Simple Test Time Scaling for Machine Translation: Kaze-MT at the WMT25 General Translation Task

Shaomu Tan **Christof Monz**
Language Technology Lab
University of Amsterdam
{s.tan, c.monz}@uva.nl

Abstract

This paper describes the Kaze-MT submission to the WMT25 General Machine Translation task (Japanese–Chinese). Our system deliberately adopts a minimalist Test-Time Scaling (TTS) pipeline with three stages—*Sampling*, *Scoring*, and *Selection*—while avoiding any task-specific fine-tuning, in-context exemplars, or bespoke decoding heuristics. In the sampling stage, we use the zero-shot `Qwen2.5-72B-Instruct` model to generate 512 candidate translations under a fixed temperature schedule designed to encourage lexical and syntactic diversity without sacrificing fluency. In the scoring stage, each candidate is evaluated by multiple reference-free quality estimation (QE) models—KIWI-22, MetricX-24 Hybrid-XXL, and Remedy-24-9B. The selection stage aggregates metric-specific rankings and chooses the candidate with the lowest mean rank, which we found more stable than averaging raw scores across heterogeneous ranges. We submit to both constrained and unconstrained tracks with minimal configuration changes. According to official preliminary results, our submissions are competitive on automatic metrics; in human evaluation, Kaze-MT falls within the 8–13 cluster, delivering performance comparable to CommandA-WMT and DeepSeek-V3 and outperforming other large LLM baselines such as Mistral-Medium and other extensively tuned MT systems.

1 Introduction

Allocating additional computation at inference time—commonly referred to as *Test-Time Scaling (TTS)* or *Best-of-N (BoN)*—can improve quality without the overhead of scaling training to ever larger models (Snell et al., 2024; Wu et al., 2025; Muennighoff et al., 2025). In machine translation, TTS has a long history via candidate reranking using quality estimation (QE) metrics (Neubig et al., 2015; Mizumoto and Matsumoto, 2016; Lee et al., 2021). Rather than optimizing a particular reranking recipe, Tan et al. (2025) study scaling laws for TTS-MT and find that scaling N for Best-of-N brings performance improvements for high-resource languages.

We adopt this minimalist perspective. Our submission, **Kaze-MT**, relies on a strong, off-the-shelf LLM for diverse candidate generation and on robust, reference-free QE models for selection. Our submission targets the WMT25 Japanese–Chinese track and intentionally avoids any task-specific parameter updates or domain adaptation. The pipeline is deliberately simple: (i) *Sampling*, (ii) *Scoring*, and (iii) *Selection*—yet competitive against substantially engineered systems. Beyond reporting official results, we discuss metric–human preference gaps and practical considerations for scaling TTS under realistic compute constraints.

On the official WMT25 Japanese→Chinese evaluation, **Kaze-MT** attains a strong position under automatic metrics and competitive human judgments despite using no fine-tuning or in-context exemplars. In AutoRank (an ensemble of KIWI-XL, GEMBA-ESA-CMDA, GEMBA-ESA-GPT-4.1, MetricX-24 Hybrid-XL, and XCOMET-XL), our primary system ranks **4/41** submissions (Table 2), outperforming several large closed LLM baselines (e.g., GPT-4.1, Claude-4, DeepSeek-V3, Mistral-Medium). Even though there is no exact the same metric used for both TTS setup and AutoRank evaluation, we acknowledge that potential metric interference (Pombal et al., 2025) may exist.

In the final official human evaluation, Kaze-MT falls in the **8–13** cluster (Table 1), comparable to CommandA-WMT and DeepSeek-V3 and ahead of models such as Mistral-Medium and Qwen3-235B. We note a modest gap between AutoRank and human ranking, which indicates that Quality Estimation as signal for improving translation quality remain a unclear problem for the future study. Developing human preference aligned MT metrics, therefore, hold a great promise for machine transla-

Proceedings of the Tenth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 651–656
November 8–9, 2025.

tion.

