Discourse Dependency: A Continuous Criterion for Translation Difficulty

Ahrii Kim[1] **Chanjun Park**[2] ⋆ **Seong-heum Kim**[1,3] ⋆
[1]AI-Bio Convergence Research Inst. [2]School of Software
[3]Dept. of Intelligent Semiconductors
Soongsil University
{ahriikim,chanjun.park,seongheum}@ssu.ac.kr

Abstract

Recent calls for harder machine translation benchmarks have not clarified what difficulty should mean. We argue that one meaningful and currently unmeasured axis is *referential reach*, the distance a segment must look back into its document to resolve the entities and pronouns it contains. We formalize this as *discourse dependency* (DDP), a metric-free, source-side measure computed from named entity re-mentions and pronominal coreference. Validated against gold coreference, DDP errs one-sidedly in 99.2% of segments, so a high-DDP segment is certified to require long-range context. Applying DDP to WMT24++ and WMT25 shows that both are heavily skewed toward low-DDP segments, which domain labels do not distinguish. Building on DDP, we compare five context injection strategies in an English-Korean post-editing setup, varying context size and selection. As DDP grows, no strategy keeps pace with human post-editing. On segments with DDP ≥ 15 raters prefer human translations, while automatic metrics register no difference. As frontier systems saturate aggregate scores, DDP shifts evaluation from *how well models score* to *how far they can reach*.

1 Introduction

Translation quality has improved dramatically with large language models (LLMs), to the point where general-domain benchmarks can no longer reliably distinguish state-of-the-art systems (Kocmi et al., 2024, 2025; Akhtar et al., 2026). The machine translation (MT) community has responded by seeking harder evaluation data, either by selecting difficult instances from existing datasets or by generating them synthetically (Zouhar et al., 2025; Proietti et al., 2025; Zouhar et al., 2026). These efforts share a common premise: difficulty is estimated from translation quality scores, using automatic metrics as a proxy for how much a sentence challenges a model, a premise worth revisiting now that aggregate scores no longer differentiate frontier systems.

This premise has two limitations. First, automatic metrics are unreliable for the subtle quality differences that distinguish frontier systems (Mathur et al., 2020; Freitag et al., 2022; Lavie et al., 2025), so using them as difficulty estimators risks optimizing evaluation toward what metrics can measure rather than what genuinely challenges models. Second, the approach treats difficulty as a property of individual segments, while translation difficulty is often not intrinsic to a sentence but relational, depending on how far the information required to translate it faithfully lies within the document (Halliday and Hasan, 1976; Hardmeier, 2012).

A parallel line of work has long recognized this relational nature of translation. Document-level MT studies show that discourse phenomena such as coreference, lexical cohesion, and tense consistency require information beyond the sentence (Voita et al., 2019; Bawden et al., 2018; Fernandes et al., 2021), and more recent efforts have moved toward systematic identification of context-dependent segments through automatic taggers or rule-based extraction pipelines (Fernandes et al., 2023; Wicks and Post, 2023). These evaluations remain phenomenon-specific: a sentence either falls under a target phenomenon or it does not, and coverage depends on per-language, per-phenomenon rule engineering. What is missing is a continuous, document-level measure that quantifies how much contextual reach a segment requires, applicable uniformly across genres and datasets to characterize a benchmark's overall difficulty profile.

We therefore reframe difficulty around a single document-level property, **discourse dependency** (DDP): the distance within a document at which discourse-relevant information appears relative to

⋆Corresponding authors.
https://github.com/trotacodigos/ddp.git

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 9887–9912
October 24-29, 2026 ©2026 Association for Computational Linguistics

the current segment, operationalized as the maximum distance between successive mentions of the same entity. DDP is metric-free and source-side, which lets it serve as a difficulty criterion independent of any model or scoring function. It also recasts the role of domain in MT evaluation: rather than a topical label, domain becomes a coarse proxy for the discourse dependency that DDP measures directly.

To test whether DDP tracks translation difficulty, we examine how models exploit context in automatic post-editing (APE) for English-Korean (En-Ko), a language pair known to be challenging both linguistically and for translation evaluation (Park and Padó, 2024; Choi et al., 2018; Lee et al., 2025). Our main contributions are as follows:

- **Discourse dependency (DDP).** A linguistically grounded, metric-free difficulty criterion computed from entity re-mentions and pronominal coreference, exposing within-domain variation that domain labels conflate.
- **A DDP-stratified evaluation framework.** Five context injection strategies paired with DDP-stratified analysis, providing a discourse-level account of when context utilization breaks down.
- **A diagnostic of automatic metrics.** DDP stratification exposes a blind spot in standard metrics, providing a concrete axis for auditing metric reliability beyond aggregate scores.
- **DDP-annotated benchmarks.** DDP statistics for WMT24++ and WMT25, showing that current evaluation data is heavily skewed toward low-DDP segments.

Organizing evaluation along this axis supports test sets stratified by discourse progression rather than instance-level difficulty selection, targeting the inter-sentential reasoning that current systems still lack. It also exposes how much within-domain variation topical labels conceal (Section 3.3).

2 Related Work

Difficulty in MT evaluation. As frontier MT systems converge on general-domain benchmarks, recent work has shifted toward harder evaluation data (Kocmi et al., 2024, 2025). One line selects difficult instances from existing corpora using metric-based filtering or model disagreement (Zouhar et al., 2025; Proietti et al., 2025), while another generates challenging instances synthetically, often through perturbation or targeted construction (Zouhar et al., 2026). Both rely on automatic quality scores to define difficulty, inheriting the known limitations of these metrics at the high end of the quality scale (Mathur et al., 2020; Freitag et al., 2022; Lavie et al., 2025). DDP departs from this framing by deriving difficulty from a source-side, document-level discourse property, independent of any model or metric. Metric-derived estimates require system outputs, so they can filter an existing test set but cannot inform its design, whereas DDP is available before any translation exists. We therefore do not expect the two to correlate, since a source-side criterion that reproduced quality scores would add nothing (Section 5).

Discourse phenomena in MT evaluation. Sentence-level evaluation obscures discourse-sensitive errors, prompting two complementary responses. WMT human evaluation has progressively expanded rater context from neighboring sentences to paragraph-level pilots (Barrault et al., 2019; Akhbardeh et al., 2021; Kocmi et al., 2022, 2023), though ratings still tend toward ceiling effects (Kocmi et al., 2024). In parallel, contrastive challenge sets and WMT test suites isolate phenomena such as pronoun resolution, lexical cohesion, and ellipsis through binary preference on minimal pairs (Müller et al., 2018; Voita et al., 2019; Manakhimova et al., 2024; Bhattacharjee et al., 2024), while recent work moves toward systematic identification of context-dependent segments (Fernandes et al., 2023; Wicks and Post, 2023) and toward measuring whether models actually use the supplied context (Mohammed and Niculae, 2024). A shared finding across these lines is that standard WMT test sets contain few context-dependent sentences, with document-level gains visible only on discourse-dense subsets (Post and Junczys-Dowmunt, 2024). Yet discourse remains operationalized through phenomenon-specific annotations or binary classifications, leaving open the question of *how much* contextual reach a segment requires.

Context-aware MT. Approaches to incorporating document context have evolved along two axes: how much context to provide, and how to select it. Early neural systems concatenate a fixed window of k preceding segments (Tiedemann and Scherrer, 2017; Junczys-Dowmunt, 2019), with hierarchical and cache-based variants encoding surrounding sentences through separate modules (Miculi-

cich et al., 2018; Maruf et al., 2019). This sequential window remains the de facto standard, though architectural extensions yield diminishing returns once context is long enough (Sun et al., 2022; Post and Junczys-Dowmunt, 2024). The shift to LLM-based MT has reframed the problem from architecture to prompting. Frontier LLMs can ingest full documents directly (Wang et al., 2023; Karpinska and Iyyer, 2023), yet they do not exploit extended input uniformly and often degrade well before the nominal limit (Liu et al., 2024; Mohammed and Niculae, 2024). Our results offer a discourse-level account of this failure. The context utilization degrades systematically as DDP grows, indicating that the bottleneck is not input length but the linguistic reach a segment demands.

3 DDP: Discourse Dependency

3.1 Theoretical Background

The difficulty of translating a segment depends not only on its internal structure but on where in the document its required information lies. A linguistically simple segment can be contextually demanding when its referents are introduced many segments earlier. Consider the following passage:

(1) John arrived.
(2) He looked tired.
(3) The manager greeted him.
(4) John smiled.
(5) She offered him coffee.

Translating segment (5) requires knowing that *she* refers to *the manager* from (3), and that *him* refers to *John* from (1). A system that sees only the immediately preceding segment cannot resolve either reference.

We formalize this intuition as *discourse dependency* (DDP). Each segment contains *anchors*, entity mentions whose interpretation depends on a prior segment (Hobbs, 1978; Grosz and Sidner, 1986), and DDP records the distance from each anchor to its antecedent according to two linguistically motivated rules: ① **a named entity is usually the first mention in the document**, since it typically introduces new information. If a prior mention exists, the distance is measured to it (Prince, 1981). ② **A pronoun is anchored to the most recent named entity or head noun of compatible gender and number**, following the accessibility hierarchy of referential expressions (Clark and Haviland, 1977; Grosz et al., 1995). The segment-level DDP is the maximum anchor distance within the segment. Table 1 illustrates the computation on the passage above. DDP measures one axis, referential reach, and is a *lower bound* on contextual dependency rather than a general model of translation difficulty (Section 3.2).

Seg	Anchor	Antecedent	d
(1)	*John*	— (first mention)	0
(2)	*he*	*John* (1)	1
(3)	*manager*	— (first mention)	0
	him	*John* (1)	2
		$\mathrm{ddp}_{(3)} = \max(0, 2)$	**2**
(4)	*John*	*John* (1)	3
(5)	*she*	*manager* (3)	2
	him	*John* (1)†	4
		$\mathrm{ddp}_{(5)} = \max(2, 4)$	**4**

Table 1: DDP computation on the example passage. Each anchor is traced to its antecedent; the segment DDP is the maximum anchor distance (d). †*him* is chained through its most recent named entity (*John*, seg 4), whose own distance from its first mention (seg 1) is accumulated: $(5-4) + 3 = 4$.

3.2 Operationalization

Anchor extraction. We extract two types of discourse anchors from the English source: named entities and referential pronouns.

Named entities are identified by a named entity recognizer (NER), restricted to four categories — PERSON, ORG, GPE, LOC — which are the types most likely to persist across segments and require discourse-level resolution.

Pronouns are identified through part-of-speech (POS) tagging (`pos_ == PRON`) and restricted to referential uses, and expletive pronouns (e.g., *it* in *it is raining*) are excluded via dependency parsing (`dep_ == expl`). Pronouns are grouped by gender and number into four classes: MASC (*he, him, his*), FEM (*she, her, hers*), NEUT (*it, its*), and PLUR (*they, them, their*).[1] Head nouns in subject or object position (`dep_` $\in$ `{nsubj, dobj, nsubjpass}`, `pos_ == NOUN`) that are not captured by NER serve as fallback antecedents.

Distance computation. We process each document sentence by sentence, aligning the computation with the conventional evaluation unit. For each anchor in sentence s_i, we compute a distance $d(s_i, a)$ according to Rules ① and ②.

[1] Singular gender-neutral *they* (e.g., for non-binary referents) is not separately classified and falls into the PLUR group.

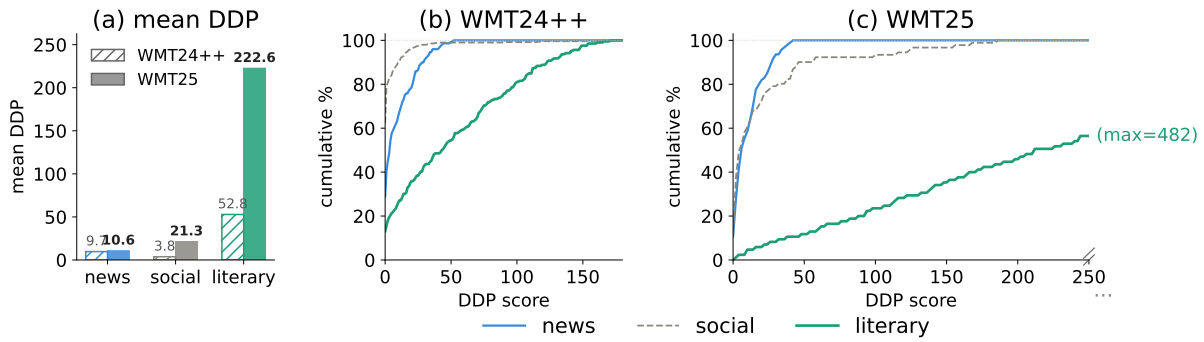


Figure 1: DDP score distributions across domains in WMT24++ and WMT25. **(a)** Mean DDP per domain: literary texts exhibit substantially higher dependency than news and social across both benchmarks, with WMT25 showing a marked increase across all domains (μ=222.6 for literary vs. 52.8 in WMT24++) due to the inclusion of long-form narrative texts and longer conversational threads. **(b–c)** Cumulative distribution of DDP scores: news segments saturate quickly in both benchmarks, while literary scores spread widely, reaching $\mathrm{ddp} = 171$ in WMT24++ and $\mathrm{ddp} = 482$ in WMT25. Note the different x-axis ranges in (b) and (c).

