

ASEAN-8B: A Prompted Qwen3.5-9B System for Chinese–Southeast Asian Web Translation

Haobin Wen

Student, South China University of Technology
2578277105@qq.com

Abstract

This paper describes ASEAN-8B, our training-free submission to the WMT26 Chinese–Southeast Asian Multilingual Machine Translation Task. ASEAN-8B uses the publicly available Qwen3.5-9B model with one direction-explicit prompt, non-thinking decoding, and a single generation pass. We first compare 17 systems on a balanced 210-example model-selection subset covering all 14 bidirectional directions between Chinese and Indonesian, Khmer, Lao, Malay, Burmese, Thai, and Vietnamese. We then evaluate ASEAN-8B on a separate 1,400-example balanced validation subset, with 100 examples per direction. Following the task-specific convention, we average sentence-level sacreBLEU with the FLORES-200 tokenizer and combine it equally with COMET-22. On the larger subset, ASEAN-8B obtains 24.86 BLEU, 83.23 COMET, and a quality score of 54.04. A local eight-GPU run processes 3.12 segments/s; reference lengths imply approximately 172 reference-length-equivalent output tokens/s per H100 80GB GPU. The deliberately simple system provides a reproducible baseline under the task’s 20B-parameter constraint.

1 Introduction

Machine translation between Chinese and Southeast Asian languages combines two practical difficulties. Several target languages remain comparatively low-resource, while real web text contains navigation menus, duplicated fragments, markup, names, dates, advertisements, and encoding artifacts that are uncommon in clean sentence-level benchmarks. The WMT26 task emphasizes both translation quality and efficient inference and limits submitted systems to at most 20 billion parameters (WMT26 Chinese–Southeast Asian Task Organizers, 2026).

Large language models offer a low-cost alternative to training a dedicated multilingual translation

model. We ask a deliberately narrow engineering question: how far can a publicly available 9B model go without parameter updates, retrieval, or multi-pass refinement?

Our submitted system, ASEAN-8B, uses Qwen3.5-9B (Qwen Team, 2026). We disable reasoning output and ask the model to return only the translation. No task-specific training, model merging, external quality-estimation model, or test-time candidate reranking is used. Our contributions are: (1) a reproducible prompted baseline covering all 14 task directions; (2) a controlled comparison with general-purpose and translation-specific models in one local pipeline; (3) a 1,400-example follow-up evaluation of the selected system; and (4) an analysis of failure modes in noisy website text.

2 Related Work

Prompted language models have become credible translation systems even without task-specific parameter updates. A WMT25 system study found that an off-the-shelf Qwen model with a fixed prompt forms a useful translation baseline (Grozea and Verbitsky, 2025). This motivates treating prompt, decoding configuration, and checkpoint identity as the complete system rather than as incidental implementation details.

Other WMT systems improve a base generator with retrieval augmentation (Guttmann et al., 2025) or iterative self-refinement (Grubisic and Korencic, 2025). Such methods can improve individual hypotheses, but they add retrieval indices, extra model calls, or reranking stages. Because WMT26 explicitly includes throughput in the final score, we instead study the simplest single-pass alternative and make its engineering path fully reproducible.

Metric reporting is also consequential in this setting. BLEU depends on tokenization, smooth-

ing, and aggregation choices (Post, 2018), while COMET can reward semantic adequacy when a hypothesis differs lexically from a single reference (Rei et al., 2022). We therefore report the task-specified sentence-averaged BLEU as the primary measure and add corpus-level BLEU as a sensitivity check.

3 Task and Data

The task covers translation from Chinese into seven Southeast Asian languages and translation from each of them back into Chinese. We use the codes id, km, lo, ms, my, th, vi, and zh. The released validation material contains 2,000 bilingual pairs for each Chinese–Southeast Asian language pair. It is derived from websites and spans domains such as news, health, finance, and entertainment.

For rapid system selection, we sampled 15 bilingual pairs per language pair with random seed 42 and evaluated both orientations, producing 210 directional examples in total. Each of the 14 directions therefore has equal weight in the macro average. This balanced model-selection subset is not a replacement for official blind evaluation.

After system selection, we drew a second balanced validation subset with 100 examples per direction, or 1,400 directional examples in total. This larger subset is used for the detailed ASEAN-8B results. It is kept separate from the 210-example comparison because we did not regenerate all 17 candidate systems on the larger sample.

