MetaDocEval: A Contrastive Framework for Evaluating Machine Translation Metrics at the Document-Level

Nicolas Dahan[♢♠] **Rachel Bawden**[♢] **François Yvon**[♠]
[♢]Inria Paris, France
[♠]Sorbonne Université, CNRS, ISIR, Paris, France
nicolas.dahan@inria.fr rachel.bawden@inria.fr yvon@isir.upmc.fr

Abstract

Recent advances in neural machine translation (MT) have spurred increased interest in evaluating translations beyond the sentence level, making it possible to assess discourse-level phenomena related to coherence and consistency. While existing sentence-level metrics can be applied to multi-sentence spans, it remains unclear whether their scores truly capture document-level quality. We introduce MetaDocEval, a reusable framework for generating contrastive document-level test sets, together with an instantiated test set covering three language pairs (en–fr, en–es, en–de). The framework includes automatic perturbation generation, quality-control filtering and sliding-window scoring, so that new corpora, language pairs, or perturbation types can be added with minimal manual intervention. The released test set targets a range of discourse-level phenomena and potential problems linked to translation at the document level. To evaluate how metrics behave as a function of context size, we apply them under a sliding-window protocol, varying the input from single sentences up to full documents. Our experiments show that none of the metrics tested genuinely capture document-level coherence: reference-based metrics overfit lexical overlap, reference+source metrics gain little from added context, reference-free encoders show brief context sensitivity before degrading on longer spans, and LLM-based scorers collapse beyond short inputs. A key finding is that reference access can be actively harmful for detecting discourse-level errors. Using short windows ($\approx$ 3 sentences) offers the best trade-off between discourse error detection and score dilution.

1 Introduction

The machine translation (MT) community is increasingly shifting its focus beyond sentence-level translation towards the document level (Maruf et al., 2021; Post and Junczys-Dowmunt, 2024). Neural MT systems achieve remarkable fluency and adequacy at the sentence level, but their limitations become apparent when translations are looked at in their broader context (Läubli et al., 2018; Toral et al., 2018). This is because many critical linguistic phenomena, related to document coherence and cohesion, can span across sentences. As attention turns to these document-level challenges, it is becoming increasingly pressing to develop automatic evaluation tools and metrics to assess translation quality beyond the sentence level. However, most metrics are still developed for the sentence level.

A straightforward response is to apply existing metrics to longer spans. Surface-based scores such as BLEU (Papineni et al., 2002) and chrF (Popović, 2015), embedding-based metrics such as BERTScore (Zhang et al., 2020) and COMET (Rei et al., 2020), and LLM-based metrics such as GEMBA-MQM (Kocmi and Federmann, 2023) can all process text inputs spanning multiple sentences. However, it remains unclear whether their scores truly capture long-distance coherence. Our main question is the following: can existing metrics be reliably used to evaluate the translation quality of segments spanning multiple sentences?

Metrics are typically judged based on their corre-

lation with human scores (Kocmi et al., 2021; Freitag et al., 2021a; Thompson et al., 2024; Freitag et al., 2024; Lavie et al., 2025). In this perspective, answering our main question would require to collect document-level judgments, an expensive and cognitively demanding task (Castilho, 2021). Contrastive challenge sets sidestep this by injecting controlled errors and checking whether metrics rank the original segments higher (Karpinska et al., 2022; Amrhein et al., 2022). Existing challenge sets, however, mostly target short-range context, typically one or two neighbouring sentences (Karpinska et al., 2022) and therefore are not adapted to evaluating metrics on long-distance discourse phenomena.

To fill this gap, we introduce `MetaDocEval`, a contrastive framework targeting the evaluation of translation errors affecting coherence and cohesiveness beyond the sentence level for English to French (en→fr), Spanish (en→es) and German (en→de). Contrastive examples are created by automatically perturbing MT outputs to produce versions of lesser quality. We include two complementary kinds of perturbations: (i) **lexical perturbations** that break coherence across sentences while leaving each individual sentence locally well-formed (e.g., inconsistent verb tenses, broken lexical or anaphoric chains, weakened conjunctions); and (ii) **structural perturbations** that introduce locally visible errors (e.g., a duplicated, removed or shuffled sentence) within an otherwise coherent document, in order to assess whether metrics remain sensitive to local errors when forced to evaluate multi-sentence spans. We evaluate eight representative metrics (reference-based and reference-free), computing document-level scores using SLIDE (Wicks and Post, 2022) with various window lengths to score increasingly long spans of text. We compare results using accuracy to assess whether (i) metrics are able to penalize document-level perturbations and (ii) they are robust to longer inputs. We provide a fine-grained view of each metric's strengths and weaknesses when applied to long spans of text, reporting results both in aggregate across perturbation types and broken down per perturbation type to isolate the specific document-level phenomenon each captures. We publicly release our test set.[1]

[1] https://github.com/nicolasdahan/metadoceval-testset

2 Related Work

Recent advances in fine-tuned and LLM-based metrics (Rei et al., 2020; Fernandes et al., 2023; Kocmi and Federmann, 2023; Juraska et al., 2024a; Freitag et al., 2024) have made it feasible to score translations over longer contexts, prompting growing interest in document-level evaluation. However, little research has examined how reliably they can be used to evaluate translated documents. Recent work has approached document-level MT and its evaluation from complementary angles: Kim (2025b) propose a holistic human-evaluation framework for document-level MT, complementary to our automatic contrastive approach; Kim (2025a) revisit how context affects document-level MT judgments through human evaluation; Luo et al. (2025) also explore sliding-window aggregation of sentence-level metrics across 16 language pairs; and O'Brien et al. (2025) release DocHPLT, a massively multilingual document-level translation dataset.

A related line of work involves designing metrics specifically for document-level evaluation. Doc-COMET (Vernikos et al., 2022) extends COMET with document context, and BlonDe (Jiang et al., 2022) targets discourse-related spans for ZH↔EN. Our focus is deliberately on the widely deployed sentence-level metrics that practitioners still apply to documents; Raunak et al. (2024) also report that sliding-window COMET correlates with human judgments better than Doc-COMET, motivating our choice of SLIDE as the document-level scoring protocol.

The evaluation of MT metrics, known as meta-evaluation, typically uses one of two approaches. The first measures correlation with human judgments of translation quality (Kocmi et al., 2021; Freitag et al., 2021a; Thompson et al., 2024; Freitag et al., 2024), the *de facto* gold standard. However, obtaining document-level annotations are costly and cognitively demanding (Castilho, 2021), and averaging sentence-level scores across a document fails to capture long-range dependencies (Deutsch et al., 2023). Correlations may also reflect spurious patterns learned during training rather than genuine translation quality (Perrella et al., 2024), and most large-scale human evaluations do not explicitly target document-level phenomena (Wicks and Post, 2023). The second approach relies on contrastive challenge sets, which introduce controlled perturbations and test whether metrics assign higher scores to the originals (Karpinska et al.,

2022; Amrhein et al., 2022; Avramidis et al., 2023; Lo et al., 2023; Zouhar et al., 2024; Wang et al., 2024; Avramidis et al., 2024). Such datasets offer fine-grained insights into metrics' sensitivity to specific linguistic phenomena without requiring human scoring. A related line outside MT uses minimally perturbed test suites to probe whether neural language models find incoherent continuations more surprising than coherent ones (Beyer et al., 2021). Within the WMT Metrics shared tasks (Freitag et al., 2021b; Freitag et al., 2022; Freitag et al., 2023), several challenge sets have targeted issues such as negation, named entities, and antonym replacement (Macketanz et al., 2022; Alves et al., 2022), helping diagnose weaknesses in sentence-level evaluation. However, existing challenge sets remain limited to short-range discourse, typically involving one or two sentences. For example, ACES (Amrhein et al., 2022) and TQ-AutoTest (Macketanz et al., 2018) cover coordination, ellipsis, and connectives, while DEMETR (Karpinska et al., 2022) extends to anaphoric phenomena. Yet none provides whole documents or multi-sentence spans, and therefore they cannot test whether metrics detect inconsistencies that arise only across sentences or paragraphs. Moreover, document-level MT systems are known to produce structural errors such as sentence omission (Wu et al., 2024; Wang et al., 2025) and re-segmentation (Yan et al., 2024), none of which are covered by existing challenge sets.

3 `MetaDocEval`: a Contrastive Framework for Document-level MT Evaluation

3.1 Overview

`MetaDocEval` is both an instantiated contrastive test set and a reusable framework for generating such test sets. Our goal is to test whether metrics can detect translation errors whose impact extends beyond the sentence level. The framework integrates perturbation generation, quality-control filtering, and sliding-window scoring into a single automated pipeline, allowing new corpora, language pairs, or perturbation types to be added with minimal manual intervention. The released test set is the result of running this pipeline on existing parallel corpora.

`MetaDocEval` currently covers three directions (*en→fr*, *en→es*, *en→de*), starting from parallel corpora (source, reference, and system outputs) in which system outputs are systematically perturbed to simulate document-level errors: perturbations are designed to preserve local meaning and coherence at the sentence level, while degrading consistency and cohesion across the document. Segments span several sentences, ranging from a paragraph to a full document.

Perturbations differ in their applicability: some changes are language-agnostic, while others depend on language-specific morphology or tense systems. Some are direction-specific, arising from source→target asymmetries (e.g., gendered plural pronouns in *en→fr*/*en→es* but not in *en→de*, where the plural form is gender-neutral). As a result, not every perturbation is instantiated for every pair.

A key aspect of this pipeline is the use of LLMs[2] for perturbations that are difficult to produce heuristically. Many discourse phenomena are rare, long-range, and hard to trigger with hand-crafted rules (e.g., maintaining lexical consistency, propagating agreement across sentences, or manipulating tense/mood). Rule-based rewriting often fails due to brittle morphological and syntactic constraints. In contrast, LLMs can condition on multiple sentences and produce targeted, minimal edits that preserve grammaticality and local meaning while subtly degrading global coherence. They also enable fine-grained control via prompts, scale across languages, and result in diverse perturbations, making them a practical tool for generating document-level contrastive examples.

3.2 Perturbations

We use two types of perturbations: (i) structural, which alter the document structure by adding, removing, or reordering sentences, and (ii) lexical, which modify words to affect the document's overall coherence and cohesion (e.g., changing tenses, swapping the gender of anaphoric pronouns, or introducing lexical inconsistencies by replacing words with synonyms that break cross-sentence consistency), thus introducing linguistic, document-level errors whose impact on meaning and coherence is primarily visible at the document level rather than within individual sentences. Table 1 summarizes the number of documents per perturbation type and language pair.

Structural perturbations We consider three structural changes designed to degrade translation quality: (i) sentence repetition, (ii) sentence shuffling, (iii) sentence removal. We also use one perturbation designed to preserve translation

[2] `gpt-4o-mini` is used for LLM-generated edits.

Perturbation	en→fr	en→es	en→de
Lexical perturbations			
LEX	59	70	55
TENSE	85	70	76
CONJ	82	96	110
PRO-SG	9	0	4
PRO-PL	10	1	0
Structural perturbations			
Sentence Removal	150	143	150
Sentence Repetition	150	143	150
Sentence Shuffling	150	143	150
Sentence Splitting	122	122	115

Table 1: Number of source documents per perturbation type and language pair (union of documents that received the perturbation under at least one of the two systems, AYA23 or GEMINI).

quality: (iv) sentence splitting,[3] implemented via a spaCy-based heuristic (Honnibal et al., 2020) that splits at most one comma per sentence only if the comma separates two independent clauses, each containing a finite verb; we then insert a period and capitalize the next token, e.g.:

ORIG (fr): Le monde évolue dans un état de changement constant, **c'est** une réalité.
PERT (fr): Le monde évolue dans un état de changement constant. **C'est** une réalité.
(gloss: 'The world evolves in a state of constant change, this is a reality.')

Lexical perturbations All lexical perturbations are generated through LLM-based edits. These edits are designed to preserve sentence-level fluency while degrading document-level quality. We target four discourse-sensitive phenomena: (i) tense consistency, (ii) lexical consistency, (iii) gender of anaphoric pronouns and (iv) conjunction substitution. Prompts are crafted to induce minimal and coherent changes, and all outputs are subsequently filtered for quality (Section 3.3).

Tense consistency (TENSE) This perturbation targets translations of the English simple past. We modify past tense verbs in the target language by switching between available past forms (*passé composé*, *imparfait*, and, where applicable, *passé simple* in French). We only consider documents containing at least four sentences in the English simple past; in such cases, we alter 50% of the aligned verbs in the target translation. Candidate verbs are identified in the source. We rewrite only the verb morphology while leaving the surrounding context unchanged. The modified sentences are designed to remain grammatical in isolation, since each substituted form is a valid past tense in the target language and the surrounding syntax is left unchanged. The random alternation of tense forms is therefore intended to introduce temporal incoherence that is only detectable with broader context. Our manual evaluation (Appendix A.6) confirms this for GEMINI outputs (94% of perturbed sentences remain acceptable in isolation) but reveals a non-trivial fraction of sentence-level degradation for AYA23 (54% acceptable), which we discuss when interpreting results (Section 6).

SRC (en): Nuclear power **was** neither clean nor safe nor cheap. Indeed, the opposite **was** true.
REF (fr): L'énergie nucléaire n'**était** ni propre, ni sûre, ni bon marché. En fait, c**était** tout le contraire.
HYP (fr): L'énergie nucléaire ne **n'était** ni propre, ni sûre, ni bon marché. En fait, ce **fut** tout le contraire.

Lexical consistency (LEX) We simulate lexical consistency[4] errors by identifying repeated content words (nouns, verbs, adjectives and adverbs identified by SpaCy) in source documents, aligning them with their target translations using SimAlign (Jalili Sabet et al., 2020) and substituting them with a random synonym.

SRC (en): As the economic **crisis** deepens and widens, the world has been searching [...] At the start of the **crisis**, many people likened it to 1982 or 1973.
ORIG (fr): Alors que la **crise** économique s'étend et s'aggrave, le monde cherche [...] Au début de la **crise**, beaucoup de gens faisaient le rapprochement avec 1982 ou 1973.
PERT (fr): Alors que la **crise** économique s'étend et s'aggrave, le monde cherche [...] Au début de la **tourmente**, beaucoup de gens faisaient le rapprochement avec 1982 ou 1973.

Gender of anaphoric pronouns (PRO) The English anaphoric pronouns *it* and *they* translate into gendered forms in French, Spanish and German, depending on the grammatical gender of their antecedent (French *il/elle* and *ils/elles*, Spanish *él/ella* and *ellos/ellas*). For German, we restrict this perturbation to *it*, as it maps to gendered forms (*er/sie/es*) determined by the grammatical gender of the antecedent noun. Following the methodology of ContraPro (Müller et al., 2018), we first compute coreference chains within each document

[3] Each perturbation affects roughly 20% of a document's sentences (minimum one).

[4] Simply understood here as the repeated use of the same lexical items across a document.

using Cofr[5] (Wilkens et al., 2020). We identify referential 3rd person pronouns that (a) have antecedent nouns in the reference chain and (b) align with the pronoun *it* or *they* in the source. To systematically alter gendered references, we extract the corresponding sentences and modify not only the pronouns but also any gender-dependent words (e.g., adjectives and past participles).

SRC (en): [...] Last week **Tony Blair, Jacques Chirac, and Gerhard Schroeder** met in Berlin. **They** departed pledging to revive Europe's growth.

ORIG (fr): [...] La semaine dernière, **Tony Blair, Jacques Chirac, et Gerhard Schroeder** se sont rencontrés à Berlin. **Ils** se sont **quittés** sur la promesse de relancer la croissance en Europe.

PERT (fr): [...] **Tony Blair, Jacques Chirac, et Gerhard Schroeder** se sont rencontrés à Berlin. **Elles** se sont **quittées** sur la promesse de relancer la croissance en Europe.

Conjunction (CONJ) Conjunctions explicitly mark discourse relations (causality, contrast, temporal succession, condition, comparison, etc.) and are deliberately chosen to ensure coherent progression across a document. We work from per-language inventories of coordinating and subordinating conjunctions grouped by discourse function (full lists in Appendix A.2 for French, A.3 for Spanish, and A.4 for German). Near-synonym substitution within a same discourse class is designed to preserve sentence-level fluency, but it weakens or distorts the discourse relations conjunctions signal. Since conjunctions appear frequently, systematically substituting several of them causes distortions to accumulate, collectively degrading document-level coherence. Importantly, because multiple conjunctions operate mainly at the discourse level rather than the sentence level, such degradation may go undetected by sentence-level metrics, making this perturbation particularly suited to probe document-level evaluation. We apply this perturbation only to documents containing at least five conjunctions, to ensure sufficient accumulation of discourse-level distortions.

