LaoBench: A Large-Scale Multidimensional Lao Benchmark for Large Language Models

Anonymous ACL submission

Abstract

The rapid advancement of large language models (LLMs) has not been matched by their evaluation in low-resource languages, especially Southeast Asian languages like Lao. To fill this gap, we introduce **LaoBench**, the first large-scale, high-quality, and multidimensional benchmark for assessing LLM language understanding and reasoning in Lao. LaoBench contains **17,000+** expert-curated samples across three dimensions: culturally grounded knowledge application, curriculum-aligned K12 education, and bilingual translation among Lao, Chinese, and English. It includes open-source and held-out subsets, where the held-out portion enables secure black-box evaluation via a controlled service to improve fairness and data security. We construct LaoBench with a hybrid pipeline that combines expert authoring with agent-assisted verification, ensuring linguistic accuracy, cultural relevance, and educational validity. We evaluate diverse state-of-the-art open-source and closed-source LLMs, and find that even strong multilingual models lag behind human experts, particularly in culturally grounded reasoning and translation fidelity. We hope LaoBench will catalyze research on Lao and other underrepresented Southeast Asian languages for more inclusive multilingual evaluation.

1 Introduction

Large language models (LLMs) have achieved strong performance on reasoning, dialogue, and translation (OpenAI et al., 2024), yet evaluation remains heavily skewed toward high-resource languages. This gap is particularly severe in Southeast Asia, where linguistic diversity is high but benchmark coverage is limited (Goyal et al., 2021; Adelani et al., 2021). While evaluation resources exist for Thai, Vietnamese, and Indonesian (Yang et al., 2022; Xu et al., 2025; Zhang et al., 2025b; Raja and Vats, 2025), Lao is still largely absent from large-scale multilingual benchmarks (Table 1).

Existing SEA-focused benchmarks also have limitations beyond language coverage. Many emphasize high-level multilingual reasoning or rely on translation from English, which often misses Lao-specific cultural grounding and linguistic properties. Lao's *scriptio continua* writing system introduces tokenization ambiguity and can distort generation and translation metrics. Curriculum-aligned education evaluation is under-explored, despite being central to native proficiency. Finally, publicly released benchmarks are increasingly vulnerable to contamination and leaderboard overfitting, yet Lao lacks a held-out black-box evaluation service for fair and sustainable comparison. These gaps motivate LaoBench as a native, multidimensional, and contamination-resistant evaluation suite for Lao.

Lao is the official language of Laos and is used by millions of speakers, yet it remains low-resource in NLP due to limited digitized corpora and scarce labeled data. Beyond data scarcity, Lao poses challenges for LLMs, including ambiguous tokenization under *scriptio continua*, a complex tonal system, and frequent Pali–Sanskrit loanwords. Existing Lao NLP resources are mostly task-specific, such as morphological analysis (Eskander et al., 2019) and limited translation datasets (Haulai and Hussain, 2023; Geigle et al., 2024), which are insufficient for evaluating general-purpose LLMs. Secure evaluation is also increasingly important as public benchmarks face leakage and overfitting risks, motivating held-out test sets and black-box services (Deng et al., 2024).

To address these gaps, we introduce **LaoBench**, the first large-scale, high-quality, and multidimensional benchmark for evaluating LLMs in Lao. LaoBench contains **17,000+** instances spanning three dimensions: (1) *Knowledge Application* grounded in Lao society, culture, politics, history, and science; (2) *K12 Foundational Education* aligned with Lao's national curriculum; and (3) *Bilingual Translation* among Lao, Chinese, and

Benchmark	Year	SEA Focus	Lao	Native	Knowledge	K12/Exam	Translation	Open Set	Held-out/Black-box
SEA-Region Benchmarks (Holistic / Cultural / Applications)									
SeaEval (Wang et al., 2024)	2024	✓	†	†	✓	†	†	✓	✓
SEA-HELM (Susanto et al., 2025)	2025	✓	†	†	✓	†	†	✓	✓
SeaExam & SeaBench (Liu et al., 2025)	2025	✓	†	✓	✓	✓	†	✓	†
Language-Specific Benchmarks (Localized / Monolingual)									
VMLU (Bui et al., 2025)	2025	†	†	✓	✓	†	†	✓	†
LORAXBENCH (Aji and Cohn, 2025)	2025	✓	†	✓	✓	†	✓	✓	†
Broader Low-Resource / Multi-Script Benchmarks (Context)									
M3Exam (Zhang et al., 2023)	2023	✓	†	†	✓	✓	†	✓	†
CIF-Bench (Li et al., 2024)	2024	†	†	✓	†	†	†	✓	†
MiLiC-Eval (Zhang et al., 2025a)	2025	†	†	✓	✓	†	†	✓	†
This Work									
LaoBench (ours)	2026	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison between LaoBench and recent SEA-focused or low-resource evaluation benchmarks. **Native** indicates whether the benchmark is originally constructed in the target language(s) rather than translated from English. **Open Set** indicates whether a publicly released evaluation set is available for reproducible research. **Held-out / Black-box** indicates whether the benchmark provides hidden test evaluation (e.g., held-out test sets or official black-box services) to mitigate contamination and leaderboard overfitting.

English.

LaoBench includes three subsets: **Lao-7k** (7,000 open-source MCQs), **Lao-10k** (10,000+ held-out MCQs for black-box evaluation), and **Lao-500** (500 open-ended prompts selected using a BenchBuilder-inspired pipeline (Li et al., 2025)). We construct LaoBench with a rigorous hybrid pipeline combining expert authoring and agent-assisted verification (e.g., duplicate detection, semantic consistency checks, context-independence filtering, and sensitivity screening).

We benchmark diverse state-of-the-art open-source and closed-source LLMs on LaoBench and find a substantial gap to human experts, especially in culturally grounded reasoning and translation fidelity. LaoBench provides a standardized and secure evaluation suite to support reliable comparison and future research on Lao and other underrepresented Southeast Asian languages.

In summary, our contributions are:

- **First multidimensional Lao benchmark.** We introduce **LaoBench**, the first large-scale benchmark for evaluating LLMs in Lao across (i) culturally grounded knowledge application, (ii) curriculum-aligned K12 education, and (iii) trilingual translation (Lao–Chinese–English). LaoBench contains 17,000+ expert-curated instances.
- **Contamination-resistant held-out evaluation.** We construct a held-out subset of 10,000+ MCQs and design a black-box evaluation protocol to mitigate test leakage and leaderboard overfitting. We will deploy the protocol as an online evaluation service upon publication.
- **Rigorous hybrid construction pipeline.** We develop a scalable pipeline combining expert authoring with agent-assisted verification, including duplicate detection, semantic consistency checks, context-independence filtering, and sensitivity screening.
- **Comprehensive benchmarking and key findings.** We evaluate diverse state-of-the-art open-source and closed-source LLMs on LaoBench and reveal persistent gaps to human experts, especially in culturally grounded reasoning and high-fidelity translation.

2 LaoBench

LaoBench is a large-scale, multidimensional benchmark designed to systematically evaluate large language models (LLMs) in Lao, a low-resource Southeast Asian language. LaoBench provides comprehensive evaluation coverage across three core dimensions: **Knowledge Application**, **K12 Foundational Education**, and **Bilingual Translation** among Lao, Chinese, and English.

