

MakotoAI at WMT26: Per-Pair Model Selection with Deterministic Output Normalization for the General Machine Translation Task

Chris Nishihama
Independent Researcher
chris248ma@gmail.com

Abstract

We describe MakotoAI, an unconstrained submission to the WMT26 General Machine Translation shared task covering all 23 language pairs (4,897 documents). The system performs no training or fine-tuning: it is a lightweight orchestration layer over six off-the-shelf commercial LLMs, in which each language pair is routed to the single model that won a blinded pre-competition shootout for that pair. Translation is document-level with the full task instruction provided in-prompt. Structured output (JSON and HTML domains) is repaired by a small deterministic normalization pass rather than by model-based retry: an ablation on real test data found that a post-hoc validate-and-retry instruction layer was inert on clean documents and net-negative on JSON documents, while deterministic fence and preamble stripping achieved the intended effect at zero inference cost. Quality assurance combined an exhaustive deterministic instruction-adherence scan, a two-judge LLM evaluation whose self-grading bias we measured to be directional by model family, and the organizers’ official alignment checker, against which the final submission scores 1,914 of 1,914 structurally checkable documents aligned. Four defective documents (three truncations, one repetition-loop degeneration) were detected and regenerated, one requiring escalation to a different system model; all four are disclosed. The entire system runs on commodity hardware; total API cost to produce and verify the submission was just under \$25. We argue that for API-orchestrated MT, verification discipline and deterministic post-processing are higher-leverage than added inference-time machinery.

1 Introduction

Modern LLMs translate well individually, but no single model is best across all language pairs, and API-based systems introduce failure modes—output-format drift, degenerate generation, structural misalignment—that sit outside the model’s

translation quality per se. MakotoAI’s WMT26 submission takes a deliberately minimal position: select the best available model per pair empirically, interfere with generation as little as possible, and concentrate effort on verification and deterministic repair of the output.

This is also a study in what a single independent participant can field. The submission was produced by one person on commodity hardware with no local inference and no training, competing in the unconstrained track against systems built on the same frontier models by their own vendors. The methodological question the paper documents is therefore not “how to train a better model” but “how to select, drive, verify, and repair existing models so that the delivered artifact is free of mechanical defects.” By *defect* we mean a discrete, machine-detectable failure of the deliverable—truncation, degenerate generation, format breakage, structural misalignment, instruction violation—as distinct from translation errors (a continuous quality attribute) and hallucinations (unsourced content), which verification can bound and sample but cannot certify absent. We claim that the submitted system contains none of the former; for the latter we claim only measured estimates.

Our contributions: (1) a per-pair model-selection protocol based on blinded two-judge shootouts; (2) evidence, from an ablation on real WMT26 test data, that a validate-and-retry instruction-enforcement layer is dominated by deterministic output normalization; (3) a measured, directional characterization of LLM self-grading bias relevant to anyone using production models as judges; and (4) a verification pipeline (deterministic scans plus the official alignment checker) that took the submission from four latent content defects and seven structural misalignments to a fully aligned state with zero remaining mechanical defects before the deadline.

2 System

2.1 Per-pair model selection

The routing table maps each of the 23 pairs to one production model. Candidates were drawn from six commercial LLMs—Claude Opus 4.6 (Anthropic), GPT-5.4 and GPT-5.5 (OpenAI), Gemini 2.5 Flash and Gemma-4-31B-it (Google), and Qwen3.7-Max (Alibaba)—all accessed through a single API aggregator (OpenRouter), which is also the metering source for the cost figures in §7. Selection was empirical: for each pair, candidate models translated a common evaluation set and were compared in blinded pairwise A/B judgments by two independent judges, with the winner locked into the routing table.

The resulting assignment maps the 23 directions across the six models, with one model carrying the largest share (nine directions) and the remainder distributed among the other five. We do not disclose the per-pair assignment or per-pair development scores. The mapping is the distilled output of the selection methodology and constitutes the system’s commercial know-how; publishing it would allow exact replication of the submission at negligible cost, while the official per-pair results appear in the findings paper regardless (Kocmi et al., 2026). We instead characterize the selection statistically, which we consider more informative than the mapping itself.

