Quit While You're Ahead: QUIT for Efficient Candidate Generation in Machine Translation Reranking

Guangyu Chen[*1], **Boxuan Lyu** 吕博轩[*1],
Hidetaka Kamigaito[2], **Kotaro Funakoshi**[1], and **Manabu Okumura**[1]
[1]Institute of Science Tokyo, [2]Nara Institute of Science and Technology
{co_gy,lyu,funakoshi,oku}@lr.first.iir.isct.ac.jp
kamigaito.h@is.naist.jp

arXiv:2609.00588v3 [cs.CL] 10 Sep 2026

Abstract

Reranking methods, such as Minimum Bayes Risk (MBR) decoding and Quality Estimation (QE) reranking, have been widely used in modern neural machine translation (NMT) to select an output from a set of candidate hypotheses. However, the performance gains come at the cost of high inference latency. Existing acceleration methods target MBR decoding and reduce only the reranking computation, leaving QE reranking unaddressed and candidate generation—which can be the larger computational bottleneck—largely untouched. In this work, we propose **QUIT** (**Q**uantifying **U**ncertainty for **I**ncremental **T**ermination), a novel early-stopping strategy for the entire generation–reranking pipeline. QUIT treats candidate generation as a sequential decision-making process under uncertainty. It incrementally generates and reranks candidates, stopping when the best reranking score stabilizes. Comprehensive experiments with three NMT models across 19 language pairs show that QUIT achieves end-to-end speedups of 1.47–2.66× for MBR decoding and 3.43–4.12× for QE reranking while preserving translation quality for nearly all external quality metrics.

1 Introduction

Reranking is increasingly common in modern neural machine translation (NMT): a reranker selects the final output from a set of hypotheses generated by an NMT model. Reranking offers two advantages: (1) it uses external criteria rather than relying solely on model probability, which has been shown to correlate poorly with translation quality (Eikema and Aziz, 2020; Freitag et al., 2022), and (2) it flexibly adapts output selection to different inference-time objectives without requiring model retraining (Saunders et al., 2022). In recent WMT General Translation Shared Tasks, nearly all top-performing systems have used reranking during inference or for training-data construction (Kocmi et al., 2024, 2025). The predominant approaches are Minimum Bayes Risk (MBR) decoding (Kumar and Byrne, 2004; Tromble et al., 2008; Eikema and Aziz, 2020, 2022; Freitag et al., 2022; Fernandes et al., 2022) and Quality Estimation (QE) reranking (Fernandes et al., 2022), along with their variants.

Despite its effectiveness, the reranking pipeline is computationally expensive, as it requires generating many hypotheses and repeatedly invoking the reranker. Although several methods have recently been proposed to accelerate MBR decoding (Cheng and Vlachos, 2023; Vamvas and Sennrich, 2024; Deguchi et al., 2024), they are specific to MBR and optimize only the reranking stage, without reducing the need for large candidate sets. To the best of our knowledge, no acceleration method has been developed for QE reranking. Moreover, as large language model (LLM)-based NMT systems become more common, candidate generation can be substantially more costly than reranking.

In a pilot study on WMT25 with Qwen3, we found that the minimum candidate set needed to match the quality of full 512-candidate MBR varies widely across source sentences (Figure 1). This variability motivates an early-stopping signal to determine when to stop candidate generation for each source sentence.

Our goal is to stop with low *early-stopping risk*: the absolute difference in reference-based quality between the reranker's current output and its output from the full candidate set. This quantity can only be measured retrospectively, since the full-budget output is unavailable during incremental generation. We therefore use recent variation in the best reference-based quality within the growing candidate set as a local uncertainty signal. Because this signal also requires a human reference translation, we approximate it using variation in the best reranking score at inference time. We therefore propose **QUIT** (**Q**uantifying **U**ncertainty for

[*]Equal contribution.

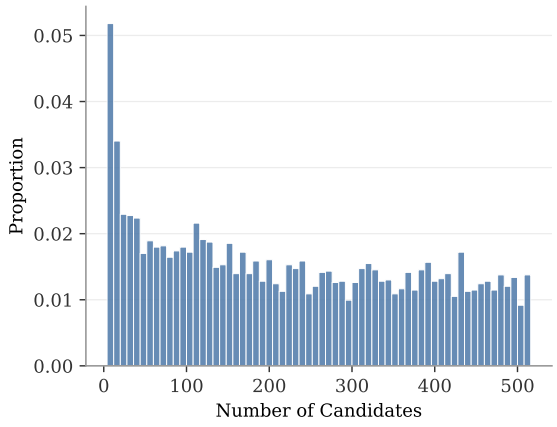


Figure 1: Distribution of the smallest candidate-set size N satisfying $|\,\mathrm{xCOMET}(y_N) - \mathrm{xCOMET}(y_{512})| \leq 10^{-3}$, where y_N denotes the MBR output selected from N candidates. WMT25, Qwen3-8B, COMET-22 utility.

Incremental **T**ermination), an uncertainty-guided early-stopping strategy for the entire generation–reranking pipeline. QUIT generates candidates incrementally (e.g., eight at a time) and updates their reranking scores. It uses recent variation in the best reranking score as a practical proxy for this uncertainty and stops when that variation is at or below a prespecified convergence threshold. This strategy therefore reduces both candidate-generation and reranking costs.

Across three NMT models and the WMT24 (Kocmi et al., 2024) and WMT25 test sets, spanning 19 language pairs, QUIT achieves end-to-end speedups of 1.47–2.66× for MBR decoding and 3.43–4.12× for QE reranking while remaining statistically equivalent to the unaccelerated baseline on nearly all external quality metrics. These results demonstrate that adaptive candidate generation is key to reducing end-to-end reranking latency.

2 NMT Reranking

Reranking selects a translation according to the preferences of an external reranker, rather than solely according to the probabilities estimated by the NMT model. An NMT system with reranking consists of two stages. First, the NMT model generates multiple candidates, for example through beam search or sampling. Second, the reranker scores these candidates, and the system returns the one with the best reranking score. Decoupling generation from selection allows output selection to target criteria that may align better with translation quality than model probability.

Formally, given a source sentence s, an NMT model with parameters θ defines a distribution $P_\theta(\cdot \mid s)$ over the hypothesis space $\mathcal{H}$. A generation method constructs a candidate set $\mathcal{C} = \{h_1, \ldots, h_N\} \subset \mathcal{H}$ of size N. Let $s_r(h; s, \mathcal{C})$ denote the score that an external reranker assigns to candidate $h \in \mathcal{C}$, where a higher score indicates a stronger preference. The system selects:

$$\hat{y}_r = \operatorname*{argmax}_{h \in \mathcal{C}} s_r(h; s, \mathcal{C}). \tag{1}$$

The resulting performance depends on both the coverage of $\mathcal{C}$ and the accuracy of s_r. Increasing N makes a high-quality translation more likely to appear in $\mathcal{C}$, but also raises the cost of both stages.

2.1 Minimum Bayes Risk Decoding

MBR decoding is a reranking method with a long history in statistical machine translation (Kumar and Byrne, 2004; Tromble et al., 2008). It uses hypotheses sampled from the model as "pseudo-references", called *support hypotheses*, and scores each candidate by its average utility against them. The candidate with the lowest estimated Bayes risk—or, equivalently, the highest expected utility—is returned as the final translation.

Formally, MBR instantiates the generic score s_r in Equation 1 as an expected-utility score s_{MBR}. Let $\mathcal{S} = \{h_s^{(m)}\}_{m=1}^{M}$ denote a set of M independently sampled support hypotheses, where $h_s^{(m)} \sim P_\theta(\cdot \mid s)$. Let $u(h, h_s)$ denote the utility of candidate h when support hypothesis $h_s \in \mathcal{S}$ serves as a pseudo-reference. The Monte Carlo MBR score and the corresponding selected output are:

$$s_{\mathrm{MBR}}(h; \mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{h_s \in \mathcal{S}} u(h, h_s), \tag{2}$$

$$\hat{y}_{\mathrm{MBR}} = \operatorname*{argmax}_{h \in \mathcal{C}} s_{\mathrm{MBR}}(h; \mathcal{S}). \tag{3}$$

This uniform average is the standard Monte Carlo approximation to the expectation under the model distribution. In standard MBR decoding, the candidate and support sets are often identical, i.e., $\mathcal{C} = \mathcal{S}$. Neural evaluation metrics can serve as u, allowing selection to target translation quality more directly than model probability does. This benefit, however, comes at a substantial computational cost: scoring all candidate–support pairs requires $|\mathcal{C}|\,|\mathcal{S}|$ utility evaluations and becomes quadratic in the candidate-set size when $\mathcal{C} = \mathcal{S}$.

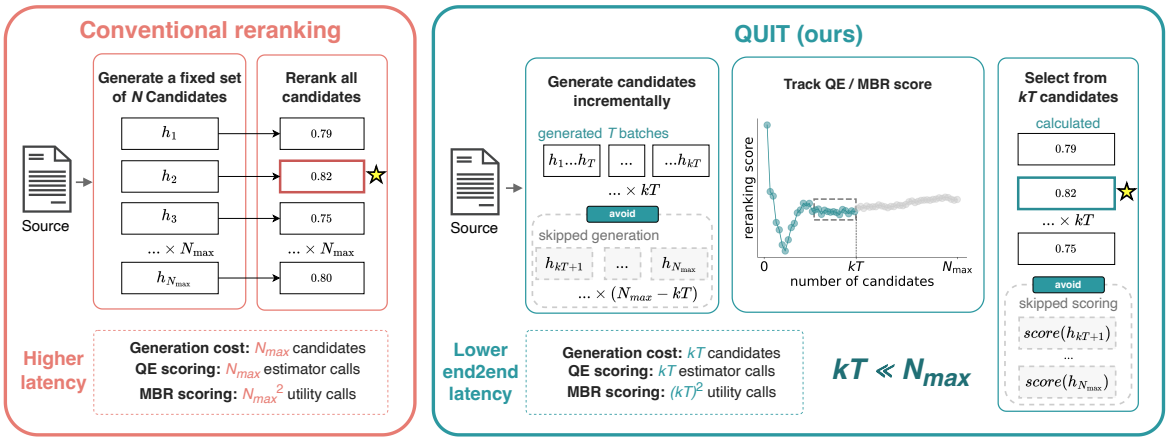


Figure 2: Overview of conventional reranking and QUIT. Conventional reranking generates the full candidate set and scores every candidate, whereas QUIT generates candidates incrementally, tracks the best reranking score, and stops once that score stabilizes. It thereby avoids generating and scoring additional candidates and reduces end-to-end latency.