Rule ① (named entities). Let e be a named entity mention in s_i, and let s_j ($j < i$) be the earliest prior sentence containing a mention of the same entity:

$$d(s_i, e) = \begin{cases} 0 & \text{if no prior mention exists,} \\ i - j & \text{otherwise.} \end{cases} \tag{1}$$

Rule ② (pronouns). A pronoun p in s_i of gender-number group g cannot be a first mention, so it is anchored to the most recent named entity or head noun of compatible group g in some prior sentence s_k ($k < i$).

$$d(s_i, p) = \begin{cases} (i - k) + d(s_k, e_k) & \text{named entity } e_k, \\ i - k & \text{head noun,} \\ 1 & \text{no antecedent.} \end{cases} \tag{2}$$

When the antecedent is a named entity, the chain accumulates its own distance $d(s_k, e_k)$ from its first mention. A head noun outside NER carries no such distance, so no accumulation applies. When no compatible antecedent exists, $d = 1$ reflects a minimal dependency on the immediately preceding sentence.

Each anchor in s_i is computed independently, and the sentence-level DDP is the maximum distance across all anchors $\mathcal{A}_i$:

$$\mathrm{ddp}(s_i) = \max_{a \in \mathcal{A}_i} d(s_i, a). \tag{3}$$

Sentences with no anchors, or only first-mention anchors, receive $ddp = 0$, indicating that no prior context is required for their translation.

Evaluation unit. Distances are counted in sentences and searched over the whole document, so an antecedent may lie anywhere before s_i regardless of segment boundaries. Segments are only the unit at which we report. When a segment spans several sentences we take the maximum,

$$\mathrm{ddp}(\mathrm{seg}) = \max_{s_i \in \mathrm{seg}} \mathrm{ddp}(s_i). \tag{4}$$

Segmentation therefore affects only how the sentence-level distances are grouped for reporting, not the distances themselves.

Implementation. NER, POS tagging, dependency parsing, and sentence segmentation are performed jointly with spaCy `en_core_web_trf`, a transformer-based pipeline (RoBERTa backbone) that achieves state-of-the-art accuracy on OntoNotes (Hovy et al., 2006). Appendix A.3 confirms that DDP is robust to the choice of NER and POS tagger. Replicating the full pipeline with the architecturally distinct Flair tagger (Akbik et al., 2019) reproduces segment-level DDP at Pearson $r = 0.89$ with 69% exact agreement, and preserves the domain ordering. All computations run on the English source, making DDP independent of the target language and translation system.

Scope and validity as a lower bound. Unmodeled phenomena such as lexical cohesion and bridging anaphora, and links missed by the tagger, lower DDP but do not inflate it, since inflation would require a spurious *early* entity mention. The guarantee is thereforeone-sided. A high-DDP segment is certified to require long-range context, while a low-DDP segment is *not* certified to be easy, and

Model	Params	Context	Cutoff	En-X	Reasoning	MoE	Thinking
Open-source							
GEMMA-4-31B (Gemma Team, 2026)	30.7B	256K→128K	Jan 2025	✓	✓	–	✓
QWEN3.5-27B (Qwen Team, 2026)	27B	256K→128K	–	✓	–	–	✓
HYPERCLOVA-X-SEED-THINK-32B (NAVER, 2026)♦	32B	128K	–	✓	✓	–	✓
DEEPSEEK-V4-PRO (DeepSeek-AI, 2026)	1.6T (A49B)	1M→128K	–	–	✓	✓	✓
Closed							
GPT-5.4 (Singh et al., 2026)	–	1.05M	Aug 2025	✓	✓	–	✓
GEMINI-3-FLASH (Google DeepMind, 2026)	–	1M	Jan 2025	✓	✓	–	✓

Table 2: LLMs evaluated across input configurations. *Cutoff* indicates the knowledge cutoff date as publicly disclosed; "–" denotes undisclosed. *En-X* indicates whether the model was explicitly trained on X∈ {Ko, Zh}. *Reasoning* = general reasoning capability, *MoE* = Mixture of Experts architecture, *Thinking* = a chain-of-thought mode is *available*. It is disabled in all our runs for comparability (Section 4). ♦ Korean-specialized models. Context lengths are restricted to 128K.

every claim in this paper rests on that asymmetry. To check that the bound is tight enough to be useful, we validate DDP against gold coreference annotations from OntoNotes 5.0 on 80 documents (Appendix A.1). Segment-level DDP correlates with gold distance at Pearson $r = 0.81$, and the error is one-sided in 99.2% of segments, with the residual 0.8% attributable to NER false positives.

3.3 WMT Sets Through the Lens of DDP

To show that DDP captures variation that domain labels miss, we analyze two widely used MT benchmarks: WMT24++ (Deutsch et al., 2025) and WMT25 (Kocmi et al., 2025). Figure 1 shows the segment-level DDP distribution across domains, and Table 5 reports summary statistics.

Across both benchmarks, literary texts exhibit substantially higher DDP than news and social domains (Figure 1a). In WMT24++, literary segments have a mean $\mathrm{ddp} = 52.8$ ($\sigma = 47.0$), compared to 9.7 for news and 3.8 for social. WMT25 follows the same ordering, with literary reaching a mean of 222.6 (max 482) due to long-form narratives with character references spanning hundreds of sentences. WMT25 also shifts social upward to $\mu = 21.3$, reflecting longer conversational threads than in WMT24++. Within WMT24++, social falls below news ($\mu = 3.8$ vs. 9.7), against the assumption that conversational texts carry rich discourse structure. *Discourse dependency thus varies within domains as much as across them*: a news article with persistent entity references can be harder to translate than a literary passage of self-contained sentences, yet both carry the same domain label.

The cumulative distributions (Figure 1b–c) reveal a structural bias in both benchmarks: news segments saturate well below $\mathrm{ddp} = 50$, while literary segments extend far into the high-dependency range. WMT25 covers substantially more high-DDP texts than WMT24++, yet the bulk of segments in both remain at low to moderate dependency. Current MT benchmarks thus systematically undersample contextually difficult cases, providing quantitative grounding for Post and Junczys-Dowmunt (2024)'s observation that document-level gains surface mainly on discourse-dense subsets. This limitation becomes critical when evaluating document-level systems.

4 Experimental Setup

4.1 Data

We use WMT24++ (Deutsch et al., 2025), which provides professionally post-edited translations (HPE) across four domains. We focus on English-Korean (En-Ko), a target language particularly sensitive to discourse. Korean's pro-drop structure requires inter-sentential restoration of omitted subjects and objects (Lee et al., 2025), and its honorific and named-entity systems demand cross-segment consistency, making En-Ko challenging for both translation and evaluation (Park and Padó, 2024; Choi et al., 2018). After excluding the speech domain and documents with fewer than two segments, we retain 59 En-Ko documents spanning literary, news, and social. Dataset statistics are in Appendix B.1. We also report English-Chinese (En-Zh) results in Appendix A.4 for cross-lingual consistency.

Why post-editing. We evaluate in an APE setup, in which a model revises an existing draft, rather than translating from scratch. This isolates our research question, whether models *use* the context they are given, from raw generation quality, which would otherwise confound not-using-context with not-translating-well. It also reduces the benefit of

memorized references (Limitations). Since DDP is computed source-side, the stratification itself is unaffected by this choice.

4.2 Models

We employ six state-of-the-art LLMs spanning open-weight and proprietary systems, selected to represent diversity in parameter scale, architecture, and language specialization (Table 2). All models are prompted with a uniform 128K context limit, thinking mode disabled for comparability (the *Thinking* column of Table 2 records whether a chain-of-thought mode is *available*, not whether it is used), and fixed decoding parameters (`temperature=1.0`, `top_p=0.95`, `seed=42`), following recent LLM-based MT evaluation protocols (Kocmi et al., 2025).

4.3 Context Injection Strategies

APE_{seg} Size

No context is provided. Null baseline against all context-aware strategies.

APE_{full} Size

The entire document, excluding the current segment pair. Maximizes context quantity while preserving sequential structure.

APE_{seq} Size Selection

The k segment pairs immediately preceding s_i, replicating the de facto standard in prior APE work.

APE_{rel} Selection

The k segment pairs most semantically similar to s_i, retrieved via embedding-based search.

APE_{exp} Selection

Structured declarative knowledge of genre, participant relationships, and register, extracted from the source document by prompting an LLM.

If DDP captures genuine contextual difficulty, models should benefit more from broader context as DDP grows, where the information required for faithful translation lies beyond the immediate window. We test this by varying *context injection*, the deliberate selection and structuring of information placed in the context window for post-editing. Given a document of n segments, let s_i denote the i-th segment under APE and k the number of context segments provided. We design five strategies organized around two assumptions, marked in the boxes above: Size, that more context is better, and Selection, that proximity implies relevance. APE_{seq} serves as the shared pivot against which both are tested.

Size is tested by comparing APE_{seg}, APE_{seq}, and APE_{full}, which provide zero, k, and full-document context. Selection is tested by comparing APE_{seq}, APE_{rel}, and APE_{exp}, which fix context size (k=5) and vary the selection criterion. Prompt structure is held identical across strategies, with only the context block varying (Appendix B.3), and construction details are in Appendix B.2.

4.4 Evaluation Protocol

Human ranking. We conduct relative ranking on two non-overlapping subsets of ≈500 segments, one per assumption. Annotators rank four candidate translations (including HPE) from 1 to 4 with ties allowed. Each segment receives two independent judgments. For analysis, ranks are reversed to preference scores on a 1–4 scale, so that higher values indicate stronger preference. Tied candidates receive the average of the ranks they span, so that the four scores always sum to 10. Inter-annotator agreement measured by Kendall's τ is 0.37 on average, consistent with prior MT ranking studies (Callison-Burch et al., 2007). Details are in Appendix B.5.

Automatic metrics. We report TER (Snover et al., 2006) to quantify edit size, four segment-level metrics (xCOMET-XXL: Guerreiro et al. 2024, xCOMETKIWI-XXL: Rei et al. 2023, METRICX-24-XXL, and METRICX-24-QE-XXL: Juraska et al. 2024), and three document-level metrics (d-BLEU: Liu et al. 2020, SLIDE: Raunak et al. 2023a, doc-COMET: Vernikos et al. 2022) for translation quality. Outputs with TER > 100 are excluded as generation failures (Raunak et al., 2023b).