2.1 Dataset Overview and Subsets

LaoBench contains more than **17,000** carefully curated instances. To support both transparent research usage and secure benchmarking, the benchmark is organized into three complementary subsets, as summarized in Table 2.

Lao-7k (open-source multiple-choice). Lao-7k contains 7,000 expert-written multiple-choice questions released publicly to enable reproducible

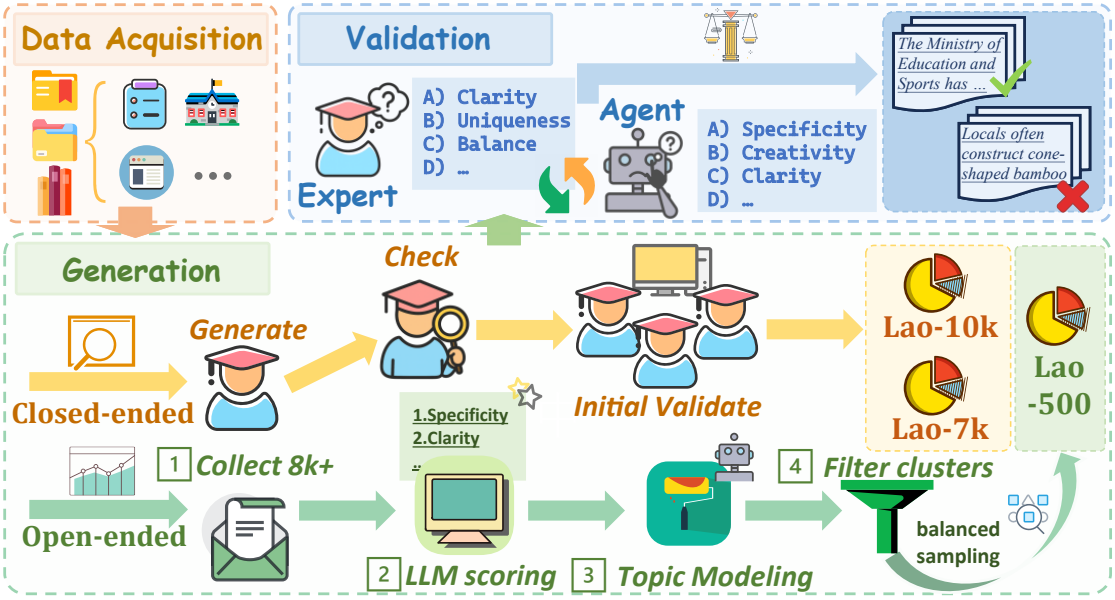

Figure 1: Overview of the LaoBench construction pipeline. We collect raw materials from authoritative Lao sources, construct closed-form multiple-choice questions and open-ended prompts using different strategies, and perform multi-stage validation with expert review and agent-assisted verification to ensure linguistic correctness, cultural relevance, and educational validity.

Subset	Size	Type
Lao-7k	7,000	Multiple-choice, open-source
Lao-10k	10,000+	Multiple-choice, closed-source
Lao-500	500	Open-ended prompts, open-source

Table 2: Summary of LaoBench dataset subsets.

benchmarking. These questions span all three evaluation dimensions and are designed to assess both factual understanding and reasoning ability in Lao.

Lao-10k (closed-source multiple-choice). Lao-10k contains over 10,000 additional multiple-choice questions reserved for black-box evaluation via a controlled black-box service. We keep this subset hidden and return only aggregated scores to mitigate test leakage and leaderboard overfitting. The full black-box evaluation protocol is described in Appendix H.

Lao-500 (open-source open-ended). Lao-500 is a set of 500 open-ended prompts intended for evaluating long-form generation and open-domain reasoning in Lao. These prompts are automatically selected from a large candidate pool using a pipeline inspired by BenchBuilder (Li et al., 2025), which supports continuous benchmark expansion while maintaining diversity and quality.

Design rationale of the three subsets. We design LaoBench as a combination of open-source and closed-source subsets to balance transparency, fairness, and long-term benchmark reliability.

2.2 Task Categories and Formats

Each LaoBench instance belongs to one of the following three categories, designed to capture complementary aspects of Lao language competence:

Knowledge Application. This category evaluates domain knowledge grounded in Lao society and culture, covering subdomains. Questions are designed to require contextual reasoning and culturally specific knowledge rather than simple pattern matching.

K12 Foundational Education. This category aligns with Lao's national K12 curriculum and evaluates foundational knowledge and reasoning skills. Items emphasize educational validity and reflect real classroom-style knowledge requirements.

Bilingual Translation. This category evaluates translation capabilities among Lao, Chinese, and English in practical domains. Each instance contains a source sentence and a professionally written reference translation. We report corpus-level BLEU under a standardized SacreBLEU configuration with Lao-aware tokenization, and provide full evaluation details in Appendix C.

2.3 Construction Pipeline

LaoBench is constructed through a three-stage pipeline consisting of **raw data acquisition**,

Figure 2: Example cases from LaoBench, illustrating the three task types: Knowledge Application, K12 Education, and Bilingual Translation.

dataset construction, and **validation and quality assurance**, as illustrated in Figure 1. The pipeline integrates expert human curation with agent-assisted verification to ensure both high quality and scalability.

Raw Data Acquisition. To ensure that LaoBench reflects realistic language use in Laos, we collect materials from diverse authoritative sources, including K12 textbooks and curriculum guidelines, government and legal documents, encyclopedic and educational publications, as well as culturally grounded articles and local knowledge resources. These sources cover both formal and informal registers, enabling the benchmark to evaluate not only factual recall but also culturally grounded reasoning and practical language understanding. By grounding questions in authentic Lao contexts, LaoBench reduces the risk of constructing synthetic or overly translation-based evaluation data, which is a common limitation in low-resource benchmarks.

Dataset Construction. We adopt two complementary construction strategies depending on the data format. For Lao-7k and Lao-10k, expert linguists and domain specialists create multiple-choice questions by selecting key knowledge points, writing question stems in Lao, and designing plausible distractors with one correct option. Difficulty is calibrated through iterative refinement to reduce ambiguity and ensure meaningful reasoning requirements. For Lao-500, we begin with a large candidate prompt pool derived from the same sources and apply an automated selection procedure inspired by BenchBuilder (Li et al., 2025). Candidate prompts are scored by an LLM annotator on quality dimensions such as specificity, clarity, domain depth, and creativity. We further apply topic-based clustering to promote diversity, discard low-quality or redundant clusters, and sample evenly from the remaining clusters to form a balanced set of 500 open-ended prompts.

Validation and Quality Assurance. All items undergo multi-stage validation combining expert review and agent-assisted verification. Human experts verify factual correctness, linguistic fluency, cultural appropriateness, and educational validity. For multiple-choice questions, experts also assess distractor plausibility and remove ambiguous or underspecified items. For translation instances, experts verify semantic alignment between source and reference translations and correct unnatural phrasing. Automated agents further support quality control by detecting duplicates and near-duplicates, checking semantic consistency, and screening potentially sensitive or harmful content. Items failing any validation checks are revised or removed through iterative refinement.