2.2 Quality Estimation Reranking

QE reranking dispenses with support hypotheses entirely. Instead, a QE model predicts translation quality directly from the source sentence and a candidate, without access to a human reference translation.

Formally, QE instantiates s_r as a reference-free score $s_{\mathrm{QE}}(h; s)$. This score represents the predicted quality of candidate h given the source sentence s and, unlike the MBR score, does not depend on the other candidates in $\mathcal{C}$. The QE-selected output is:

$$\hat{y}_{\mathrm{QE}} = \operatorname*{argmax}_{h \in \mathcal{C}} s_{\mathrm{QE}}(h; s). \tag{4}$$

QE reranking scores each candidate independently given the source sentence and therefore requires $|\mathcal{C}|$ QE-model evaluations. Nevertheless, both methods conventionally generate the full candidate set before selecting a translation. They therefore use the full candidate budget even when the current candidate set is sufficient for reliable output selection. QUIT addresses this limitation by adaptively determining when to stop expanding the candidate set for each source sentence.

3 Proposed Method: QUIT

Conventional reranking uses the same fixed candidate budget for every source sentence, even when additional candidates cease to improve translation quality. We propose QUIT to address this inefficiency. Its goal is to stop while the quality of the selected output is close to that of the full-budget reranker's output. Since this difference is unavailable during incremental generation, QUIT uses recent reranking score stability as an uncertainty signal for deciding when to stop.

Formally, let $N_{\max}$ denote the maximum candidate budget. Candidates are generated in batches of k, so after generation step i, a total of ik candidates have been generated. Assuming that k divides $N_{\max}$, the candidate count grows as follows:

$$0 < k < 2k < \cdots < N_{\max}. \tag{5}$$

For $i \in \{1, \ldots, N_{\max}/k\}$, define the partial candidate set after step i as:

$$\mathcal{C}_i = \{h_1, \ldots, h_{ik}\}, \qquad |\mathcal{C}_i| = ik. \tag{6}$$

For MBR in our experiments, we use the current candidate set as the support set, i.e., $\mathcal{S}_i = \mathcal{C}_i$, so both sets expand at each generation step. Let s_r be any reranking function for which higher scores are better. Let $\hat{y}_i$ denote the translation selected after generation step i:

$$\hat{y}_i = \operatorname*{argmax}_{h \in \mathcal{C}_i} s_r(h; s, \mathcal{C}_i). \tag{7}$$

Early-stopping Risk Given a window size w, let $\mathcal{B}_b = \{(b-1)w + 1, \ldots, bw\}$ denote the b-th non-overlapping window of generation steps, where $b \in \{1, \ldots, \lfloor N_{\max}/(kw) \rfloor\}$. Let $q(h, y_s^{\mathrm{ref}})$ evaluate candidate h against a human reference translation y_s^{ref}, with higher scores indicating better translation quality. We define the early-stopping risk at the end of window b as:

$$\begin{aligned}\mathcal{L}_b(s) \coloneqq \big| & q(\hat{y}_{bw}, y_s^{\mathrm{ref}}) \\ & - q(\hat{y}_{N_{\max}/k}, y_s^{\mathrm{ref}}) \big|.\end{aligned} \tag{8}$$

Algorithm 1: QUIT

procedure QUIT$(s, s_r, k, w, \alpha, N_{\max})$
  ▷ Initialize the candidate set and score buffer
  $\mathcal{C} \leftarrow \varnothing$
  $\mathcal{R} \leftarrow [\,]$
  $i \leftarrow 0$
  while $ik < N_{\max}$ **do**
    $i \leftarrow i + 1$
    Generate k new candidates and add them to $\mathcal{C}$
    $\hat{y} \leftarrow \operatorname{argmax}_{h \in \mathcal{C}} s_r(h; s, \mathcal{C})$
    $R \leftarrow \max_{h \in \mathcal{C}} s_r(h; s, \mathcal{C})$
    Append R to $\mathcal{R}$
    if $|\mathcal{R}| = w$ **then**
      $\Delta \leftarrow \max \mathcal{R} - \min \mathcal{R}$
      if $\Delta \leq \alpha$ **then**
        return $\hat{y}$
      else
        ▷ Start a new non-overlapping window
        $\mathcal{R} \leftarrow [\,]$
  return $\hat{y}$

This nonnegative quantity measures the absolute quality difference between two outputs selected by the same reranker; a small value means that stopping preserves the full-budget output's quality. During incremental inference, the full-budget output is unavailable, so we need a local uncertainty signal to guide stopping.

Within-window Uncertainty If a human reference were available, we define the best reference-based quality in the candidate set at step i as:

$$Q_i(s) \coloneqq \max_{h \in \mathcal{C}_i} q(h, y_s^{\text{ref}}). \tag{9}$$

The trajectory $\{Q_i(s)\}$ traces how the best reference-based quality evolves as the candidate set expands. We measure its within-window variation by the range:

$$\Delta_b^Q(s) = \max_{j \in \mathcal{B}_b} Q_j(s) - \min_{j \in \mathcal{B}_b} Q_j(s). \tag{10}$$

A large $\Delta_b^Q(s)$ indicates that recent candidates continue to improve the best available reference-based quality; a small value indicates local stability, suggesting that further sampling may have limited benefit. This is a local uncertainty signal, not the early-stopping risk $\mathcal{L}_b(s)$: $Q_i(s)$ measures the best quality in the candidate set, rather than the quality of the reranker's selected output, and recent stability does not guarantee stability through the full budget.

Practical Proxy At inference time, the human reference translation—and therefore the reference-based quality variation $\Delta_b^Q(s)$—is unavailable. QUIT instead tracks the best reranking score:

$$R_i(s) \coloneqq \max_{h \in \mathcal{C}_i} s_r(h; s, \mathcal{C}_i). \tag{11}$$

Its within-window variation serves as a reference-free proxy for $\Delta_b^Q(s)$:

$$\Delta_b^R(s) = \max_{j \in \mathcal{B}_b} R_j(s) - \min_{j \in \mathcal{B}_b} R_j(s). \tag{12}$$

Given a convergence threshold α, QUIT stops at:

$$T = \min \left\{ bw : \Delta_b^R(s) \leq \alpha \right\}, \tag{13}$$

and returns $\hat{y}_T$. If the condition is never satisfied, generation continues until the full candidate budget $N_{\max}$ is reached, and QUIT returns $\hat{y}_{N_{\max}/k}$. The threshold controls score stability, rather than imposing a bound on $\mathcal{L}_b(s)$; Section 5.1 evaluates how well this signal ranks early-stopping risk. Algorithm 1 summarizes the procedure.

4 Experiments

4.1 Experimental Setup

Datasets and NMT Models We used the complete test sets from the WMT24 (Kocmi et al., 2024) and WMT25 (Kocmi et al., 2025) General Translation Shared Tasks, which together cover 19 language pairs. We evaluated three NMT models across two categories. Qwen3 (`Qwen3-8B`) (Yang et al., 2025) is a general-purpose LLM without translation-specific fine-tuning, whereas TranslateGemma (`translategemma-12b-it`) (Finkelstein et al., 2026) and Hy-MT2 (`Hy-MT2-30B-A3B`) (Zheng et al., 2026) are translation-specialized models. For each source sentence, we set the maximum candidate budget to $N_{\max} = 512$ and used ancestral sampling with a temperature of 1.2.

Acceleration Baselines For MBR, we also compared against PruneMBR (Cheng and Vlachos, 2023) and PMBR (Trabelsi et al., 2024), which reduce MBR reranking computation while retaining the full candidate budget.

NMT Model	Reranker	Acceleration	Speedup ↑	Rerank. ↑	ChrF++ ↑	xCOMET ↑	MetricX ↓	GEMBA ↑
Qwen3	MBR	Unaccelerated (512)	1.00	.8357	42.09	.7520	5.17	-7.81
		PruneMBR	1.05	.8355††	42.07††	.7519††	5.17††	-7.77††
		PMBR	1.03	.8352††	42.11††	.7516††	5.17††	-7.79††
		QUIT ($w = 8$)	1.74	.8344††	42.04††	.7497††	5.19††	-7.86††
		QUIT ($w = 12$)	1.33	.8352††	42.04††	.7509††	5.17††	-7.85††
		QUIT ($w = 15$)	1.19	.8355††	42.06††	.7518††	5.17††	-7.79††
	QE	Unaccelerated (512)	1.00	.8223	39.61	.7585	4.96	-7.93
		QUIT ($w = 8$)	3.64	.8133	40.46†	.7524†	5.13†	-7.95††
		QUIT ($w = 12$)	2.33	.8165	40.34†	.7554††	5.05†	-7.86††
		QUIT ($w = 15$)	1.84	.8180	40.24†	.7570††	5.01††	-7.85††
TranslateGemma	MBR	Unaccelerated (512)	1.00	.8570	44.90	.8521	3.15	-3.13
		PruneMBR	1.04	.8570††	44.88††	.8520††	3.15††	-3.12††
		PMBR	1.03	.8566††	44.91††	.8518††	3.15††	-3.14††
		QUIT ($w = 8$)	2.66	.8566††	44.88††	.8509††	3.16††	-3.15††
		QUIT ($w = 12$)	1.89	.8568††	44.88††	.8511††	3.16††	-3.14††
		QUIT ($w = 15$)	1.59	.8569††	44.87††	.8514††	3.15††	-3.15††
	QE	Unaccelerated (512)	1.00	.8376	43.59	.8570	3.14	-3.10
		QUIT ($w = 8$)	4.01	.8327	44.37†	.8553††	3.15††	-2.98†
		QUIT ($w = 12$)	2.57	.8344	44.29†	.8562††	3.14††	-3.00†
		QUIT ($w = 15$)	2.02	.8352†	44.23†	.8570††	3.14††	-2.99†
Hy-MT2	MBR	Unaccelerated (512)	1.00	.8731	49.86	.8867	2.92	-1.77
		PruneMBR	1.02	.8732††	49.87††	.8868††	2.92††	-1.79††
		PMBR	1.02	.8728††	49.87††	.8866††	2.93††	-1.79††
		QUIT ($w = 8$)	2.66	.8728††	49.86††	.8867††	2.93††	-1.77††
		QUIT ($w = 12$)	1.87	.8730††	49.85††	.8865††	2.93††	-1.76††
		QUIT ($w = 15$)	1.58	.8730††	49.82††	.8865††	2.93††	-1.78††
	QE	Unaccelerated (512)	1.00	.8366	47.67	.8801	2.98	-2.37
		QUIT ($w = 8$)	4.12	.8317	48.54†	.8821††	2.96††	-2.07
		QUIT ($w = 12$)	2.65	.8332	48.34†	.8819††	2.96††	-2.10
		QUIT ($w = 15$)	2.08	.8340†	48.24†	.8816††	2.97††	-2.16†

Table 1: Main results on WMT24. The reranking score (Rerank.) is computed using COMET-22 for MBR decoding and CometKiwi-22 for QE reranking. † marks scores that are statistically equivalent to the unaccelerated baseline within a prespecified margin of 0.05σ, and †† marks equivalence within a stricter margin of 0.02σ. Speedup is measured end to end relative to that baseline.

