

Beyond Prompts: A Systematic Study of LLM-Based Machine Translation Evaluators

Dmitry Popov[♣], Roman Bokhyan[◇], Ekaterina Enikeeva[♣],
Artem Mekhraliev[♣], Stepan Vysotsky[♣], Nikolay Karpachev[♣]
♣ Yandex, ◇ HSE University

Abstract

LLM-based evaluators have rapidly become a standard tool for machine translation evaluation, yet the design choices that shape their behavior remain poorly understood. We present a systematic study of how instruction design, few-shot demonstrations, backbone model choice, and multi-step workflows affect span-level and ranking-based MT evaluation. We collect new expert RATE annotations on WMT24 and WMT25 English–Russian translations, with transfer validation on MQM- and ESA-annotated benchmarks across five additional language pairs. Our experiments reveal four robust findings: instruction structure shapes evaluator behavior, high-quality demonstrations matter more than quantity, frontier backbones exhibit distinct evaluator profiles (recall-, precision-, or ranking-oriented), and fusing outputs across diverse backbones outperforms single-evaluator multi-call setups. Although evaluator design exposes a large combinatorial space, several discovered design principles transfer consistently across annotation protocols and language pairs. We synthesize these findings into ORBIT, a family of LLM-based MT evaluators. Our strongest configurations achieve expert-level performance in error span detection and in Accuracy and Fluency ranking at roughly 20× lower cost than expert human annotation.

1 Introduction

Machine translation (MT) quality has improved rapidly in recent years, with frontier LLM-based systems approaching professional human translation quality on many language pairs and domains (Kocmi et al., 2024a). As translation quality improves, evaluation increasingly becomes the bottleneck: identifying subtle translation errors requires trained annotators, substantial annotation time, and careful quality control. Automatic metrics remain useful at scale, but still lag behind expert human

System	Span-level					Ranking
	R	P	F1	R ₄₊	P ₄₊	MQM
Human _E	46.8	44.3	45.5	65.3	65.3	75.3
Human _{A-SbS}	39.3	47.1	42.8	55.7	63.2	68.0
Human _{A-PW}	31.8	46.4	37.7	49.8	57.1	65.0
Human _{WMT}	12.3	61.9	20.6	26.6	62.8	42.7
GEMBA-MQM	24.6	43.0	31.3	45.0	51.0	58.7
GEMBA-ESA	24.9	41.3	31.0	44.6	45.7	57.3
MQM-AE	24.0	51.8	32.8	45.4	54.5	60.0
ESA-AE	27.0	54.8	36.2	49.5	58.6	53.7
UMB-v5	27.8	44.5	34.2	48.7	56.4	57.3
xCOMET-XXL	18.3	36.4	24.3	33.5	36.8	58.3
ORBIT-SC	51.5	46.2	48.7	74.7	63.5	67.0
ORBIT-MC	60.3	45.8	52.1	82.0	70.6	69.3

Table 1: Overview of human baselines, automatic baselines, and our configurations. Prompt-based baselines and ORBIT-SC use GPT-5.4. ORBIT-MC is a multi-call setup that fuses three evaluator backbones with an LLM fuser.

judgment on fine-grained translation assessment, particularly for span-level error localization and diagnostic analysis (Lavie et al., 2025; Freitag et al., 2021).

Large language models have recently emerged as strong MT evaluators (Kocmi and Federmann, 2023b). Prior work has explored both score-based approaches, which predict a single quality score for a translation (Juraska et al., 2024), and span-based approaches, which attempt to identify concrete translation errors following human evaluation protocols such as MQM (Freitag et al., 2021) and ESA (Kocmi et al., 2024b). However, evaluator design choices — prompt structure, demonstrations, workflow decomposition, and aggregation — are typically treated as implementation details, despite substantial evidence from in-context learning research that demonstration design strongly affects downstream behavior (Liu et al., 2022; Agrawal et al., 2023). PrExMe explores prompt variation for

scalar-score MT evaluation (Leiter and Eger, 2024). We instead study protocol-grounded evaluator design for fine-grained span-level and ranking-based assessment against expert annotations.

This leaves a central question largely unexplored: which evaluator design choices robustly improve LLM-based MT evaluation quality across datasets, protocols, and annotator settings?

In this work, we study LLM-based MT evaluation as a structured evaluator design problem rather than as a fixed prompting configuration. We conduct a systematic study across four major design dimensions: instruction design, few-shot demonstration design, backbone model selection, and multi-step evaluator workflows. Our focus is span-level MT evaluation under the RATE protocol (Popov et al., 2025), which provides fine-grained error spans, severity annotations, and separate Accuracy and Fluency judgments.

To support this study, we collect new expert annotations on English–Russian WMT24 and WMT25 translations under the RATE protocol. Each translation is independently annotated by multiple professional linguists, enabling direct evaluation of error span detection against human annotations rather than relying only on scalar quality-score correlations. Span-level evaluation is particularly important for diagnostic MT assessment, since it localizes concrete translation failures and provides actionable feedback for both system developers and human editors.

Our experiments reveal four robust findings: instruction structure shapes evaluator behavior, high-quality demonstrations matter more than quantity, frontier backbones exhibit distinct evaluator profiles (recall-, precision-, or ranking-oriented), and multi-model fusion substantially outperforms single-call evaluation. Although evaluator design exposes a large combinatorial space, several discovered design principles transfer consistently across annotation protocols and language pairs. Across these experiments, our strongest configurations consistently exceed assessor-level agreement and match expert-level error span detection performance (Table 1).

We encode these design principles in ORBIT, a family of evaluator configurations instantiating the most robust choices identified in our study. We refer to the resulting single-call and multi-call configurations as ORBIT-SC and ORBIT-MC respectively. Our strongest configurations achieve expert-level error span detection performance. ORBIT-

MC also reaches expert-level Accuracy and Fluency ranking (Table 7) and exceeds assessor agreement on MQM ranking (Table 1). We additionally provide a preliminary study of how expert annotators interact with LLM-generated reports in collaborative annotation workflows (Appendix H).

Our contributions are: (i) A systematic study of evaluator design for fine-grained LLM-based MT evaluation. (ii) Evidence that several evaluator design principles transfer across annotation protocols (RATE, MQM, ESA), annotator pools, and five additional WMT language directions: en-de, zh-en, cs-de, en-it, and en-zh. (iii) New expert RATE annotations on WMT24 and WMT25 English–Russian translations spanning multiple annotator expertise levels.¹ (iv) ORBIT, a family of single- and multi-call MT evaluators whose strongest configuration achieves expert-level error span detection performance and surpasses assessor-level translation ranking while costing roughly $20\times$ less than expert human annotation.

2 Background and Related Work

2.1 Human MT Evaluation Protocols

MQM (Lommel et al., 2014; Freitag et al., 2021) and ESA (Kocmi et al., 2024b) serve as de facto gold standards in WMT evaluation campaigns. Both ask annotators to identify error spans in translations. MQM additionally assigns each span an error category and severity, which are aggregated into a final score, while ESA removes categories and instead uses spans as guidance for a final 0–100 quality score.

RATE (Popov et al., 2025) extends this setup with five-point severity labels, a streamlined semantic/non-semantic taxonomy, separate Accuracy and Fluency scores, and an overall score derived from span severities.

2.2 LLM-Based MT Evaluation

LLM-based evaluation has emerged as a general paradigm for natural language assessment (Zheng et al., 2023; Gu et al., 2026; Liu et al., 2023; Bavaresco et al., 2025). In MT evaluation, existing approaches can be broadly divided into score-based, span-based, and supervised methods.

Score-based approaches such as GEMBA-DA (Kocmi and Federmann, 2023b) and MetricX (Juraska et al., 2024) predict a single qual-

¹<https://huggingface.co/datasets/yandex/orbit-mt-eval>

ity score for a translation segment. Prompt-based span evaluators more closely follow human annotation workflows by asking LLMs to identify translation errors directly. Canonical examples include GEMBA-MQM (Kocmi and Federmann, 2023a) and AutoMQM (Fernandes et al., 2023), which prompt LLMs using MQM-style instructions and a small set of demonstrations. Subsequent work extends this paradigm with decomposition, post-editing validation, and rubric-based prompting (Lu et al., 2025; Kim, 2025; Lu et al., 2024). Multi-agent approaches such as M-MAD (Feng et al., 2025) further decompose evaluation into interacting specialist agents.

PrExMe (Leiter and Eger, 2024) systematically explores more than 720 prompt templates for open-source LLM-based MT and summarization metrics, showing that seemingly minor prompt and output-format choices can alter metric behavior. Our study has a different focus: we measure direct agreement with multiple professional experts on error localization and severity, and jointly study demonstrations, backbone choice, and multi-call workflows in addition to instruction variation.