ORIG (fr): **Comme** des années 2000, [...], des passe-temps et tout le reste, mais **avant** et depuis, ce n'a été que des choses comme Warhammer...

PERT (fr): **À l'instar** des années 2000, [...], des passe-temps et tout le reste, mais **auparavant** et depuis, ce n'a été que des choses comme Warhammer...

3.3 Perturbations: Quality Control

While heuristic perturbations (e.g., sentence shuffling or deletion) are deterministic and easily verifiable, LLM-based perturbations require strict quality control to ensure transformations are both successful and minimal. We adopt a multi-step filtering pipeline to discard noisy or hallucinated outputs. First, we discard any sample whose Levenshtein distance ratio exceeds a perturbation-specific threshold (see Appendix A.1 for details), ensuring that minimal edits (such as synonym substitutions) remain close to the original while hallucinated (e.g., explanations) or paraphrased outputs are filtered out. Across all perturbation types, samples where no modification occurred are discarded. We then apply targeted checks depending on the perturbation type. For lexical consistency, the modified word is aligned with its original counterpart and compared using BERTScore; samples with low similarity are removed as contextually invalid. For tense consistency, we prompt an LLM to verify that the tense actually changed; samples where original and perturbed are judged to be in the same tense (indicating the perturbation failed) are discarded.[6] For conjunction substitutions, samples where the output is identical to the input are similarly rejected. Finally, for the PRO perturbation, another LLM-based check ensures that the gender expressed actually differs between the two variants; if not, the perturbation is rejected.

4 Evaluating Metrics using `metadoceval`

We use the `metadoceval` challenge set to assess how existing MT evaluation metrics behave when applied at the document level. Our goal is to quantify their sensitivity to the controlled perturbations of Section 3.2, and to determine how this sensitivity changes as the available context length increases.

4.1 Computing Document-Level Scores

To obtain document-level scores, we apply each metric to contiguous blocks of sentences taken from system outputs or their perturbed variants. We follow the sliding-window approach of (Raunak et al., 2024), denoted SLIDE$(w, 1)$, where w is the

[5] https://boberle.com/projects/coreference-resolution-with-cofr

[6] We perform a manual evaluation of the four LLM-based perturbations (tense consistency, lexical consistency, conjunction substitution, and gender of anaphoric pronouns (PRO)) to verify that they preserve sentence-level acceptability in Appendix A.6. Across the four perturbations, 73–98% of perturbed outputs preserve sentence-level acceptability overall. The only notable system-specific outlier is tense for AYA23 (54%), where the perturbation appears to conflate sentence-level and discourse-level effects. The per-system breakdown in Appendix A.5 shows that the patterns and conclusions of our aggregated analysis hold for GEMINI alone, where sentence-level acceptability is uniformly high (71–100%).

window size in sentences and the stride is 1. This ensures that every sentence in a document appears in multiple overlapping windows, maximizing the chances that a perturbation is observed with enough surrounding context. For a given document, we compute metric scores for each individual window (*chunk-level scores*) and average the scores to give a single *document-level score*. Finally, we average over the entire test set to obtain a *corpus-level score*. Using $w = 1$ produces a sentence-level evaluation in which document-level errors (e.g., all lexical and structural perturbations outside the targeted sentences) are inaccessible to the metric. Larger window sizes progressively incorporate more context, allowing us to investigate whether metrics can exploit longer spans to detect discourse-level errors. This setup reveals how metrics react to both perturbation type and input span.

4.2 Comparison of Scores

We compare the metric scores assigned to the original MT outputs with those assigned to their perturbed counterparts. For each metric, perturbation type, and window size, we compute the mean difference in document-level scores, aggregated across language pairs and systems; per-language-pair breakdowns are provided in Appendix A.5. A metric that is sensitive to a given perturbation should consistently assign better scores to the original outputs. We quantify this sensitivity in two ways. First, we run two-sided paired t-tests over per-document score differences to assess whether the observed gaps are statistically significant. Second, we compute *accuracy*, the proportion of document pairs where the initial translation is scored higher than its perturbed variant. Accuracy gives a direct measure of ranking quality, independent of score scales.[7] A metric's ability to detect a perturbation will depend on three interacting factors: whether it uses the reference, the source, or both; whether the perturbation is structural or lexical; and the amount of accessible context.

[7] Absolute score values across window sizes cannot be compared directly, as some metrics exhibit systematic drifts as the window size increases.

5 Experimental settings

5.1 Metrics

We evaluate a range of MT metrics: BLEU[8] and chrF (surface-level metrics based on n-gram overlap), BERTScore[9] (an embedding-based metric), fine-tuned metrics: COMET[10] and METRICX-24[11](Juraska et al., 2024b) and their reference-free versions COMETKIWI[12] and METRICX-24-QE [11] and GEMBA-MQM,[13] a GPT-based metric, which relies on structured prompts.

5.2 WMT24++ Datasets

We perform our evaluation using the WMT24++ parallel test sets (Deutsch et al., 2025) for three language pairs: English–French (en→fr), English–Spanish (en→es), and English–German (en→de). These test sets, of which the original documents were drawn from diverse sources for the WMT shared task (Kocmi et al., 2025) (literary, speech, news and social) include human references, and multiple system outputs.

We only keep documents containing at least three sentences, and focus on the outputs of two representative systems: GEMINI-1.5 PRO (Team et al., 2024), ranked among the top-tier systems in WMT24++, and AYA23 (Aryabumi et al., 2024), ranked among the lowest-tier systems for all three language pairs, giving us a diverse range of MT outputs to test the metrics on.

To ensure consistency in evaluation and avoid biases stemming from segment length variability (e.g., documents with single long segments vs. multiple short ones), we realign all documents (source, reference, and system outputs) using `bertalign` (Liu and Zhu, 2023).[14] This produces a consistent sentence-level segmentation across all inputs, where each document is treated as a sequence of aligned segments, and each segment corresponds to a sentence. Only documents for which this triple alignment (source, reference, system output) succeeds are retained for analysis, resulting in a reli-

[8] We use SacreBLEU (Post, 2018) with signature: `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.4.3`. BLEU is computed per window: the w consecutive sentences in each window are concatenated into a single text and scored as a single unit using SacreBLEU's `corpus_bleu`.

[9] `facebook/bart-large-mnli`

[10] `Unbabel/wmt22-comet-da`

[11] `google/metricx-24-hybrid-large-v2p6`

[12] `Unbabel/wmt22-cometkiwi-da`

[13] QE version using `gpt-4o-mini`

[14] `github.com/bfsujason/bertalign`

able set of well-aligned documents with an average length of 11.3 sentences. We exclude the PRO perturbation from our evaluation, as there are too few gendered instances in the WMT24++ to generate enough contrastive examples (Table 1).

6 Results and Analyses

Figure 1 presents document-level accuracy for each metric and perturbation, micro-averaged across three language pairs (*en→fr*, *en→es*, *en→de*) and two MT systems (GEMINI-1.5 PRO and AYA23); per-language-pair and per-system results are provided in Appendix A.5. We report accuracy with window sizes of 1, 3, 6 and 9; chance performance is 50%.

6.1 Two Distinct Failure Modes

A fundamental asymmetry separates structural and lexical perturbations, and understanding it is key to interpreting all subsequent results.

Structural perturbations are easy to detect at the sentence level but less easy with context. Sentence removal, repetition, and shuffling are easily detected at the sentence-level, since they inevitably result in comparisons across aligned sentences of sentences that are no longer translations (at w=1, nearly all metrics achieve 90–100% accuracy). As w increases, accuracy drops across the board, as the chance of observing correct translations within the window being evaluated (albeit in the wrong order or only partially) increases. The underlying reason differs by perturbation type. For sentence removal and repetition, the distortion signal is diluted by surrounding coherent sentences scored jointly within the window. For sentence shuffling, the drop may reflect a different effect: at w=1, each inverted sentence is individually penalized, whereas at larger w, the metric sees a mix of correctly and incorrectly ordered sentences within the same window, which can partially offset the error signal. A notable outlier is METRICX-24-QE on sentence repetition: at w=1, it scores *below* the 50% chance baseline, initially *preferring* duplicated outputs. This likely stems from its MQM-based training objective, which conflates input length with error severity: at w=1, a duplicated sentence doubles the output length, producing a lower (i.e. seemingly better) MQM score. At w=3, the relative length difference shrinks and the redundancy signal becomes detectable, correcting the bias.

Sentence splitting stands apart from the other structural perturbations as it is designed to preserve translation quality while altering the structure, so we would hope that the metrics are less sensitive to the perturbation than the others. However most metrics do penalize splits at w=1 (95–97% accuracy for BLEU, chrF, BERTScore). More concerning, COMETKIWI increasingly penalizes splits as context grows (68%→85%), and COMET follows a similar trend (76%→89%), suggesting that these metrics amplify false positives with broader context rather than recognizing that the translation remains adequate.

Lexical perturbations expose a reference bias at the sentence level. By design, tense consistency, lexical consistency, and conjunction substitution perturbations are intended to preserve sentence-level fluency and adequacy while degrading global coherence. Our manual evaluation (Appendix A.6) confirms this design across the board (73–98% overall), with the only striking exception of tense consistency for AYA23 (54%). We would therefore expect accuracy at w=1 to be near chance (50%). This expected pattern is broadly observed for reference-free metrics (and for METRICX-24 and COMET for conjunction substitution), which mostly score between 40% and 60% at w=1, though not uniformly: GEMBA-MQM reaches 61% on tense consistency, suggesting some sentence-level sensitivity to tense changes, and METRICX-24-QE reaches 61% on lexical consistency. A notable exception is COMETKIWI, which systematically scores below 40% across all three lexical perturbations (34% for conjunction substitution, 32% for tense consistency, 38% for lexical consistency), indicating that it actively favors the perturbed variants at the sentence level. Reference-based metrics (BLEU, chrF, BERTScore), however, already score well above chance at w=1, reaching over 75% for conjunction substitution and around 80% or above for lexical consistency. This is not evidence of discourse sensitivity: as shown in Table 2, the original modified fragment appears in the reference in over 50% of cases for lexical consistency and conjunction substitution, and 24% for tense consistency. Reference-based metrics therefore inevitably penalize the perturbation at the sentence level as it often no longer matches the reference text.

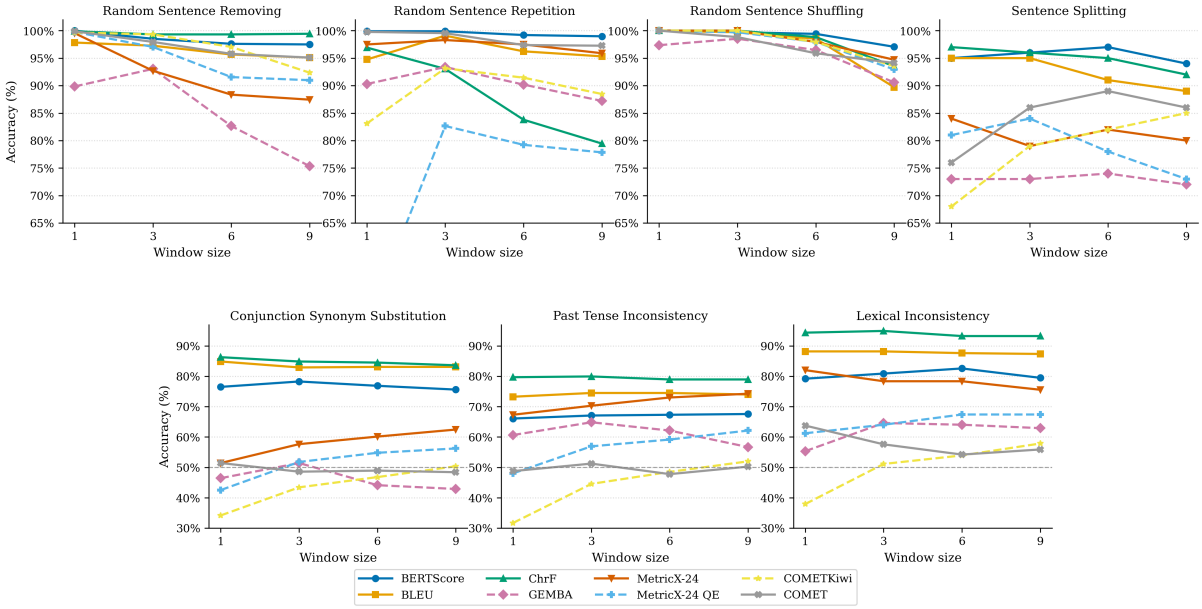

Figure 1: Average document-level accuracy (%) for each evaluation metric, computed for varying window sizes for SLIDE(w,1) with $w \in \{1, 3, 6, 9\}$, aggregated across three language pairs and two MT systems. Dashed lines represent QE metrics. Structural perturbations are on the top row, lexical perturbations on the bottom row. Detailed results per language pair and system are in Appendix A.5.

Perturbation	n	***orig*∈*ref***	***pert*∈*ref***
LEX	956	56.28	4.50
TENSE	956	24.58	3.24
CONJ	1532	53.39	6.92

Table 2: For each lexical perturbation type the percentage of cases where the original modified fragment appears in the reference (*orig*∈*ref*) and the perturbed fragment appears in the reference (*pert*∈*ref*).

6.2 Metric Families Reveal Systematic Patterns

The most informative way to read Figure 1 is by metric family, as structural design determines behavior more than any other factor.

Reference-only metrics are biased, not accurate. BLEU, CHRF, and BERTScore achieve the highest accuracies on lexical perturbations, but for the wrong reason. Their reliance on surface overlap with the reference means that in cases where the perturbation introduces divergence from the reference (Table 2), lexical consistency and tense consistency perturbations are penalized at w=1 regardless of their discourse validity. As context grows, their trajectories remain flat across all three metrics, confirming that added context provides no additional signal. BLEU and CHRF are structurally limited by their n-gram formulations ($n \leq 4$), which cannot capture discourse-level dependencies. BERTScore, despite relying on contextual embeddings rather than n-grams, exhibits the same flat behavior: on lexical consistency, it scores 79% at w=1 and 79% at w=9; on tense consistency, 66% and 68%; on conjunction substitution, 77% and 76%. This suggests that even though its underlying encoder has a broad context window, BERTScore's token-level matching mechanism fails to exploit document-level information, reducing it in practice to a surface-overlap metric similar to its n-gram counterparts.

Reference+source metrics add grounding but stall on document context. COMET and METRICX-24 aim to reduce reference-overlap bias through source access, yet their behaviors diverge by perturbation type. On lexical consistency, both start with high accuracy at w=1 ($\approx$ 64% and $\approx$ 82% respectively), consistent with the reference divergence effect discussed above, and both decrease as w grows. On conjunction substitution and tense consistency, COMET hovers near chance across all window sizes, suggesting it gains nothing from either the reference or the source for these phenomena. METRICX-24, by contrast, starts near 50% on conjunction substitution but increases continuously to 62% at w=9, and similarly increases on tense consistency (67% $\rightarrow$ 74%), suggesting it can exploit broader context to detect these discourse errors. Notably, on conjunction substitution and

tense consistency, METRICX-24 and METRICX-24-QE exhibit nearly identical variation patterns across all window sizes, with their mean accuracies differing by an approximately constant offset, a direct consequence of their shared underlying model weights trained on a mixture of reference-based and reference-free examples.

COMET and COMETKIWI are mirror images. On all three lexical perturbations, COMETKIWI consistently rises from well below chance as w increases, while COMET's behavior depends on the perturbation type. The clearest mirror appears on lexical consistency, where COMET drops from 64% to 56% while COMETKIWI climbs from 38% to 58%. On conjunction substitution and tense consistency, COMET scores near chance at $w{=}1$ ($\approx$49–51%), which is expected since these perturbations preserve sentence-level quality, but it remains flat as context grows. COMETKIWI starts well below chance (34%, 32%) but rises steadily to 50% and 52%. These contrasting trajectories suggest that reference access plays different roles: on lexical consistency, it is *actively harmful*, inflating accuracy at $w{=}1$ through surface overlap then diluting with context; on conjunction substitution and tense consistency, it is simply uninformative. In all cases, COMETKIWI is the only metric whose accuracy consistently increases with context. However, even at $w{=}9$ it barely reaches chance level (50–58%).