QUIT Hyperparameters We fixed the generation batch size at $k = 8$ and the convergence threshold at $\alpha = 10^{-3}$, and evaluated window sizes $w \in \{3, 5, 8, 12, 15, 20, 30\}$. The main results report $w \in \{8, 12, 15\}$.

Rerankers We considered two widely used reranking paradigms in NMT: MBR decoding and QE reranking. For MBR, we used COMET-22 (`wmt22-comet-da`) (Rei et al., 2022a) as the pairwise utility function; for QE reranking, we used CometKiwi-22 (`wmt22-cometkiwi-da`) (Rei et al., 2022b) as the QE model. Because newer rerankers are substantially larger, evaluating them at our experimental scale would be prohibitively expensive; we therefore excluded them.

Quality Evaluation We evaluated translation quality using three complementary classes of automatic metrics that correlate strongly with human judgments. We used ChrF++ (Popović, 2015) as a surface-form metric; xCOMET (`XCOMET-XXL`) (Guerreiro et al., 2024) and MetricX (`metricx-24-hybrid-xl-v2p6`) (Juraska et al., 2024) as learned neural metrics; and GEMBA-MQM (hereafter GEMBA) (Kocmi and Federmann, 2023), instantiated with Gemma-4-31B (`gemma-4-31B-it`) (Team et al., 2026), as an LLM-based metric. Because GEMBA is methodologically distinct from the COMET-family rerankers, it provides an independent check and reduces the risk that the evaluator favors outputs optimized by a similar metric. We also report the COMET-22 and CometKiwi-22 reranking scores to measure how closely early stopping preserves each reranker's preferences. To assess whether acceleration affects translation quality, we ran paired-bootstrap equivalence tests using two one-sided tests (TOST)

(Schuirmann, 1987) with 10,000 resamples against the unaccelerated baseline for every metric. For each metric, we defined σ as the baseline segment-level standard deviation and adopted equivalence margins of 0.05σ and, as a stricter criterion, 0.02σ.

Efficiency Evaluation For efficiency, we report end-to-end speedup relative to the unaccelerated baseline. End-to-end time includes both candidate generation and reranking; the accelerated and unaccelerated systems were measured under identical hardware conditions.

4.2 Experimental Results

Tables 1 and 2 summarize the main quality–efficiency trade-off on the two test sets. Among the window sizes shown in the main tables, $w = 8$ yielded the largest speedups: 1.47–$2.66\times$ for MBR decoding and 3.43–$4.12\times$ for QE reranking across the evaluated test-set–model combinations. On nearly all external quality metrics, the resulting scores were statistically equivalent to those of the unaccelerated baseline, with several scores improving over the baseline. Some QE reranking scores did not pass the equivalence test. For QE, each candidate's score is independent of the other candidates, so the best reranking score is nondecreasing with candidate-set size and inherently favors the full 512-candidate set. Increasing the window size to 12 or 15 generally reduced speedup but brought reranking scores and external quality scores closer to those of the unaccelerated baseline. In contrast, PruneMBR and PMBR achieved an end-to-end speedup of at most $1.05\times$ because candidate generation remained unchanged.

5 Discussion

5.1 Does the uncertainty score identify early-stopping risk?

Here, we examine whether QUIT's uncertainty score $\Delta_b^R(s)$ ranks the early-stopping risk $\mathcal{L}_b(s)$ defined in Equation 8. Following the prediction-rejection ratio (PRR) framework (Vashurin et al., 2025), we treat each source–window pair as an instance, using $\mathcal{L}_b(s)$ as the risk and $\Delta_b^R(s)$ as the uncertainty score. Rejecting instances in descending order of uncertainty gives a curve of the mean early-stopping risk among retained instances. We compute PRR as follows:

$$\mathrm{PRR} = \frac{\mathrm{AUC}_{\mathrm{rnd}} - \mathrm{AUC}_{\mathrm{uns}}}{\mathrm{AUC}_{\mathrm{rnd}} - \mathrm{AUC}_{\mathrm{oracle}}}, \qquad (14)$$

where $\mathrm{AUC}_{\mathrm{uns}}$ is the area under this curve, $\mathrm{AUC}_{\mathrm{oracle}}$ uses the ranking by $\mathcal{L}_b(s)$ itself, and $\mathrm{AUC}_{\mathrm{rnd}}$ uses random ranking. A value of 1 indicates oracle-equivalent early-stopping-risk ranking, whereas 0 indicates random ranking. This evaluation is retrospective; neither the full-budget output nor a human reference is used to make stopping decisions. Appendix D gives the evaluation procedure and corresponding results using $\Delta_b^Q(s)$ as the uncertainty score.

Table 3 reports pooled PRR for $w = 8$ across WMT24 and WMT25, all three NMT models. PRR is positive in all 24 settings. The signal is more informative on average for MBR than for QE, but remains well below oracle ranking for both. These results support $\Delta_b^R(s)$ as an informative, imperfect indicator of early-stopping risk.

5.2 Where does QUIT stop?

Figure 3 shows that QE generally stops with relatively small candidate sets, whereas MBR stopping positions are more broadly distributed across the candidate budget. This pattern is consistent with how the two reranking scores evolve: QE scores are unaffected by later candidates, allowing the best score to stabilize quickly, while expanding MBR's support set changes candidate scores and can delay stabilization. Larger windows generally delay stopping, particularly for MBR, producing behavior closer to full-budget reranking at the cost of lower speedups.

5.3 Where do the efficiency gains come from?

Table 4 separates the cost of candidate generation from that of MBR reranking. For the unaccelerated MBR baseline, candidate generation requires 53.52 seconds per segment on average, compared with only 1.40 seconds for reranking. The results reveal that candidate generation, rather than reranking, is the dominant bottleneck, accounting for nearly all end-to-end latency in unaccelerated MBR. Consequently, PMBR and PruneMBR substantially reduce the cost of reranking itself but yield little end-to-end improvement because they still generate the full candidate set. This highlights a broader limitation of reranking-only acceleration: as NMT generators become more expensive, optimizing the comparatively small reranking stage has diminishing system-level returns.

NMT Model	Reranker	Acceleration	Speedup ↑	Rerank. ↑	ChrF++ ↑	xCOMET ↑	MetricX ↓	GEMBA ↑
Qwen3	MBR	Unaccelerated (512)	1.00	.7345	34.06	.3696	10.86	-18.11
		PruneMBR	1.02	.7344††	34.04††	.3692††	10.86††	-18.10††
		PMBR	1.01	.7339††	34.07††	.3689††	10.88††	-18.17††
		QUIT ($w = 8$)	1.47	.7322††	33.98††	.3671††	10.95††	-18.17††
		QUIT ($w = 12$)	1.20	.7336††	34.01††	.3683††	10.90††	-18.17††
		QUIT ($w = 15$)	1.12	.7340††	34.04††	.3687††	10.88††	-18.11††
	QE	Unaccelerated (512)	1.00	.7106	30.32	.3705	10.10	-18.24
		QUIT ($w = 8$)	3.43	.6935	31.28	.3613†	10.54	-18.40†
		QUIT ($w = 12$)	2.19	.6996	31.05†	.3646†	10.38	-18.33†
		QUIT ($w = 15$)	1.74	.7029	30.87†	.3662†	10.29†	-18.27††
TranslateGemma	MBR	Unaccelerated (512)	1.00	.8129	36.25	.5731	6.06	-6.72
		PruneMBR	1.01	.8129††	36.25††	.5735††	6.06††	-6.72††
		PMBR	1.01	.8125††	36.22††	.5717††	6.08††	-6.85†
		QUIT ($w = 8$)	2.61	.8112††	36.18††	.5690†	6.12†	-6.86†
		QUIT ($w = 12$)	1.89	.8119††	36.16††	.5704††	6.10††	-6.82†
		QUIT ($w = 15$)	1.59	.8123††	36.20††	.5714††	6.09††	-6.85†
	QE	Unaccelerated (512)	1.00	.7721	36.17	.5670	6.11	-6.81
		QUIT ($w = 8$)	3.55	.7657	36.16††	.5645††	6.18†	-6.86†
		QUIT ($w = 12$)	2.29	.7680	36.17††	.5651††	6.16††	-6.84††
		QUIT ($w = 15$)	1.80	.7692†	36.20††	.5656††	6.14††	-6.80††
Hy-MT2	MBR	Unaccelerated (512)	1.00	.8298	41.19	.6281	5.62	-4.22
		PruneMBR	1.01	.8298††	41.20††	.6278††	5.62††	-4.19††
		PMBR	1.01	.8294††	41.19††	.6279††	5.63††	-4.16††
		QUIT ($w = 8$)	2.36	.8289††	41.18††	.6262††	5.65††	-4.24††
		QUIT ($w = 12$)	1.72	.8292††	41.16††	.6278††	5.64††	-4.22††
		QUIT ($w = 15$)	1.46	.8295††	41.17††	.6277††	5.64††	-4.17††
	QE	Unaccelerated (512)	1.00	.7786	40.78	.6219	5.71	-4.84
		QUIT ($w = 8$)	3.59	.7725	40.90††	.6185†	5.77†	-4.72†
		QUIT ($w = 12$)	2.29	.7747	40.88††	.6197††	5.75††	-4.76†
		QUIT ($w = 15$)	1.81	.7759†	40.84††	.6208††	5.75††	-4.74†

Table 2: Main results on WMT25. The reporting conventions follow Table 1.

