

Algebras FluencyScore at WMT26 Automated Translation Quality Evaluation

Airana Mongush

Algebras

aira@algebras.ai

Dmitrii Pukhov

Algebras

dima@algebras.ai

Abstract

We describe the Algebras **FluencyScore** submissions to the WMT26 Automated Translation Quality Evaluation shared task (Tasks 1–2) (Lavie et al., 2026). FluencyScore is a reference-free structured LLM judge: five fluency sub-dimensions, a fixed-weight aggregate, and a diagnostic error record (issue type, severity, problem phrase) (Mongush and Pukhov, 2026). This system paper adapts that judge, rather than replacing it. For Task 1 the diagnostic record is specialized into a multi-error character-span annotator with deterministic phrase-to-offset alignment, merge / whole-segment post-processing, and a conservative add-only overlay. For Task 2 the calibrated scalar (`fluency201`) is mapped onto the cESA scale with a WMT25-fitted affine adapter and moment matching (`aira_mo`). Submitted systems use Gemini 3.5 Flash and cover all 21 official pairs plus challenge sets. As of 9 September 2026, organizers have not released official human-cESA rankings for this task; we therefore report no ranks and no Codabench verification (pseudo-gold) scores.

1 Introduction

The WMT26 Automated Translation Quality Evaluation task asks systems to (1) annotate character-level error spans with severity, (2) predict segment-level cESA scores in $[0, 100]$, and (3) detect error-free segments (Lavie et al., 2026). Human gold for ranking is the General MT cESA protocol (Kocmi et al., 2024). We participate in Tasks 1–2 with **FluencyScore**, the structured LLM fluency judge introduced in the companion study (Mongush and Pukhov, 2026).¹

The companion paper evaluates FluencyScore as a scalar against human ESA/MQM and documents the diagnostic error record. It does not pro-

duce WMT26 Task 1 spans or Task 2 cESA files. This paper is the adaptation:

1. **Task 1.** Expand the single-issue error record into $0..N$ major/minor spans; align phrases to character offsets; apply deterministic post-processing; optionally add a second conservative overlay pass.
2. **Task 2.** Keep the FluencyScore-cal map, then apply a WMT26-only affine adapter and moment match fitted on WMT25 human ESA (not on WMT26 test data).

On a fair WMT25 ESA mask the research judge is statistically tied *in Spearman ρ* with matched GEMBA-ESA on the same Gemini backbone (§2). That result motivates treating FluencyScore as a quality signal to adapt, not as a shared-task score by itself.

Findings (for organizers). Algebras (`aira_mo`) submitted two Task 1 systems and one Task 2 system: a multi-error span annotator, a conservative add-only overlay on empty hypotheses, and a cESA-scale score file. All three are reference-free Gemini 3.5 Flash judges, not fine-tuned and not ensembles, and cover the full official 21-pair set plus challenge sets on `mteval-test26`. We did not train a dedicated Task 3 model; the Task 2 upload name encodes an inclusive threshold of 70 for organizer auto-conversion only. Table 1 lists the files.

Constraints: reference-free (source + hypothesis only); prompt and post-process, no fine-tuning; Gemini 3.5 Flash via Google AI Studio for submitted systems (research backbone: Gemini 3 Flash; §2).

2 Background: FluencyScore Judge

Structured FluencyScore. The judge scores five fluency sub-dimensions on a 1–10 scale—

¹Organization / model card: **Algebras**. Submission IDs in Table 1.