LLM-based scoring is unreliable beyond short spans. GEMBA-MQM follows a characteristic pattern on most perturbations: it stays flat or improves slightly at $w{=}3$, then drops at $w{=}6$ and $w{=}9$. This is most pronounced on sentence removal, where it starts with the *lowest* accuracy among all metrics at $w{=}1$ ($\approx$90%), peaks at $w{=}3$ (93%), then drops to 75% at $w{=}9$. A similar pattern appears on sentence shuffling, tense consistency, and conjunction substitution. On conjunction substitution, it reaches the lowest accuracy among all metrics at $w{=}9$ (36.2% for en$\rightarrow$fr). However, this pattern is not universal: on lexical consistency, GEMBA-MQM improves sharply at $w{=}3$ ($55\%{\rightarrow}65\%$) but then plateaus rather than declining. Overall, these results suggest that LLM-based scoring with short prompts cannot reliably process extended inputs, despite the model's large context window.

No single window size works universally. Structural errors are best detected at $w{=}1$, where they are isolated and directly measurable. Lexical discourse errors require broader context to become visible. These opposing demands explain why no metric achieves uniformly high accuracy across all conditions. The most reliable compromise is $w{=}3$: it provides enough local context for reference-free metrics to begin detecting discourse errors ($+10$–45% gains over $w{=}1$) while limiting the dilution that degrades all metrics at $w{=}9$. Importantly, for most metrics the accuracy drop at larger window sizes is not an artifact of metrics becoming inherently unreliable with longer inputs: scoring the same unperturbed outputs across window sizes shows consistent shifts without increased variance, preserving system rankings (Appendix A.5.4). The drop is therefore attributable to perturbation signal dilution, not metric degradation. However, METRICX-24 and METRICX-24-QE exhibit a stronger length-dependent score drift due to their MQM-based training objective, which conflates input length with error severity, adding an additional confound to their cross-window comparisons.

6.3 Practical Recommendations

None of the evaluated metrics were designed to assess the targeted document-level phenomena, and their behaviors clearly reflect this limitation. Nevertheless, several actionable guidelines emerge:

1. **Avoid large evaluation windows.** For the document lengths in our evaluation (11.3 sentences on average), all metrics degrade at $w{=}9$: structural errors (removal, repetition, shuffling) are penalized less when scored jointly with many surrounding coherent sentences.

2. **Do not score sentences independently either.** Genuine discourse-level inconsistencies (tense consistency, lexical consistency, conjunction substitution) are designed to preserve sentence-level fluency and adequacy, and are thus largely undetectable at $w{=}1$, mainly for reference-free metrics.

3. **Use short windows of $\approx$3 sentences.** This provides the best trade-off: enough context for discourse error detection without excessive signal dilution.

4. **When references are available, use them but be aware of their bias.** Reference-based

and reference+source metrics achieve the highest absolute accuracies, particularly on structural perturbations, and remain the most reliable option when references exist. However, their high scores on lexical perturbations are largely artifactual: they penalize any surface deviation from the reference, including valid alternatives. Practitioners should therefore not interpret high reference-based scores as evidence that document-level coherence has been assessed.

5. **No current metric genuinely captures document-level phenomena.** Among reference-free metrics, COMETKIWI and METRICX-24-QE are the only metrics whose accuracy *increases* with context on lexical perturbations, showing that they can exploit the source to detect discourse disruptions. Yet even at their best (w=9), they barely exceed chance ($\approx 50\%$), meaning the signal is real but far too weak for practical use. METRICX-24-QE is the most stable across conditions but similarly limited in absolute terms. These results point to a clear research gap rather than a usable solution.

7 Conclusion

We presented `MetaDocEval`, an automatically constructed contrastive test set for evaluating MT metrics on document-level phenomena. By introducing controlled discourse-sensitive perturbations into MT system outputs and scoring them under a sliding-window protocol with increasing context lengths, we exposed how current metrics behave when moving beyond the sentence level. Our central finding is that no evaluated metric genuinely captures our targeted document-level phenomena, though the reasons differ instructively by metric family. We evaluate metrics by type of perturbation (structural or lexical) and discuss the implications of carrying out quality estimation versus reference-based evaluation. Finally, we provide some practical recommendations based on our observations and how existing metrics can be used to evaluate longer spans of text.

8 Limitations

We acknowledge several limitations of this work.

Perturbation coverage. Our perturbations target a specific set of discourse phenomena (tense consistency, lexical consistency, gender of anaphoric pronouns, conjunction substitution, and structural edits). Important document-level challenges are not yet covered, notably lexical ambiguities whose resolution depends on document context (Bawden et al., 2018), coreference beyond pronoun gender, and discourse-level register or style shifts.

Language and data scope. The linguistic perturbations are tailored to three Western European language pairs (en→fr, en→es, en→de), which share substantial typological overlap. Languages with richer morphology, freer word order, or different discourse-structuring devices (e.g., pro-drop languages, SOV languages) may expose different metric behaviors. The underlying dataset (WMT24++) is also relatively small, and the PRO perturbation had to be excluded due to insufficient instances, limiting our analysis of anaphoric phenomena.

Metric configuration. For GEMBA-MQM, we used `gpt-4o-mini` rather than the larger model recommended by the original authors, and did not explore recent advances in prompt design for long-context evaluation (Agrawal et al., 2024). The collapse we observe beyond w=3 may therefore partly reflect suboptimal prompting rather than a fundamental limitation of LLM-based scoring. Similarly, all metrics were used with default configurations; task-specific tuning (e.g., adjusting BERTScore's layer selection for longer inputs) might alter some findings.

Evaluation protocol. Our sliding-window approach treats all sentences within a window equally, but perturbations affect only a subset of them. A metric that assigns correct local scores but averages them with unaffected sentences will appear to degrade, an effect we attribute to "dilution" but that is partly an artifact of uniform averaging. Alternative aggregation strategies (e.g., weighting sentences by perturbation density) could yield different conclusions and deserve investigation.

Sentence-level artifacts in LLM-generated perturbations. LLM-generated perturbations may, in addition to the intended document-level change, introduce unintended sentence-level defects (awkward synonyms, locally infelicitous verb forms, agreement errors), so that a metric could appear sensitive to a perturbation for the wrong reason. To gauge this risk we inspect the per-system breakdown of the metric analysis in Appendix A.5; our

manual evaluation (Appendix A.6) shows that GEMINI's perturbed outputs preserve sentence-level acceptability uniformly well across all four perturbations (71–100% *Both valid*), limiting the potential contamination from sentence-level artifacts. The patterns and conclusions of our aggregated analysis hold when focusing on GEMINI outputs alone.

Acknowledgements

This work was supported by the French national research agency (ANR) as part of the MaTOS project (grant number: ANR-22-CE23-0033).[15] Rachel Bawden was also partly funded by her chair position in the PRAIRIE institute funded by ANR as part of the "Investissements d'avenir" programme under reference ANR19-P3IA-0001. The authors are grateful to the anonymous reviewers for their insightful comments and suggestions and to Ziqian Peng and Paul Lerner for their review and feedback on a preliminary draft of this work.

References

Agrawal, Sweta, Amin Farajian, Patrick Fernandes, Ricardo Rei, and André F. T. Martins. 2024. Assessing the role of context in chat translation evaluation: Is context helpful and under what conditions? *Transactions of the Association for Computational Linguistics*, 12:1250–1267.

Alves, Duarte, Ricardo Rei, Ana C Farinha, José G. C. de Souza, and André F. T. Martins. 2022. Robust MT evaluation with sentence-level multilingual augmentation. In Koehn, Philipp, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 469–478, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

Amrhein, Chantal, Nikita Moghe, and Liane Guillou. 2022. ACES: Translation accuracy challenge sets for evaluating machine translation metrics. In Koehn, Philipp, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 479–513, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

Aryabumi, Viraat, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress.

Avramidis, Eleftherios, Shushen Manakhimova, Vivien Macketanz, and Sebastian Möller. 2023. Challenging the state-of-the-art machine translation metrics from a linguistic perspective. In Koehn, Philipp, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 713–729, Singapore, December. Association for Computational Linguistics.

Avramidis, Eleftherios, Shushen Manakhimova, Vivien Macketanz, and Sebastian Möller. 2024. Machine translation metrics are better in evaluating linguistic errors on LLMs than on encoder-decoder systems. In Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 517–528, Miami, Florida, USA, November. Association for Computational Linguistics.

Bawden, Rachel, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In Walker, Marilyn, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313, New Orleans, Louisiana, June. Association for Computational Linguistics.

Beyer, Anne, Sharid Loáiciga, and David Schlangen. 2021. Is Incoherence Surprising? Targeted Evaluation of Coherence Prediction from Language Models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4164–4173, Online. Association for Computational Linguistics.

Castilho, Sheila. 2021. Towards document-level human MT evaluation: On the issues of annotator agreement, effort and misevaluation. In Belz, Anya, Shubham Agarwal, Yvette Graham, Ehud Reiter, and Anastasia Shimorina, editors, *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 34–45, Online, April. Association for Computational Linguistics.

[15] http://anr-matos.fr/

Deutsch, Daniel, Juraj Juraska, Mara Finkelstein, and Markus Freitag. 2023. Training and meta-evaluating machine translation evaluation metrics at the paragraph level. In Koehn, Philipp, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 996–1013, Singapore, December. Association for Computational Linguistics.

Deutsch, Daniel, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. WMT24++: Expanding the Language Coverage of WMT24 to 55 Languages & Dialects.

Fernandes, Patrick, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In Koehn, Philipp, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore, December. Association for Computational Linguistics.

Freitag, Markus, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Freitag, Markus, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In Barrault, Loic, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online, November. Association for Computational Linguistics.

Freitag, Markus, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In Koehn, Philipp, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

Freitag, Markus, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In Koehn, Philipp, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore, December. Association for Computational Linguistics.

Freitag, Markus, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs Breaking MT Metrics? Results of the WMT24 Metrics Shared Task. In Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA, November. Association for Computational Linguistics.

Honnibal, Matthew, Ines Montani, Sofie van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in python.

Jalili Sabet, Masoud, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In Cohn, Trevor, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online, November. Association for Computational Linguistics.

Jiang, Yuchen Eleanor, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022. BlonDe: An automatic evaluation metric for document-level machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1550–1565, Seattle, United States. Association for Computational Linguistics.

Juraska, Juraj, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024a. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA, November. Association for Computational Linguistics.

Juraska, Juraj, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024b. MetricX-24: The Google

submission to the WMT 2024 metrics shared task. In Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA, November. Association for Computational Linguistics.

Karpinska, Marzena, Nishant Raj, Katherine Thai, Yixiao Song, Ankita Gupta, and Mohit Iyyer. 2022. DEMETR: Diagnosing evaluation metrics for translation. In Goldberg, Yoav, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9540–9561, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Kim, Ahrii. 2025a. Context is ubiquitous, but rarely changes judgments: Revisiting document-level MT evaluation. In *Proceedings of the Tenth Conference on Machine Translation*, pages 81–97, Suzhou, China. Association for Computational Linguistics.

Kim, Ahrii. 2025b. Falcon: Holistic framework for document-level machine translation evaluation. TechRxiv.

Kocmi, Tom and Christian Federmann. 2023. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In Koehn, Philipp, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore, December. Association for Computational Linguistics.

Kocmi, Tom, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In Barrault, Loic, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online, November. Association for Computational Linguistics.

Kocmi, Tom, 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 Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinthór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China, November. Association for Computational Linguistics.

Läubli, Samuel, Rico Sennrich, and Martin Volk. 2018. Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. In Riloff, Ellen, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium, October-November. Association for Computational Linguistics.

Lavie, Alon, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhujan, 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 Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Tenth Conference on Machine Translation*, pages 436–483, Suzhou, China, November. Association for Computational Linguistics.

Liu, Lei and Min Zhu. 2023. Bertalign: Improved word embedding-based sentence alignment for Chinese–English parallel corpora of literary texts. *Digital Scholarship in the Humanities*, 38(2):621–634, May.

Lo, Chi-kiu, Samuel Larkin, and Rebecca Knowles. 2023. Metric score landscape challenge (MSLC23): Understanding metrics' performance on a wider landscape of translation quality. In Koehn, Philipp, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 776–799, Singapore, December. Association for Computational Linguistics.

Luo, Yuanchang, Jiaxin Guo, Daimeng Wei, Hengchao Shang, Zongyao Li, Zhiqiang Rao, Jinlong Yang, Zhanglin Wu, Xiaoyu Chen, and Hao Yang. 2025. HW-TSC's submissions to the WMT 2025 segment-level quality score prediction task. In *Proceedings of the Tenth Conference on Machine Translation*, pages 969–973, Suzhou, China. Association for Computational Linguistics.

Macketanz, Vivien, Renlong Ai, Aljoscha Burchardt, and Hans Uszkoreit. 2018. TQ-AutoTest – an automated test suite for (machine) translation quality. In Calzolari, Nicoletta, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Macketanz, Vivien, Eleftherios Avramidis, Aljoscha Burchardt, He Wang, Renlong Ai, Shushen Manakhimova, Ursula Strohriegel, Sebastian Möller, and Hans Uszkoreit. 2022. A linguistically motivated test suite to semi-automatically evaluate German–English machine translation output. In Calzolari, Nicoletta, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 936–947, Marseille, France, June. European Language Resources Association.

Maruf, Sameen, Fahimeh Saleh, and Gholamreza Haffari. 2021. A Survey on Document-Level Neural Machine Translation: Methods and Evaluation. *ACM Comput. Surv.*, 54(2), March. Place: New York, NY, USA Publisher: Association for Computing Machinery.

Müller, Mathias, 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 Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium, October. Association for Computational Linguistics.

O'Brien, Dayyán, Bhavitvya Malik, Ona de Gibert, Pinzhen Chen, Barry Haddow, and Jörg Tiedemann. 2025. DocHPLT: A massively multilingual document-level translation dataset. In *Proceedings of the Tenth Conference on Machine Translation*, pages 286–300, Suzhou, China. Association for Computational Linguistics.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Isabelle, Pierre, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Perrella, Stefano, Lorenzo Proietti, Alessandro Scirè, Edoardo Barba, and Roberto Navigli. 2024. Guardians of the machine translation meta-evaluation: Sentinel metrics fall in! In Ku, Lun-Wei, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16216–16244, Bangkok, Thailand, August. Association for Computational Linguistics.

Popović, Maja. 2015. chrF: character n-gram F-score for automatic MT evaluation. In Bojar, Ondřej, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

Post, Matt and Marcin Junczys-Dowmunt. 2024. Escaping the sentence-level paradigm in machine translation.

Post, Matt. 2018. A call for clarity in reporting BLEU scores. In Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October. Association for Computational Linguistics.

Raunak, Vikas, Tom Kocmi, and Matt Post. 2024. SLIDE: Reference-free evaluation for machine translation using a sliding document window. In Duh, Kevin, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 205–211, Mexico City, Mexico, June. Association for Computational Linguistics.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In Webber, Bonnie, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November. Association for Computational Linguistics.