Supervised approaches instead train models on MQM-style annotations to predict scores, spans, or both. For example, xCOMET (Guerreiro et al., 2024) combines sentence-level scoring with span detection, while recent work such as GemSpanEval (Juraska et al., 2025) adapts frontier LLMs to span-level evaluation tasks.

Span-level results are sensitive to annotation quality and annotator expertise, while systematic comparisons against multiple professional experts remain limited (Freitag et al., 2021; Lavie et al., 2025).

Concurrent work by Perrella et al. (2026) proposes alternative span-matching formulations, which we discuss in Appendix E.

3 Experimental Setup

3.1 Design Study Protocol

Evaluator design exposes a large combinatorial space of potentially interacting decisions, including instruction structure, few-shot demonstration selection, workflow decomposition, aggregation, and backbone model choice. Exhaustive joint optimization across all dimensions is impractical. We therefore study design dimensions sequentially, fixing variables after each stage to isolate the effect of subsequent decisions. Our objective is not to

find a universally optimal evaluator configuration, but to identify design choices that consistently improve span-level MT evaluation quality and transfer across protocols and language pairs.

3.2 Dataset and Annotation Setup

We collect new annotations on 600 source-translation pairs from WMT24 (Kocmi et al., 2024a) and WMT25 (Kocmi et al., 2025) English-Russian translations, extending the RATE-annotated coverage to newer frontier LLM translation systems. Detailed dataset statistics are in Appendix A.

Sources and translations. Source segments are sampled from WMT24 and WMT25 across multiple domains. We select 300 source segments and pair each with two translations, enabling segment-level pairwise ranking accuracy (Section 3.4) alongside span-level evaluation. Rather than sampling uniformly, we apply diversity sampling to ensure coverage across translation systems and quality levels, producing a challenging evaluation set.

Annotation layers and human baselines. Each translation receives annotation at three levels of expertise. Professional expert linguists annotate the dataset under the full RATE protocol in implicit side-by-side mode, yielding three independent expert annotations per segment (**Expert** $\times$ **3**). Assessors — Russian-native annotators with verified C1 English proficiency but no formal translation training — follow the same protocol (**Assessor**). We additionally include standard ESA annotations from the WMT evaluation campaign (**WMT annotator**), providing a link to the public benchmark setting.

For evaluation, we keep two of the three expert annotations as gold references. Human_E represents pairwise expert agreement, computed by evaluating the third expert against each gold annotation and averaging — the same procedure used for all evaluated systems. Human_{A-SbS} and Human_{A-PW} correspond to assessors evaluated in implicit side-by-side and point-wise setups respectively, while Human_{WMT} follows standard ESA annotation conditions. Full annotator criteria and statistics are in Appendix C.

Few-shot pool. Separately from the evaluation set, we construct a dedicated few-shot calibration pool consisting of 8 source segments across multiple domains, each translated by 12 systems. Pro-

fessional linguists annotate the pool collaboratively under the RATE protocol with consensus reconciliation, yielding more reliable demonstrations than single-annotator outputs. This pool is used exclusively for few-shot selection (Section 4.2) and does not overlap with the evaluation set.

Multilingual validation. To test whether discovered evaluator design choices transfer beyond English–Russian and the RATE protocol, we additionally evaluate our best single-call configuration on five WMT language directions with publicly available annotations. From the WMT23 Metrics Shared Task (Freitag et al., 2023), we use MQM annotations on en-de and zh-en, sampling 450 examples per direction. From the WMT25 evaluation shared task (Lavie et al., 2025), we use ESA annotations on cs-de, en-it, and en-zh, sampling 300 examples per direction. Full results are in Section 5.

3.3 LLM Baselines

We compare against baselines representing several major evaluator paradigms: concise prompting, detailed protocol prompting, decomposition-based workflows, and supervised span-level evaluation.

Prompt-based baselines include GEMBA-MQM (Kocmi and Federmann, 2023a), GEMBA-ESA, a reference-free LLM judge used in the WMT25 General MT shared task (Kocmi et al., 2025), and UMB-v5, our abbreviation for the unified-mqm-boosted-v5 schema released with Perrella et al. (2026). The latter is a detailed prompt that thoroughly describes the annotation task. We additionally include MQM-AE and ESA-AE as our own adaptations of MQM and ESA human annotation guidelines into LLM prompts, keeping deviations from the originals minimal. As a supervised baseline we include xCOMET-XXL (Guerreiro et al., 2024), trained on MQM-style annotations to predict both scores and spans. All evaluation prompts are available in our [Hugging Face release](#). Omitted methods are discussed in Appendix D.

3.4 Evaluation Metrics

We evaluate MT evaluators along two dimensions: span-level error detection and segment-level pairwise ranking accuracy. Throughout the paper, MQM scores are derived from RATE annotations following Popov et al. (2025): RATE severity levels 2–3 map to MQM minor errors and levels 4–5

to major errors. As in the original RATE-to-MQM conversion, RATE severity 1 (“negligible or arguable” errors) is excluded because standard MQM contains no equivalent severity level.

Span-level metrics. Predicted error spans are matched to gold annotations via bipartite matching with character overlap (MP with $\tau=1$). We report micro-averaged recall R , precision P , and harmonic mean F_1 , following Perrella et al. (2026). Segment-level macro-averaging gives equal weight to each segment regardless of its number of annotated errors and can therefore overstate performance for low-recall evaluators. To evaluate severe-error behavior, we additionally report two complementary diagnostics: R_{4+} , recall over gold spans of severity ≥ 4 regardless of predicted severity, and P_{4+} , precision for predictions assigned severity ≥ 4 against gold major-error spans. For MQM- and ESA-annotated data, R_{4+} and P_{4+} are computed over major and critical errors. We verify robustness under alternative span-matching formulations in Appendix E.

Ranking accuracy. For each pair of translations of the same source, we measure whether the evaluator correctly predicts which translation is better across Accuracy (Acc), Fluency (FL), and MQM dimensions.

System	Span-level					Ranking		
	R	P	F1	R_{4+}	P_{4+}	MQM	Acc	FL
Human _E	46.8	44.3	45.5	65.3	65.3	75.3	67.3	70.0
GEMBA-MQM	19.9	42.1	27.0	39.6	51.1	60.7	–	–
GEMBA-ESA	20.8	39.6	27.3	39.3	53.0	60.7	–	–
MQM-AE	22.2	53.8	31.4	42.9	64.4	63.0	–	–
ESA-AE	26.3	54.5	35.5	47.5	61.9	59.3	–	–
UMB-v5	22.0	47.0	30.0	42.7	57.9	55.0	–	–
RATE-AE _{brief}	26.5	48.3	34.2	46.3	58.6	59.7	61.7	64.3
RATE-AE _{full}	23.9	44.7	31.2	42.2	55.2	58.3	61.5	64.9
RATE-AE _{full+ex}	30.9	53.8	39.3	52.9	66.3	56.7	60.3	65.3

Table 2: Effect of instruction design (zero-shot, GPT-5.2 backbone). GEMBA systems use their original few-shot demonstrations. All other systems are zero-shot.

4 Systematic Study of LLM Evaluation Design

4.1 Instruction Design

How closely should evaluator instructions mirror human annotation protocols? Existing prompt-based evaluators adopt substantially different assumptions: GEMBA-MQM and AutoMQM rely on

concise language-agnostic prompts, while UMB-v5 follows the opposite strategy with a detailed protocol-oriented specification. Neither line of work ablates the instruction content itself, leaving the relationship between protocol detail and evaluator behavior unmeasured.

We compare three families of protocol-adapted prompts (MQM-AE, ESA-AE, RATE-AE), with RATE-AE ablated into three variants of decreasing detail: RATE-AE_{full+ex} contains full category definitions with inline error-span examples (as category-boundary disambiguators, not few-shot demonstrations), RATE-AE_{full} keeps definitions and severity anchors without examples, and RATE-AE_{brief} compresses to a minimal task description, severity scale, and output format. All experiments in this and the following section (Section 4.2) use GPT-5.2 with temperature 0.

Table 2 shows large differences across instruction variants, with span F1 ranging from 27.0 (GEMBA-MQM) to 39.3 (RATE-AE_{full+ex}). MQM- and ESA-style prompts achieve high precision but low recall, including on high-severity errors — indicating genuinely missed problems rather than annotation granularity differences. GEMBA systems exhibit similar behavior despite using few-shot demonstrations. The strongest RATE variant improves span F1 and major-error recovery over the protocol baselines in this GPT-5.2 comparison. Appendix G further shows that all tested RATE variants outperform the protocol baselines on F1 for Claude Opus 4.6 and Gemini 3 Flash. This effect is not specific to RATE-annotated gold: RATE-AE_{brief} also outperforms MQM-AE on F1 and R_{4+} when evaluated against MQM-annotated WMT data (Section 5).