DDP-stratified analysis. For each segment, we compute its DDP and the mean preference score per strategy, converted from rank to a preference scale (higher = better). We fit LOWESS smoothing (Cleveland, 1981) with bandwidth frac $= 0.4$ to the {DDP, preference} pairs, a non-parametric local regression that estimates the conditional mean at each DDP value without assuming a global functional form, suited to the unevenly distributed and

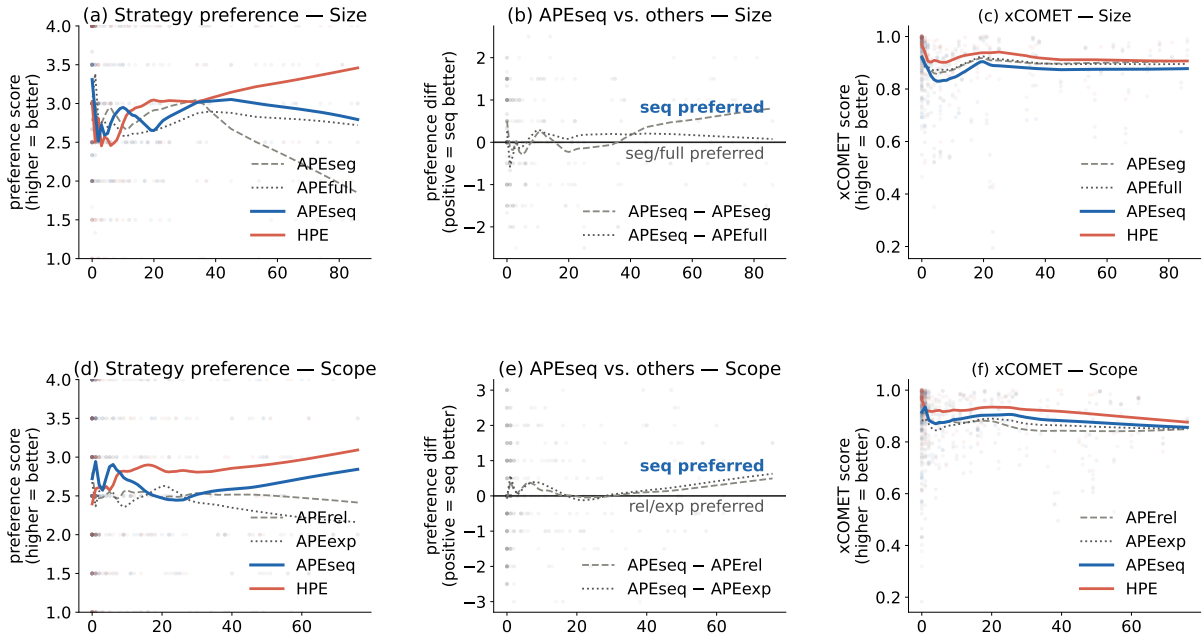


Figure 2: LOWESS-smoothed preference scores and pairwise differences as a function of segment-level DDP. **Left panels** show the mean preference score (higher = better) of each strategy across the DDP spectrum. **Middle panels** show the preference difference between APE_{seq} and each competing strategy (positive = APE_{seq} preferred). **Right panels** show the corresponding XCOMET scores for comparison with human judgments.

potentially non-linear DDP spectrum. We visualize both the absolute preference of each strategy and its difference relative to APE_{seq}, and report standard metric-based performance stratified by DDP to show how an established evaluation frame yields new findings when organized along discourse dependency. Post-edit volume is reported as an auxiliary diagnostic with TER. A qualitative analysis of cases where context-aware post-editing fails and where DDP captures the underlying difficulty is provided in Appendix C.2.

5 Results

5.1 DDP as a Predictor of Context Injection

Strategy preference by human. We conduct relative ranking on two non-overlapping subsets of ≈500 segments, one per assumption. Because documents are selected by inter-condition divergence rather than uniformly (Appendix B.5), these subsets over-represent the high-dependency region. 163 of the 498 Size segments (32.7%) have DDP ≥ 15, against 212 of 849 (25.0%) in the full benchmark.

Figure 2 reveals consistent patterns across both assumptions. At low DDP ($\lesssim$10), multiple APE strategies surpass HPE, indicating that for segments requiring little discourse context, LLMs produce post-edits that annotators prefer over professional human revisions. From DDP ≈10 onward, HPE shows a monotonically increasing preference across the entire spectrum, while no APE strategy catches up.

Within APE strategies, APE_{seq} consistently outperforms APE_{full} across nearly the full DDP range (Figure 2b). Providing the entire document does not help models exploit context more effectively than providing only the k preceding segments, indicating that current LLMs do not extract useful signal from long-range document context even when it is available. APE_{seg} occasionally matches or outperforms APE_{seq} at low DDP despite providing no context, which is consistent with the definition of low-DDP segments as self-contained: injecting sequential context introduces noise rather than signal when no signal is needed.

In Figure 2e, APE_{seq} is preferred over both APE_{rel} and APE_{exp} across the DDP spectrum. The advantage over APE_{exp} shows that providing context in its natural sentence form is more effective than abstracting it into structured declarative knowledge, and the near-identical trajectories of APE_{rel} and APE_{exp} further suggest that what matters is not how segments are selected but the form in which context is delivered.

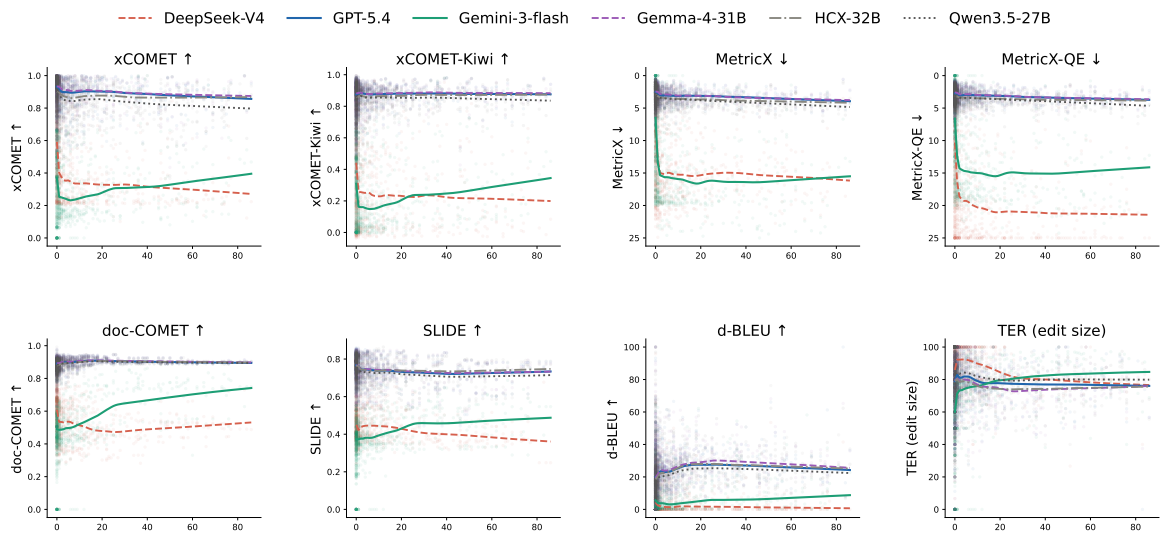


Figure 3: LOWESS-smoothed metric scores per model as a function of segment-level DDP ($n = 229{,}230$ segment–model–strategy–metric observations). Models cluster into two groups across all metrics regardless of DDP. TER (bottom right) decreases as DDP grows across most models, reflecting more conservative outputs in contextually demanding segments.

Strategy preference by metrics. Across all automatic metrics (Figure 10 in Appendix C), strategy preferences remain largely insensitive to DDP, in contrast to human annotators whose preferences sharpen as discourse dependency grows (Figure 2c,f). Two patterns hold throughout. First, metric scores show little directional change as DDP increases, indicating that automatic metrics do not register the growing contextual difficulty that human evaluators respond to. Second, all APE strategies receive near-identical scores across the DDP spectrum, indicating that metrics cannot distinguish strategies that humans clearly differentiate. QE metrics (xCOMETKIWI, METRICX-24-QE) exhibit the flattest trajectories, lacking access to references and missing even the surface-level differences that reference-based metrics detect. XCOMET further ranks APE_{full} above APE_{seq} across the entire DDP spectrum, reversing the human preference.

Statistical confirmation. The widening human preference for 👤✎ HPE holds under significance testing. Over the annotated segments the per-segment preference gap between 👤✎ HPE and APE_{seq} correlates positively with DDP (Spearman $\rho = 0.28$, $p = 1.2 \times 10^{-9}$, $n = 498$), so the effect is not confined to the tail. Restricting to DDP ≥ 15, raters prefer 👤✎ HPE over APE_{seq} by a mean of $+0.35$ on the 1–4 scale (Wilcoxon signed-rank, one-sided $p = 3.0 \times 10^{-5}$, bootstrap 95% CI $[+0.18, +0.52]$, with 👤✎ HPE ranked higher in 60% of segments), whereas below DDP $= 15$ the gap is $+0.05$ and not significant. On the identical segments XCOMET cannot separate the two (Wilcoxon $p = 0.40$, mean gap -0.005) and shows no trend in DDP (Spearman $\rho = 0.03$, $p = 0.50$). Automatic metrics should therefore be interpreted with caution in document-level APE evaluation, particularly at high DDP, and discourse-stratified analyses provide a useful complement to aggregate scores.

5.2 DDP as an Analytical Lens

DDP-level performance. Treating automatic metrics as a useful signal for system-level comparison, DDP reveals patterns that aggregate scores obscure (Figure 3). Models cluster into two groups regardless of DDP: GPT-5.4, GEMMA-4-31B, HYPERCLOVA-X-SEED-THINK-32B, and QWEN3.5-27B outperform the rest across the entire spectrum. Within the upper cluster, the four models produce nearly indistinguishable scores, suggesting that current metrics lack the resolution to rank closely matched systems. GEMINI-3-FLASH is one exception, showing a mild upward trajectory on COMET-series as DDP grows, while DEEPSEEK-V4-PRO trends in the opposite direction. Separating segment- and document-level metrics reveals no qualitatively different patterns, except that d-BLEU yields rankings consistent with neural metrics despite operating at the surface level.

Model	Overall (%)	By DDP stratum (%)			
		$= 0$	1–4	5–14	15+
DEEPSEEK-V4-PRO	3.2	5.1	1.9	1.9	0.7
HCX	7.2	10.1	6.4	4.5	2.8
GEMMA-4-31B	10.6	13.2	12.3	6.4	4.2
GPT-5.4	11.5	13.8	16.0	6.4	2.8
QWEN3.5-27B	13.8	17.7	13.2	10.0	7.0
GEMINI-3-FLASH	26.9	44.2	16.9	15.5	4.9

Table 3: Generation failure rate (TER > 100) per model and DDP stratum. Failure rates consistently decrease as DDP increases across all models.

Post-editing volume. TER decreases as DDP grows across most models and stabilizes beyond DDP $\approx$20 (Figure 3, bottom right), indicating that models produce more conservative outputs for contextually demanding segments. GEMINI-3-FLASH exhibits the highest generation failure rate (26.9%), concentrated at low DDP (44.2% at $\mathrm{ddp} = 0$) and dropping sharply at higher dependency. Since failures are excluded as generation errors and are concentrated at low DDP, part of the decline in TER reflects this differential filtering. We therefore read TER as a diagnostic of edit volume rather than as a quality signal.

6 Conclusion

We introduced discourse dependency (DDP), a metric-free, source-side measure of contextual difficulty in MT, and used it to revisit how context interacts with translation quality. Three implications follow. First, the gap is not a matter of input length. Models can ingest the full document but cannot selectively attend to the antecedents that resolve referential dependencies, while k-preceding injection often omits it entirely. Discourse-aware context selection, not context size, is the bottleneck. Second, metric blindness to DDP is not incidental. Reference- and QE-based metrics are trained on parallel data dominated by sentence-level alignments, where inter-sentential phenomena such as coreference and entity tracking are rare and weakly supervised. Third, organizing evaluation by discourse dependency rather than topical domain admits a sharper accounting of progress, separating segments that current systems handle from those that genuinely demand discourse-level reasoning, before rather than after system outputs are produced. By making discourse dependency explicit in the source, DDP offers a principled axis for evaluating what frontier MT systems actually understand about the documents they translate.

Limitations

Single language pair. We focus on English-Korean as the primary testbed, a typologically distant pair where discourse phenomena such as pro-drop, honorifics, and named-entity rendering make cross-segment context consequential. A partial replication on English-Chinese (Appendix A.4) shows broadly consistent metric trends, suggesting that the insensitivity to DDP is not specific to Korean. Conclusive evidence requires human evaluation on additional pairs, which we leave for future work. How DDP-stratified patterns generalize to other typologically distant pairs, to more closely related pairs, and to non-English source languages remains an important open question.

Anchor coverage. DDP operationalizes contextual dependency through two anchor types: named-entity re-mentions and pronominal coreference. These cover a substantial portion of inter-segmental discourse dependency in English, but other phenomena such as lexical cohesion, tense and aspect consistency, ellipsis, and bridging anaphora fall outside the current definition. Our gender-number grouping also collapses singular gender-neutral *they* (e.g., for non-binary referents) into the plural class, since surface form alone cannot disambiguate the two uses. The variation DDP already exposes within and across domains suggests that even this minimal definition is informative; extending the anchor inventory and pronoun typology is a natural direction for future work.

Dependence on the NLP toolchain. Given a fixed anchor definition, DDP relies on automatic NER, POS tagging, and dependency parsing. Pipeline errors are strongly one-directional, and a missed link shortens a chain and lowers DDP, whereas inflation requires a spurious *early* mention, which is far rarer. Our validation against gold coreference (Appendix A.1) confirms this in practice. DDP under- or exactly estimates the gold distance in 99.2% of segments, and the residual 0.8% is traceable to NER false positives. The lower-bound reading is therefore an empirical near-guarantee rather than an absolute one, and it is tight enough that high-DDP segments can be treated as certified to require long-range context. The ablation in Appendix A.3 further shows that architecturally distinct taggers yield strongly correlated values (Pearson $r = 0.89$ for spaCy vs. Flair) and preserve the domain-level ordering.

