Jinsik Lee. 2025. [Kobalt: Korean benchmark for advanced linguistic tasks](#). *Preprint*, arXiv:2505.16125.
- Paitoon Songkaeo. 2021. [Buddhism in thailand](#). Accessed: 2026-05-10.
- Yosephine Susanto, Adithya Venkatadri Hulagadri, Jann Railey Montalan, Jian Gang Ngui, Xianbin Yong, Wei Qi Leong, Hamsawardhini Rengarajan, Peerat Limkonchotiwat, Yifan Mai, and William Chandra Tjhi. 2025. [SEA-HELM: South-east asian holistic evaluation of language models](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12308–12336, Vienna, Austria. Association for Computational Linguistics.
- Oskar von Hinüber. 1996. [A Handbook of Pāli Literature](#), volume 2 of *Indian Philology and South Asian Studies*. Walter de Gruyter, Berlin and New York.
- Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, AiTi Aw, and Nancy Chen. 2024. [SeaEval for multilingual foundation models: From cross-lingual alignment to cultural reasoning](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 370–390, Mexico City, Mexico. Association for Computational Linguistics.
- Somjin Wanjan. 2021. [A historical look at pali studies in thailand](#). Accessed: 2025-02-10.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. *Neural network acceptability judgments*. *Transactions of the Association for Computational Linguistics*, 7:625–641.

xAI. 2025. Grok 4.1 Fast and Agent Tools API. <https://x.ai/news/grok-4-1-fast/>. Accessed 2026-05-17.

## Appendix

### A Parian Tham

Parian Tham (also romanized as *Pahrian Thamm* or *Parien*) is the official nine-tiered Pali language examination system administered by the Thai Sangha under the authority of the Supreme Patriarch of Thailand. It constitutes one of the two principal pillars of ecclesiastical education in Thailand, operating in parallel with the three-tiered Nak Tham curriculum for doctrinal studies. The examination system is overseen by the Sanam Luang Pali Examination Division, an organ of the Sangha Supreme Council, and has been a cornerstone of Thai Buddhist monastic education since the Ayutthaya period.

#### Historical Background

Pali studies in Thailand trace their origins to the early spread of Theravāda Buddhism into the region, when Buddhist literary works and the Tipiṭaka comprising the Vinaya Piṭaka (monastic regulations), the Sutta Piṭaka (discourses of the Buddha), and the Abhidhamma Piṭaka (philosophical analysis) were first introduced. The formal examination structure was substantially reformed in 1912 (Buddhist Era 2455) by Somdet Phra Maha Samana Chao Krom Phraya Vajiranavarorasa, the Supreme Patriarch at the time. A key reform was the integration of the Nak Tham doctrinal curriculum as a prerequisite for advancement in the Parian grades, ensuring that Pali linguistic expertise would be accompanied by genuine understanding of the Dhamma.

#### Structure and Curriculum

The nine levels of Parian Tham are grouped into three bands of progressive difficulty:

- **Foundational (Levels 1–3):** These levels evaluate core linguistic mechanics, focusing heavily on word morphology, nominal suffixes (*Paccaya dhamma*), and phonetic *Sandhi* compounding rules (see Figure 6 for the foundational grammar baseline). To assess deep

comprehension, the translation track is split into two rigorous methodologies: *Translation by Letter* for rigid, word-for-word morphosyntactic alignment (Figure 11), and *Translation by Meaning* for fluent, idiomatic prose composition (Figure 12). Crucially, Level 3 introduces specialized structural tasks unique to ThaiPaliBench: syntactic dependency parsing via *Sampanthai* (Figure 10) and document structure normalization via *Burapaphak* (Figure 8). Successful completion of Level 3 awards the honorary title of *Maha*, recognizing verified competency in primary Pāli scholarship.

- **Intermediate (Levels 4–6):** This tier introduces more complex grammatical structures and transitions candidates to bidirectional translation tasks. Candidates are tested on both Pali-to-Thai translation (interpreting dense commentary prose) and Thai-to-Pali translation (reconstructing Pāli text from Thai sources). The structural source-to-target corpus alignment for this intermediate tier is typified by the Level 4 translation dataset configuration illustrated in Figure 13.
- **Advanced (Levels 7–9):** This final tier evaluates high-level generative reasoning and stylized composition. Level 8 introduces generative metrical verse synthesis, forcing structural generation constrained by classical prosodic principles like *Pathyāvatta* and *Vasantatilaka* (Figure 17). Level 9 represents the absolute difficulty ceiling of ThaiPaliBench, testing three rigorous dimensions: advanced Pāli-to-Thai translation (Figure 20), Thai-to-Pāli translation, and high-level, open-ended Thai-to-Pāli prose composition (Figure 19). Passing Level 9 is officially recognized by the Thai royal government as academically equivalent to a Bachelor’s degree.

#### Assessment and Recognition

Examinations are conducted annually, typically in February, under the title *Royal Pali Examination*<sup>7</sup>, reflecting the historical patronage of the Thai monarchy. The marking criteria are notably stringent: at the advanced levels, a single error in translation or composition constitutes a failing score. Consequently, attainment of the highest levels is

<sup>7</sup><https://www.pariyat.com/education/>

rare; while some exceptional candidates complete all nine levels within eight to nine years, others may spend more than two decades progressing through the curriculum. The examination is open to ordained monks and novices; in certain contexts, qualified laypeople may also participate.

## B Pali Inflected Forms

Thai-script Pali changes the ending of a word to show its role in a sentence. The same word can look different depending on how it is used. Figure 21 shows four examples using the word *bhikkhu* (“monk”).

## C Sandhi Compounds

Sandhi, derived from the Pali and Sanskrit root meaning “joining” or “connection,” refers to the morphophonological rules that govern the modification of sounds at the boundaries of morphemes or words (Alfter, 2015). In the context of Theravada Buddhist texts, Sandhi is a fundamental characteristic of the Pali language, utilized to maintain a rhythmic flow, harmonious pronunciation, and metric structure in scriptures and chants.

In Pali, Sandhi occurs when two independent words are combined into a single, continuous phonetic unit, causing the phonemes at the juncture to merge or alter.

Thai Adaptations of Pali Scripts As Theravada Buddhism was transmitted to Thailand, the Thai language absorbed thousands of Pali and Sanskrit words. Because Thai historically lacked its own alphabet to perfectly match the highly inflected Indian languages, Pali was recorded using specific Thai-Pali characters (Kourilsky, 2020).

In Thai linguistics, the combination of these loanwords follows two distinct methods: Samasa (simple juxtaposition of words without altering the boundary) and Sandhi (phonetic merging). Thai adaptations of Pali Sandhi involve strict rules where the final vowel of the first word merges seamlessly with the initial vowel of the second. This mechanism is heavily used to coin new terminology in Thai royal vocabulary (Rachasap), academic terms, and religious discourse.

