

# Lanigo at WMT26 Video Subtitle Translation Shared Task: Multi-Objective Fusion of Pareto-Optimal Candidates

Kamil Guttmann<sup>1,2</sup>, Artur Nowakowski<sup>1,2</sup>

<sup>1</sup> Lanigo, Poznań, Poland

<sup>2</sup> Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poznań, Poland  
{name}. {surname}@lanigo.com

## Abstract

This paper describes Lanigo’s submission to the WMT26 Video Subtitle Translation Shared Task, translating subtitles from Simplified Chinese into English, Thai, Indonesian, Malay, and Traditional Chinese under a 20B-parameter, open-license model constraint. Our system optimizes translation quality and subtitle compliance entirely at inference time, without fine-tuning, by combining hedged multi-prompt candidate generation, language-identification and compliance pruning, multi-objective Pareto reranking, reasoning-based candidate fusion, and a fallback/compression step. We screened four open-weight models and compared the two strongest, Qwen3.5-9B and Gemma-4-12B-it, across the full pipeline on the complete test corpus. Gemma-4-12B-it reached significantly higher compliance rates in every target language, thus it was chosen as our final system.

## 1 Introduction

We describe Lanigo’s submission to the WMT26 Video Subtitle Translation Shared Task. The shared task constrained model size to under 20B parameters in total and required participants to release their models under an open license, but placed no constraints on the methods or data used. We participated in all five translation directions, translating video subtitles from Simplified Chinese into English, Thai, Indonesian, Malay, and Traditional Chinese, across 100 videos of varying length. Each video was accompanied by additional metadata.

Subtitle translation differs from general-domain machine translation (MT) in that focusing on adequacy and fluency alone is not sufficient. Viewers must be able to read each subtitle within the time it is displayed, so translations are subject to a maximum reading speed, measured in characters per second (CPS). Subtitles must also fit on screen, subject to a maximum number of characters per line (CPL) and no more than two lines per cue.

These constraints compete with translation quality: a semantically faithful translation that exceeds the character or line limits is not a usable subtitle, and neither is a translation compressed to the point of losing its meaning. Producing subtitles therefore requires jointly optimizing translation quality and compliance with hard length and display constraints.

We address this joint optimization problem entirely at inference time, on top of an off-the-shelf instruction-tuned Large Language Model (LLM), without any fine-tuning. Our pipeline consists of the following steps:

1. **Multi-prompt candidate generation:** We generate multiple candidate translations per subtitle cue, combining four prompt variants with hedged sampling to obtain a diverse candidate pool.
2. **Compliance pruning:** We discard wrong-language candidates and those that violate the CPS/CPL constraints the most.
3. **Multi-objective Pareto Reranking:** We compute the Pareto non-dominated set of remaining candidates across translation-quality and compliance objectives.
4. **LLM candidate fusion:** A further LLM inference pass selects, rewrites, or merges candidates from the Pareto set into a single translation.
5. **Fallback and compression:** A final pass re-translates or compresses any cue that still fails a hard requirement.

## 2 Related Work

Prior work treats reading-speed and display constraints as objectives to optimize jointly with translation quality rather than applying them post hoc.

Matusov et al. (2019) learn subtitle segmentation from human-authored breaks alongside industry length and duration constraints, while Petrick et al. (2025) integrate length control into the translation model itself and optimize every component of their IWSLT 2025 subtitling track (Abdulmumin et al., 2025) submission against reading-speed and line-length limits.

We instead impose these constraints at decoding time, leaving the model untouched. Minimum Bayes Risk (MBR) decoding selects the candidate that minimizes expected risk with respect to a single utility metric, computed over a set of sampled translations (Kumar and Byrne, 2004). This framing optimizes for one reward at a time. In earlier work, we introduced Pareto-Optimal Decoding (Guttmann et al., 2025), a reranking strategy that instead selects the candidate that jointly satisfies multiple, potentially competing objectives by computing the Pareto-non-dominated set over several metrics. In this paper, we generalize Pareto-Optimal Decoding from balancing translation quality against terminology accuracy to balancing translation quality against other constraints.

