Lost in Morphology at WMT26: A Korean–English Test Suite for LLM Translation Blind Spots

Deunsol Lee* **and Daewoong Kim*** **and Yunha Seo***
Soongsil University
{leedeunsol,8330kdw,bobo9245}@soongsil.ac.kr

Abstract

We present ***lost-in-morphology***, a 3,536-item Korean→English challenge suite spanning nine linguistic phenomena and evaluated on 23 systems submitted to WMT26. The suite targets ellipsis punctuation, stacked particles, wordplay, topic/subject marking, evidentiality, scrambling, auxiliary constructions, connective endings, and ideophones. Rather than forcing all phenomena through one metric, we match each to its observable evidence: deterministic checks for visible or lexically constrained contrasts, structured semantic observations for discourse and operator meanings, contrastive or distributional scoring for multi-variant frames, and a strategy-aware panel for wordplay. Learned judges and classifiers take task-specific admission tests. Scores from those that fail are marked provisional. The evaluation reveals a recurring failure mode hidden by aggregate MT quality: translations often preserve the core proposition while neutralising morphology-induced distinctions. Distinct scalar particles collapse into bare English *even*, evidential and auxiliary meanings are frequently omitted, and connective and ideophone contrasts are flattened. Wordplay remains the hardest subset, with the best system scoring 0.385 against 0.546 for curated references. The evaluator study further shows that task-specific validation is necessary: a conflict-free wordplay judge fails through ceiling saturation, while three judges that are also evaluated systems pass without detectable family preference.

1 Introduction

General-purpose language models now match or surpass conventional MT systems on many high-resource directions (Kocmi et al., 2024; Manakhimova et al., 2023). Yet strong corpus-level scores do not imply linguistic completeness. A translation may remain fluent and propositionally plausible while deleting a discourse contrast, neutralizing an information source, collapsing scalar meanings, or omitting a verbal operator. Such losses are difficult to detect with aggregate metrics and are consistent with prior findings that MT can reduce linguistic richness (Vanmassenhove et al., 2019, 2021).

Korean→English is a revealing setting because Korean packages grammatical, information-structural, and pragmatic distinctions into particles, connective endings, auxiliaries, evidential constructions, and flexible word order. English often requires these distinctions to be unpacked into separate lexical or syntactic choices. Ideophones, ellipsis punctuation, and wordplay add expressive contrasts without stable one-to-one equivalents. Challenge sets are well suited to isolating such divergences (Isabelle et al., 2017; Macketanz et al., 2018), but the phenomena differ too much for a single scorer to be appropriate.

We therefore present ***lost-in-morphology***, a 3,536-item Korean→English suite spanning nine subsets and evaluated on 23 WMT26 systems (WMT Organizers, 2026). The suite combines naturally occurring examples, controlled frames, minimal-variant groups, and expert-authored items. Its central design principle is **phenomenon-matched measurement**: visible features use deterministic rules; semantic distinctions use structured observations whose scores are derived by code; multi-variant frames use contrastive or distributional preservation; and wordplay uses a fixed strategy-aware rubric.

The resulting analysis exposes a common pattern across otherwise different phenomena: systems often retain the event or proposition while suppressing the distinction contributed by Korean morphology or discourse marking. Particle stacks are funnelled into generic English cues, evidence source and auxiliary operators disappear, and connective or ideophone variants become less distin-

* Equal contribution.

Proceedings of the Eleventh Conference on Machine Translation, pages 1555–1567
October 28-29, 2026 ©2026 Association for Computational Linguistics

guishable. At the same time, some failures remain strongly phenomenon- or family-specific, such as deletion of ellipsis-bearing turns.

A second contribution is methodological. Because learned evaluators vary substantially across tasks and models (Zheng et al., 2023; Bavaresco et al., 2025), we treat them as measurement instruments rather than ground truth. Each learned judge or classifier takes a subset-specific admission test, and scores from one that fails are marked provisional. We report the resulting instrument failures alongside the translation failures they are intended to detect.

2 Related Work

MT challenge sets isolate targeted linguistic contrasts rather than averaging over heterogeneous test data (Isabelle et al., 2017; Macketanz et al., 2018). Regex- and rule-scored morphology has been studied in Morpheval (Burlot and Yvon, 2017; Burlot et al., 2018), while contrastive and minimal-pair designs have been used for grammatical and discourse phenomena (Sennrich, 2017; Bawden et al., 2018; Müller et al., 2018). Our output-side collapse diagnostics are motivated by the limits of minimal-pair scoring when systems are merely insensitive to the tested contrast (Vamvas and Sennrich, 2021). ACES (Amrhein et al., 2022) and MQM (Lommel et al., 2014) provide broader error taxonomies; our error codes are MQM-compatible.

For Korean, K-SNACS (Hwang et al., 2020) informs the particle feature inventory, while prior work motivates the discourse and grammatical distinctions tested here. Wordplay evaluation builds on SemEval-2017 Task 7 (Miller et al., 2017) and the JOKER tasks (Ermakova et al., 2023, 2025). LLM judging has shown strong system-level correspondence with human MT evaluation (Kocmi and Federmann, 2023), but judge reliability is task-dependent; this motivates our admission tests, reference pseudo-systems, negative anchors, agreement checks, and attrition reporting.

3 The *lost-in-morphology* Suite

3.1 Suite Overview

Table 1 summarizes the nine subsets. Full construction details, inventories, and rubric examples are provided in Appendix A.

Every item includes a machine-readable `judge_spec`. Where a three-level verdict is appropriate, scores are PASS=1, PARTIAL=0.5, and FAIL=0. Deterministic rules may emit WARN, in which case a learned adjudicator handles the residue. System names are withheld from learned judges, chat-style outputs are normalized before scoring, and subset scores are item micro-averages unless a frame-level metric is explicitly defined.

All scorers are also run on curated references as pseudo-systems. This gives an empirical false-negative floor before system ranking: 0.999 for the ELLIPSIS rules, 0.892 for the PARTICLES checklist, 0.546 for the admitted WORDPLAY panel, 0.967 for the connective classifier, and 0.733 for the ideophone classifier. These anchors are diagnostic rather than claims that references should receive perfect scores.

3.2 Phenomenon-Matched Measurement

Visible or lexically constrained evidence. Ellipsis is scored by category-specific checks over a localized focus span: presence, position, pause count, terminal combinations, quote flanking, and overgeneration. Stacked particles use feature-specific cue checklists, because several English realisations may be valid. Checklist accuracy alone cannot detect when distinct Korean inputs receive the same English output, so particle scoring is paired with a contrast-collapse rate over gold-distinct members of the same minimal-variant group.

Structured semantic observations. Topic/subject marking, evidentiality, scrambling, and auxiliaries are judged through strict structured outputs rather than free-form scalar ratings. A judge observes, for example, proposition correctness, realized discourse status, evidence source, typed argument roles, or semantic operators. Deterministic code then maps those observations to the final score. This separates proposition preservation from the morphological or discourse distinction under test.

