

Findings of the WMT26 Chinese–Southeast Asian Multilingual MT Shared Task

Tianyu Dong¹, Ziyang Chen², Jingsong Liu², Shaolin Zhu^{1*}

¹Tianjin University

²Newtranx

{tydong, zhushaolin}@tju.edu.cn

Abstract

We present the findings of the WMT26 Chinese–Southeast Asian Multilingual Machine Translation shared task. The task evaluates fourteen translation directions between Chinese and Thai, Vietnamese, Lao, Burmese, Khmer, Indonesian, and Malay. Organizers released manually aligned parallel data, domain-matched monolingual data, validation data, and a blind test set. The published leaderboard contains eleven systems and combines translation quality, measured by the equal-weight average of sacreBLEU and COMET, with a throughput bonus measured on a shared four-GPU platform. Jiutian-MT ranks first overall, while AiBabelMT achieves the highest throughput. We describe the task design, data resources, evaluation protocol, and published aggregate results, then examine how quality and throughput shape the final ranking. The results also motivate more detailed reporting by language direction and domain.

1 Introduction

Machine translation between Chinese and Southeast Asian languages serves cross-border communication, public information access, and commercial applications. The language pairs in this task differ substantially in the amount of digital text and parallel training data available to researchers. For languages with limited parallel corpora, model quality depends on architecture, data selection, multilingual transfer, and effective use of monolingual text (Haddow et al., 2022). Practical deployment also depends on memory use and latency.

The WMT26 Chinese–Southeast Asian Multilingual MT task brings these concerns into one evaluation campaign (Chen et al., 2026). It covers Chinese paired with Thai, Vietnamese, Lao, Burmese, Khmer, Indonesian, and Malay, in both translation directions. Organizers provide a common set of

training and evaluation resources and assess system outputs on a held-out test set. The ranking combines automatic translation quality with inference throughput, reflecting the task’s interest in compact and efficient translation as well as accuracy.

This paper follows the findings-paper format used by WMT shared task reports. We document the task definition and released resources, explain the scoring rules, reproduce the final aggregate results, and examine how quality and throughput shape the ranking. Our analysis compares the campaign-wide system scores and their quality and throughput components.

The main observations are as follows. First, Jiutian-MT has the highest reported translation-quality score and the highest final score, with a third-place throughput rank. Second, throughput can materially change the ordering: AiBabelMT ranks third overall with the fastest measured throughput, while BridgeMT has a slightly higher quality score. Third, systems with identical quality scores can finish at different ranks: ASEAN and KDA_MT tie on quality, yet their different throughput bonuses place them fifth and seventh in the final ranking.

2 Related Work

Shared tasks make comparisons more interpretable by fixing a common test set, submission protocol, and scoring procedure. Overview papers normally describe the resources and task organization before analyzing participating systems. The WMT25 Creole machine translation report combines system and data tracks, describes participating approaches, and discusses the resulting resource and translation contributions (Robinson et al., 2025). The WMT25 low-resource Indic translation report organizes its language pairs into data availability categories and presents results across language directions (Pakray et al., 2025). These papers provide useful mod-

*Corresponding author.

els for reporting low-resource campaigns: dataset provenance and evaluation detail are part of the findings, rather than background alone.

The present task shares the low-resource motivation of those campaigns, while evaluating seven Chinese–Southeast Asian pairs bidirectionally and adding an inference-throughput bonus to the automatic quality score. Differences in language pairs, training data, metrics, test domains, and scoring scales make the WMT25 reports most useful here as structural and methodological references.

Automatic MT evaluation continues to use reference-based lexical metrics and learned metrics with complementary properties. SacreBLEU provides a standardized implementation and reporting signature for BLEU (Post, 2018); COMET uses learned representations to estimate translation quality with respect to source and reference information (Rei et al., 2020). The WMT26 task combines the two in an equal-weight average. This gives the campaign a single quality score for ranking, while leaving the individual metric values visible in the result table.