2.4 Data Statistics

LaoBench contains more than 17,000 instances across three primary categories: Knowledge Application, K12 Foundational Education, and Bilingual Translation. Each category is further organized into subdomains. The overall distribution of samples is shown in Figure 3. Further details are shown in Appendix A.

Model	K12 Accuracy (%)					Translation BLEU				Knowledge Application Accuracy (%)			
	Nat. Sci.	Soc. Sci.	Think. & Phil.	Hum. & Arts	Health & Env.	Soc. & Law	Cult. & Hist.	Inter. Aff.	Env. & Dev.	Pol. & Law	Soc. & Cult.	Hist. & Dev.	Nat. Sci.
Blind Evaluation													
Random Choice	25.00	25.00	25.00	25.00	25.00	–	–	–	–	25.00	25.00	25.00	25.00
Open-Source Models													
Qwen3-Next-80B-A3B-Instruct	76.15	69.94	68.32	91.15	93.43	16.03	23.94	38.48	22.95	68.78	67.67	60.00	55.73
Qwen3-235B-A22B	82.75	76.48	77.09	93.11	97.65	20.39	27.15	36.64	28.01	71.51	71.55	62.35	57.64
Qwen3-235B-A22B-Instruct-2507	86.45	76.36	74.74	94.75	98.59	21.81	29.02	41.47	29.14	74.82	73.85	59.53	61.46
DeepSeek-V3.2-Exp (Thinking)	84.69	79.19	75.39	91.48	94.84	20.57	28.29	34.96	27.29	70.94	72.32	63.76	69.43
DeepSeek-V3.2-Exp (Non-Thinking)	79.96	72.48	68.06	91.80	95.31	22.52	29.77	37.16	29.49	67.63	70.32	64.47	58.28
Ministral-8B-Instruct-2410	23.52	30.18	25.79	34.75	27.23	0.83	2.09	8.61	2.09	22.16	26.53	24.00	23.89
Ling-mini-2.0	35.16	35.15	34.55	39.34	40.38	0.69	2.56	9.82	2.11	27.91	32.33	31.76	28.98
Closed-Source Models													
GPT-5-High	90.03	82.91	80.76	95.08	99.53	20.96	28.52	38.59	26.91	79.42	77.74	66.59	75.80
Qwen3-Max	87.35	77.82	76.31	94.75	97.65	21.70	30.07	39.71	28.87	75.11	76.33	61.41	63.38
Qwen3-Plus (Non-Thinking)	86.00	76.61	74.08	95.41	99.06	21.69	28.83	40.95	28.91	75.54	74.73	59.29	60.51
Qwen3-Plus (Thinking)	84.88	76.85	74.48	95.41	99.06	20.32	28.42	39.46	28.25	75.83	73.85	60.47	60.51
Gemini-2.5-Pro	88.69	82.79	82.20	95.08	99.06	26.22	34.31	40.71	33.68	77.55	79.51	67.29	70.38
Claude-Sonnet-4.5-20250929-thinking	89.36	80.85	79.71	95.41	98.59	22.76	29.96	37.50	28.41	77.12	77.21	63.29	69.11
Claude-Opus-4.1-20250805	89.03	80.00	78.27	95.74	96.71	24.78	32.08	38.52	32.38	77.84	78.80	68.47	68.47
Human Evaluation													
Human Experts	98.34	98.14	97.29	98.93	99.92	–	–	–	–	99.21	98.78	98.97	97.98

Table 3: Detailed performance of evaluated models on Lao-7k. We report accuracy (%) for K12 Education and Knowledge Application, and BLEU for Translation. Highlighted cells indicate the best result in each column. Dimension-level averages for the three core dimensions are reported in Figure 4.

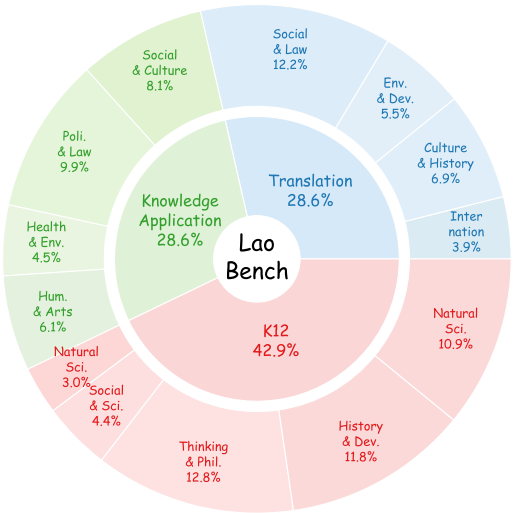

Figure 3: Distribution of Lao-7k samples across the three main categories—Knowledge Application, K12 Education, and Translation—and their subdomains.

3 Experiment

We evaluate LLMs on LaoBench using two complementary protocols. First, we perform closed-form multiple-choice evaluation on **Lao-7k**, enabling transparent and reproducible benchmarking with objective metrics. Second, we conduct open-ended generation evaluation on **Lao-500** using an Arena-style pairwise comparison framework. Together, these evaluations provide a holistic view of LLM capabilities in Lao.

3.1 Experimental Setup

Models. We select representative SOTA closed-source models and diverse open-source families covering different scales and instruction styles Open-source models include Qwen3-Next-80B-A3B-Instruct (Yang et al., 2025), Qwen3-235B-A22B (Team, 2025), Qwen3-235B-A22B-Instruct-2507, DeepSeek-V3.2-Exp (DeepSeek-AI, 2025) (Thinking and Non-Thinking), Ministral-8B-Instruct-2410 (The Mistral AI Team, 2024), and Ling-mini-2.0 (inclusionAI, 2025). Closed-source models include GPT-5-High (OpenAI, 2025), Qwen3-Max, Qwen3-Plus (Thinking/Non-Thinking), Gemini-2.5-Pro (Comanici et al., 2025), Claude-Sonnet-4.5-20250929-thinking (Claude), and Claude-Opus-4.1-20250805. All API evaluations were conducted during Oct.–Dec. 2025 with deterministic decoding (temperature $= 0$ where supported).

Prompting and inference. All models are evaluated in a **zero-shot** setting without fine-tuning on LaoBench. For multiple-choice questions, we present the question stem and four options in Lao and require models to output exactly one option. For models that support explicit reasoning, we additionally evaluate chain-of-thought (CoT) style prompting (denoted as *Thinking*) and compare against direct-answer prompting (*Non-Thinking*).