2 Task Overview

The WMT25 Japanese–Chinese track evaluates systems with both automatic metrics and human judgments. The *constrained* track limits models and resources (e.g., parameter count <20B and approved data), whereas the *unconstrained* track permits any publicly available model or data. Our pipeline fits both settings with minor differences in the scoring configuration (e.g., which QE variants are permitted).

We submitted a *primary* system built on `Qwen2.5-72B-Instruct` and a *contrastive* system built on `Qwen2.5-14B-Instruct`(Hui et al., 2024). The contrastive run was not included in AutoRank or human evaluation by the organizers; thus, we only report the 72B primary system in this paper.

3 Data

Because Kaze-MT is purely zero-shot, no pretraining or fine-tuning data are used beyond the official test set. The WMT25 materials contain multiple domains and are provided at the document level. Very long contexts may degrade generation performance and stability; therefore, we segment documents into paragraph units simply using a double-newline delimiter (`\n\n`). We retain original sentence order within each paragraph and do not apply additional filtering or normalization beyond standard Unicode cleanup.

4 Methodology

4.1 Sampling

We generate $N=512$ candidates per source paragraph with `Qwen2.5-72B-Instruct` (Hui et al., 2024) in zero-shot mode. Decoding with top-$p=0.95$ and a fixed temperature $t=1.0$ across all candidates to produce lexical and structural variety. The maximum generation length is 1500 tokens with EOS-based stopping. We implement inference with `vLLM` (Kwon et al., 2023), employing data parallelism on 4× NVIDIA H100 NVL GPUs. We observed that holding t fixed while sampling many candidates yields more predictable diversity than annealing schedules in this setting. Figure 4.1 demonstrates our translation generation prompt.

Why using N=512? Following Tan et al. (2025), who evaluate Best-of-N with $N \in \{1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024\}$, we select $N=512$ as a sweet point on the TTS Pareto frontier. Empirically, the quality–compute curve exhibits clear diminishing returns: as N grows, candidate diversity increases sublinearly, and the marginal utility of additional samples is increasingly limited by redundancy among high-probability modes. In that regime, $N=512$ lies very close (in terms of automatic quality) to the performance obtained with $N=1024$, yet it requires roughly *half* the sampling and scoring budget. In sum, $N=512$ captures most of the attainable TTS benefit identified by prior scaling studies while maintaining a favorable quality–latency trade-off for production-style constraints.

Translation Prompt Template

You are a helpful translation assistant. Now translate the following src_lang text into natural, fluent tgt_lang sentence while preserving the original meaning.

––
Source: $SOURCE

4.2 Scoring

Each candidate is scored by three reference-free QE models spanning different capacities and training paradigms:

- **KIWI-22** (0.5B): a lightweight, widely deployed QE model trained on synthetic and human annotations (Rei et al., 2022b).
- **MetricX-24 Hybrid-XXL** (13B): a strong WMT24 metric that combines synthetic judgments and curated references (Juraska et al., 2024).
- **Remedy-24 (9B)**: a recent SOTA QE model emphasizing robustness to domain and format variation (Tan and Monz, 2025).

Why ensemble? Individual QE metrics differ in architecture, training data, and inductive biases (e.g., sensitivity to literalness, tolerance to stylistic risk, robustness to domain drift). In practice, these differences induce complementary error profiles: one metric may down-weight fluent paraphrases, another may reward stylistic richness but underpenalize subtle adequacy errors.

An ensemble therefore acts as a variance-reduction mechanism, stabilizing selection across domains and styles. As emprical studied in Rei et al. (2022a); Freitag et al. (2024), ensembling the same metric model with different random seed achieves more robust results and ensembling of different metric models like Comet and MetricX outperforms both of them on WMT24 metric shared task.