The source and reference fields contain realistic website noise, including HTML iframe tags, navigation text, repeated captions, incomplete final sentences, and visible character-encoding corruption. We retain the supplied strings and perform no manual reference correction.

4 System Description

4.1 Base Model

ASEAN-8B is the registered name of our submitted system; the name is not a claim that its checkpoint contains exactly eight billion parameters. Its base model is Qwen3.5-9B, an open-weight post-trained causal language model (Qwen Team, 2026) with 9,653,104,368 parameters. It satisfies the task’s 20B ceiling. The model supports a non-thinking mode, which is useful for translation because visible reasoning can contaminate the hypothesis.

Development translations were obtained through the SiliconFlow endpoint exposing the identifier Qwen/Qwen3.5-9B. The endpoint did not expose the Git revision of the weights actually served, so we do not claim bitwise identity between the hosted outputs and a particular public snapshot. For official local reproduction, our package pins the Hugging Face model revision to the full Git commit c202236235762e1c871ad0ccb60c8ee5ba337b9a. This revision, together with the prompt, decoding parameters, and output-cleaning code, defines the submitted local system without dependence on an external API.

4.2 Prompt and Decoding

We use one prompt for every direction, substituting language names from a fixed mapping:

```
You are a professional translator.
Translate the following text from
{source language} to {target language}.
Output only the translation, without any
explanations or extra text.
```

The user message contains only the source text. The prompt avoids chain-of-thought requests, domain labels, and few-shot examples. Thinking is disabled through `enable_thinking=false`. Temperature is 0.1 and the maximum output is 2,048 tokens. Requests are issued concurrently with up to 15 in flight and failed calls are retried at most three times.

The client removes any `<think>... </think>` block and matching outer quotes. The submitted local runner uses Transformers in BF16, deterministic greedy decoding, and one independent model replica per GPU. Input batches contain text only.

4.3 Interruption Recovery

The pipeline stores one JSON object per sentence, including sample identifier, direction, source, reference, hypothesis, backend, model, and status. Successful identifiers are skipped on restart. Long runs therefore continue without repeating completed generation. Progress is reported per direction.

The local runner assigns stable, disjoint record-index shards to independent GPU workers. Each worker appends JSONL predictions and batch-level metrics to disk, so completed records survive worker or job interruption.

5 Evaluation

We report sacreBLEU (Post, 2018) and COMET (Rei et al., 2022). Following the task-organizer clarification supplied to participants, the primary BLEU score is computed separately for every sentence using sacreBLEU 2.6.0 with one reference, mixed case, exponential smoothing, effective order, and the FLORES-200 tokenizer. Scores are averaged within each direction and then macro-averaged over 14 directions. The saved signature is:

```
nrefs:1|case:mixed|eff:yes|
tok:flores200|smooth:exp|version:2.6.0
```

COMET uses Unbabel/wmt22-comet-da. Segment scores are averaged and multiplied by 100. The task quality score is

$$Q = 0.5 \text{ BLEU} + 0.5 \text{ COMET}. \quad (1)$$

Because sentence-level BLEU depends heavily on smoothing, we additionally compute corpus BLEU separately for each direction and macro-average the 14 direction scores. This secondary analysis uses the same FLORES-200 tokenizer and does not replace the task-specified primary metric.

For efficiency, we record a local eight-H100 run over 28,000 validation segments: 8,976.8 seconds wall-clock time, or 3.12 segments/s in aggregate and 0.38–0.40 segments/s per GPU. The run used batch size 8 per replica and a 512-token generation cap. Its logs did not retain generated-token counts, so a stratified 1,400-segment tokenizer sample with a mean reference length of 440.5 target tokens provides an approximate reference-length-equivalent throughput of 172 output tokens/s per GPU. Linear scaling projects four identical GPUs at 1.56 segments/s and 687 reference-length-equivalent output tokens/s. The four-GPU configuration was not run; it is a deployment projection, and the organizer-side throughput measurement is authoritative.