Task	Submission	ID
1 (dev.)	fluencyscore-algebras-spans	870122
1	fluencyscore-multispan-v2	871898
1	fluencyscore-overlay	873483
2	fluencyscore-algebras-esa-t70	872328

Table 1: Submitted systems. Task 2 public name `aira_mo`. IDs are Codabench submission identifiers, not ranks.

idiomatic, collocational, discourse, pragmatic, and calque—and aggregates them with *fixed* weights into one scalar (Mongush and Pukhov, 2026). The judge prompt and sub-dimension weights are proprietary; the public claim is the framework (schema, taxonomy, evaluation protocol). For every segment the judge also returns a diagnostic error record: `issue_type` $\in$ {`calque`, `register_mismatch`, `false_friend`, `idiom_literal`, `collocation`, `discourse_marker`, `none`}, severity, problem phrase, and a suggested fix. The judge is reference-free (source + MT only).

FluencyScore-cal. FluencyScore-cal (deployment name `fluency201`) is a fixed display map from the raw 1–10 aggregate onto [1,9]. It is continuous and non-decreasing, so rank statistics (Spearman, Kendall, pairwise accuracy) are unchanged except at ties. Parameters were frozen in development and are not refit on WMT26 test data. The map is

$$f(x) = 1 + ((\text{clip}(x, 1, 10) - 1)/9)^{1.8} \cdot 8,$$

rounded to two decimals ($\gamma=1.8$, $\text{cap}=9$). Example: $f(8.0)=6.09$; $f(10.0)=9.00$. Compression reduces saturation at the top of the raw 1–10 scale. This map is part of the FluencyScore framework, not a WMT26 adapter.

Companion protocol (WMT25, not WMT26).

Research runs average $k=2$ independent calls; a single-pass control changes ρ by only +0.007 (CI covers zero). On a fair WMT25 ESA mask ($n=3526$), FluencyScore with Gemini 3 Flash reaches pooled Spearman $\rho=0.613$ and pairwise accuracy 0.722, statistically tied in ρ with matched GEMBA-ESA on the same backbone ($\Delta\rho=+0.007$, CI covering zero; matched GEMBA is slightly ahead on pairwise accuracy). Register/dialect mismatches and fluent-but-wrong hypotheses remain known failure modes. WMT26

submissions use a single Gemini 3.5 Flash pass for cost, not the Gemini 3 Flash research backbone.²

WMT26 outputs. Task 1 requires 0.. N character spans; Task 2 requires one cESA-scale number in [0,100]. Submitted systems are closed-source, not fine-tuned, not ensembles, and reference-free.

3 Task 1: Error Spans

Official Task 1 scoring is severity-weighted micro MPP F1 (Perrella et al., 2026).³ MPP is a maximum one-to-one matching of predicted spans to gold spans, with partial overlap credit. Each match is weighted by severity (minor=0.25, major=1.0). The official number is the F1 of that weighted overlap, pooled over language pairs, against *human* cESA spans. Codabench test-period scores instead compare systems to an automatic pseudo-gold; we omit those numbers.

A single predicted span can recover at most one gold span. If a hypothesis contains several independent errors, a one-phrase judge is structurally recall-limited under MPP, regardless of alignment quality. That limitation, not a claim about human-gold accuracy, is why we moved from a single-span baseline to a multi-error prompt.

3.1 v1 — Single-span baseline

The research FluencyScore record names *one* problem phrase per segment. v1 maps that phrase to a single character interval [*start*, *end*) in the hypothesis by string search, copies the severity, and writes an empty list if the phrase cannot be found.

Example. Hypothesis: *I make a photo of the cathedral*. FluencyScore returns problem phrase *make a photo*, severity `minor`. v1 emits one span covering *make a photo*. A second error in the same sentence (e.g., a calque on *cathedral*) cannot be emitted.

Development system: `fluencyscore-algebras-spans` (ID 870122).

3.2 v2 — Multi-error judge

We keep FluencyScore as the judge family and change only the diagnostic channel: a multi-error prompt that returns 0.. N spans

²Prose names: Gemini 3.5 Flash (submissions) and Gemini 3 Flash (companion research primary). API strings in our logs are `gemini-3.5-flash` and `gemini-3-flash-preview`; companion tables also write `gemini-3-flash` for that research primary.

³Official WMT26 implementation: sted19/span-mt-metrics-eval.