Within the RATE family, the relationship between instruction detail and evaluator quality is non-monotonic: removing the inline examples is costly, with RATE-AE_{full} falling behind even the much shorter RATE-AE_{brief}. This indicates that detailed taxonomy definitions alone overconstrain evaluator behavior, while compact instructions preserving only task-critical guidance generalize more effectively. At the same time, the strong performance of RATE-AE_{full+ex} indicates that detailed protocols remain useful when augmented with concrete examples, foreshadowing the demonstration-based calibration we study in Section 4.2. Ranking accuracy remains stable across variants, indicating that segment-level judgments are less sensitive to instruction structure than to the backbone (Sec-

tion 4.3).

Overall, human annotation guidelines transfer effectively to LLM evaluators, but exhaustive taxonomy descriptions become less important once high-quality demonstrations or compact task specifications are available. We use RATE-AE_{full+ex} as our primary zero-shot setup and revisit this choice in Section 4.2.

Few-shot source	Span-level			Ranking		
	F1	R_{4+}	P_{4+}	MQM	Acc	FL
RATE-AE _{full+ex} (0-shot)	39.3	52.9	66.3	56.7	60.3	65.3
WMT	37.0 _[36.4, 37.7]	50.1 _[49.8, 50.4]	69.6 _[69.2, 70.0]	61.1 _[60.7, 62.0]	64.5 _[63.2, 66.0]	61.5 _[60.7, 62.7]
RATE-WMT24	38.8 _[37.3, 40.3]	52.8 _[51.0, 55.1]	67.9 _[66.3, 70.1]	61.6 _[61.0, 62.3]	64.1 _[63.0, 65.3]	63.9 _[63.7, 64.6]
Ours	41.2 _[40.9, 41.7]	56.8 _[56.0, 57.5]	68.1 _[67.8, 68.7]	62.4 _[60.0, 67.3]	63.4 _[63.0, 64.3]	65.3 _[64.3, 67.6]

Table 3: Effect of few-shot annotation quality. Each value is the mean over 3 runs. Brackets show the corresponding minimum and maximum.

4.2 Few-Shot Design

4.2.1 Does Annotation Quality Matter?

We compare three sources of few-shot demonstrations: WMT25 shared task ESA annotations (manually mapped to the RATE taxonomy and severity scale by our expert annotators), single-expert RATE annotations on WMT24 from Popov et al. (2025), and our collaboratively curated annotations produced by multiple experts with consensus reconciliation.

Table 3 shows that demonstration quality matters substantially more than demonstration presence alone. WMT annotations push F1 below the zero-shot baseline across all three runs (37.0 vs 39.3 mean), suggesting that noisy demonstrations can actively mislead the evaluator about what constitutes a translation error. RATE-WMT24 broadly matches the zero-shot baseline on mean F1. Only our collaboratively curated annotations consistently improve both overall span recovery and major-error detection.

In contrast, ranking metrics improve over the zero-shot baseline more uniformly across demonstration sources, indicating that span-level evaluator behavior is substantially more affected by demonstration design than pairwise ranking judgments. We therefore fix our curated annotation pool for all subsequent few-shot experiments.

4.2.2 How Many Examples?

With the demonstration source fixed, we next vary the number of examples ($k \in \{1, 2, 4, 8\}$, Figure 1).

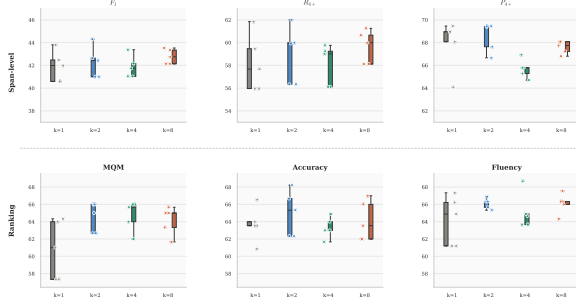


Figure 1: Effect of few-shot example count (k) across span-level and ranking metrics. Each condition uses 5 random draws. Boxes show IQR, points show individual runs.

Increasing k produces only modest changes in mean performance. Instead, the primary effect of larger few-shot sets is variance reduction: $k=1$ exhibits noticeably less stable behavior, while $k \geq 2$ yields broadly consistent results across independently sampled configurations. The most consistent mean trend is a slight improvement in R_{4+} , suggesting that additional demonstrations primarily improve major-error recovery rather than general error detection. We adopt $k=2$ for the remaining experiments, revisiting selection strategy in Section 4.2.3.

4.2.3 Coverage-Based Selection

The previous experiments showed that evaluator behavior depends strongly on demonstration design, motivating a more structured selection strategy. We construct compact few-shot sets that span the protocol’s error categories and severity range. We evaluate $k \in \{2, 3, 4\}$, corresponding to approximately 85%, 96%, and 100% category/severity coverage.

Few-shot config	Span-level					Ranking		
	R	P	F1	R_{4+}	P_{4+}	MQM	Acc	FL
RATE-AE _{full+ex} (0-shot)	30.9	53.8	39.3	52.9	66.3	56.7	60.3	65.3
Coverage-based selection (varying k), with RATE-AE _{full+ex}								
$k=2$	39.8	51.1	44.8	63.4	67.8	64.0	62.7	69.3
$k=3$	35.8	53.4	42.9	59.8	66.5	60.3	62.6	64.6
$k=4$	37.8	53.2	44.2	61.6	68.1	63.7	65.7	65.3
Coverage-based $k=2$, varying instruction								
↪ RATE-AE _{full+ex}	39.8	51.1	44.8	63.4	67.8	64.0	62.7	69.3
↪ RATE-AE _{full}	39.5	52.2	44.9	63.6	66.6	63.3	64.7	69.3
↪ RATE-AE _{brief} (ORBIT-SC)	42.4	49.6	45.7	64.8	67.5	64.0	65.3	64.3
↪ GEMBA-MQM	31.0	39.4	34.7	49.5	49.1	65.3	—	—
↪ GEMBA-ESA	31.8	37.0	34.2	51.5	50.0	63.3	—	—

Table 4: Effect of coverage-based few-shot selection. Upper block: varying k with RATE-AE_{full+ex}. Lower block: $k=2$ with varying instruction.

Table 4 shows that compact coverage-based selection already substantially improves span-level quality over the zero-shot setup. Performance is

broadly comparable across k , with a mild recall-precision trade-off: smaller k favors recall, while larger k improves severity-aware precision at the cost of recall. This is consistent with the earlier observation that demonstration quantity has limited effect on mean performance, primarily reducing run-to-run variance.

The coverage criterion itself is defined without optimizing metric scores. We select $k=2$ as the smallest configuration with comparable validation performance, balancing quality, prompt complexity, and inference cost.

Revisiting instruction choice under this setup reverses the zero-shot result from Section 4.1: RATE-AE_{brief} now slightly surpasses RATE-AE_{full+ex} on span F1, suggesting that high-quality few-shot demonstrations partially substitute for detailed taxonomy descriptions. Instruction design nevertheless remains important: replacing the RATE-style prompts with GEMBA-MQM or GEMBA-ESA substantially reduces span-level quality even under the same few-shot configuration.

We refer to this single-call configuration — RATE-AE_{brief} paired with coverage-based $k=2$ demonstrations — as ORBIT-SC. Section 4.3 evaluates this configuration across a range of frontier LLM backbones.

Model	Span-level					Ranking		
	R	P	F1	R_{4+}	P_{4+}	MQM	Acc	FL
Human _E	46.8	44.3	45.5	65.3	65.3	75.3	67.3	70.0
GPT-4.1	51.7	45.6	48.5	<u>72.7</u>	61.7	61.0	59.0	57.0
GPT-5.2	42.4	49.9	45.9	64.7	65.4	65.0	62.3	66.0
GPT-5.4	51.5	46.2	48.7	74.7	63.5	67.0	65.0	65.0
Claude Sonnet 4.6	45.8	45.6	45.7	67.1	63.5	64.0	59.5	62.9
Claude Opus 4.6	47.8	46.9	47.3	<u>72.2</u>	68.3	65.7	68.3	70.7
Gemini 3 Flash	38.9	49.1	43.4	62.2	66.8	68.0	59.3	67.0
Gemini 3.1 Pro	45.0	51.2	<u>47.9</u>	67.8	66.9	74.7	67.1	68.5
Grok-4	<u>50.4</u>	35.8	<u>41.8</u>	71.0	52.5	56.7	55.9	49.8
DeepSeek-V3.2	41.8	17.4	24.6	60.7	15.3	53.0	54.1	52.7
DeepSeek-R1	39.6	35.0	37.2	59.2	46.6	53.3	60.5	55.9
Kimi K2 Thinking	40.9	42.1	41.5	59.5	50.4	57.7	57.2	63.2
Qwen-235B	38.9	39.4	39.2	59.3	50.7	52.7	53.5	46.8

Table 5: Model comparison using ORBIT-SC across frontier models.