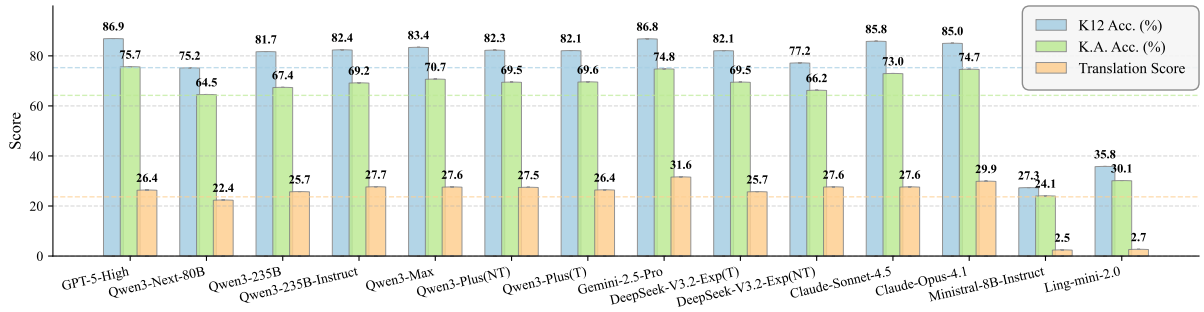

Figure 4: Overall **dimension-level averages** (*K12 Avg / Translation Avg / Knowledge Avg*) of evaluated models on Lao-7k. Each score is averaged over its corresponding subdomains in Table 3.

For translation evaluation, we use a fixed SacreBLEU configuration with Lao-aware tokenization and additionally report chrF++(Popović, 2017) in Appendix C.

Output normalization. For multiple-choice evaluation, we apply a unified post-processing rule that maps model outputs to a single option label (A/B/C/D). If a model produces multiple labels, non-option text, or an unparseable answer, it is counted as incorrect.

3.2 Closed-Form Evaluation on Lao-7k

We first conduct closed-form evaluation on **Lao-7k**. This subset enables transparent and reproducible benchmarking with objective metrics.

Metrics. For K12 Education and Knowledge Application multiple-choice questions, we report **Accuracy**. For Translation tasks, we compute corpus-level **BLEU** against expert-written references under a standardized SacreBLEU configuration. Since Lao is written in a scriptio continua style without explicit word boundaries, BLEU can be sensitive to segmentation; we therefore apply Lao-aware tokenization using LaoNLP before scoring, and additionally report chrF++ in Appendix C.

Overall results. Table 3 reports detailed results across subdomains for all evaluated models. A clear gap emerges between open-source and closed-source models: closed-source systems consistently dominate across all dimensions, while open-source models show larger variance. Among closed-source models, GPT-5-High achieves the strongest overall accuracy on K12 and Knowledge Application, while Gemini-2.5-Pro leads in Translation BLEU, indicating stronger multilingual generation fidelity. Among open-source models, Qwen3-235B-A22B and DeepSeek-V3.2 are the most competitive, whereas smaller models significantly lag behind in all categories.

Fine-grained analysis by subdomain. Figure 4 summarizes model performance aggregated by the three core dimensions. Across nearly all systems, **K12 Education** is generally easier than **Knowledge Application**. For instance, many strong models exceed 90% accuracy on K12 subdomains such as Health & Environment and Humanities & Arts (Table 3), suggesting that structured curriculum-aligned content is relatively accessible to multilingual LLMs. In contrast, Knowledge Application subdomains yield notably lower accuracy, reflecting the need for culturally grounded and domain-specific reasoning. Even GPT-5-High shows a substantial drop from K12 to Knowledge Application accuracy (Table 3), highlighting the intrinsic difficulty of culturally grounded Lao reasoning.

Translation performance remains limited. Translation BLEU scores are modest for all models. Even the strongest closed-source systems only achieve mid-30s BLEU in the best subdomains, indicating persistent challenges in translation fidelity and fluency. We also observe topic-dependent differences: Culture & History and Social & Law tend to be harder because they contain culturally specific expressions and formal terminology that require precise lexical choice. These results suggest that translation involving Lao remains challenging even for state-of-the-art multilingual models.

Effect of Chain-of-Thought prompting. Figure 5 highlights the impact of Chain-of-Thought (CoT) prompting by comparing Thinking and Non-Thinking variants. We observe consistent improvements from CoT prompting, especially on subdomains requiring multi-step reasoning and culturally

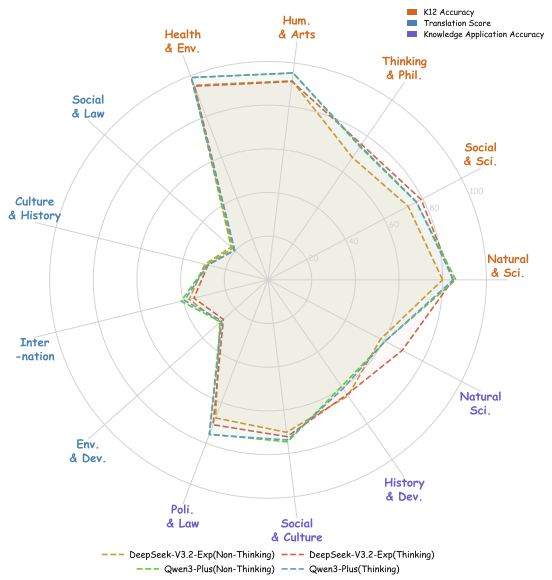

Figure 5: Radar chart comparing performance of models with and without Chain-of-Thought (CoT) prompting on Lao-7k. Models with CoT (Thinking) consistently outperform their Non-Thinking counterparts, especially in complex reasoning subdomains such as Thinking & Philosophy and Knowledge Application.

grounded Knowledge Application questions. However, improvements are smaller on more factual or formulaic K12 subdomains, suggesting that CoT primarily benefits complex reasoning rather than straightforward recall.

Gap to human performance. Human experts achieve near-perfect accuracy across all K12 and Knowledge Application subdomains establishing a strong upper bound. The persistent gap between human performance and state-of-the-art models, especially on Knowledge Application and Translation, indicates that LaoBench remains challenging and provides meaningful headroom for future research.

Overall, LaoBench reveals a consistent performance drop from curriculum-aligned knowledge to culturally grounded reasoning and bilingual translation, indicating that Lao remains a challenging low-resource setting for modern LLMs.

3.3 Open-Ended Arena-Style Evaluation on Lao-500

Closed-form multiple-choice evaluation cannot fully reflect user-facing generation quality such as coherence, explanation quality, and long-form instruction following. We therefore evaluate open-ended prompts in Lao-500 using an Arena-style pairwise protocol, which compares model outputs by preference rather than absolute scoring.

Baseline and pairwise comparison. For each prompt x_i, we generate one response from a candidate model M, denoted as y_i^M, and one response from a fixed baseline model B (GPT-5-High), denoted as y_i^B. A judge model J is asked to decide which response is better with respect to correctness, completeness, reasoning quality, clarity, and Lao fluency. To mitigate potential position bias, we randomize the response ordering and repeat evaluation with swapped positions, then average the two outcomes for each prompt.

Self-preference and judge bias control. To reduce potential self-preference effects, we avoid using a judge model to evaluate comparisons where it is also a candidate model whenever applicable, and we additionally report judge-specific scores and cross-judge gaps (Appendix F). This design helps mitigate correlated preferences between judge and candidate model families.

Win-rate score. Let $w_i^J(M) \in [0, 1]$ denote the (bias-corrected) win signal of model M on prompt i under judge J, where ties are assigned a fractional win. In practice we treat ties as half-wins. We compute the judge-specific win rate of model M against baseline B as:

$$S_J(M) = \frac{1}{N} \sum_{i=1}^{N} w_i^J(M). \tag{1}$$

This score is reported as a percentage.