3 Task Overview

3.1 Language directions and objective

The task includes seven language pairs: Chinese–Thai, Chinese–Vietnamese, Chinese–Lao, Chinese–Burmese, Chinese–Khmer, Chinese–Indonesian, and Chinese–Malay. Each pair is evaluated in both directions, giving fourteen directions in total. Figure 1 summarizes the coverage. The organizers frame the task around data-efficient multilingual translation, model efficiency, and inference speed for practical use, including real-time and mobile scenarios (Chen et al., 2026).

The shared task supplies a common benchmark and leaves model architecture to participants. They can investigate supervised learning from the released parallel corpus, multilingual transfer, data augmentation, model compression, or decoding optimization. The task page specifically allows parallel and monolingual data from the WMT26 General MT task to be used for augmentation. This makes the evaluation relevant to systems that combine task-provided resources with broader training data. System descriptions can pair leaderboard scores with the resources used by each team.

3.2 Participation and submission constraints

The task page specifies at most one translation output per language-pair direction for each participating team. Submitted systems must remain within the model-size limit of 20 billion parameters. For mixture-of-experts models, the limit applies to the total parameter count, including the full expert set when only a subset is selected for an input. These rules place a common upper bound on model scale while allowing different architectures and optimization choices.

Teams are asked to provide test-set outputs and a Hugging Face model repository link containing model weights. The task page also lists a system description paper deadline as part of the WMT26 submission process. These artifacts support inspection and reuse beyond the final score. Reading the leaderboard alongside the corresponding system descriptions connects each team’s score with its model design, training data, and decoding configuration.

3.3 Campaign timeline

Figure 2 shows selected milestones on the task page: the announcement, release of training and validation data, system-paper deadline, test-output and model-repository deadline, and publication of the final results. This sequence separates development time from official evaluation: teams can use the released training and validation resources to build systems, then submit outputs for organizer-side scoring against the hidden references. We show selected milestones rather than every date listed on the webpage.

4 Data Resources

4.1 Collection and released splits

The organizers report that the data were collected from copyright-compliant resources including OPUS, Tatoeba, the UN Corpus, and manually proofread corpora. The released material covers general, news, healthcare, finance, and other practical domains. The organizers further report that sourcing and quality-control work took eight months and involved approximately EUR 30,000 in manual alignment and native-speaker verification (Chen et al., 2026). This investment is relevant context for interpreting the size and intended quality of the benchmark: the task prioritizes a modest amount of curated parallel data rather than a very large automatically collected corpus.

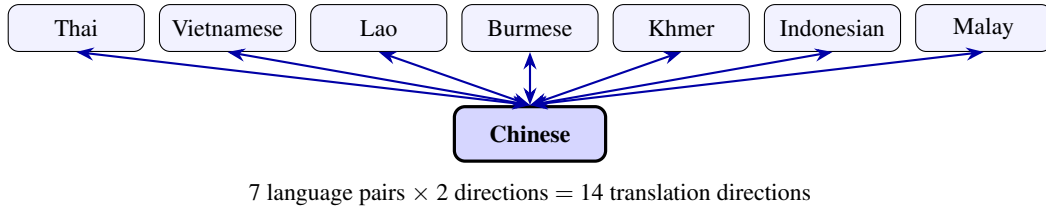


Figure 1: Language coverage of the WMT26 Chinese–Southeast Asian task.

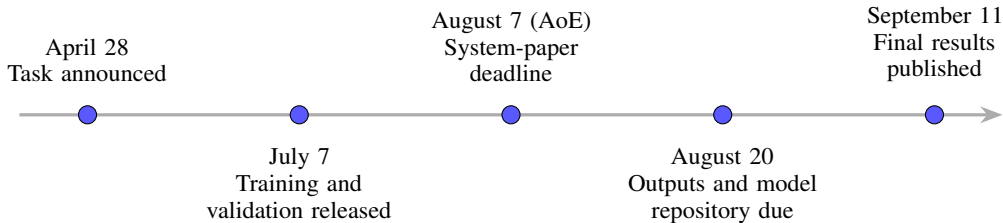


Figure 2: Selected milestones listed on the WMT26 task page.