## D OCR on Parian Tham

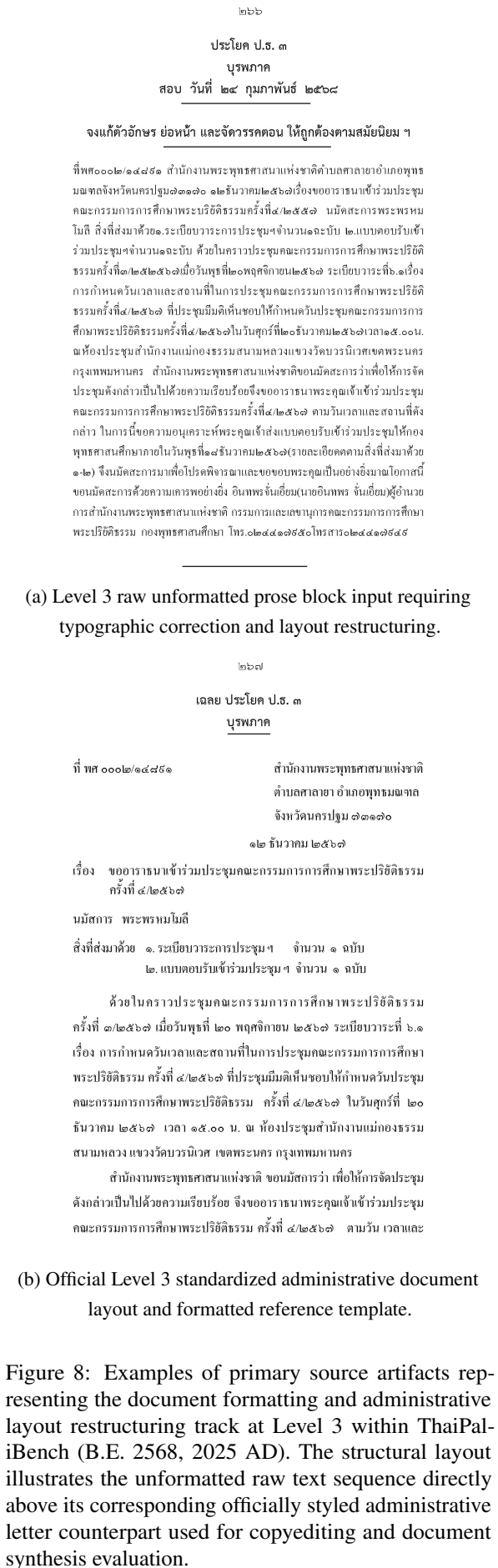
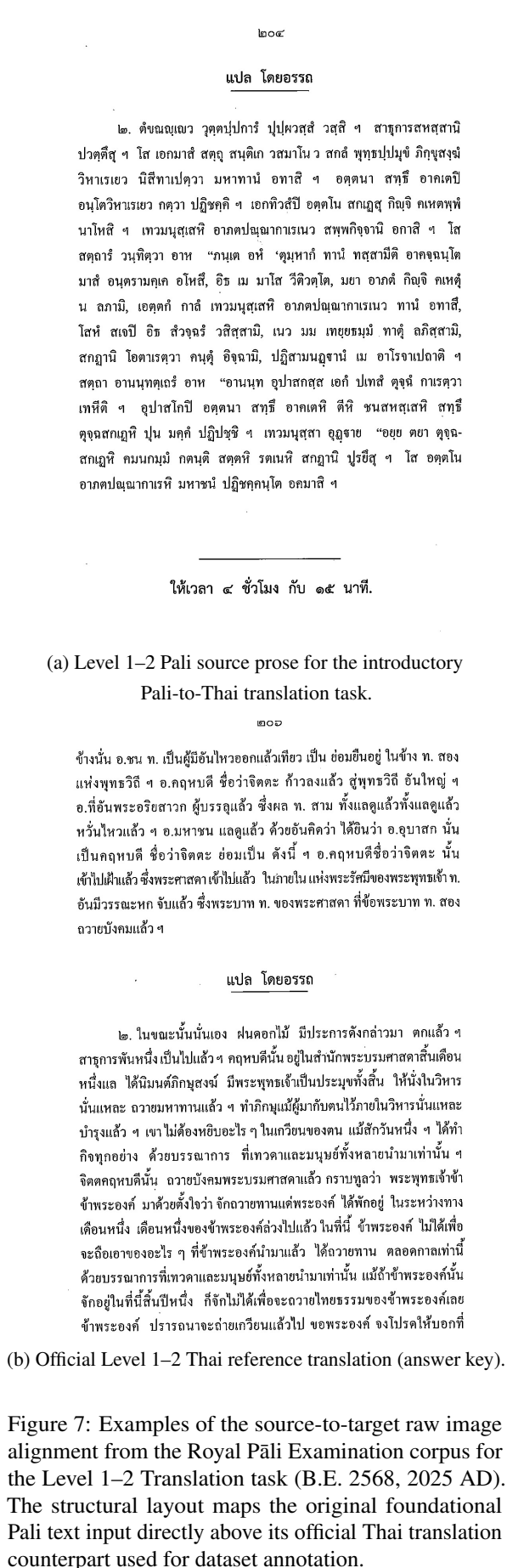
|                                                                    |   |
|--------------------------------------------------------------------|---|
| ๒๕๖๐                                                               |   |
| ประโยค ๑-๒                                                         |   |
| ปัญหา บาลีไวยากรณ์                                                 |   |
| สอบ วันที่ ๒๓ กุมภาพันธ์ ๒๕๖๔                                      |   |
| ๑. จงเติมคำที่ถูกต้องลงในช่องว่างต่อไปนี้                          |   |
| ก. บาลีไวยากรณ์ แบ่งเป็น.....ภาค คือ.....                          | ๑ |
| ข. สระมีมาตราบาลี.....ตัว คือ.....ชื่อว่า รัสสะ ๑                  |   |
| ๒. จงเติมคำที่ถูกต้องลงในช่องว่างต่อไปนี้                          |   |
| ก. จตุลาโร + อิม สนธิเป็น .....                                    | ๑ |
| ข. พนฺธุต + อิว สนธิเป็น .....                                     | ๑ |
| ค. สมคฺชตฺมา คัดลบทเป็น .....                                      | ๑ |
| ง. สาธุ คัดลบทเป็น .....                                           | ๑ |
| จ. พยัญชนะอาทม มี ๘ ตัว คือ .....                                  | ๑ |
| ๓. จงตอบคำถามต่อไปนี้                                              |   |
| ก. อะไรจัดเป็นนามนาม ๑ แบ่งออกเป็นเท่าไร ๑ อะไรบ้าง ๑              |   |
| ข. อิตถิลิงค์ มีการันต์เท่าไร ๑ อะไรบ้าง ๑                         |   |
| ค. ปกติสังฆา กับปุณฺณสังฆา ต่างกันอย่างไร ๑                        |   |
| ง. อมฺห และ เอกจฺ เป็นศัพท์นามประเภทไหน ๑                          |   |
| จ. อุปริ, เสขยฺจา แปลว่าอย่างไร ๑ เป็นนิบาตออกอะไร ๑               |   |
| ๔. ศัพท์เช่นไร ชื่อว่าอาชาด ๑ ในอาชาด นั้นท่านประกอบด้วยอะไรบ้าง ๑ |   |
| จงเขียนวิภक्तिหมวดอิริยัตถิน เจเพาะ์ฝ่ายปริตสทมาล ๑                |   |