Multi-judge aggregation. We note that judge models may exhibit systematic preferences that correlate with model families. We therefore employ two judges, Gemini-2.5-Pro and Qwen3-Max, and define the overall score as the average across judges:

$$S(M) = \frac{1}{|\mathcal{J}|} \sum_{J \in \mathcal{J}} S_J(M), \tag{2}$$

where $\mathcal{J}$ denotes the set of judges. To mitigate single-judge bias, we aggregate scores across two independent judges and report judge-specific results in Appendix F.

Bootstrap confidence intervals. To quantify uncertainty under a limited prompt budget, we estimate confidence intervals via bootstrap resampling over prompts. For each bootstrap trial t, we sample a multiset of prompts $\mathcal{I}^{(t)}$ with replacement and recompute the score:

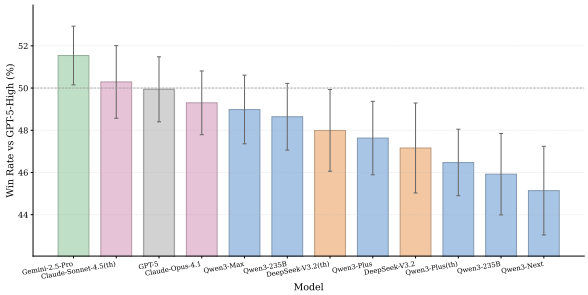

Figure 6: Arena-style open-ended evaluation results on Lao-500. Scores are win rates against GPT-5-High with 95% bootstrap confidence intervals (CI), aggregated over two judges (Gemini-2.5-Pro and Qwen3-Max).

$$S^{(t)}(M) = \frac{1}{|\mathcal{J}|} \sum_{J \in \mathcal{J}} \left(\frac{1}{|\mathcal{I}^{(t)}|} \sum_{i \in \mathcal{I}^{(t)}} w_i^J(M) \right). \tag{3}$$

The reported 95% confidence interval is obtained from the percentile interval of the bootstrap distribution $S^{(t)}(M)$.

Main results. Figure 6 shows the aggregated Lao-500 win-rate results (Avg Score) and 95% bootstrap confidence intervals. To make judge sensitivity explicit, we report judge-specific win rates, confidence intervals, and cross-judge gaps in Appendix F. We further report judge agreement statistics and a human sanity-check subset in Appendix D to quantify judge sensitivity and potential bias.

Analysis of Lao-500 results. We highlight three key observations.

(1) Rankings differ from closed-form evaluation. Model rankings under Lao-500 are not strictly identical to Lao-7k accuracy rankings, indicating that open-ended evaluation captures additional aspects such as explanation quality and coherence beyond option selection.

(2) Confidence intervals enable uncertainty-aware comparison. Models with close scores often exhibit overlapping confidence intervals, suggesting their differences may not be statistically significant under the sampled prompt set. Non-overlapping intervals indicate robust gaps.

(3) Judge sensitivity is non-negligible. While the two judges agree on major trends, we observe systematic shifts for certain model families. For example, Qwen3-Max judge tends to assign higher win rates to Qwen-family models compared to Gemini-2.5-Pro, indicating judge-dependent preference patterns.

4 Related Work

Benchmarks for low-resource languages. Recent efforts have expanded evaluation resources for underrepresented languages, including African and Indic language benchmarks (Adelani et al., 2021; Kakwani et al., 2020). For Southeast Asia, benchmarks have been proposed for Thai, Vietnamese, and Indonesian (Yang et al., 2022; Xu et al., 2025; Zhang et al., 2025b; Raja and Vats, 2025). However, Lao remains largely absent from large-scale multidimensional evaluation suites, limiting systematic assessment of modern multilingual LLMs in this language.

Multilingual evaluation and translation benchmarks. Multilingual benchmarks such as FLORES (Goyal et al., 2021) and large-scale MT evaluation datasets provide broad language coverage, but they often focus on translation or surface-level understanding and typically exclude Lao or provide limited Lao-specific evaluation. Moreover, multilingual benchmarks rarely include culturally grounded knowledge reasoning or education-aligned tasks that reflect real-world language use in Laos.

Dataset construction and quality assurance. High-quality benchmarks increasingly combine human expertise with automated verification to ensure scalability and reliability (Hendrycks et al., 2021; Luo et al., 2024; Wang et al., 2025; Kang et al., 2025). In particular, recent work highlights the importance of held-out test sets and black-box evaluation to mitigate contamination (Deng et al., 2024). LaoBench follows this trend by integrating expert curation with agent-assisted validation and providing a large closed-source subset.

5 Conclusion

We present **LaoBench**, the first large-scale multidimensional benchmark for evaluating LLMs in Lao, covering culturally grounded knowledge, K12 education, and bilingual translation. LaoBench includes both open-source subsets for reproducible research and a large closed-source subset for secure black-box evaluation. Experiments on Lao-7k and Lao-500 reveal that even strong multilingual models still lag behind human experts, especially on knowledge-intensive reasoning and high-fidelity translation. We hope LaoBench will facilitate more reliable evaluation and drive progress for Lao and other underrepresented Southeast Asian languages.

Limitations

LaoBench has several limitations. First, a large portion of the benchmark uses multiple-choice questions, which may not fully reflect open-ended reasoning ability and can allow partial gains from test-taking strategies. Second, translation evaluation relies primarily on reference-based metrics such as BLEU, which can under-estimate valid paraphrastic translations and can be sensitive to tokenization for Lao script. Third, Arena-style evaluation depends on LLM-based judges and a fixed baseline model, which may introduce preference bias or baseline anchoring effects, although we mitigate these via multi-judge aggregation and sanity checks. In particular, a judge model may favor outputs that resemble its own instruction style or that come from the same model family. We partially mitigate this by using two independent judges and reporting judge-specific scores and cross-judge gaps in the appendix. Finally, the online black-box evaluation service for Lao-10k is under active development and will be released upon publication.

Ethics Statement

Data sourcing and licensing. We collect materials from publicly accessible authoritative sources such as textbooks, curriculum guidelines, and government documents. For open-source releases, we only distribute derived benchmark items and do not redistribute copyrighted raw texts. We comply with institutional policies and ensure that released content does not contain identifiable private information.

Annotator welfare. All human annotators and reviewers are compensated at competitive local market rates. They receive clear task instructions, are informed of potential sensitive topics, and may opt out of sensitive content annotation.

Sensitive content and harm mitigation. Since Knowledge Application may involve political, legal, and cultural topics, we apply multi-stage sensitivity screening using both agent-assisted filtering and expert review. We remove or revise items that may promote hate, discrimination, misinformation, or political persuasion, and we avoid collecting or releasing personal data.

Intended use and reporting mechanism. LaoBench is intended for research and evaluation of language technologies for Lao. We discourage use for surveillance, targeted manipulation, or misinformation generation.

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, and 42 others. 2021. Masakhaner: Named entity recognition for african languages. *Preprint*, arXiv:2103.11811.

Alham Fikri Aji and Trevor Cohn. 2025. LORAXBENCH: A multitask, multilingual benchmark suite for 20 Indonesian languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17432–17457, Suzhou, China. Association for Computational Linguistics.