Both MBR decoding and reranking benefit from diverse candidate pools, but constructing that diversity at scale is costly. Epsilon sampling (Freitag et al., 2023) produces high-quality diverse candidates, but requires sampling budgets in the hundreds to thousands of candidates per input to be effective. Khairi et al. (2025) show that this budget can be reduced substantially without sacrificing robustness via hedged sampling: combining a small number of temperature-sampled candidates with one greedy decoded candidate protects against the variance introduced by high-temperature sampling, particularly for non-English languages. We adopt hedged sampling and extend it across multiple prompt variants (e.g., with and without preceding context or metadata) as an additional, cheap source of candidate diversity and robustness against noise introduced by this contextual information.

Selecting a single best candidate from a generated pool discards potentially useful information contained in the remaining candidates. Khairi et al. (2026) propose Fusion-of-N (FusioN) as an alternative to Best-of-N (BoN) selection: instead of picking one candidate, an LLM judge synthesizes their complementary strengths into a new output. On WMT24++ (Deutsch et al., 2025), FusioN outperforms BoN across languages, and even surpasses oracle selection in German, Russian, and Chinese.

In machine translation specifically, the Hunyuan-MT-Chimera model (Zheng et al., 2025) adopts the same principle, training a dedicated fusion model to synthesize several translation candidates generated under varying decoding settings into one output. We adopt this method but, unlike Zheng et al. (2025), use the same off-the-shelf model for both candidate generation and fusion additionally conditioning on the candidates’ compliance and quality scores to enforce subtitle-length constraints.

### 3 Approach

#### 3.1 Candidate Generation

For each subtitle cue, we generate a total of 25 candidate translations distributed across four prompt variants: a base sentence-level prompt (V1), a variant that adds the album synopsis<sup>1</sup> as metadata (V2), a variant that adds the preceding cues as context (V3), and a batch variant that translates a document-level window of 5 cues in a single call (V4). Following Khairi et al. (2025), we use hedged sampling to form the 25-candidate pool: 4 candidates come from one greedy decode per prompt variant, while the remaining 21 candidates are sampled across the variants at temperature 0.7 and min-p 0.2. We do not enable reasoning at this stage and reserve it for the candidate fusion step, where it has the most leverage, since applying it to the generation of every candidate would multiply the compute cost drastically.

#### 3.2 Compliance Pruning

High-temperature sampling occasionally yields candidates in the wrong language – typically English – regardless of the requested target. QE metrics cannot catch this: they are not conditioned on the target language and will score a fluent English candidate as highly as, or higher than, the correct-language candidates for the same source segment.

We therefore prune the candidate pool with two filters applied in sequence. The first filter, based on the `lingua-py`<sup>2</sup> library, keeps a candidate only if its target language is among the three top-ranked detected languages with confidence threshold of 0.3. Indonesian and Malay, being mutually intelligible and often confused by the identifier, are scored jointly and pass if either clears the threshold. Candidates under five characters are exempted since

<sup>1</sup>A short description of the video series to which a given episode belongs.

<sup>2</sup><https://github.com/pemistahl/lingua-py>

language identification is unreliable at that length. A length-based filtering step runs second, removing the least CPS/CPL-compliant of the remaining candidates up to 50% of the original candidate set size.

### 3.3 Multi-Objective Reranking

For each cue, we compute the Pareto-non-dominated set of the remaining candidates over four objectives: xCOMET-XL-QE (Guerreiro et al., 2024), MetricX-24-XL-QE (Juraska et al., 2024), and CPS/CPL compliance scores. The two compliance scores are zero for a fully compliant candidate and negative otherwise, in proportion to how far the candidate’s reading speed (CPS) or longest display line (CPL) exceeds the per-language cap.

Averaged over the full test corpus, the non-dominated set contains 1.9 candidates per cue (minimum 1, maximum 11), out of a mean of 13 candidates surviving pruning.

### 3.4 Candidate Fusion

No single candidate in the Pareto set is guaranteed to satisfy both the quality and the compliance objectives. We therefore pass the whole set to the model, annotated with each candidate’s quality and compliance scores, and prompt it to rewrite or merge candidates into one translation that combines their strengths.