4.3 Model Comparison

How do frontier LLM backbones compare as MT evaluators? Using the ORBIT-SC configuration, we evaluate a range of frontier models as evaluator backbones. Table 5 reports the main results. Extended metrics are provided in Appendix F.

Different frontier models exhibit distinct evaluator profiles, with no single backbone domi-

nating all evaluation dimensions. Recent GPT models are recall-oriented: GPT-5.4 achieves the strongest span-level recovery and major-error detection, while GPT-4.1 provides similar recall behavior at substantially lower cost. Claude Opus 4.6 exhibits the strongest precision-oriented profile, leading on severity-aware precision and direct-assessment ranking while remaining competitive on span F1. Gemini 3.1 Pro combines ranking and precision strengths: it achieves the strongest MQM ranking accuracy together with the highest overall precision among span-level evaluators.

At the other end of the spectrum, Grok-4 and DeepSeek-V3.2 substantially over-predict errors at a large precision cost, with DeepSeek-V3.2 in particular flagging more than twice as many errors per segment as Human_E. The remaining open-weight models (DeepSeek-R1, Kimi K2 variants, Qwen-235B) sit consistently below the top tier across both span-level and ranking-based evaluation.

Overall, strong evaluator quality is multi-dimensional: different models favor different trade-offs between recall, precision, severity calibration, and ranking behavior. We adopt GPT-5.4 as the default backbone in the remaining experiments. Appendix G repeats the matched design comparison on Claude Opus 4.6 and Gemini 3 Flash to test whether the broader design conclusions transfer beyond GPT-5.4.

4.4 Point-Wise vs. Side-by-Side Evaluation

Human MT evaluation is commonly conducted in an implicit side-by-side (SbS) setting, where one annotator sees multiple translations of the same source (Freitag et al., 2021). Our main ORBIT-SC setup is point-wise (PW): each call evaluates one source–translation pair independently. We adapt it to SbS by placing two translations of the same source in one call and returning a separate report for each. The few-shot sources remain fixed, with the second translation added to each demonstration.

Table 6 shows a clear trade-off. SbS improves Accuracy and Fluency ranking for all three backbones and improves or matches MQM ranking, but does not improve span F_1 . Paired context shifts predictions toward higher precision and lower recall. PW is therefore convenient for evaluating an individual MT system, whereas SbS is better suited to ranking a fixed set of systems. Reversing the presentation order produced no statistically significant differences.

Backbone	Format	Span-level					Ranking		
		R	P	F1	R ₄₊	P ₄₊	MQM	Acc	FL
GPT-5.4	PW	51.5	46.2	48.7	74.7	63.5	67.0	65.0	65.0
	SbS	48.7	48.8	48.8	73.5	64.2	68.3	66.7	70.3
Claude Opus 4.6	PW	47.8	46.9	47.3	72.2	68.3	65.7	68.3	70.7
	SbS	42.1	50.5	45.9	66.8	71.9	65.7	71.0	73.3
Gemini 3 Flash	PW	38.9	49.1	43.4	62.2	66.8	68.0	59.3	67.0
	SbS	32.5	53.7	40.5	56.4	70.8	73.3	67.0	73.9

Table 6: Point-wise (PW) ORBIT-SC and its side-by-side (SbS) adaptation.

4.5 Multi-Call Workflows

We next study multi-call workflows as a design dimension: do additional LLM calls improve evaluator quality? We compare three approaches: task decomposition, iterative self-review, and fusion across multiple evaluators (Table 7).

4.5.1 Task decomposition

Inspired by M-MAD-style decomposition (Feng et al., 2025), we split annotation into semantic and non-semantic passes. Concatenating both outputs improves recall, particularly for major errors, but substantially hurts precision due to overlapping and conflicting annotations. LLM fusion recovers precision at the cost of recall. In our setting, decomposition alone provides limited benefit over the single-call baseline.

4.5.2 Self-review

We add a second-pass reviewer that re-examines the first-pass output, either only adding missed spans or also editing existing predictions. Self-review consistently improves recall, especially on major errors, but also introduces borderline predictions that reduce precision. Allowing the reviewer to edit existing spans partially recovers this precision loss and yields a small F1 gain over the single-call setup.

4.5.3 Model-diversity fusion

Fusion across diverse evaluator backbones produces the strongest overall gains. We refer to this multi-call configuration — fusing Claude Opus 4.6, GPT-5.4, and Gemini 3.1 Pro outputs via Claude Opus 4.6 as fuser — as ORBIT-MC. ORBIT-MC matches or exceeds expert-to-expert agreement on span F1 and major-error recovery, while maintaining strong precision on high-severity errors.

Unlike fixed-label ensembles, these structured reports do not provide a canonical unit for majority voting: models may identify the same error with

different span boundaries and descriptions, or assign conflicting categories, severities, and segment scores. A deterministic rule would therefore first require a separate span-matching policy and arbitrary conflict-resolution choices. The semantic fuser is used precisely to perform this reconciliation.

Fusion quality depends strongly on the fuser backbone itself. Using the same three constituent models, GPT-5.4 as fuser falls several F1 points below Claude, while Gemini 3.1 Pro as fuser produces the strongest MQM ranking accuracy. Claude substantially outperforms GPT-5.4 as a fusion model despite weaker standalone recall (Section 4.3), suggesting that strong evaluator generation and strong evaluator aggregation are distinct capabilities.

System	Span-level					Ranking		
	R	P	F1	R ₊₊	P ₊₊	MQM	Acc	FL
Human _E	46.8	44.3	45.5	65.3	65.3	75.3	67.3	70.0
Single-call								
ORBIT-SC (GPT-5.4)	51.5	46.2	48.7	74.7	63.5	67.0	65.0	65.0
Multi-call: dimension split								
Sem/non-sem concatenated	54.8	37.6	44.6	76.9	52.0	69.7	64.1	63.7
Sem/non-sem fused	41.0	51.1	45.5	64.1	64.9	63.3	62.3	64.0
Multi-call: self-review								
Add missing spans	58.4	40.5	47.8	80.3	59.4	65.7	63.0	67.0
Add + edit spans	52.2	46.3	49.1	76.5	62.1	65.0	62.3	65.0
Multi-call: model diversity fusion (C+G+Gem, varying fuser)								
↔ Claude Opus 4.6 fuser (ORBIT-MC)	60.3	45.8	52.1	82.0	70.6	69.3	67.7	73.7
↔ GPT-5.4 fuser	48.0	48.0	48.0	68.5	66.0	64.0	66.7	70.7
↔ Gemini 3.1 Pro fuser	51.7	49.7	50.7	74.1	71.6	70.0	65.9	71.1
Multi-call: varying constituent set (Claude Opus 4.6 fuser)								
2 models (C+G)	55.4	46.5	50.5	78.6	69.3	68.3	70.3	69.3
5 models (C+G+Gem+Grok+Kimi)	62.6	42.1	50.3	84.3	66.2	67.3	70.3	72.3
Multi-call: other fusion strategies (Claude Opus 4.6 fuser)								
Temperature diversity (3 runs)	57.0	45.9	50.8	78.7	66.1	68.3	65.3	70.0
Strategy diversity	56.0	45.1	50.0	78.7	67.1	68.0	65.0	69.3

Table 7: Multi-call strategies. C = Claude Opus 4.6, G = GPT-5.4, Gem = Gemini 3.1 Pro. Fused = LLM-merged report.

4.5.4 Sources of diversity

Within model-diversity fusion, expanding from two to three constituent models improves overall quality, while expanding the fusion set to five models by adding recall-oriented backbones (Grok-4, Kimi-K2) further increases recall but reduces precision and F1, consistent with the over-prediction tendencies observed in Section 4.3.

Temperature and strategy diversity also improve over the single-call baseline but produce consistently smaller gains than model diversity, indicating that complementary evaluator profiles are more valuable than repeated sampling from the same backbone.

Overall, multi-call workflows benefit most from genuine evaluator diversity rather than from iteration or decomposition over a single backbone. The multi-call configuration ORBIT-MC therefore

Evaluator setup	Format	USD/10k seg.
<i>Human, one annotator pass</i>		
Professional expert RATE	SbS	\$48,920
Assessor RATE	SbS	\$6,500
Assessor RATE	PW	\$6,865
ORBIT-SC, varying backbone and format		
Gemini 3 Flash	PW	\$20
GPT-4.1	PW	\$60
GPT-5.4	PW	\$85
GPT-5.4	SbS	\$75
Claude Opus 4.6	PW	\$345
Gemini 3.1 Pro	PW	\$385
ORBIT-MC (3 models + fuser)	PW	\$1,720

Table 8: API cost and estimated cost of one human annotator pass per 10,000 evaluated segments.

adopts three-model fusion (Claude Opus 4.6, GPT-5.4, Gemini 3.1 Pro) via Claude as fuser, achieving the strongest span-detection profile across all tested configurations.