Data contamination. WMT24++ was released in February 2025 and may overlap with the training data of recent LLMs. We mitigate this risk by adopting an APE setup. The models receive an initial machine translation and revise it rather than generating from the source alone, which reduces the benefit of memorized references. We cannot fully rule out residual contamination, particularly at low DDP where surface memorization is most plausible. The main finding of this paper, that human preference for HPE grows monotonically with DDP, is least likely to be driven by contamination, since contamination would inflate model output quality rather than diminish it relative to human revisions.

Acknowledgment

This work was supported by the G-LAMP Program of the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education (No. RS-2025-25441317); the Ministry of Science and ICT (MSIT), and the National IT Industry Promotion Agency (NIPA) through the Advanced GPU Utilization Support Program (02-26-01-0499). The MSIT under the Convergence security core talent training business support program (IITP-2024 2024-RS-2024-00426853) supervised by the IITP(Institute of Information & Communications Technology Planning & Evaluation). This work was supported by Korea Internet & Security Agency(KISA) grant funded by the Korea government (PIPC) (No.RS-2026-25526342, Development of Technologies for Preventing Sensitive Information Inference and Risk Assessment in Foundation Model Operations)

Ethics Statement

Our study involves expert human annotators. Participation was voluntary, and annotators were compensated at a fair market rate. All annotations were anonymized, and no personally identifiable information was collected.

Licenses of artifacts. All datasets, models, and software used in this study are publicly available for research purposes. The WMT24++ and WM25 test set is distributed under CC-BY 4.0; the automatic metrics are released under Apache 2.0. GPT-5.4 is used via the OpenAI API under OpenAI's terms of service. Our benchmark and code will be released under CC-BY 4.0 and MIT, respectively.

Intended use. Our use of the WMT24++ and WM25 test set and automatic metrics is consistent with their intended use for MT evaluation research. The annotations and datasets we release are intended for research use only, consistent with the access conditions of the source data.

We acknowledge the use of AI assistants (CLAUDE OPUS 4.6) for writing refinement and code review during paper preparation.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, and 17 others. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Mubashara Akhtar, Anka Reuel, Prajna Soni, Sanchit Ahuja, Pawan Sasanka Ammanamanchi, Ruchit Rawal, Vilém Zouhar, Srishti Yadav, Chenxi Whitehouse, Dayeon Ki, Jennifer Mickel, Leshem Choshen, Marek Šuppa, Jan Batzner, Jenny Chim, Jeba Sania, Yanan Long, Hossein A. Rahmani, Christina Knight, and 18 others. 2026. When ai benchmarks plateau: A systematic study of benchmark saturation. *Preprint*, arXiv:2602.16763.

Anthropic. 2025. Claude Opus 4.6 system card. Accessed: 2026-05-17.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, Volume 1 (Long Papers), pages 1304–1313, New Orleans, Louisiana. Association for Computational Linguistics.

Soham Bhattacharjee, Baban Gain, and Asif Ekbal. 2024. Domain dynamics: Evaluating large language models in English-Hindi translation. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 169–177, AU-KBC Research Centre, Chennai, India. NLP Association of India (NLPAI).

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Association for Computational Linguistics.

Sung-Kwon Choi, Gyu-Hyeun Choi, and Youngkil Kim. 2018. Automatic evaluation of English-to-Korean and Korean-to-English neural machine translation systems by linguistic test points. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

Herbert H. Clark and Susan E. Haviland. 1977. Comprehension and the given-new contract. In Roy O. Freedle, editor, *Discourse Production and Comprehension*, volume 1 of *Discourse Processes: Advances in Research and Theory*, chapter 1, pages 1–40. Ablex Publishing Corporation, Norwood, NJ.

William S. Cleveland. 1981. LOWESS: A program for smoothing scatterplots by robust locally weighted regression. *The American Statistician*, 35(1):54.

DeepSeek-AI. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence.

Daniel Deutsch, Eleftheria Briakou, Isaac Rayburn 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. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12257–12284, Vienna, Austria. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Patrick Fernandes, Kayo Yin, Emmy Liu, André Martins, and Graham Neubig. 2023. When does translation require context? a data-driven, multilingual exploration. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 606–626, Toronto, Canada. Association for Computational Linguistics.

Patrick Fernandes, Kayo Yin, Graham Neubig, and André F. T. Martins. 2021. Measuring and increasing context usage in context-aware machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6467–6478, Online. Association for Computational Linguistics.

Markus Freitag, 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 *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Gemma Team. 2026. Gemma 4: Byte for byte, the most capable open models. Google DeepMind Blog.

Google DeepMind. 2026. Gemini 3 flash. Accessed: 2026-05-16.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Comput. Linguist.*, 12(3):175–204.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

M. A. K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*, 1st edition. Routledge, London.

Christian Hardmeier. 2012. Discourse in statistical machine translation: A survey and a case study. *Discours*, 11.

Jerry R. Hobbs. 1978. Resolving pronoun references. *Lingua*, 44(4):311–338.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA. Association for Computational Linguistics.

Marcin Junczys-Dowmunt. 2019. Microsoft translator at WMT 2019: Towards large-scale document-level neural machine translation. In *Proceedings of the*

Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), pages 225–233, Florence, Italy. Association for Computational Linguistics.

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

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451, Singapore. Association for Computational Linguistics.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.

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

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, and 3 others. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Alon Lavie, 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 *Proceedings of the Tenth Conference on Machine Translation*, pages 436–483, Suzhou, China. Association for Computational Linguistics.

Minjae Lee, Youngbin Noh, and Seung Jin Lee. 2025. A testset for context-aware LLM translation in Korean-to-English discourse level translation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1632–1646, Abu Dhabi, UAE. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pretraining for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Shushen Manakhimova, Vivien Macketanz, Eleftherios Avramidis, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian Möller. 2024. Investigating the linguistic performance of large language models in machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 355–371, Miami, Florida, USA. Association for Computational Linguistics.

Sameen Maruf, André F. T. Martins, and Gholamreza Haffari. 2019. Selective attention for context-aware neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3092–3102, Minneapolis, Minnesota. Association for Computational Linguistics.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium. Association for Computational Linguistics.

Wafaa Mohammed and Vlad Niculae. 2024. On measuring context utilization in document-level MT systems. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1633–1643, St. Julian's, Malta. Association for Computational Linguistics.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium. Association for Computational Linguistics.

Cloud HyperCLOVA X Team NAVER. 2026. Hyperclova x 32b think. *Preprint*, arXiv:2601.03286.

Dojun Park and Sebastian Padó. 2024. Multidimensional machine translation evaluation: Model evaluation and resource for Korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11723–11744, Torino, Italia. ELRA and ICCL.

Matt Post and Marcin Junczys-Dowmunt. 2024. Escaping the sentence-level paradigm in machine translation. *Preprint*, arXiv:2304.12959.

Ellen F. Prince. 1981. Toward a taxonomy of given-new information. In P. Cole, editor, *Syntax and semantics: Vol. 14. Radical Pragmatics*, pages 223–255. Academic Press, New York.

Lorenzo Proietti, Stefano Perrella, Vilém Zouhar, Roberto Navigli, and Tom Kocmi. 2025. Estimating machine translation difficulty. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 24261–24285, Suzhou, China. Association for Computational Linguistics.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Vikas Raunak, Tom Kocmi, and Matt Post. 2023a. Evaluating metrics for document-context evaluation in machine translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 812–814, Singapore. Association for Computational Linguistics.

Vikas Raunak, Amr Sharaf, Yiren Wang, Hany Awadalla, and Arul Menezes. 2023b. Leveraging GPT-4 for automatic translation post-editing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12009–12024, Singapore. Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André F. T. Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

Aaditya Singh, Adam Fry, Adam Perelman, and 1 others. 2026. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Zewei Sun, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Shujian Huang, Jiajun Chen, and Lei Li. 2022. Rethinking document-level neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3537–3548, Dublin, Ireland. Association for Computational Linguistics.

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Giorgos Vernikos, 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.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.

Rachel Wicks and Matt Post. 2023. Identifying context-dependent translations for evaluation set production. In *Proceedings of the Eighth Conference on Machine Translation*, pages 452–467, Singapore. Association for Computational Linguistics.

Vilém Zouhar, Wenda Xu, Parker Riley, Juraj Juraska, Mara Finkelstein, Markus Freitag, and Daniel Deutsch. 2026. Generating difficult-to-translate texts. In *Proceedings of the First Workshop on Multilingual Multicultural Evaluation*, pages 204–219, Rabat, Morocco. Association for Computational Linguistics.

Vilém Zouhar, Peng Cui, and Mrinmaya Sachan. 2025. How to select datapoints for efficient human evaluation of nlg models? *Transactions of the Association for Computational Linguistics*, 13:1789–1811.

Appendix

A DDP Computation

A.1 Validation against Gold Coreference

To assess how much of true inter-sentential dependency our two-anchor formulation captures, we compare DDP against gold coreference chains from OntoNotes 5.0 (Hovy et al., 2006). We select 80 English documents from the news (n=45) and narrative (n=35) portions whose genre overlaps with our evaluation data, yielding 2,143 segments.

Gold-distance computation. For each segment s_i, we extract all mention spans whose coreference chain has at least one antecedent in a prior segment s_j ($j < i$). The gold dependency distance is defined symmetrically to DDP:

$$d_{\text{gold}}(s_i) \;=\; \max_{m \in M_i} (i - j_m), \tag{5}$$

where M_i is the set of mentions in s_i with a prior antecedent and j_m is the segment index of the most distant reachable chain link. Segments without prior-chain mentions receive $d_{\text{gold}} = 0$.

Agreement. Segment-level DDP correlates with gold distance at Pearson $r = 0.81$ (Spearman $\rho = 0.78$), with stronger agreement on news ($r = 0.86$) than narrative ($r = 0.74$). The narrative gap reflects the denser presence of bridging anaphora and lexical chains, which fall outside our anchor inventory.

One-sided error. Consistent with its construction, DDP underestimates gold distance in 86.4% of segments and matches it in 12.8%. Overestimation occurs in 0.8% and is fully attributable to NER false positives that introduce spurious early mentions. This confirms the lower-bound guarantee in practice.

Implications. Both empirical claims in this paper, that current benchmarks undersample high-dependency segments and that human-metric divergence grows with them, depend on the presence rather than the exhaustive enumeration of high-dependency cases. Since DDP only undercounts, segments it identifies as high-dependency are also high-distance under gold annotation. The moderate gap on narrative further suggests that the divergence pattern may be *underestimated* by DDP-stratified analysis, as some segments treated as low-DDP carry hidden bridging or lexical-chain

Domain	spaCy (trf)				Flair				spaCy (sm, weak)			
	μ	Med	Max	%=0	μ	Med	Max	%=0	μ	Med	Max	%=0
News	9.7	4	52	28.9	9.2	4	50	32.9	11.2	4	50	26.2
Social	3.8	0	166	62.4	4.1	0	130	63.0	4.4	0	82	60.9
Literary	52.8	44	171	13.1	52.2	33	200	16.0	44.6	32	162	15.5

Table 4: Per-domain DDP statistics under spaCy (`en_core_web_trf`), Flair, and a deliberately weaker extractor (`en_core_web_sm`). The Literary > News > Social ordering and the ordering of ddp =0 shares are preserved under all three, including the weakened pipeline.

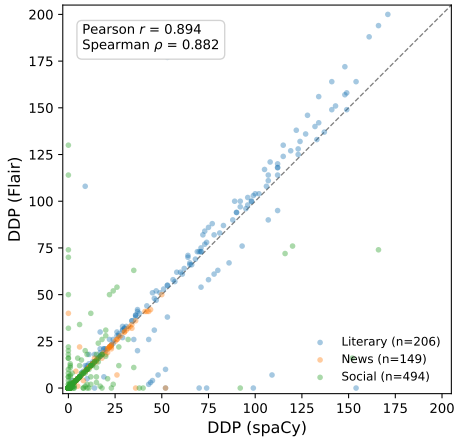


Figure 4: Segment-lveel DDP from spaCy vs Flair on WMT24++. Each point is a unique source segment, colored by domain. Most segmetns lie on or near the identity line.

Domain	WMT24++		WMT25	
	μ	max	μ	max
News	9.7	52	10.6	41
Social	3.8	166	21.3	186
Literary	52.8	171	222.6	482

Table 5: Mean and maximum segment-level DDP scores across domains in WMT24++ and WMT25.

dependencies that would reinforce the observed effect if measured. Extending the anchor inventory to bridging and lexical cohesion is a natural direction for future work.

A.2 Worked examples on WMT24++

We illustrate DDP computation on three documents from WMT24++, one per domain in Figure 7. Documents are chosen to display the operation of both anchor rules, and their DDP statistics are close to or above the respective domain means, allowing the figures to show meaningful anchor activity without depicting extreme tails.

