When Reranking Hurts: Uncertainty-Based Gating for Few-Shot Reranking

Anonymous ACL submission

Abstract

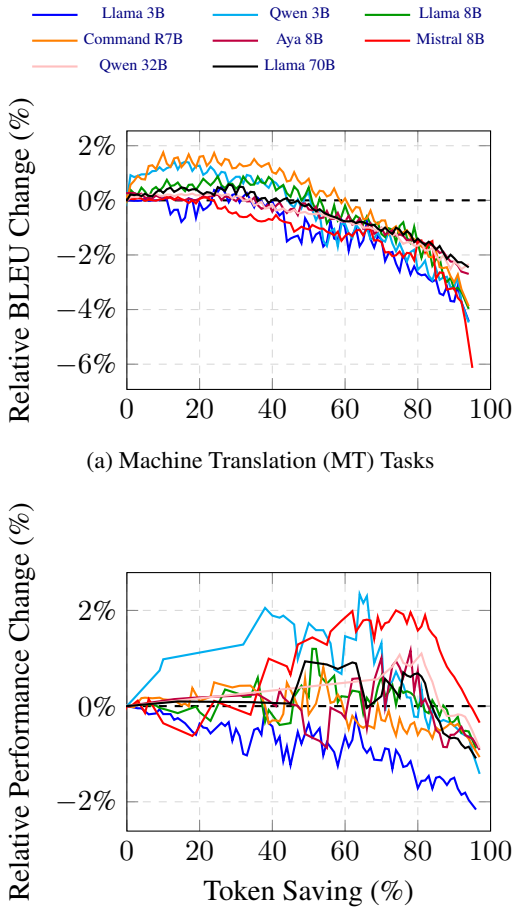
Few-shot selection typically assumes that reranking retrieved examples always improves performance. We challenge this view by identifying that the expensive reranking step can in fact degrade performance. Instead, we propose *Training-Free Gated Reranking*, which decides whether to rerank the few-shot examples based on the model's uncertainty. Extensive experiments across 8 LLMs, covering 7 NLU datasets and 9 MT domain-language combinations, demonstrate that our approach reduces computational costs by 15%-80% while improving average performance by up to 2%. These findings indicate that higher computational cost does not guarantee better performance, and that reranking is most beneficial when targeted at high-uncertainty instances.

1 Introduction

Recent work has found that adaptive few-shot selection can improve the performance of LLMs on in-context learning tasks (Agrawal et al., 2023; Chitale et al., 2024; Liu et al., 2022). In particular, several works adopt a "retrieve-then-rerank" approach, fetching a broad pool of candidates and using trained cross-encoders (Li et al., 2023; Wang et al., 2024b; Rubin et al., 2022) or training-free scorers to select demonstrations for a specific inference sample (Wu et al., 2023; Peng et al., 2024). However, reranking with repeated LLM calls increases computational costs, total token consumption can increase by 13.4× for MT and by 29.5× for NLU.[1]

To address these concerns, we propose *Training-Free Gated Reranking*, a simple yet effective approach, which uses the perplexity of the LLM's initial generation to rerank examples only when the model is uncertain. Rather than reranking every instance, we apply the reranking step exclusively when the input's perplexity falls above a predefined threshold.

Our approach improves efficiency while matching or slightly surpassing performance across various models and datasets. Figure 1 shows that tun-



(a) Machine Translation (MT) Tasks

(b) Natural Language Understanding (NLU)

Figure 1: Relative performance impact versus computational token savings achieved via selective reranking for Machine Translation (MT, measured in BLEU) and Natural Language Understanding (NLU) tasks. The majority of models (3B–70B) exhibit performance gains, though the optimal savings window shifts from 15–50% for MT to roughly 50–80% for most NLU models.

[1] Calculated per task, as the total sum of tokens across all models and datasets.

ing the perplexity threshold for reranking yields computational savings of 15%-80% while maintaining or surpassing baseline performance. Our manual analysis confirms that reranking provides clear gains for difficult instances, but can degrade performance when applied to easy instances. Overall, our approach presents a convenient tradeoff between resources and performance across 8 models ranging from 3B to 70B parameters, covering 9 combinations of language pairs and specialized domains, as well as 7 NLU datasets.[2]

2 Perplexity-Based Reranking of Few-Shot Examples

We propose a simple approach that adaptively allocates computational resources to choose few-shot examples based on the model's intrinsic uncertainty.