Team, Gemini, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, DJ Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He,

Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh, Aakanksha Chowdhery, Yang Xu, Mehran Kazemi, Ehsan Amid, Anastasia Petrushkina, Kevin Swersky, Ali Khodaei, Gowoon Chen, Chris Larkin, Mario Pinto, Geng Yan, Adria Puigdomenech Badia, Piyush Patil, Steven Hansen, Dave Orr, Sebastien M. R. Arnold, Jordan Grimstad, Andrew Dai, Sholto Douglas, Rishika Sinha, Vikas Yadav, Xi Chen, Elena Gribovskaya, Jacob Austin, Jeffrey Zhao, Kaushal Patel, Paul Komarek, Sophia Austin, Sebastian Borgeaud, Linda Friso, Abhimanyu Goyal, Ben Caine, Kris Cao, Da-Woon Chung, Matthew Lamm, Gabe Barth-Maron, Thais Kagohara, Kate Olszewska, Mia Chen, Kaushik Shivakumar, Rishabh Agarwal, Harshal Godhia, Ravi Rajwar, Javier Snaider, Xerxes Dotiwalla, Yuan Liu, Aditya Barua, Victor Ungureanu, Yuan Zhang, Bat-Orgil Batsaikhan, Mateo Wirth, James Qin, Ivo Danihelka, Tulsee Doshi, Martin Chadwick, Jilin Chen, Sanil Jain, Quoc Le, Arjun Kar, Madhu Gurumurthy, Cheng Li, Ruoxin Sang, Fangyu Liu, Lampros Lamprou, Rich Munoz, Nathan Lintz, Harsh Mehta, Heidi Howard, Malcolm Reynolds, Lora Aroyo, Quan Wang, Lorenzo Blanco, Albin Cassirer, Jordan Griffith, Dipanjan Das, Stephan Lee, Jakub Sygnowski, Zach Fisher, James Besley, Richard Powell, Zafarali Ahmed, Dominik Paulus, David Reitter, Zalan Borsos, Rishabh Joshi, Aedan Pope, Steven Hand, Vittorio Selo, Vihan Jain, Nikhil Sethi, Megha Goel, Takaki Makino, Rhys May, Zhen Yang, Johan Schalkwyk, Christina Butterfield, Anja Hauth, Alex Goldin, Will Hawkins, Evan Senter, Sergey Brin, Oliver Woodman, Marvin Ritter, Eric Noland, Minh Giang, Vijay Bolina, Lisa Lee, Tim Blyth, Ian Mackinnon, Machel Reid, Obaid Sarvana, David Silver, Alexander Chen, Lily Wang, Loren Maggiore, Oscar Chang, Nithya Attaluri, Gregory Thornton, Chung-Cheng Chiu, Oskar Bunyan, Nir Levine, Timothy Chung, Evgenii Eltyshev, Xiance Si, Timothy Lillicrap, Demetra Brady, Vaibhav Aggarwal, Boxi Wu, Yuanzhong Xu, Ross McIlroy, Kartikeya Badola, Paramjit Sandhu, Erica Moreira, Wojciech Stokowiec, Ross Hemsley, Dong Li, Alex Tudor, Pranav Shyam, Elahe Rahimtoroghi, Salem Haykal, Pablo Sprechmann, Xiang Zhou, Diana Mincu, Yujia Li, Ravi Addanki, Kalpesh Krishna, Xiao Wu, Alexandre Frechette, Matan Eyal, Allan Dafoe, Dave Lacey, Jay Whang, Thi Avrahami, Ye Zhang, Emanuel Taropa, Hanzhao Lin, Daniel Toyama, Eliza Rutherford, Motoki Sano, HyunJeong Choe, Alex Tomala, Chalence Safranek-Shrader, Nora Kassner, Mantas Pajarskas, Matt Harvey, Sean Sechrist, Meire Fortunato, Christina Lyu, Gamaleldin Elsayed, Chenkai Kuang, James Lottes, Eric Chu, Chao Jia, Chih-Wei Chen, Peter Humphreys, Kate Baumli, Connie Tao, Rajkumar Samuel, Cicero Nogueira dos Santos, Anders Andreassen, Nemanja Rakićević, Dominik Grewe, Aviral Kumar, Stephanie Winkler, Jonathan Caton, Andrew Brock, Sid Dalmia, Hannah Sheahan, Iain Barr, Yingjie Miao, Paul Natsev, Jacob Devlin, Feryal Behbahani, Flavien Prost, Yanhua Sun, Artiom Myaskovsky, Thanumalayan Sankaranarayana Pillai, Dan Hurt, Angeliki Lazaridou, Xi Xiong, Ce Zheng, Fabio Pardo, Xiaowei Li, Dan Horgan, Joe Stanton, Moran Ambar, Fei Xia, Alejandro Lince, Mingqiu Wang, Basil Mustafa, Albert Webson, Hyo Lee, Rohan Anil, Martin Wicke, Timothy Dozat, Abhishek Sinha, Enrique Piqueras, Elahe Dabir, Shyam Upadhyay, Anudhyan Boral, Lisa Anne Hendricks, Corey Fry, Josip Djolonga, Yi Su, Jake Walker, Jane Labanowski, Ronny Huang, Vedant Misra, Jeremy Chen, RJ Skerry-Ryan, Avi Singh, Shruti Rijhwani, Dian Yu, Alex Castro-Ros, Beer Changpinyo, Romina Datta, Sumit Bagri, Arnar Mar Hrafnkelsson, Marcello Maggioni, Daniel Zheng, Yury Sulsky, Shaobo Hou, Tom Le Paine, Antoine Yang, Jason Riesa, Dominika Rogozinska, Dror Marcus, Dalia El Badawy, Qiao Zhang, Luyu Wang, Helen Miller, Jeremy Greer, Lars Lowe Sjos, Azade Nova, Heiga Zen, Rahma Chaabouni, Mihaela Rosca, Jiepu Jiang, Charlie Chen, Ruibo Liu, Tara Sainath, Maxim Krikun, Alex Polozov, Jean-Baptiste Lespiau, Josh Newlan, Zeyncep Cankara, Soo Kwak, Yunhan Xu, Phil Chen, Andy Coenen, Clemens Meyer, Katerina Tsihlas, Ada Ma, Juraj Gottweis, Jinwei Xing, Chenjie Gu, Jin Miao, Christian Frank, Zeynep Cankara, Sanjay Ganapathy, Ishita Dasgupta, Steph Hughes-Fitt, Heng Chen, David Reid, Keran Rong, Hongmin Fan, Joost van Amersfoort, Vincent Zhuang, Aaron Cohen, Shixiang Shane Gu, Anhad Mohananey, Anastasija Ilic, Taylor Tobin, John Wieting, Anna Bortsova, Phoebe Thacker, Emma Wang, Emily Caveness, Justin Chiu, Eren Sezener, Alex Kaskasoli, Steven Baker, Katie Millican, Mohamed Elhawaty, Kostas Aisopos, Carl Lebsack, Nathan Byrd, Hanjun Dai, Wenhao Jia, Matthew Wiethoff, Elnaz Davoodi, Albert Weston, Lakshman Yagati, Arun Ahuja, Isabel Gao, Golan Pundak, Susan Zhang, Michael Azzam, Khe Chai Sim, Sergi Caelles, James Keeling, Abhanshu Sharma, Andy Swing, YaGuang Li, Chenxi Liu, Carrie Grimes Bostock, Yamini Bansal, Zachary Nado, Ankesh Anand, Josh Lipschultz, Abhijit Karmarkar, Lev Proleev, Abe Ittycheriah, Soheil Hassas Yeganeh, George Polovets, Aleksandra Faust, Jiao Sun, Alban Rrustemi, Pen Li, Rakesh Shivanna, Jeremiah Liu, Chris Welty, Federico Lebron, Anirudh Baddepudi, Sebastian Krause, Emilio Parisotto, Radu Soricut, Zheng Xu, Dawn Bloxwich, Melvin Johnson, Behnam Neyshabur, Justin Mao-Jones, Renshen Wang, Vinay Ramasesh, Zaheer Abbas, Arthur Guez, Constant Segal, Duc Dung Nguyen, James Svensson, Le Hou, Sarah York, Kieran Milan, Sophie Bridgers, Wiktor Gworek, Marco Tagliasacchi, James Lee-Thorp, Michael Chang, Alexey Guseynov, Ale Jakse Hartman, Michael Kwong, Ruizhe Zhao, Sheleem Kashem, Elizabeth Cole, Antoine Miech, Richard Tanburn, Mary Phuong, Filip Pavetic, Sebastien Cevey, Ramona Comanescu, Richard Ives, Sherry Yang, Cosmo Du, Bo Li, Zizhao Zhang, Mariko Iinuma, Clara Huiyi Hu, Aurko Roy, Shaan Bijwadia, Zhenkai Zhu, Danilo Martins, Rachel Sapu-

tro, Anita Gergely, Steven Zheng, Dawei Jia, Ioannis Antonoglou, Adam Sadovsky, Shane Gu, Yingying Bi, Alek Andreev, Sina Samangooei, Mina Khan, Tomas Kocisky, Angelos Filos, Chintu Kumar, Colton Bishop, Adams Yu, Sarah Hodkinson, Sid Mittal, Premal Shah, Alexandre Moufarek, Yong Cheng, Adam Bloniarz, Jaehoon Lee, Pedram Pejman, Paul Michel, Stephen Spencer, Vladimir Feinberg, Xuehan Xiong, Nikolay Savinov, Charlotte Smith, Siamak Shakeri, Dustin Tran, Mary Chesus, Bernd Bohnet, George Tucker, Tamara von Glehn, Carrie Muir, Yiran Mao, Hideto Kazawa, Ambrose Slone, Kedar Soparkar, Disha Shrivastava, James Cobon-Kerr, Michael Sharman, Jay Pavagadhi, Carlos Araya, Karolis Misiunas, Nimesh Ghelani, Michael Laskin, David Barker, Qiujia Li, Anton Briukhov, Neil Houlsby, Mia Glaese, Balaji Lakshminarayanan, Nathan Schucher, Yunhao Tang, Eli Collins, Hyeontaek Lim, Fangxiaoyu Feng, Adria Recasens, Guangda Lai, Alberto Magni, Nicola De Cao, Aditya Siddhant, Zoe Ashwood, Jordi Orbay, Mostafa Dehghani, Jenny Brennan, Yifan He, Kelvin Xu, Yang Gao, Carl Saroufim, James Molloy, Xinyi Wu, Seb Arnold, Solomon Chang, Julian Schrittwieser, Elena Buchatskaya, Soroush Radpour, Martin Polacek, Skye Giordano, Ankur Bapna, Simon Tokumine, Vincent Hellendoorn, Thibault Sottiaux, Sarah Cogan, Aliaksei Severyn, Mohammad Saleh, Shantanu Thakoor, Laurent Shefey, Siyuan Qiao, Meenu Gaba, Shuo yiin Chang, Craig Swanson, Biao Zhang, Benjamin Lee, Paul Kishan Rubenstein, Gan Song, Tom Kwiatkowski, Anna Koop, Ajay Kannan, David Kao, Parker Schuh, Axel Stjerngren, Golnaz Ghiasi, Gena Gibson, Luke Vilnis, Ye Yuan, Felipe Tiengo Ferreira, Aishwarya Kamath, Ted Klimenko, Ken Franko, Kefan Xiao, Indro Bhattacharya, Miteyan Patel, Rui Wang, Alex Morris, Robin Strudel, Vivek Sharma, Peter Choy, Sayed Hadi Hashemi, Jessica Landon, Mara Finkelstein, Priya Jhakra, Justin Frye, Megan Barnes, Matthew Mauger, Dennis Daun, Khuslen Baatarsukh, Matthew Tung, Wael Farhan, Henryk Michalewski, Fabio Viola, Felix de Chaumont Quitry, Charline Le Lan, Tom Hudson, Qingze Wang, Felix Fischer, Ivy Zheng, Elspeth White, Anca Dragan, Jean baptiste Alayrac, Eric Ni, Alexander Pritzel, Adam Iwanicki, Michael Isard, Anna Bulanova, Lukas Zilka, Ethan Dyer, Devendra Sachan, Srivatsan Srinivasan, Hannah Muckenhirn, Honglong Cai, Amol Mandhane, Mukarram Tariq, Jack W. Rae, Gary Wang, Kareem Ayoub, Nicholas FitzGerald, Yao Zhao, Woohyun Han, Chris Alberti, Dan Garrette, Kashyap Krishnakumar, Mai Gimenez, Anselm Levskaya, Daniel Sohn, Josip Matak, Inaki Iturrate, Michael B. Chang, Jackie Xiang, Yuan Cao, Nishant Ranka, Geoff Brown, Adrian Hutter, Vahab Mirrokni, Nanxin Chen, Kaisheng Yao, Zoltan Egyed, Francois Galilee, Tyler Liechty, Praveen Kallakuri, Evan Palmer, Sanjay Ghemawat, Jasmine Liu, David Tao, Chloe Thornton, Tim Green, Mimi Jasarevic, Sharon Lin, Victor Cotruta, Yi-Xuan Tan, Noah Fiedel, Hongkun Yu, Ed Chi, Alexander Neitz, Jens Heitkaemper, Anu Sinha, Denny Zhou, Yi Sun, Charbel Kaed, Brice Hulse, Swaroop Mishra, Maria Georgaki, Sneha Kudugunta, Clement Farabet, Izhak Shafran, Daniel Vlasic, Anton Tsitsulin, Rajagopal Ananthanarayanan, Alen Carin, Guolong Su, Pei Sun, Shashank V, Gabriel Carvajal, Josef Broder, Iulia Comsa, Alena Repina, William Wong, Warren Weilun Chen, Peter Hawkins, Egor Filonov, Lucia Loher, Christoph Hirnschall, Weiyi Wang, Jingchen Ye, Andrea Burns, Hardie Cate, Diana Gage Wright, Federico Piccinini, Lei Zhang, Chu-Cheng Lin, Ionel Gog, Yana Kulizhskaya, Ashwin Sreevatsa, Shuang Song, Luis C. Cobo, Anand Iyer, Chetan Tekur, Guillermo Garrido, Zhuyun Xiao, Rupert Kemp, Huaixiu Steven Zheng, Hui Li, Ananth Agarwal, Christel Ngani, Kati Goshvadi, Rebeca Santamaria-Fernandez, Wojciech Fica, Xinyun Chen, Chris Gorgolewski, Sean Sun, Roopal Garg, Xinyu Ye, S. M. Ali Eslami, Nan Hua, Jon Simon, Pratik Joshi, Yelin Kim, Ian Tenney, Sahitya Potluri, Lam Nguyen Thiet, Quan Yuan, Florian Luisier, Alexandra Chronopoulou, Salvatore Scellato, Praveen Srinivasan, Minmin Chen, Vinod Koverkathu, Valentin Dalibard, Yaming Xu, Brennan Saeta, Keith Anderson, Thibault Sellam, Nick Fernando, Fantine Huot, Junehyuk Jung, Mani Varadarajan, Michael Quinn, Amit Raul, Maigo Le, Ruslan Habalov, Jon Clark, Komal Jalan, Kalesha Bullard, Achintya Singhal, Thang Luong, Boyu Wang, Sujeevan Rajayogam, Julian Eisenschlos, Johnson Jia, Daniel Finchelstein, Alex Yakubovich, Daniel Balle, Michael Fink, Sameer Agarwal, Jing Li, Dj Dvijotham, Shalini Pal, Kai Kang, Jaclyn Konzelmann, Jennifer Beattie, Olivier Dousse, Diane Wu, Remi Crocker, Chen Elkind, Siddhartha Reddy Jonnalagadda, Jong Lee, Dan Holtmann-Rice, Krystal Kallarackal, Rosanne Liu, Denis Vnukov, Neera Vats, Luca Invernizzi, Mohsen Jafari, Huanjie Zhou, Lilly Taylor, Jennifer Prendki, Marcus Wu, Tom Eccles, Tianqi Liu, Kavya Kopparapu, Francoise Beaufays, Christof Angermueller, Andreea Marzoca, Shourya Sarcar, Hilal Dib, Jeff Stanway, Frank Perbet, Nejc Trdin, Rachel Sterneck, Andrey Khorlin, Dinghua Li, Xihui Wu, Sonam Goenka, David Madras, Sasha Goldshtein, Willi Gierke, Tong Zhou, Yaxin Liu, Yannie Liang, Anais White, Yunjie Li, Shreya Singh, Sanaz Bahargam, Mark Epstein, Sujoy Basu, Li Lao, Adnan Ozturel, Carl Crous, Alex Zhai, Han Lu, Zora Tung, Neeraj Gaur, Alanna Walton, Lucas Dixon, Ming Zhang, Amir Globerson, Grant Uy, Andrew Bolt, Olivia Wiles, Milad Nasr, Ilia Shumailov, Marco Selvi, Francesco Piccinno, Ricardo Aguilar, Sara McCarthy, Misha Khalman, Mrinal Shukla, Vlado Galic, John Carpenter, Kevin Villela, Haibin Zhang, Harry Richardson, James Martens, Matko Bosnjak, Shreyas Rammohan Belle, Jeff Seibert, Mahmoud Alnahlawi, Brian McWilliams, Sankalp Singh, Annie Louis, Wen Ding, Dan Popovici, Lenin Simicich, Laura Knight, Pulkit Mehta, Nishesh Gupta, Chongyang Shi, Saaber Fatehi, Jovana Mitrovic, Alex Grills, Joseph Pagadora, Tsendsuren Munkhdalai, Dessie Petrova, Danielle Eisenbud, Zhishuai Zhang, Damion Yates, Bhavishya Mittal, Nilesh Tripuraneni, Yannis Assael, Thomas Brovelli, Prateek Jain, Mihajlo Velimirovic, Canfer Akbulut, Jiaqi Mu, Wolfgang Macherey, Ravin Kumar, Jun Xu, Haroon Qureshi,