Additionally, fusion is applied over a window of 12 consecutive cues in a single prompt, with an instruction to keep the fused translations coherent across the window. To preserve coherence across window boundaries, each prompt also includes the last 3 cues already fused in the previous window as read-only context. Crucially, fusion operates strictly on text assignments within existing cue time slots: the original timecodes are strictly preserved, ensuring no audio-visual timing mismatch for the viewer.

### 3.5 Fallback & Compression

If the fusion output for a cue is empty, contains untranslated source text, or is identified as being in a wrong language, we fall back to the best compliant candidate from the Multi-Objective Reranking step, selected by an average of the four normalized scores. If this still fails, we retry generation: for wrong-language cues, we show the model its own broken translation and ask it to redo the translation in the correct language, while empty cues

are regenerated from scratch with a fresh greedy decode.

A compression pass is applied if the resulting text still exceeds the CPS cap by a margin of 30%. Separately, to satisfy the line-length and line-count constraints, any line still over the per-line cap is split into two based on punctuation or the nearest word boundary to the midpoint.

## 4 Experimental Setup

### 4.1 Data

For preliminary experiments, we used the demo episode released by the task organizers, consisting of 231 cues. The test set consists of 100 videos, 41,837 source segments in total (mean 418.4, min 41, max 2188 segments per video). No reference translations were released to the public. Each video is accompanied by metadata used to identify and locate it on the source streaming platform: a video ID and title, an album ID and title, an album description, internal content identifiers, and overseas/domestic URLs linking to the hosted video on streaming platforms. Among these fields, the album description is the one source of context relevant to translation, so it is the only field we use in our approach.

### 4.2 Models

The shared task’s constrained track required participants to use openly licensed models with fewer than 20B parameters. We screened four models meeting these requirements:

- **Hy-MT2-7B**<sup>3</sup> – a translation-specialized model built on Tencent’s Hunyuan architecture, suggested by the task organizers.
- **Qwen3.5-9B**<sup>4</sup> – a general-purpose multilingual LLM from the Qwen family, with substantial Chinese-language training data.
- **Sailor2-20B**<sup>5</sup> – built on Qwen2.5, developed specifically for Southeast Asian languages, including Thai, Indonesian, and Malay.
- **Gemma-4-12B-it**<sup>6</sup> – Google’s general-purpose instruction-tuned multilingual LLM.

<sup>3</sup><https://huggingface.co/tencent/Hy-MT2-7B>

<sup>4</sup><https://huggingface.co/Qwen/Qwen3.5-9B>

<sup>5</sup><https://huggingface.co/sail/Sailor2-20B>

<sup>6</sup><https://huggingface.co/google/gemma-4-12b-it>

| Language | CPS cap | CPL cap |
|----------|---------|---------|
| EN       | 20      | 42      |
| ID       | 20      | 42      |
| MS       | 20      | 42      |
| TH       | 17      | 38      |
| ZH-TW    | 9       | 16      |

Table 1: Per-language CPS and CPL caps used in this work, maximum 2 lines per cue.

| Model          | CPS%        | CPL%        | QE           | Fail%       |
|----------------|-------------|-------------|--------------|-------------|
| Hy-MT2-7B      | 50.5        | 80.6        | 0.881        | 1.49        |
| Qwen3.5-9B     | <b>66.4</b> | <b>90.8</b> | <b>0.888</b> | 0.16        |
| Sailor2-20B    | 61.4        | 84.9        | 0.887        | 0.32        |
| Gemma-4-12B-it | 60.9        | 88.0        | 0.877        | <b>0.00</b> |

Table 2: Model screening: xCOMET-QE (QE), CPS/CPL compliance, and parse-failure rate, averaged over six prompt configurations and all five target languages (dev episode, 231 cues each). Per-prompt and per-language breakdowns in Appendix A.

All four are used off the shelf as instruction-tuned checkpoints, with no additional fine-tuning.