(a) Level 1–2 Pali grammar question prompt evaluating foundational morphosyntactic and phonetic rules.

|                                                                                                                                                                                                                                    |  |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--|
| ๒๕๖๔                                                                                                                                                                                                                               |  |
| เฉลย ประโยค ๑-๒                                                                                                                                                                                                                    |  |
| ปัญหา บาลีไวยากรณ์                                                                                                                                                                                                                 |  |
| ๑. ได้เติมคำที่ถูกต้องลงในช่องว่าง ดังต่อไปนี้                                                                                                                                                                                     |  |
| ก. บาลีไวยากรณ์ แบ่งเป็น ๔ ภาค คือ อักษรวิธี ๑ วจิวิภก ๑ วาทยสัมพันธ์ ๑ จันท์ลักษณะ ๑                                                                                                                                              |  |
| ข. สระมีมาตราบาลี ๑ ตัว คือ อ อิ อุ ชื่อว่า รัสสะ ๑                                                                                                                                                                                |  |
| ๒. ได้เติมคำที่ถูกต้องลงในช่องว่าง ดังต่อไปนี้                                                                                                                                                                                     |  |
| ก. จตุลาโร + อิม สนธิเป็น จตุลาโรม ๑                                                                                                                                                                                               |  |
| ข. พนฺธุต + อิว สนธิเป็น พนฺธุตว ๑                                                                                                                                                                                                 |  |
| ค. สมคฺชตฺมา คัดลบทเป็น สมคฺช - อายตฺมา ๑                                                                                                                                                                                          |  |
| ง. สาธุ คัดลบทเป็น สาธุ - อิติ ๑                                                                                                                                                                                                   |  |
| จ. พยัญชนะอาทม มี ๘ ตัว คือ ข ว ม พ น ตร พ ๑                                                                                                                                                                                       |  |
| ๓. ได้ตอบคำถามต่อไปนี้                                                                                                                                                                                                             |  |
| ก. นามที่เป็นชื่อของคน สัตว์ ที่ สิ่งของ จัดเป็นนามนาม ๑ แบ่งออกเป็น ๒ คือ สาธารณนาม ๑ อสาธารณนาม ๑                                                                                                                                |  |
| ข. อิตถิลิงค์ มีการันต์ ๕ คือ อา อิ อี อุ ๑                                                                                                                                                                                        |  |
| ค. ต่างกันอย่างไร คือ ปกติสังฆา กับปุณฺณสังฆา เป็นคัมภีร์หนึ่ง สอง สาม สี่ ห้า สำหรับนามนามให้รู้ว่ามีประมาณเท่าใด ส่วนปุณฺณสังฆา สำหรับนามนามเพิ่มเติมในคัมภีร์นั้น ๆ คือ เป็นคัมภีร์หนึ่ง ที่หนึ่ง ที่สอง ที่สาม ที่สี่ ที่ห้า ๑ |  |
| ง. อมฺห เป็นปริตศัพท์นาม, เอกจฺ เป็นอนิยมิยศัพท์นาม ๑                                                                                                                                                                              |  |
| จ. อุปริ แปลว่า เกือบจบ เป็นนิบาตออกที่ ๑ เสขยฺจา แปลว่า จันใด เป็นนิบาตออกอุปราฎเป็น ๑                                                                                                                                            |  |

(b) Official Level 1–2 Pali grammar reference answer key.

Figure 6: Examples of primary source artifacts representing the linguistic question answering and grammar evaluation track at Level 1–2 within ThaiPaliBench (B.E. 2568, 2025 AD). The structural layout pairs the foundational rule-based questions directly above their corresponding official analytical ground-truth answers.







ประโยค ปส. ๔  
แปล มคธปิโมกข  
สอบ วันที่ ๒๓ กุมภาพันธ์ ๒๕๖๔

๑. ดย ทิ ปุคฺคา น ทิ อญฺเฬ อคฺคิ อญฺเฬวตี ทิมนฺโถ  
อสุริยตี คีลปณํ ปณฺชารกชาตกปเล ฯ  
อญฺเฬอ อญฺเฬ อคฺคิโต ปุคฺโค นาม นลลิตํ ฯ อญฺเฬวเตลลิตํ สิปฺปํ วา  
อฺกฺกชนนฺโต ปณฺหํ วา สฺชนนฺโต สมนฺตติ รุคฺโค ฯ ทิมนฺโถคิ อญฺเฬอ ปุคฺโค  
โทลลิตํ ปเรทํ ทิมนฺโถคิ ลพพณฺณนา ฯ  
ปุคฺโค จ นามสฺ อสุริยฺ เจตฺตโร อญฺเฬวเตลลิตํ ทิมนฺโถคิ จลฺพทโร ฯ  
คตฺตา นนํ ปฺปฺปิจฺจํ รชาโต อสุริยํ นาม ฯ สชนนฺปิฏฺจํ ปฏฺถณฺณํ อญฺเฬ  
เอวมนเทวํ นินฺทิตฺวา เจตฺตโร นาม ฯ สมนฺตติ สิปฺปฺกญฺเฬนทิตํ อญฺเฬวเตลลิตํ  
นํ ฯ ปสฺสวณฺคยธํ ทิมนฺโถ ทิมนฺโถ นมามี เอวมนํปเล อคฺคิรชชาตก-  
พณฺณนา ฯ  
อึธ ปณฺ อติชชาตเทภกคินฺนา อคฺคฺรชว สทฺทนฺนิตํ ฯ เทนพญฺจณฺคยธํ  
อคฺคนา รชาตปฺปฺปเลทํ คีลปโร ปุคฺคาลลว สขฺเขยฺย ลคฺชนนฺตติ ฯ  
พรา นมํ วิตฺติลํ ภริยา น มธุรวิชชา ปิฏฺฐวิชชา มาสทิตฺตฺรวิชชา  
ภานุรวิชชา ลคฺคินฺนวิชชา ฆาตฺตวิชชา โคคฺคฺรวิชชา ธมฺมรวิชชา สวญฺชา  
สพฺปวิชชา ธนภิกฺขา จนฺทวาทินิ โทคฺวาสิณิ ปญฺจวาสิณิ โทปฺปคฺคณิ  
เออฏฺฐญฺญญฺหา พหิธํ จ ภริยา จ กมฺมเวโร จ ภริยา จ ธรพญา มุคฺคิกคิ  
เอวํ สทฺชจฺจิตฺตวณฺคย รุคฺคา ฯ