6 Related Work

Reranking for NMT Although selecting the most probable candidate is a natural decoding strategy, model probability is often poorly aligned with translation quality (Stahlberg and Byrne, 2019; Ott et al., 2018; Freitag et al., 2022), motivating the use of external criteria for output selection. Against this background, MBR decoding, a decision rule widely studied in statistical machine translation, has been revisited for NMT. Early work on neural MBR decoding typically used surface-form metrics (Eikema and Aziz, 2020, 2022), such as sentence-level BLEU (Papineni et al., 2002) or ChrF (Popović, 2015, 2017), as utility functions for comparing candidates with sampled support hypotheses. As learned metrics have shown stronger agreement with human judgments, more recent work has increasingly adopted neural metrics as MBR utilities, allowing the decision rule to capture semantic similarity beyond lexical overlap (Freitag et al., 2022; Fernandes et al., 2022; Lyu et al., 2025; Kamigaito et al., 2025). In parallel, advances in reference-free QE have enabled QE reranking, which directly scores each source–hypothesis pair and selects the highest-scoring candidate without relying on support hypotheses (Fernandes et al., 2022).

Efficient Methods for NMT Reranking Despite their effectiveness, both MBR decoding and QE reranking can be prohibitively expensive in latency-sensitive settings because they conventionally require a large, fully generated candidate set. Existing efficiency methods have focused primarily on MBR decoding by reducing the number or cost of utility-function evaluations, for example by pruning or approximating candidate–support comparisons (Trabelsi et al., 2024; Deguchi et al., 2024; Cheng and Vlachos, 2023; Jinnai and Ariu, 2024). These methods optimize the reranking stage but generally retain the full candidate budget. This limitation is increasingly consequential as LLM-based systems grow larger: candidate generation

Test set	NMT model	MBR		QE	
		xCOMET	GEMBA	xCOMET	GEMBA
WMT24	Qwen3	0.430	0.391	0.195	0.216
WMT24	Hy-MT2	0.455	0.579	0.233	0.158
WMT24	TranslateGemma	0.356	0.429	0.207	0.207
WMT25	Qwen3	0.196	0.145	0.163	0.194
WMT25	Hy-MT2	0.360	0.376	0.178	0.145
WMT25	TranslateGemma	0.260	0.287	0.141	0.159
Average		**0.343**	**0.368**	**0.186**	**0.180**

Table 3: Pooled PRR with uncertainty $\Delta_b^R(s)$ and early-stopping risk $\mathcal{L}_b(s)$ (Equation 8). Higher PRR is better.

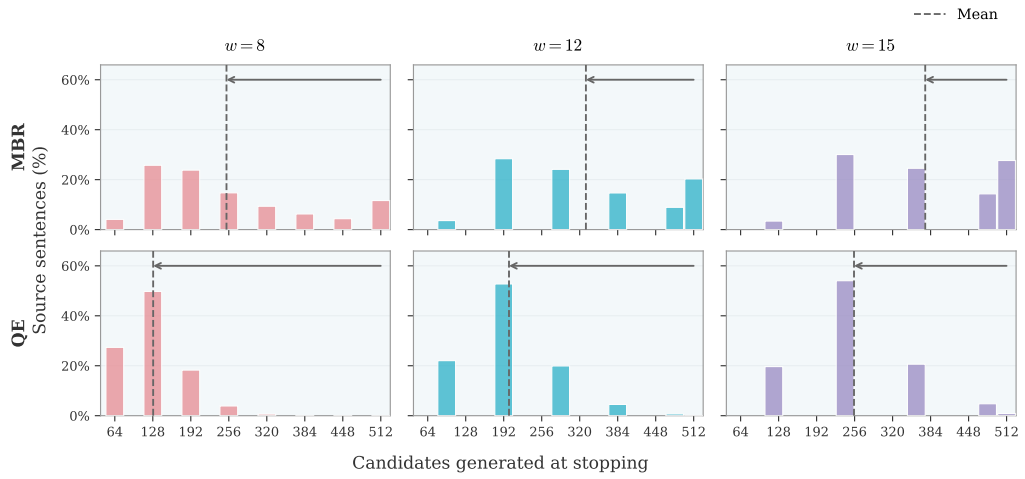


Figure 3: Distribution of candidate-set sizes when QUIT stops for MBR decoding and QE reranking. Bars show the percentage of source sentences for each window size. The bar at 512 includes sentences for which generation reaches the full candidate budget.

Method	w	Time per segment (s) ↓		
		Gen.	Rerank.	Total
Unaccelerated MBR	∞	53.52	1.40	54.92
PMBR	–	53.52	0.53	54.05
PruneMBR	–	53.52	0.33	53.85
QUIT	8	24.69	0.43	25.13
QUIT	12	33.06	0.67	33.73
QUIT	15	38.04	0.83	38.87

Table 4: Average component-wise MBR runtime per source segment. Times are pooled over WMT24 and WMT25 and the three NMT models, weighting each test set by its number of source segments. Complete aggregate wall-clock times are given in Appendix E.

can dominate end-to-end inference time, even when reranking is performed naively. Moreover, QE reranking avoids MBR's quadratic pairwise comparisons but still pays the generation and scoring costs for every candidate in the fixed budget. To the best of our knowledge, QUIT is the first acceleration method for NMT reranking that adaptively stops candidate generation and thereby reduces both candidate-generation and reranking costs.

7 Conclusions and Future Work

We introduced QUIT, an uncertainty-guided early-stopping method that jointly reduces candidate-generation and reranking costs. It uses local reranking-score stability to decide when to stop, with the aim of keeping the selected output's quality close to that of full-budget reranking. Across three NMT models and 19 language pairs, QUIT substantially reduced end-to-end latency while remaining within the tested equivalence margins on nearly all external quality metrics.

Future work will explore early-stopping criteria that do not require manually specified hyperparameters and evaluate the effectiveness of QUIT on tasks beyond machine translation.

Limitations

First, although our experiments cover three NMT models and 19 language pairs from WMT24 and WMT25, our conclusions remain tied to these mod-

els and test distributions. The observed speedups depend on the relative costs of candidate generation and reranking, which are influenced by model size and data characteristics. If candidate generation becomes substantially cheaper in future NMT systems, reranking may instead dominate end-to-end latency, and the speedups achieved by QUIT could differ considerably from those reported here.

Second, QUIT assumes that the candidate set can be generated incrementally, with newly sampled candidates appended to the existing set. This assumption does not hold directly for generation procedures such as beam search, in which hypotheses are jointly expanded and the resulting candidate sets are not necessarily nested. Adapting QUIT to such procedures requires further investigation.

Finally, using the uncertainty score in Equation 12 assumes that reranking scores obtained from different candidate sets are comparable. Although this assumption holds for the reranking methods considered in our experiments, future reranking objectives may produce scores whose scales vary with the candidate set. In such cases, the current stopping criterion may require score normalization or calibration. Even with comparable scores, local stability cannot guarantee low early-stopping risk: later candidates or changes in the reranker's selection may still alter output quality.

References

Julius Cheng and Andreas Vlachos. 2023. Faster minimum Bayes risk decoding with confidence-based pruning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12473–12480, Singapore. Association for Computational Linguistics.

Hiroyuki Deguchi, Yusuke Sakai, Hidetaka Kamigaito, Taro Watanabe, Hideki Tanaka, and Masao Utiyama. 2024. Centroid-based efficient minimum Bayes risk decoding. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11009–11018, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Bryan Eikema and Wilker Aziz. 2020. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Bryan Eikema and Wilker Aziz. 2022. Sampling-based approximations to minimum Bayes risk decoding for neural machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10978–10993, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412, Seattle, United States. Association for Computational Linguistics.

Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, Eleftheria Briakou, Elizabeth Nielsen, Jiaming Luo, Kat Black, Ryan Mullins, Sweta Agrawal, Wenda Xu, Erin Kats, Stephane Jaskiewicz, and 2 others. 2026. Translategemma technical report. *Preprint*, arXiv:2601.09012.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

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.

Yuu Jinnai and Kaito Ariu. 2024. Hyperparameter-free approach for faster minimum Bayes risk decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8547–8566, Bangkok, Thailand. 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.

Hidetaka Kamigaito, Hiroyuki Deguchi, Yusuke Sakai, Katsuhiko Hayashi, and Taro Watanabe. 2025. Diversity explains inference scaling laws: Through a case study of minimum Bayes risk decoding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29060–29094, Vienna, Austria. Association for Computational Linguistics.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard 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 and Christian Federmann. 2023. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Shankar Kumar and Bill Byrne. 2004. Minimum bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176.

Boxuan Lyu, Hidetaka Kamigaito, Kotaro Funakoshi, and Manabu Okumura. 2025. Unveiling the power of source: Source-based minimum Bayes risk decoding for neural machine translation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2976–2994, Vienna, Austria. Association for Computational Linguistics.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *International Conference on Machine Learning*.

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

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Danielle Saunders, Rosie Sallis, and Bill Byrne. 2022. First the worst: Finding better gender translations during beam search. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3814–3823, Dublin, Ireland. Association for Computational Linguistics.

Donald J. Schuirmann. 1987. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15:657–680.

Felix Stahlberg and Bill Byrne. 2019. On NMT search errors and model errors: Cat got your tongue? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3356–3362, Hong Kong, China. Association for Computational Linguistics.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, and 304 others. 2026. Gemma 4 technical report. *Preprint*, arXiv:2607.02770.

Firas Trabelsi, David Vilar, Mara Finkelstein, and Markus Freitag. 2024. Efficient minimum bayes risk decoding using low-rank matrix completion algorithms. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Roy Tromble, Shankar Kumar, Franz Och, and Wolfgang Macherey. 2008. Lattice Minimum Bayes-Risk decoding for statistical machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 620–629,

Honolulu, Hawaii. Association for Computational Linguistics.

Jannis Vamvas and Rico Sennrich. 2024. Linear-time minimum Bayes risk decoding with reference aggregation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 790–801, Bangkok, Thailand. Association for Computational Linguistics.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, Maxim Panov, and Artem Shelmanov. 2025. Benchmarking uncertainty quantification methods for large language models with LM-polygraph. *Transactions of the Association for Computational Linguistics*, 13:220–248.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Mao Zheng, Zheng Li, Tao Chen, Bo Lv, Mingrui Sun, Mingyang Song, Jinlong Song, Hong Huang, Decheng Wu, Hai Wang, Yifan Song, Yanfeng Chen, and Guanwei Zhang. 2026. Hy-mt2: A family of fast, efficient and powerful multilingual translation models in the wild. *Preprint*, arXiv:2605.22064.