Contrast and distribution preservation. Connective endings and ideophones use multi-variant frames. Pointwise label accuracy can reward a classifier or system that repeatedly defaults to a majority label, so we additionally measure frame-level preservation with

$$\mathrm{SRP} = 1 - \mathrm{JSD}(p_{\mathrm{gold}}, p_{\mathrm{pred}}),$$

where the distributions are taken over the labels realized within a controlled frame. This makes se-

Table 1: Overview of the nine test-suite subsets. The main paper groups phenomena by measurement type; detailed item construction is in Appendix A.

Phenomenon	n	Test structure	Primary measurement	Key diagnostic
Ellipsis	601	corpus + templated items	deterministic rules + adjudication	preservation / omission
Stacked particles	579	172 minimal-variant groups	cue checklist + pair comparison	semantic fidelity / collapse
Wordplay	65	attested + expert-authored	admitted LLM panel	strategy-aware score
Topic/subject	387	controlled minimal pairs	structured semantic judge	contrast preservation
Evidentiality	400	100 four-way frames	structured semantic judge	source preservation / neutralisation
Scrambling	599	argument-order variants	typed role recovery	role exact match / discourse preservation
Auxiliaries	541	operator variants	structured semantic judge	operator recall / deletion
Connectives	194	discourse-relation variants	admitted relation classifier	distributional preservation
Ideophones	170	manner/intensity variants	admitted fine-grained classifier	distributional preservation

mantic collapse observable even when individual outputs are fluent.

Open-ended strategy-aware evaluation. Wordplay cannot be decided from closed surface cues. Each item therefore specifies a pun pivot, sense inventory, and curated strategy, and admitted judges apply a fixed three-level rubric: 1.0 for recreated wordplay or equivalent humorous dual-sense preservation, 0.5 when one sense is faithful and the other remains recoverable, and 0.0 when the pun is ignored or the pivot mistranslated. The reported score is the median of admitted judges.

3.3 Admission of Learned Evaluators

Learned evaluators are admitted per task rather than by architecture. Wordplay judges are tested on curated-reference positives whose expected rubric strata are known a priori and on mismatched-reference negatives. A judge must preserve the expected positive ordering, stay below the rubric ceiling, and reject negatives. Connective and ideophone label classifiers use their own calibration instruments. Full prompts, tables, and agreement analyses are provided in Appendix C.

The most important admission result is counter-intuitive: EXAONE-4.0-32B, the only conflict-free wordplay judge in the candidate panel, fails because it assigns 1.00 to every positive stratum and to the references overall. Gemma4-31B, gpt-oss-20b, and Qwen3.6-27B pass, reproduce the expected strategy ordering, and reject negatives. The wordplay score therefore uses the median of those three judges. For connective endings, a rule-based classifier is selected as primary on a held-out calibration set; for ideophones, the reference-as-pseudo-system protocol selects the primary classifier.

4 Results

4.1 Overall Pattern: Fluent Neutralisation

The dominant cross-phenomenon failure is often *neutralisation* rather than an obviously false proposition. Evidential and auxiliary constructions are the clearest examples. Evidentiality spans 0.538–0.968 across systems (mean 0.798), but pooled evidential-feature recall is only 0.712. Among proposition-correct outputs, neutralisation averages 0.204, while conflicting source shift occurs in 0.081 of all outputs. Plain assertions and inferential *geot gatda* are nearly saturated (0.984/0.960), whereas reportative *-dae* falls to 0.690 and direct-perceptual *-teora* to 0.557.

Auxiliary constructions show the widest admitted spread: Dubformer reaches 0.935 while SalamandraTA scores 0.195 (system mean 0.584). Operator precision remains high at 0.946, but recall is 0.624 and the operator-deletion rate is 0.376. The main predicate therefore frequently survives after the contribution of the auxiliary disappears. The try/experience auxiliary *boda* is especially difficult (0.243), while unwanted-result *-go malda* is much easier (0.888).

The same tendency appears in multi-variant phenomena. Connective-ending $\mathrm{SRP}_{\mathrm{JSD}}$ ranges 0.466–0.671 (mean 0.566), and ideophone $\mathrm{SRP}_{\mathrm{JSD}}$ ranges 0.474–0.634 (mean 0.551). Pointwise accuracy does not reproduce the connective ranking, because majority-relation defaults can score adequately while collapsing the other

Table 2: Results for 23 systems. ELLIPSIS and PARTICLES are rule-based scores with LLM-adjudicated WARN cases; WORDPLAY is the median score of admitted judges. Collapse is lower-better; all other scores are higher-better. Bold marks the best system per column.