Cuc Thi Bui, Nguyen Truong Son, Truong Van Trang, Lam Viet Phung, Pham Nhut Huy, Hoang Anh Le, Quoc Huu Van, Phong Nguyen-Thuan Do, Van Le Tran Truc, Duc Thanh Chau, and Le-Minh Nguyen. 2025. VMLU benchmarks: A comprehensive benchmark toolkit for Vietnamese LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11495–11515, Vienna, Austria. Association for Computational Linguistics.

Claude. The claude 3 model family: Opus, sonnet, haiku.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

DeepSeek-AI. 2025. Deepseek-v3.2-exp: Boosting long-context efficiency with deepseek sparse attention.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2024. Investigating data contamination in modern benchmarks for large language models. 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 8706–8719, Mexico City, Mexico. Association for Computational Linguistics.

Ramy Eskander, Judith Klavans, and Smaranda Muresan. 2019. Unsupervised morphological segmentation for low-resource polysynthetic languages. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 189–195, Florence, Italy. Association for Computational Linguistics.

Gregor Geigle, Radu Timofte, and Goran Glavaš. 2024. Babel-imagenet: Massively multilingual evaluation of vision-and-language representations. *Preprint*, arXiv:2306.08658.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

Thangkhanhau Haulai and Jamal Hussain. 2023. Construction of mizo: English parallel corpus for machine translation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(8).

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *Preprint*, arXiv:2103.03874.

inclusionAI. 2025. Ling-mini-2.0. `https://huggingface.co/inclusionAI/Ling-mini-2.0`.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.

Zhaolu Kang, Junhao Gong, Jiaxu Yan, Wanke Xia, Yian Wang, Ziwen Wang, Huaxuan Ding, Zhuo Cheng, Wenhao Cao, Zhiyuan Feng, Siqi He, Shannan Yan, Junzhe Chen, Xiaomin He, Chaoya Jiang, Wei Ye, Kaidong Yu, and Xuelong Li. 2025. Hss-bench: Benchmarking humanities and social sciences ability for multimodal large language models. *Preprint*, arXiv:2506.03922.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. 2025. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. In *Forty-second International Conference on Machine Learning*.

Yizhi Li, Ge Zhang, Xingwei Qu, Jiali Li, Zhaoqun Li, Noah Wang, Hao Li, Ruibin Yuan, Yinghao Ma, Kai Zhang, Wangchunshu Zhou, Yiming Liang, Lei Zhang, Lei Ma, Jiajun Zhang, Zuowen Li, Wenhao Huang, Chenghua Lin, and Jie Fu. 2024. CIF-bench: A Chinese instruction-following benchmark for evaluating the generalizability of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12431–12446, Bangkok, Thailand. Association for Computational Linguistics.

Chaoqun Liu, Wenxuan Zhang, Jiahao Ying, Mahani Aljunied, Anh Tuan Luu, and Lidong Bing. 2025. SeaExam and SeaBench: Benchmarking LLMs with local multilingual questions in Southeast Asia. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6119–6136, Albuquerque, New Mexico. Association for Computational Linguistics.

Fuwen Luo, Chi Chen, Zihao Wan, Zhaolu Kang, Qidong Yan, Yingjie Li, Xiaolong Wang, Siyu Wang, Ziyue Wang, Xiaoyue Mi, Peng Li, Ning Ma, Maosong Sun, and Yang Liu. 2024. CODIS: Benchmarking context-dependent visual comprehension for multimodal large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10639–10659, Bangkok, Thailand. Association for Computational Linguistics.

OpenAI. 2025. Gpt-5. `https://openai.com/index/introducing-gpt-5/`.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Rahul Raja and Arpita Vats. 2025. Parallel corpora for machine translation in low-resource indic languages: A comprehensive review. *Preprint*, arXiv:2503.04797.

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: Southeast 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.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

The Mistral AI Team. 2024. Ministral-8B-Instruct-2410. `https://huggingface.co/mistralai/Ministral-8B-Instruct-2410`.

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.

Xiaolong Wang, Zhaolu Kang, Wangyuxuan Zhai, Xinyue Lou, Yunghwei Lai, Ziyue Wang, Yawen Wang, Kaiyu Huang, Yile Wang, Peng Li, and Yang Liu. 2025. Mucar: Benchmarking multilingual cross-modal ambiguity resolution for multimodal large language models. *Preprint*, arXiv:2506.17046.

Guixian Xu, Zeli Su, Ziyin Zhang, Jianing Liu, XU Han, Ting Zhang, and Yushuang Dong. 2025. Cmhg: A dataset and benchmark for headline generation of minority languages in china. *Preprint*, arXiv:2509.09990.

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, Junyang Lin, Kai Dang, Kexin Yang, Le Yu, Mei Li, Minmin Sun, Qin Zhu, Rui Men, Tao He, and 9 others. 2025. Qwen2.5-1m technical report. *arXiv preprint arXiv:2501.15383*.

Ziqing Yang, Zihang Xu, Yiming Cui, Baoxin Wang, Min Lin, Dayong Wu, and Zhigang Chen. 2022. CINO: A Chinese minority pre-trained language model. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3937–3949, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Chen Zhang, Mingxu Tao, Zhiyuan Liao, and Yansong Feng. 2025a. MiLiC-eval: Benchmarking multilingual LLMs for China's minority languages. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11086–11102, Vienna, Austria. Association for Computational Linguistics.

Chen Zhang, Mingxu Tao, Zhiyuan Liao, and Yansong Feng. 2025b. Milic-eval: Benchmarking multilingual llms for china's minority languages. *arXiv preprint arXiv:2503.01150*.

Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Preprint*, arXiv:2306.05179.

This appendix provides detailed dataset construction, validation protocols, evaluation configurations, and additional analyses that support the main paper.

A Detailed Dataset Statistics

A.1 Exact subset sizes

LaoBench consists of three subsets:

- **Lao-7k**: **7,000** open-source multiple-choice questions.
- **Lao-10k**: **10k** closed-source multiple-choice questions used for secure black-box evaluation.
- **Lao-500**: **500** open-ended prompts for generation evaluation.

A.2 Category and subdomain distribution

Table 4 reports the distribution of instances across categories and subdomains. We ensure balanced coverage by enforcing minimum instance counts per subdomain during construction.

Category	Subdomain	# Instances
Knowledge Application		
History & Development		425
Politics & Law		695
Society & Culture		566
Nature & Science		314
K12 Education		
Humanities & Arts		305
Health & Environment		213
Thinking & Philosophy		764
Social Sciences		825
Natural Sciences		893
Translation		
International Affairs		276
Culture & History		483
Environment & Development		388
Society & Law		853

Table 4: Detailed distribution of LaoBench instances across categories and subdomains.

Fine-grained taxonomy. Beyond the coarse-grained subdomain categorization, we further define a three-level hierarchical taxonomy (*dimension* → *category* → *subcategory*) to ensure consistent data construction and coverage control. Table 5 provides the full taxonomy used in LaoBench, including all third-level subcategories.