A.3 Ablation: Anchor Extraction Tool

We replicate DDP with Flair NER and POS tagging (Akbik et al., 2019) in place of spaCy. Sentence segmentation, dependency parsing for head-noun extraction, and downstream distance computation are held constant to isolate the effect of the NER and POS components.

Agreement on DDP values. Figure 4 plots segment-level DDP from the two pipelines against each other. The values are strongly correlated (Pearson $r = 0.89$, Spearman $\rho = 0.88$), with 69% of segments receiving identical DDP and 85% agreeing within ± 5. Disagreement concentrates in segments with longer referential chains, where small differences in entity recognition propagate through chain accumulation.

Domain-level statistics. Table 4 reports per-domain DDP under both tools. Domain means shift by less than one point in absolute terms, and the ordering Literary > News > Social is preserved. The qualitative findings of Section 3.3, that benchmarks skew toward low-DDP segments and that domain labels conflate within-domain variation, hold under both tools.

A deliberately weaker tagger. Agreement between two strong taggers leaves open whether the ordering depends on tagger quality. We therefore repeat the replication with `en_core_web_sm`, holding sentence segmentation and distance computation constant. Tagger quality affects the two rules asymmetrically. Under Rule ①, a missed mention can only shorten a chain, since a first mention cannot be invented. Under Rule ②, a pronoun is anchored to the *most recent* compatible antecedent, so a dropped intervening entity moves the antecedent further back and lengthens the distance. Neither error produces the Literary > News > Social ordering, though individual segments may move in either direction.

The ordering is preserved (Table 4). Segment-level agreement with spaCy is high (Pearson $r =$

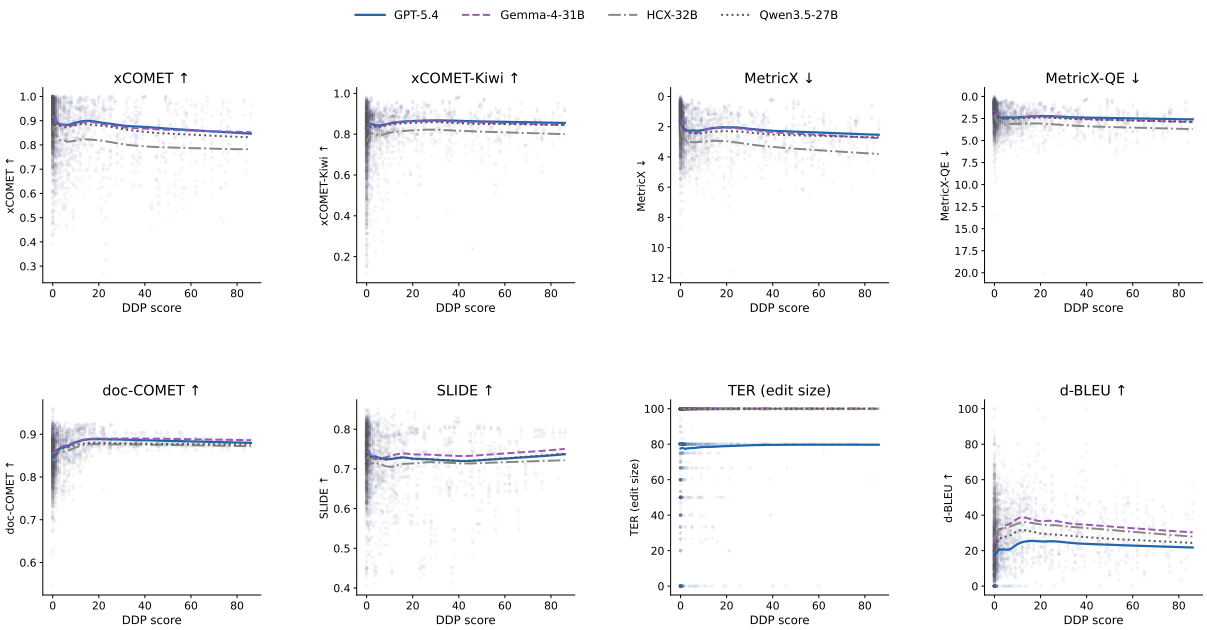


Figure 5: LOWESS-smoothed metric scores per model as a function of segment-level DDP (**En–Zh**, all injection strategies averaged; four models available). As in En–Ko (Figure 3), metric scores remain largely flat across the DDP spectrum and no model shows consistent improvement or degradation as discourse dependency increases, confirming that the insensitivity of automatic metrics to DDP is not specific to Korean.

0.84, Spearman $\rho = 0.88$), with 73.4% exact agreement and 83.8% within ± 5, and the ordering of $\mathrm{ddp} = 0$ shares is also retained. Literary means fall (52.8 $\rightarrow$ 44.6), where long chains offer the most links to miss, while news and social rise slightly (9.7 $\rightarrow$ 11.2, 3.8 $\rightarrow$ 4.4) through the Rule ② effect. DDP decreases on 14.7% of segments and increases on 11.9%. The lower bound of Section 3.2 therefore holds strictly for entity re-mentions and approximately for pronominal chains, which does not affect the domain-level results but bears on certifying individual segments as high-dependency.

A.4 Cross-lingual Consistency: En–Zh

Since DDP is computed from the English source, the same scores apply to any target language. We report automatic metric results for En–Zh on the four models for which outputs were available (GPT-5.4, GEMMA-4-31B, QWEN3.5-27B, HYPERCLOVA-X-SEED-THINK-32B). Metric scores remain largely flat across the DDP spectrum as in En–Ko (Figure 5), so metric insensitivity to discourse dependency is not specific to Korean. HYPERCLOVA-X-SEED-THINK-32B ranks lower here than in En–Ko, consistent with its Korean-specialized training.

Across injection strategies (Figure 11), most metrics again fail to separate them, and the gap between APE and 👤✎ HPE is smaller than in En–Ko. Two readings are available. The evaluated models, HYPERCLOVA-X-SEED-THINK-32B aside, have stronger Chinese coverage, which would narrow the true APE–👤✎ HPE gap below what current metrics can resolve; alternatively, Chinese discourse properties such as pervasive zero-anaphora and topic-drop may leave fewer surface traces for metrics to detect. Distinguishing the two requires human evaluation, and whether human preferences follow the same DDP-stratified pattern across language pairs remains open.

B Experimental Details

B.1 Dataset statistics

We use the English-Korean subset of WMT24++ (Deutsch et al., 2025) for all experiments. The released data spans three domains (literary, news, social) across 59 documents and 849 unique source segments. Table 6 reports the per-domain breakdown along with DDP statistics computed on the English source. The three domains differ sharply in their dependency profiles: 62.3% of social segments have $\mathrm{ddp} = 0$ compared with 13.1% of literary segments, and the literary mean (52.8) is over an order of magnitude larger than the social mean (3.8). News falls between the two extremes (mean 9.7, 28.9% at $\mathrm{ddp} = 0$). All documents in the subset contain at least two segments, so no further filtering is applied.

Domain	Docs	Segs	ddp				%=0
			mean (std)	med	p95	max	
News	17	149	9.7 (12.4)	4	35	52	28.9
Social	34	494	3.8 (14.3)	0	17	166	62.3
Literary	8	206	52.8 (46.9)	44	141	171	13.1
Total	59	849	16.7 (33.2)	1	101	171	44.5

Table 6: Statistics of the WMT24++ En–Ko subset used in our experiments. %=0 reports the share of segments with $\mathrm{ddp} = 0$, for which no prior discourse context is required. Social texts skew strongly toward self-contained segments, while literary texts span a wide range of dependencies.

B.2 Context Injection Strategy: Construction

$\mathrm{APE}_{\mathrm{rel}}$. We encode all source segments $\{s_j\}_{j \neq i}$ in each document using LaBSE (Feng et al., 2022), a multilingual sentence encoder that produces language-agnostic representations. The current segment s_i is excluded from the index to prevent self-retrieval. Given s_i as query, we select the top-k segments by cosine similarity and pair them with their target segments. Retrieved pairs are reordered by their original document position before being injected into the prompt, preserving sequential structure across conditions.

$\mathrm{APE}_{\mathrm{exp}}$. For each document, we prompt CLAUDE OPUS 4.6 (Anthropic, 2025) to extract structured annotations covering three discourse dimensions that affect target-language editing decisions:

- **Genre and domain**: text type and subject domain (e.g., *news report*, *literary dialogue*).
- **Participant relationships**: roles and social relationships between discourse participants where identifiable (e.g., *interviewer-interviewee*, *narrator-character*).
- **Register and formality**: expected formality level and register conventions, such as honorific speech level choices in Korean and written vs. colloquial distinctions in Chinese.

Annotations are produced in two stages. We first prompt CLAUDE OPUS 4.6 on the full source document to generate a JSON annotation per document via greedy decoding. It is not among the evaluated models, avoiding circular dependency between annotation and evaluation. A trained annotator then reviews each annotation and corrects errors, ambiguities, and cases where the model fails to identify participant relationships or register conventions. The annotation guidelines and inference prompt are in Figure 8.

B.3 Context Injection Prompts

A unified prompt structure is used across all conditions (Figure 9). The system prompt is identical and contains only a role definition, so any performance difference reflects context content rather than prompt formulation. All task instructions and quality guidelines are placed in the user prompt, where they are processed jointly with the provided context.

The context block is the only element that varies across conditions. $\mathrm{APE}_{\mathrm{seg}}$ receives no context. $\mathrm{APE}_{\mathrm{seq}}$, $\mathrm{APE}_{\mathrm{rel}}$, and $\mathrm{APE}_{\mathrm{full}}$ each receive a source and target context block populated according to the curation strategy. $\mathrm{APE}_{\mathrm{exp}}$ receives a structured document-information block in place of raw context, with the register instruction adjusted to signal that the relevant information is supplied explicitly rather than requiring inference. No instruction is given regarding the relevance or quality of the provided context, so the model's sensitivity to context content can be observed without interference.

B.4 Pilot Study: Choice of k

The number of retrieved segments k controls the amount of context injected in $\mathrm{APE}_{\mathrm{seq}}$ and $\mathrm{APE}_{\mathrm{rel}}$. To determine an appropriate value, we conduct a pilot study on a held-out subset of ≈270 segments from WMT25 En–Ko (Kocmi et al., 2025), using GEMMA-4-31B (Gemma Team, 2026) as the primary model with $k \in \{1, 3, 5, 10\}$.

We select k based on two criteria. First, we measure the divergence between $\mathrm{APE}_{\mathrm{seq}}$ and $\mathrm{APE}_{\mathrm{rel}}$

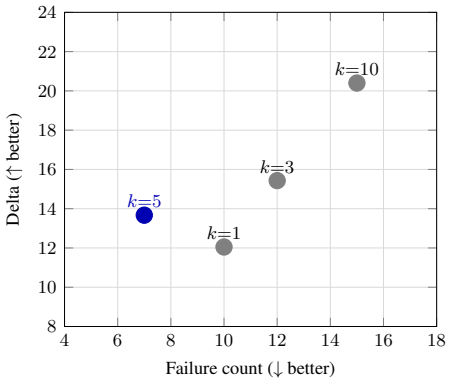


Figure 6: Delta vs. Failure count per k. k=5 (blue) achieves the lowest failure count with moderate delta.

via TER:

$$\Delta\text{TER}(k) = \left|\text{TER}_{\text{seq}} - \text{TER}_{\text{rel}}\right| \quad (6)$$

A larger $\Delta\text{TER}(k)$ indicates that the two injection strategies produce meaningfully different outputs, exposing the contrast between sequential and relevance-based selection. Second, we track the generation failure rate, defined as the proportion of segments with TER > 100.

We select the k that maximizes $\Delta\text{TER}(k)$ subject to a low failure rate. As shown in Figure 6, k=10 yields the largest divergence but also the highest failure rate, making it unreliable. k=5 achieves the best trade-off: sufficient divergence between strategies with the lowest hallucination rate across all settings. We therefore fix k=5 for all subsequent experiments.

B.5 Human Evaluation Details

Sampling Criteria. To ensure that human evaluation focuses on documents where context curation produces measurable variation, we select documents based on inter-condition divergence rather than uniform random sampling. For each document, we compute the mean pairwise TER distance among all curation strategy outputs, excluding segments with TER > 100 as generation failures (Raunak et al., 2023b). Documents are ranked by mean divergence in descending order, and the top-k documents per domain are retained as candidates. The final subset is stratified by domain and capped at ≈500 segments per assumption, yielding 1,000 segments in total, with no overlap between **Size** and **Selection** subsets. Within each selected document, APE outputs are anonymized and presented in randomized order to mitigate position bias.

Annotation Platform. We use Label Studio as the annotation platform.[2] Each task presents the current segment alongside its immediately preceding and following source segments for discourse context. Annotators may freely navigate between segments within the same document for reference and may revise their rankings at any time before submission.