A Complete Window-size Sweep

Figures 4–7 report the complete sweep over $w \in \{3, 5, 8, 12, 15, 20, 30\}$ for every test set, NMT model, and reranking paradigm. We show end-to-end speedup and all reference-based evaluation metrics reported in the main tables, while omitting the reranking score. Solid lines show QUIT; gray dashed lines show the corresponding unaccelerated MBR or QE baseline. The speedup baseline is $1\times$ in every panel.

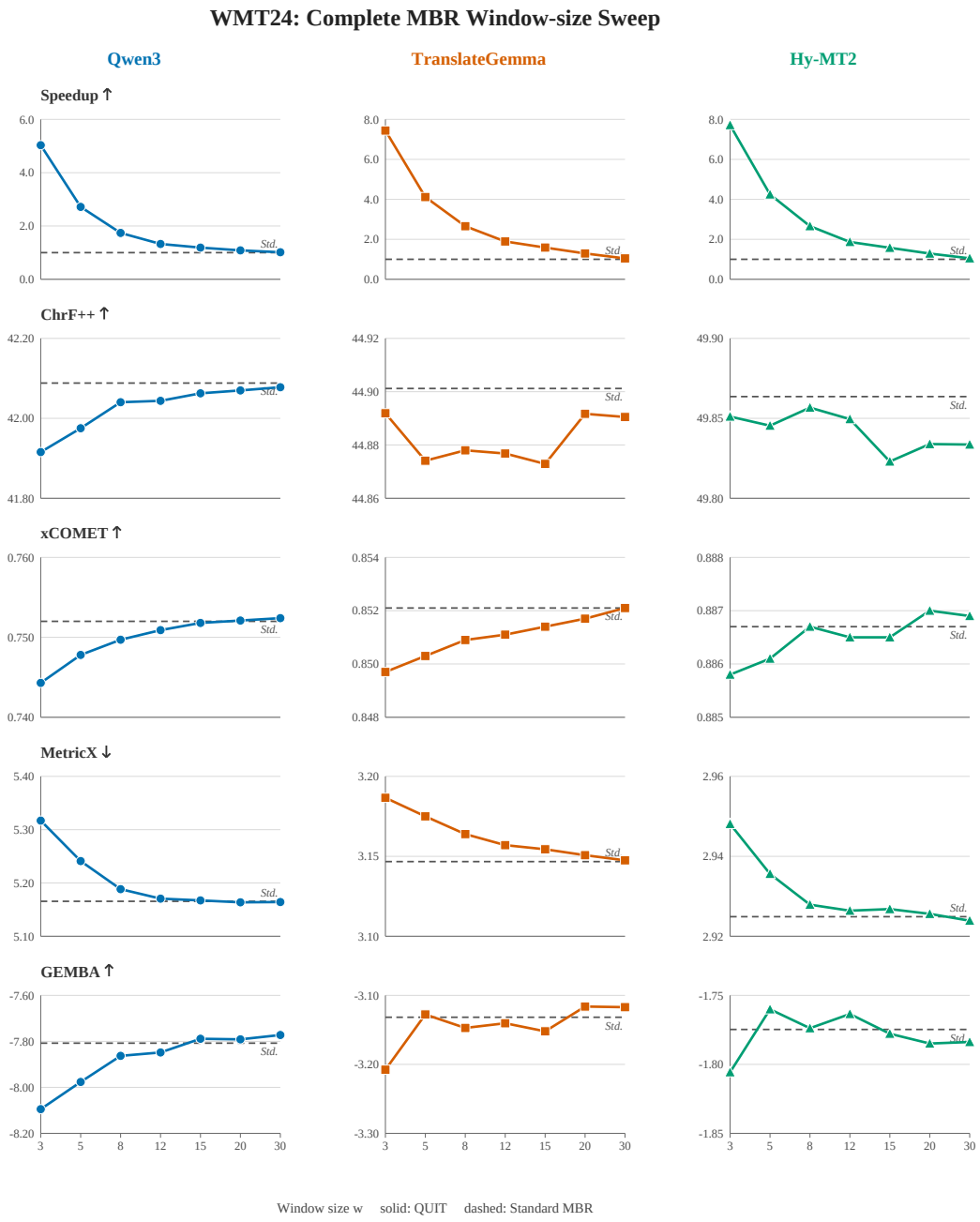


Figure 4: Complete MBR window-size sweep on WMT24. Gray dashed lines mark the per-model unaccelerated MBR performance; the gray dashed line in the speedup panel marks the $1\times$ unaccelerated baseline.

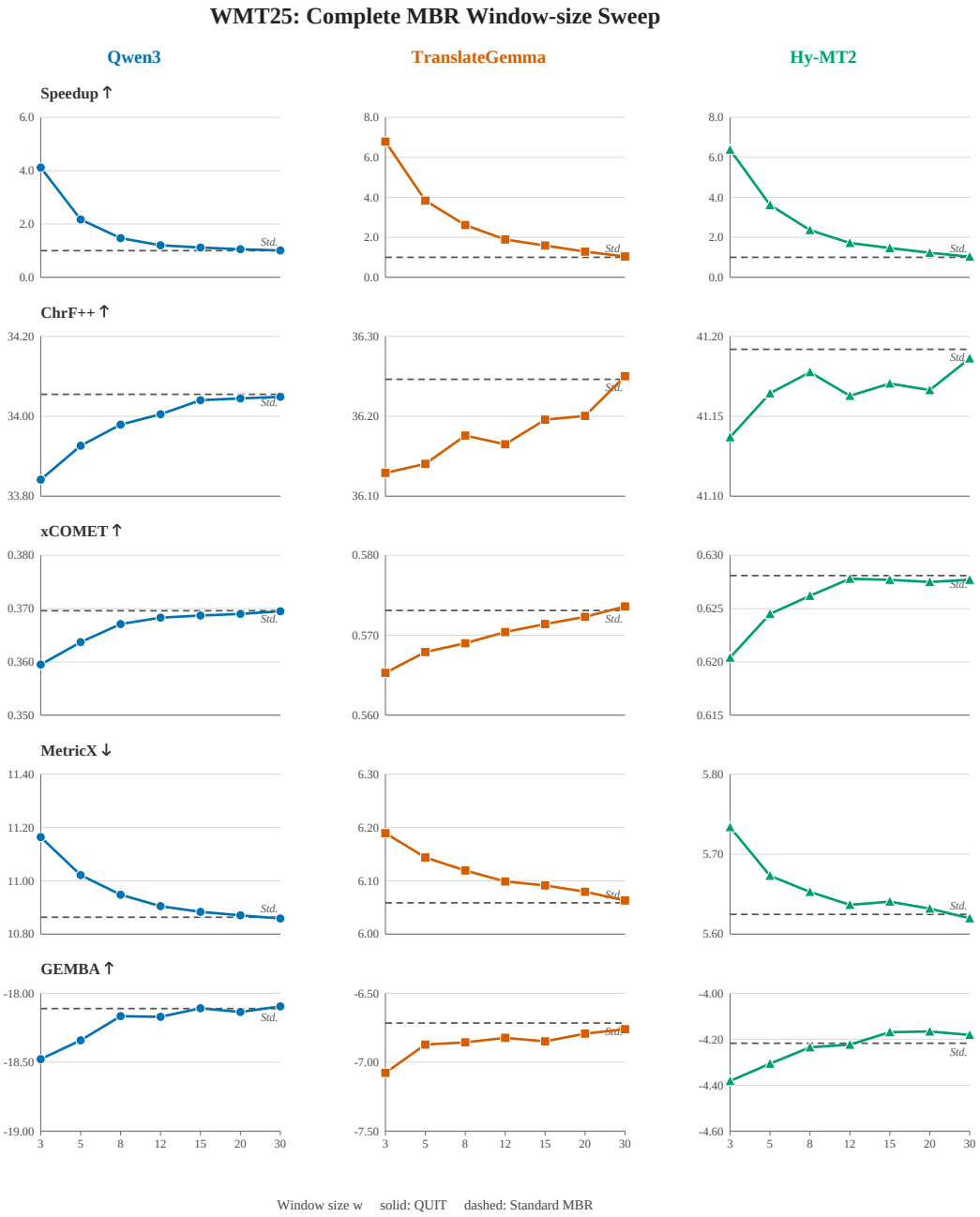


Figure 5: Complete MBR window-size sweep on WMT25, following the conventions of Figure 4.