(`problematic_phrase`, `occurrence`, `severity`, `category`, `confidence`), plus omission and garbled (Appendix A). Policies, checked against the official scorer:

- **Prefer minor when unsure.** A false-positive major costs 4× a false-positive minor in weighted precision (weights 1.0 vs. 0.25). Example: *make a photo* may be a collocation error or acceptable informal English—label `minor`.
- **Empty lists are valid.** If the hypothesis looks fine, return no spans.
- **Whole-segment major only if garbled** (nonsense or wrong language). Omissions are not spans: missing text has no hypothesis offsets.
- **Language-pair / domain notes** in the prompt: CJK over-flagging; under-generation on low-resource targets; WMT26 domains (news, social, speech, software, education).

Phrase-to-offset alignment. The judge returns a string, not indices. We locate `problematic_phrase` in the hypothesis: (1) **exact** substring search; (2) **normalized** NFKC + casefold + punctuation, then map indices back (typical path for CJK / Arabic / Thai); (3) **bounded fuzzy** edit distance in a window (typos, hyphenation); disabled for forced whole-segment majors in the submitted systems. The `occurrence` index selects among repeated phrases. Claimed intervals prevent two errors from collapsing onto the same offsets. Unit tests cover CJK, Thai, Arabic, curly quotes, and NBSP.

Post-processing. Merge overlapping spans (union, max severity) and adjacent same-severity spans within a 1-character gap (splitting one gold span in half scores $F=0.5$). Drop spans covering $\geq 90\%$ of the MT unless garbled. Cap at six spans by severity mass. Forced fuzzy for unlocatable majors is off in the submitted systems. API or safety failures fall back to the v1 span for that key. Official and challenge slots are filled.

Output profile (not an evaluation metric). Table 2 describes *our* Task 1 files on official hypotheses: how many spans we emit, not agreement with human gold. Denser output is a prerequisite for MPP recall; it is not evidence of higher Task 1

System	Spans/hyp.	Empty%	Mass
v1 single-span	0.67	32.9	0.48
v2 multi-span	1.71	26.3	0.86
overlay on v2	+2583 spans	—	—

Table 2: Descriptive output statistics on 252,923 official hypotheses (our files). Mass is severity-weighted span mass per hypothesis (`minor=0.25`, `major=1`). Not human-gold MPP, not Codabench verification scores.

accuracy. On 252,923 official hypotheses, v2 averages 1.71 errors per hypothesis, 26.3% empty, severity mass 0.86 per hypothesis (v1: 0.67 errors, 32.9% empty, mass 0.48). Empty rate ranges from 12.9% (English–Latin) to 42.5% (English–Chinese Simplified). Table 3 gives the same profile on the eight organizer-priority pairs. Submitted primary: `fluencyscore-multispan-v2` (ID 871898), all 21 official pairs.

Locator events on the built files are almost all exact substring matches (439,237), with rare normalized (1,495) and bounded fuzzy (1,621) hits. Unlocatable phrases are dropped (798); 15,724 overlapping or adjacent spans are merged.

3.3 Overlay — Conservative second pass

An A/B on WMT25 human gold did not clear a +0.02 macro-F gate for switching the backbone to Gemini 3 Flash, so we kept Gemini 3.5 Flash and v2. Overlay is a second pass that may only *add* spans, and only on official hypotheses that v2 left empty. Existing v2 spans are frozen: overlay cannot edit, delete, or overlap them. It is allowed at most two additions per hypothesis, only at high confidence, and almost never as `major` (127 proposed majors demoted, 2 kept). Of 35,774 empty official targets, 2,441 hypotheses received new spans (2,583 spans total); 33,184 stays empty after a second look. Rejections: 837 below confidence, 30 over the add cap. Second submitted system: `fluencyscore-overlay` (ID 873483).