Table 1 summarizes the four resource types described on the official task page. Parallel training data comprise 20,000 sentence pairs for each of the seven language pairs. The 700,000 monolingual sentences are distributed across the seven Southeast Asian languages, with 100,000 for each. Validation data provide 2,000 sentences per Southeast Asian language. The blind test set has 2,000 source sentences for each of eight source languages: Chinese and the seven Southeast Asian languages. In total, the test set contains 16,000 source sentences.

The task page describes the data as multi-domain and states that the validation set is unseen and independent of the training data. It also describes the blind test set as held out and independent of both training and validation data. Test inputs are made available to registered participants after approval; domain labels accompany the source text, while human references remain with the organizers for official scoring. Participants can develop systems on the shared inputs while the reference translations remain private until official evaluation.

Figure 3 summarizes the collection process and counts reported by the organizers. Additional documentation of domain-level counts, alignment procedures, and native-speaker verification criteria would further support corpus reproduction and domain-specific analysis.

4.2 Use and scope of the resources

The parallel corpus offers equal pair-level totals across the seven Chinese–Southeast Asian pairs, while the monolingual component supports methods such as back-translation or forward-translation.

The language pairs also differ in script, tokenization, lexical overlap with Chinese, and access to pretrained resources, creating distinct modeling conditions across the fourteen directions.

The task also permits WMT26 General MT data for augmentation. This choice reflects practical system development, where training data may combine a small task-specific corpus with broader resources. For reproducible comparisons, system descriptions can identify external corpora, pretrained checkpoints, filtering rules, and synthetic data generated from the released monolingual sentences. These details clarify how data choices affect translation quality and model size.

5 Evaluation Protocol

5.1 Translation quality

The organizers score output against hidden human references using sacreBLEU and COMET. Both metrics are reported on a 0–100 scale and receive equal weight (Chen et al., 2026). The translation-quality score is

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

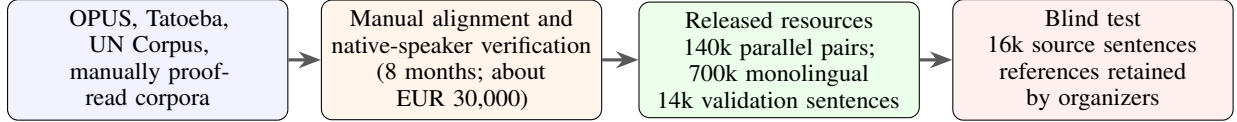
The displayed quality score is therefore easy to reproduce from the two columns in the results table, subject to rounding. Keeping both component metrics visible is useful because a single average can hide cases where systems perform differently under lexical overlap and learned semantic scoring.

5.2 Throughput and final ranking

Throughput is measured by the organizers on a shared platform equipped with four 80-GB GPUs.

Resource	Size	Distribution and role
Parallel training	140,000 pairs	20,000 aligned pairs for each of seven Chinese–Southeast Asian language pairs.
Monolingual training	700,000 sentences	100,000 sentences for each Southeast Asian language; available for augmentation.
Validation	14,000 sentences	2,000 per Southeast Asian language, with a human reference.
Blind test	16,000 sentences	2,000 source sentences for Chinese and each of the seven Southeast Asian languages; references are hidden.

Table 1: Released resources reported on the WMT26 task page.



Reported domains: general, news, healthcare, finance, and other practical fields

Figure 3: Resource overview based on the official task description. The blind test references are withheld for centralized evaluation.