Formal definitions. Given an LLM M, an input x and a large pool of candidate demonstrations $\mathcal{C} = ((x_1, y_1), \ldots, (x_n, y_n))$, each consisting of an input and a corresponding gold label, ranked according to their relevance to x, we are looking for a function:

$$T(M, x, \mathcal{C}) \mapsto \{c_1, \ldots, c_k | c_i \in \mathcal{C}\} \quad (1)$$

That is, a selection of k demonstration samples from the pool according to x.

Method. We start by generating a prediction without reranking $\hat{y} = M(x, c_1, \ldots, c_k)$, and use it to compute a normalized uncertainty score $U \in [0, 1]$. This score is derived from the inverse perplexity of the generation:

$$\mathrm{U}(M, x, \mathcal{C}) = 1 - \mathrm{PPL}_M(\hat{y} \mid x, c_1, \ldots, c_k)^{-1} \quad (2)$$

Where, and PPL computes the conditional perplexity of $\hat{y}$ on x and the first k examples. U is in the range $[0, 1]$, where larger values indicate that the model is less certain in its own prediction.

We then define our demonstration selection function T based on a predefined uncertainty threshold τ:

$$T(M, x, \mathcal{C}) = \begin{cases} ((x_1, y_1), \ldots (x_k, y_k)) & U \leq \tau, \\ \mathrm{rerank}(M, x, \mathcal{C}) & U > \tau \end{cases} \quad (3)$$

where $rerank$ is defined as the conditional entropy rerank (Peng et al., 2024):

$$\mathrm{rerank}(M, x, \mathcal{C}) = \underset{(x_i, y_i) \in \mathcal{C}}{\text{Top-k}} \left(-\mathrm{PPL}_M(x \mid (x_i, y_i)) \right) \quad (4)$$

In practice, we calibrate the threshold τ to maximize performance on a development set. During this calibration phase, we apply a moving average with a window size of 5 to smooth local variance.

3 Evaluation

3.1 Experimental Setup

We list below key experimental details, see more details in the Appendix.

Tasks and models. We evaluate our approach across 8 LLMs spanning 3B–70B parameters, compared to a full reranking baseline (Peng et al., 2024), assessing performance on both natural language understanding (NLU) and machine translation (MT) tasks. Specifically, we adopt the NLU benchmarks used by Peng et al. (2024): SST-2, SST-5 (Socher et al., 2013), CR, AgNews, Subj (Wang et al., 2019), MNLI (Williams et al., 2018), and QNLI (Wang et al., 2019). For MT, we use three domain specific corpora (Koehn and Knowles, 2017): EMEA (medical) (Tiedemann, 2012), JRC-Acquis (legal) (Steinberger et al., 2006), and KDE (technical) (Tiedemann, 2012). We consider both translation directions between English and Spanish, Portuguese, and German, yielding six directions per domain. We sample 1,000 parallel sentence pairs for each domain-direction combination, partitioned into 200 development examples for threshold tuning and 800 test examples for evaluation. All results are averaged over 20 random splits for statistical robustness.

Baselines. Baselines include retrieval-based k-shot selection with BM25 (Robertson and Zaragoza, 2009) and dense retrieval using e5-base-v2 or multilingual-e5-base depending on the language (Wang et al., 2022, 2024a), and always-rerank ($\tau = 0$, reranking for every input at $O(N)$). We also compare gating signals to simple input proxies (e.g., source length, source entropy)

[2] Code and experimental data are provided in the supplementary material for review, and will be made publicly available upon acceptance.