First, selection margins are mostly within noise. Of the 20 directions with head-to-head shootout data, only two winners were separated at the confidence-interval level (+9.34 and +5.78 chrF; Popović, 2015, computed with sacreBLEU, Post, 2018); the remaining margins overlap. Per-pair selection is therefore a pragmatic tie-break among strong candidates, not a claim that one frontier model dominates another. One direction fields its shootout runner-up (−2.52 chrF) because the nominal winner was unavailable through our API provider at production time; two directions were selected from provisional $N=50$ shootouts whose planned re-confirmation on the released test data was not recorded before the production run.

Second, development data can lie. During selection we found the English–Northern Sámi development set contaminated: roughly half of it was biblical text present in the models’ training data, so the measured 52.22 chrF partly reflected memorization rather than translation. A same-day re-evaluation on real-domain government text ($N=80$) scored

the same winning model at 32.58—about twenty points lower—without changing the routing decision. We report both numbers as a caution against trusting development figures in low-resource pairs.

Selection was locked on 14 June 2026. No consensus, ensembling, or reranking was applied at inference time: one pair, one model, one generation. The broader MakotoAI platform includes terminology-enforcement components, but the WMT26 task supplies no glossaries, so no terminology machinery was active in this run, and no proprietary or client data of any kind entered the pipeline.

2.2 Translation pipeline

Translation is document-level (the blind set default). For every domain, the full task instruction is prepended verbatim to the prompt, and the model output is taken as the hypothesis. There is no fine-tuning, no adapters, no retrieval, and no translation memory in this run; models are used exactly as served by their APIs. Every specialist call uses temperature 0.3 and a 32,000-token output ceiling; all other sampling parameters are left at the provider default, with no streaming and no reasoning modes. A 953k-pair translation-memory cache present in the codebase recorded zero hits during production (verified from the run logs), so every document was generated by its routed model.

2.3 Deterministic output normalization

Two systematic output-format artifacts appeared in the raw generations: (a) 112 documents in the JSON (software) domain arrived wrapped in Markdown code fences, and (b) 2 documents echoed a leading “Translate into X:” instruction line. Both are format drift, not translation errors. A small deterministic `normalize_output()` pass strips a single wrapping code fence and a leading instruction-echo line. After normalization, the submission contains zero fenced hypotheses, and the organizers’ checker parses every JSON document (the 11 documents whose JSON carries a trailing comma mirror a trailing comma present in the source, i.e., they are faithful).

2.4 Structural alignment

WMT26 aligns HTML documents by paragraph (`<p>`) count and JSON documents by key–value count; 1,914 of our 4,897 documents carry such structure and are therefore checkable, the remaining 2,983 being plain-text doc-

uments for which the checker reports no structural verdict. Running the organizers’ corrected `genmt_check_alignment.py` against our normalized submission initially reported 1,907 of 1,914 aligned; all 7 failures were HTML paragraph-count drift caused by models reshaping `<br>/<p>` boundaries. Six of the seven carried complete, correct text and were repaired deterministically by re-flowing the existing translation into the source’s paragraph structure, under the invariant that the tag-stripped text be byte-identical before and after repair—no translated character changed. The seventh is the English–Thai document discussed in §3.3, which required regeneration rather than re-flow. The final submission scores 1,914 checked / 1,914 aligned / 0 misaligned under the official script.

3 Quality assurance

3.1 Deterministic instruction-adherence scan

Every document (4,897) was checked mechanically against the task instructions: translation-only output (no commentary or preamble), preservation of `{{placeholders}}`, JSON keys, URLs and `@-` handles, and HTML tag fidelity. Adherence was effectively perfect: 100% translation-only, ~100% placeholder/key preservation, with only cosmetic emphasis-tag drift in a handful of documents. Notably, this adherence was achieved with the instruction supplied purely in-prompt (see §4).

3.2 Two-judge quality evaluation and self-grading bias

Twelve documents per pair were scored for accuracy and fluency by (a) the pair’s own production model and (b) an independent Claude cross-read of the identical documents. Because judge (a) grades its own output, we compared the two columns and found the self-grading bias to be real and directional by family: GPT-5.4 graded its own output harshly relative to the independent judge, while Gemma/Gemini graded leniently. Reading both judges together brackets the truth; we suggest practitioners using production models as judges measure this direction rather than assume a uniform inflation. Most pairs scored in the 4.3–4.9 band; a weak tier (English into Ladin, Northern Sámi, and Ligurian) at 3.2–3.9 exhibited language-purity intrusions from dominant neighbours (Italian, Finnish), which we declared honestly in the model card (§5).