System	Ellipsis	Particles	Collapse	Wordplay	Connective	Ideophone	Topic	Evidentiality	Scrambling	Auxiliary	Structured Mean	Mean	COMET	BLEU	chrF
Reference (calibration)	0.999	0.892	0.019	0.546	0.967	0.733	1	1	0.957	1	0.989	0.813	–	–	–
Gemini 3.1 Pro	0.973	0.843	0.087	**0.385**	**0.671**	**0.634**	0.953	0.885	0.985	0.713	0.884	**0.734**	0.638	6.69	34.55
TRIVE	**1.000**	0.832	0.238	0.292	0.594	0.578	0.975	0.875	0.998	0.723	0.893	0.708	0.798	36.76	63.46
Wayfinder	0.995	0.79	0.186	0.3	0.566	0.563	0.993	0.875	**1.000**	0.73	0.899	0.695	0.799	33.56	62.47
GPT 5.5	0.984	0.793	0.227	0.304	0.543	0.531	0.993	0.86	**1.000**	0.81	0.916	0.694	0.799	33.81	62.42
DeepSeek V4 Pro	0.985	0.833	0.189	0.235	0.551	0.625	0.993	0.882	0.993	0.775	0.911	0.684	0.794	34.63	62.28
Dubformer	0.992	**0.850**	0.127	0.2	0.554	0.578	0.973	**0.968**	0.998	**0.935**	**0.968**	0.68	0.788	31.81	61.84
Lumen	0.992	0.778	0.351	0.195	0.517	0.543	0.985	0.87	0.995	0.725	0.894	0.655	0.795	32.66	61.64
GoogleTranslate	0.912	0.819	0.325	0.188	0.476	0.584	**0.995**	0.875	0.998	0.42	0.822	0.64	0.798	36.01	63.1
Cohere CAT+	0.988	0.758	0.249	0.156	0.56	0.579	0.968	0.945	0.983	0.677	0.893	0.634	0.786	28.86	58.58
Qingqiu-MT-9B	0.978	0.767	0.265	0.146	0.536	0.517	0.98	0.912	0.945	0.343	0.795	0.63	0.78	30.12	58.48
Mistral Medium 3.5	0.948	0.762	0.282	0.12	0.5	0.505	0.99	0.66	0.953	0.525	0.782	0.61	0.786	31.24	59.01
Qwen 3.6 - 27b	0.958	0.712	0.062	0.146	0.584	0.581	0.985	0.823	0.935	0.67	0.853	0.605	0.785	25.87	56.86
Command A+	0.928	0.754	0.068	0.108	0.602	0.523	0.968	0.848	0.932	0.675	0.856	0.596	0.788	25.92	55.57
Tower 9B	0.956	0.738	0.244	0.085	0.501	0.54	**0.995**	0.63	0.975	0.522	0.781	0.593	0.785	28.02	57.17
Qwen 3.5 - 9b	0.932	0.708	**0.017**	0.135	0.616	0.587	0.927	0.807	0.86	0.655	0.812	0.592	0.78	23.61	54.14
GPT OSS 120B	0.869	0.788	0.121	0.115	0.608	0.61	0.93	0.875	0.953	0.5	0.814	0.591	0.76	23.61	56.21
GLM 4 - 9B	0.933	0.732	0.066	0.082	0.59	0.474	0.935	0.59	0.882	0.482	0.723	0.582	0.767	25.36	54.57
SalamandraTA	0.965	0.71	0.137	0.067	0.466	0.476	0.745	0.555	0.482	0.195	0.494	0.581	0.757	22.68	49.86
Ministral 3 - 14B	0.879	0.703	0.134	0.091	0.658	0.486	0.978	0.635	0.912	0.468	0.748	0.557	0.779	23.77	53.39
GPT OSS 20B	0.806	0.749	0.028	0.08	0.636	0.591	0.983	0.85	0.91	0.63	0.843	0.545	0.774	20.76	52.43
Tiny Aya Global	0.893	0.618	0.156	0.064	0.513	0.498	0.782	0.537	0.562	0.23	0.528	0.525	0.756	23.39	51.47
Gemma 4 - 31B	0.439	0.775	0.284	0.153	0.597	0.589	0.983	0.875	0.985	0.762	0.901	0.456	0.782	33.49	57.26
Gemma 4 - 4B	0.423	0.781	0.221	0.1	0.578	0.483	0.983	0.713	0.965	0.268	0.732	0.435	0.775	32.78	56.24

distinctions. Ideophone collapse averages 0.370 and likewise does not track SRP monotonically.

4.2 Lexical Funneling and Contrast Collapse

Stacked particles expose a lexical version of the same failure. Mean score by stack depth is 0.95 / 0.83 / 0.75 / 0.70 / 0.74 for depths 1–5; the depth-5 rebound also appears in the references and is therefore best read as a template-composition effect rather than improved depth handling.

More importantly, distinct scalar particles are funnelled into bare English *even*. Pooled over systems, *-cocha* and *-mace* are rendered as bare *even* in 88.6% and 87.9% of realisations, while the references use differentiators in 57% of cases. Because *even* satisfies the cue lexicon for multiple scalar labels, a checklist can reward the very collapse the suite is designed to expose. The group-level metric corrects this: Lumen collapses 35.1% of gold-distinct pairs, whereas Qwen 3.5-9B collapses only 1.7% despite a mediocre checklist score of 0.708. Low collapse therefore does not itself imply semantic correctness; it must be interpreted jointly with the feature score.

4.3 Phenomenon-Specific Failure Modes

Ellipsis is largely saturated: 19 of 23 systems score at least 0.87 (median 0.96), and TRIVE preserves all 601 items. The Gemma 4 family is a striking exception at 0.42–0.44. Its failures are not mainly typographic: in 88% (31B) and 87% (4B) of failing cases, the entire ellipsis-bearing utterance is absent. The profile is category-specific: pure silence and most trail-off cases survive, while response turns combined with other punctuation approach zero. Appendix A gives the category breakdown.

Scrambling is comparatively robust once role recovery is separated from surface order. The admitted score ranges 0.483–1.000 (mean 0.922), pooled role-level accuracy is 0.974, strict three-role exact match is 0.942, and explicit role reversal is only 0.009. Object fronting is the hardest variant (0.881), consistent with the need to recover both thematic role and contextual focus rather than imitate Korean word order.

Topic/subject marking similarly looks easy pointwise but becomes more diagnostic under paired comparison. Item scores average 0.956, although the judge fails the negative-anchor criterion and the number is therefore provisional. Jointly judged minimal pairs average 0.863 contrast accuracy and 0.116 semantic collapse, with system-level collapse ranging from 0.02 to 0.53.

4.4 Wordplay and Evaluator Reliability

Wordplay remains the hardest subset. The best system scores 0.385 against 0.546 for curated references, and the top eight systems lie between 0.20–0.39. With only 65 items and confidence-interval half-widths around ± 0.07, the subset supports coarse system-level distinctions rather than fine ranking. The 15 attested items from translation-studies literature are harder for systems (mean 0.139) than the 50 expert-authored items (0.177), so the authored majority does not appear to inflate difficulty.

The evaluator study is itself diagnostic. The conflict-free EXAONE judge fails through ceiling saturation, whereas the three admitted judges are also evaluated systems but show no detectable kin inflation. Their strategy-to- score mappings are similar, yet they agree on the perceived strategy label for only 42% of judgments. Thus the main source of disagreement is often whether surviving wordplay is perceived at all, not how a shared label is valued. We consequently use the panel median only as a calibrated diagnostic for coarse comparison, not as a substitute for human wordplay judgment.

5 Conclusion and Limitations

We introduced *lost-in-morphology*, a 3,536-item Korean→English test suite covering nine grammatical, discourse, and expressive phenomena, and evaluated 23 WMT26 systems with phenomenon-matched instruments. Across the suite, the recurring failure is not necessarily a false proposition but loss of the distinction encoded by Korean morphology or discourse marking: ellipsis-bearing turns are deleted, scalar particles funnel into generic English cues, evidence sources and auxiliary operators disappear, and connective or ideophone contrasts collapse.

The suite also shows why learned evaluators should be validated as task-specific measurement instruments. A conflict-free judge can still be unusable because of ceiling saturation, and judges that nominally share a rubric can disagree mainly in perception. Admission anchors, attrition reporting, agreement checks, and independent classifiers reduce these risks but do not establish equivalence to human evaluation.

Several limitations bound the claims. The suite covers only Korean→English and mixes corpus-mined, controlled, template-generated, and expert-authored items, so category balance is not an estimate of naturally occurring error prevalence. Items within minimal-variant frames are dependent, and psychometric estimates based on 23 systems remain noisy. Wordplay lacks independent human MQM labels; connective tier assignment has low inter-judge agreement; per-judge tier decisions were not retained for the full ideophone pool; and some judge models are also evaluated systems. Future work should add independent human validation, revise saturated and near-dead items, repeat the analysis on new system and judge pools, and extend matched phenomena to additional language directions.