6 Results

6.1 Overall Comparison

Table 1 presents the ten strongest of 17 complete systems in the existing local evaluation; Appendix A gives the full ranking. All rows use the same 210 examples and scoring script. ASEAN-8B obtains the best point estimate, 54.50. Qwen3-14B has slightly higher BLEU but lower COMET,

Model	Params	BLEU	COMET	Q
ASEAN-8B	9.65B	25.19	83.82	54.50
Qwen3-14B	14B	25.76	82.40	54.08
Qwen3-8B-DS	8B	23.71	79.23	51.47
Qwen3-8B	8B	21.52	78.24	49.88
GLM-4-9B	9B	21.94	76.49	49.21
Qwen2.5-14B	14B	22.55	75.87	49.21
Qwen3-4B-Instruct	4B	14.94	73.25	44.10
Hunyuan-MT-7B	7B	12.91	73.00	42.95
HY-MT1.5-1.8B	1.8B	13.71	71.07	42.39
Qwen3-1.7B	1.7B	12.24	68.85	40.54

Table 1: Top ten macro-average development results among 17 complete systems. Every model is evaluated on the same 210 examples.

Direction	BLEU	COMET	Q
id→zh	31.81	82.36	57.09
km→zh	21.83	84.83	53.33
lo→zh	18.11	83.09	50.60
ms→zh	30.18	85.06	57.62
my→zh	17.35	85.98	51.67
th→zh	20.34	83.27	51.81
vi→zh	30.00	87.12	58.56
zh→id	37.02	85.68	61.35
zh→km	16.83	80.48	48.65
zh→lo	7.02	62.60	34.81
zh→ms	32.29	86.60	59.45
zh→my	7.78	82.47	45.13
zh→th	33.12	84.82	58.97
zh→vi	44.34	90.85	67.60
SEA→zh avg.	24.23	84.53	54.38
zh→SEA avg.	25.49	81.93	53.71

Table 2: ASEAN-8B (Qwen3.5-9B) results on the larger validation subset. Each direction contains 100 examples.

leaving its combined score 0.42 points behind. The remaining systems trail ASEAN-8B by at least 3.04 combined points in this model-selection subset.

6.2 Direction-Level Results

Table 2 reports the 1,400-example follow-up evaluation and shows large direction differences. Chinese-to-Vietnamese is strongest at 67.60 combined, followed by Chinese-to-Indonesian at 61.35. Chinese-to-Lao is weakest at 34.81, with 7.02 BLEU and 62.60 COMET. Chinese-to-Burmese is the next weakest direction.

The two broad sides remain close: 54.38 for Southeast Asian languages into Chinese and 53.71 for Chinese into Southeast Asian languages. Their metric composition differs. Chinese-to-Southeast-Asian has 1.26 points higher BLEU, driven by Vietnamese, Indonesian, Malay, and Thai, while Southeast-Asian-to-Chinese has COMET that is

2.60 points higher.

6.3 Metric Trade-offs

In the 210-example selection experiment, Qwen3-14B leads ASEAN-8B by 0.57 BLEU but trails by 1.42 COMET, so ASEAN-8B has the higher equal-weight point estimate. On the larger ASEAN-8B subset, Burmese-to-Chinese obtains only 17.35 BLEU while retaining 85.98 COMET. Chinese-to-Burmese is more extreme at 7.78 BLEU and 82.47 COMET. In both cases, the neural metric is much less pessimistic than surface overlap.

These gaps matter for system selection. Optimizing only adequacy may not improve the task score if the output moves farther from the single reference. Conversely, literal reproduction of damaged website strings may raise BLEU while reducing semantic quality. We therefore use the official equal-weight combination as the primary local criterion and retain the components for diagnosis.

6.4 Selection Stability and Metric Aggregation

Table 3 compares the ASEAN-8B macro average from the 15-example-per-direction model-selection subset with the separate 100-example-per-direction evaluation. Increasing the sample sevenfold changes BLEU by -0.33 , COMET by -0.59 , and the combined score by -0.46 points. Thus, the ASEAN-8B aggregate estimate is comparatively stable even though individual low-resource directions move more substantially. The 210-example experiment remains the only basis for cross-model ranking, while all direction-level analysis uses the larger subset.

The top-two ordering itself is uncertain. We perform 100,000 paired bootstrap resamples over the 14 direction-level Q differences (seed 42). ASEAN-8B exceeds Qwen3-14B by 0.424 points on average, but the 95% interval is $[-1.199, 2.128]$ and the bootstrap probability of a positive difference is 68.6%. Each model wins seven directions. We therefore treat ASEAN-8B as the selected engineering point estimate, not as a statistically significant winner. As an aggregation sensitivity check, macro-averaged direction-level corpus BLEU is 25.93 for ASEAN-8B and 26.40 for Qwen3-14B; the corresponding equal-weight scores are 54.87 and 54.40, so the point ordering is unchanged.