4.6 Cost and Deployment Trade-offs

Table 8 compares human and API costs at a common scale of 10,000 evaluated segments. We use this volume as a representative data-collection campaign for training or adapting an automatic evaluator rather than a one-off evaluation set. A segment is one source–translation pair. PW is the standard ORBIT-SC setup, with one pair per call. Its SbS adaptation places two translations of the same source in one call and returns a separate report for each. The human PW and SbS conditions use the same visibility distinction. Thus, 300 SbS tasks and 600 PW tasks both yield 600 segment-level reports.

Human estimates use observed aggregate labor for one annotator pass rather than extrapolating median completion times. Scaling these totals to 10,000 reports gives roughly 3,542 expert-hours in SbS, 1,625 assessor-hours in SbS, and 1,717 assessor-hours in PW. Monetary estimates are rounded to the nearest \$. The ORBIT-MC total includes three evaluators and the fuser.

At this scale, one pass by a professional expert costs about \$48,920. An assessor pass costs \$6,500 in SbS and \$6,865 in PW. ORBIT-MC costs roughly \$1,720, corresponding to a 4× cost reduction from the least expensive assessor setup and a 28× reduction from professional expert RATE. Point-wise ORBIT-SC costs \$20–\$385 depending on the backbone, a further 4–86× reduction relative to ORBIT-MC while retaining the strong results in Table 5.

System	en-de ^{MQM}					zh-en ^{MQM}					cs-de ^{ESA}					en-it ^{ESA}					en-zh ^{ESA}				
	R	P	F1	R ₄₊	MQM	R	P	F1	R ₄₊	MQM	R	P	F1	R ₄₊	MQM	R	P	F1	R ₄₊	MQM	R	P	F1	R ₄₊	MQM
Human ₂	31.8	40.7	35.7	48.7	60.7	44.5	41.5	42.9	63.2	68.7	21.8	36.0	27.2	29.3	45.3	25.5	20.3	22.6	37.2	48.0	15.5	18.1	16.7	17.8	41.3
GEMBA-MQM	29.9	35.5	32.5	43.2	60.0	29.3	44.3	35.2	46.6	63.8	17.0	22.4	19.3	23.8	46.7	17.8	10.5	13.2	29.8	43.3	33.2	12.8	18.5	43.9	50.7
MQM-AE	30.1	45.5	36.2	46.2	56.9	31.6	51.7	39.3	52.0	58.0	29.1	29.5	29.3	38.4	32.0	43.8	14.7	22.0	60.3	50.0	38.8	14.4	21.0	46.7	42.0
ESA-AE	29.8	46.8	36.4	45.2	59.8	33.2	60.6	42.9	55.7	62.2	37.7	27.9	32.1	45.4	52.0	45.1	15.1	22.7	63.6	46.0	42.1	14.6	21.7	46.7	36.0
UMB-v5	32.5	42.6	36.9	48.0	61.3	37.7	49.1	42.6	58.3	63.1	23.5	28.9	25.9	32.7	50.0	39.9	13.6	20.3	60.3	50.0	35.5	13.1	19.1	44.9	43.3
RATE-AE _{brief}	35.4	43.2	38.9	51.3	61.8	41.8	55.7	47.7	62.6	64.9	32.1	27.6	29.7	41.4	52.0	51.0	14.2	22.2	66.9	52.7	49.9	12.9	20.4	54.2	44.0
RATE-AE _{brief} + DS	38.0	40.3	39.1	53.8	57.1	47.8	49.0	48.4	67.1	65.1	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
ORBIT-SC	41.5	37.1	39.2	58.3	60.7	50.6	49.5	50.0	69.0	66.9	39.6	22.1	28.3	49.0	49.3	60.1	11.6	19.4	74.4	53.3	61.5	10.7	18.3	58.9	40.0
ORBIT-MC	45.2	35.5	39.7	62.1	63.8	51.7	48.4	49.9	68.0	67.3	45.3	22.9	30.4	54.1	48.0	70.1	10.5	18.2	78.5	49.3	64.5	9.2	16.1	59.8	45.3

Table 9: Multilingual transfer on WMT23 MQM (en-de, zh-en) and WMT25 ESA (cs-de, en-it, en-zh) data. DS = direction-specific demonstrations: two few-shot examples in the target translation direction.

For GPT-5.4, SbS costs roughly \$75 compared with \$85 for separate PW calls. The corresponding quality comparison is reported in Section 4.4.

These dollar totals still understate the practical gap. Human evaluation requires recruiting and screening annotators who are native speakers of the target language and have C1/C2 proficiency in the source language. For expert pools, formal training or a degree in linguistics or translation is also preferred. Protocol training, calibration, scheduling, and ongoing quality control add further overhead (Freitag et al., 2021; Kocmi et al., 2024b; Popov et al., 2025). Much of this infrastructure must be established again for each language direction, which delays evaluation and makes procedures harder to standardize across pools. API-based evaluation can reuse a common workflow across directions, although model quality may still vary across languages and domains.

5 Multilingual Transfer

We test whether the evaluator design choices identified on English–Russian transfer across annotation protocols, annotation densities, and language pairs. We evaluate ORBIT-SC and ORBIT-MC on WMT23 MQM (en-de, zh-en) and WMT25 ESA (cs-de, en-it, en-zh) benchmarks using the same inference-time configurations developed on RATE English–Russian data. Human₂ denotes a second annotator scored against the gold annotation.

Across both MQM and ESA benchmarks (Table 9), the main findings of our design study transfer consistently beyond the English–Russian setting. ORBIT-SC remains competitive across all evaluated language pairs, while ORBIT-MC preserves its strong recall-oriented profile, improving major-error recovery over ORBIT-SC on en-de and remaining comparable on zh-en.

Compact RATE-style prompting remains competitive on MQM. Replacing the en-ru few-shot

pool with direction-specific demonstrations (DS) provides no substantial gain over ORBIT-SC on either MQM direction.

Notably, the en-ru pool matches direction-specific demonstrations on en-de span F_1 while improving R_{4+} and MQM, and improves all three metrics on zh-en. These results suggest that the advantage comes not from Russian itself but from how the pool was constructed. The pool covers diverse error categories and severities, with annotations collaboratively reconciled by multiple experts.

Transfer to ESA benchmarks exhibits substantially lower precision across all evaluators, likely reflecting the much sparser annotation density of ESA compared to MQM and RATE. On span F_1 , ORBIT-MC provides smaller gains than on MQM and is slightly below ORBIT-SC on en-it and en-zh, reflecting that multi-model fusion expands prediction sets in a regime where additional spans are heavily penalized. Nevertheless, transfer gains remain clearest on major-error recovery, where ORBIT-MC consistently exceeds Human₂ across all ESA directions.

6 Conclusion

We systematically study how instruction design, demonstrations, backbone choice, and multi-call workflows shape LLM-based MT evaluation across RATE, MQM, ESA, and multiple language pairs. Curated, collaboratively reconciled demonstrations can match or outperform larger pools, model families show distinct evaluator profiles, and fusion yields the strongest overall results. ORBIT instantiates these findings as single- and multi-call evaluators. Its strongest configurations achieve expert-level error span detection and Accuracy/Fluency ranking, while ORBIT-MC exceeds assessor agreement on MQM ranking.

Limitations

Language coverage. Our main design study is conducted on a single high-resource language pair (English–Russian). We validate the resulting ORBIT-SC and ORBIT-MC configurations on five additional WMT directions (en-de and zh-en for MQM, and cs-de, en-it, and en-zh for ESA), but coverage of low-resource pairs, non-Latin scripts beyond Russian and Chinese, and language families outside Indo-European and Sino-Tibetan remains limited. Findings on instruction language and few-shot example selection in particular may require adaptation for typologically distant target languages.

Closed-source dependencies. Our strongest configurations (ORBIT-SC with GPT-5.4, ORBIT-MC fusing Claude Opus 4.6, GPT-5.4, and Gemini 3.1 Pro) depend on closed-source, API-only models whose weights and training procedures are not public. This affects reproducibility (exact behaviour may shift across silent model updates) and longevity (specific model snapshots may be deprecated). We mitigate the reproducibility concern through the broad cross-model comparison in Section 4.3, showing that qualitative conclusions hold across model families.