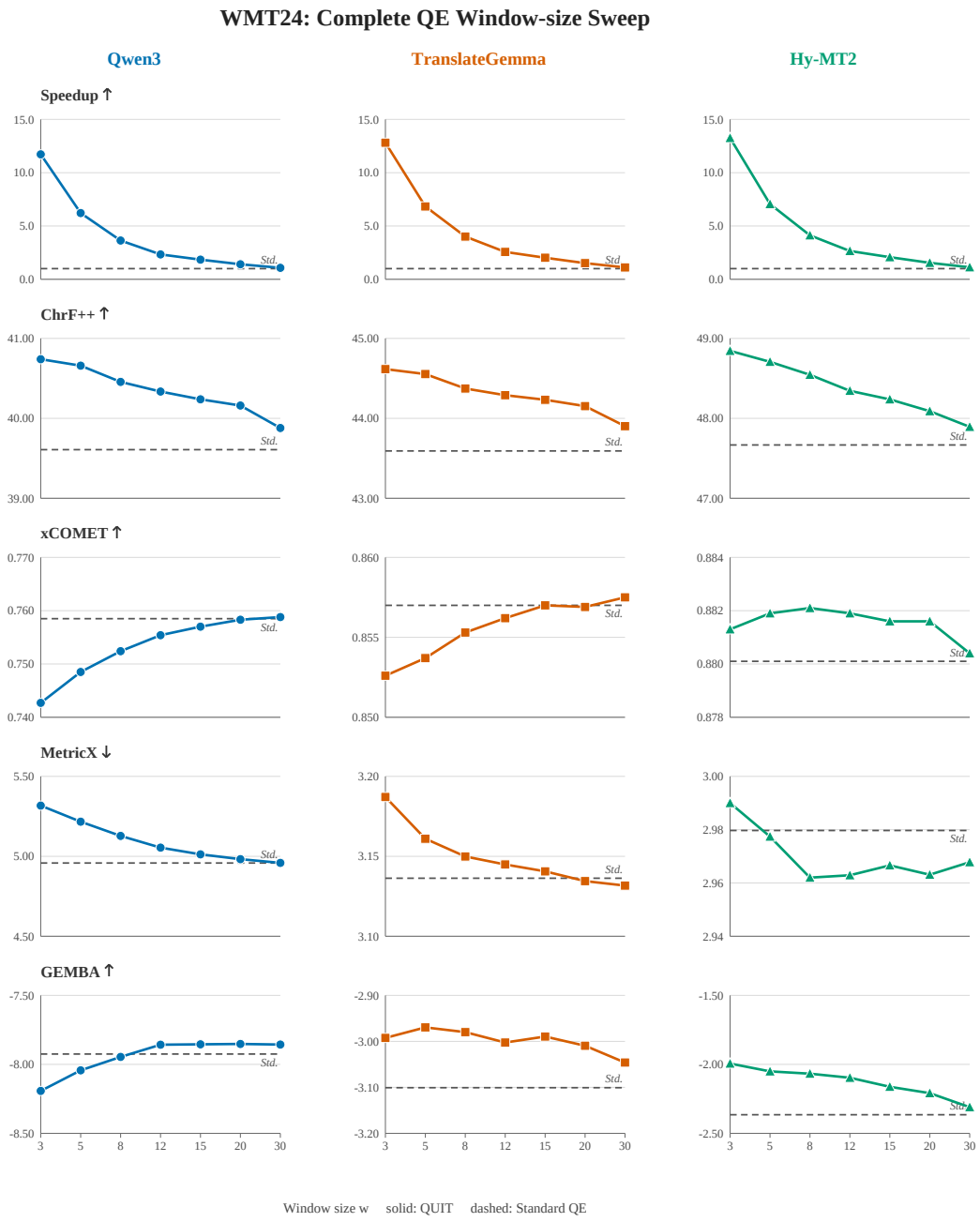


Figure 6: Complete QE window-size sweep on WMT24. Gray dashed lines mark the per-model unaccelerated QE performance; the gray dashed line in the speedup panel marks the $1\times$ unaccelerated baseline.

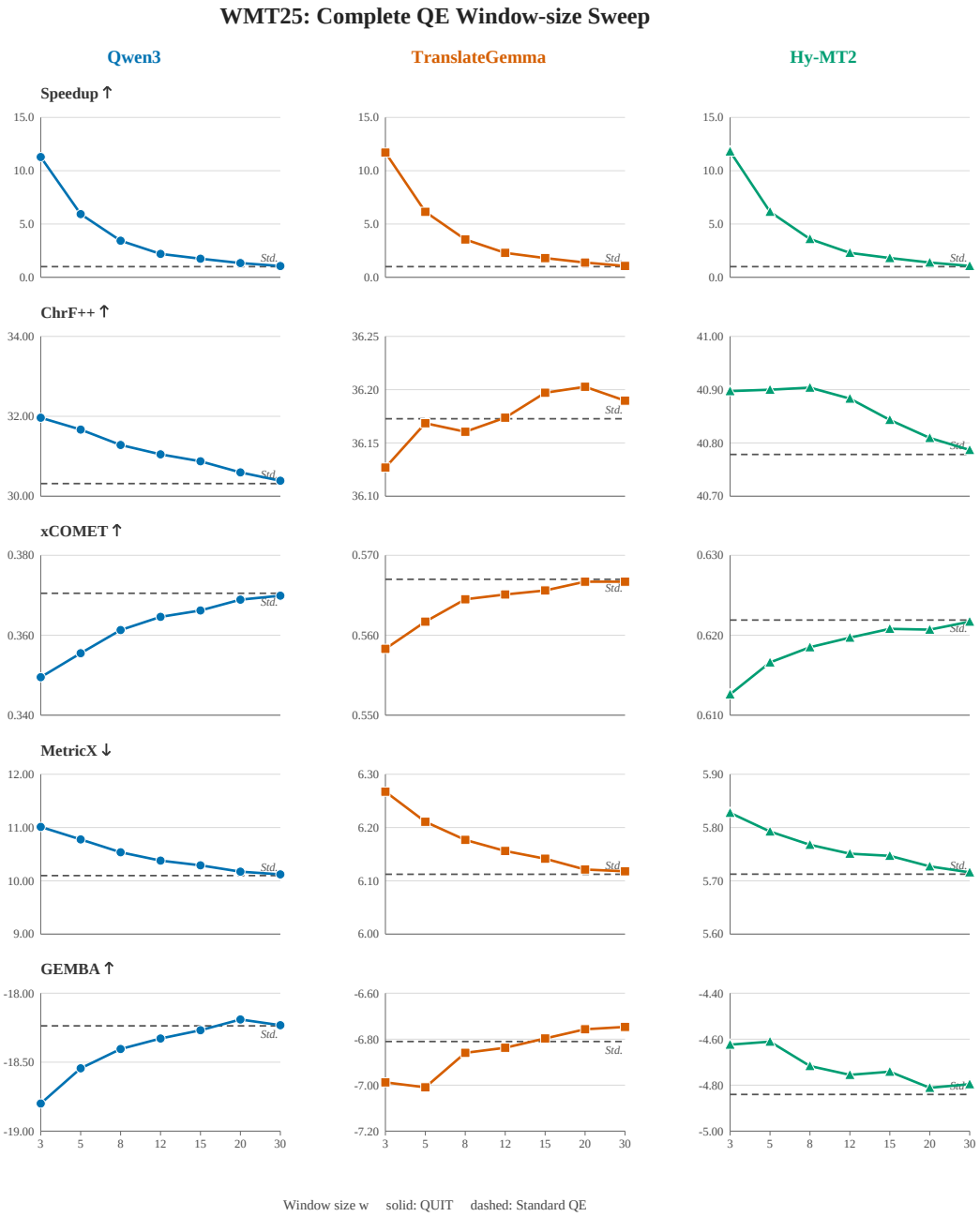


Figure 7: Complete QE window-size sweep on WMT25, following the conventions of Figure 6.

B Comparison with a Fixed-Budget Baseline

We compare QUIT at $w \in \{8, 12, 15\}$ and the default $\alpha = 10^{-3}$ with a fixed-budget baseline, Fixed-N^*, which generates and reranks only the first N^* candidates. For each test set, NMT model, reranker, and window size, we select N^* from among multiples of eight to match QUIT's end-to-end cost as closely as possible. We otherwise follow the evaluation setup in Section 4.1 and use paired-bootstrap 95% confidence intervals to compare the two methods.

Tables 5–7 report the complete results. The two methods achieve nearly identical speedups by construction. Across most evaluation metrics, their quality differences are not significant; the significant differences that remain are small and show no consistent winner.

This comparison favors Fixed-N^* because N^* is selected post hoc for each test set, NMT model, and reranker to match the cost of a QUIT configuration that has passed the equivalence tests on nearly all external quality metrics. For each evaluated window size, QUIT instead uses a shared (w, α) configuration and makes reference-free stopping decisions for each source sentence. Indeed, the oracle N^* varies from 128 to 352 for $w = 8$ and from 248 to 456 for $w = 15$. These results suggest that a shared, adaptive stopping criterion can replace reference-dependent, setting-specific candidate-budget tuning.

Test set / NMT model	Reranker	System	Speedup	Rerank.	ChrF++	xCOMET	MetricX	GEMBA
WMT24 Qwen3	MBR	Unaccelerated (512)	1.00	.8357	42.09	.7520	5.17	−7.81
		QUIT	1.74	**.8344**	42.04	.7497	**5.19**	−7.89
		Fixed-N^* ($N^* = 304$)	1.73	.8340	41.99	.7499	5.20	−7.87
	QE	Unaccelerated (512)	1.00	.8223	39.61	.7585	4.96	−7.94
		QUIT	3.64	.8133	**40.46**	.7524	5.13	−7.96
		Fixed-N^* ($N^* = 144$)	3.56	**.8143**	40.19	.7529	5.13	−8.02
WMT24 TranslateGemma	MBR	Unaccelerated (512)	1.00	.8570	44.90	.8521	3.15	−3.17
		QUIT	2.66	.8566	44.88	.8509	3.16	−3.18
		Fixed-N^* ($N^* = 200$)	2.64	.8564	44.91	.8511	3.16	−3.18
	QE	Unaccelerated (512)	1.00	.8376	43.59	.8570	3.14	−3.12
		QUIT	4.01	.8327	**44.37**	.8553	3.15	−3.01
		Fixed-N^* ($N^* = 128$)	4.00	**.8333**	44.19	.8549	3.15	−3.02
WMT24 Hy-MT2	MBR	Unaccelerated (512)	1.00	.8731	49.86	.8867	2.92	−1.81
		QUIT	2.66	.8728	49.86	.8867	2.93	−1.79
		Fixed-N^* ($N^* = 200$)	2.62	.8728	49.87	.8864	2.93	−1.78
	QE	Unaccelerated (512)	1.00	.8366	47.67	.8801	2.98	−2.39
		QUIT	4.12	.8317	**48.54**	**.8821**	**2.96**	−2.08
		Fixed-N^* ($N^* = 128$)	4.00	**.8323**	48.29	.8811	2.98	−2.13
WMT25 Qwen3	MBR	Unaccelerated (512)	1.00	.7345	34.06	.3696	10.86	−18.11
		QUIT	1.47	.7322	33.98	.3671	10.95	−18.21
		Fixed-N^* ($N^* = 352$)	1.46	**.7328**	**34.03**	.3670	10.93	−18.27
	QE	Unaccelerated (512)	1.00	.7106	30.32	.3705	10.10	−18.25
		QUIT	3.43	.6935	31.28	.3613	10.54	−18.43
		Fixed-N^* ($N^* = 152$)	3.37	**.6948**	31.29	.3617	10.51	−18.36
WMT25 TranslateGemma	MBR	Unaccelerated (512)	1.00	.8129	36.25	.5731	6.06	−6.78
		QUIT	2.61	.8112	36.18	.5690	6.12	−6.89
		Fixed-N^* ($N^* = 200$)	2.58	.8113	36.15	.5700	6.11	−6.88
	QE	Unaccelerated (512)	1.00	.7721	36.17	.5670	6.11	−6.83
		QUIT	3.55	.7657	36.16	.5645	6.18	−6.89
		Fixed-N^* ($N^* = 144$)	3.56	**.7662**	36.19	.5650	6.18	−6.96
WMT25 Hy-MT2	MBR	Unaccelerated (512)	1.00	.8298	41.19	.6281	5.62	−4.25
		QUIT	2.36	.8289	41.18	.6262	5.65	−4.23
		Fixed-N^* ($N^* = 216$)	2.38	.8290	41.20	.6264	5.66	−4.15
	QE	Unaccelerated (512)	1.00	.7786	40.78	.6219	5.71	−4.92
		QUIT	3.59	.7725	40.90	.6185	5.77	−4.75
		Fixed-N^* ($N^* = 144$)	3.56	**.7730**	40.89	.6187	5.76	−4.79

Table 5: Budget-matched comparison for $w = 8$. Boldface indicates a significantly better score in the comparison between QUIT and Fixed-N^*, based on paired-bootstrap 95% confidence intervals. Lower MetricX scores are better; higher scores are better for all other metrics.