Annotators. We recruit four professional translators with at least three years of experience in En-Ko translation. Each annotator evaluated ≈500 segments. Annotators were compensated at a competitive industry rate and consented to the release of their annotations.

Annotation Guidelines. Annotators were instructed to rank the four candidate translations from 1 (best) to 4 (worst) based on overall translation quality, with ties permitted when candidates were judged indistinguishable. They were asked to consider not only fluency and adequacy, but also discourse attributes, and were encouraged to consult surrounding segments when assessing referential coherence.

Inter-Annotator Agreement. IAA is measured using Kendall's τ computed between the two independent rankings per segment, averaged across all evaluated segments. The overall τ was 0.37, consistent with prior relative ranking evaluations in MT (Callison-Burch et al., 2007), and indicates moderate agreement.

Remuneration. Professional translators are compensated at competitive industry rates commensurate with their expertise and time. All annotators provided informed consent for the use and publication of their annotations.

B.6 Inference Overhead

To assess the computational overhead of each context injection strategy, we record three efficiency metrics for every model–condition pair: (1) **input tokens**, which vary directly with context size and injection strategy; (2) **output tokens**, reflecting editing verbosity; and (3) **latency**, measured as wall-clock time per segment from prompt submission to response completion.

Input token counts are deterministic given the injection strategy and document, and are reported as averages over the full evaluation set. Latency

[2] https://labelstud.io

Model	Strategy	Input (tok)	Input (ratio)	Output (tok)	Latency (sec)	Latency (ratio)
GPT-5.4	APE_{seg}	216	1.0×	68	–	–
	APE_{seq}	644	3.0×	60	–	–
	APE_{rel}	730	3.4×	62	–	–
	APE_{exp}	347	1.6×	72	–	–
	APE_{full}	10,688	49.5×	62	–	–
GEMINI-3-FLASH	APE_{seg}	208	1.0×	32	–	–
	APE_{seq}	636	3.1×	32	–	–
	APE_{rel}	718	3.5×	32	–	–
	APE_{exp}	343	1.6×	32	–	–
	APE_{full}	17,556	84.4×	34	–	–
DEEPSEEK-V4-PRO	APE_{seg}	218	1.0×	880	–	–
	APE_{seq}	686	3.1×	927	–	–
	APE_{rel}	781	3.6×	952	–	–
	APE_{exp}	364	1.7×	934	–	–
	APE_{full}	12,080	55.4×	906	–	–
GEMMA-4-31B	APE_{seg}	243	1.0×	57	2.90	1.00×
	APE_{seq}	665	2.7×	54	3.26	1.12×
	APE_{rel}	747	3.1×	56	3.40	1.17×
	APE_{exp}	372	1.5×	57	3.27	1.13×
	APE_{full}	10,371	42.7×	57	3.89	1.34×
HYPERCLOVA-X-SEED-THINK-32B	APE_{seg}	220	1.0×	47	4.09	1.00×
	APE_{seq}	613	2.8×	44	7.45	1.82×
	APE_{rel}	690	3.1×	46	7.90	1.93×
	APE_{exp}	343	1.6×	46	2.81	0.69×
	APE_{full}	8,947	40.7×	49	13.18	3.22×
QWEN3.5-27B	APE_{seg}	222	1.0×	53	2.49	1.00×
	APE_{seq}	631	2.8×	51	2.79	1.12×
	APE_{rel}	708	3.2×	53	2.95	1.18×
	APE_{exp}	358	1.6×	54	2.60	1.04×
	APE_{full}	9,471	42.7×	56	12.33	4.95×

Table 7: Inference overhead per model and curation strategy, averaged over all evaluation segments (En–Ko). Input/latency ratios are relative to APE_{seg} (null baseline). Latency for API-based models (GPT-5.4, GEMINI-3-FLASH, DEEPSEEK-V4-PRO) includes network round-trip time and is not directly comparable to open-weight models (reported as –). APE_{full} requires 40–84× more input tokens than APE_{seg}, yet provides no consistent improvement in translation quality (Section 5).

is measured on a single NVIDIA RTX PRO 6000 Blackwell GPU under a fixed batch size of one to isolate per-segment processing time, averaged over three runs to reduce variance. API-based models (GPT-5.4, GEMINI-3-FLASH, DEEPSEEK-V4-PRO) include network round-trip time and are reported separately from open-weight models to avoid confounding infrastructure differences. Models that do not support a `seed` parameter are noted where applicable. Results for these models may exhibit minor run-to-run variation.

Table 7 shows that APE_{full} incurs the highest input token cost across all models, requiring 40 - 84× more tokens than APE_{seg}, yet yields no consistent improvement in translation quality (Section 5). APE_{exp} is the most token-efficient context strategy (≈1.6× overhead), as structured declarative knowledge is considerably shorter than raw segments, but it consistently underperforms APE_{seq} in human evaluation. APE_{seq} and APE_{rel} incur comparable input overhead (≈3×), with APE_{rel} marginally higher due to embedding-based retrieval of longer segments.

In terms of latency, open-weight models show modest increases from APE_{seg} to APE_{seq} (≈1.1 - 1.8×), while APE_{full} causes substantially higher latency on some models (up to 5× for QWEN3.5-27B). Notably, DEEPSEEK-V4-PRO generates substantially more output tokens than other models (880 - 952 vs. 32 - 72), likely reflecting tokenizer differences in Korean text generation. Taken together, APE_{seq} offers the best trade-off between context effectiveness and computational cost.

C Additional Results

C.1 Automatic metric results across DDP

Figure 10 shows automatic metric scores across the DDP spectrum for all strategies in Size and Selection. Two consistent patterns emerge across all metrics and both assumptions. First, strategy scores remain largely flat as DDP increases: the curves for APE_{seg}, APE_{seq}, and APE_{full} (or APE_{rel} and APE_{exp}) run nearly parallel throughout the dependency spectrum, indicating that automatic metrics do not register the growing contextual difficulty reflected in human judgments (Figure 2). Second, all APE strategies cluster tightly together regardless of DDP, making it impossible to distinguish between injection strategies on the basis of automatic scores alone. Notably, HPE scores also remain largely stable across DDP, suggesting that automatic metrics assess professional post-edits similarly regardless of how contextually demanding a segment is.

Two observations are worth noting. Document-level metrics (doc-COMET, SLIDE, d-BLEU) do not provide more reliable signal than segment-level metrics. doc-COMET shows near-identical scores across all strategies throughout the DDP spectrum, while SLIDE shows a counter-intuitive upward trend for APE strategies in Size as DDP grows, contradicting human judgments. In terms of post-edit volume measured by TER, its sharp spike at low DDP reflects high generation failure rates in context-heavy conditions (Table 3), and its subsequent stabilization indicates that models produce more conservative outputs as discourse dependency increases, rather than improving in quality.

C.2 Qualitative case studies

Table 8 illustrates why translation difficulty is relational rather than intrinsic. What makes a segment hard to translate faithfully is how far into the document one must look to find the information it requires.

NER inconsistency. The character name *Aquilo* is transliterated as *아킬로*, *아쿠일로*, or *아퀼로* depending on the strategy and segment, reflecting the absence of a stable anchor when context is limited or mismatched. Since DDP is computed from the source text, it identifies such segments as high-dependency regardless of whether the model resolves them correctly.

Pronoun and coreference failure. In Seg. 925 (ddp = 54), *they* refers to Ivory, a non-binary character for whom Korean has no direct equivalent pronoun. All APE strategies default to the plural *그들*, while the HPE resolves the referent by proper pronoun (*아이보리는*), drawing on character information established far earlier in the narrative. Likewise, *their silent friend* is contextually Princess Kari (*카리 공주*), but APE strategies render it generically (*말없는 친구*), while the HPE supplies the name directly. Both cases exceed what any fixed-window context strategy can capture, and both are precisely the cases DDP should but cannot flag as contextually difficult.

Register inconsistency. All APE strategies shift between formal (*-었습니다*, *-이에요*) and informal (*-었다*) endings across segments without discourse motivation, and frequently fail to reflect the status relationship between Ivory and Aquilo that a consistent register would encode. This pattern persists across all context sizes and selection criteria, suggesting that register coherence requires document-wide grounding that current LLMs do not reliably achieve even when context is provided.

ID	ddp	**(a) Literary** *(Forever Snow)*
871	0	Prologue
872	2	"Please!" Queen Eirwen$^{[0]}$ begged, "don't send her$^{[\rightarrow 1]}$ away! She's$^{[\rightarrow 1]}$ just a child!" The Ice King$^{[0]}$ looked down at his$^{[\rightarrow 1]}$ wife from his$^{[\rightarrow 1]}$ throne, his$^{[\rightarrow 1]}$ royal advisor next to him$^{[\rightarrow 1]}$. She$^{[\rightarrow 2]}$ had given birth to a girl. He$^{[\rightarrow 2]}$ scowled at the thought.
873	6	The girl$^{[0]}$ in question looked a lot like her$^{[\rightarrow 1]}$ mother; pale blue eyes and white hair decorated with blue strands. The Ice King$^{[4]}$ thought a daughter made him$^{[\rightarrow 5]}$ look weak. After all, he$^{[\rightarrow 6]}$ had only had sons up until this point.
874	11	"The only options you have are to kill the child or send it$^{[\rightarrow 1]}$ out to the wilderness. This is her$^{[\rightarrow 4]}$ punishment for your wrongdoing, Eirwen$^{[9]}$." He$^{[\rightarrow 10]}$ spat out her$^{[\rightarrow 5]}$ name as if it$^{[\rightarrow 2]}$ was poison. He$^{[\rightarrow 11]}$ always did.
876	6	The South$^{[\rightarrow 2]}$ was desperate, and their$^{[\rightarrow 2]}$ royal family$^{[0]}$ would do anything to stop the bloodshed. They$^{[\rightarrow 2]}$ offered their only princess to be queen of the North$^{[\rightarrow 4]}$ and bear a son to be the future king. Then, the South$^{[\rightarrow 4]}$ would join with the North$^{[\rightarrow 5]}$ to make one kingdom. The North$^{[\rightarrow 6]}$ accepted, and Princess Eirwen$^{[0]}$ married Prince Akull$^{[0]}$.

ID	ddp	**(b) News** *(seattle_times.800119, "How to find out if you're flying on a Boeing 737 MAX")*
132	0	How to find out if you're flying on a Boeing 737 MAX$^{[0]}$
133	2	For most travelers, the aircraft model isn't a deciding factor when booking a flight. But with the majority of Boeing 737 MAX$^{[\rightarrow 1]}$ 9 jets grounded around the country after an Alaska Airlines$^{[0]}$ fuselage blowout on Jan. 6, some prospective passengers may want to know how to tell what type of plane they'll$^{[\rightarrow 2]}$ be on and which models are the safest.
134	1	The airplane model is typically noted in the booking information for a flight. On Google Flights$^{[0]}$, it's$^{[\rightarrow 1]}$ included in the flight details when a listing is expanded, and it's$^{[\rightarrow 1]}$ usually listed on individual airlines' reservation pages. If, for some reason, the plane type isn't apparent on Google Flights$^{[1]}$ or the reservation page, sites like Expert Flyer$^{[0]}$ and Seat Guru$^{[0]}$ aggregate flight information like seat maps and aircraft type.
136	3	Another useful resource the FAA$^{[0]}$ links to is the Boeing Worldwide Statistical Summary of Commercial Jet Airplane Accidents$^{[0]}$ 1959-2022, which assesses plane-type safety by breaking down accidents based on the aircraft model. According to the report, the Boeing 787$^{[5]}$ and Airbus A350$^{[0]}$ are the aircraft models with the fewest total hull losses. There are a few additional models in the report that have never experienced a hull loss, including the double-decker behemoth Airbus A380$^{[0]}$. But those$^{[\rightarrow 3]}$ models had accumulated fewer than 1 million departures at the time of the report.
137	12	The report also includes a comparison of hull loss accident rates per million departures, which helps account for the fact that some of the models are more common or have been around longer than others. The Airbus A310$^{[0]}$, which was introduced in 1983, has the highest rate of hull losses among the models that are still in service as passenger aircraft. The Airbus 320 family$^{[0]}$, which includes the A321$^{[0]}$ and A319neo$^{[0]}$, has the lowest rate. The Boeing 737 MAX$^{[\rightarrow 12]}$ line has a rate of 1.48.