Span-matching methodology. Span-level F_1 depends on the choice of span-matching formulation. We use MP with $\tau=1$ as the primary formulation because our main question is whether the evaluator identifies the annotated error location. MPP assigns partial credit in proportion to character overlap and is therefore more sensitive to span-boundary agreement. Appendix E reports MPP and LLM-based reconciliation as complementary sensitivity analyses.

Severity calibration. Appendix F shows that severity assignment is a distinct capability from error detection, with both human experts and LLM evaluators substantially under-rating major errors when stricter severity thresholds are required. We report detection-oriented R_{4+}/P_{4+} as primary metrics in the main paper because they isolate the more fundamental question of error localization, but downstream use cases that depend on accurate severity assignment (e.g., quality estimation thresholds, automatic gating) require additional severity calibration that is not the focus of this work.

Human–LLM collaboration study. Our human–LLM collaboration experiments (Appendix H) are exploratory in scope: they measure annotation revision rather than ground-truth correctness, use only ORBIT-MC as the reference configuration, and were conducted with a limited annotator group on a single language pair. Whether the observed shifts in annotation behavior correspond to improved annotation quality, and how the effects generalize across configurations and annotator pools, remains to be studied.

Subjectivity of MT evaluation. Translation quality assessment remains partly subjective even among trained experts. The agreement between our expert annotators (Table 1) provides a human-agreement reference point for how closely any evaluator — human or LLM-based — can be expected to match a single “ground truth”. We treat expert agreement as the headline target rather than a hard ceiling, and recognize that ranking differences within this band may not reflect meaningful quality differences.

Ethics Statement

Annotator compensation and consent. RATE annotators received compensation at standard commercial translation rates for their region, with expert annotators earning approximately double the hourly wage of assessors, reflecting the higher qualifications required. No personal data was collected during the annotation process, and all source segments were screened for potentially disturbing material prior to annotation. All annotators were informed about how their annotations would be used — including for the development and evaluation of LLM-based MT evaluators — prior to beginning the tasks.

Human–LLM collaboration study. The collaborative annotation study in Appendix H followed the same compensation and consent procedures. Annotators were informed that they would be reviewing LLM-generated reports as part of their workflow, and that their pre- and post-hint annotations would be compared. No identifying information about annotators is reported in our analyses.

Annotator feedback. Following the annotation, we administered a feedback questionnaire. The vast majority of annotators reported a positive experience and found the instructions clear.

Use of closed-source models. Our strongest configurations rely on closed-source LLMs accessed via commercial APIs. We discuss reproducibility and longevity implications of this choice in the Limitations section.

References

- Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*.
- 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 Suggia, 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 (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.
- Zhaopeng Feng, Jiayuan Su, Jiamei Zheng, Jiahao Ren, Yan Zhang, Jian Wu, Hongwei Wang, and Zuozhu Liu. 2025. M-MAD: Multidimensional multi-agent debate for advanced machine translation evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.
- Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André F. T. Martins, Graham Neubig, Ankush Garg, Jonathan H. Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, Singapore. Association for Computational Linguistics.
- Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.
- Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frédéric 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, Singapore. Association for Computational Linguistics.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Zhouchi Lin, Bowen Zhang, Lionel Ni, Wen Gao, Yuanzhuo Wang, and Jian Guo. 2026. [A survey on LLM-as-a-judge](#). *The Innovation*, 7(6):101253.
- 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)*.
- Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation (WMT)*.
- Juraj Juraska, Tobias Domhan, Mara Finkelstein, Tetsumi Nakagawa, Geza Kovacs, Daniel Deutsch, Piding Wang, and Markus Freitag. 2025. MetricX-25 and GemSpanEval: Google Translate submissions to the WMT25 evaluation shared task. In *Proceedings of the Tenth Conference on Machine Translation (WMT)*, Suzhou, China. Association for Computational Linguistics.
- Ahrii Kim. 2025. RUBRIC-MQM: Span-level LLM-as-judge in machine translation for high-end models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 147–165, Vienna, Austria. Association for Computational Linguistics.
- Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakounga, Jessica M. Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. [Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets](#). In *Proceedings of the Tenth Conference on Machine Translation (WMT)*, Suzhou, China. Association for Computational Linguistics.
- 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. 2024a. 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 (WMT)*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

- 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, Singapore. Association for Computational Linguistics.
- 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, Tampere, Finland.
- Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024b. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation (WMT)*, Miami, Florida, USA. Association for Computational Linguistics.
- Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deghare, Archchana Sindhuja, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. 2025. [Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 436–483, Suzhou, China. Association for Computational Linguistics.
- Christoph Leiter and Steffen Eger. 2024. [PrExMe! large scale prompt exploration of open source LLMs for machine translation and summarization evaluation](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11481–11506, Miami, Florida, USA. Association for Computational Linguistics.
- Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. [Multidimensional quality metrics \(MQM\): A framework for declaring and describing translation quality metrics](#). *Tradumàtica*, pages 455–463. No. 12.
- Qingyu Lu, Liang Ding, Kanjian Zhang, Jinxia Zhang, and Dacheng Tao. 2025. MQM-APE: Toward high-quality error annotation predictors with automatic post-editing in LLM translation evaluators. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*. Association for Computational Linguistics.
- Qingyu Lu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2024. Error analysis prompting enables human-like translation evaluation in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8801–8816. Association for Computational Linguistics.
- Stefano Perrella, Eric Morales Agostinho, and Hugo Zaragoza. 2026. Span-level machine translation meta-evaluation. *arXiv preprint arXiv:2603.19921*.
- Dmitry Popov, Vladislav Negodin, Ekaterina Enikeeva, Iana Matrosova, Nikolay Karpachev, and Max Ryabinin. 2025. Refined assessment for translation evaluation: Rethinking machine translation evaluation in the era of human-level systems. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22079–22095, Suzhou, China. Association for Computational Linguistics.
- 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 (NeurIPS)*.

A Dataset Statistics

The design study dataset comprises 600 source-translation pairs constructed from 300 English source segments sampled from WMT24–WMT25 English–Russian shared task data, each paired with two system translations. Source selection prioritizes challenging segments where candidate translations disagree, ensuring evaluators must discriminate between subtle quality differences rather than identify obvious failures.

The few-shot pool uses separately selected sources and does not overlap with this 600-pair evaluation set.

Domain coverage. The dataset covers four domains (Table 10), with literary, news, and speech segments slightly over-represented relative to social media text.

Translation systems. The 600 segments span 14 translation systems (Table 11) including frontier closed-source LLMs (Claude-3.5/4, GPT-4/4.1, Gemini 2.5), production MT systems (Dub-

Domain	Segments
Literary	166
News	146
Speech	146
Social	142
Total	600

Table 10: Domain distribution of source–translation pairs.

former, ONLINE-G, Yandex Translate), open-weight LLMs (Llama3-70B, DeepSeek-V3), dedicated MT systems (Unbabel-Tower70B), and human reference translations from the WMT campaign.

Translation system	Segments
Gemini 2.5	69
Claude-3.5	55
Dubformer	55
GPT-4	53
ONLINE-G	53
Unbabel-Tower70B	53
Yandex Translate	52
Reference (human)	52
Hunyuan	33
DeepSeek-V3	29
Llama3-70B	27
Claude 4	25
Internal experimental	25
GPT-4.1	19
Total	600

Table 11: Translation system distribution.

Annotation density. Each translation is annotated independently by three expert linguists under the RATE protocol (Expert $\times$ 3). Combining two annotations, translations contain on average 10.64 error spans per segment (Table 12), with severity 2 (minor errors) most frequent followed by severity 3 and severity 4 (major). Critical errors (severity 5) and trivial errors (severity 1) together account for roughly 11% of annotated spans. The per-individual annotator rate is approximately halved (5.48 spans per segment), consistent with partial overlap across independent expert annotations.

Category distribution. Table 13 reports the distribution across RATE error categories. Fluency errors dominate (6.06 spans per segment, combined), followed by Mistranslation (2.34). All other categories average below 1 span per segment. This long-tailed distribution motivates coverage-based few-shot selection (Section 4.2): uniform random

Severity	Combined	Annotator A	Annotator B
1 (trivial)	0.65	0.37	0.29
2 (minor)	4.16	2.16	2.00
3 (minor)	3.60	1.83	1.77
4 (major)	1.71	0.85	0.86
5 (critical)	0.53	0.28	0.25
Total	10.64	5.48	5.22

Table 12: Mean error spans per segment by severity. “Combined” aggregates two independent expert annotations. Per-annotator columns show individual outputs.

sampling would oversample Fluency and Mistranslation while underrepresenting rarer categories.

Category	Spans / segment
Fluency	6.06
Mistranslation	2.34
Undertranslation	0.61
Overtranslation	0.44
Grammar	0.32
Other	0.31
Named Entity	0.30
Style	0.29
Omission	0.01

Table 13: Mean error spans per segment by RATE category, aggregated across the two expert gold annotations.