Gheorghe Comanici, Jeremy Wiesner, Zhitao Gong, Anton Ruddock, Matthias Bauer, Nick Felt, Anirudh GP, Anurag Arnab, Dustin Zelle, Jonas Rothfuss, Bill Rosgen, Ashish Shenoy, Bryan Seybold, Xinjian Li, Jayaram Mudigonda, Goker Erdogan, Jiawei Xia, Jiri Simsa, Andrea Michi, Yi Yao, Christopher Yew, Steven Kan, Isaac Caswell, Carey Radebaugh, Andre Elisseeff, Pedro Valenzuela, Kay McKinney, Kim Paterson, Albert Cui, Eri Latorre-Chimoto, Solomon Kim, William Zeng, Ken Durden, Priya Ponnapalli, Tiberiu Sosea, Christopher A. Choquette-Choo, James Manyika, Brona Robenek, Harsha Vashisht, Sebastien Pereira, Hoi Lam, Marko Velic, Denese Owusu-Afriyie, Katherine Lee, Tolga Bolukbasi, Alicia Parrish, Shawn Lu, Jane Park, Balaji Venkatraman, Alice Talbert, Lambert Rosique, Yuchung Cheng, Andrei Sozanschi, Adam Paszke, Praveen Kumar, Jessica Austin, Lu Li, Khalid Salama, Bartek Perz, Wooyeol Kim, Nandita Dukkipati, Anthony Baryshnikov, Christos Kaplanis, XiangHai Sheng, Yuri Chervonyi, Caglar Unlu, Diego de Las Casas, Harry Askham, Kathryn Tunyasuvunakool, Felix Gimeno, Siim Poder, Chester Kwak, Matt Miecnikowski, Vahab Mirrokni, Alek Dimitriev, Aaron Parisi, Dangyi Liu, Tomy Tsai, Toby Shevlane, Christina Kouridi, Drew Garmon, Adrian Goedeckemeyer, Adam R. Brown, Anitha Vijayakumar, Ali Elqursh, Sadegh Jazayeri, Jin Huang, Sara Mc Carthy, Jay Hoover, Lucy Kim, Sandeep Kumar, Wei Chen, Courtney Biles, Garrett Bingham, Evan Rosen, Lisa Wang, Qijun Tan, David Engel, Francesco Pongetti, Dario de Cesare, Dongseong Hwang, Lily Yu, Jennifer Pullman, Srini Narayanan, Kyle Levin, Siddharth Gopal, Megan Li, Asaf Aharoni, Trieu Trinh, Jessica Lo, Norman Casagrande, Roopali Vij, Loic Matthey, Bramandia Ramadhana, Austin Matthews, CJ Carey, Matthew Johnson, Kremena Goranova, Rohin Shah, Shereen Ashraf, Kingshuk Dasgupta, Rasmus Larsen, Yicheng Wang, Manish Reddy Vuyyuru, Chong Jiang, Joana Ijazi, Kazuki Osawa, Celine Smith, Ramya Sree Boppana, Taylan Bilal, Yuma Koizumi, Ying Xu, Yasemin Altun, Nir Shabat, Ben Bariach, Alex Korchemniy, Kiam Choo, Olaf Ronneberger, Chimezie Iwuanyanwu, Shubin Zhao, David Soergel, Cho-Jui Hsieh, Irene Cai, Shariq Iqbal, Martin Sundermeyer, Zhe Chen, Elie Bursztein, Chaitanya Malaviya, Fadi Biadsy, Prakash Shroff, Inderjit Dhillon, Tejasi Latkar, Chris Dyer, Hannah Forbes, Massimo Nicosia, Vitaly Nikolaev, Somer Greene, Marin Georgiev, Pidong Wang, Nina Martin, Hanie Sedghi, John Zhang, Praseem Banzal, Doug Fritz, Vikram Rao, Xuezhi Wang, Jiageng Zhang, Viorica Patraucean, Dayou Du, Igor Mordatch, Ivan Jurin, Lewis Liu, Ayush Dubey, Abhi Mohan, Janek Nowakowski, Vlad-Doru Ion, Nan Wei, Reiko Tojo, Maria Abi Raad, Drew A. Hudson, Vaishakh Keshava, Shubham Agrawal, Kevin Ramirez, Zhichun Wu, Hoang Nguyen, Ji Liu, Madhavi Sewak, Bryce Petrini, DongHyun Choi, Ivan Philips, Ziyue Wang, Ioana Bica, Ankush Garg, Jarek Wilkiewicz, Priyanka Agrawal, Xiaowei Li, Danhao Guo, Emily Xue, Naseer Shaik, Andrew Leach, Sadh MNM Khan, Julia Wiesinger, Sammy Jerome, Abhishek Chakladar, Alek Wenjiao Wang, Tina Ornduff, Folake Abu, Alireza Ghaffarkhah, Marcus Wainwright, Mario Cortes, Frederick Liu, Joshua Maynez, Andreas Terzis, Pouya Samangouei, Riham Mansour, Tomasz Kępa, François-Xavier Aubet, Anton Algymr, Dan Banica, Agoston Weisz, Andras Orban, Alexandre Senges, Ewa Andrejczuk, Mark Geller, Niccolo Dal Santo, Valentin Anklin, Majd Al Merey, Martin Baeuml, Trevor Strohman, Junwen Bai, Slav Petrov, Yonghui Wu, Demis Hassabis, Koray Kavukcuoglu, Jeff Dean, and Oriol Vinyals. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.

Thompson, Brian, Nitika Mathur, Daniel Deutsch, and Huda Khayrallah. 2024. Improving statistical significance in human evaluation of automatic metrics via soft pairwise accuracy. In Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 1222–1234, Miami, Florida, USA, November. Association for Computational Linguistics.

Toral, Antonio, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the unattainable? reassessing claims of human parity in neural machine translation. In Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 113–123, Brussels, Belgium, October. Association for Computational Linguistics.

Vernikos, Giorgos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 118–128, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Wang, Jiayi, David Ifeoluwa Adelani, and Pontus Stenetorp. 2024. Evaluating WMT 2024 metrics shared task submissions on AfriMTE (the African challenge set). In Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 505–516, Miami, Florida, USA, November. Association for Computational Linguistics.

Wang, Yutong, Jiali Zeng, Xuebo Liu, Derek F. Wong, Fandong Meng, Jie Zhou, and Min Zhang. 2025. Delta: An online document-level translation agent based on multi-level memory.

Wicks, Rachel and Matt Post. 2022. Does sentence segmentation matter for machine translation? In Koehn, Philipp, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus

Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 843–854, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

Wicks, Rachel and Matt Post. 2023. Identifying context-dependent translations for evaluation set production. In Koehn, Philipp, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 452–467, Singapore, December. Association for Computational Linguistics.

Wilkens, Rodrigo, Bruno Oberle, Frédéric Landragin, and Amalia Todirascu. 2020. COFR: COreference resolution tool For FRench, February.

Wu, Minghao, Thuy-Trang Vu, Lizhen Qu, George Foster, and Gholamreza Haffari. 2024. Adapting large language models for document-level machine translation.

Yan, Jianhao, Pingchuan Yan, Yulong Chen, Judy Li, Xianchao Zhu, and Yue Zhang. 2024. Gpt-4 vs. human translators: A comprehensive evaluation of translation quality across languages, domains, and expertise levels.

Zhang, Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

Zouhar, Vilém, Shuoyang Ding, Anna Currey, Tatyana Badeka, Jenyuan Wang, and Brian Thompson. 2024. Fine-tuned machine translation metrics struggle in unseen domains. In Ku, Lun-Wei, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 488–500, Bangkok, Thailand, August. Association for Computational Linguistics.

A Appendix

A.1 Levenshtein Distance Ratio for Quality Filtering

To filter out noisy LLM outputs, we use the Levenshtein distance ratio, defined as:

$$r(s, s') = \frac{d_{\text{lev}}(s,\ s')}{\max(|s|,\ |s'|)}$$

where $d_{\text{lev}}(s, s')$ is the character-level Levenshtein distance between the original segment s and the perturbed segment s', and $|\cdot|$ denotes the number of characters. The ratio $r \in [0, 1]$ thus measures the proportion of characters that need to be changed relative to the longer of the two strings, making it invariant to segment length. A low ratio indicates a minimal, targeted edit; a high ratio indicates heavy rewriting or hallucination.

We apply a perturbation-specific threshold τ and discard any sample for which $r(s, s') > \tau$. The thresholds were chosen arbitrarily: $\tau = 0.20$ for perturbations expected to affect only a small number of tokens (lexical consistency, gender of anaphoric pronouns, conjunction substitution), and $\tau = 0.30$ for tense modifications, which can alter inflectional morphology across multiple verbs within the same segment. These values reflect a pragmatic, unsystematic choice with no formal optimization: loose enough to retain legitimate minimal edits, yet strict enough to discard obvious hallucinations such as full paraphrases or outputs where the LLM added explanatory content.

A.2 French Conjunctions

We list below the French coordinating and subordinating conjunctions used by the CONJ perturbation. Conjunctions are grouped by discourse function; substitutions are sampled uniformly at random within each group.

- **Coordinating Conjunctions:** Coordinating conjunctions link words, phrases, or clauses of equal syntactic importance, creating compound structures within a sentence. They serve to indicate relationships such as addition, contrast, choice, and consequence. In French, the primary coordinating conjunctions are: *et* (and), *ou* (or), *mais* (but), *donc* (so, therefore), *ni* (neither/nor), *car* (for/because), and *or* (yet).

- **Subordinating Conjunctions:** A subordinating conjunction links a subordinate clause to a main clause, establishing a relationship of dependency between the two. These conjunctions are essential for structuring complex sentences and expressing relationships such as cause, purpose, condition, time, and contrast. In French, common subordinating conjunctions include:

  - **Cause and Consequence:** *À un tel point que, À tel point que, À ce que, Au point que, Comme quoi, De telle manière que, De telle sorte que, De manière que, De*

sorte que, De façon que, En sorte que, Si bien que, Tellement que, À ce que

- **Purpose:** *À seule fin que, Afin que, De telle manière que, De telle sorte que, De manière que, De sorte que, De façon que, De crainte que, De peur que, Pour que*
- **Condition:** *À condition que, À moins que, Autant que, Au cas où, Dans la mesure où, Du moment que, En cas que, Pour peu que, Pourvu que, Selon que, Suivant que, Si*
- **Comparison and Manner:** *À mesure que, Ainsi que, Au fur et à mesure que, Aussi bien que, Comme si, De telle manière que, De telle sorte que, De manière que, De sorte que, De façon que, De même que, En sorte que, Sans que, Selon que*
- **Time:** *Alors que, Cependant que, Après que, Aussitôt que, Avant que, Cependant que, Comme, Dès lors que, Dès que, D'ici que, Du moment que, Durant que, En attendant que, Lors même que, Pendant que, Sitôt que, Tandis que, Tant que, Une fois que, Quand, Lorsque*
- **Cause and Justification:** *Comme, D'autant plus que, D'autant que, Du fait que, Étant donné que, Parce que, Puisque, Vu que*
- **Concession and Opposition:** *Bien que, Encore que, Loin que, Même si, Quand bien même que, Quoique, Qui que, Si ce n'est que, Où que, Mis à part le fait que, Sauf que*

A.3 Spanish Conjunctions

We list below the Spanish coordinating and subordinating conjunctions used by the CONJ perturbation, grouped by discourse function. Substitutions are sampled uniformly at random within each group.

- **Coordinating Conjunctions:** Coordinating conjunctions in Spanish link elements of equal syntactic status, such as words, phrases, or clauses. They express relationships like addition, contrast, or consequence. Common coordinating conjunctions include: *y* (and), *o* (or), *pero* (but), *ni* (neither/nor), *pues* (so), *sin embargo* (however), *sino* (but rather), *aunque* (although).
- **Subordinating Conjunctions:** Subordinating conjunctions introduce subordinate clauses and express logical relations such as cause, purpose, condition, time, and contrast. In Spanish, common subordinating conjunctions include:
  - **Cause and Consequence:** *Hasta tal punto que, A tal punto que, A lo que, Al punto que, Como para que, De tal manera que, De tal forma que, De manera que, De modo que, De forma que, De suerte que, Tan bien que, Tanto que*
  - **Purpose:** *Con el único fin de que, A fin de que, Para que, Por miedo a que, Por temor a que*
  - **Condition:** *A condición de que, A menos que, En tanto que, En caso de que, En la medida en que, Desde que, En caso que, Con tal de que, Siempre que*
  - **Comparison and Manner:** *Según, Conforme, A medida que, Así como, A la vez que, Tan bien como, Como si, Igual que, Sin que*
  - **Time:** *Después de que, Tan pronto como, Antes de que, Desde entonces que, En cuanto, De aquí a que, Durante que, Mientras tanto que, Apenas que, Mientras, Una vez que, Cuando*
  - **Cause and Justification:** *Como, Cuanto más que, Cuanto que, Debido a que, Dado que, Porque, Ya que, Visto que*
  - **Concession and Opposition:** *Aunque, Aun cuando, Lejos de que, Incluso si, A pesar de que, Quienquiera que, Salvo que, Dondequiera que, Aparte del hecho de que, Excepto que, Hasta que*

A.4 German Conjunctions

We list below the German coordinating and subordinating conjunctions used by the CONJ perturbation, grouped by discourse function. Substitutions are sampled uniformly at random within each group.

- **Coordinating Conjunctions:** Coordinating conjunctions in German link clauses or elements of equal rank, and are not followed by a change in word order. Common examples include: *und* (and), *oder* (or), *aber* (but), *denn* (because/for), *sondern* (but rather), *doch* (however), *jedoch* (yet/still).

- **Subordinating Conjunctions:** These introduce dependent clauses and affect the word order of the clause (verb to the end). They express relationships such as cause, purpose, condition, time, or contrast. Examples include:
  - **Cause and Consequence:** *Sodass, So sehr, dass, In dem Maße, dass, Insofern, Derart, dass, Infolgedessen, Da, Weil, Denn, Zumal, Angesichts der Tatsache, dass*
  - **Purpose:** *Damit, Um zu, Zu dem Zweck, dass, Mit dem Ziel, dass, Mit der Absicht, dass, Aus Angst, dass, Aus Furcht, dass*
  - **Condition:** *Unter der Bedingung, dass, Es sei denn, dass, Soviel, Falls, Insoweit, Für den Fall, dass, Vorausgesetzt, dass, Sofern, Je nachdem, ob*
  - **Comparison and Manner:** *Während, Gleichzeitig wie, Ebenso wie, Als ob, In einer Weise, dass, Auf eine Weise, dass, Ohne dass*
  - **Time:** *Nachdem, Sobald, Bevor, Während, Solange, Kaum dass, Wenn, Als, Immer wenn, Seitdem, Bis*
  - **Cause and Justification:** *Da, Weil, Denn, Zumal, Angesichts der Tatsache, dass*
  - **Concession and Opposition:** *Obwohl, Auch wenn, Weit davon entfernt, dass, Selbst wenn, Obgleich, Wer auch immer, Außer dass, Wo auch immer, Abgesehen davon, dass, Es sei denn, Bis, Wohingegen*

A.5 Results

Each figure shows results for one system and one language pair, with GEMINI and AYA23 shown separately. The commentary below highlights patterns that are specific to each condition and complement the aggregated analysis in Section 6.

A.5.1 English–French

English–French analysis. For tense consistency, the COMET/COMETKIWI mirror pattern is most pronounced in en–fr. COMETKIWI starts at 30% for AYA23 (Figure 2) and at 20% for GEMINI (Figure 3) at w=1, then climbs to 60% and 52% respectively at w=9. This is the sharpest COMETKIWI ascent across all language pairs, consistent with French offering a rich tense system (*passé composé*, *imparfait*, *passé simple*) that creates more detectable inconsistencies when mixed. GEMBA-MQM shows its most severe collapse on conjunction substitution for en–fr: it drops from 38% (AYA23) and 36% (GEMINI) at w=1 to 40% and 35% at w=9, remaining consistently below chance. On sentence repetition (Figure 2), METRICX-24-QE exhibits the sharpest reversal for AYA23: from 40% at w=1 to 81% at w=3, confirming that this metric initially prefers duplicated content when seen in isolation. The sentence splitting false-positive effect is visible for both systems: COMET increases from 73%/85% at w=1 to 91%/100% at w=9, increasingly penalizing quality-preserving splits.

