

SpinPop: A Fast Spin Metric for the WMT26 Metrics Shared Task

Jia Xu
8AI LLC

Abstract

We describe our submission to the WMT26 Shared Task on Automated Translation Quality Evaluation. Our primary system is a per-language-pair, z -normalized, weight-free ensemble that combines six complementary signals: our novel metric SpinPop, COMET-22, and four large language model judges.

SpinPop is training-free, constant-cost, tokenization-agnostic, broadly multilingual, and deterministic. Within its encoder’s language coverage it scores every segment, with or without a reference, using a single frozen-encoder pass that runs locally and requires no paid API calls.

Our system achieves a mean per-language-pair Pearson correlation of 0.8534, ranking second on the WMT26 Task 2 leaderboard. The results demonstrate that a simple unsupervised ensemble of a spin-code metric, a neural metric, and multiple LLM judges can achieve top-tier performance with equal weights.

1 Introduction

Task 2 of the WMT26 Shared Task on Automated Translation Quality Evaluation asks systems to predict continuous segment-level translation quality scores from a source sentence and its translation, optionally using a reference translation. Systems are evaluated by the mean segment-level Pearson correlation with human cESA scores (Kocmi et al., 2024), averaged over 21 language pairs.

We submitted two closely related systems based on a simple, unsupervised ensemble of complementary evaluation signals: our lightweight byte-code metric SpinPop, the neural metric COMET-22, and multiple zero-shot LLM judges. They differ only in their use of Gemini-3.1-Pro: System 1 applies it to the nine language pairs where the base ensemble is weakest, whereas System 2 uses it for all language

pairs.¹ The results show that a simple ensemble of heterogeneous evaluation signals can achieve top-tier translation quality estimation without learning combination weights.

The remainder of this paper is organized as follows. Section 2 presents the system architecture, covering the base ensemble, the judge prompting, the fusion rule, and our handling of low-resource pairs. Section 3 reports results on the WMT26 blind test together with component ablations on a WMT25 development slice. Section 4 situates our approach within prior work on quality metrics and ensembles, and Section 5 concludes.

2 System Architecture

The base ensemble. Figure 1 gives an overview of the system. Both submissions are unsupervised ensembles built from the same five per-segment signals: SpinPop, the neural metric COMET-22, and three LLM judges.

Our novel metric SpinPop, based on SPINHASH (Xu et al., 2026), represents each token by a compact code derived from a frozen multilingual encoder. Its design significantly advances prior byte-based representation work (Zhang and Xu, 2022; Li et al., 2025), which it adapts to translation quality evaluation. It requires only a single frozen encoder pass followed by a fast byte-level comparison, making it substantially more cost- and time-efficient than a learned metric or an LLM judge.

SpinPop in detail. SpinPop scores a hypothesis with BERTScore-style token alignment, but over compact binary codes rather than dense embeddings. Each token is embedded by a frozen multilingual mBERT encoder (Devlin et al., 2019), the `bert-base-multilingual-cased` checkpoint of ~ 178 M parameters, and its frozen layer-9

¹Submitted to Codabench as submission 875949 for System 1 and submission 876231 for System 2, our primary submission.

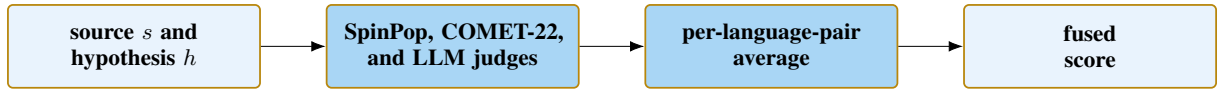


Figure 1: **System overview.** The five base signals, SpinPop, COMET-22, and three LLM judges, with an optional judge, Gemini-3.1-Pro, added on some or all language pairs, each score the source–hypothesis pair reference-based or, when no reference is present, against the source. Per language pair the present signals are z -normalized and averaged with no learned weights.