Systems are ordered from fastest to slowest. Let N be the number of systems in the comparison and r the throughput rank, where rank 1 is fastest. For $N > 1$, the task assigns the bonus

$$B(r; N) = 0.5 + \frac{N - r}{N - 1} \left(\frac{N}{2} - 0.5 \right). \quad (2)$$

The fastest system receives $N/2$ points and the slowest receives 0.5 points. If only one system is evaluated, the bonus is 0.5. The final score is

$$S = Q + B(r; N). \quad (3)$$

For $N > 1$, Equation 2 simplifies to $B(r; N) = (N - r + 1)/2$. Thus, each one-place change in throughput rank changes the bonus by exactly 0.5 points. With eleven systems, the fastest-to-slowest bonus spread is five points. The final score therefore combines translation quality with a campaign-relative reward for throughput rank. For the eleven systems on the current results page, the bonus ranges from 0.5 to 5.5 points. Reporting both components alongside absolute speed measurements gives readers the full efficiency profile.

Figure 4 depicts the two scoring paths. The design makes translation quality and runtime efficiency contribute in separate, visible ways. Two systems with similar quality can be separated by throughput, and the quality leader can receive a smaller bonus than a faster competitor. The shared four-GPU measurement provides a campaign-level efficiency comparison; per-language throughput and measurements on target devices would extend the analysis to specific deployment settings.

6 Results

The official page lists eleven systems and provides one aggregate BLEU score, COMET score, throughput rank, bonus, quality score, and final score for each. Table 2 reproduces those values as published on September 11, 2026 (Chen et al., 2026). We preserve the system names and official final ranking, and analyze the quality and throughput components across systems.

Jiutian-MT ranks first with a quality score of 59.01 and a throughput rank of 3, producing a final score of 63.51. Xiaoyu-MT follows at 61.80. The 1.71-point gap between them consists of a 0.71-point quality difference and a one-point throughput-bonus difference. Thus, the top result combines the strongest published quality with a third-place throughput rank. The two highest quality scores are also well separated from the next group in the final ranking.

AiBabelMT is the fastest system, receives the maximum 5.5-point bonus, and ranks third overall. BridgeMT has a 0.20-point higher quality score than AiBabelMT, but its throughput bonus is 0.5 points lower; AiBabelMT therefore finishes 0.30 points ahead. This example shows how the combined score changes the quality-only ordering: the throughput advantage offsets AiBabelMT’s small quality-score gap under the task’s scoring rule.

ASEAN and KDA_MT each have a quality score of 55.14. Their throughput ranks are 7 and 11, respectively, so their bonuses are 2.5 and 0.5 points. The four-point bonus difference yields a two-place difference in final rank, with ASEAN fifth and

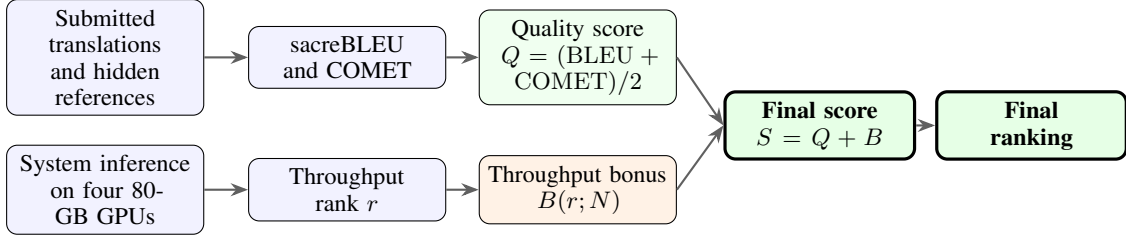


Figure 4: The official evaluation combines reference-based quality and relative throughput into a final score.

Final rank	System	BLEU	COMET	Throughput rank	Bonus	Quality Q	Final score
1	Jiutian-MT	35.34	82.68	3	4.50	59.01	63.51
2	Xiaoyu-MT	34.73	81.87	5	3.50	58.30	61.80
3	AiBabelMT	27.61	79.28	1	5.50	53.45	58.95
4	BridgeMT	27.74	79.56	2	5.00	53.65	58.65
5	ASEAN	29.06	81.22	7	2.50	55.14	57.64
6	WMT26_ours2	26.37	78.97	4	4.00	52.67	56.67
7	KDA_MT	29.17	81.11	11	0.50	55.14	55.64
8	zzuli_software	26.58	78.68	6	3.00	52.63	55.63
9	luke	27.88	80.42	10	1.00	54.15	55.15
10	FlashMT	26.71	78.43	8	2.00	52.57	54.57
11	my_mt5	25.89	77.83	9	1.50	51.86	53.36

Table 2: Published aggregate results. Quality is the equal-weight average of sacreBLEU and COMET; final score adds the throughput bonus.