B Data Construction and Validation Protocols

B.1 Data sources

We curate raw materials from authoritative Lao sources, including: (i) national K12 textbooks and

Dimension	Category	Subcategory
Knowledge Application	History and Development	Historical Figures and Events Wars and Conflicts Political Changes and Development Education and Progress
	Politics and Law	Peace Agreements and Civil War Resolution Diplomacy and International Relations Political and Administrative Systems Laws and Conventions Borders and Territorial Demarcation
	Society and Culture	Religion and Beliefs Culture and Traditions Ethnic Migration and Social Development Commemorative Days and Cultural Heritage
	Nature and Science	Geography and Environment Archaeological Discoveries and Anthropology
K12 Foundational Education	Humanities and Arts	Writing and Media Behavioral Norms and Etiquette
	Health and Environment	Health and Psychology Environmental Protection and Sustainability
	Thinking and Philosophy	Mathematics and Logical Thinking Moral Philosophy and Common Sense
	Social Sciences	History and Geography Historical Events and Culture Local Culture and History Social Sciences and Political Science Social Responsibility and Civic Awareness
	Natural Sciences	Zoology and Botany Physics and Technological Inventions Matter and Chemistry Fundamentals of Biology
Translation	International Affairs	International Relations and Treaties Impact of Border Agreements
	Culture and History	Religion and Festival Customs Lao History and Culture Archaeology and Historical Research Natural Landscapes and Cultural Heritage
	Environment and Development	Education System and Social Equity Natural Resources and Environmental Protection
	Society and Law	Civil Rights and Development History Laws, Regulations, and Education Society and Mental Health Economic Development and Legal Systems Ethics and Legal Responsibility

Table 5: Full hierarchical taxonomy of LaoBench (dimension → category → subcategory), used to guide data construction and ensure balanced coverage across fine-grained topics.

curriculum guidelines, (ii) government and legal documents, (iii) encyclopedic and educational publications, (iv) culturally grounded articles and local knowledge resources. All sources are documented and archived internally. We release metadata summaries and source-type statistics for the open-source subset.

B.2 Expert annotation roles and training

Our annotation team includes Lao linguists, education experts, and subject-matter specialists. Before annotation, we conduct a standardized training session covering: question writing standards, distractor design principles, difficulty calibration, and sensitivity guidelines.

Annotator statistics. We employ 11 Lao-native annotators and 4 expert reviewers. Each item is authored by one annotator and reviewed by at least 2 independent experts. Disagreements are resolved through adjudication by a senior reviewer.

B.3 Multiple-choice question construction guidelines

For each MCQ:

- The stem is written in natural Lao and references grounded Lao contexts.
- Four options are provided with exactly one correct answer.
- Distractors are designed to be plausible and semantically close to the correct option, avoiding trivial elimination.
- We prohibit "all of the above" and "none of the above" to reduce ambiguity.
- We remove items whose correct option depends on missing context or relies on overly specific memorization.

B.4 Translation instance construction

Each translation instance is constructed by: (i) selecting representative sentences from authoritative sources, (ii) producing a reference translation by professional translators, (iii) verifying semantic alignment and Lao fluency via expert review. We cover translation directions among Lao, Chinese, and English according to the benchmark taxonomy.

B.5 Agent-assisted verification and filtering

We employ agent-based verification to support scalability and consistency. Agents are used for:

- **Duplicate / near-duplicate detection:** lexical overlap (character n-grams) and semantic similarity (embedding retrieval).
- **Semantic consistency:** verifying that the correct option is uniquely supported by the stem and that distractors are not also correct.
- **Context independence:** removing items requiring external context not provided in the prompt.
- **Sensitivity screening:** detecting potentially harmful, private, or politically sensitive content, followed by expert review.

B.6 Inter-annotator agreement and adjudication

To estimate reliability, we sample 500 items and ask 3 annotators to independently label the correct answer. We report Fleiss' κ:

$$\kappa = \frac{p_o - p_e}{1 - p_e},$$

where p_o is observed agreement and p_e is chance agreement. Our measured agreement is $\kappa = 0.87$, indicating substantial consistency.

C Translation Evaluation Configuration

C.1 Tokenization and segmentation

We evaluate translation outputs under a standardized reference-based protocol. Since Lao is written in a *scriptio continua* style without explicit whitespace word boundaries, both BLEU and chrF++ can be sensitive to segmentation. We therefore perform Lao-aware word segmentation using LaoNLP (v0.7) on both system outputs and references prior to metric computation. For non-Lao languages (Chinese and English), we apply standard whitespace tokenization (English) and character-aware tokenization (Chinese) as provided by SacreBLEU.

C.2 BLEU

We compute corpus-level BLEU using SacreBLEU with a unified configuration across all systems. We report BLEU scores on the tokenized text, which improves stability for Lao evaluation.

C.3 ChrF++

In addition to BLEU, we report **chrF++** to capture character n-gram overlap and reduce sensitivity to tokenization. chrF++ computes a character-level F-score with word boundary awareness, which is particularly suitable for low-resource and scriptio continua languages such as Lao. We compute chrF++

Model	Soc.& Law	Cult.& Hist.	Inter.& Aff.	Env.& Dev.
GPT-5-High	50.31	59.92	66.08	57.41
Gemini-2.5-Pro	52.84	56.62	66.35	60.12
Qwen3-Max	52.38	58.04	65.21	58.93

Table 6: chrF++ scores on LaoBench translation subsets. Higher is better.

using SacreBLEU with the default setting (character n-gram order up to 6 and word n-gram order up to 2), and report corpus-level scores for each translation subdomain.

D Arena-Style Evaluation Reliability

We enforce a strict JSON-only output format for judges to ensure deterministic parsing and prevent explanation leakage.

D.1 Judge agreement

We compute rank correlation between judges (Gemini-2.5-Pro and Qwen3-Max) using Spearman's ρ and Kendall's τ. Our measured agreement is:

- Spearman ρ: 0.83
- Kendall τ: 0.71

D.2 Human sanity check

We randomly sample 50 prompts from Lao-500 and ask Lao-native human evaluators to judge pairwise outputs. We compare human preferences with LLM judges and observe 84% agreement, supporting judge reliability.

D.3 Baseline anchoring analysis

To quantify baseline effects, we repeat evaluation using an alternative baseline model (Claude-Opus-4.1) and measure ranking stability. We find that while absolute win-rates shift, relative rankings remain largely stable.

E Arena Prompt Templates and Output Parsing

To ensure transparency and reproducibility, we provide the exact prompt templates used for Arena-style pairwise evaluation on Lao-500. Each comparison consists of a user prompt and two candidate answers (one from the baseline model and one from a challenger model), evaluated by an independent judge model. We enforce a strict JSON-only output format to enable deterministic parsing and to avoid explanation leakage. To mitigate position bias, we randomize answer ordering and run each comparison twice with swapped positions, then average the outcomes.

E.1 Generation and Judge Prompt Template

For each Lao-500 prompt, we generate one response from the baseline model and one response from a challenger model using the same generation template. To reflect realistic user-facing usage in Laos and to avoid code-switching into English or Chinese, we explicitly enforce Lao-only outputs. All models are evaluated in a zero-shot setting with deterministic decoding (temperature set to 0 when supported).

Generation prompt template. Table 7 shows the prompt template used to generate model responses. The template consists of (i) a system instruction enforcing Lao-only outputs, and (ii) the user prompt drawn from Lao-500. For models that do not support system prompts, the system instruction is prepended to the user message.