3.3 Defect detection and targeted regeneration

Exhaustive scanning surfaced four content defects in 4,897 documents (0.08%), all of which we disclose:

1. **English–Japanese, one document** (24368-jpn_Jpan): near-total truncation—the hypothesis was 8 characters against a source of roughly 900. Regenerated with the pair’s own routed model.
2. **English–Belarusian, one document** (00706-bel_Cyrl): near-total truncation—a 2-character hypothesis. Regenerated with the pair’s own routed model.
3. **English–Belarusian, one document** (00770-bel_Cyrl): truncated to roughly 20% of the source. Regenerated with the pair’s own routed model; the repair was verified for paragraph count, length ratio, and entity preservation.
4. **English–Thai, one document** (45038-tha_Thai): degenerate repetition-loop output that also dropped a source paragraph, and the seventh alignment failure of §2.4. Regeneration with the pair’s production model reproduced the loop (worse), so this single document was regenerated with a different model (Claude Opus 4.6), verified paragraph-by-paragraph against the source. This is the only document in the submission not produced by its pair’s routed model, and it is disclosed in our model card.

The first two defects were caught in the first verification pass over the assembled submission; the third and fourth in later passes, together with the alignment repairs of §2.4. Taken end to end, exactly 10 of the 4,897 documents differ between the raw production output and the submitted file—the four regenerations above and the six deterministic paragraph re-flows of §2.4—and the remaining 4,887 are byte-identical. A re-run of every deterministic check after the final repair reported zero broken outputs.

We note that a defect count is only as trustworthy as the baseline it is measured from. Our first internal statement of this figure was taken against an

intermediate snapshot rather than the raw production output, and consequently undercounted; the numbers above are measured against the generated output itself.

4 Ablation: validate-and-retry versus deterministic normalization

The system includes an optional instruction-enforcement layer: post-generation validation of checkable constraints (translation-only, JSON validity, HTML tag balance) with up to three critique-guided retries on failure. A standing design question was whether to enable it for the submission. We ran the gate on real WMT26 test data: 24 documents spanning clean prose, HTML domains, and the JSON domain, generated with the layer on, then compared pairwise-blinded against the layer-off (normalized) outputs by the same two-judge method.

The result was decisive in both directions of interest. On the 14 clean documents the layer was inert—one attempt, zero spurious retries, no degradation. On all 10 JSON documents the layer’s retry loop fired, failed all three attempts (the model faithfully re-emits a code fence that mirrors the source formatting, which strict validation rejects), tripled the inference cost, and still delivered fenced output—strictly worse than the deterministic normalization already in place, which repairs the same defect at zero cost. The independent judge preferred the layer-off version on 9 of 10 JSON documents. We therefore shipped with the layer off. The general lesson we draw: when an output defect is deterministic in form (fences, echoes, tag drift), repair it deterministically; reserve model-based retry for defects that are actually semantic.

5 Language coverage

We submitted all 23 pairs but focused engineering effort (model shootouts, alignment repair, QA depth) on 17 target languages, declared as primary. Three micro-languages—Ladin, Ligurian, and Northern Sámi—were submitted essentially zero-shot for coverage and declared contrastive (not focused on) in the model card, consistent with their weaker measured quality (§3.2).

6 Results

Internal evaluation (two-judge, §3.2) and the official alignment check (1,914/1,914) are reported above. Official human-evaluation results were

Direction	Docs	chrF
Czech–German [†]	315	79.95
English–Egyptian Arabic	198	75.27
Czech–Vietnamese [†]	315	72.48
English–German	198	66.33
English–Estonian	198	66.12
English–Icelandic	198	65.42
English–Czech	198	62.93
English–Northern Sámi [‡]	198	58.04
English–Ligurian [‡]	198	53.20
English–Ladin [‡]	198	52.19
English–Japanese	198	52.01
Chinese–Japanese	190	42.36
Macro-average	2,602	62.19

Table 1: Post-hoc chrF of the submitted output against the references released by the organizers, for the 12 of 23 directions that have them. [†]two references, scored multi-reference. [‡]contrastive micro-language (§5). Not a ranking; the official measure for this task is human evaluation.

not available at the time of writing and will appear in the WMT26 findings paper (Kocmi et al., 2026). Per-pair development figures remain withheld (§2.1).

After the shared task released reference translations for part of the test set, we scored our submitted output against them with chrF (Popović, 2015), computed with sacreBLEU (Post, 2018); Table 1 reports the result. These are *post-hoc* measurements: the references were published after the submission deadline and played no part in model selection, generation, or any repair described above. References exist for 12 of the 23 directions (2,602 of 4,897 documents); the remaining 11 have none. Czech–German and Czech–Vietnamese carry two references each and are scored multi-reference.