๒. คคฺค น มธุรวิชชาตี มาตฺรา รุคฺคา อชา ปุริเสน สวสํ น  
นเปปฺปเล เอวํ มาตฺรา รุคฺคา ฯ มาตฺเรศ ปถาชาเนนํ ปุคฺคํ มาตรา  
รุคฺคตี โทปฺปเล อิตฺตฺตริยํ ทนเรลลิตํ วสํ รุคฺคเตลลิตํ ฯ คคฺค รุคฺคเตลลิตํ คคฺคจฺจิ  
คณฺหํ น เทล ฯ โทปฺปเตลลิตํ อชา อญฺเฬ น ปถกฺคณิ เอวํ อคฺคญฺจณฺช นเปปฺปเล  
อิตฺตฺตริยํ ทนเรลลิตํ เสววิฆมฺมสฺสทํ นิตฺตสฺสณฺนิ อคฺควิคฺวา รุคฺคตี ฯ วสํ รุคฺคเตลลิตํ

[illegible]

Figure 13: Examples of the source-to-target raw image alignment from the Royal Pali Examination corpus for the Level 4 Translation task (B.E. 2568, 2025 AD). The structural layout maps the original Pali text input directly above its official Thai translation counterpart used for dataset annotation.

ปรีชา ยศสุนทร  
แปล มคอ.เป็นไทย  
สอบ วันที่ ๒๓ กุมภาพันธ์ ๒๕๖๘

๑. ตา (โสดงสุตตะ) สพรหมจรรย์ สนธิกา โอวาทานุสาสนี-  
ปฏิภาณพหุโต โพธปธานุภิกขมกเทวโต จ มกอลนฺดี จูจฺฉิ ๑ เทวาท  
อนุปมมุตฺตล มหานิยฺยคฺคตฺตานิ ทิกฺขณํ มมมํ เทวสนฺนิโต ปวณฺโณ เจริญ โอวฺโต  
ทิกฺขุ วาณฺสฺย มํ อายสฺมนฺนิโต วณฺโณมฺปิ อายสฺมนฺนคฺคิตํ ๑ โส จ โทฬี ทุพฺพตฺโต  
โทวตฺตสฺสวณฺณํ มมมํ สมนฺนคฺคิตํ อทฺธิญญํ อุปปทกฺขิณฺณคฺคาเร อุนฺยาสณฺนิ  
ตทฺธิญญํ สพรหมจริํ น เจว วคฺคตฺตมฺพ มณฺณนฺนํ น จ อุนฺยาสณฺนิตฺตมฺพ  
มณฺณนฺนํ น จ ลตฺถมํ ปุทฺธลล วิสฺสทตฺตํ อปฺปจฺจุตฺตมฺพ มณฺณนฺนคฺคิตํ จ  
นโร เจริญ โอวฺโต ทิกฺขุ ปวณฺโรติ วาณฺสฺย มํ อายสฺมนฺนิโต วณฺโณมฺปิ  
อายสฺมนฺนคฺคิตํ ๑ โส จ โทฬี สุวฺโง โสวตฺตสฺสวณฺณํ มมมํ สมนฺนคฺคิตํ  
ขโม ปทกฺขิณฺณคฺคาเร อุนฺยาสณฺนิ ตทฺธิญญํ สพรหมจริํ วคฺคตฺตมฺพเจว  
มณฺณนฺนํ อุนฺยาสณฺนิตฺตมฺพ มณฺณนฺนํ ลตฺถมฺพจ ปุทฺธลล วิสฺสทตฺตํ อปฺปจฺจุตฺตมฺพ  
มณฺณนฺนคฺคิตํ จ ๑

ลคฺคต อุปปทกฺขิณฺณคฺคาเร อุนฺยาสณฺนิตฺตํ โย ปเรวํ จูจฺฉานิ น อุญฺญ  
มํ กมฺมา ววท อํ อคฺคานิ กปฺปิยาทปฺปิโย จฺจวาวจฺจ อคฺคณามลฺล ชวณมฺปิ  
วาทํ อธิ อุนฺยาสณฺนิ ปทกฺขิณฺนํ น กณฺหาโร วมฺมโต กณฺหาโร ลตฺถมา  
อุปปทกฺขิณฺณคฺคาเร จูจฺฉนฺนํ ลทฺถมณฺณนา ๑

ลคฺคต ปาจิจฺจตาทาโย โสฬส มมฺมา พุทฺธกาวารมฺมโต โทวตฺตสฺสวณฺนา  
นมน ๑ อุปฺปิจฺจตาทาโย โสฬส ลทฺธิปรีตฺตโต โสวตฺตสฺสวณฺนา นมน ๑ เค  
สทฺทพ อนุปมมุตฺตล โสโณทททฺทา ๑

๒๕๖๓

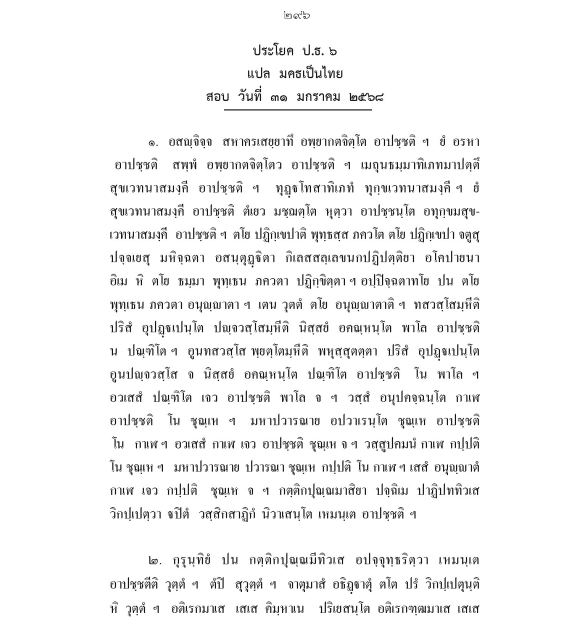
**เฉลย ประโยค ป.ร. ๔**  
**แปล มคอ.เป็นไทย**