Judge prompt template. Table 8 shows the judge prompt used for pairwise evaluation. Given the same user prompt, the judge receives two anonymized candidate responses (Response A and Response B) and decides which is better. The judge is instructed to evaluate correctness, completeness, reasoning quality, clarity, and Lao fluency. To mitigate position bias, we evaluate each pair twice by swapping the positions of A and B and then average the outcomes.

E.2 Tie Handling and Scoring

A tie is assigned a half-win (0.5) for each model. We report the final win rate of each challenger against the baseline across all prompts.

E.3 Position Bias Mitigation

For each prompt, we run the judge twice with swapped ordering (A/B). We compute the bias-corrected win signal $w_i^J(M)$ by averaging the two outcomes.

E.4 Output Parsing Rules

We parse the judge output as a JSON object with a single field `winner`, whose value must be exactly one of `"A"`, `"B"`, or `"Tie"`. Outputs that fail JSON parsing or contain invalid values are re-queried once. If the output remains invalid, we assign a tie to avoid introducing systematic bias.

Type	Generation Prompt Template (Visualized Layout)
System	You are a helpful assistant for Lao-speaking users. Answer the user prompt in **fluent and natural Lao**. Do **not** switch to English or Chinese unless explicitly requested. **Rules:** • Respond only in Lao. • Prioritize correctness and clarity. • Avoid unnecessary verbosity; be concise but complete.
User	**[User Prompt]**

Table 7: Visualized generation prompt template used to obtain candidate responses on Lao-500. Both the baseline model and each challenger model are prompted with the same template to ensure fair comparison.

Type	Arena Judge Prompt Template (Visualized Layout)
System	You are a strict and fair evaluator for **Lao-language** answers. You will be shown a user prompt and two candidate answers (Answer A and Answer B). Your task is to decide which answer is better for a Lao-speaking user. **Evaluation Criteria:** (1) Correctness, (2) Completeness, (3) Reasoning Quality, (4) Clarity & Structure, (5) Lao Fluency & Appropriateness. **Rules:** • Do **not** favor verbosity. Prefer concise but complete answers. • If both answers are similarly good or similarly bad, output **Tie**. • The answer order is randomized. Do not assume A is better than B. • Do **not** reveal your reasoning or analysis. • Output must follow the required JSON format exactly. **Output Format (must be exact):** `{"winner": "A"}` or `{"winner": "B"}` or `{"winner": "Tie"}`
User	**[User Prompt]** **Answer A:** **[Answer A]** **Answer B:** **[Answer B]**

Table 8: Visualized Arena judge prompt template used in Lao-500 evaluation. The judge outputs only a JSON decision (`A/B/Tie`) to support deterministic parsing and reduce judge bias.

F Judge-Specific Arena Results and Cross-Judge Gap

Table 9 reports judge-specific win rates and 95% bootstrap confidence intervals for each model on Lao-500. We also report Δ(G$-$Q), the difference between Gemini-2.5-Pro and Qwen3-Max judges, and **Gap**=$|\Delta|$ to quantify judge sensitivity. We observe that some model families show systematic preference shifts under different judges, motivating our multi-judge aggregation protocol in the main paper. We further verify judge consistency by reporting cross-judge rank correlation (Spearman ρ and Kendall τ) in Appendix D.

G Contamination and Overlap Checks

We perform two forms of contamination analysis on the open-source subset (Lao-7k):

- **Web overlap retrieval:** searching question stems and key phrases via web search and checking for exact matches.
- **N-gram overlap:** measuring overlap between benchmark text and public corpora (e.g., Lao Wikipedia, news dumps).

We observe that 6.2% of items have potential overlap candidates. Manual inspection suggests that most cases are attributable to common factual statements rather than direct test leakage.

H Black-box Evaluation Protocol for Lao-10k

Black-box evaluation protocol. To enable secure and contamination-resistant evaluation, we keep Lao-10k hidden and evaluate models through a black-box protocol. Participants submit either (i) predicted option labels for a provided set of item IDs, or (ii) an inference API endpoint that follows

Model	Gemini	Gemini 95% CI	Qwen	Qwen 95% CI	Δ(G$-$Q)	Gap
Gemini-2.5-Pro	54.22	[52.51, 55.93]	48.85	[46.60, 51.10]	+5.37	5.37
Claude-Sonnet-4.5-20250929-thinking	50.50	[48.35, 52.65]	50.08	[47.51, 52.65]	+0.42	0.42
GPT-5-High (baseline)	49.94	[48.00, 51.88]	49.94	[47.99, 51.89]	+0.00	0.00
Claude-Opus-4.1-20250805	50.96	[48.91, 53.01]	47.64	[45.26, 50.02]	+3.32	3.32
Qwen3-Max	45.16	[43.16, 47.16]	52.80	[49.86, 55.74]	-7.64	7.64
Qwen3-235B-A22B-Instruct-2507	45.53	[43.50, 47.56]	51.75	[48.95, 54.55]	-6.22	6.22
DeepSeek-V3.2-Exp(Thinking)	48.69	[46.48, 50.90]	47.29	[44.67, 49.91]	+1.40	1.40
Qwen3-Plus(Non-Thinking)	45.91	[43.91, 47.91]	49.35	[46.19, 52.51]	-3.44	3.44
DeepSeek-V3.2-Exp(Non-Thinking)	47.96	[45.41, 50.51]	46.36	[43.07, 49.65]	+1.60	1.60
Qwen3-Plus(Thinking)	47.74	[45.15, 50.33]	45.20	[43.31, 47.09]	+2.54	2.54
Qwen3-235B-A22B	46.31	[44.15, 48.47]	45.53	[42.65, 48.41]	+0.78	0.78
Qwen3-Next-80B-A3B-Instruct	44.55	[41.85, 47.25]	45.73	[42.48, 48.98]	-1.18	1.18

Table 9: Judge-specific Arena win rates (%) on Lao-500 with 95% bootstrap confidence intervals (CI). Δ(G$-$Q) denotes the score difference between Gemini-2.5-Pro and Qwen3-Max judges, and **Gap** is the absolute difference.

our standardized prompt template. We return only aggregated scores (overall and subdomain-level accuracies) without per-item feedback, and enforce rate limits on submissions to reduce leaderboard overfitting. We will deploy this protocol as an online evaluation service upon publication.

I Error Analysis

I.1 Knowledge Application failure modes

We categorize model errors into:

1. **Cultural grounding errors:** misunderstanding Lao-specific conventions or institutions.
2. **Reasoning errors:** failing multi-step inference even with correct knowledge.
3. **Lexical confusion:** confusion caused by loanwords, named entities, or polysemy.

I.2 Translation error types

We analyze translation errors and identify: (i) terminology mistranslation, (ii) omission or hallucination, (iii) incorrect formal register, (iv) word-order and fluency degradation. We find that culturally grounded and legal/administrative domains exhibit the highest error rates.

J Human Performance Protocol

Human expert performance is measured by assigning 3 Lao-native experts to answer MCQ items without external tools. Each item is answered independently. We report mean accuracy and standard deviation across experts. Human accuracy exceeds 97% across all subdomains, confirming benchmark validity and meaningful headroom for future research.