ID	ddp	**(c) Social** *(MLS game reactions thread, 1121113858483918 72)*
397	0	Absolute rocket from Stroud$^{[0]}$ on a perfect half turn to goal. #DCU$^{[0]}$ #MLS$^{[0]}$ #MastodonFC$^{[0]}$
398	1	Really tough not to buy that the league tells officials to help Miami$^{[0]}$ when that's$^{[\rightarrow 1]}$ not a penalty. #DCU$^{[\rightarrow 1]}$ #MLS$^{[\rightarrow 1]}$ #mastodonfc
399	2	Probably Pirani's$^{[0]}$ best game this season, but he$^{[\rightarrow 1]}$ still just needs to be better. Like the aggressive call to sub him$^{[\rightarrow 1]}$ out for an offensive minded player like Fletcher$^{[0]}$. #DCU$^{[\rightarrow 2]}$ #MLS$^{[\rightarrow 2]}$ #mastodonfc
400	4	Amazing how there's two PKs #DCU$^{[\rightarrow 3]}$ should have had and got neither. The fix is so in for Miami$^{[0]}$. #MLS$^{[\rightarrow 4]}$ #mastodonfc
401	5	Well, perhaps ball doesn't lie. What an awful give away by Klich$^{[0]}$. Can't win when your$^{[\rightarrow 1]}$ DP plays that sloppy. #DCU$^{[\rightarrow 4]}$ #MLS$^{[\rightarrow 5]}$ #mastodonfc
403	7	Very curious to see how this$^{[\rightarrow 6]}$ #DCU$^{[\rightarrow 5]}$ team responds to a game that's best forgotten. They've$^{[\rightarrow 6]}$ had a bright start and sometimes bad games just happen, but they$^{[\rightarrow 6]}$ looked completely outclasses as the game went on. #MLS$^{[\rightarrow 7]}$ #mastodonfc
404	0	@user7 who give a crap, go Badgers Women's hockey$^{[0]}$! Hahaha

Figure 7: DDP computation on three example documents from WMT24++, one per domain. Some segments are omitted for space. Green marks named-entity anchors (Rule ①); blue marks pronominal anchors (Rule ②). Superscripts indicate the distance to the antecedent in sentences, with [0] denoting first mentions; segment-level DDP is the maximum anchor distance within the segment.

	`seg_id=923` (ddp = 6)	`seg_id=924` (ddp = 53)	`seg_id=925` (ddp = 54)
Source	They opened the door, expecting the normal sight. Paint splotches everywhere, pencils and papers scattered across the room, easels covering up every last inch of wall. Unexpectedly, there was a different sight to be found. It was all clean, and there was a red velvet background and a stool in front of it. On the stool was Crown Prince Aquilo, getting his portrait done.	"Ah, good morning, Ivory! Care to watch your future king get his portrait done?" Aquilo inquired, winking at Ivory.	"Sorry, princey," they said, partially ignoring him and searching for the things that their silent friend wanted. "I'm just here to get art supplies."
MT (GEMINI-1)	그들은 평소의 광경을 예상하며 문을 열었다. 사방에 흩어진 페인트 얼룩, 방바닥에 어지럽게 널린 연필과 종이들, 벽의 모든 공간을 차지한 이젤들. 그러나 예상치 못한 광경이 눈앞에 펼쳐졌다. 방은 깨끗했고, 붉은 벨벳 배경과 그 앞에 놓인 의자가 있었다. 의자에는 아퀼로 왕세자 가 앉아 초상화를 그리는 중이었다.	"아, 좋은 아침이야, 아이보리! 네 미래의 왕이 초상화를 그리는 걸 구경할래?" 아퀼로 가 아이보리에게 윙크하며 물었다.	"미안해, 꼬마야," 그들은 그를 반쯤 무시하며 조용한 친구가 원하는 물건들을 찾았다. "나 그냥 미술 용품 사러 온 거야."
HPE	그는 평범한 광경을 기대하며 문을 열었다. 곳곳에 물감이 얼룩져 있고 연필과 종이가 방 전체에 흩어져 있으며 이젤이 벽 구석구석을 뒤덮고 있을 것이라 생각했다. 그러나 예상과는 다른 광경이 펼쳐졌다. 모든 것이 깨끗하게 정리되어 있었고 붉은 벨벳을 배경으로 앞에 의자가 놓여 있었다. 그 의자에는 아퀼로 왕세자 가 앉아 있었고, 그의 초상화가 그려지고 있었다.	"아, 좋은 아침이야, 아이보리! 네 미래 왕의 초상화가 완성되는 모습을 보고 싶나?" 아퀼로 가 아이보리에게 윙크하며 물었다.	"죄송합니다, 왕자님" 아이보리는 그를 약간 무시하고 카리 공주 가 원하는 것을 찾으며 말했다. "저는 그저 미술용품을 구하러 온 것뿐입니다."
APE_{seg}	그들은 평범한 광경을 기대하며 문을 열었습니다. 사방에 페인트 자국이 가득하고, 연필과 종이가 방 안에 흩어져 있으며, 벽에는 이젤이 가득 차 있었습니다. 하지만 예상치 못한 다른 광경이 펼쳐졌습니다. 모든 것이 깨끗하게 정리되어 있었고, 빨간 벨벳 배경과 그 앞에 놓인 의자가 있었습니다. 그 의자 위에는 아킬로 왕자 가 초상화를 그리고 있었습니다.	"아, 좋은 아침이에요, 아이보리! 미래의 왕이 초상화를 그리는 걸 보실래요?" 아퀼로 가 아이보리에게 윙크하며 물었다.	"미안해, 왕자님." 그들은 그를 부분적으로 무시하며 말없는 친구가 원하는 것을 찾고 있었다. "그냥 미술 용품을 사러 왔어요."
APE_{seq}	그들은 평범한 광경을 기대하며 문을 **열었습니다**. 사방에 흩어진 페인트 자국, 연필과 종이, 벽마다 세워진 이젤들. 하지만 예상과 달리 완전히 정돈된 **모습이었습니다**. 빨간 벨벳 배경 앞에 의자가 놓여 있었고, 그 의자에 아쿠일로 왕세자 가 초상화를 그리고 **있었습니다**.	"아, 좋은 **아침이에요**, 아이보리! 내 미래의 왕이 초상화를 그리는 걸 **볼래요**" 아킬로 가 아이보리에게 윙크하며 **말했다**.	"**미안해요**, 왕자님." 그들은 그를 일부러 외면하며 말 없는 친구가 원하는 것을 찾으려 **했습니다**. "그냥 미술 용품을 찾으러 **왔어요**."
APE_{rel}	그들은 평범한 광경을 예상하며 문을 열었습니다. 페인트 자국이 사방에 있고, 연필과 종이가 방 안에 흩어져 있으며, 이젤이 벽 전체를 덮고 있는 모습이죠. 하지만 그들의 예상과 달리 완전히 다른 광경이 펼쳐졌습니다. 모든 것이 깔끔하게 정돈되어 있었고, 빨간 벨벳 배경 앞에는 의자가 놓여 있었습니다. 그리고 그 의자 위에는 아쿠일로 왕세자 가 초상화를 그리는 중이었습니다.	"아, 좋은 아침이에요, 아이보리! 미래의 왕이 초상화를 그리는 걸 보실래요?" 아킬로 가 아이보리에게 윙크하며 물었다.	"미안해요, 왕자님." 그들은 그를 부분적으로 무시한 채로 말없는 친구가 원하는 물건들을 찾았습니다. "그냥 미술 용품을 사러 왔어요."
APE_{exp}	그들은 평범한 광경을 기대하며 문을 열었지만, 예상치 못한 광경이 펼쳐졌다. 사방에 흩어진 페인트 얼룩, 연필과 종이들 대신, 모든 것이 깔끔했고 빨간 벨벳 배경 앞에는 의자가 놓여 있었다. 그 의자 위에 아킬로 왕세자 가 초상화를 그리고 있었다.	"아, 좋은 아침, 아이보리! 미래의 왕이 초상화를 그리는 걸 볼 생각 없냐?" 아킬로 가 윙크를 하며 말했다.	"미안해, 왕자님," 그들은 그를 부분적으로 무시한 채로 말없는 친구 가 원하는 것을 찾으려 애썼다. "그냥 미술 용품을 구하러 왔어."
APE_{full}	그들은 평소의 광경을 예상하며 문을 열었다. 사방에 흩어진 페인트 얼룩, 방 안에 어지럽게 널린 연필과 종이, 벽을 빈틈없이 메운 이젤들. 그러나 예상과 달리 눈앞에 펼쳐진 것은 전혀 다른 모습이었다. 방은 깨끗하게 정돈되어 있었고, 붉은 벨벳 배경 앞에 의자가 놓여 있었다. 그리고 그 의자에는 아퀼로 왕세자 가 앉아 초상화를 그리고 있었다.	"아, 좋은 아침, 아이보리! 네 미래의 왕이 초상화를 그리는 걸 구경할래?" 아퀼로 가 아이보리에게 윙크하며 물었다.	"미안해, 왕자님." 그들은 그를 반쯤 무시하며 조용한 친구가 원하는 것을 찾고 있다고 말했다. "나는 그냥 미술 용품을 사러 왔어."

Table 8: Qualitative comparison of five context injection strategies on three consecutive segments from a literary document (`doc_id`: `forever_snow_chunk_2`, `seg_id`: 923–925) translated by GEMINI-1 and post-edited by HYPERCLOVA-X-SEED-THINK-32B. DDP increases from 6 to 53 and 54 across segments. Highlights focus on NER anchors and pronoun resolution: correct NER mention or pronoun (consistent with HPE), incorrect or inconsistent NER transliteration, correct pronoun resolution, incorrect pronoun resolution, source pronoun requiring discourse-level resolution. **Bold blue** and **bold black** indicate formal and informal register respectively; inconsistency within a single translation reflects the failure to maintain discourse-level formality, a pattern observed across all injection strategies.

APE_{exp} Construction

System: You are a discourse analyst. Your task is to analyze a document and extract its key discourse attributes. Be concise and precise. Output ONLY valid JSON.

User:

```
The following source document belongs to the "{genre}" genre. Analyze its content and extract the
three discourse attributes listed below. Base your analysis solely on the document content.
[Source Document]
{src_doc}
Use the genre label to select the most appropriate text_type and domain from the examples provided,
then extract all attributes and output a JSON object.
Genre: {genre}
  text_type examples: {text_type_examples}
  domain examples:    {domain_examples}
Output the following JSON structure:
{
  "genre_and_domain": {
    "text_type": "<select from examples above>",
    "domain":    "<select from examples above>"
  },
  "participant_relationships": {
    "identified": <true | false>,
    "participants": [
      {
        "role": "<e.g. narrator, interviewer, character,
                 customer, agent>",
        "description": "<brief description if identifiable>"
      }
    ],
    "relationship_type": "<e.g. interviewer--interviewee,
                           narrator--character,
                           customer--agent,
                           none identifiable>"
  },
  "register_and_formality": {
    "overall_register": "<one of: formal | semi-formal
                         | informal | mixed>",
    "notes": "<any additional register observations
              relevant to translation>"
  }
}
Rules:
- Use the provided genre label as the primary guide for text_type and domain selection.
- You may use values outside the examples if none fits, but prefer the listed options.
- If participant relationships cannot be identified, set "identified" to false and leave
"participants" as an empty list.
- Do not infer information not present in the document.
- Output ONLY the JSON object, no preamble or explanation.
```

Genre-specific examples:

```
news:
  text_type: news report, investigative article, opinion piece, editorial, press release,
             interview, feature article
  domain:    politics, diplomacy, economics, finance, science, technology, environment,
             health, sports, culture, society, crime, international affairs, military, education
social:
  text_type: social media post, blog post, online comment, forum thread, product review,
             newsletter, personal essay
  domain:    lifestyle, travel, food, fashion, entertainment, gaming, parenting, health
             and wellness, personal finance, technology, politics, sports
literary:
  text_type: literary narrative, literary dialogue, short story, novel excerpt, poetry,
             drama, personal memoir, historical fiction
  domain:    coming-of-age, romance, historical, thriller, family, war, social commentary,
             philosophical, fantasy, biographical
```

Figure 8: Prompt template used for APE_{exp} construction.

APE$_{seq}$ / APE$_{rel}$

System: You are a professional post-editor. Your task is to revise the draft translation of the given source segment so that it is faithful to the source, fluent in {tgt_lang}, and coherent with the surrounding segments. Pay particular attention to consistency in terminology and named entities, stylistic and tonal coherence appropriate to the genre, formality and register as specified in the document information, and natural discourse flow including connectives and reference expressions. Output ONLY the corrected {tgt_lang} sentence between <pe> and </pe>.

User:

```
[Source Context]
{src_ctx}

[Target Context]
{tgt_ctx}

[{src_lang} Source]
{src_seg}

[{tgt_lang} Draft Translation]
{tgt_seg}

MUST output in this format:
<pe>{tgt_lang} corrected sentence only</pe>
```

APE$_{seg}$

System: You are a professional post-editor. Your task is to revise the draft translation of the given source segment so that it is faithful to the source and fluent in {tgt_lang}. Pay particular attention to consistency in terminology and named entities, stylistic and tonal coherence appropriate to the genre, formality and register given the speaker relationships and discourse context, and natural discourse flow including connectives and reference expressions. Output ONLY the corrected {tgt_lang} sentence between <pe> and </pe>.