Method	Global (All)		Small Models		Bigger Models	
	Perf. ↑	Save. %	Perf. ↑	Save. %	Perf. ↑	Save. %
Machine Translation (BLEU / COMET)						
No Reranking	37.09/81.05	100.0	33.37/79.46	100.0	38.33/81.57	100.0
Full Reranking	38.32/81.37	0.0	34.77/79.89	0.0	39.51/81.86	0.0
Gating (dev-set calibrated)	38.42†/81.39	20.86	34.95†/79.94	20.52	39.58†/81.88	20.98
Gating (test-set calibrated)	**38.69**†/**81.43**	17.42	**35.19**†/**79.98**	19.22	**39.85**†/**81.92**	16.82
Natural Language Understanding (Accuracy)						
No Reranking	79.88	100.0	77.16	100.0	80.78	100.0
Full Reranking	80.73	0.0	78.45	0.0	81.50	0.0
Gating (dev-set calibrated)	80.81	54.3	78.60	50.2	81.55	55.6
Gating (test-set calibrated)	**81.41**†	47.16	**79.30**†	46.69	**82.11**†	47.32

Table 1: Performance summary across tasks and model sizes. Performance (Perf.) shows BLEU/COMET for MT and Accuracy for NLU. **Bold** indicates row maxima; underlined values exceed Full Reranking performance. Gating (dev-set calibrated) uses dev-set τ, while Gating (test-set calibrated) uses test-set τ. † denotes a statistically significant improvement over No Reranking ($p < 0.05$). Exact p-values are provided for transparency.

and an uncertainty proxy (token-level logit gap - TARG) (Wang et al., 2025).

Hyperparameters and metrics. We fix $k = 5$ demonstrations and a candidate pool of $N = 100$, report MT quality with BLEU (Papineni et al., 2002) and COMET (wmt22-comet-da) (Rei et al., 2022), and NLU quality with accuracy. Following scaling laws for inference compute (Kaplan et al., 2020) and standard open-weight serving metrics (Griggs et al., 2024), we evaluate efficiency using the unweighted sum of input and output tokens across all processing stages: (a) initial draft generation, (b) second-pass generation after reranking for triggered cases, and (c) all conditional perplexity computations over the $N = 100$ candidates. We report relative savings against an always-reranking baseline.

3.2 Results

Our experiments yield several interesting observations.

Always reranking is not the optimal policy for few-shot selection in MT and NLU. Applying reranking to every query is highly inefficient and unnecessary for maintaining quality. Table 1 demonstrates that our practical Empirical Gating method reduces average computational costs by 20.86% and 54.3% for MT and NLU, while achieving performance on par with Full Reranking (38.42 vs 38.32 avg BLEU; 80.81 vs 80.73 avg Accuracy). Furthermore, using a post-hoc test-set calibrated gating threshold achieves comparable cost reductions (17.42% / 47.16%) while slightly increasing Full Reranking performance (+0.37 avg BLEU, +0.68 avg Accuracy), demonstrating there is room for improvement.

Selective reranking via uncertainty can improve performance, especially for smaller models. Selective reranking remains highly competitive with the Full Reranking baseline across model scales. Table 1 highlights that this dynamic is most pronounced for small models, Full Reranking performance (34.77 avg BLEU / 78.45 avg Accuracy) is slightly improved upon by our retrospective Gating (test-set calibrated) (+0.42 avg BLEU / +0.85 avg Accuracy) and practical Gating (dev-set calibrated) results (+0.18 avg BLEU / +0.15 avg Accuracy) while yielding average token savings of 19.22% / 46.69% and 20.52% / 50.2%. Larger models in the optimal Gating (test-set calibrated) settings show smaller yet positive margins (+0.34 avg BLEU / +0.61 avg Accuracy), while the Gating (dev-set calibrated) yields strong savings (20.98% and 55.6%) and comparable performance (+0.07 avg BLEU / +0.05 avg Accuracy) demonstrating that selective application remains a highly efficient alternative even at scale.

Selective reranking via uncertainty provides efficient way to optimize performance with minimal impact on task performance. As Figure 1 illustrates, adjusting the gating threshold provides a way to balance cost and quality. This tradeoff relies on the correlation between the LLM's initial uncertainty and the utility of the reranker. This al-

Strategy	Task	Score	Savings
No Reranking	MT	37.09	100.0%
	NLU	79.88	100.0%
Full Reranking	MT	38.32	0.0%
	NLU	80.73	0.0%
TARG	MT	38.15	8.6%
	NLU	80.46	35.6%
Source Entropy	MT	38.19	11.8%
	NLU	80.46	42.2%
Word Length	MT	38.11	17.1%
	NLU	80.37	38.1%
Uncertainty (Ours)	MT	**38.42**	20.9%
	NLU	**80.81**	54.3%

Table 2: Comparison of gating indicators averaged across all models for MT and NLU. Thresholds (τ) for each strategy are calibrated on a development set and evaluated on a test set. Uncertainty achieves the best performance-efficiency balance.