Subset	$n/\text{dir.}$	Total	BLEU	COMET	Q
Selection	15	210	25.19	83.82	54.50
Larger	100	1,400	24.86	83.23	54.04

Table 3: ASEAN-8B macro averages on two separately sampled balanced subsets.

7 Analysis

7.1 Strong and Weak Directions

Vietnamese, Indonesian, and Malay use Latin script and have substantial online text, which likely contributes to stronger lexical realization. Lao and Burmese are more difficult. For Chinese-to-Lao, both lexical overlap and semantic evaluation deteriorate, suggesting that the problem is not only tokenization. Chinese-to-Burmese has relatively high COMET but very low BLEU, consistent with a fluent paraphrase or reference-style mismatch.

Khmer is a middle case. Khmer-to-Chinese reaches COMET 84.83 but BLEU 21.83. Chinese-to-Khmer falls to 16.83 BLEU and 80.48 COMET. The model often appears to preserve broad meaning while differing from the supplied reference at the surface level.

7.2 Website Noise and Case Studies

The material differs from clean news sentences. One Indonesian example begins with an iframe, a Google Tag Manager URL, and a long list of navigation categories before the article headline. ASEAN-8B sometimes translates only the salient headline and omits navigation fragments. This is reasonable for a user-facing translator but is penalized when the reference preserves boilerplate.

Visible encoding corruption creates a second failure mode. Some Chinese strings contain mojibake-like sequences, punctuation damage, or truncated phrases. A general-purpose model may infer intended meaning and produce a clean target sentence, whereas BLEU rewards matching the literal reference style. This tension helps explain why COMET remains high in directions with low BLEU.

We inspect two examples from the 1,400-example validation subset (Table 4). The first, id2zh-003, is a concatenation of entertainment headlines, navigation labels, popularity counters, and a truncated final item. The reference preserves most of this page structure. The hypothesis translates the salient headlines but compresses the list,

ID	Dir.	BLEU	COMET	Main behavior
id2zh-003	id→zh	15.79	74.06	Boilerplate compression and omissions
th2zh-010	th→zh	10.08	84.04	Fluent normalization of a noisy reference

Table 4: Two representative examples from the larger validation subset. Scores are segment-level.

omits several navigation fragments, and ends with the same incomplete context. This behavior produces only 15.79 BLEU and also lowers COMET to 74.06. The example shows that website translation is partly a content-selection problem: a concise user-facing rendering can lose points when the reference retains boilerplate.

The second example, th2zh-010, is a first-person account of a farmer who survived through agriculture but accumulated wealth through land appreciation. The supplied Chinese reference is highly literal, contains abrupt pronoun shifts, and includes the artifact *n n*. ASEAN-8B instead produces a coherent narrative, preserves the age, 30-year duration, medical expenses, land sale, and final contrast between farming and land speculation, while smoothing the damaged phrasing. Its BLEU is only 10.08, but COMET reaches 84.04. Together, the cases illustrate why both metrics are needed: omissions damage both scores, whereas fluent normalization of a noisy reference can retain semantic credit while sharply reducing lexical overlap.

7.3 Why a Single-Pass System?

Best-of- N generation, self-refinement, and neural reranking can improve some systems, but they multiply latency and may add model parameters. WMT26 explicitly rewards throughput. ASEAN-8B therefore performs one call per input. The simple design also isolates the capabilities of Qwen3.5-9B and reduces the components that organizers must reproduce.

8 Limitations and Ethical Considerations

The cross-model comparison contains only 15 examples per direction. The larger ASEAN-8B-only subset contains 100 per direction, but it still does not establish an expected official ranking and cannot validate the relative ordering of all 17 systems. The paired bootstrap interval includes zero, con-

firmed that the ASEAN-8B–Qwen3-14B ordering is uncertain. The data contain long website fragments, so mean sentence-level BLEU should not be compared directly with conventional corpus BLEU; we report corpus BLEU only as a sensitivity analysis. Development outputs were obtained through a hosted endpoint, so provider implementation details can affect them even when the public model identifier is unchanged. Our analysis is metric-based and does not include native-speaker human evaluation.

The released web material may contain personal names, political content, stereotypes, or damaged text. Model output can omit source material or turn corrupted strings into plausible but unsupported statements. The system is not a certified translator and should not be used for medical, legal, financial, or safety-critical decisions without qualified human review.