Source length and overall quality. The average source segment length is 377 characters. The mean RATE overall score (a derived 0–10 quality score) is 9.30, indicating that the dataset contains publication-grade translations rather than failure-mode outputs. Mean annotated span coverage per translation is 62 characters, roughly 16% of source length, with errors concentrated in localized regions rather than spread diffusely.

B Artifact Use and Licenses

WMT24 General MT data is released by the WMT organizers under terms permitting research use. WMT23 MQM annotations (Freitag et al., 2023) are released under the Apache 2.0 license via the google/wmt-mqm-human-evaluation repository. WMT25 General MT data, including ESA annotations (Lavie et al., 2025), is publicly released by the WMT organizers via the wmt-conference/wmt25-general-mt repository for research purposes. xCOMET-XXL (Guerreiro et al., 2024) is released under CC-BY-NC-SA-4.0. We use it for non-commercial research evaluation, consistent with its license terms. Closed-source LLM APIs (GPT, Claude, Gemini) were accessed

System	R			P			F_1		
	MP	MPP	LLM	MP	MPP	LLM	MP	MPP	LLM
Human _E	46.8	41.9	54.1	44.3	38.6	52.1	45.5	40.2	53.1
Human _{A-Sbs}	39.3	34.8	47.6	47.1	40.8	55.2	42.8	37.5	51.1
ORBIT-SC GPT-5.4	51.5	49.9	55.7	46.2	26.7	51.2	48.7	34.8	53.4
ORBIT-SC Claude	47.8	45.9	57.1	46.9	30.6	51.4	47.3	36.8	54.1
ORBIT-SC Gemini	45.0	42.8	47.6	51.2	36.9	56.7	47.9	39.6	51.7
ORBIT-MC	60.3	58.2	66.2	45.8	28.6	49.1	52.1	38.4	56.4

Table 14: Span-level metrics under three span-matching formulations: MP with $\tau=1$ (primary), MPP (Perrella et al., 2026) (partial overlap with partial credit), and an LLM reconciliation proxy using GPT-5.4. All metrics are micro-averaged.

in accordance with their respective terms of service.

Our use of all existing artifacts is consistent with their intended research and non-commercial evaluation purposes. The expert-annotated dataset, evaluation prompts, and accompanying paper artifacts are available at <https://huggingface.co/datasets/yandex/orbit-mt-eval>.

C Annotator Pool

Expert annotators. Our Expert tier consists of professional translators who are native Russian speakers with verified C1/C2 English proficiency and substantial industry experience. A university degree in linguistics or translation studies was preferred. Candidates passed a multi-stage selection process: a post-editing test that required context analysis and source fact-checking, followed by an interview. After selection, experts received structured training on the RATE protocol — error categorization, severity calibration, and the Accuracy/Fluency scoring scales — with concrete examples for each category.

Assessor. Our Assessor tier consists of in-house crowd-level evaluators: native Russian speakers with verified C1 English proficiency (screened via a standardized proficiency test). We do *not* require linguistics or translation degrees or professional translation experience. Assessors received the same RATE-protocol training materials as experts but were not subject to the post-editing selection test.

Pool size and workload. The pool comprises **35** expert translators and **82** assessors. Each segment is annotated independently by three experts from the pool in implicit side-by-side mode (Expert $\times 3$), and by assessors in two setups: im-

plicit side-by-side (Human_{A-Sbs}) and point-wise (Human_{A-PW}). For evaluation we hold two expert annotations as gold and score the third against each of them and average, giving the Human_E agreement reported in Table 1 (Section 3.3). Average workload per annotation pass was **17** segments per expert and **22** segments per assessor. Median annotation times were **14** minutes per segment for experts and **5** minutes for assessors, compared to the ~ 68 seconds per segment that Kocmi et al. (2025) report for English–Russian WMT25 ESA annotation.

Training and dispute resolution. Before annotation began, all annotators participated in a calibration session in which a small set of pre-annotated examples was reviewed and edge cases were discussed collaboratively. For the few-shot pool used in Section 4.2, annotators worked in overlapping groups on subsets of segments and reconciled disagreements through discussion to produce consensus annotations more reliable than any single-annotator output.

D Methods Not Included in Main Comparison

We do not include several recent MT evaluation approaches in our primary comparison due to differences in evaluation setup and workflow design. M-MAD (Feng et al., 2025) explores debate-style multi-agent evaluation with substantially different orchestration strategies and computational budgets. We draw on its decomposition idea in our multi-call ablations (Section 4.5) but do not reproduce the full multi-agent system. EAPrompt (Lu et al., 2024) produces error analyses through iterative prompting. A controlled re-evaluation by Feng et al. (2025) reports lower span-level F_1 than GEMBA-MQM, which is already included in our comparison. MQM-APE (Lu et al., 2025) relies on automatic post-editing as a validation step, requiring multiple sequential LLM passes per segment and following a substantially different multi-stage workflow. More broadly, our goal is not to exhaustively benchmark every existing MT evaluation system, but to study evaluator design principles under a controlled experimental setup.

E Sensitivity to Span Matching

Perrella et al. (2026) recommend MPP with micro-averaging to penalize imprecise span boundaries and avoid the preference of MP for longer spans.

Model	Spans / segment			Severity-restricted metrics			
	Non-trivial	Major	Critical	R_{4+}	P_{4+}	$R_{4+}^\bullet$	$P_{4+}^\bullet$
Human _E	5.28	1.01	0.28	65.3	65.3	54.0	52.5
GPT-4.1	5.67	0.93	0.28	72.7	61.7	53.1	56.1
GPT-5.2	4.26	0.90	0.27	64.8	67.5	45.4	59.1
GPT-5.4	5.57	1.55	0.49	74.7	63.5	61.3	55.7
Claude Sonnet 4.6	5.02	1.10	0.33	67.1	63.5	56.8	53.7
Claude Opus 4.6	5.09	0.77	0.34	72.2	68.3	56.8	57.4
Gemini 3 Flash	3.96	0.66	0.23	62.2	66.8	47.6	60.7
Gemini 3.1 Pro	4.39	1.09	0.24	67.8	66.9	56.8	61.0
Grok-4	7.03	1.36	0.29	71.0	52.5	59.0	42.7
DeepSeek-V3.2	11.96	6.12	1.59	60.7	15.3	53.4	17.4
DeepSeek-R1	5.65	1.68	0.40	59.2	46.6	52.1	39.3
Kimi K2	3.49	0.70	0.18	39.4	50.5	32.6	39.2
Kimi K2 Thinking	4.84	1.09	0.31	59.5	50.4	43.7	46.8
Qwen-235B	4.94	1.66	0.58	59.3	50.7	53.0	44.4

Table 15: Extended model comparison using ORBIT-SC (RATE-AE_{brief} with coverage-based $k=2$ demonstrations). Non-trivial, Major, and Critical report mean predicted spans per segment at severity ≥ 2 , ≥ 4 , and $= 5$ respectively. R_{4+} measures recovery of gold major spans regardless of predicted severity, while P_{4+} uses predictions assigned severity ≥ 4 . The bullet variants use predicted severity ≥ 3 , providing one-level tolerance around the major threshold.

Our target places more emphasis on error-identity agreement, so we retain MP with $\tau=1$ as the primary formulation and compare it with MPP and an LLM-based reconciliation proxy. All metrics in this section are micro-averaged.

E.1 Formal Definitions

We compare three matching formulations.

MP with $\tau=1$ (primary). Predicted error spans are matched to gold annotations via bipartite matching that maximizes the number of matched pairs. A predicted–gold pair is eligible for matching iff their character overlap is ≥ 1 . Each matched pair contributes a credit of 1.0 to both precision (over predicted spans) and recall (over gold spans).

MPP (Perrella et al., 2026). Eligible predicted–gold pairs must overlap by at least one character. Among eligible one-to-one matchings, MPP selects the matching that maximizes the resulting MPP F_1 . Credit is then proportional to overlap. For a matched pair with k overlapping characters, precision credit (over the predicted span of length ℓ_p) is k/ℓ_p , and recall credit (over the gold span of length ℓ_g) is k/ℓ_g .

LLM-based reconciliation proxy. For each translation, we present the predicted spans and gold annotations to GPT-5.4 and instruct it to match annotations that describe the same error, allowing character boundaries to differ. Each matched pair

contributes a credit of 1.0. We treat this as an exploratory reconciliation proxy rather than human adjudication.

E.2 Cross-Formulation Comparison

Table 14 reports all three formulations across our main systems.