token embedding is converted to a binary spin code following SPINHASH (Xu et al., 2026), that is, the sign of a fixed set of random linear projections, giving a 1,024-bit code, i.e., 128 bytes per token. The projections are drawn once and never trained. Token-to-token similarity is the fraction of agreeing bits, computed by XOR and popcount, a binary analogue of cosine similarity. Given the code sequences of a hypothesis and its comparison text, each hypothesis token is greedily matched to the comparison token with the highest bit agreement, and the matched similarities are aggregated into precision, recall, and F_1 , with F_1 reported as the segment score. In *reference-based* mode the comparison text is the reference; in *reference-free* (QE) mode it is the source, so SpinPop scores cross-lingually against the source whenever no reference is present. SpinPop is *training-free*, adding no learned parameters and never fine-tuned on quality-evaluation data, unlike supervised metrics such as COMET-22, its only pretrained component being the frozen encoder, exactly as in BERTScore. It is *tokenization-agnostic*, adding no metric-specific vocabulary and inheriting whatever segmentation the frozen encoder uses.

COMET-22 (Rei et al., 2022a) is a reference-based neural metric. The three LLM judges are GPT-5.4, Gemini-3.5-Flash, and GPT-5.6-luna. We selected them on a random sample of 500 WMT25 segments to achieve the best quality-to-cost trade-off, excluding higher-scoring but substantially more expensive models, including GPT-5.6-sol and GPT-5.5.

Judge prompting. All LLM judges receive the same system prompt: act as a professional translation quality evaluator and assign a single score from 0 to 100, where 0 denotes a completely wrong or unintelligible translation and 100 a perfect one, balancing adequacy and fluency. The prompt then provides the source sentence, the candidate translation, and the reference whenever available. We generate a single response per segment and parse the returned integer as the quality score.

All judges are used zero-shot, without in-context examples or task-specific fine-tuning. Gemini-3.5-Flash and GPT-5.6-luna are run with thinking disabled, GPT-5.4 uses its default reasoning, and Gemini-3.1-Pro additionally uses a small explicit thinking budget.

Fusion and coverage. The shared-task metric is Pearson correlation, which is invariant to affine transformations. We therefore combine the signals without learning weights. Intuitively, since Pearson ignores any global scale or offset, the signals’ differing ranges can be aligned by z -normalization instead of learned weights, and the fused score needs no further calibration. For each language pair, we z -normalize every available signal across its segments, average the normalized scores for each segment, and linearly rescale the result. The per-signal normalization statistics, each signal’s mean and standard deviation, are computed for each language pair over all of that pair’s evaluated segments; because Pearson is invariant to affine transforms and no human scores enter the computation, this transductive normalization fits nothing to the gold labels. A signal missing on a segment, such as COMET-22 or any reference-based signal when no reference is available, is excluded from that segment’s average, which is taken over the signals present.

Reference availability varies widely across language pairs. Some, such as CS-DE and ZH-JA, are almost entirely reference-based, whereas others, such as CS-UK, are almost entirely reference-free, with only about 3% of segments providing a reference. When a reference is available, we use COMET-22 together with the reference-based versions of SpinPop and the LLM judges. Otherwise, SpinPop and the judges operate in QE mode against the source. Because SpinPop works with or without a reference, reference availability never leaves a supported segment unscored.

Low-resource pairs. Three target languages, Ladin, Ligurian, and Northern Sami, fall outside

the coverage of both the multilingual encoder underlying SpinPop and COMET-22, so we exclude both signals for these language pairs. Instead, we rely on LLM judging alone: Gemini-3.1-Pro for Ligurian and Northern Sami, and Gemini-3.1-Pro combined with a SONAR-based pivot signal (Duquenne et al., 2023) for Ladin, where the hypothesis is first mapped into a supported language before scoring. As Section 3 shows, these three language pairs remain our weakest and account for most of the gap to the top-ranked system.