4.3 Selection

Since different metrics provides quality scores in different ranges, e.g., MetricX outputs [-25,0] while KIWI and Remedy-24 outputs in the range of [0,100]. Therefore, we aggregate rankings from the three QE models and select the candidate with the lowest mean rank as the final translation. This rank-based approach proved more stable than averaging raw scores to avoid the range difference.

Foormally, let $\mathcal{C} = \{c_1, \ldots, c_N\}$ denote the N sampled candidates for a source segment and $\mathcal{M} = \{1, \ldots, M\}$ the set of QE metrics. We denote raw metric scores by $s_m(c)$ and (ascending) ranks by $r_m(c) \in \{1, \ldots, N\}$. Our default selector is the *mean-rank* rule:

$$\bar{r}(c) = \frac{1}{M} \sum_{m \in \mathcal{M}} r_m(c), \quad c^\star = \arg\min_{c \in \mathcal{C}} \bar{r}(c).$$

This avoids scale incompatibilities across s_m and is less brittle to heavy-tailed score distributions than direct averaging of raw scores.

5 Results

5.1 WMT25 AutoRank

Table 2 reports the official preliminary automatic results for Japanese→Chinese. AutoRank is computed by ensembling multiple metrics (KIWI-XL, GEMBA-ESA-CMDA, GEMBA-ESA-GPT-4.1, MetricX-24 Hybrid-XL, and XCOMET-XL). Our system ranks 4th out of 41 valid submissions, outperforming several large closed models (e.g., GPT-4.1, Claude-4, DeepSeek-V3, Mistral-Medium). Because our selection stage employs metrics related to those in AutoRank, some metric coupling is possible (Pombal et al., 2025); we therefore treat absolute deltas with caution and emphasize the human evaluation below.

5.2 WMT25 Human Evaluation

Table 1 presents the final human evaluation results, adopted from the official WMT25 findings (Kocmi et al., 2025a). As shown in the table, our submission system, Kaze-MT ranked 8th, slightly lagging behind the massive closed LLMs like DeepSeek-V3, CommandA-WMT while still outperforming systems like Mistral-Medium. Notably, the human evaluation presents lower ranking compared to the Autorank, presenting the automatic translation metric still presents unaligned preference as humans.

Japanese→Chinese

Rank	System	Human	AutoRank
1-1	**Human**	-3.5	
2-2	Gemini-2.5-Pro	-4.4	3.3
3-6	Algharb	-5.8	4.3
3-7	Claude-4	-5.9	6.4
3-7	Shy-hunyuan-MT	-6.1	1.0
3-7	GPT-4.1	-6.2	4.5
4-7	Wenyiil	-6.9	4.5
8-10	CommandA-WMT	-7.7	5.2
8-10	DeepSeek-V3	-8.1	6.5
8-13	Kaze-MT	-8.6	3.9
10-13	Mistral-Medium✗	-10.0	6.6
10-13	In2x✗	-10.0	3.0
10-13	Qwen3-235B	-10.9	7.6
14-15	GemTrans	-10.9	6.6
14-15	NTTSU	-11.3	5.9
16-17	Yolu	-12.6	7.1
16-17	TowerPlus-9B[M]	-13.3	11.5
18-18	IRB-MT	-13.9	12.4
19-19	Laniqo	-18.3	11.3

Table 1: The official WMT25 Human Evaluation results adopted from Kocmi et al. (2025a). The human score is the micro-average of human judgements across all domains and double annotations. AutoRank is calculated from automatic metrics as per (Kocmi et al., 2025b). Significance testing is done using a Wilcoxon signed rank test with a p-value threshold of 5%. Ranks from row in two directions until they reach a system that is significantly different. Clusters are created such that they do not overlap with ranks. Systems are either constrained (white), or unconstrained (gray). Systems that do not officially support the language pair are marked with ✗.