E.3 Boundary Sensitivity under MPP

The matching formulation affects exact system rankings and the comparison between human and model agreement. Under MPP, Human_E drops by 4.9 recall points and 5.7 precision points. Across the three ORBIT-SC backbones, recall changes by only 1.6–2.2 points, while precision falls by 14.3–19.5 points. ORBIT-MC shows the same asymmetry. This follows directly from MPP’s partial-credit rule when a predicted span extends beyond the gold boundaries.

Boundary fidelity is relevant for some downstream uses. However, MPP can under-credit error-identity agreement when systems and experts use different span granularities.

E.4 LLM Reconciliation Proxy

The LLM reconciliation proxy provides a complementary view. It places Human_E at $F_1=53.1$, the three ORBIT-SC configurations at 51.7–54.1, and ORBIT-MC at 56.4. These results support the main error-identity comparison, while MPP shows that

Configuration	Span-level					Ranking		
	R	P	F1	R_{4+}	P_{4+}	MQM	Acc	FL
GEMBA-MQM	21.9	43.8	29.2	42.6	55.0	53.0	—	—
GEMBA-ESA	24.7	43.0	31.4	46.7	58.3	53.7	—	—
MQM-AE	24.5	48.0	32.4	46.7	63.7	54.0	—	—
ESA-AE	26.7	51.3	35.1	50.4	62.7	57.3	—	—
UMB-v5	23.5	42.5	30.3	46.1	61.8	54.7	—	—
RATE-AE _{brief}	29.5	54.4	38.2	54.4	72.4	59.3	67.7	65.3
RATE-AE _{full}	26.6	55.8	36.0	52.2	78.3	61.0	68.7	64.3
RATE-AE _{full+ex}	29.2	52.4	37.5	54.4	74.2	56.3	69.0	62.7
RATE-AE _{full+ex} + coverage, $k=2$	45.1	49.4	47.2	70.3	68.5	65.7	71.0	70.0
RATE-AE _{full+ex} + coverage, $k=4$	48.4	46.4	47.4	72.1	65.5	63.7	65.0	68.0
ORBIT-SC	47.8	46.9	47.3	72.2	68.3	65.7	68.3	70.7

(a) Claude Opus 4.6.

Configuration	Span-level					Ranking		
	R	P	F1	R_{4+}	P_{4+}	MQM	Acc	FL
GEMBA-MQM	20.3	45.4	28.1	40.2	60.5	59.3	—	—
GEMBA-ESA	16.3	47.2	24.2	36.1	57.9	52.3	—	—
MQM-AE	15.3	56.2	24.0	36.5	63.1	56.7	—	—
ESA-AE	19.5	60.1	29.4	42.0	69.4	60.0	57.7	—
UMB-v5	20.6	47.9	28.8	43.7	68.6	56.3	—	—
RATE-AE _{brief}	21.7	54.9	31.1	45.1	73.5	58.0	53.7	61.7
RATE-AE _{full}	20.8	59.7	30.9	42.6	76.6	59.0	55.7	60.3
RATE-AE _{full+ex}	20.7	59.6	30.7	43.4	75.8	58.7	58.7	61.3
RATE-AE _{full+ex} + coverage, $k=2$	33.9	52.3	41.1	58.2	69.7	65.3	59.7	62.3
RATE-AE _{full+ex} + coverage, $k=4$	31.0	54.4	39.5	54.8	70.0	67.3	59.3	65.0
ORBIT-SC	38.9	49.1	43.4	62.2	66.8	68.0	59.3	67.0

(b) Gemini 3 Flash.

Table 16: Matched design comparison across two backbones. ORBIT-SC uses RATE-AE_{brief} with curated coverage ($k=2$).

the exact human–model ordering is sensitive to span boundaries.

GPT-5.4 is both the reconciler and one of the evaluated backbones, so the proxy may share biases with the systems being judged. We therefore use it as complementary evidence rather than as the primary meta-evaluation method.

E.5 Methodological Takeaway

We retain MP with $\tau=1$ as primary because our main question is whether an evaluator identifies the annotated error location. MPP provides a boundary-sensitive view, while the LLM reconciliation proxy emphasizes error identity. The three formulations support the same broad conclusion that ORBIT is a strong evaluator, but exact system ordering and the human–model comparison depend on the operational definition of a span match.

F Model Comparison: Extended Results

Table 15 extends the model comparison from Section 4.3 with per-model span density (mean predicted spans per segment, broken down by severity) and severity-calibrated variants of the major-error metrics.

Severity-aware metrics. The main R_{4+} metric measures recovery of gold spans with severity ≥ 4 regardless of predicted severity. The main P_{4+} metric measures precision for predictions assigned severity ≥ 4 against the same gold major-error spans. For $R_{4+}^\bullet$ and $P_{4+}^\bullet$, only predictions assigned severity ≥ 3 are eligible for matching. These are severity-aware variants with a one-level tolerance around the major threshold. The gap between R_{4+} and $R_{4+}^\bullet$ therefore measures how often real major errors are detected but assigned severity ≤ 2 .

Severity calibration is a separate axis of evaluator behavior. As shown in Table 15, the gap between R_{4+} and $R_{4+}^\bullet$ reveals substantial under-rating of major errors across both human and LLM evaluators. Human_E itself drops 11 points ($65.3 \rightarrow 54.0$), confirming that severity assignment is genuinely hard even for trained experts. LLM evaluators show gaps of comparable magnitude: GPT-5.4 drops from 74.7 to 61.3 (-13.4), Claude Opus 4.6 from 72.2 to 56.8 (-15.4), and Gemini 3.1 Pro from 67.8 to 56.8 (-11.0). This indicates that detection and severity assignment are partially independent capabilities, and that the difficulty of severity calibration is not specific to LLM evaluators — trained experts exhibit the same pattern. We adopt the detection-oriented R_{4+}/P_{4+} as primary metrics in the main paper because they isolate the comparatively more fundamental question of whether the evaluator finds the right spans, but report severity-calibrated variants here as a complementary diagnostic.

Span density patterns. The mean spans-per-segment columns clarify the recall–precision trade-offs visible in the span-level metrics (Table 5). Backbones with strong precision (Gemini 3.1 Pro, Claude Opus 4.6) predict at or below the expert non-trivial rate (4.39 and 5.09 vs. 5.28), while recall-oriented backbones (GPT-5.4, Grok-4) predict above it but with major-severity counts that remain plausible (1.55 and 1.36 vs. 1.01). DeepSeek-V3.2 is the clear outlier: 11.96 non-trivial spans and 6.12 major errors per segment — roughly $6\times$ the expert major-error rate — explaining its severe precision collapse ($P_{4+} = 15.3$). Kimi-K2 sits at the opposite extreme, under-predicting both non-trivial (3.49) and major (0.70) spans, consistent with its low overall recall ($R = 22.0$).

G Cross-Backbone Design Replication

We evaluate the same 11 configurations with Claude Opus 4.6 and Gemini 3 Flash using the final metric pipeline. This matched set covers published baselines, RATE instruction variants, coverage-based demonstrations, and the final ORBIT-SC configuration. The full comparison is reported in Table 16.

On Gemini, ORBIT-SC is strongest on span F_1 and R_{4+} . On Claude, it has the strongest R_{4+} and is within 0.1 F_1 of the strongest configuration. RATE-based variants also outperform the protocol baselines on F_1 for both backbones, supporting transfer of the overall design while leaving individual instruction effects backbone-dependent.

H Human-LLM Collaborative Annotation

We conduct a preliminary study of how expert annotators interact with reports generated by the ORBIT-MC multi-call configuration. On a held-out set of 600 English–Russian segments (disjoint from the design study data), we compare two expert groups. Human_{E1} first produced independent annotations and then re-annotated the same segments after reviewing ORBIT-MC reports (post-hint, within-subject design). Human_{E2} received the reports before its initial annotation (pre-hint).

In aggregate, access to ORBIT-MC reports increases annotation density and shifts annotators toward lower overall quality scores (Table 17). Annotators do not adopt reports uncritically: post-hint annotations retain fewer spans than the raw report contains, and slider scores shift toward but do not match ORBIT-MC’s more severe assessments, indicating partial rather than wholesale integration of report signals. These results are exploratory — the study measures annotation revision rather than ground-truth correctness.

Annotator	Spans / segment			Mean scores	
	All	Major	Crit.	Acc	FL
Human _{E1} (baseline)	5.60	1.00	0.28	76.6	74.8
Human _{E1} + post-hint	7.04	1.11	0.30	72.3	72.1
Δ within-subject	+1.44	+0.11	+0.02	−4.3	−2.7
Human _{E2} + pre-hint	6.00	1.03	0.26	74.1	73.0
ORBIT-MC (raw hint)	8.09	0.78	0.24	60.2	59.7

Table 17: Effect of ORBIT-MC hints on expert annotation (held-out 600-segment set).