Test set / NMT model	Reranker	System	Speedup	Rerank.	ChrF++	xCOMET	MetricX	GEMBA
WMT24 Qwen3	MBR	Unaccelerated (512)	1.00	.8357	42.09	.7520	5.17	−7.81
		QUIT	1.33	**.8352**	42.04	.7509	**5.17**	−7.87
		Fixed-N^* ($N^* = 392$)	1.32	.8348	42.00	.7506	5.18	−7.87
	QE	Unaccelerated (512)	1.00	.8223	39.61	.7585	4.96	−7.94
		QUIT	2.33	.8165	**40.33**	.7554	5.05	−7.88
		Fixed-N^* ($N^* = 216$)	2.37	**.8170**	40.13	.7547	5.06	−7.96
WMT24 TranslateGemma	MBR	Unaccelerated (512)	1.00	.8570	44.90	.8521	3.15	−3.17
		QUIT	1.89	.8568	44.88	.8511	3.16	−3.18
		Fixed-N^* ($N^* = 280$)	1.87	.8568	44.91	.8515	3.16	−3.15
	QE	Unaccelerated (512)	1.00	.8376	43.59	.8570	3.14	−3.12
		QUIT	2.57	.8344	**44.29**	.8562	3.14	−3.03
		Fixed-N^* ($N^* = 200$)	2.56	**.8347**	44.10	.8561	3.15	−3.06
WMT24 Hy-MT2	MBR	Unaccelerated (512)	1.00	.8731	49.86	.8867	2.92	−1.81
		QUIT	1.87	.8730	49.85	.8865	2.93	−1.80
		Fixed-N^* ($N^* = 280$)	1.86	.8731	49.85	.8868	2.93	−1.77
	QE	Unaccelerated (512)	1.00	.8366	47.67	.8801	2.98	−2.39
		QUIT	2.65	.8332	**48.34**	**.8819**	2.96	−2.12
		Fixed-N^* ($N^* = 192$)	2.67	**.8337**	48.05	.8810	2.97	−2.19
WMT25 Qwen3	MBR	Unaccelerated (512)	1.00	.7345	34.06	.3696	10.86	−18.11
		QUIT	1.20	.7336	34.01	.3683	10.90	−18.22
		Fixed-N^* ($N^* = 424$)	1.21	.7334	**34.05**	.3687	10.91	−18.19
	QE	Unaccelerated (512)	1.00	.7106	30.32	.3705	10.10	−18.25
		QUIT	2.19	.6996	31.05	.3646	10.38	−18.36
		Fixed-N^* ($N^* = 232$)	2.21	**.7006**	31.02	.3652	10.36	−18.29
WMT25 TranslateGemma	MBR	Unaccelerated (512)	1.00	.8129	36.25	.5731	6.06	−6.78
		QUIT	1.89	.8119	36.16	.5704	6.10	−6.84
		Fixed-N^* ($N^* = 272$)	1.89	.8120	36.16	.5706	6.09	−6.85
	QE	Unaccelerated (512)	1.00	.7721	36.17	.5670	6.11	−6.83
		QUIT	2.29	.7680	36.17	.5651	6.16	−6.88
		Fixed-N^* ($N^* = 224$)	2.29	**.7684**	36.17	.5650	6.14	−6.89
WMT25 Hy-MT2	MBR	Unaccelerated (512)	1.00	.8298	41.19	.6281	5.62	−4.25
		QUIT	1.72	.8292	41.16	.6278	5.64	−4.23
		Fixed-N^* ($N^* = 296$)	1.74	**.8294**	41.19	.6281	5.63	−4.20
	QE	Unaccelerated (512)	1.00	.7786	40.78	.6219	5.71	−4.92
		QUIT	2.29	.7747	40.88	.6197	5.75	−4.82
		Fixed-N^* ($N^* = 224$)	2.29	**.7751**	40.88	.6198	5.76	−4.82

Table 6: Budget-matched comparison for $w = 12$, following the conventions of Table 5.

Test set / NMT model	Reranker	System	Speedup	Rerank.	ChrF++	xCOMET	MetricX	GEMBA
WMT24 Qwen3	MBR	Unaccelerated (512)	1.00	.8357	42.09	.7520	5.17	−7.81
		QUIT	1.19	**.8355**	42.06	.7518	5.17	−7.84
		Fixed-N^* ($N^* = 432$)	1.20	.8352	42.06	.7514	5.17	−7.84
	QE	Unaccelerated (512)	1.00	.8223	39.61	.7585	4.96	−7.94
		QUIT	1.84	.8180	**40.24**	.7570	**5.01**	−7.86
		Fixed-N^* ($N^* = 280$)	1.83	**.8186**	40.01	.7561	5.04	−7.93
WMT24 TranslateGemma	MBR	Unaccelerated (512)	1.00	.8570	44.90	.8521	3.15	−3.17
		QUIT	1.59	.8569	44.87	.8514	3.15	−3.17
		Fixed-N^* ($N^* = 328$)	1.59	.8570	**44.92**	.8517	3.15	−3.15
	QE	Unaccelerated (512)	1.00	.8376	43.59	.8570	3.14	−3.12
		QUIT	2.02	.8352	**44.23**	**.8570**	3.14	−3.01
		Fixed-N^* ($N^* = 256$)	2.00	**.8356**	44.02	.8562	3.15	−3.06
WMT24 Hy-MT2	MBR	Unaccelerated (512)	1.00	.8731	49.86	.8867	2.92	−1.81
		QUIT	1.58	.8730	49.82	.8865	2.93	−1.81
		Fixed-N^* ($N^* = 328$)	1.58	**.8732**	**49.85**	**.8871**	2.92	−1.79
	QE	Unaccelerated (512)	1.00	.8366	47.67	.8801	2.98	−2.39
		QUIT	2.08	.8340	**48.24**	**.8816**	2.97	−2.18
		Fixed-N^* ($N^* = 248$)	2.06	**.8345**	47.94	.8809	2.97	−2.21
WMT25 Qwen3	MBR	Unaccelerated (512)	1.00	.7345	34.06	.3696	10.86	−18.11
		QUIT	1.12	.7340	34.04	.3687	10.88	−18.13
		Fixed-N^* ($N^* = 456$)	1.13	.7339	34.07	.3689	10.89	−18.15
	QE	Unaccelerated (512)	1.00	.7106	30.32	.3705	10.10	−18.25
		QUIT	1.74	.7029	30.87	.3662	10.29	−18.29
		Fixed-N^* ($N^* = 296$)	1.73	**.7038**	30.81	.3670	**10.26**	−18.23
WMT25 TranslateGemma	MBR	Unaccelerated (512)	1.00	.8129	36.25	.5731	6.06	−6.78
		QUIT	1.59	.8123	36.20	.5714	6.09	−6.85
		Fixed-N^* ($N^* = 328$)	1.57	.8122	36.19	.5713	6.08	−6.81
	QE	Unaccelerated (512)	1.00	.7721	36.17	.5670	6.11	−6.83
		QUIT	1.80	.7692	36.20	.5656	6.14	−6.85
		Fixed-N^* ($N^* = 288$)	1.78	**.7695**	36.20	.5660	6.14	−6.82
WMT25 Hy-MT2	MBR	Unaccelerated (512)	1.00	.8298	41.19	.6281	5.62	−4.25
		QUIT	1.46	.8295	41.17	.6277	5.64	−4.19
		Fixed-N^* ($N^* = 352$)	1.46	**.8296**	41.18	.6280	5.64	−4.23
	QE	Unaccelerated (512)	1.00	.7786	40.78	.6219	5.71	−4.92
		QUIT	1.81	.7759	40.84	.6208	5.75	−4.82
		Fixed-N^* ($N^* = 280$)	1.83	**.7761**	40.85	.6206	**5.73**	−4.80

Table 7: Budget-matched comparison for $w = 15$, following the conventions of Table 5.

C Convergence-threshold Ablation

We evaluate $\alpha \in \{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$ with $w \in \{8, 12, 15\}$ on the complete WMT24 and WMT25 test sets, while keeping all other settings unchanged from Section 4.1. We report the mean number of generated candidates, the proportion of source sentences for which generation reaches the full candidate budget of 512, end-to-end speedup, and the same reranking and translation quality metrics as in the main experiments.

Tables 8–13 show macro-averages over the six test-set–model combinations. We use paired-bootstrap tests with 10,000 resamples against both the default $\alpha = 10^{-3}$ and the unaccelerated baseline. The final column counts how often the tested configuration performs significantly worse or better than the default configuration across the 30 test-set–model–metric comparisons. MetricX is sign-reversed for testing but reported in its original lower-is-better orientation. "Rerank." denotes the COMET-22 score for MBR and the CometKiwi-22 score for QE.

For MBR, $\alpha = 10^{-3}$ marks a turning point in the trade-off between speedup and translation quality. Tightening it by one order of magnitude causes generation to reach the full candidate budget for more than 80% of source sentences and reduces speedup to approximately $1\times$, without any significant improvement over the unaccelerated baseline. Relaxing it to 10^{-2} more than doubles speedup relative to the default, but yields significantly worse values in 20–21 of the 30 comparisons; the mean effect remains small at approximately -0.01σ.

QE is less sensitive to small positive convergence thresholds. For QE, the best reranking score $R_i(s)$ is a running maximum. If it does not increase within a window, the uncertainty score $\Delta_b^R(s)$ is exactly zero, so the stopping criterion is satisfied for any $\alpha > 0$. Consequently, even $\alpha = 10^{-6}$ retains much of the speedup. A detectable but small degradation, approximately -0.006σ on average, emerges only when the threshold is relaxed to 10^{-2}. Overall, QE is relatively insensitive to convergence thresholds below the default, whereas MBR shows a stronger trade-off between speedup and translation quality.