6 Discussion

6.1 On Metric Bias and Coupling Effects

When the selection ensemble and the official evaluation share metric families, *metric coupling* can inflate automatic rankings. In our case, AutoRank includes metrics related to our selectors (e.g., KIWI-22 and MetricX variants), which may partially explain why our AutoRank position exceeds our human-evaluation cluster. This is a form of

Japanese-Simplified Chinese									
System Name	LP Supported	Params. (B)	Humeval?	AutoRank ↓	Kiwi-XL ↑	GEMBA-ESA-CMDA ↑	GEMBA-ESA-GPT4.1 ↑	MetricX-24-Hybrid-XL ↑	XCOMET-XL ↑
Shy-hunyuan-MT	✓	7	✓	1.0	0.577	85.1	85.5	-4.2	0.629
In2x		72	✓	3.0	0.624	77.0	77.7	-4.7	0.618
Gemini-2.5-Pro	✓		✓	3.2	0.549	84.8	84.8	-4.6	0.596
Kaze-MT	✓	72	✓	3.8	0.569	81.5	81.8	-4.8	0.605
Algharb	✓	14	✓	4.2	0.547	83.5	84.1	-4.8	0.583
GPT-4.1	✓		✓	4.4	0.549	83.8	84.7	-5.1	0.582
Wenyiil	✓	14	✓	4.5	0.555	81.4	81.9	-4.8	0.591
CommandA-WMT	✓	111	✓	5.1	0.558	80.2	79.7	-4.7	0.575
NTTSU	✓	14	✓	5.8	0.563	77.5	74.8	-4.6	0.577
bb88				6.1	0.551	80.1	78.9	-5.2	0.573
Claude-4	✓		✓	6.2	0.545	82.9	83.7	-5.6	0.556
DeepSeek-V3	✓	671	✓	6.3	0.534	82.9	80.9	-5.1	0.552
Mistral-Medium			✓	6.4	0.546	81.1	81.1	-5.4	0.558
GemTrans	✓	27	✓	6.5	0.556	76.0	74.9	-4.8	0.579
Yolu	✓	14	✓	6.9	0.578	74.6	73.6	-5.0	0.565
Qwen3-235B	✓	235	✓	7.5	0.549	78.4	77.0	-5.4	0.555
CommandA	✓	111		7.6	0.54	79.4	77.6	-5.5	0.556
UvA-MT	✓	12		8.3	0.564	73.9	75.2	-5.6	0.561
TowerPlus-72B[M]	✓	72		9.7	0.537	76.5	75.0	-5.9	0.536
AyaExpanse-32B	✓	32		10.7	0.537	73.2	72.0	-5.8	0.521
Laniqo	✓	9	✓	11.1	0.579	63.1	62.1	-5.4	0.557
TowerPlus-9B[M]	✓	9	✓	11.2	0.535	71.9	69.8	-5.8	0.523
IRB-MT	✓	12	✓	12.1	0.521	72.2	70.4	-6.0	0.509
Gemma-3-27B	✓	27		12.8	0.526	70.4	70.2	-6.2	0.503
Llama-4-Maverick	✓	400		13.1	0.524	71.5	66.1	-6.3	0.518
Qwen2.5-7B	✓	7		13.6	0.524	68.9	67.4	-6.3	0.502
IR-MultiagentMT				13.7	0.523	67.8	68.5	-6.2	0.492
SRPOL		12		13.8	0.56	63.8	62.5	-6.4	0.522
EuroLLM-22B-pre.[M]	✓	22		14.7	0.521	66.4	66.2	-6.3	0.486
AyaExpanse-8B	✓	8		15.5	0.518	65.6	64.4	-6.4	0.472
ONLINE-B	✓			16.2	0.499	63.7	63.2	-6.2	0.472
Gemma-3-12B	✓	12		17.1	0.509	65.0	64.1	-7.1	0.465
CommandR7B	✓	7		18.4	0.496	59.8	58.5	-6.9	0.486
TranssionTranslate				18.8	0.488	59.9	60.6	-6.7	0.45
Llama-3.1-8B		8		20.2	0.507	58.8	57.3	-7.2	0.423
EuroLLM-9B[M]	✓	9		20.8	0.479	59.4	57.2	-7.6	0.461
ONLINE-W				25.2	0.456	52.3	52.9	-7.9	0.387
Mistral-7B		7		32.8	0.445	42.9	43.4	-9.8	0.317
SalamandraTA	✓	8		33.1	0.426	36.5	38.0	-8.6	0.328
ONLINE-G	✓			40.8	0.352	39.5	39.8	-12.1	0.28
NLLB	✓	1		41.0	0.371	35.5	35.8	-12.1	0.303