We report these for transparency, not as a ranking, and make no comparison against other participants’ systems. One observation is worth recording. The three contrastive micro-languages (§5) score 52–58 chrF, overlapping the band of several focused directions, yet our two-judge evaluation placed them in a distinctly weaker tier (3.2–3.9 of 5, §3.2). We read this as chrF rewarding character overlap with a dominant related language: output drifting from Ligurian or Ladin toward Italian retains substantial character n-gram agreement with the reference while reading poorly to a speaker. Where the failure mode is language-purity drift toward a close neighbour, character-level metrics are a weak quality signal, and the divergence between the two measurements is itself the finding.

Relatedly, English–Northern Sámi scores 58.04 here against 52.22 on the contaminated development set and 32.58 on real-domain government text (§2.1); three measurements of one direction spanning twenty-six chrF points is a further caution against trusting any single development figure in low-resource pairs.

7 Cost and reproducibility

The submission was produced by a single participant on commodity hardware with no local inference. The full 23-pair production run took 49 minutes of wall-clock time and cost \$22.12, measured as the provider balance delta across the run (\$81.82 to \$59.70). The run executed as two invocations five minutes apart: the first completed two pairs (\$2.62) before being stopped and restarted with the spend cap raised from \$35 to \$50; the second completed the remaining 21 pairs (\$19.50) in 44 minutes, skipping the two already-finished pairs. We report the whole-run figure rather than the second invocation’s self-measured \$19.50, which excludes the first two pairs.

Generation is one API call per document (4,897 calls), with no batching; the only additional generation was the four targeted regenerations of §3.3 and the deterministic re-flows of §2.4, which required no inference. The quality-assurance passes added approximately \$2.58 (two-judge evaluation \$1.68; instruction-layer ablation \$0.80; targeted regenerations \$0.10), for an end-to-end total just under \$25. Same-day calibration and selection batches run before the production window cost roughly \$25 more and are excluded from these figures as development rather than production spend. We note the total not as a benchmark but as an existence proof: frontier-model orchestration with serious verification is within reach of independent researchers.

8 Conclusion

MakotoAI’s WMT26 submission is an argument for minimalism plus discipline: empirically routed single models, untouched generation, deterministic repair, and verification run to exhaustion. The two findings we believe travel beyond this submission are the ablation of §4—deterministic normalization dominated a validate-and-retry layer on real data—and the directional self-grading bias of §3.2. Both are cheap to measure and directly actionable for any API-orchestrated MT system.

Limitations

The system’s quality ceiling is that of the underlying commercial models, which we neither control nor can fully characterize (parameter counts undisclosed). All six models were reached through a single API aggregator, so provider-side routing or version drift is a source of variance we cannot observe. Per-pair selection was performed once before the test period; model updates after selection could shift the optimum. The two-judge evaluation uses LLMs as judges and, despite the cross-read, remains weakest exactly where the translations are weakest (low-resource pairs). One document was produced by a non-routed model (§3.3), disclosed. Results on the three contrastive micro-languages should not be read as focused attempts. Throughout, “defect-free” refers strictly to the mechanical class defined in §1: the submission certainly contains translation errors of the ordinary kind, and we make no claim of hallucination-freedom—semantic fidelity was sampled and judged, not certified.

Acknowledgments

We thank the WMT26 organizers for the shared task, for the corrected alignment tooling released during the evaluation period, and for their responsiveness on the task mailing list. System development made extensive use of AI-assisted engineering: large language models served as coding, analysis, and drafting assistants throughout the build and in the preparation of this paper, under the verification discipline the paper describes. All design decisions, all disclosures, and the final submission are the author’s, who reviewed and takes responsibility for every claim.

References

- Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Sergey Dukanov, Anton Dvorkovich, Mark Fishel, Markus Freitag, Samuel Frontull, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Stan Kharevich, Philipp Koehn, Zheng Li, Jean Maillard, Christof Monz, Alexander Murauski, and 12 others. 2026. Findings of the WMT26 general machine translation shared task: Contrastive dynamic human evaluation at scale. In *Proceedings of the Eleventh Conference on Machine Translation*. Association for Computational Linguistics.
- Maja Popović. 2015. chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the*

Tenth Workshop on Statistical Machine Translation,
pages 392–395.

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