α	$\overline{N}$	Full (%)	Speedup	Rerank.	ChrF++	xCOMET	MetricX	GEMBA	$n_\downarrow/n_\uparrow$
10^{-6}	504	98.1	1.00	.8239	41.39	.6770	5.63	−6.97	0/20
10^{-5}	504	98.1	1.00	.8239	41.39	.6770	5.63	−6.97	0/20
10^{-4}	481	82.6	1.06	.8239	41.38	.6770	5.63	−6.97	0/20
$\mathbf{10^{-3}}$	**252**	**12.1**	**2.25**	**.8227**	**41.35**	**.6749**	**5.67**	**−7.03**	–
10^{-2}	95	0.0	5.83	.8198	41.30	.6716	5.74	−7.16	21/0

Table 8: α ablation for MBR with $w = 8$. "Full (%)" is the percentage of source sentences that reach 512 candidates; $n_\downarrow$ and $n_\uparrow$ count comparisons in which the tested configuration performs significantly worse and better than the default, respectively.

α	$\overline{N}$	Full (%)	Speedup	Rerank.	ChrF++	xCOMET	MetricX	GEMBA	$n_\downarrow/n_\uparrow$
10^{-6}	505	98.4	1.00	.8239	41.39	.6770	5.63	−6.97	0/19
10^{-5}	505	98.3	1.00	.8239	41.39	.6770	5.63	−6.97	0/19
10^{-4}	499	91.6	1.02	.8239	41.38	.6770	5.63	−6.97	0/18
$\mathbf{10^{-3}}$	**331**	**20.7**	**1.65**	**.8233**	**41.35**	**.6758**	**5.65**	**−7.02**	–
10^{-2}	143	0.0	3.84	.8212	41.32	.6732	5.71	−7.10	20/0

Table 9: α ablation for MBR with $w = 12$, following Table 8.

α	$\overline{N}$	Full (%)	Speedup	Rerank.	ChrF++	xCOMET	MetricX	GEMBA	$n_\downarrow/n_\uparrow$
10^{-6}	506	98.5	1.00	.8239	41.39	.6770	5.63	−6.97	0/16
10^{-5}	506	98.5	1.00	.8239	41.39	.6770	5.63	−6.97	0/16
10^{-4}	503	95.1	1.01	.8239	41.39	.6770	5.63	−6.97	0/16
$\mathbf{10^{-3}}$	**376**	**27.8**	**1.42**	**.8235**	**41.36**	**.6762**	**5.64**	**−7.00**	–
10^{-2}	179	0.1	3.06	.8218	41.35	.6741	5.69	−7.06	20/0

Table 10: α ablation for MBR with $w = 15$, following Table 8.

α	$\overline{N}$	Full (%)	Speedup	Rerank.	ChrF++	xCOMET	MetricX	GEMBA	$n_\downarrow/n_\uparrow$
10^{-6}	144	0.0	3.41	.7852	40.29	.6727	5.62	−7.19	0/8
10^{-5}	144	0.0	3.41	.7852	40.29	.6727	5.62	−7.19	0/8
10^{-4}	142	0.0	3.45	.7852	40.29	.6726	5.62	−7.19	1/8
$\mathbf{10^{-3}}$	**133**	**0.0**	**3.72**	**.7849**	**40.29**	**.6723**	**5.62**	**−7.19**	–
10^{-2}	97	0.0	5.25	.7834	40.32	.6711	5.65	−7.21	14/2

Table 11: α ablation for QE with $w = 8$, following the conventions of Table 8.

α	$\overline{N}$	Full (%)	Speedup	Rerank.	ChrF++	xCOMET	MetricX	GEMBA	$n_\downarrow/n_\uparrow$
10^{-6}	225	0.4	2.18	.7880	40.17	.6739	5.57	−7.19	1/7
10^{-5}	225	0.4	2.18	.7880	40.17	.6739	5.57	−7.19	1/7
10^{-4}	223	0.3	2.21	.7880	40.17	.6738	5.57	−7.19	1/7
$\mathbf{10^{-3}}$	**208**	**0.1**	**2.39**	**.7877**	**40.18**	**.6738**	**5.57**	**−7.19**	–
10^{-2}	151	0.0	3.38	.7863	40.19	.6725	5.60	−7.20	13/0

Table 12: α ablation for QE with $w = 12$, following Table 8.

α	$\overline{N}$	Full (%)	Speedup	Rerank.	ChrF++	xCOMET	MetricX	GEMBA	$n_{\downarrow}/n_{\uparrow}$
10^{-6}	284	2.5	1.73	.7894	40.10	.6749	5.55	-7.17	1/9
10^{-5}	284	2.5	1.73	.7894	40.10	.6749	5.55	-7.17	1/8
10^{-4}	282	2.2	1.75	.7894	40.10	.6749	5.55	-7.17	1/9
$\mathbf{10^{-3}}$	**263**	**1.0**	**1.88**	**.7892**	**40.10**	**.6747**	**5.55**	$\mathbf{-7.17}$	–
10^{-2}	192	0.1	2.66	.7878	40.13	.6738	5.57	-7.18	12/3

Table 13: α ablation for QE with $w = 15$, following Table 8.

D PRR Evaluation

We evaluate a 10% sample of source sentences, replaying the full 512-candidate trajectories with $k = w = 8$. Each source–window pair (s, b) is one instance, for $b = 1, \ldots, 7$, corresponding to candidate counts 64 through 448. We exclude the final window because it already uses the full candidate pool and thus has zero early-stopping risk by construction. We evaluate all seven windows, irrespective of where QUIT would stop, and pool instances within each test-set–model–reranker–quality-metric combination.

The risk is $\mathcal{L}_b(s)$: the absolute difference between the reference-based quality of the output selected at step bw and that of the output selected from all 512 candidates by the same reranker. We compute it separately with xCOMET and GEMBA. The uncertainty score is $\Delta_b^R(s)$ in Table 3 and $\Delta_b^Q(s)$ in Table 14; in the latter case, the same quality metric is used for both the uncertainty score and the risk.

For each uncertainty score, we progressively reject instances in descending score order and record the mean $\mathcal{L}_b(s)$ among the retained instances. The oracle curve instead rejects instances in descending order of $\mathcal{L}_b(s)$, while the expected random-ranking curve stays at the overall mean risk. We integrate these curves and compute PRR with Equation 14.

Table 14 evaluates the reference-based local uncertainty signal $\Delta_b^Q(s)$ from Equation 10 against the same early-stopping risk. All values are positive, but access to reference-based within-window quality does not yield oracle risk ranking. Indeed, $\Delta_b^R(s)$ has higher mean PRR in three of the four reranker–metric groups. This comparison illustrates the gap between local changes in the best available candidate quality and the eventual quality difference between reranker-selected outputs: even $\Delta_b^Q(s)$ is only a local indicator, not an oracle score for $\mathcal{L}_b(s)$.

E Component-wise Runtime

Tables 15 and 16 report the complete wall-clock measurements underlying Table 4. The Generation and Reranking columns report candidate-generation time and MBR reranking time, respectively, in seconds. The aggregate measurements cover 12,010 source segments per model for WMT24 and 5,232 per model for WMT25; the per-segment averages in Table 4 pool all 51,726 segment–model runs.

Test set	NMT model	MBR		QE	
		xCOMET	GEMBA	xCOMET	GEMBA
WMT24	Qwen3	0.231	0.171	0.231	0.149
WMT24	Hy-MT2	0.258	0.097	0.208	0.053
WMT24	TranslateGemma	0.217	0.127	0.216	0.089
WMT25	Qwen3	0.148	0.138	0.189	0.164
WMT25	Hy-MT2	0.253	0.102	0.161	0.073
WMT25	TranslateGemma	0.108	0.099	0.153	0.090
Average		**0.203**	**0.122**	**0.193**	**0.103**

Table 14: Pooled PRR with reference-based uncertainty $\Delta_b^Q(s)$ and the same early-stopping risk $\mathcal{L}_b(s)$ as in Table 3. The sampling, windows, and column order match the main-text table. This reference-based signal is used only for retrospective analysis.

Method	w	Qwen3		TranslateGemma		Hy-MT2	
		Generation	Reranking	Generation	Reranking	Generation	Reranking
QUIT	3	55,125.1	1,140.17	44,194.7	729.717	62,264	513.179
QUIT	5	100,523	3,719.60	79,026.1	2,216.81	112,471	1,618.09
QUIT	8	155,297	7,524.66	121,476	4,395.06	178,592	3,587.20
QUIT	12	202,580	10,963.6	169,529	7,143.55	253,174	6,200.59
QUIT	15	225,530	12,749.1	201,717	9,021.52	299,299	8,053.21
QUIT	20	246,466	14,558.6	246,847	11,795.0	365,539	10,868.0
QUIT	30	262,880	16,181.4	304,428	15,520.8	446,931	14,798.9
Unaccelerated MBR	∞	266,352	16,803.0	317,548	16,803.0	467,484	16,803.0
PMBR	–	266,352	7,301.66	317,548	6,079.76	467,484	5,497.37
PruneMBR	–	266,352	2,978.32	317,548	4,977.67	467,484	8,002.77

Table 15: Complete candidate-generation and MBR reranking wall-clock times (seconds) on WMT24 with a maximum budget of 512 candidates.

Method	w	Qwen3		TranslateGemma		Hy-MT2	
		Generation	Reranking	Generation	Reranking	Generation	Reranking
QUIT	3	107,026	530.372	76,505.1	283.198	121,034	316.009
QUIT	5	202,400	1,818.97	135,144	806.749	213,211	869.407
QUIT	8	297,718	3,641.88	197,999	1,478.28	326,227	1,706.51
QUIT	12	363,642	5,136.86	273,125	2,525.07	447,977	2,862.44
QUIT	15	390,797	5,846.22	325,003	3,377.58	525,377	3,775.98
QUIT	20	414,240	6,550.99	400,981	4,812.30	626,408	5,093.12
QUIT	30	432,436	7,183.47	495,210	6,833.80	744,970	6,900.09
Unaccelerated MBR	∞	435,665	7,320.00	514,263	7,320.00	767,318	7,320.00
PMBR	–	435,665	3,030.74	514,263	2,637.34	767,318	2,741.44
PruneMBR	–	435,665	284.149	514,263	295.944	767,318	423.401

Table 16: Complete candidate-generation and MBR reranking wall-clock times (seconds) on WMT25 with a maximum budget of 512 candidates.