Table 2: The official WMT25 AutoRank results adopted from Kocmi et al. (2025b). Our submission, Kaze-MT ranked 4th out of 41 valid submissions. Note that our submission is based on the best-of-N reranking using KIWI22, MetricX24-QE-XXL, and Remedy24-QE, thus such approach could deliver biased results when using the same model for reranking and evaluation.

evaluation-on-the-features bias: the system is optimized for the very signals (or close proxies) used to score it.

6.2 Potential discrepancy between human and automatic metrics

Another potential reason is the gap between human judgments and current automatic metrics. Most widely used metrics are black-box models: they output a single overall score without exposing intermediate decisions or confidence. Without explanations, these scores can reflect surface cues (e.g., lexical overlap, length) rather than the properties humans care about (translation accuracy, register). They may also miss context-sensitive errors (tone, pragmatics) and discourse links across sentences.

As a result, a system optimized to rank well under such metrics can improve automatic scores without a matching gain in human preference. This points to the need for more interpretable evaluation models.

7 Conclusion

We presented Kaze-MT, a simple yet competitive TTS system for Japanese–Chinese MT. By pairing diverse zero-shot sampling from a strong LLM with robust QE-based selection, we achieve strong results without any fine-tuning or in-domain resources. The modest gap between AutoRank and human ranking of our submission indicates that evaluation-on-the-features bias may exist, and TTS approach largely depends on the quality and robustness of quality estimation metrics. Reduce metric coupling and improving alignment of quality estimation methods with human preferences remains an important future work.

Acknowledgments

This research was funded in part by the Netherlands Organization for Scientific Research (NWO) under project number VI.C.192.080.

References

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. Metricx-24: The google submission to the wmt 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica M. Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinþór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025a. Findings of the wmt25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinþór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025b. Preliminary ranking of wmt25 general machine translation systems.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626.

Ann Lee, Michael Auli, and Marc'Aurelio Ranzato. 2021. Discriminative reranking for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7250–7264.

Tomoya Mizumoto and Yuji Matsumoto. 2016. Discriminative reranking for grammatical error correction with statistical machine translation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1133–1138.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.

Graham Neubig, Makoto Morishita, and Satoshi Nakamura. 2015. Neural reranking improves subjective quality of machine translation: Naist at wat2015. In *Proceedings of the 2nd Workshop on Asian Translation (WAT2015)*, pages 35–41.

José Pombal, Nuno M Guerreiro, Ricardo Rei, and André FT Martins. 2025. Adding chocolate to mint: Mitigating metric interference in machine translation. *arXiv preprint arXiv:2503.08327*.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022a. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

Ricardo Rei, Marcos Treviso, Nuno M Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José GC de Souza, Taisiya Glushkova, Duarte Alves, Luísa Coheur, et al. 2022b. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Shaomu Tan, Ryosuke Mitani, Ritvik Choudhary, and Toshiyuki Sekiya. 2025. Investigating test-time scaling with reranking for machine translation.

Shaomu Tan and Christof Monz. 2025. Remedy: Learning machine translation evaluation from human preferences with reward modeling. *arXiv preprint arXiv:2504.13630*.

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. 2025. Inference scaling laws: An empirical analysis of compute-optimal inference for llm problem-solving. In *The Thirteenth International Conference on Learning Representations*.