References

Chantal Amrhein, Nikita Moghe, and Liane Guillou. 2022. ACES: Translation accuracy challenge sets for evaluating machine translation metrics. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 479–513, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K. Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 1304–1313.

Franck Burlot, Yves Scherrer, Vinit Ravishankar, Ondřej Bojar, Stig-Arne Grönroos, Maarit Koponen, Tommi Nieminen, and François Yvon. 2018. The WMT'18 morpheval test suites for English-Czech, English-German, English-Finnish and Turkish-English. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 546–560, Belgium, Brussels. Association for Computational Linguistics.

Franck Burlot and François Yvon. 2017. Evaluating the morphological competence of machine translation systems. In *Proceedings of the Second Conference on Machine Translation*, pages 43–55, Copenhagen, Denmark. Association for Computational Linguistics.

Sookhee Chae. 2015. Reported evidentiality in Korean. *Korean Semantics*, 50:69–98.

Lee J. Cronbach. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3):297–334.

Dirk Delabastita. 1994. Focus on the pun: Wordplay as a special problem in translation studies. *Target*, 6(2):223–243.

Dirk Delabastita. 1996. Introduction. *The Translator*, 2(2):127–139. Special issue: Wordplay and Translation.

Liana Ermakova, Tristan Miller, Anne-Gwenn Bosser, Victor Manuel Palma Preciado, Grigori Sidorov, and Adam Jatowt. 2023. Overview of JOKER 2023 automatic wordplay analysis task 3 – pun translation. In *Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum*.

Liana Ermakova, Tristan Miller, Anne-Gwenn Bosser, et al. 2025. Overview of the CLEF 2025 JOKER task 2: Wordplay translation. In *Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum*.

Gemma Team. 2026. Gemma 4 technical report.

Jeonghan Han. 2010. The cases of the more informative proposition triggers in Korean: Focusing on the Korean auxiliary particles -mace, -cocha, -lato, -na. *Korean Journal of Linguistics (한국언어학)*, 35(1):291–319.

Hyun Kyung Hwang and John Whitman. 2021. Korean verb-verb sequences. In Taro Kageyama, Peter E. Hook, and Prashant Pardeshi, editors, *Verb-Verb Complexes in Asian Languages*, pages 193–220. Oxford University Press.

Jena D. Hwang, Hanwool Choe, Na-Rae Han, and Nathan Schneider. 2020. K-SNACS: Annotating Korean adposition semantics. In *Proceedings of the Second International Workshop on Designing Meaning Representations*, pages 53–66, Barcelona, Spain (online). Association for Computational Linguistics.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496, Copenhagen, Denmark. Association for Computational Linguistics.

Youngjun Jang. 2025. The syntax and semantics of the evidential marker –te in Korean. *Eoneo [Korean Journal of Linguistics]*, 50(2):683–713.

Do-Hun Kim. 2011. 영어 문장부호 일립시스(Ellipsis)와 한국어 문장부호 줄임표의 비교 및 일립시스의 번역 전략 [a comparison of the English ellipsis mark and the Korean ellipsis mark, and translation strategies for the ellipsis]. *English21*, 12(4).

Kong-on Kim. 1977. Sound symbolism in Korean. *Journal of Linguistics*, 13(1):67–75.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. 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*, pages 193–203, Tampere, Finland.

Chungmin Lee. 2015. Focus, topic, and contrast. In Lucien Brown and Jaehoon Yeon, editors, *The Handbook of Korean Linguistics*, pages 179–195. Wiley-Blackwell.

Min-chul Lee. 2024. Kiwi: Korean intelligent word identifier. `https://github.com/bab2min/Kiwi`. Version 0.18.

Sunyoung Lee and Soonyoung Kim. 2015. 문체표지로서의 말줄임표와 의성어 · 의태어 번역 [an analysis of the translations of ellipsis and onomatopoeic and mimetic words as style markers]. *Journal of Translation Studies*, 16(2):7–31.

LG AI Research. 2025. EXAONE 4.0: Unified large language models integrating non-reasoning and reasoning modes. arXiv:2507.11407.

Dong-Hoon Lim. 2004. 한국어 조사의 하위 부류와 결합 유형 [subclasses and combination types of Korean particles]. *Kwukehak [Korean Linguistics]*, 43:119–154.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica tecnologies de la traducció*, pages 455–463.

Vivien Macketanz, Eleftherios Avramidis, Aljoscha Burchardt, and Hans Uszkoreit. 2018. Fine-grained evaluation of German-English machine translation based on a test suite. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 578–587, Belgium, Brussels. Association for Computational Linguistics.

Shushen Manakhimova, Eleftherios Avramidis, Vivien Macketanz, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian Möller. 2023. Linguistically motivated evaluation of the 2023 state-of-the-art machine translation: Can ChatGPT outperform NMT? In *Proceedings of the Eighth Conference on Machine Translation*, pages 224–245, Singapore. Association for Computational Linguistics.

Tristan Miller, Christian Hempelmann, and Iryna Gurevych. 2017. SemEval-2017 task 7: Detection and interpretation of English puns. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 58–68, Vancouver, Canada. Association for Computational Linguistics.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third*

Conference on Machine Translation (WMT): Research Papers, pages 61–72.

National Institute of Korean Language. 2014. 문장 부호 해설 [commentary on punctuation marks]. Seoul: National Institute of Korean Language.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. arXiv:2508.10925.

Hyun S. Park. 1995. Mapping scrambled Korean sentences into English using synchronous TAGs. In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 317–319. Association for Computational Linguistics.

Qwen Team. 2026. Qwen3.6-35B-A3B: Agentic coding power, now open to all.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. Evaluation examples are not equally informative: How should that change NLP leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4486–4503. Association for Computational Linguistics.

Rico Sennrich. 2017. How grammatical is character-level neural machine translation? assessing MT quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382, Valencia, Spain. Association for Computational Linguistics.

Jannis Vamvas and Rico Sennrich. 2021. On the limits of minimal pairs in contrastive evaluation. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 58–68.

Clara Vania, Phu Mon Htut, William Huang, Dhara Mungra, Richard Yuanzhe Pang, Jason Phang, Haokun Liu, Kyunghyun Cho, and Samuel R. Bowman. 2021. Comparing test sets with item response theory. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1141–1158. Association for Computational Linguistics.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online. Association for Computational Linguistics.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 222–232, Dublin, Ireland.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The Penn Discourse TreeBank 3.0 annotation manual. Linguistic Data Consortium, University of Pennsylvania. LDC2019T05.

WMT Organizers. 2026. WMT26 subtask: MT test suites — mind the gap: Exposing LLM translation blind spots. `https://www2.statmt.org/wmt26/testsuite-subtask.html`.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Datasets and Benchmarks Track*.

A Detailed Test-Suite Construction

This appendix records the phenomenon-level construction details and examples that are useful for reproduction but not required to follow the main narrative.