4 Task 2: Quality Scores

Task 2 requires one number in $[0, 100]$ per hypothesis, ranked against human cESA (Lavie et al., 2026). The FluencyScore framework stops at FluencyScore-cal; steps 3–4 below are WMT26 adapters. WMT26 test data are not used to fit any parameter. Running example: raw aggregate 8.0.

1. **Judge** (FluencyScore). Gemini 3.5 Flash, sin-

Pair	Spans/hyp.	Empty%	Mass
cs-de	1.59	28.4	0.71
cs-vi	1.61	27.3	0.80
zh-ja	1.77	24.1	0.99
en-arz	1.52	31.2	0.76
en-et	2.03	22.9	1.08
en-is	2.53	16.5	1.43
en-id	1.36	25.5	0.59
en-sme	1.82	15.4	1.24

Table 3: v2 output profile on the eight high-priority official pairs (Lavie et al., 2026). Same disclaimer as Table 2.

gle pass, five sub-dimensions $\rightarrow$ raw aggregate on 1–10 (example: 8.0).

2. **FluencyScore-cal** (framework). Apply f from §2 ($f(8.0)=6.09$). Ranks among hypotheses are unchanged.
3. **Affine cESA adapter** (WMT26). Fit $\widehat{\text{cESA}} = a + b \cdot \text{fluency}_{201}$ on WMT25 human ESA (development data, not WMT26 test; $n=4559$). Spearman $\rho=0.614$ on that fitting set (ordinary-least-squares $R^2=0.098$: the map sets scale, not rank). Coefficients omitted; they are a shared-task adapter, not part of FluencyScore-cal.
4. **Moment match, then clip** (WMT26). After step 3 the *distribution* need not match WMT25 human ESA (too many 100s, wrong spread). We rescale unclipped predictions so their mean and standard deviation match the WMT25 calibration set ($\mu=62.03$, $\sigma=10.02$), then clip to $[0, 100]$. The linear rescale is monotone and does not improve rank correlation except through ties; clipping can add ties at the bounds. The step is for scale calibration, not for a higher Spearman ρ . Effect on the fitting distribution: the share of scores at ESA=100 falls from $\approx 31\%$ to 1.4% .
5. **Missing hypotheses**. API failures are filled with 50.

The Codabench upload is named `esa-t70` (inclusive threshold **70** in the filename). Task 2 scores remain continuous values in $[0, 100]$; the threshold is only for organizer Task 3 auto-conversion (a score ≥ 70 would count as error-free in that conversion). We did not submit a separate Task 3 model. Submitted system: `aira_mo` / ID 872328, all 21 official pairs. On the official Task 2 checkpoint, 241,170 of 242,216 hypothe-

ses are scored (99.6%); a handful of pairs sit at 96–99% and are filled with 50 on API failure.

We omit Codabench verification correlations: they measure agreement with an automatic pseudo-gold, not with human cESA, and are not a substitute for the WMT25 fair-mask $\rho=0.613$ (Mongush and Pukhov, 2026). Illustrative chain (not a measured item): raw 8.0 $\rightarrow f(8.0)=6.09 \rightarrow$ adapter and moment match $\rightarrow$ a clipped score such as 79.

5 Coverage and Challenge Sets

We score every official pair, including the eight high-priority pairs listed by the organizers and Czech–Ukrainian (distributed without references; analyzed separately in the planned Findings) (Lavie et al., 2026). English–Japanese and English–Korean were removed from the official set on 23 July 2026 and are not in our official files.

Challenge-set items use the same judge, locator, and Task 2 map. They are returned to set authors and are not part of the 21-pair ranking. The challenge checkpoint contains 110,319 rows (104,952 scored, 95.1%).