A.5.2 English–Spanish

English–Spanish analysis. A notable difference from the aggregated results is the behavior of METRICX-24 on tense consistency for AYA23 (Figure 4): it achieves 86% at w=1 and remains above 82% across all window sizes. This is substantially higher than for en–fr or en–de, suggesting that the reference-based signal is particularly strong for Spanish tense perturbations in AYA23 outputs. This contrasts with GEMINI (Figure 5), where METRICX-24 starts at 70% and shows moderate context sensitivity (65–72%). COMETKIWI on tense consistency follows the expected climbing pattern but with a weaker signal than en–fr: from 40% to 54% for AYA23, and from 31% to only 40% for GEMINI. On lexical consistency (Figure 4), AYA23 shows systematically higher CHRF accuracy (99%–96%) than the other language pairs, indicating stronger surface overlap between original AYA23 outputs and the reference for en–es. GEMBA-MQM on conjunction substitution collapses similarly to en–fr, reaching 41% for AYA23 and 42% for GEMINI at w=9.

A.5.3 English–German

English–German analysis. English–German reveals the strongest context sensitivity for METRICX-24 and METRICX-24-QE on tense consistency. For AYA23 (Figure 6), METRICX-24 rises from 52% at w=1 to 73% at w=9; for GEMINI (Figure 7), from 49% to 71%. METRICX-24-QE follows a parallel trajectory (from 38%/37% to 72%/62%). These are the largest context-driven gains observed across all conditions, suggesting that German tense perturbations pro-

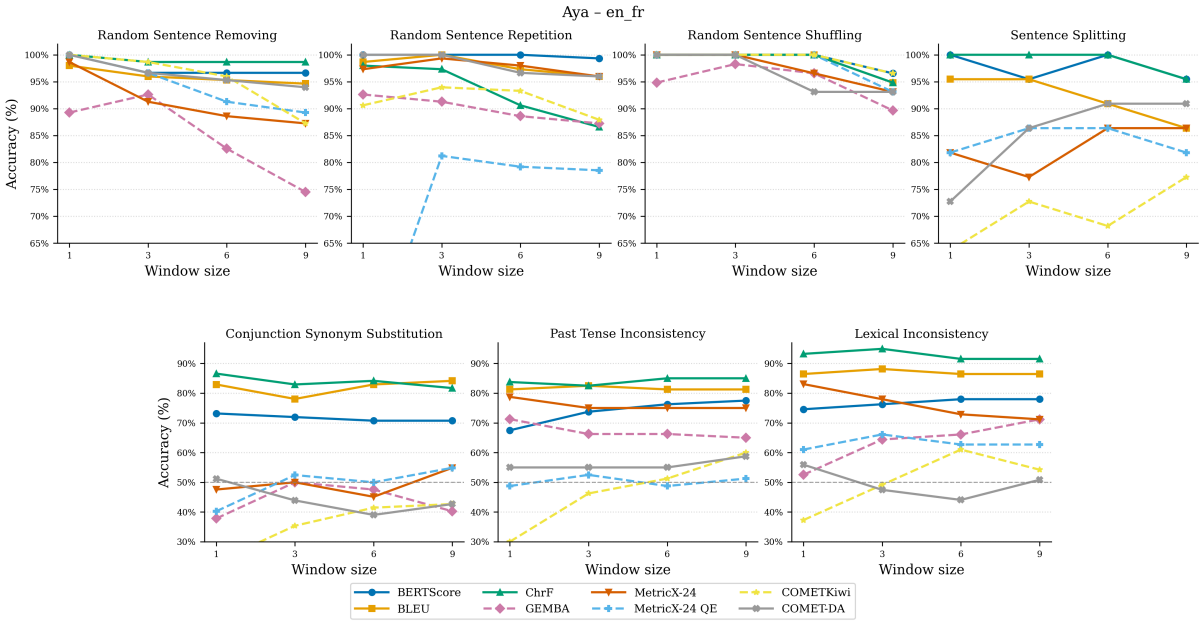

Figure 2: Average document-level accuracy (%) for each evaluation metric, computed for varying window sizes for SLIDE(w,1) with $w \in \{1, 3, 6, 9\}$ for the AYA23 system and the en–fr language pair.

duce a stronger signal for these metrics, possibly because the German tense system (*Präteritum* vs. *Perfekt*) interacts with word order in ways that accumulate more detectable patterns across sentences. On conjunction substitution, COMETKIWI shows weaker sensitivity for en–de than for the Romance languages: it starts at 46%/44% at w=1 and barely reaches 50%/49% at w=9, suggesting that conjunction perturbations are harder to detect in German, where conjunctions can affect clause structure (verb-final order in subordinate clauses). On lexical consistency (Figure 7), COMET shows the sharpest decline for en–de: from 56% (AYA23) and 67% (GEMINI) at w=1 down to 50% and 47% at w=9, illustrating how reference access becomes increasingly uninformative with context for this language pair. METRICX-24-QE on sentence repetition (Figure 6) starts lower for en–de (57%/59% at w=1) compared to other pairs, and the recovery at w=3 is less sharp (82%/86%), indicating that the initial preference for duplicated content is weaker for German.

A.5.4 Metric Behaviour on Unperturbed Outputs Across Window Sizes

In addition to the per-pair perturbation analyses above, we examine the average scores of unperturbed MT outputs under different sliding-window settings. Table 3 shows that mean scores for AYA23 shift noticeably as w grows, even though the translations are identical.

Metric	Δ 1→3	Δ 3→6	Δ 6→9
BLEU	3.19	3.31	0.53
CHRF	3.07	2.69	1.05
BERTSCORE	-1.23	0.00	0.00
METRICX-24	60.09	33.78	15.63
METRICX-24-QE	56.80	28.57	12.70
GEMBA-MQM	-17.92	22.07	6.15
COMET	-4.76	-1.22	0.00
COMETKIWI	-1.20	-1.22	-2.47

Table 3: Variation (%) in system-level scores of the original **AYA-23** system for the **English–French (en–fr)** pair, when moving across successive context windows: from SLIDE(1, 1) to SLIDE(3, 1) (Δ 1→3), then from SLIDE(3, 1) to SLIDE(6, 1) (Δ 3→6), and from SLIDE(6, 1) to SLIDE(9, 1) (Δ 6→9).

Raw scores are not directly comparable across window sizes. In theory, a metric should give the same score to the same output regardless of segmentation. In practice, score drift occurs with longer windows, confirming that absolute values across w cannot be compared directly. The effect is particularly pronounced for METRICX-24, METRICX-24-QE and GEMBA-MQM, while it is almost absent for BERTSCORE.

A natural concern is whether metrics become inherently less reliable as the sliding window grows, independently of any perturbation effect. If increasing context size degrades metric quality on clean translations, then the apparent drop in perturbation-detection accuracy at larger window sizes (Figure 8) could simply reflect this loss of reliability rather

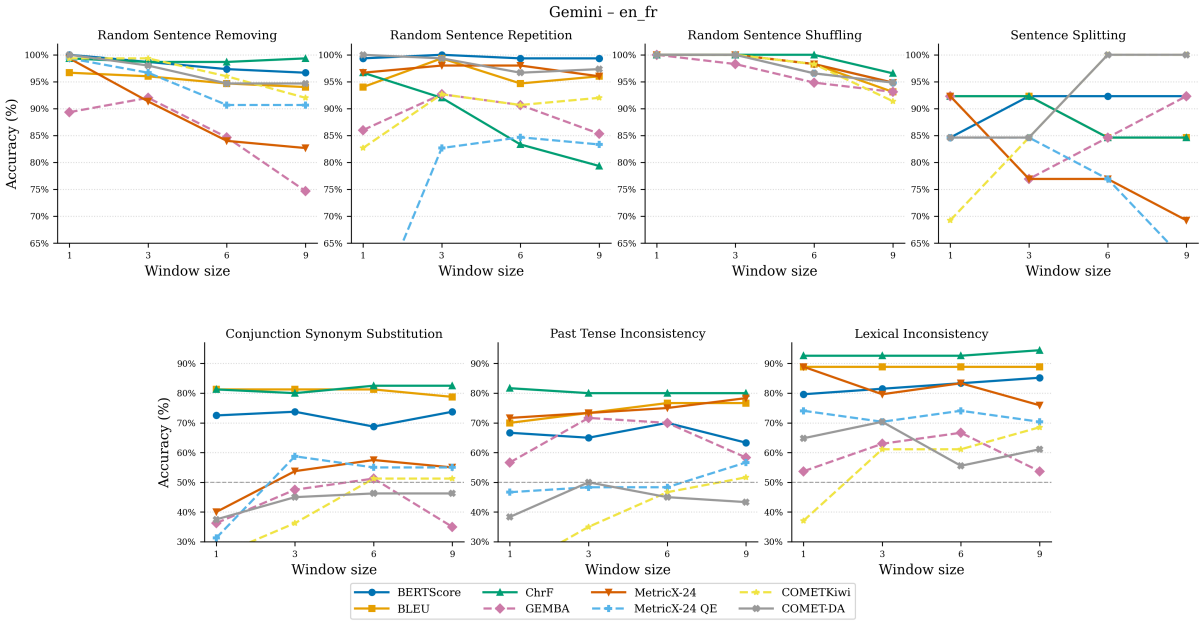

Figure 3: Average document-level accuracy (%) for each evaluation metric, computed for varying window sizes for SLIDE(w,1) with $w \in \{1, 3, 6, 9\}$ for the GEMINI system and the en–fr language pair.

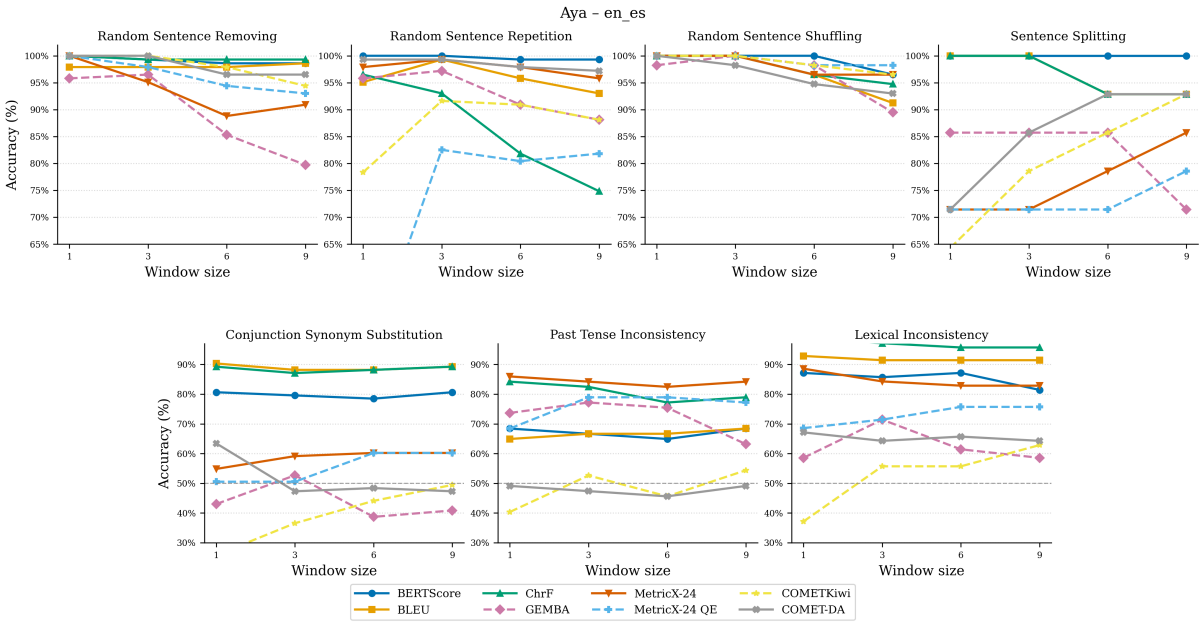

Figure 4: Average document-level accuracy (%) for each evaluation metric, computed for varying window sizes for SLIDE(w,1) with $w \in \{1, 3, 6, 9\}$ for the AYA23 system and the en–es language pair.

than a genuine evaluation signal.

To test this, we score the same unperturbed translated documents with different window sizes ($w \in \{1, 3, 6, 9\}$) and examine how scores and their distributions shift. Table 4 reports the mean metric scores per window size, aggregated over all systems and language pairs.

Figure 8 shows the full score distributions. Most metrics exhibit only a mild, systematic shift: the distributions move as a block without becoming wider or more dispersed, preserving the relative ordering of documents. BERTSCORE is nearly invariant, while COMET and COMETKIWI drift only slightly downward.

The notable exception is METRICX-24 and METRICX-24-QE , which show a markedly stronger upward drift. This is consistent with their training objective: both models are fine-tuned to regress on MQM error counts, meaning they learn to accumulate error signals over the input span. As

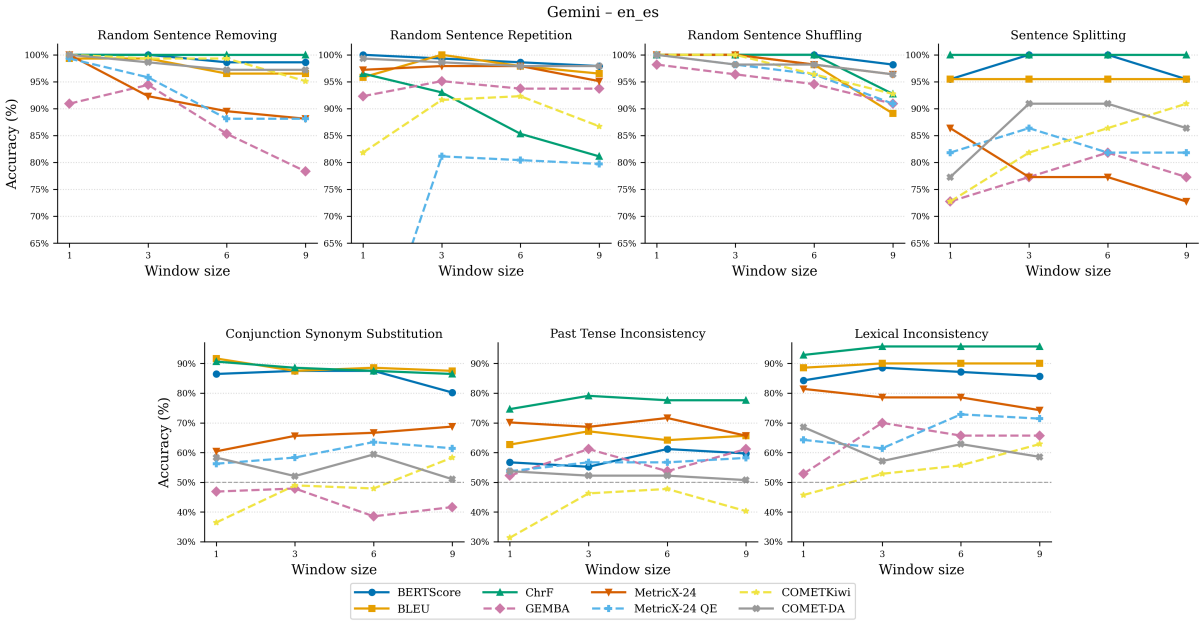

Figure 5: Average document-level accuracy (%) for each evaluation metric, computed for varying window sizes for SLIDE(w,1) with $w \in \{1, 3, 6, 9\}$ for the GEMINI system and the en–es language pair.