A.1 ELLIPSIS: Ellipsis Punctuation (601 items)

Korean writes reticence, silence, hesitation, and editorial omission with the ellipsis mark (. . . , ...), and the 2015 orthographic norms enumerate four core functions (National Institute of Korean Language, 2014). Prior Korean translation studies show that English and Korean ellipses are graphically parallel but functionally divergent (Kim, 2011), and that professional literary Korean→English translation preserves the mark in 83 of 85 cases while frequently shifting function (Lee and Kim, 2015). Our nine categories are those functions plus mark+?/! combinations.

We restrict this subset to dialogue and broadcast-style register because punctuation preservation is genre-dependent. In the licensed broadcast-subtitle corpus used for mining, professional translators retain the mark in 85.3% of Korean-ellipsis pairs (215/252), consistent in direction with the literary figure above.

Of the 601 items, 182 are mined by regex and retained only when the professional English side also contains a mark; 419 are template-generated with carrier contexts written by Qwen3.6-35B-A3B (Qwen Team, 2026) and manually screened.

Table 3: ELLIPSIS categories, after the four norm functions plus terminal combinations.

Category	n	Expected treatment
HESITATION	94	keep pauses, any glyph
ELL_Q	90	keep both … and ?
TRAIL_OFF	89	utterance-final mark
EMOTIVE_PAUSE	86	keep affective pause
SILENCE	74	silent turn stays mark-only
MULTI_PAUSE	59	preserve pause count
ELL_EX	58	keep both … and !
QUOTE_OMISSION	41	editorial … inside quotes
MID_PHRASE	10	pause stays medial

Glyph variants (…, ……, ...) are rotated row-by-row so category is not confounded with one Unicode form. Every prompt explicitly instructs preservation.

Scoring details. Glyph variants are canonicalized before category-specific checks over a localized focus span. The focus span is found by line alignment against the source (72.2% of judgments), fuzzy match against the reference focus (23.6%), or whole-output fallback (4.2%). When localization fails and the context itself contains an ellipsis, a would-be PASS is escalated rather than credited. Word-based compensation such as "he paused" is escalated rather than automatically failed.

Rubric example (ELL_Q, mark+?).

KO "이제 다 끝났어." / "정말…… 끝난 거야?"
REF "It's all over now." / "Really... is it over?"
PASS TRIVE: "Is it really... over?" — mark and ? both kept, medial position tolerated.
FAIL (E3_FUNC_SHIFT) Gemma 4-31B: "It's all over now." — the question turn is absent entirely.
FAIL (E4_COMBO_LOSS) GPT OSS 20B: "What's all of this sudden?" — question kept, mark stripped.

Additional diagnostics: utterance-final ellipses are hardest (0.81 vs. 0.93 medial); the six-dot form …… (0.87) and ASCII . . . (0.88) are harder than single … (0.92); and the corpus arm is easier than the synthetic arm (0.95 vs. 0.88). Word-based compensation is nearly absent (≤5/601 per system). Sixteen of 23 systems prefer ASCII . . . over U+2026.

A.2 PARTICLES: Stacked Particles (579 items)

Korean post-nominal particles stack under ordering constraints (Lim, 2004), and each layer contributes meaning. Scalar particles such as 도/까지/조차/마저 are semantically distinct (Han, 2010) even though they often gravitate toward English *even*. Each particle is annotated with a semantic feature informed by K-SNACS (Hwang et al., 2020).

The design contains 172 minimal-variant groups. Each group fixes a base frame and host noun while varying only the particle stack at depth 1–5. Stacks were checked for well-formedness and ordering with the Kiwi analyser (Lee, 2024); carrier contexts were LLM-written and reviewed.

Each row separates *checkable* features (11 of 17 labels, with a reliable English cue lexicon; 564/579 items) from weak features such as case or contrastive topic, which go to adjudication. The rule score is the fraction of checkable cues realized in the focus sentence. For every pair of variants whose feature sets differ, we additionally test whether the English focus sentences are identical, yielding the contrast-collapse rate over 724 pairs.

Rubric example (group `grp_0001`, host 선생님 'teacher').

도 *-to* (ADDITIVE) → REF "The teacher **also** entered…"; Lumen "The teacher also entered…" ✓
까지도 *-kkaci-to* (SCALAR_EXTENT+ADDITIVE) → REF "**Even** the teacher entered…"; Lumen "Even the teacher entered…" ✓
조차도 *-cocha-to* (SCALAR_LEAST_LIKELY+ADDITIVE) → REF "Even the teacher **(of all people)** entered…"; Lumen "Even the teacher entered…" ✓
All three pass the cue checklist, yet the last two English outputs are identical and therefore collapsed.

Contrast-family stacks (은/는) are the hardest family (0.66 vs. additive 0.86), while role-class hosts such as 로서 *-lose* are the hardest host class (0.59). Bare-*even* rate correlates $r = 0.82$ with checklist score across systems, partly mechanically. The reference attains the highest checklist score despite the lowest bare-*even* rate (0.43). Full-scale judge validation further shows that rule-PASS depth-5 rows average only 0.62 under the judge, motivating future differentiator-aware cue features.

A.3 WORDPLAY: Wordplay (65 items)

Fifteen items are attested puns with published professional translations from the Korean translation-studies literature; 50 are expert-authored. Items cover native and Sino-Korean homophones, idiom literalization, polysemy, and culture-specific play. Each item records the pun pivot and sense inventory, following SemEval-style annotation (Miller et al., 2017), and a curated reference strategy in Delabastita's typology (Delabastita, 1994, 1996): pun-to-non-pun (33), new pun (14), literal-plus-explanation (11), or target-language pun (7).

The fixed rubric assigns 1.0 when wordplay is recreated or both senses are conveyed with humor-

ous effect, 0.5 when one sense is faithful and the other is recoverable, and 0.0 when the pun is ignored or the pivot mistranslated. Judges receive the pivot, its sense inventory, and the curated reference marked as one acceptable strategy. Structured output includes both strategy label and score.

Rubric example (pivot 꼬리/이야기, *kkoli/iyaki*, 'tail'/'tale').

KO "꼬리에 꼬리를 무는 길고도 슬픈 이야기지!" / "꼬리가 길긴 하네."
REF "It's a long sad tail-of-a-tale, one tail biting another!" / "That's certainly a long tail."
1.0 Command A+: "...a long, sad tale that just keeps dragging on!" / "Well, the tail is long..."
0.0 Cohere CAT+: "...a long and sad tale that keeps going on and on!" / "Well, it certainly is a long one."
The same Command A+ output receives 1.0 from one admitted judge and 0.5 from another, motivating a panel median.

A.4 Topic/Subject Marking (387 items)

In the canonical contrast tested here, Korean *i/ga* (이/가) marks a subject that answers the current question, whereas *eun/neun* (은/는) establishes or contrasts a discourse topic. English may realise the distinction through clefts, prosodic emphasis, *as for*, contrastive connectives, or contextually licensed neutral SVO, making simple keyword rules unreliable.