6 Implementation

A multithreaded Gemini client writes a per-run `manifest.json` (model, prompt/schema hashes, decoding settings, repair counts, token medians) and resumes from the last unfinished hypothesis. OpenAI / OpenRouter adapters exist but were not used for submitted systems. On HTTP 400 the runner drops an unsupported flag (`temperature`, `thinkingConfig`, `few-shot prefill`) and retries. Older A/B runs pinned `no-temperature` and `thinking-level=low`; submitted v2 used the provider default. Official v2 wrote 241,463 multi-error judge records; overlay targeted 35,774 empty official hypotheses. API or safety failures on Task 1 fall back to the v1 single-span record for that key; official and challenge slots are filled.

7 Discussion

Task 1 is an adaptation of the diagnostic channel: several problem phrases per hypothesis instead of one, then deterministic alignment. Overlay and a gated trial of Gemini 3 Flash were conservative—the backbone did not clear the WMT25 +0.02 macro-F gate, and overlay may only add spans on empty slots—consistent with the companion

finding that, at a matched backbone, changing the scoring framework moves agreement only a little (Mongush and Pukhov, 2026).

Task 2 uses one global cESA adapter for all 21 pairs. Pair-conditioned mapping, and the register/dialect fixes from the companion error analysis, are the obvious next steps; that study already finds per-pair reliability to be predictable from typological distance.

We submitted full 21-pair files rather than partial-coverage runs. The structured error record is what Task 1 consumes; the scalar is what Task 2 consumes.

8 Limitations

As of 9 September 2026 we find no official WMT26 MT-Evaluation human-cESA rankings or Findings tables. The General MT campaign released preliminary human annotations (finalize date stated as 16 September 2026) (Zouhar et al., 2026); that release scores translation systems, not metric submissions, and we do not derive unofficial ranks from it. Codabench test-period numbers are pseudo-gold and are omitted. Table 2 is an output profile of our span files, not human-gold span accuracy. Task 2 uses a single affine adapter and moment-match statistics shared across 21 pairs; no pair-specific calibration is included. Overlay cannot repair incorrect spans on non-empty hypotheses. Known FluencyScore failure modes still apply: fluent-but-wrong hypotheses often receive none, and `register_mismatch` remains weak. A single closed-source Gemini backbone implies provider and snapshot drift. The judge prompt and sub-dimension weights are proprietary, so the systems are not fully reproducible from this paper. References were unused by design.

Acknowledgments

We thank the WMT26 Automated Translation Quality Evaluation organizers for the shared task and the cESA protocol, and Google for Gemini API access. Any remaining errors are ours.

Ethics Statement

Submitted systems score machine-translation output only. They use a closed-source LLM API; prompts and weights are not released. No personal data beyond shared-task test text are collected. Au-

tomatic quality scores should not replace human review for high-stakes publication.

A Task 1 Response Schema (Abbreviated)

Each hypothesis is judged to JSON (line-wrapped for column width):

```
{ "errors": [
  { "problematic_phrase": "...",
    "occurrence": 1,
    "severity": "major|minor",
    "category": "...",
    "confidence": "high|med|low" } ],
  "omission": "major|minor|null",
  "garbled": false }
```

Omission is never emitted as a span (missing text has no target offsets). When `garbled=true`, post-processing emits a single whole-segment major span.

References

- Tom Kocmi et al. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation (WMT 2024)*.
- Alon Lavie et al. 2026. WMT 2026 shared task: Automated translation quality evaluation systems. <https://www2.statmt.org/wmt26/mteval-task.html>. Accessed 2026-09-09.
- Airana Mongush and Dmitrii Pukhov. 2026. FluencyScore: A structured LLM-judge for fluency in machine translation. Under review.
- Stefano Perrella, Eric Morales Agostinho, and Hugo Zaragoza. 2026. [Span-level machine translation meta-evaluation](#). *Preprint*, arXiv:2603.19921.
- Vilém Zouhar et al. 2026. WMT 2026 general MT human evaluation data. <https://github.com/wmt-conference/wmt26-general-mt/releases/tag/humeval>. Preliminary annotations; organizers state finalization by 16 September 2026. Accessed 2026-09-09.