Error Cause	MT	NLU	Overall
Reranking errors			
Structural Template	21	23	44
High Variance	4	7	11
Other	5	0	5
No Reranking errors			
Domain/Term Shift	15	8	23
High Variance	9	9	18
Relational Mapping	0	9	9
Bleu Misleading	4	0	4
Other	2	4	6

Table 3: Distribution of error causes when reranking degrades performance compared to baseline vs. when reranking improved performance on translation and NLU tasks.

lows us us to to reduce average token consumption from 15% to 80%.

Uncertainty gating outperforms other gating methods. Alternative gating signals fail to achieve the same balance of quality and efficiency. As shown in Table 2, where we compare our approach against several alternative baseline metrics used for the same gating purpose: the perplexity of the input text (Source Entropy), the count of words (Word Length), along with TARG (Wang et al., 2025).

4 Manual Qualitative Analysis

To understand the mechanics driving the performance trade-offs of uncertainty-based gating, we manually annotate and analyze instances with the highest and lowest uncertainty scores. Specifically, we examine some of the top 30 instances (highest uncertainty, reranking beneficial) and bottom 30 instances (lowest uncertainty, reranking degraded) across both Machine Translation (MT) and Natural Language Understanding (NLU) tasks to diagnose how the few-shot examples influence the model's generation. See Table 3 and Appendix Table 7.

High uncertainty reveals weak baseline retrieval. Under high uncertainty, the baseline frequently retrieves misleading demonstrations, such as out-of-domain examples in MT (15 instances) or missed Relational Mappings in NLU (9 instances). The reranker resolves these issues by correctly identifying the task's structural logic and domain constraints. By successfully retrieving examples that enforce low variance, specialized terminology or exact relationships (e.g., mapping a neighborhood to a borough), reranking effectively grounds the model and reduces hallucinations.

Low uncertainty indicates strong baseline retrieval. For low-uncertainty instances, the baseline already retrieves highly relevant, format-preserving contexts. Here, reranking can actually degrade performance. The primary cause across both MT (21 instances) and NLU (23 instances) is the disruption of Structural Templates. While the baseline retrieves near-exact syntactic templates, the reranker often overrides them with examples that are topically relevant but structurally varied. Furthermore, reranking can introduce high-variance examples with divergent meanings or lexical distractions that confuse the model. (Note: A small subset of MT errors were "BLEU Misleading," where the reranker's valid translation was unfairly penalized by the metric

5 Conclusion

We show that reranking few-shot examples is not always beneficial and can even hurt performance. Instead, we introduce *Training-Free Gated Reranking*, which selectively ranks when the model is uncertain, reducing computation by 15%–80% while improving average performance by up to 2%. Overall, our findings suggest that targeted reranking can break the usual tradeoff between higher computational cost and better downstream quality.

6 Limitations

Our approach presents three primary limitations. First, the gating mechanism relies on an empirically calibrated threshold from a development set; distribution shifts between development and test

data can cause suboptimal thresholding. Second, calculating the uncertainty score requires generating an initial, non-reranked prediction. For tasks with long output sequences, the latency of this initial step may reduce the efficiency gained from skipping the reranking phase. Finally, the method requires access to token-level perplexity, limiting its use to models and APIs that expose logits.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.

Pranjal A. Chitale, Jay Gala, and Raj Dabre. 2024. An empirical study of in-context learning in llms for machine translation.

Tyler Griggs, Xiaoxuan Liu, Jiaxiang Yu, Doyoung Kim, Wei-Lin Chiang, Alvin Cheung, and Ion Stoica. 2024. Mélange: Cost efficient large language model serving by exploiting gpu heterogeneity.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu. 2023. Unified demonstration retriever for in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4644–4668, Toronto, Canada. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

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.

Keqin Peng, Liang Ding, Yancheng Yuan, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2024. Revisiting demonstration selection strategies in in-context learning.