We instantiate 195 controlled lexical frames and vary the question under discussion together with the marked subject while holding the proposition fixed. The resulting 387 items contain 194 *ga*-focus and 193 *neun*-contrast items; 192 frames form complete minimal pairs.

Minimal-pair example (subject focus vs. contrastive topic).

Context "누가 왔어?" / KO "철수가 왔어."
REF "Cheolsu came." or "It was Cheolsu who came."
Context "지수는 안 왔어. 철수는?" / KO "철수는 왔어."
REF "As for Cheolsu, he came." / "Cheolsu did come." / "Cheolsu came, though."
A plain "Cheolsu came" can be licensed by either context; a who-cleft used as the only rendering of the second member does not itself preserve the contrast.

The structured judge records proposition correctness, realized focus span, required-feature labels, and information-structure status. Code assigns PASS only when proposition and required features are preserved with no spurious feature; PARTIAL covers a weakened distinction; and FAIL covers wrong focus/contrast assignment, direction reversal, or proposition error. A separate joint judgment over 100 selected pairs yields pairwise contrast accuracy and semantic-collapse rate.

A.5 Evidentiality (400 items)

Korean *-teora* (-더라) is tested in direct-perceptual contexts, colloquial *-dae* (-대) as reportative (Jang, 2025; Chae, 2015), *-(eul) geot gatda* (-(으)ㄹ 것 같다) as inferential/epistemic, and a plain assertion as control. One hundred propositional frames are each realized as a four-member set.

Table 4: Balanced evidentiality strata.

Evidence type	n	Required interpretation
direct perception (-더라)	100	_saw / noticed / when I checked_
reportative (-대)	100	_they say / I heard / apparently_
inference (것 같다)	100	_looks like / seems / probably_
unspecified assertion	100	no added source or uncertainty

The judge records proposition correctness independently from predicted evidence type and flags conflicting source shifts. PASS requires the correct proposition and exact evidence class; a marked source weakened to an unmarked assertion is PARTIAL; source replacement or proposition error is FAIL.

Rubric example (direct perception vs. source loss/shift).

Context "직접 봤어?" / KO "비가 오더라."
PASS "I saw that it was raining."
PARTIAL "It was raining."
FAIL "I heard that it was raining."

A.6 Scrambling (599 items)

Korean canonical transitive order is SOV, but overt case particles permit several non-canonical orders. English typically restores the same roles in SVO order, so the task is semantic transfer rather than source-order imitation (Park, 1995). Scrambling is also discourse-sensitive (Lee, 2015).

We create 100 three-argument transfer frames annotated with subject (S), indirect object/recipient (IO), object/theme (O), and verb, then generate canonical, argument-fronting, subject-after-object, right-dislocation, and contrastive-object variants. Ninety-nine frames contain all six variants and one contains five, for 599 items.

Table 5: Scrambling variants.

Variant	n	Korean order	Focused role
canonical	100	S–IO–O–V	S
object fronting	100	O–S–IO–V	O
IO fronting	100	IO–S–O–V	IO
subject after object	100	O–IO–S–V	S
right dislocation	100	S–IO–V–O	O
contrastive object	99	O–S–IO–V	O

The structured judge fills a typed assignment for every gold role: gold entity, corresponding candidate span or MISSING, and same-entity status. Missing arguments, wrong recipients, or role reversals are failures even when the remaining English is fluent. When all roles are correct, discourse preservation determines PASS versus PARTIAL.

Rubric example (object focus after scrambling).

Context "철수는 현우에게 뭘 주었어?" / KO "책을 철수가 현우에게 주었어."
PASS "Cheolsu gave Hyunwoo a book."
PASS "It was a book that Cheolsu gave Hyunwoo."
FAIL "Hyunwoo gave Cheolsu a book."

A.7 Auxiliary Constructions (541 items)

Korean productive V–V sequences combine a lexical predicate with secondary verbs after *-a/eo*; in auxiliary use these verbs contribute meanings such as attempt, experience, benefactive, preparation, unwanted outcome, and achievement (Hwang and Whitman, 2021). English typically lexicalises these meanings with constructions such as *managed to*, *ahead of time*, or *as a favour*.

The released suite isolates one auxiliary construction per item. Of 91 lexical frames, 88 contain all six operator variants, two contain five, and one contains three, yielding 541 items.

Table 6: Auxiliary-construction inventory.

Subtype	n	Auxiliary	Required semantics
completion/regret	91	버리다	completion + regret/unexpectedness
preparatory retention	91	두다	in advance + retained result
benefactive	90	주다	action performed for another
unwanted result	90	말다	unwanted outcome + resignation
manage to	90	내다	achievement + success
try/experience	89	보다	attempt + experience

The judge evaluates the main predicate, each exact semantic operator, added operators, and operator scope separately. PASS requires complete operator preservation with correct scope; PARTIAL covers weakening or deletion with the core proposition intact; and reversed meaning, polarity, modality, or main predicate is FAIL.

Rubric example (achievement/success operator).

Context "어렵다며, 결국 성공했어?" / KO "철수는 케이크를 먹어 냈어."
PASS "Cheolsu managed to eat the cake."
PARTIAL "Cheolsu ate the cake."
FAIL "Cheolsu failed to eat the cake."

A.8 Connective Endings (194 items)

Korean clausal connective endings fuse discourse-relation marking into a bound morpheme. Each item is labelled with a PDTB 3.0 discourse relation (Webber et al., 2019). Minimal variants hold the clause pair fixed while rotating only the connective ending.

여생을 보내기로 작정**해서** 재산을 나눠주었다
→ *Because he decided..., he gave away...*
여생을 보내기로 작정**하면서** 재산을 나눠주었다 → *While deciding..., he gave away...*
여생을 보내기로 작정**하고** 재산을 나눠주었다
→ *Having decided..., he gave away...*
여생을 보내기로 작정**하자** 재산을 나눠주었다
→ *As soon as he decided..., he gave away...*

Table 7: Connective-ending relation inventory.

Relation	n	Marker example	Required distinction
CONTINGENCY-Condition	13	-(으)면	first clause is a condition for the second
TEMPORAL-Synchronous	12	-(으)면서	simultaneous non-backgrounded

Three LLMs also assigned a four-point source-acceptability rubric, yielding 16 primary, 139 marginal, and 39 excluded items. Inter-judge agreement on tier assignment was low (Fleiss $\kappa \approx 0.05$), so tiering is treated as a limitation rather than hidden. A post-hoc analysis over all 194 items preserves the load-bearing ranking claims but produces larger mid-table shifts, up to 12 ranks for some systems.