### 4.3 Evaluation

We report referenceless quality estimation (QE) scores using xCOMET-XL-QE and MetricX-24-XL-QE. Pombal et al. (2025) show that reusing the same metric for both decoding and evaluation can overstate quality gains, so we additionally report scores from ReMedy-9B-22 (Tan and Monz, 2025) as a non-interfering metric. We additionally report compliance with reading-speed and line-length constraints as the percentage of cues within the applicable characters-per-second (CPS) and characters-per-line (CPL) caps, using the per-language caps in Table 1, chosen to reflect common subtitling practice.

## 5 Results

### 5.1 Model screening

We screened all four models on the dev episode across six prompt configurations – V1–V3, and V4 at window sizes 5, 10, and 15 – in all five target languages, measuring xCOMET-QE, CPS and CPL compliance, and response parse failure rate. Averaged results are described in Table 2 while detailed per-prompt and per-language results are given in Appendix A.

Hy-MT2-7B is weakest on both compliance axes and has the highest parse-failure rate. Despite roughly double the parameter count, Sailor2-20B

trails Qwen3.5-9B on both compliance metrics and only matches it on quality. Qwen3.5-9B leads on average compliance and xCOMET-QE, while Gemma-4-12B-it has the lowest average xCOMET-QE of the four but is the only model that never fails to provide a parsable response. We therefore carried Qwen3.5-9B and Gemma-4-12B-it forward.

### 5.2 Multi-Objective MT Fusion

We compared five candidate-selection strategies on top of the same Qwen3.5-9B Pareto-filtered candidate pool: max-xCOMET-QE selection, max-averaged-metrics selection, base document-level fusion with no metric information shown to the model, fusion with metrics, and fusion with metrics plus reasoning (Table 3).

Max-xCOMET-QE selection, by definition, reaches the highest xCOMET-QE of all variants, but only the second-lowest CPS compliance, since selecting purely for translation quality does not account for length restrictions. Max-averaged-metrics selection leads on MetricX-QE – again one of its own selection criteria – and reaches 81.9% CPS compliance, making it the strongest strategy that requires no second inference pass. Base fusion, which sees no scores, has the lowest CPS compliance of the five, suggesting that the model does not infer translation length from the text alone. Including the scores raises CPS compliance to 83.7%, and enabling reasoning to 88.1%.

ReMedy-QE – the only metric used nowhere in the pipeline – orders the fusion variants almost inversely, scoring base fusion highest. A manual check showed that this may be an artifact of per-cue evaluation, as fusion occasionally reorders adjacent cues to translate multi-cue sentences more naturally. Finally, we adopt fusion with metrics and enabled reasoning, trading 0.02 ReMedy-QE for 14 points of CPS compliance.

### 5.3 Final results

Table 4 reports compliance and quality at each pipeline stage. Both models gain most of their CPS compliance between V1-greedy and Fusion, with a smaller further gain at Final, and CPL compliance reaches exactly 100% at Final by construction. The three quality metrics do not agree on the pipeline’s overall effect: xCOMET-QE and MetricX-QE improve from V1 to Final for both models, while ReMedy-QE declines – likely reflecting Fusion’s window-level rewriting, which redis-

| Variant                    | xCOMET-QE | MetricX-QE | ReMedy-QE | CPS%  | CPL%   |
|----------------------------|-----------|------------|-----------|-------|--------|
| Max-xCOMET selection       | 0.9354    | 3.583      | 0.8821    | 79.31 | 100.00 |
| Max-avg-metrics selection  | 0.9193    | 3.257      | 0.8999    | 81.90 | 100.00 |
| Base fusion                | 0.8956    | 3.459      | 0.9085    | 74.03 | 97.06  |
| Fusion + metrics           | 0.9114    | 3.515      | 0.9026    | 83.72 | 100.00 |
| Reasoning fusion + metrics | 0.9184    | 3.595      | 0.8880    | 88.14 | 100.00 |

Table 3: Multi-objective MT fusion ablation: xCOMET-QE, MetricX-QE, ReMedy-QE, CPS, and CPL compliance for each candidate-selection strategy, averaged across all 5 languages (Qwen3.5-9B candidates, dev episode, 231 cues).