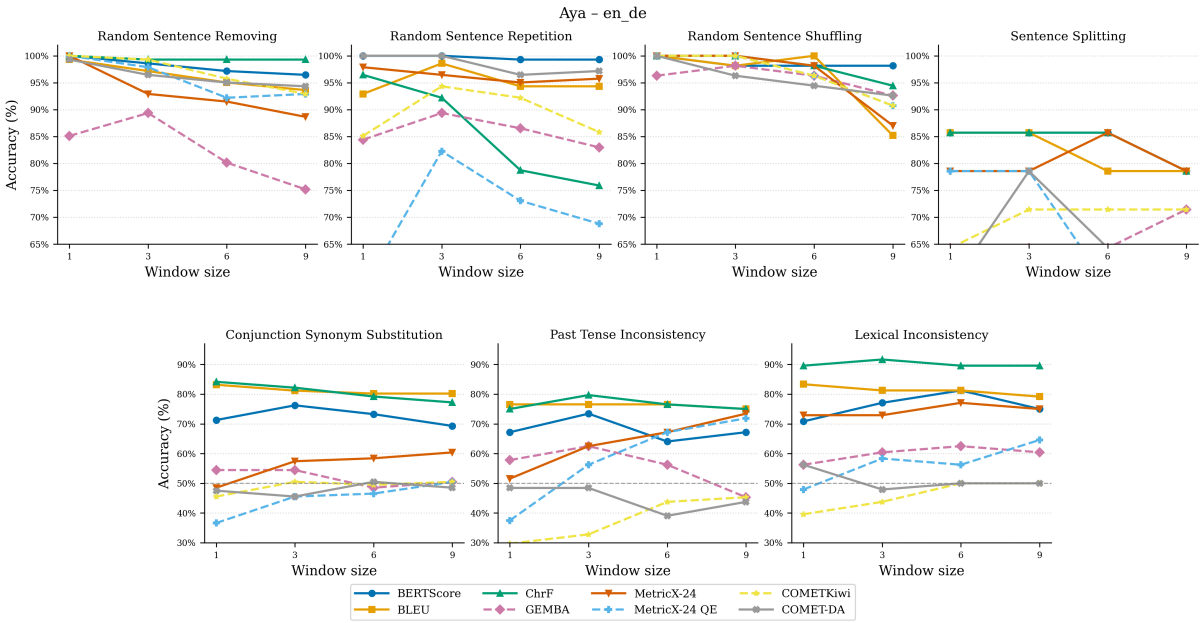

Figure 6: Average document-level accuracy (%) for each evaluation metric, computed for varying window sizes for SLIDE(w,1) with $w \in \{1, 3, 6, 9\}$ for the AYA23 system and the en–de language pair.

the sliding window grows, the input presented to the model becomes longer, and the probability of matching error-like patterns increases mechanically, even when the underlying translation quality is unchanged. In other words, these metrics conflate input length with error severity. While this property may be desirable in a setting where longer spans genuinely expose more errors (e.g., document-level evaluation with a fixed segmentation), it becomes problematic when comparing the *same* output under different segmentation strategies, as the score difference reflects context size rather than quality.

These results indicate that, for most metrics, larger windows introduce a *consistent offset* rather than random noise, and system rankings on unperturbed outputs remain stable. The accuracy drop observed in perturbation detection at larger w is therefore not attributable to metrics becoming unreliable, but rather to the perturbation signal being diluted by the additional unperturbed context included in

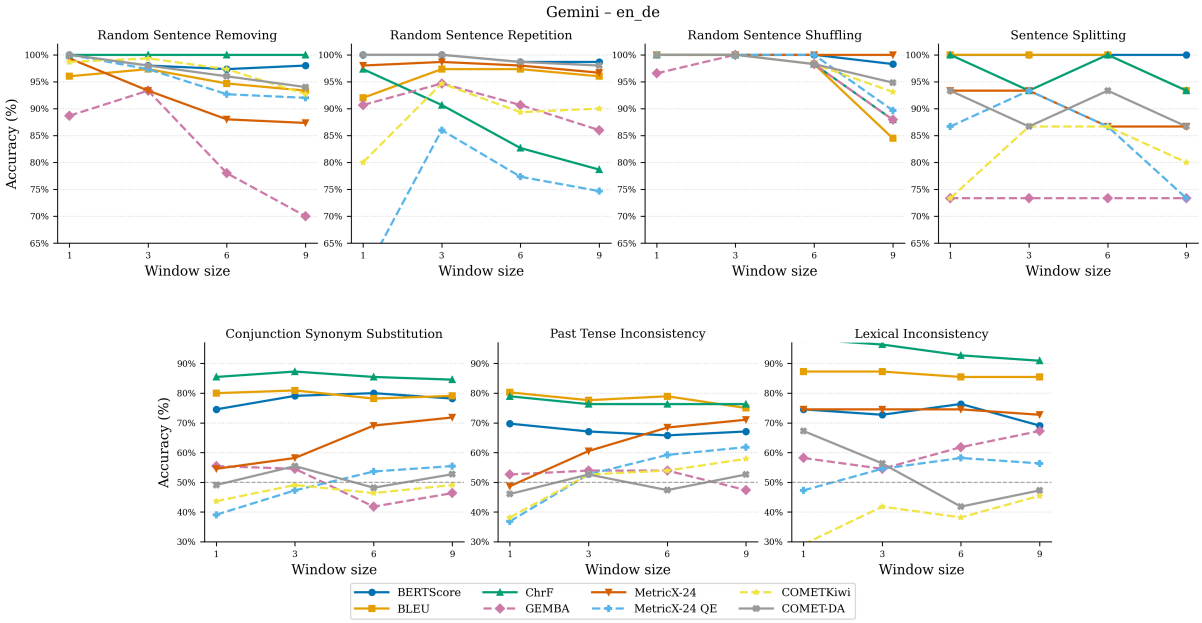

Figure 7: Average document-level accuracy (%) for each evaluation metric, computed for varying window sizes for SLIDE(w,1) with $w \in \{1, 3, 6, 9\}$ for the GEMINI system and the en–de language pair.

Metric	w=1	w=3	w=6	w=9
BLEU	36.40	37.80	38.79	39.03
CHRF	62.08	63.64	65.05	65.61
BERTSCORE	0.83	0.82	0.82	0.82
METRICX-24	1.86	3.03	4.17	4.85
METRICX-24-QE	2.05	3.30	4.47	5.10
GEMBA-MQM	-5.66	-4.50	-5.00	-5.42
COMET	0.85	0.82	0.81	0.82
COMETKIWI	0.82	0.81	0.80	0.78

Table 4: Mean metric scores per window size, aggregated over all systems and language pairs.

each window. For METRICX-24 and METRICX-24-QE , however, the length-dependent drift adds an additional confound that should be accounted for when interpreting cross-window comparisons.

A.6 Manual Evaluation of LLM-based Perturbations

We manually evaluate the four LLM-based perturbation types (tense consistency, lexical consistency, conjunction substitution, and gender of anaphoric pronouns (PRO)) to verify that they preserve sentence-level acceptability and that the reported metric behaviors reflect genuine document-level effects rather than sentence-level artifacts. Structural perturbations are deterministic and verifiable by construction, so they are not part of this manual evaluation. The PRO perturbation is reported here for completeness even though it is excluded from the main evaluation due to insufficient instances. For each perturbation type, an annotator was shown the English source (*SRC*) and two anonymous candidate translations, the original MT output (*SYS*) and its LLM-perturbed version (*PERT*), presented in random order, and judged which candidates were acceptable translations of the source.

Tense consistency. The tense consistency perturbation modifies past-tense verb forms in the target language by switching between available past tenses (*passé composé*, *imparfait*, and *passé simple* in French), while leaving all other words unchanged (see Section 3.2). By design, each perturbed sentence remains grammatical in isolation; the incoherence only emerges at the discourse level through the random alternation of tense forms across the document.

We conducted the manual evaluation on a random sample of 31 perturbed sentences (en→fr; 13 AYA23 and 18 GEMINI). Both translations are judged valid in 77% of cases overall, with a marked system asymmetry (94% for GEMINI versus 54% for AYA23), suggesting that for AYA23 the perturbation may conflate sentence-level and discourse-level effects. The full breakdown is reported in Table 5.

Lexical consistency. The lexical consistency perturbation replaces repeated content words in the target language by random synonyms aligned with the corresponding source tokens (Section 3.2). Be-

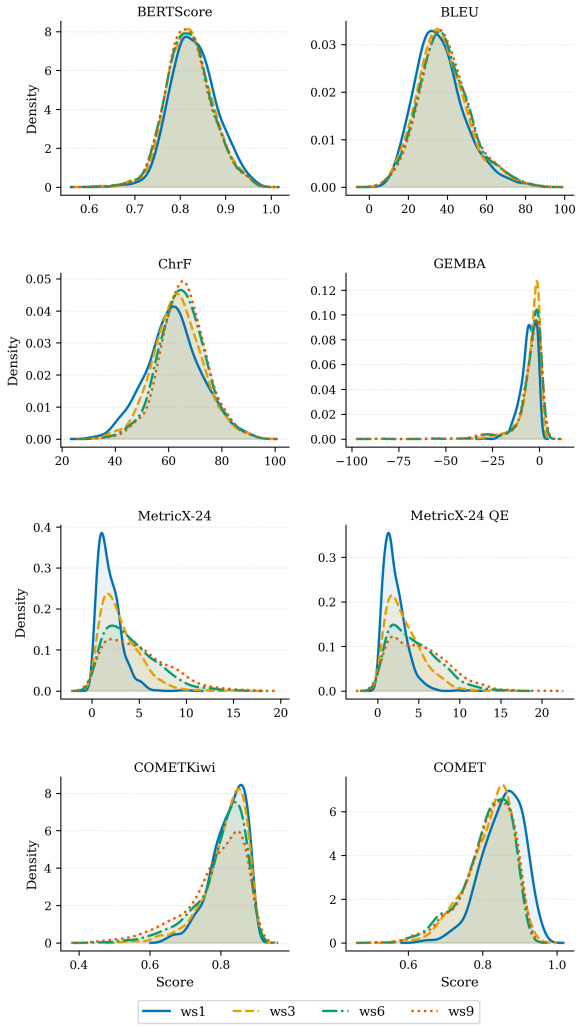

Figure 8: Score distributions across sliding window sizes ($w \in \{1, 3, 6, 9\}$) for the same unperturbed translated documents, aggregated over all systems and language pairs. Distributions shift systematically but preserve their shape, indicating that larger windows introduce a consistent offset rather than increased unreliability.

cause each substitution is a single near-synonym replacement, the edit surface is much narrower than for tense, and the perturbation is expected to leave each sentence grammatical and locally adequate.

We conducted the same manual evaluation protocol as for tense on a random sample of 26 perturbed sentences (en→fr; 9 AYA23 and 17 GEMINI). Both translations are judged valid in 73% of cases overall, with SYS alone preferred in 19% of cases, comparable to tense consistency. The effect is slightly more pronounced for GEMINI (24% *Only SYS valid*) than for AYA23 (11%). The full breakdown is reported in Table 6.

Conjunction substitution. The conjunction substitution perturbation replaces conjunctions by near-synonyms within the same broad discourse class. The edit again touches a single token per substitution, and the perturbation is designed so that each sentence remains grammatical in isolation while the inter-sentence discourse relations are weakened or distorted across the document.

We conducted the same manual evaluation protocol as for tense on a random sample of 19 perturbed sentences (en→fr; 11 AYA23 and 8 GEMINI). Both translations are judged valid in 95% of cases, with a single AYA23 output for which only SYS was judged acceptable, confirming that the perturbation rarely affects sentence-level acceptability. The full breakdown is reported in Table 7.

Gender of anaphoric pronouns. The PRO perturbation flips singular (PRO-SG, e.g. *il*/*elle*) and plural (PRO-PL, *ils*/*elles*) anaphoric pronouns in the target translation, producing a sentence that remains grammatical in isolation but disagrees with the source antecedent across the discourse. Because the perturbation depends on the target morphology,

Outcome	Aya23		Gemini		Overall	
	n	%	*n*	%	*n*	%
Both valid	7	54%	17	94%	24	77%
Only *PERT* valid	0	0%	1	6%	1	3%
Only *SYS* valid	6	46%	0	0%	6	19%
Neither valid	0	0%	0	0%	0	0%
Total	13	100%	18	100%	31	100%

Table 5: Manual evaluation of LLM tense perturbations (en→fr). Given an English source segment (*SRC*), an annotator judged whether the original MT output (*SYS*) and its LLM-perturbed version (*PERT*) were each acceptable translations, presented as two anonymous candidates in random order. *Both valid*: both translations are acceptable (perturbation preserved quality). *Only PERT valid*: only the perturbed version is acceptable. *Only SYS valid*: perturbation degraded the translation.

Outcome	Aya23		Gemini		Overall	
	n	%	*n*	%	*n*	%
Both valid	7	78%	12	71%	19	73%
Only *PERT* valid	1	11%	0	0%	1	4%
Only *SYS* valid	1	11%	4	24%	5	19%
Neither valid	0	0%	1	6%	1	4%
Total	9	100%	17	100%	26	100%

Table 6: Manual evaluation of LLM lexical-consistency perturbations (en→fr). Given an English source segment (*SRC*), an annotator judged whether the original MT output (*SYS*) and its LLM-perturbed version (*PERT*) were each acceptable translations, presented as two anonymous candidates in random order. *Both valid*: both translations are acceptable (perturbation preserved quality). *Only PERT valid*: only the perturbed version is acceptable. *Only SYS valid*: perturbation degraded the translation.

Outcome	Aya23		Gemini		Overall	
	n	%	*n*	%	*n*	%
Both valid	10	91%	8	100%	18	95%
Only *PERT* valid	0	0%	0	0%	0	0%
Only *SYS* valid	1	9%	0	0%	1	5%
Neither valid	0	0%	0	0%	0	0%
Total	11	100%	8	100%	19	100%

Table 7: Manual evaluation of LLM conjunction perturbations (en→fr). Given an English source segment (*SRC*), an annotator judged whether the original MT output (*SYS*) and its LLM-perturbed version (*PERT*) were each acceptable translations, presented as two anonymous candidates in random order. *Both valid*: both translations are acceptable (perturbation preserved quality). *Only PERT valid*: only the perturbed version is acceptable. *Only SYS valid*: perturbation degraded the translation.

Outcome	Aya23		Gemini		Overall	
	n	%	*n*	%	*n*	%
Both valid	24	96%	15	100%	39	98%
Only *PERT* valid	0	0%	0	0%	0	0%
Only *SYS* valid	0	0%	0	0%	0	0%
Neither valid	1	4%	0	0%	1	2%
Total	25	100%	15	100%	40	100%

Table 8: Manual evaluation of LLM perturbations of the gender of anaphoric pronouns (PRO, en→fr). Given an English source segment (*SRC*), an annotator judged whether the original MT output (*SYS*) and its LLM-perturbed version (*PERT*) were each acceptable translations, presented as two anonymous candidates in random order. *Both valid*: both translations are acceptable (perturbation preserved quality). *Only PERT valid*: only the perturbed version is acceptable. *Only SYS valid*: perturbation degraded the translation.

it is essentially confined to en→fr in our data, with too few instances in en→de and en→es to support a reliable analysis (the reason it is excluded from the main evaluation).

We conducted the same manual evaluation protocol as for tense on the 40 perturbed sentences available for en→fr (25 AYA23 and 15 GEMINI). Both translations are judged valid in 98% of cases, with no instance where one candidate was clearly preferred over the other and a single AYA23 output rated as invalid in both versions. As expected from a perturbation editing a single closed-class token, sentence-level acceptability is essentially preserved. The full breakdown is reported in Table 8.

Summary. Across the four LLM-based perturbation types, the manual evaluation indicates that PRO (98% *Both valid*) and conjunction substitution (95%) preserve sentence-level acceptability almost universally. Tense consistency (77%) and lexical consistency (73%) show somewhat more sentence-level degradation, the lowest validity being observed for AYA23 tense outputs (54%). To examine whether sentence-level artifacts could drive our results, we inspect the per-system breakdown of the metric analysis in Appendix A.5. On the en→fr subset where the manual evaluation was conducted, GEMINI's sentence-level validity is uniformly high (71–100% across all four perturbations), and we observe the same patterns and conclusions as in the aggregated analysis (Section 6).

A.7 Perturbation Generation Prompts

We list below the prompts used to generate each LLM-based perturbation. Each block specifies the system message and the user instruction passed to `gpt-4o-mini` for one (perturbation, language) combination. Prompts were crafted to produce minimal edits while preserving sentence-level grammaticality, and to instruct the model to return `None` when

no valid edit was possible (this is part of the quality-control pipeline described in Section 3.3).

A.8 Computational Resources and Cost Considerations

A.8.1 ChatGPT API Usage and Cost

In this study, we utilized the ChatGPT API for the perturbation tasks and the GEMBA-MQM inference. The total cost to run these API-based processes was approximately $50.