A 70-item held-out calibration set shows that pointwise majority prediction already reaches 27–29% accuracy. We therefore report $\mathrm{SRP}_{\mathrm{JSD}}$ over multi-variant frames rather than relying on accuracy alone.

Rubric example (CONTINGENCY-Condition, -면).

KO "비치면 이케 왔다 갔다 하거덩."
REF "If the light shines on it, it moves back and forth like this."
PASS Lumen: "When you unfold it, it goes back and forth like this." — relation preserved despite a lexical error.
FAIL Command A+: "It would be like going back and forth like that." — no connective is realized.

A.9 Ideophones (170 items)

Korean ideophones encode manner and intensity iconically; fine phonological changes can shift repetition, force, or speed (Kim, 1977). We define seven coarse semantic categories and embed one ideophone from a 52-item inventory in a fixed carrier sentence, rotating only the ideophone across variants of the same category.

Tier construction follows the same three-judge acceptability family as connectives: 119 primary / 45 marginal / 6 exclude. The fine-grained target is a label such as PRECIPITATION-high, and frame-level $\mathrm{SRP}_{\mathrm{JSD}}$ measures whether fine distinctions remain distributed as in the gold frame.

Table 8: Ideophone category inventory. Last three rows (n=4–5) are exploratory only.

Coarse category	n	Example pair	Intensity axis
MOTION-REPEATED	54	깡총깡총 vs. 폴짝폴짝	force of repeated motion
MANNER-PACE	40	천천히 vs. 빨리빨리	speed/manner of an action
EMOTION-BODY	29	두근두근 vs. 벌렁벌렁	intensity of a bodily/affective state
LIGHT-VISUAL	28	반짝반짝 vs. 번쩍번쩍	brightness/suddenness of a visual effect
SOUND-IMPACT	5	쿵 vs. 쾅	force of an impact sound
TEXTURE	4	보들보들 vs. 까슬까슬	surface texture quality
PRECIPITATION	4	보슬보슬 vs. 주룩주룩	intensity of rainfall

Rubric example: category and intensity confusion.

For 살랑살랑 (MOTION-REPEATED.LOW), Tower 9B's "while gently moving its front paws" preserves low-intensity repeated motion, while a classifier reads Qwen 3.6-27B's equally plausible "gently wiggled" as the wrong category. For 두근두근 (EMOTION-BODY.MEDIUM), a faithful "thump-thump" rendering preserves medium intensity, while an intensified cue elsewhere in the sentence can push the classifier to a high-intensity label. These examples motivate classifier validation and frame-level reporting rather than blind reliance on pointwise labels.

B Structured Semantic Evaluation Protocol

Topic/subject marking, evidentiality, scrambling, and auxiliaries use one shared structured protocol. We draw 200 items per phenomenon with seed 42, stratified by subtype; topic/subject instead samples 100 complete minimal pairs. The same source sample is scored for all 23 systems, yielding 18,400 item judgments, and the topic pairs add 2,300 joint judgments.

Qwen3.6-27B receives the Korean context and source, references, the annotated feature inventory, and one normalized candidate, but not the system name. Temperature is zero and output is strict JSON. Deterministic code, not the judge, maps observations to three-level scores. Free-form `extra_features` are diagnostic only unless they match an item's declared `forbidden_features`; translating the supplied context is not penalized.

Before system scoring, curated-reference positives and mismatched-reference negatives probe the same contract. Full-release counts describe construction; system scores on these four subsets estimate performance on the shared 200-item samples.

C Judge Models as Measurement Instruments

All wordplay judging uses unmodified open-weight checkpoints through one OpenAI-compatible endpoint with a fixed protocol: system+user prompt, temperature 0, strict-JSON request, and 2000-token budget. Candidate judges are EXAONE-4.0-32B (LG AI Research, 2025), Qwen3.6-27B (Qwen Team, 2026), Gemma4-31B (Gemma Team, 2026), and gpt-oss-20b (OpenAI, 2025). Three are themselves among the 23 evaluated systems; EXAONE is not.

C.1 Wordplay Judge Admission

With no human MQM labels for wordplay, each judge is validated against items whose score is known a priori: curated references whose expected rubric stratum follows from strategy labels, plus mismatched-reference negative anchors. A judge is admitted if it (i) orders positive strata as expected, (ii) keeps references at or below the 0.746 rubric ceiling, and (iii) scores negatives below 0.1.

Table 9: Wordplay judge admission.

Anchor (expected)	EXAONE-4.0-32B	Qwen3.6-27B	Gemma4-31B	gpt-oss-20b
new pun (~1.0)	1.00	0.80	0.86	0.77
target-lang pun (~1.0)	1.00	0.33	0.86	0.58
literal+expl. (0.5–1.0)	1.00	0.62	0.64	0.59
pun-to-non-pun (~0.5)	1.00	0.26	0.32	0.43
mismatched (~0.0)	0.01	0.02	0.00	0.00
reference overall (ceiling 0.746)	1.00	0.47	0.55	0.55
admitted	**no**	**yes**	**yes**	**yes**

EXAONE rejects negatives but assigns 1.00 to every positive stratum and the reference overall, making it unusable for ranking. Gemma4-31B, gpt-oss-20b, and Qwen3.6-27B reproduce the expected ordering, keep the reference below the ceiling (0.55, 0.55, 0.48), and score negatives at 0.00–0.02.

C.2 Behaviour Profiles and Agreement

EXAONE places 77% of its score mass on 1.0, whereas the three admitted judges place 69–83% on 0.0. Output-format attrition varies from 0% (EXAONE, Gemma4-31B) to 29% (Qwen3.6-27B), showing that a fixed judge protocol is not model-agnostic.

Because each judge returns a strategy label as well as a score, we can separate valuation from perception disagreement.

The label-to-score mappings are similar, but the judges agree on the strategy label in only 42% of judgments and on the score in 60%. Of 399 score disagreements, 41% use the same label but different score, while 59% use different labels. The share of outputs perceived as containing surviving wordplay ranges from 3.2% to 9.8%.

The three judges that are also evaluated systems show negative mean preference residuals toward their own model families (Gemma4-31B: -0.34;

Table 10: Judge behaviour on shared WORDPLAY judgments. Attrition is the share with no parseable score under the fixed protocol.

Judge	Scored system?	Attrition	Mean	%0	%.5	%1	Kin Δz	max|Δz|
EXAONE-4.0-32B	no	0%	0.82	13%	11%	76%	—	1.34
Qwen3.6-27B	yes	29%	0.15	75%	19%	6%	-0.10	1.04
Gemma4-31B	yes	0%	0.14	83%	7%	10%	-0.34	1.11
gpt-oss-20b	yes	5%	0.21	69%	20%	12%	-0.10	1.13

Table 11: Mean score each admitted judge assigns given its own strategy label, with perceived surviving-wordplay rate.