9 Reproducibility and Submission

The intended repository contains the exact model revision, inference script, dependency specification, prompt, language mapping, and output validator. The pinned base model is Qwen/Qwen3.5-9B at Git revision c202236235762e1c871ad0ccb60c8ee5ba337b9a; the same value is machine-readable in `reproducibility.json`. The complete model has 9,653,104,368 parameters. No LoRA, auxiliary language model, reranker, or external service is required during official evaluation.

Generation uses Transformers with BF16 weights, writes append-only JSONL records, and displays progress. On restart, completed sample identifiers are loaded and skipped. The 19–20GB weights fit on one 80GB GPU, and independent replicas process stable data shards without tensor-parallel communication.

10 Conclusion

We presented ASEAN-8B, a training-free system based on Qwen3.5-9B, for 14 Chinese–Southeast Asian translation directions. On a balanced 210-example model-selection subset, it obtains the highest point estimate among 17 complete systems in our local comparison, although paired bootstrap resampling does not establish a statistically stable advantage over Qwen3-14B. On a separate 1,400-example evaluation, ASEAN-8B obtains 24.86 sentence-averaged FLORES-200

sacreBLEU, 83.23 COMET, and a combined score of 54.04. Performance is strongest for Chinese-to-Vietnamese and Chinese-to-Indonesian and weakest for Chinese-to-Lao and Chinese-to-Burmese. The system requires only a fixed prompt, non-thinking decoding, and lightweight output cleanup. Its open weights, 9.65B parameters, and single-pass inference make it a suitable reproducible entry for an efficiency-aware shared task.

A Complete 17-Model Results

Table 5 gives every complete system in the balanced model-selection experiment. All entries cover the same 14 directions and 210 examples. The ranking excludes Qwen3.6-27B, which exceeds the task parameter ceiling and completed only 33 examples, and Qwen2.5-7B, which is missing one example. API and local systems are retained because the purpose of this subset is engineering selection rather than a hardware-controlled efficiency comparison.

Rank	Model	Params	BLEU	COMET	<i>Q</i>
1	ASEAN-8B	9.65B	25.19	83.82	54.50
2	Qwen3-14B	14B	25.76	82.40	54.08
3	Qwen3-8B-DS	8B	23.71	79.23	51.47
4	Qwen3-8B	8B	21.52	78.24	49.88
5	GLM-4-9B	9B	21.94	76.49	49.21
6	Qwen2.5-14B	14B	22.55	75.87	49.21
7	Qwen3-4B-Instruct	4B	14.94	73.25	44.10
8	Hunyuan-MT-7B	7B	12.91	73.00	42.95
9	HY-MT1.5-1.8B	1.8B	13.71	71.07	42.39
10	Qwen3-1.7B	1.7B	12.24	68.85	40.54
11	SeaLLMs-v3-7B	7B	12.03	67.19	39.61
12	SeaLLMs-v3-1.5B	1.5B	11.13	67.38	39.26
13	NLLB-200-3.3B	3.3B	10.69	66.98	38.84
14	NLLB-200-dist.-1.3B	1.3B	10.08	66.56	38.32
15	NLLB-200-dist.-600M	0.6B	7.69	61.42	34.55
16	Qwen3-0.6B	0.6B	6.45	57.33	31.89
17	MADLAD-400-3B-MT	3B	6.29	46.76	26.52

Table 5: Complete macro-average ranking on the 210-example model-selection subset.

References

- Cristian Grozea and Oleg Verbitsky. 2025. [Evaluation of Qwen3 for English to Ukrainian translation](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 594–598.
- Ivan Grubisic and Damir Korencic. 2025. [IRB-MT at WMT25 translation task: A simple agentic system using an off-the-shelf LLM](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 753–764.
- Kamil Guttman, Zofia Rostek, Adrian Charkiewicz, Antoni Solariski, Mikolaj Pokrywka, and Artur

Nowakowski. 2025. [Lanigo at WMT25 general translation task: Self-improved and retrieval-augmented translation](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 778–788.

Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Qwen Team. 2026. [Qwen3.5-9B model card](#). Hugging Face model repository. Accessed 2026-08-05.

Ricardo Rei, Jose G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C. Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and Andre F. T. Martins. 2022. [COMET-22: Unbabel-IST 2022 submission for the metrics shared task](#). In *Proceedings of the Seventh Conference on Machine Translation*, pages 578–585.

WMT26 Chinese–Southeast Asian Task Organizers. 2026. [WMT26 chinese–southeast asian multilingual MT task](#). Accessed 2026-08-05.