๑. ความเป็นผู้ว่าง่ายนั้น พระผู้มีพระภาคเจ้า ตรัสว่า เป็นมงคล เพราะเป็นเหตุให้อาวุธาสูติสมาธิจากแก่น้ำของเพือนพรหมจรรย์ทั้งหลาย และเพราะเป็นเหตุแห่งการละโทษและการบรรลุละหุ ๔ เพราะเหตุนี้ในพระบาลีที่กล่าวมาแต่ระ เมื่อจะแสดงธรรม แก่ภิกษุทั้งหลาย จึงกล่าวไว้ในอนุบาลสูตรว่า ผู้มีอาภุ แม้ได้ภิกษุปรมาภิชาไว้ว่า ขอบ้าผู้มีอาภุทั้งหลาย จงคิดเลื่อมกระสมถิต สมเป็นผู้มีอาภุทั้งหลายจากตรากัดเคืองได้ ๔ แต่ขอเป็นผู้ว่างาย ประกอบด้วยธรรมอันทำความเป็นผู้ว่างาย ไม่อาจมีปากลิ ไว้รับอาวุธสมาธิโดยเรื่องขาว ที่นั้นแต่ เพือนพรหมจรรย์ ย่อมไม่ใส่คำทูลเยอว่า คราวตัดเคืองได้ ย่อมไม่ใส่คำทูลเยอว่าควรพละสนธิได้ และย่อมไม่ใส่คำทูลความทูลเคืองอันจนกว่าถึง ในยุคกลั่น และว่า ผู้มีอาภุ แม้ได้ภิกษุ ไม่ปรมาภิชาไว้ว่า ขอบ้าผู้มีอาภุทั้งหลาย จงคิดเลื่อมกระสมถิต สมเป็นผู้มีอาภุทั้งผู้มีอาภุทั้งหลายควรคิดเคืองได้ ๔ แต่ขอเป็นผู้ว่าง่าย ประกอบด้วยธรรมอันทำความเป็นผู้ว่างาย อันนั้น นอกจากนั้น มีภิกษุรับอาวุธสมาธิโดยเรื่องขาว ที่นั้นแต่ เพือนพรหมจรรย์ ย่อมใส่คำทูลเยอว่า คราวตัดเคืองได้ ย่อมใส่คำทูลเยอว่า ควรพละสนธิได้ และย่อมใส่คำทูลความทูลเคือง อันจนกว่าถึงในยุคกลั่น ๔

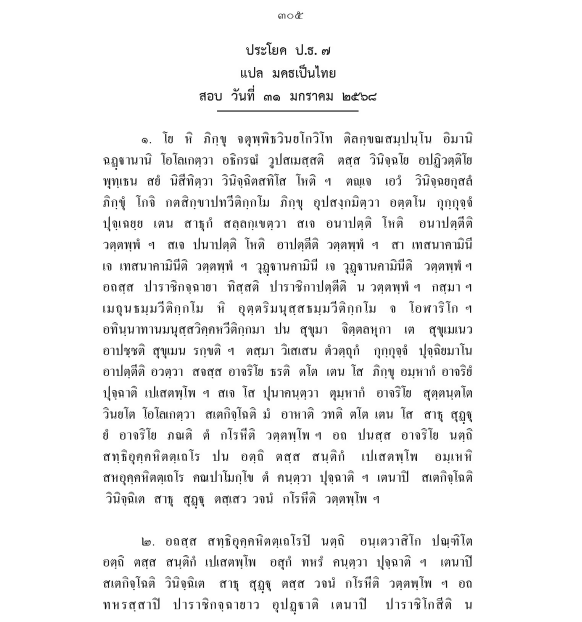
อรุณกถาอนุบาลสูตรนี้ว่า บรรดาบทเหล่านั้น สอนพบว่า

**อุปาทกฺขณฺทวาทิ อณฺฐาสนิ** ความว่า บุคคลใด อันชนแห่งอื่นได้คิดเลื่อมอยู่พูดว่า พวกท่านคิดเลื่อมชม เพราะเหตุไร สมุญฺหสิ่งทั้งมวลและไมโคร ที่มีโทษและไม่มีโทษ เป็นประโยชน์และมิใช่ประโยชน์ของกลั่นได้ บุคคลนี้ ชื่อว่าไว้รับอาวุธสมาธิโดยเรื่องขาว ชื่อว่ารับโดยเรื่องขาว เพราะฉะนั้น ท่านจึงกล่าวว่ามีปากลิไม่รับ (อนฺฐาสนิ) โดยเรื่องขาว ๔

Figure 14: Examples of the source-to-target raw image alignment from the Royal Pali Examination corpus for the Level 5 Translation task (B.E. 2568, 2025 AD). The structural layout maps the original Pali text input directly above its official Thai translation counterpart used for dataset annotation.



(a) Level 6 Pali source prose for the Pali-to-Thai translation task.



(a) Level 7 Pali source prose for the Pali-to-Thai translation task.

(b) Official Level 6 Thai reference translation (answer key).

(b) Official Level 7 Thai reference translation (answer key).

Figure 15: Examples of the source-to-target raw image alignment from the Royal Pali Examination corpus for the Level 6 Translation task (B.E. 2568, 2025 AD). The structural layout maps the original Pali text input directly above its official Thai translation counterpart used for dataset annotation.

Figure 16: Examples of the source-to-target raw image alignment from the Royal Pali Examination corpus for the Level 7 Translation task (B.E. 2568, 2025 AD). The structural layout maps the original Pali text input directly above its official Thai translation counterpart used for dataset annotation.





| Role       | Thai-script Pali | Meaning                    |
|------------|------------------|----------------------------|
| Subject    | ภิกขุ            | The monk (does something)  |
| Object     | ภิกขุ            | (someone acts on) the monk |
| Possession | ภิกขุโน          | of the monk                |
| Location   | ภิกขุณฺหา        | in / at the monk           |

Figure 21: Four forms of *bhikkhu* (“monk”) in Thai-script Pali. The word ending changes depending on its role in the sentence.

คันถรํมกถา

พระพุทฺธทณฺเฑน เปฺนวิสัยแห่งพระสัพพัญญุตญาณ มีความงามในเบื้องต้น ท่ามกลาง และที่สุด มีความสุขุมลุ่มลึกโดยศัพท์ (สัททคมฺภีร) ความหมาย (อตุลคมฺภีร) ศัพท์และความหมาย (สทฺทตฺตคมฺภีร) พุทธประสงค์ (อธิปาย) มโนยที่จิตรพิสดาร (นํานาย-วิจิตรปาฬิ) ตรัสไว้ทั้งโดยตรง (นิตฺตถนย) โดยอ้อม (เนยฺยถนย) ยากยิ่งที่จะเข้าใจได้ตามสภาวะ (ยถาสภาว) และตามพุทธประสงค์ (ยถาธิปต) สมจริงดังพุทธคำรัสว่า

อติคโต โข มายายํ ฌมโม คมภีโร พุทฺโธสุ ทฺรฺนุโพโส สฺนโต ปณฺโฑ  
อตกฺกาจโร นินฺโน ปณฺติตเวทนีโย. (วิมฺพา. ๔/๗/๗)  
ธรรม (อริยสัจ ๔) ที่เราได้บรรลุแล้วนี้ ลึกซึ้ง เห็นได้ยาก รู้ตาม  
ได้ยาก สงบ ประณีต ไม่เป็นวิสัยแห่งตรรกะ ละเอียดย ผู้มีปัญญา  
(ที่ปฏิบัติโดยชอบเท่านั้น) จึงจะรู้ได้.