| Stage                 | CPS%↑ | CPL%↑  | xCOMET↑ | MetricX↓ | ReMedy↑ |
|-----------------------|-------|--------|---------|----------|---------|
| <b>Gemma-4-12B-it</b> |       |        |         |          |         |
| V1                    | 75.40 | 96.11  | 0.9191  | 2.9705   | 0.9407  |
| Fusion                | 92.73 | 99.69  | 0.9304  | 2.9485   | 0.9330  |
| Final                 | 93.37 | 100.00 | 0.9296  | 2.9669   | 0.9312  |
| <b>Qwen3.5-9B</b>     |       |        |         |          |         |
| V1                    | 65.86 | 92.97  | 0.9173  | 2.9686   | 0.9394  |
| Fusion                | 84.49 | 98.87  | 0.9341  | 2.8209   | 0.9318  |
| Final                 | 85.95 | 100.00 | 0.9321  | 2.8684   | 0.9269  |

Table 4: Compliance and quality at each pipeline stage (V1-greedy, raw Fusion, Final), averaged over the five target languages (full test corpus, 41,837 cues per language).

tributes content across adjacent cues in ways a per-cue reference-free metric penalizes even when the combined meaning is preserved. The metrics do agree on where quality is lost, scoring Fusion above Final, consistent with the fallback and compression stages trading quality for length compliance.

Table 5 compares the two models at the Final stage, with a paired bootstrap significance test on each metric. Gemma-4-12B-it wins CPS compliance significantly in every language (+0.7 to +12.2 points). The three quality metrics disagree with each other: xCOMET-QE and MetricX-QE mostly favor Qwen3.5-9B (both favor Gemma-4-12B-it only on Thai), while ReMedy-QE favors Gemma-4-12B-it on Indonesian and Malay too.

We chose Gemma-4-12B-it for submission because its CPS advantage was significant and consistent across every language, while no quality metric consistently favored either model – though ReMedy-QE, the one quality metric not used anywhere in reranking or fusion and therefore free of metric interference, favored Gemma-4-12B-it on three of the five language pairs.

## 6 Conclusion

We described Laniqo’s submission to the WMT26 Video Subtitle Translation Shared Task, a pipeline

that optimizes translation quality jointly with reading-speed and line-length constraints entirely at inference time, on top of an off-the-shelf instruction-tuned model. Our submitted system, based on Gemma-4-12B-it, reaches 91–100% CPS and 100% CPL compliance across the five target languages, against 69–98% CPS for single-cue greedy decoding.

Two findings generalize beyond this task. First, most off-the-shelf reference-free quality metrics are target-language agnostic and fail to penalize off-target generations, frequently ranking fluent wrong-language outputs above valid translations. Incorporating dedicated language identification as an additional evaluation dimension is therefore essential to complement bare QE metrics. Second, enforcing hard compliance carries a clear quality cost: all three of our quality metrics score the pre-compression output at or above the final output in every language, as post-hoc compression forces a direct trade-off between adequacy and reading-speed limits.

## References

Idris Abdulmumin, Victor Agostinelli, Tanel Alumäe, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Fethi Bougares, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, William Chen,

| Model                        | CPS% $\uparrow$ | xCOMET-QE $\uparrow$ | MetricX-QE $\downarrow$ | ReMedy-QE $\uparrow$ |
|------------------------------|-----------------|----------------------|-------------------------|----------------------|
| <b>English</b>               |                 |                      |                         |                      |
| Gemma-4-12B-it               | <b>92.10*</b>   | 0.9403               | 2.8951                  | 0.9008               |
| Qwen3.5-9B                   | 83.59           | <b>0.9446*</b>       | <b>2.7154*</b>          | <b>0.9102*</b>       |
| <b>Thai</b>                  |                 |                      |                         |                      |
| Gemma-4-12B-it               | <b>90.99*</b>   | <b>0.9170*</b>       | <b>3.0752*</b>          | <b>0.9342*</b>       |
| Qwen3.5-9B                   | 86.72           | 0.9159               | 3.1133                  | 0.9090               |
| <b>Indonesian</b>            |                 |                      |                         |                      |
| Gemma-4-12B-it               | <b>92.28*</b>   | 0.9092               | 2.9999                  | <b>0.9234*</b>       |
| Qwen3.5-9B                   | 80.87           | <b>0.9137*</b>       | <b>2.7887*</b>          | 0.9207               |
| <b>Malay</b>                 |                 |                      |                         |                      |
| Gemma-4-12B-it               | <b>91.64*</b>   | 0.9040               | 3.0144                  | <b>0.9204*</b>       |
| Qwen3.5-9B                   | 79.45           | <b>0.9081*</b>       | <b>2.8809*</b>          | 0.9162               |
| <b>Chinese (Traditional)</b> |                 |                      |                         |                      |
| Gemma-4-12B-it               | <b>99.83*</b>   | 0.9774               | 2.8500                  | 0.9774               |
| Qwen3.5-9B                   | 99.14           | <b>0.9782*</b>       | <b>2.8440</b>           | <b>0.9785*</b>       |