The two configurations. The submissions differ only in how widely a fourth judge, Gemini-3.1-Pro, is added to the base ensemble for the non-low-resource language pairs.

- **System 1**, submission 875949, adds Gemini-3.1-Pro only for the nine language pairs where the base ensemble was least competitive: CS-UK, EN-UK, ZH-JA, EN-TH, EN-HY, EN-DE, EN-CS, EN-RU, and CS-DE. All remaining pairs use the base ensemble alone.
- **System 2**, submission 876231, adds Gemini-3.1-Pro for every non-low-resource language pair.

Although Gemini-3.1-Pro achieved only a modest development score as a standalone judge, extending it from the nine weakest pairs to all eighteen non-low-resource pairs raised the mean Pearson correlation from 0.8496 to 0.8534 on the blind test. The two-system split was primarily operational rather than a controlled ablation. Because Gemini-3.1-Pro costs roughly four times as much as the three base judges combined, a ratio we report in Table 6, we initially applied it only to the language pairs where we trailed competing teams most on the intermediate leaderboard, yielding System 1. Once additional API quota became available, we extended the same configuration to all language pairs, yielding System 2. System 2 is our intended primary submission, while System 1 is an earlier checkpoint retained as the designated second submission.

3 Experiments

Leaderboard results. Our system is unsupervised, and nothing is fit to the blind test: the base-judge selection of Section 2 was the only choice made on development data. Table 1 reports our primary submission on the blind test. System 2 attains a mean per-language-pair Pearson of **0.8534**,

1. ammonsh		0.9399	
2. ours		0.8534	
3. 1020 Group		0.8474	
LP	r	LP	r
cs-de	0.861	en-is	0.925
cs-uk	0.869	en-id	0.864
cs-vi	0.851	en-kk	0.885
zh-ja	0.883	en-lll	0.645
en-ar	0.857	en-lij	0.701
en-hy	0.926	en-ru	0.888
en-be	0.884	en-sme	0.741
en-zh	0.891	en-th	0.912
en-zhtw	0.813	en-uk	0.870
en-cs	0.882	en-de	0.879
en-et	0.895	Mean	0.853

Table 1: **Leaderboard results.** Top: the team leaderboard. Bottom: it scores strongly across language pairs, trailing mainly on the three low-resource ones. EN-ZHTW denotes English→Traditional Chinese.

ahead of 1020 Group at 0.8474 and behind ammonsh at 0.9399, while System 1 attains 0.8496. Most pairs fall in the 0.85–0.93 range. The remaining distance to the top-ranked system comes almost entirely from the three low-resource pairs: EN-LLD at 0.645, EN-LIJ at 0.701, and EN-SME at 0.741, where only LLM judging carries information. Excluding these three, System 2 rises to 0.880 and the top-ranked system to 0.948. The gap therefore narrows from 0.086 over all 21 pairs to 0.068 over the remaining 18.

Adding components. Our ablations, Tables 2 through 5, all use a WMT25 development slice of 8 language pairs, English into Arabic, Korean, Serbian-Cyrillic, Japanese, Chinese, Russian, Ukrainian, and Czech, with 620 human-scored segments per pair, or 4,960 in total. Results are reported in both the task’s Pearson metric and the tie-calibrated pairwise accuracy acc23 (Deutsch et al., 2023), which agree closely at $r = 0.94$. Table 2 builds the ensemble up from SpinPop, whose marginal lift is largest on the weaker signals, for example chrF from 0.263 to 0.279 and COMET-22 from 0.326 to 0.338, and small on the strongest judges. In the ensemble its main roles are broad coverage and efficiency.