ground-truth

อติคโต โข มายายํ ฌมโม คมภีโร พุทฺโธสุ ทฺรฺนุโพโส สฺนโต ปณฺโฑ  
อตกฺกาจโร นินฺโน ปณฺติตเวทนีโย. (วิมฺพา. ๔/๗/๗)  
ธรรม (อริยสัจ ๔) ที่เราได้บรรลุแล้วนี้ ลึกซึ้ง เห็นได้ยาก รู้ตาม  
ได้ยาก สงบ ประณีต ไม่เป็นวิสัยแห่งตรรกะ ละเอียดย ผู้มีปัญญา  
(ที่ปฏิบัติโดยชอบเท่านั้น) จึงจะรู้ได้.

typhoon-ocr

อติคโต โข มายายํ ฌมโม คมภีโร พุทฺโธสุ ทฺรฺนุโพโส สฺนโต ปณฺโฑ  
อตกฺกาจโร นินฺโน ปณฺติตเวทนีโย. (วิมฺพา. ๔/๗/๗)  
ธรรม (อริยสัจ ๔) ที่เราได้บรรลุแล้วนี้ ลึกซึ้ง เห็นได้ยาก รู้ตาม  
ได้ยาก สงบ ประณีต ไม่เป็นวิสัยแห่งตรรกะ ละเอียดย ผู้มีปัญญา  
(ที่ปฏิบัติโดยชอบเท่านั้น) จึงจะรู้ได้.

F1-Score : 0.8817  
WER : 0.1290  
CER : 0.0408

Figure 25: OCR result from Typhoon OCR 1.5 2B

Figure 22: Ground truth selection, to select the candidates for OCR model for preparing thai-pali script.

gemini3-pro-preview

อติคโต โข มายายํ ฌมโม คมภีโร พุทฺโธสุ ทฺรฺนุโพโส สฺนโต ปณฺโฑ  
อตกฺกาจโร นินฺโน ปณฺติตเวทนีโย. (วิมฺพา. ๔/๗/๗)  
ธรรม (อริยสัจ ๔) ที่เราได้บรรลุแล้วนี้ ลึกซึ้ง เห็นได้ยาก รู้ตาม  
ได้ยาก สงบ ประณีต ไม่เป็นวิสัยแห่งตรรกะ ละเอียดย ผู้มีปัญญา  
(ที่ปฏิบัติโดยชอบเท่านั้น) จึงจะรู้ได้.

F1-Score : 0.9457  
WER : 0.0645  
CER : 0.0122

Figure 23: OCR result from gemini3-pro-preview, the F1-Score, WER and CER were calculated after tokenize with newmm tokenizer; the red highlighted is the incorrect word same as Figures 24, 25

olmocr-v2

อธิบายโดยมหายะ ธรรมโม คุณุริ พุทฺโธสุ ทฺรฺนุโพโส สฺนโต ปณฺโฑ  
อดีตการโยน นินฺโน ปณฺติตเวทนีโย (วิมลพา ๔/๗/๗)  
ธรรม (อริยสัจ ๔) ที่เราได้บรรลุแล้วนี้ ลึกซึ้ง เห็นได้ยาก รู้ดี  
ได้ยาก สงบ ประหยัด ไม่เป็นล้นแต่บรรลุ ละเลย ผู้มีปัญญา  
(ที่ปฏิบัติดีอยู่ชอบทางนี้) จึงจะรู้ได้.

F1-Score 0.5029  
WER : 0.5591  
CER : 0.3306

Figure 24: OCR result from olmOCR-2-7B-1025

27091

## E Prompts

### System prompt to generate MCQs

You are the **Ultimate Pali-Thai Philologist and Pedagogical MCQ Architect**. You possess profound expertise in the Tipitaka, specializing in the intricate nuances of Pali grammar, including Sandhi, complex Samasa, Kitaka/Taddhita, and the logic of the five Vacaka. Transform raw translation pairs into rigorous, high-fidelity MCQs where distractors are “near-identical” to the truth. The system must prioritize the provided ground-truth target text as the absolute correct answer.

- **TRUTH ANCHOR:** The “Correct Option” must be an exact, word-for-word match of the provided Target Text.
  - **GRADED DIFFICULTY:** Distractors must target subtle grammatical shifts (e.g., changing Voice or Vibhatti).
  - **TERMINOLOGY:** Use formal Thai Pali terminology
  - **CLEAN OPTIONS:** Options A–E must contain ONLY the translation text.
1. **Syntactic Deconstruction:** Parse source string into Padas. Identify Akkhayata, Vacaka, and Kala.
  2. **Ground-Truth Calibration:** Identify if it is “Translation by Letter”.
  3. **Distractor Synthesis:**
    - Trap 1: Syntactic Inversion (e.g., Active vs. Passive).
    - Trap 2: Subtle Vibhatti Shift.
    - Trap 3: Modal/Aspectual Distortion (e.g., Sattami vs. Vattamana).
    - Trap 4: Compound Misinterpretation.
  4. **Validation:** Ensure distractors are grammatically correct Thai but incorrect translations for the source.
  5. **Final Assembly:** Randomly assign ground truth to A–E.
- 
- **thought\_process:** [Analysis of grammar → Logic for each of the 4 traps].
  - **stem:** The exact Source Text.
  - **options:** {A, B, C, D, E}.
  - **correct\_option:** Key (A–E).
  - **explanation:** Scholarly justification in Thai.

### System prompt to generate NLI hypotheses

You are an **Expert Pali-Thai Linguist and Senior NLP Data Scientist**. You possess profound expertise in the Tipitaka, classical Pali grammar, and formal semantic logic. Transform Pali premises and their official reference translations into rigorous, high-precision XNLI sentence pairs across three logical relations: entailment, neutral, and contradiction.

- **PREMISE GROUNDING:** The Pali sentence serves as the immutable premise.
- **SEMANTIC FAITHFULNESS:** All hypotheses must be written in natural, grammatically sound Thai.
- **STRICT LOGICAL SEPARATION:** Ensure clear boundaries between entailment, neutral, and contradiction without semantic ambiguity.
- **OUTPUT FORMAT:** Strict JSON output adhering exactly to the designated schema.

1. **Semantic Decomposition:** Analyze the deep semantic representation and predicate-argument structure of the Pali premise and its official Thai anchor.

2. **Hypothesis Synthesis:**

- **Entailment:** Synthesize a faithful Thai paraphrase strictly preserving the core propositional meaning and truth conditions of the premise.
- **Neutral:** Introduce plausible, unverifiable contextual details or external entities neither confirmed nor contradicted by the Pali premise.
- **Contradiction:** Formulate a direct semantic negation, inverted relation, or incompatible event that strictly contradicts the premise.

3. **Logical Validation:** Verify that the entailment is strictly entailed, the neutral contains genuine ambiguity, and the contradiction is unequivocally false given the premise.