KDA_MT seventh. These examples show that the benchmark is designed to reward the quality–efficiency balance rather than quality alone. Figure 5 decomposes the final score into its two parts; Figure 6 places quality against throughput rank.

6.1 Reading the aggregate results

The quality scores span 51.86 to 59.01, while throughput ranks span all positions from 1 to 11. The scatter plot visualizes the overall quality–throughput relationship among submitted systems.

The system order also illustrates why the components should be reported alongside the combined score. A quality-only ranking would place BridgeMT slightly above AiBabelMT and would tie ASEAN and KDA_MT on the displayed quality score. The official final ranking separates these systems using throughput. Conversely, a throughput-only ranking would put systems with lower quality ahead of stronger translators. Neither component alone captures the stated goal of balancing translation quality and efficiency.

7 Discussion

7.1 Quality and efficiency as complementary objectives

The leaderboard illustrates the task’s premise that translation quality and inference throughput are

distinct objectives. Jiutian-MT and Xiaoyu-MT lead the final table through high quality scores, whereas AiBabelMT and BridgeMT receive the two largest bonuses through their first- and second-place throughput ranks. The combined score makes this trade-off explicit and provides one ordering for the campaign. Deployment teams can adapt the quality–efficiency balance to their latency targets and the quality requirements of each application.

The bonus is rank based: for a given N , each one-place throughput change has the same effect on the score. Absolute throughput, latency, batch size, sequence lengths, and hardware utilization would complement this relative ranking and make the efficiency results more actionable for real-time and edge use.

7.2 Priorities for further reporting

The task covers fourteen directions and several domains. A fuller findings report could pair BLEU and COMET by direction with test-set sizes and the aggregation rule, and summarize variation across domains. These results would help identify where additional data or language-specific adaptation may matter most.

Confidence intervals or significance tests would add context to the reported point estimates. Segment-level paired resampling can quantify score variation, while human assessment can pro-

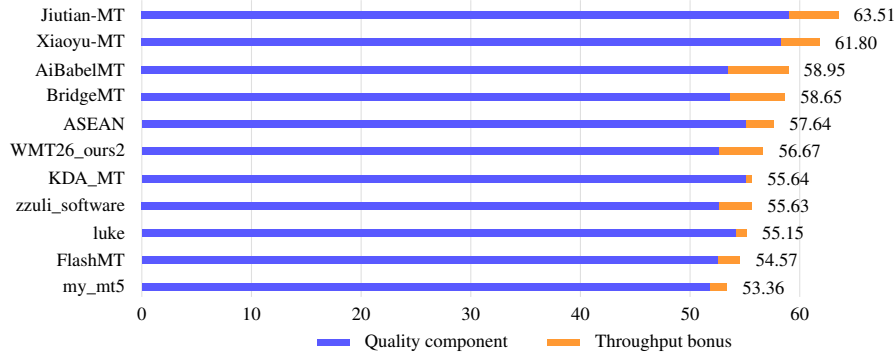


Figure 5: Final-score composition by system. Each bar shows the quality score Q plus the throughput bonus; labels at the right are final scores.