Table 5: Final-stage comparison between Gemma-4-12B-it and Qwen3.5-9B, full test corpus (41,837 cues/language). Bold marks the better score per metric per language; \* marks a statistically significant difference (paired bootstrap, 10,000 resamples, 95% CI excludes 0). CPL% is omitted here since both systems reach exactly 100.00 in every language.

Raj Dabre, Yannick Estève, Marcello Federico, Mark Fishel, Marco Gaido, Dávid Javorský, Marek Kasztelnik, and 33 others. 2025. [Findings of the IWSLT 2025 evaluation campaign](#). In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 412–481, Vienna, Austria (in-person and online). Association for Computational Linguistics.

Daniel Deutsch, Eleftheria Briakou, Isaac Rayburn Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. [WMT24++: Expanding the language coverage of WMT24 to 55 languages & dialects](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12257–12284, Vienna, Austria. Association for Computational Linguistics.

Markus Freitag, Behrooz Ghorbani, and Patrick Fernandes. 2023. [Epsilon sampling rocks: Investigating sampling strategies for minimum Bayes risk decoding for machine translation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9198–9209, Singapore. Association for Computational Linguistics.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. [xCOMET: Transparent machine translation evaluation through fine-grained error detection](#). *Transactions of the Association for Computational Linguistics*, 12:979–995.

Kamil Guttman, Adrian Charkiewicz, Zofia Rostek, Mikołaj Pokrywka, and Artur Nowakowski. 2025.

[Lanigo at WMT25 terminology translation task: A multi-objective reranking strategy for terminology-aware translation via Pareto-optimal decoding](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 1276–1283, Suzhou, China. Association for Computational Linguistics.

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, Miami, Florida, USA. Association for Computational Linguistics.

Ammar Khairi, Daniel D’souza, Marzieh Fadaee, and Julia Kreutzer. 2026. [Making, not taking, the best of n](#). In *The Fourteenth International Conference on Learning Representations*.

Ammar Khairi, Daniel D’souza, Ye Shen, Julia Kreutzer, and Sara Hooker. 2025. [When life gives you samples: The benefits of scaling up inference compute for multilingual LLMs](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27559–27583, Suzhou, China. Association for Computational Linguistics.

Shankar Kumar and William Byrne. 2004. [Minimum Bayes-risk decoding for statistical machine translation](#). In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176, Boston, Massachusetts, USA. Association for Computational Linguistics.

Evgeny Matusov, Patrick Wilken, and Yota Georgakopoulou. 2019. [Customizing neural machine](#)

translation for subtitling. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 82–93, Florence, Italy. Association for Computational Linguistics.

Frithjof Petrick, Patrick Wilken, Evgeny Matusov, Nahuel Unai Roselló Beneítez, and Sarah Beranek. 2025. [AppTek’s automatic speech translation: Generating accurate and well-readable subtitles](#). In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 222–231, Vienna, Austria (in-person and online). Association for Computational Linguistics.

José Pombal, Nuno M. Guerreiro, Ricardo Rei, and André F. T. Martins. 2025. [Adding chocolate to mint: Mitigating metric interference in machine translation](#). *Transactions of the Association for Computational Linguistics*, 13:1319–1339.