User:

```
[{src_lang} Source]
{src_seg}

[{tgt_lang} Draft Translation]
{tgt_seg}

MUST output in this format:
<pe>{tgt_lang} corrected sentence only</pe>
```

APE$_{exp}$

System: You are a professional post-editor. Your task is to revise the draft translation of the given source segment so that it is faithful to the source, fluent in {tgt_lang}, appropriate given the document information provided below. Pay particular attention to consistency in terminology and named entities, stylistic and tonal coherence appropriate to the genre, formality and register as specified in the document information, and natural discourse flow including connectives and reference expressions. Output ONLY the corrected {tgt_lang} sentence between <pe> and </pe>.

User:

```
[Document Information]
Text type: {text_type}
Domain: {domain}
Participants:
    - {role}: {description}
    ...
Relationship: {relationship_type}
Register: {overall_register}
Register notes: {notes}

[{src_lang} Source]
{src_seg}

[{tgt_lang} Draft Translation]
{tgt_seg}

MUST output in this format:
<pe>{tgt_lang} corrected sentence only</pe>
```

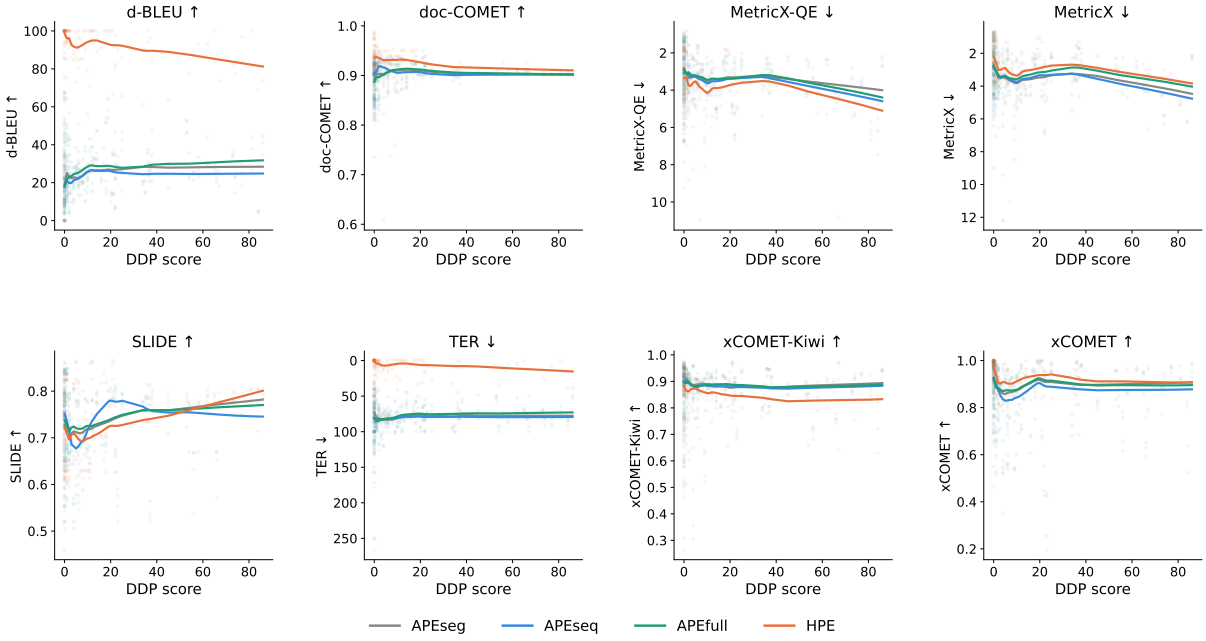
APE$_{full}$

System: You are a professional post-editor. Your task is to revise the draft translation of the given source segment so that it is faithful to the source, fluent in {tgt_lang}, fully coherent as a document. Pay particular attention to consistency in terminology and named entities, stylistic and tonal coherence appropriate to the genre, formality and register as specified in the document information, and natural discourse flow including connectives and reference expressions. Output ONLY the corrected {tgt_lang} sentence between <pe> and </pe>.

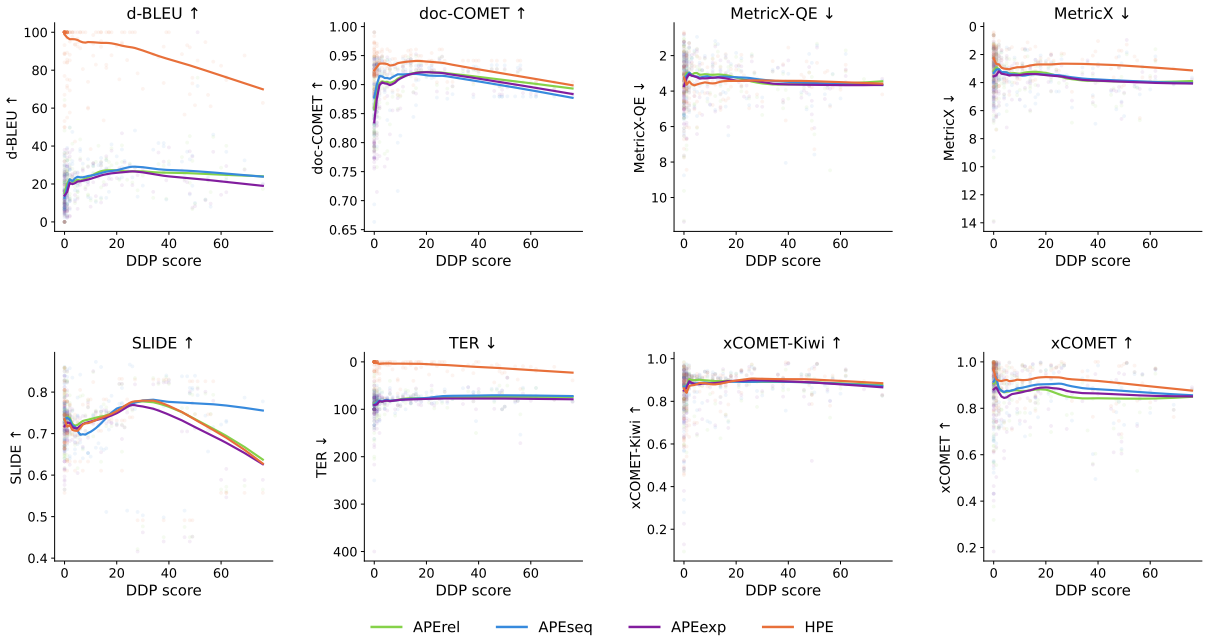
User:

```
[{src_lang} Source Document]
{src_doc}

[{tgt_lang} Draft Translation Document]
{tgt_doc}

[{src_lang} Source]
{src_seg}

[{tgt_lang} Draft Translation]
{tgt_seg}

MUST output in this format:
<pe>{tgt_lang} corrected sentence only</pe>
```

Figure 9: Prompt templates for each APE condition. The system prompt (blue) and user-side context fields differ across conditions: APE$_{seg}$ provides no context; APE$_{seq}$ and APE$_{rel}$ supply surrounding source and target segments; APE$_{exp}$ provides a document-level summary; and APE$_{full}$ passes the full source and draft translation document.

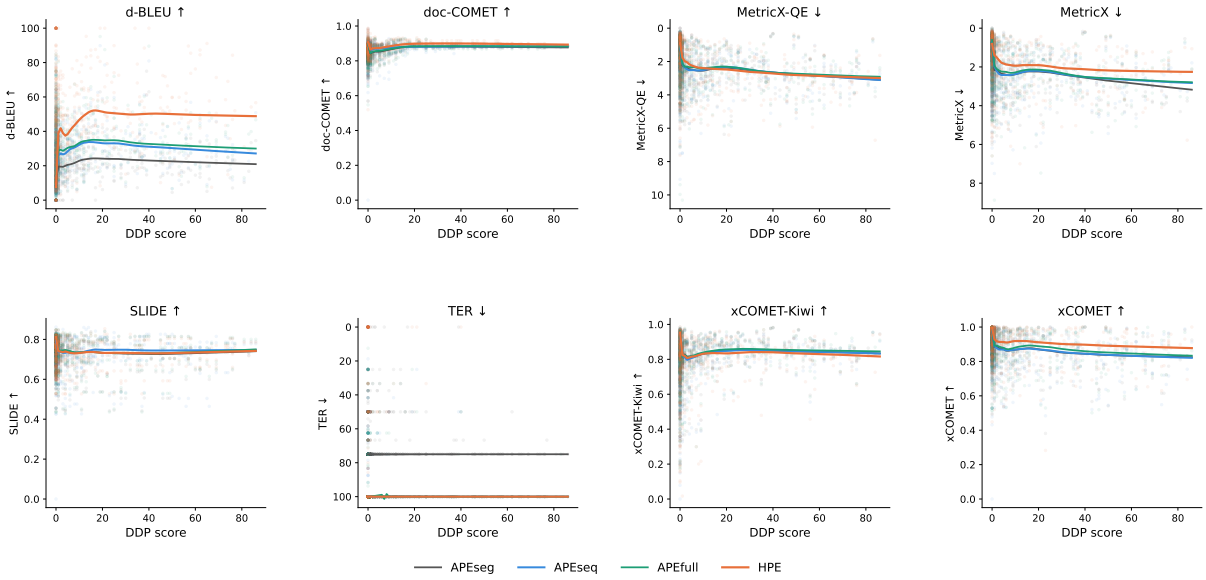


(a) **Size** : APE_{seg}, APE_{seq}, APE_{full}, HPE

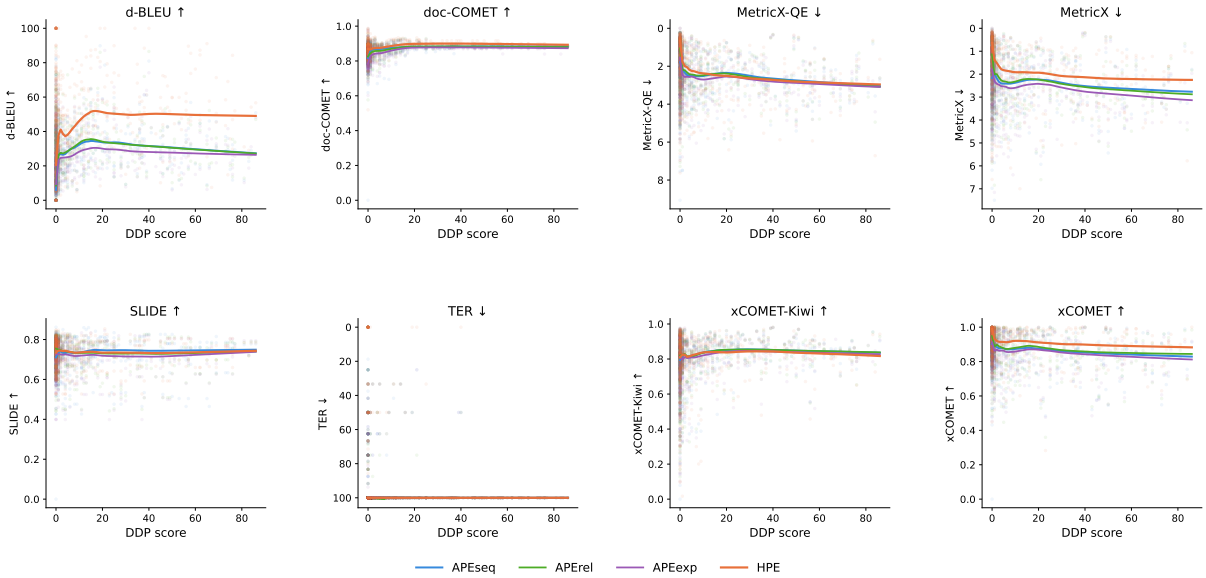


(b) **Selection** : APE_{seq}, APE_{rel}, APE_{exp}, HPE

Figure 10: LOWESS-smoothed automatic metric scores as a function of segment-level DDP (**En–Ko**, all six models averaged). Arrows indicate the preferred direction (↑ higher is better, ↓ lower is better).



(a) **Size** : APE_{seg}, APE_{seq}, APE_{full}, HPE



(b) **Selection** : APE_{seq}, APE_{rel}, APE_{exp}, HPE

Figure 11: LOWESS-smoothed automatic metric scores as a function of segment-level DDP (**En–Zh**, four models averaged). Arrows indicate the preferred direction (↑ higher is better, ↓ lower is better).