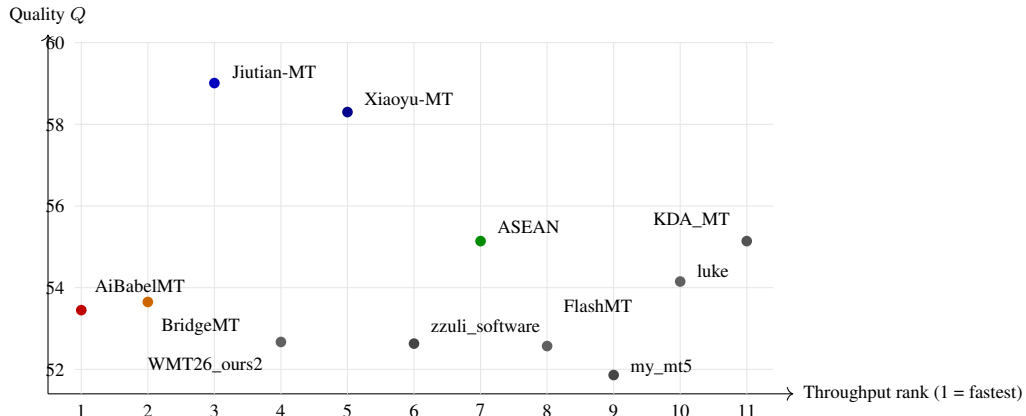


Figure 6: Aggregate quality versus throughput rank. Lower throughput rank means faster inference; each point is one system.

vide a complementary perspective on adequacy, fluency, and named entities.

System papers and repositories can complement the aggregate scores with details on architecture, pretraining, fine-tuning data, tokenization, and decoding choices. This information helps relate leaderboard performance to the methods used by each team.

8 Ethical and Data Considerations

The task organizers describe the released datasets as sourced from copyright-compliant materials and available to registered researchers for non-commercial research under the task’s stated terms (Chen et al., 2026). Per-source license references, provenance metadata, and descriptions of privacy screening would further support transparent corpus use.

The test set provides domain labels and source sentences while references remain hidden for official evaluation. Documentation of contamination checks and sensitive-text screening would further

inform use of this multilingual benchmark, particularly for its healthcare and finance domains. Applications in high-impact settings can pair benchmark results with task-specific review.

The throughput measurement uses a four-GPU platform and provides a consistent basis for the campaign’s relative speed ranking. Deployment reports can complement it with energy, memory, and latency measurements on intended hardware, together with evaluation of the relevant language directions and use cases.

9 Conclusion

The WMT26 Chinese–Southeast Asian shared task provides a common evaluation for fourteen translation directions, supported by curated parallel and monolingual data and a hidden-reference test set. Its official score combines the equal-weight average of sacreBLEU and COMET with a throughput bonus. Among the eleven systems on the published page, Jiutian-MT ranks first overall and AiBabelMT is fastest. The results illustrate the

quality–speed trade-off and provide a system-level view across the campaign.

References

- Chen, Ziyan, Jingsong Liu, and Shaolin Zhu. 2026. Chinese-Southeast Asian Multilingual MT Task. WMT26 Shared Task website. Last updated September 11, 2026; accessed September 20, 2026. <https://www2.statmt.org/wmt26/chinese-southeast-asian.html>.
- Haddow, Barry, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of Low-Resource Machine Translation. *Computational Linguistics*, 48(3):673–732. https://doi.org/10.1162/coli_a_00446.
- Pakray, Partha, Reddi Krishna, Santanu Pal, Advaita Vetagiri, Sandeep Dash, Arnab Kumar Maji, Saralin A. Lyngdoh, Lenin Laitonjam, Anupam Jamatia, Koj Sambyo, Ajit Das, and Riyanka Manna. 2025. Findings of WMT 2025 Shared Task on Low-resource Indic Languages Translation. In *Proceedings of the Tenth Conference on Machine Translation*, pages 532–553, Suzhou, China. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.wmt-1.29>.
- Post, Matt. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics. <https://doi.org/10.18653/v1/W18-6319>.
- Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2685–2702, Online. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.emnlp-main.213>.
- Robinson, Nathaniel Romney, Claire Bizon Monroc, Rasul Dent, Stefan Watson, Kenton Murray, Raj Dabre, Andre Coy, and Heather Lent. 2025. Findings of the First Shared Task for Creole Language Machine Translation at WMT25. In *Proceedings of the Tenth Conference on Machine Translation*, pages 520–531, Suzhou, China. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2025.wmt-1.28>.