A.8.2 Neural Metric Inference on GPU

For the inference of neural metrics we used a NVIDIA A40 Tensor Core GPU. This GPU was essential for efficiently processing large-scale evaluations, particularly for document-level segmentation, where longer input sequences require substantial computational resources.

```
[Anaphoric pronoun gender swap Instruction -- French:]

Voici un segment en anglais (ANG) et sa traduction en francais (FR) :

ANG : {source_segment}
FR : {target_segment}

Objectif
Creer une version contrastive de la traduction francaise en inversant le genre
d'un pronom sujet << il >> ou << elle >> **uniquement si ce pronom traduit le
pronom anaphorique << it >>** du segment anglais.

La phrase modifiee doit rester :
- grammaticale ;
- naturelle lorsqu'elle est lue isolement ;
- une traduction fidele du segment anglais.

Regles
1. Repere dans FR le **pronom sujet << il >> ou << elle >> qui correspond au
   << it >> anaphorique** d'ANG
   - **Ne considere pas** les pronoms qui traduisent << she >>, << he >>, << they >>
     ou tout autre pronom.

2. Ignore toute occurrence de << il >> utilisee dans une tournure impersonnelle
   ou figee :
   - << il faut >>, << il y a >>, << il s'agit de >>, << il convient de >>, etc.
   - verbes impersonnels : << il pleut >>, << il semble >>, etc.
   - schema << il est + ADJ + de/que >>.
   -> Dans tous ces cas, reponds exactement : 'None'

3. Ne modifie pas si le pronom possede un antecedent explicite :
   - dans la meme phrase **ou**
   - dans la phrase immediatement precedente.
   -> Reponds 'None'.

4. Sinon, inverse le genre du pronom (<< il >> <-> << elle >>) **uniquement si**
   la phrase reste fluide, correcte et fidele.

5. Si l'inversion introduit une erreur, une phrase etrange ou un contresens,
   reponds exactement : 'None'.

 Sortie attendue
- la phrase francaise modifiee avec le pronom inverse, **ou**
- 'None'
```

Listing 1: Prompt template for the swapping of anaphoric pronouns genders in French.

```
[Anaphoric pronoun gender swap Instruction -- German:]

Hier ist ein englischer Satz (ENG) und seine deutsche Ubersetzung (DE):

ENG: {source_segment}
DE : {target_segment}

Ziel
Erstelle eine kontrastive Version der deutschen Ubersetzung, indem du das Genus
des Subjektpronomens << er >> bzw. << sie >> **austauschst -- jedoch nur, wenn
dieses Pronomen das anaphorische << it >>** des englischen Satzes wiedergibt.

Der modifizierte Satz muss
- grammatisch korrekt sein,
- isoliert naturlich klingen und
- weiterhin eine getreue Ubersetzung des englischen Ausgangssatzes bleiben.

Regeln
1. Finde in DE das **Subjektpronomen << er >> oder << sie >>, das dem anaphorischen
   << it >>** in ENG entspricht.
   - **Ignoriere** Pronomen, die << he >>, << she >>, << they >> o. A. ubersetzen.

2. Ignoriere unpersonliche << es >>-Konstruktionen:
   - << es gibt >>, << es ist >>, << es scheint >>, << es handelt sich >>, usw.
   -> In all diesen Fallen antworte exakt: ‘None‘

3. Tausche nichts aus, wenn das Pronomen ein explizites Antezedens besitzt
   - im selben Satz **oder**
   - im unmittelbar vorhergehenden Satz.
   -> Antworte ‘None‘.

4. Vertausche ansonsten das Genus des Pronomens (<< er >> <-> << sie >>) **nur
   wenn** der Satz dabei flussig, korrekt und semantisch stimmig bleibt.

5. Fuhrt der Tausch zu Fehlern, Unnaturchlichkeit oder Sinnverfvalschung,
   antworte exakt: ‘None‘.

 Erwartete Ausgabe
- der modifizierte deutsche Satz mit vertauschtem Pronomen, **oder**
- ‘None‘
```

Listing 2: Prompt template for the swapping of anaphoric pronoun genders in German.

```
Aqui tienes un segmento en ingles (ENG) y su traduccion al espanol (ES):

ENG: {source_segment}
ES : {target_segment}

Objetivo
Crear una version contrastiva de la traduccion espanola invirtiendo el genero
del pronombre sujeto << el >> o << ella >> **solo si dicho pronombre traduce el
pronombre anaforico << it >>** del segmento ingles.

La frase modificada debe seguir siendo:
- gramatical,
- natural cuando se lee de forma aislada y
- una traduccion fiel del segmento ingles.

Reglas
1. Localiza en ES el **pronombre sujeto << el >> o << ella >> que corresponde al
   << it >> anaforico** de ENG.
   - **No tengas en cuenta** los pronombres que traduzcan << he >>, << she >>,
     << they >> u otros.

2. Ignora usos impersonales o formulas fijas sin pronombre explicito
   (<< hay >>, << es necesario >>, << es importante >>, etc.).
   -> En esos casos responde exactamente: ‘None‘

3. No modifiques si el pronombre tiene un antecedente explicito:
   - en la misma oracion, **o**
   - en la inmediatamente anterior.
   -> Responde ‘None‘.

4. De lo contrario, invierte el genero (<< el >> <-> << ella >>) **solo si** la
   oracion sigue siendo fluida, correcta y fiel.

5. Si la inversion produce un error, una oracion extrana o un cambio de
   sentido, responde exactamente: ‘None‘.

Salida esperada
- la oracion espanola modificada con el pronombre invertido, **o**
- ‘None‘
```

Listing 3: Prompt template for swappping of the gender of anaphoric pronouns in Spanish.

```
[Few-shot -- French:]

#1
ANG : It was a beautiful morning. It was the first to arrive.
FR : Il faisait tres beau ce matin. Elle a ete la premiere a arriver.
ANS : Il faisait tres beau ce matin. Il a ete le premier a arriver.

#2
ANG : I saw the box, it is beautiful.
FR : J'ai vu la boite, elle est belle.
ANS : None

#3
ANG : It seems that something is going wrong. It eats everything.
FR : Il semble que quelque chose ne se passe pas bien. Il mange tout.
ANS : Il semble que quelque chose ne se passe pas bien. Elle mange tout.

#4
ANG : It should be noted that the results were unexpected.
FR : Il convient de noter que les resultats etaient inattendus.
ANS : None


[Few-shot -- German:]

#1
ENG : It was a beautiful morning. It was the first to arrive.
DE : Es war ein wunderschoner Morgen. Er war der Erste, der ankam.
ANS : Es war ein wunderschoner Morgen. Sie war die Erste, die ankam.

#2
ENG : I saw the box. It is beautiful.
DE : Ich habe die Kiste gesehen. Sie ist schon.
ANS : None

#3
ENG : It seems that something is going wrong. It eats everything.
DE : Es scheint, dass etwas schieflaift. Er frisst alles.
ANS : Es scheint, dass etwas schieflaift. Sie frisst alles.

#4
ENG : It should be noted that the results were unexpected.
DE : Es sei darauf hingewiesen, dass die Ergebnisse unerwartet waren.
ANS : None


[Few-shot -- Spanish:]

#1
ENG : It was a beautiful morning. It should be noted that the results were unexpected.
ES : Era una manana preciosa. Cabe senalar que los resultados fueron inesperados.
ANS : None

#2
ENG : It was the first to arrive.
ES : El fue el primero en llegar.
ANS : Ella fue la primera en llegar.

#3
ENG : I saw the box. It is beautiful.
ES : Vi la caja. Ella es bonita.
ANS : None

#4
ENG : It seems that something is going wrong. It eats everything.
ES : Parece que algo va mal. El se lo come todo.
ANS : Parece que algo va mal. Ella se lo come todo.
```

Listing 4: Few-shot contrastive examples accompanying instruction templates for the swapping of anaphoric pronoun genders in French, German and Spanish.

```
[System Message -- Past-Tense Perturbation:]

You are a professional linguist specializing in English-to-French translation,
with expertise in distinguishing and manipulating past tense forms in French.
You will always respond in French.


[Past-Tense Perturbation Instruction -- French:]

Segment source en anglais :
‘‘‘{src}‘‘‘

Traduction francaise existante :
‘‘‘{tgt}‘‘‘

Objectif
Creer une variante correcte de la traduction francaise en changeant
uniquement le temps verbal des verbes qui sont deja conjugues au
passe compose, a l’imparfait ou au passe simple, en utilisant un autre
temps du passe parmi :
- passe compose
- imparfait
- passe simple

La phrase modifiee doit rester :
- grammaticale ;
- naturelle lorsqu’elle est lue isolement ;
- fidele au segment anglais.

Regles
1. Ne change pas le verbe (pas de synonyme, pas de reformulation).
2. Ne change aucun autre mot.
3. Ne modifie pas les verbes qui ne sont pas au passe compose,
   imparfait ou passe simple.
4. N’ajoute rien (pas d’adverbe, ni de date, etc.).
5. N’utilise pas le plus-que-parfait.

Sortie attendue
- le segment francais modifie avec un autre temps du passe, **ou**
- ‘None‘ si aucun verbe au passe compose, imparfait ou passe simple n’est present
```

Listing 5: System message and prompt template for past-tense perturbations in French.

```
[System Message -- Tense Perturbation (Spanish):]

You are a professional Spanish linguist specialized in tense alternation.
Output must be in Spanish, grammatically correct, and identical to the given
sentence except for converting preterito perfecto to preterito imperfecto
and preterito indefinido or vice-versa. Do not add, remove, or rephrase any
other word.


[Tense Perturbation Instruction -- Spanish:]

Segmento en ingles (original)
‘‘‘{src}‘‘‘

Traduccion espanola existente
‘‘‘{tgt}‘‘‘

Tarea
Genera **una unica variante** de la traduccion espanola que :

1. conserve **toda la informacion** del segmento ingles ;
2. sea **fluida e idiomatic** en espanol moderno ;
3. **modifique exclusivamente el tiempo verbal** de los verbos que esten en
   - preterito perfecto
   - preterito imperfecto
   - preterito indefinido

Conversion permitida
(Elige **una sola** conversion coherente. Si cambiar el tiempo haria la frase incoherente
con adverbios temporales o alteraria su sentido, devuelve exactamente ‘None‘.)

- Preterito perfecto -> preterito imperfecto
- Preterito imperfecto -> preterito perfecto
- Preterito imperfecto -> preterito indefinido
- Preterito indefinido -> preterito imperfecto

Reglas estrictas
1. Manten el mismo verbo y el mismo lexico siempre que sea posible ;
   pequenos ajustes de concordancia o pronombres son aceptables **solo** si mejoran
   la naturalidad sin anadir ni quitar informacion.
2. No introduzcas datos nuevos ni omitas informacion presente en el original ingles.
3. No cambies otras palabras salvo lo imprescindible para la concordancia verbal.
4. No emplees presente, futuro ni pluscuamperfecto.
5. Si hay marcadores temporales (p. ej. *hoy, esta semana, ayer*) incompatibles con la
   conversion, o si no hay verbo en los tiempos indicados, responde ‘None‘.
6. En segmentos con varias oraciones, aplica la conversion donde proceda y devuelve
   el **segmento completo** (las frases sin cambio permanecen intactas).

Salida
- **Si la modificacion es posible** : devuelve **solo** la traduccion espanola completa modificada,
    sin comentarios.
- **Si no** : devuelve exactamente ‘None‘.
```

Listing 6: System message and prompt template for tense perturbations in Spanish.

```
[System Message -- Tense Perturbation (German):]

You are a professional German linguist specialized in tense alternation.
Output must be in German, grammatically correct, and identical to the given
sentence except for converting Perfekt to Prateritum or vice-versa.
Do not add, remove, or rephrase any other word.


[Tense Perturbation Instruction -- German:]

Englischer Ausgangssatz:
‘‘‘{src}‘‘‘

Vorliegende deutsche Ubersetzung:
‘‘‘{tgt}‘‘‘

Aufgabe
Erstelle eine neue, grammatikalisch korrekte Variante der deutschen Ubersetzung.
Du darfst **ausschliesslich** die Zeitform der Verben verandern, **wenn diese aktuell im Perfekt oder
     Prateritum stehen**.

Andere dabei:
- **Perfekt -> Prateritum** oder
- **Prateritum -> Perfekt**

Strikte Regeln
1. Verwende denselben Verbstamm -- keine Synonyme, keine Umschreibung.
2. **Andere kein anderes Wort** (Partikel, Kasus, Satzstellung, Interpunktion, Grob-/Kleinschreibung)
    .
3. **Fuge nichts hinzu** (keine Adverbien, Datumsangaben, Nebensatze, Partizipialkonstruktionen).
4. **Kein Prasens, kein Futur, kein Plusquamperfekt**!
5. Enthalt der Satz **kein Verb im Perfekt oder Prateritum**, gib genau:
   ‘None‘
6. Wenn das Segment aus mehreren Satzen besteht, gib in der Ausgabe stets das
   **gesamte Segment** zuruck -- auch die Satze ohne Anderungen.

Ausgabe
- Wenn eine Anderung moglich war: **nur** die **vollstandige** modifizierte deutsche Ubersetzung (
    ganzer Segmenttext) --
**keinerlei Kommentare, Erklarungen oder Begrundungen**.
- Wenn keine Anderung moglich war: genau ‘None‘ (ohne Backticks, ohne Leerzeichen).
```

Listing 7: System message and prompt template for tense perturbations in German.

```
[System Message -- Lexical Synonym Substitution (French):]

Tu es un expert en grammaire francaise.
Quand c'est possible tu remplaces les mots demandes par des synonymes SANS changer le sens, la
    coherence ni la grammaire du segment.
Ne touche pas aux noms propres ni aux mots qui ne sont pas des noms communs.
Renvoie UNIQUEMENT le segment modifie (sans commentaire) ou, si c'est impossible, 'None'.


[Lexical Synonym Substitution -- French Prompt:]

Segment :
{sentence}

Ta tache :
1. Remplace {'les mots' if is_plural else 'le mot'} {wf} par {'des synonymes' if is_plural else 'un
    synonyme'}.
2. Le sens, la coherence et la grammaire doivent rester STRICTEMENT inchanges.
3. Ne remplace pas un mot s'il s'agit d'un nom propre ou d'un mot qui n'est pas un nom commun.
4. Si tu ne peux pas effectuer la transformation sans alterer le sens/coherence, renvoie exactement :
     None

Rappel : la sortie **ne doit contenir que le segment final modifie** (aucune explication).
```

Listing 8: System message and prompt template for lexical synonym substitution in French.

```
[System Message -- Lexical Synonym Substitution (Spanish):]

Eres un experto en gramatica espanola.
Sustituye las palabras solicitadas por sinonimos SIN cambiar el sentido, la coherencia ni la
    gramatica.
No alteres nombres propios ni palabras que no sean sustantivos comunes. Devuelve SOLO el segmento
    modificado o, si no es posible, 'None'.


[Lexical Synonym Substitution -- Spanish Prompt:]

Segmento:
{sentence}

Tarea:
1. Sustituye {'las palabras' if is_plural else 'la palabra'} {wf} por {'sinonimos' if is_plural else
    'un sinonimo'}.
2. El sentido, la coherencia y la gramatica deben permanecer EXACTAMENTE iguales.
3. No reemplaces una palabra si es un nombre propio o si no es un sustantivo comun.
4. Si no es posible sin alterar el significado o la coherencia, responde exactamente: None

Recuerda: la salida **debe contener SOLO el segmento final modificado** (sin explicaciones).
```

Listing 9: System message and prompt template for lexical synonym substitution in Spanish.

```
[System Message -- Lexical Synonym Substitution (German):]

Du bist ein Experte fur deutsche Grammatik.
Ersetze die angeforderten Worter durch Synonyme, OHNE Bedeutung, Koharenz, Grammatik oder Syntax zu
    andern.
Andere keine Eigennamen oder Worter, die keine gewohnlichen Substantive sind.
Gib NUR den modifizierten Satz oder, falls unmoglich, 'None' zuruck.


[Lexical Synonym Substitution -- German Prompt:]

Segment:
{sentence}

Aufgabe:
1. Ersetze {'die Worter' if is_plural else 'das Wort'} {wf} durch {'Synonyme' if is_plural else 'ein
    Synonym'}.
2. Bedeutung, Koharenz, Grammatik und Syntax mussen EXAKT erhalten bleiben.
3. Ersetze kein Wort, wenn es ein Eigenname ist oder kein gewohnliches Substantiv.
4. Falls dies nicht moglich ist, gib genau zuruck: None

Wichtig: Die Ausgabe **darf NUR den endgultig modifizierten Satz enthalten** (keine Erklarungen).
```

Listing 10: System message and prompt template for lexical synonym substitution in German.