Judge label	Gemma4-31B	gpt-oss-20b	Qwen3.6-27B
pun-to-pun	1.00 (60)	0.96 (138)	1.00 (20)
new pun	1.00 (21)	0.58 (12)	1.00 (13)
literal + explanation	0.59 (51)	0.36 (153)	0.34 (120)
non-pun paraphrase	0.10 (961)	0.14 (734)	0.16 (529)
omission	0.00 (110)	0.02 (241)	0.00 (187)
mistranslation	0.00 (292)	0.00 (137)	0.00 (180)
Wordplay detected	**4.7%**	**9.8%**	**3.2%**

Table 12: Pairwise judge agreement on shared WORDPLAY judgments.

Judge pair	n	Exact	Pearson	Rank ρ
exaone4-32b / qwen3.6-27b	1056	0.19	0.26	0.85
exaone4-32b / gemma4-31b	1495	0.22	0.18	0.87
exaone4-32b / gpt-oss-20b	1416	0.24	0.23	0.81
qwen3.6-27b / gemma4-31b	1056	0.79	0.55	0.83
qwen3.6-27b / gpt-oss-20b	1008	0.67	0.36	0.78
gemma4-31b / gpt-oss-20b	1416	0.70	0.38	0.81

gpt-oss-20b: -0.10; Qwen3.6-27B: -0.10). No kin inflation is detectable in this panel, although this does not rule out subtler bias.

C.3 Connective and Ideophone Classifier Admission

For connectives, a rule-based keyword classifier and two LLM candidates are compared on a 70-item calibration set stratified by cue type.

Table 13: Connective-ending classifier admission.

Cue type (n)	keyword (rule)	Qwen2.5-7B	Qwen3.6-27B
explicit (43)	**98%**	79%	**98%**
implicit (14)	57%	57%	86%
ambiguous marker (13)	85%	69%	62%
Overall (70)	87%	73%	**89%**

The rule-based classifier is adopted as primary. Qwen3.6-27B's marginally higher calibration score could not be validated at scale before the shared inference endpoint was withdrawn.

For ideophones, no comparable rule baseline exists, so the reference-as-pseudo-system protocol is used.

Table 14: Ideophone classifier admission via reference-as-pseudo-system calibration.

Classifier	Ref. SRP	Ref. pointwise	Role
Qwen2.5-7B	0.508	43.7%	rejected – inside system range
Gemini-3.5-Flash-Lite	**0.733**	**63.9%**	admitted – primary

Gemini-3.5-Flash-Lite is adopted as primary, with two independent classifiers retained for cross-validation.

C.4 Bias Probes

Across admitted wordplay judges, the within-item Spearman correlation between score and hypothesis length is small but positive ($\rho = 0.14$–0.25); correlation with chrF-style character 1–6-gram similarity to the reference is likewise weak ($\rho = 0.10$–0.21). Hypotheses with a parenthetical gloss score 0.12–0.21 points higher, but this is not an identified formatting effect because a gloss is legitimate evidence for the literal-plus-explanation strategy. These analyses are observational, not causal.

EXAONE's label behaviour reveals a different issue. It uses all seven connective labels but assigns EXPANSION-BACKGROUND to 31.9% of outputs against 11.6% in the gold distribution, whereas the other three classifiers stay within $1.1\times$ of gold. This is best interpreted as a systematic default reading rather than simple label insensitivity.

D Item Quality and Psychometrics

Following Vania et al. (2021); Rodriguez et al. (2021), we compute item difficulty, item–rest discrimination, and coefficient α (Cronbach, 1951).

ELLIPSIS has $\alpha = 0.996$, but 30% of items are saturated (mean > 0.95). QUOTE_OMISSION, ELL_Q, and EMOTIVE_PAUSE discriminate strongly (mean 0.87/0.84/0.80), whereas SILENCE, HESITATION, and TRAIL_OFF contribute little (0.05/0.11/0.20).

PARTICLES has $\alpha = 0.972$, median discrimination 0.27, and 31% saturation. WORDPLAY, using per-item median scores from admitted judges, has $\alpha = 0.896$ and median discrimination 0.37; 37/65 items have discrimination ≥ 0.2, but 20/65 are nearly dead (mean < 0.05).

Connectives and ideophones are scored at the frame level, so their psychometrics use roughly 37 multi-variant frames rather than hundreds of items. Their α values are 0.755 and 0.664, neither subset saturates, and median discrimination is 0.251 and 0.231. Connectives have one near-dead frame, whereas ideophones have nine.

With only 23 systems, item–rest correlations are noisy, and α is inflated by both test length and dependence within minimal-variant groups. We therefore use these statistics primarily for suite revision rather than fine claims about individual items.

E Additional Behavioural Analyses

E.1 System Fingerprints and Families

Representing each system by its normalized error-code distribution (nine codes plus per-subset pass rates) and clustering with cosine distance, all three same-family pairs—Gemma 4-4B/31B, Qwen 3.5/3.6, and GPT-OSS 20B/120B—are mutual nearest neighbours. Under label permutation of the observed distance matrix, the chance of this configuration is approximately 1.4×10^{-5}. The signal is not family-specific, however: the space contains seven mutual-nearest-neighbour pairs, four of them cross-family, including the tightest pair, Lumen–Wayfinder. Two features are pass rates, so part of the signal reflects matched competence rather than lineage. Dubformer's nearest profile is the REFERENCE.

Distinctive signatures include GoogleTranslate as the only substantial ellipsis overgenerator (34 inserted ellipses), Ministral and Qwen systems skewing toward scope shift, and the Gemma family being uniquely dominated by deletion plus function shift.

E.2 The Bare-*Even* Funnel

Pooling all systems, 조차 *-cocha* and 마저 *-mace* are realized as bare *even* in 88.6% and 87.9% of cases. 도 *-to* is bare *even* in 70% and *also/too* in 24%, while 까지 *-kkaci* has a broader distribution (37% *even*, 36% *only/just*, 12% *up-to/until*). Curated references use differentiators in 57% of cases; no evaluated system except Gemini 3.1 Pro (11.6%) exceeds 5%.

Because *even* satisfies the cue lexicon for multiple scalar labels, funneling is rewarded by construction. Bare-*even* rate correlates $r = 0.82$ with checklist score, though this is partly mechanical and partly confounded with general competence. The appropriate interpretation is therefore joint: checklist fidelity says whether required cues are present, while collapse and differentiator analyses say whether distinctions remain.

E.3 Additional Subset Effects

On PARTICLES, contrast collapse is highest for the shallowest contrasting pairs (0.24 at depth 2 vs. 0.17–0.18 at depths 3–5). Depth and contrast are therefore separate difficulties: deeper stacks lower cue accuracy, but shallower contrasts are more readily flattened into generic English.

On WORDPLAY, idiom literalization is the most tractable category, while Sino-Korean homophones and culture-specific items are hardest. Items whose curated reference itself resorts to pun-to-non-pun are also hardest for systems, giving a curator–system consistency signal.