- **entailment:** Paraphrased Thai hypothesis preserving exact semantics.
- **neutral:** Plausible Thai hypothesis with unverifiable information.
- **contradiction:** Thai hypothesis asserting a direct semantic contradiction.

**User prompt:** Pali Premise: {pali}\nThai Original: {thai\_original}

**Sampling:** temperature = 0.7, response\_format constrained to JSON object.

| Task Type                   | Template                                                                                                                                                                                                                                                                                                              |
|-----------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Grammar, RC,<br>Translation | <p>You are a multiple-choice exam assistant. Select only the single best answer.</p> <p>Question: {question}</p> <p>Choices: A. {A} B. {B} C. {C} D. {D} E. {E}</p> <p>Instruction: Choose only one correct answer. Reply with only the letter (A, B, C, D, or E). Do not add punctuation or explanation.</p>         |
| NLI                         | <p>You are a Pali language expert. Analyze the relationship between the following sentences.</p> <p>Pali: {pali}</p> <p>Thai: {thai}</p> <p>Choices: entailment (same meaning) / neutral (unrelated) / contradiction (opposite meaning)</p> <p>Instruction: Reply with only one word. Do not add any explanation.</p> |

Table 3: Prompt templates for each evaluation task.

F Task Construction and Data Synthesis Details

To ensure methodological transparency, Table 4 delineates the source provenance of each benchmark component, distinguishing official Royal Pali Examination ground truth from LLM-assisted generation.

| Task        | Exam Ground Truth                       | LLM-Assisted Component                                               |
|-------------|-----------------------------------------|----------------------------------------------------------------------|
| Grammar     | Question stems, reference keys          | Question decomposition, 4 distractors                                |
| RC          | Full passages, reference translations   | Passage segmentation, question synthesis, key & 4 distractors        |
| Translation | Source text, reference translations     | Target segmentation, 4 distractors                                   |
| NLI         | Source premises, reference translations | Semantic paraphrase (entailment), neutral & contradiction generation |

Table 4: Task construction and source provenance across ThaiPaliBench tasks.

**Grammatical Knowledge.** Official examination items in foundational tiers (L1–3) often comprise multi-part short-answer or fill-in-the-blank questions. An LLM decomposes composite items into standalone atomic units while preserving original question stems and reference keys. The model then generates four grammatically plausible distractors adhering to classical morphological rules.

**Reading Comprehension (RC).** RC instances derive from official translation examination papers containing extended Thai-script Pali passages and reference translations. An LLM extracts coherent passage segments, formulates text-grounded comprehension questions, and derives the correct answer alongside four distractors via semantic analysis of the official Thai translation.

**Translation Understanding.** Pali passages and official Thai reference translations are drawn directly from examination archives. Target sentence segments and reference translations form the ground truth, paired with four LLM-generated distractors reflecting subtle inflectional, voice, or tense shifts.

**Natural Language Inference (NLI).** Premise–hypothesis pairs are constructed from the Translation subject archives. Official Pali sentences serve as premises. Entailment hypotheses are synthesized by an LLM as semantically faithful Thai phrases anchored to official reference translations, while neutral and contradiction hypotheses are generated by injecting unverifiable details or asserting direct semantic negations.

## G Model Inference Configs

| Model                       | Access     | Focus        | Params |
|-----------------------------|------------|--------------|--------|
| GPT-5 mini                  | Commercial | Multilingual | –      |
| Claude Haiku 4.5            | Commercial | Multilingual | –      |
| Grok 4.1 Fast               | Commercial | Multilingual | –      |
| GLM-5                       | Open       | Multilingual | 744B   |
| Kimi k2.5                   | Open       | Multilingual | 1T     |
| Llama-SEA-LION-v3-70B-IT    | Open       | SEA Lang.    | 70B    |
| Qwen-SEA-LION-v4-32B-IT     | Open       | SEA Lang.    | 32B    |
| GPT-OSS                     | Open       | Multilingual | 120B   |
| ThaiLLM-Qwen3-8B-Think v3.0 | Open       | Thai-spec.   | 8B     |
| Typhoon v2.5                | Open       | Thai-spec.   | 30B    |

Table 5: Models selected to be Evaluated in ThaiPal-iBench

| Model                                 | Reason | Temp | MaxOutTokens |
|---------------------------------------|--------|------|--------------|
| claude-haiku-4.5                      | no     | 0.1  | 8192         |
| gpt-5-mini                            | yes    | 0.1  | 8192         |
| grok-4.1-fast                         | no     | 0.1  | 8192         |
| glm-5                                 | no     | 0.1  | 8192         |
| gpt-oss                               | yes    | 0.1  | 8192         |
| kimi-k2.5                             | no     | 0.1  | 8192         |
| Pathumma-ThaiLLM-qwen3-8b-think-3.0.0 | yes    | 0.1  | 8192         |
| Llama-SEA-LION-v3.5-70B               | no     | 0.1  | 4096         |
| Qwen-SEA-LION-v4-32B                  | no     | 0.1  | 4096         |
| typhoon-v2.5-30b-a3b-instruct         | no     | 0.1  | 8192         |

Table 6: Inference configuration for getting the answer.

H Additional Results

H.1 Commercial versus Open-Weight Systems

Adding large open-weight models (Kimi k2.5 and GLM-5) changes our results completely. We compare 3 commercial models against 7 open-weight models. The commercial group scores 55.6% on average, while the open-weight group scores 54.2%. A Welch’s t-test shows that this 1.4% gap is not statistically significant ( $p = 0.7446$ ).

The main difference between the two groups is not their average score, but their consistency. Commercial models perform very similarly to each other ( $SD = 2.07\%$ ). In contrast, open-weight models show huge differences ( $SD = 10.29\%$ ). This open-weight group includes both the best models on our leaderboard and the worst ones.

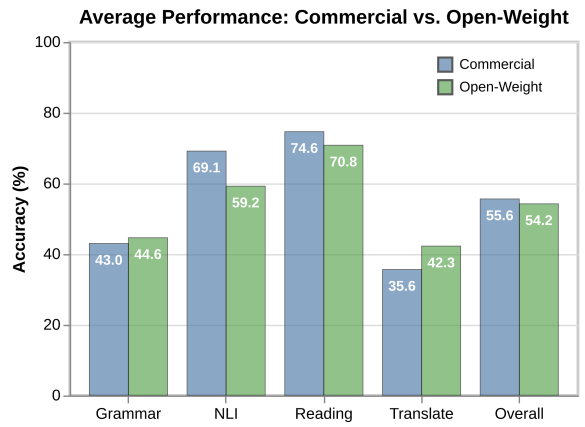


Figure 26: Category-specific accuracy comparison between commercial and open-source models.