Leave-one-out. We next drop each member from the full ensemble in turn to isolate its contribution, as reported in Table 3. On Pearson, removing Gemini-3.5-Flash hurts most, at -0.026 , while removing SpinPop slightly *helps*, at $+0.002$. On acc23, no single member is critical, with the largest

Ensemble	Pearson	acc23
SpinPop	0.269	0.569
SpinPop + COMET-22	0.338	0.582
SpinPop + GPT-5.4	0.428	0.611
SpinPop + Flash	0.447	0.623
SpinPop + luna	0.435	0.614
Full ensemble	0.498	0.634

Table 2: **Adding components.** Every added signal improves over SpinPop alone, building the base ensemble from 0.269 to 0.498. WMT25 slice, mean per-LP Pearson and acc23.

Ensemble	Pearson	acc23
– SpinPop	0.500	0.626
– COMET-22	0.493	0.628
– GPT-5.4	0.490	0.630
– Gemini-3.5-Flash	0.472	0.626
– GPT-5.6-luna	0.481	0.631
Full ensemble	0.498	0.634

Table 3: **Leave-one-out.** The full ensemble (bottom row) combines all five signals, SpinPop, COMET-22, GPT-5.4, Gemini-3.5-Flash, and GPT-5.6-luna, and each row above removes one of them. No single member is indispensable: removing Flash hurts Pearson most, while on acc23 the largest drop is shared by SpinPop and Flash. WMT25 slice.

drop, -0.008 , shared by SpinPop and Gemini-3.5-Flash. SpinPop therefore contributes little to the ensemble’s headline Pearson score, serving as an efficient, low-cost member rather than the component that most raises it.

SpinPop lifts. When we add SpinPop to three reference-based baselines, as Table 4 shows, it lifts all three on both metrics, BLEU by $+0.043$, chrF by $+0.017$, and COMET-22 by $+0.013$ in Pearson, and more on acc23, where BLEU rises by $+0.11$. We report these as descriptive deltas on the 4,960-segment development slice and do not claim statistical significance for the sub-0.02 Pearson differences; in the full ensemble SpinPop’s role is efficiency and coverage rather than a correlation gain, consistent with the leave-one-out result above.

Combination. Holding the full ensemble fixed, we vary only the fusion rule, with the alternatives summarized in Table 5. The z -normalize-then-mean rule used in our submissions reaches 0.498, versus 0.480 for an unnormalized mean and 0.480 for a median of the normalized signals, so the normalization, not only the choice of members, is what

Ref-based metric	Pearson	acc23
BLEU	0.229	0.461
+ SpinPop	0.272	0.573
chrF	0.263	0.578
+ SpinPop	0.279	0.579
COMET-22	0.326	0.547
+ SpinPop	0.338	0.582

Table 4: **SpinPop lifts.** Adding SpinPop lifts every weaker reference-based metric on both Pearson and acc23, most on BLEU, by $+0.04$ Pearson and $+0.11$ acc23. WMT25 slice.

Fusion rule	Pearson	acc23
simple mean, no z -norm	0.480	0.618
median of z -normalized	0.480	0.604
z -normalize + mean, ours	0.498	0.634

Table 5: **Combination.** The per-language z -normalize-then-mean fusion used in our submissions is best on both metrics, beating an unnormalized mean and a median. Full ensemble, WMT25 slice.

matters.

Cost. The signals differ sharply in cost, as Table 6 lays out. SpinPop needs only one pass through a frozen ~ 180 M-parameter multilingual encoder, after which each token is a 128-byte code and scoring is a fast byte-level comparison. COMET-22 is local but runs a GPU forward pass through a ~ 0.58 B-parameter model per segment, and each LLM judge costs one API call per segment, Gemini-3.1-Pro on the order of a few US dollars per thousand. SpinPop also trains nothing, whereas COMET-22 is supervised on human ratings and the judges are large pretrained models. System 2 applies Gemini-3.1-Pro on all 18 non-low-resource pairs, twice as many as System 1’s nine, so it roughly doubles the Gemini-3.1-Pro cost for a $+0.0038$ mean gain.