Shaomu Tan and Christof Monz. 2025. [ReMedy: Learning machine translation evaluation from human preferences with reward modeling](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4370–4387, Suzhou, China. Association for Computational Linguistics.

Mao Zheng, Zheng Li, Bingxin Qu, Mingyang Song, Yang Du, Mingrui Sun, and Di Wang. 2025. [Hunyuan-mt technical report](#). *Preprint*, arXiv:2509.05209.

## A Model Screening

Tables 6 and 7 report the full model-screening results, broken down by prompt configuration and by target language.

| Variant          | Hy-MT2    |           |              | Qwen3.5   |           |              | Sailor2   |           |              | Gemma-4   |           |              |
|------------------|-----------|-----------|--------------|-----------|-----------|--------------|-----------|-----------|--------------|-----------|-----------|--------------|
|                  | CPS%      | CPL%      | QE           | CPS%      | CPL%      | QE           | CPS%      | CPL%      | QE           | CPS%      | CPL%      | QE           |
| V1               | 53        | 91        | 0.886        | 65        | 90        | 0.883        | 69        | 91        | 0.888        | 77        | 98        | 0.884        |
| V2               | 48        | 81        | 0.881        | 69        | 93        | 0.880        | 70        | 93        | 0.891        | 73        | 98        | 0.883        |
| V3               | 38        | 82        | 0.861        | 65        | 94        | 0.877        | 70        | 95        | 0.872        | 70        | 97        | 0.884        |
| V4 <sub>5</sub>  | 48        | 72        | 0.888        | 64        | 88        | 0.894        | 47        | 75        | 0.886        | 49        | 80        | 0.874        |
| V4 <sub>10</sub> | 52        | 75        | 0.880        | 66        | 90        | 0.899        | 55        | 76        | 0.896        | 48        | 78        | 0.870        |
| V4 <sub>15</sub> | 65        | 82        | 0.888        | 70        | 90        | 0.892        | 58        | 79        | 0.887        | 48        | 78        | 0.868        |
| <b>Avg</b>       | <b>50</b> | <b>81</b> | <b>0.881</b> | <b>66</b> | <b>91</b> | <b>0.888</b> | <b>61</b> | <b>85</b> | <b>0.887</b> | <b>61</b> | <b>88</b> | <b>0.877</b> |

Table 6: Model screening by prompt variant: CPS%, CPL%, and xCOMET-QE (QE), averaged across all 5 languages (dev episode, 231 cues). V4 subscripts denote window size.

| Lang       | Hy-MT2    |           |              | Qwen3.5   |           |              | Sailor2   |           |              | Gemma-4   |           |              |
|------------|-----------|-----------|--------------|-----------|-----------|--------------|-----------|-----------|--------------|-----------|-----------|--------------|
|            | CPS%      | CPL%      | QE           | CPS%      | CPL%      | QE           | CPS%      | CPL%      | QE           | CPS%      | CPL%      | QE           |
| EN         | 51        | 82        | 0.897        | 57        | 86        | 0.897        | 59        | 88        | 0.892        | 64        | 96        | 0.898        |
| TH         | 51        | 87        | 0.869        | 70        | 95        | 0.872        | 71        | 95        | 0.877        | 68        | 96        | 0.854        |
| ID         | 42        | 78        | 0.857        | 54        | 88        | 0.851        | 56        | 87        | 0.866        | 62        | 95        | 0.852        |
| MS         | 45        | 82        | 0.863        | 53        | 87        | 0.855        | 60        | 89        | 0.867        | 55        | 93        | 0.849        |
| ZH-TW      | 64        | 74        | 0.918        | 98        | 98        | 0.963        | 61        | 66        | 0.931        | 54        | 61        | 0.932        |
| <b>Avg</b> | <b>50</b> | <b>81</b> | <b>0.881</b> | <b>66</b> | <b>91</b> | <b>0.888</b> | <b>61</b> | <b>85</b> | <b>0.887</b> | <b>61</b> | <b>88</b> | <b>0.877</b> |

Table 7: Model screening by target language: CPS%, CPL%, and xCOMET-QE (QE), averaged across all 6 prompt variants (dev episode, 231 cues).