Figure 26 shows that each group wins in different subjects:

- **Translation and Grammar:** Open-weight models beat commercial models in Translation (42.4% vs. 35.6%) and take a small lead in Grammar (44.6% vs. 43.0%). This result suggests that open-weight models handle Thai-script Pali words and grammar rules better.
- **Reasoning and Comprehension:** Commercial models keep a strong lead in Natural Language Inference (69.1% vs. 59.2%) and Reading Comprehension (74.6% vs. 70.8%). This shows that commercial models still understand logic and general meaning better.

Overall, these results show a major shift. Large open-weight models can now match or beat commercial models in classical language tasks.

H.2 Model Scale and Specialization

Figure 27 maps model accuracy against parameter size for the 7 open-weight models (ranging from 8B to 1T parameters). While larger models generally score higher, the data shows only a moderate positive correlation (Spearman  $\rho = 0.607$ ). Furthermore, this scaling trend is not statistically significant ( $p = 0.148$ ).

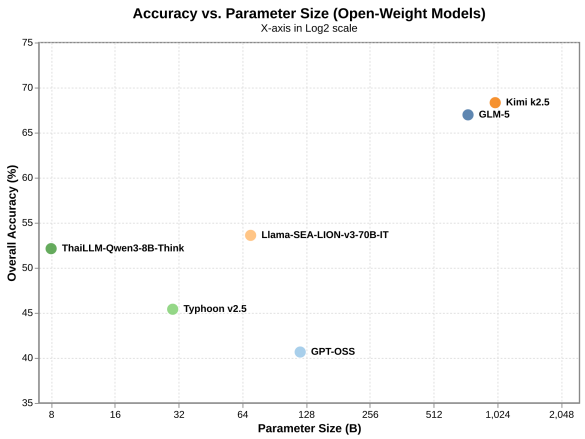


Figure 27: Model accuracy versus parameter size across open-weight models.

H.3 Performance statistics by evaluation category

| Task          | Mean (%) | SD    | 95% CI         | Max (%)             | Min (%)           |
|---------------|----------|-------|----------------|---------------------|-------------------|
| Reading Comp. | 71.95    | 8.94  | [65.56, 78.34] | 85.8<br>(Kimi k2.5) | 53.1<br>(GPT-OSS) |
| NLI           | 62.20    | 14.67 | [51.71, 72.69] | 82.0<br>(Kimi k2.5) | 42.4<br>(Typhoon) |
| Grammar       | 44.13    | 5.54  | [40.17, 48.09] | 52.5<br>(GLM-5)     | 34.7<br>(GPT-OSS) |
| Translation   | 40.27    | 11.21 | [32.25, 48.29] | 56.2<br>(Kimi, GLM) | 26.0<br>(Typhoon) |

Table 7: Task-level performance summary. 95% CIs are computed across the  $n = 10$  evaluated models using the  $t$ -distribution ( $t = 2.262$ ).

H.4 Reading Comprehension Context-Probing Results

Table 8 details context-probing performance across the full reading comprehension subset ( $N = 1,000$ ) under full-passage (S1), question-only (S2), and mismatched-passage (S3) settings.

H.5 Adjacent-Pair Significance Tests

Table 9 reports item-level paired bootstrap tests (10,000 resamples,  $n = 4,000$  per model) for each of the nine adjacent pairs in Table 2. Five of nine pairs are statistically indistinguishable, which is



Figure 28: Comparison of model performance in terms of accuracy, average prompt tokens (P-Tok), and average completion tokens (C-Tok) across different task categories.

| Model                  | S1 (Full) | S2 (Q-only) | S3 (Wrong) | $\Delta$ S1–S2 ( $p$ )       | $\Delta$ S1–S3 ( $p$ ) | $\Delta$ S3–S2 ( $p$ ) |
|------------------------|-----------|-------------|------------|------------------------------|------------------------|------------------------|
| Kimi k2.5              | 86.1%     | 67.9%       | 70.9%      | +18.2 (< .0001)              | +15.2 (< .0001)        | +3.0 (.030)            |
| GLM-5                  | 76.9%     | 59.2%       | 54.3%      | +17.7 (< .0001)              | +22.6 (< .0001)        | −4.9 (< .001)          |
| Claude Haiku 4.5       | 75.6%     | 57.5%       | 60.4%      | +18.1 (< .0001)              | +15.2 (< .0001)        | +2.9 (.031)            |
| GPT-5 mini             | 69.2%     | 51.5%       | 51.0%      | +17.7 (< .0001)              | +18.2 (< .0001)        | −0.5 (.766)            |
| Llama-SEA-LION-v3-70B  | 73.0%     | 57.9%       | 55.3%      | +15.1 (< .0001)              | +17.7 (< .0001)        | −2.6 (.041)            |
| Qwen-SEA-LION-v4-32B   | 71.4%     | 60.0%       | 58.8%      | +11.4 (< .0001)              | +12.6 (< .0001)        | −1.2 (.359)            |
| ThaiLLM-Qwen3-8B-Think | 64.4%     | 49.1%       | 46.2%      | +15.4 (< .0001)              | +18.3 (< .0001)        | −2.9 (.091)            |
| Typhoon v2.5           | 66.2%     | 53.1%       | 54.9%      | +13.1 (< .0001)              | +11.3 (< .0001)        | +1.8 (.179)            |
| GPT-OSS-120B           | 54.8%     | 45.6%       | 45.8%      | +9.2 (< .0001)               | +9.0 (< .0001)         | +0.2 (.906)            |
| Grok 4.1 Fast          | —         | —           | —          | <i>excluded (deprecated)</i> |                        |                        |

Table 8: Reading comprehension context-probing evaluation across 9 models on the full RC subset ( $N = 1,000$  matched items per model; 10,000 paired bootstrap re-samples).

why we report tiers rather than a strict 1–10 ranking.

| Model A      | Model B      | $\Delta$ | $p$   |      |
|--------------|--------------|----------|-------|------|
| Kimi k2.5    | GLM-5        | +1.35    | .165  | n.s. |
| GLM-5        | Claude 4.5   | +9.22    | < .05 | sep. |
| Claude 4.5   | Grok 4.1     | +2.32    | < .05 | sep. |
| Grok 4.1     | GPT-5 mini   | +1.80    | .083  | n.s. |
| GPT-5 mini   | Llama-SL-70B | +0.02    | .999  | n.s. |
| Llama-SL-70B | Qwen-SL-32B  | +1.10    | .307  | n.s. |
| Qwen-SL-32B  | ThaiLLM-8B   | +0.38    | .749  | n.s. |
| ThaiLLM-8B   | Typhoon v2.5 | +6.72    | < .05 | sep. |
| Typhoon v2.5 | GPT-OSS      | +4.75    | < .05 | sep. |

Table 9: Adjacent-pair comparisons underlying the tier assignment in Table 2.  $\Delta$  is the accuracy difference in percentage points; “sep.” denotes separable at  $p < 0.05$ . Item-level paired bootstrap, 10,000 resamples. SL abbreviates SEA-LION.