4 Related Work

MT evaluation metrics. Automatic MT evaluation has evolved from lexical metrics such as BLEU (Papineni et al., 2002) and chrF (Popović, 2015), through embedding-based metrics including BERTScore (Zhang et al., 2020) and BLEURT (Sellam et al., 2020), to learned neural metrics such as COMET (Rei et al., 2020, 2022a), MetricX (Juraska et al., 2023), CometKiwi (Rei et al., 2022b), and xCOMET (Guerreiro et al., 2024), which currently dominate WMT evaluations (Freytag et al., 2023). In parallel, LLM judges such as

Signal	Local?	Training	\$/1k
SpinPop	yes	none	0
COMET-22	yes	supervised	0
Gemini-3.5-Flash	no	pretrained	~0.1
GPT-5.4	no	pretrained	~0.3
GPT-5.6-luna	no	pretrained	~0.5
Gemini-3.1-Pro	no	pretrained	~4

Table 6: **Cost.** SpinPop is the only signal that is local, untrained, and free at inference time. Here, **Local** indicates on-device computation. **Training** denotes the model or training cost, ranging from none to supervised training or a large pretrained model. **\$/1k** reports the approximate US cost per 1,000 segments at submission-time API prices.

GEMBA and GEMBA-MQM (Kocmi and Federmann, 2023b,a) have shown that zero-shot prompting can match or surpass learned metrics, albeit at substantially higher inference cost. SpinPop addresses the efficiency side of this trade-off. It retains BERTScore-style token alignment while replacing dense embeddings with compact byte codes, significantly advancing prior work on byte-based representations for efficient NLP (Zhang and Xu, 2022; Li et al., 2025). In our system it serves as a lightweight ensemble signal that applies to both reference-based evaluation and quality estimation.

Ensembles of quality signals. Ensembles are common in shared-task systems, but published work on the WMT metrics and QE task has largely focused on single learned metrics (Rei et al., 2020; Guerreiro et al., 2024) or single LLM judges (Kocmi and Federmann, 2023b,a). Our contribution is a simple unsupervised ensemble: a per-language-pair z -normalized average of a byte-code metric, a neural metric, and multiple zero-shot LLM judges. This approach is competitive in WMT26 Task 2.

5 Conclusion

We described our two WMT26 Task 2 submissions: unsupervised z -normalized ensembles of the byte-code metric SpinPop, the neural metric COMET-22, and zero-shot LLM judges, which require no learned combination weights. Within the ensemble, SpinPop is the lightest, broad-coverage member: it scores every supported segment at negligible cost and lifts weaker reference-based metrics by about 0.02 in correlation on average.

Limitations

Both submissions depend on proprietary LLM judges queried through public APIs, which limits reproducibility and makes runtime cost dominated by API calls rather than by SpinPop. Our low-resource handling is a pragmatic fallback, and the three lowest-resource pairs remain far below the rest. Finally, SpinPop’s accuracy is inherited from a frozen multilingual encoder and degrades on languages that the encoder poorly supports.

References

- Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 12914–12929.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186.
- Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. SONAR: Sentence-level multimodal and language-agnostic representations. *arXiv preprint arXiv:2308.11466*.
- Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, pages 578–628.
- 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 (TACL)*, 12:979–995.
- Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, pages 756–767.
- Tom Kocmi and Christian Federmann. 2023a. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, pages 768–775.

- Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 193–203.
- Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation (WMT)*, pages 1440–1453.
- Shenglan Li, Jia Xu, and Mengjiao Zhang. 2025. Em-Byte: Decomposition and compression learning for small yet private NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 7182–7201.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.
- Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT)*, pages 392–395.
- Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. 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, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.
- Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Alon Lavie, Luisa Coheur, and André F. T. Martins. 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.
- Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7881–7892.
- Jia Xu, Jinhua Wang, and Constantin F. Aliferis. 2026. SpinHash: An $O(1)$ encoder for any vocabulary. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Mengjiao Zhang and Jia Xu. 2022. Byte-based multilingual NMT for endangered languages. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4407–4417.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*.