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. 2022. 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.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3:333–389.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş, and Dániel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *ArXiv preprint*, abs/2212.03533.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Multilingual e5 text embeddings: A technical report.

Liang Wang, Nan Yang, and Furu Wei. 2024b. Learning to retrieve in-context examples for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1752–1767, St. Julian's, Malta. Association for Computational Linguistics.

Yufeng Wang, Lu wei, and Haibin Ling. 2025. Targ: Training-free adaptive retrieval gating for efficient rag.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. 2023. Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1423–1436, Toronto, Canada. Association for Computational Linguistics.

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Rankrag: Unifying context ranking with retrieval-augmented generation in llms. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

7 Appendix

7.1 Prompt Template

We utilized the following chat template for all Training-Free Gated Reranking experiments. Dynamic fields are denoted in brackets.

MT Prompt Template

[System Message]
You are an expert translator. Task: Translate [Source Language] to [Target Language].
Rule 1: Output ONLY the translated text.
Rule 2: Do not engage in conversation or explain the translation.
Rule 3: CRITICAL: You have been provided with a "Reference Translation" in the context. You MUST copy the terminology and structure from the Reference Translation exactly. Do not rephrase.
[User Message]
Reference Examples:
Input ([Source Language]): [Retrieved Source Text 1]
Output ([Target Language]): [Retrieved Target Text 1]
Input ([Source Language]): [Retrieved Source Text 2]
Output ([Target Language]): [Retrieved Target Text 2]
...
Current Task:
Input ([Source Language]): [Input Source Text]
Output ([Target Language]):

NLU Prompt Template

[System Message]
You are an AI assistant. Complete the given task by outputting ONLY the required label. Do not explain.
[User Message]
Reference Examples:
[Retrieved Source Text 1] [Retrieved Target Label 1]
[Retrieved Source Text 2] [Retrieved Target Label 2]
...
Current Task:
[Input Source Text]

7.2 Hyperparameter Selection

To select our architectural choices, we conducted a three-stage ablation study on the KDE dataset using the `Llama 3B` model.

Retrieval Mechanism and Shot Count. First, we determined the optimal retrieval backbone. As

shown in Table 4, **Dense-based retrieval** combined with reranking consistently outperforms lexical (BM25) and random selection. Crucially, the performance gap widens as the number of shots (k) increases, peaking at $k = 5$. This validates our decision to fix $k = 5$ for the main experiments, as it maximizes the context window's utility.

Strategy	k	COMET	BLEU	ChrF
Zero-Shot	0	69.8	17.17	44.14
Random	1	**71.27**	18.74	46.05
	3	71.26	**19.16**	**46.07**
	5	70.94	19.10	45.82
Dense	1	73.56	23.28	50.78
	3	**73.98**	24.35	51.80
	5	73.79	**24.74**	**52.19**
BM25	1	73.09	23.86	50.65
	3	74.03	**25.13**	52.15
	5	**74.29**	24.97	**52.33**
BM25 Full Reranking	1	73.85	24.40	51.38
	3	74.20	26.05	52.54
	5	**74.51**	**26.21**	**52.81**
Dense Full Reranking	1	73.76	24.15	51.45
	3	74.65	26.30	53.19
	5	**74.80**	**26.79**	**53.44**

Table 4: **Impact of Retrieval Strategy and Shot Count (k).** Ablation study on the KDE (En→De) dataset. While static retrieval methods (Random, BM25, E5) show diminishing returns or saturation at higher k, the reranking strategies continue to improve, with **E5 + TopP+ConE** at $k = 5$ establishing the upper bound for performance. Bold values indicate the optimal shot count for each specific strategy.

Comparison of Reranking Algorithms. To determine the optimal training-free method, we evaluated the effectiveness of Conditional Entropy Rerank function against alternative strategies, including Source Entropy Reranking, TARG Reranking which is based on Wang et al. (2025), RankRAG Reranking which follows Yu et al. (2024) without training. As shown in Table 5, while some reranking methods offer some gains, utilizing the Conditional Entropy Rerank yields the most robust performance.

Strategy	COMET	BLEU	ChrF
Source Entropy Reranking	69.80	17.17	44.14
RankRAG Reranking	71.51	18.92	46.71
TARG Reranking	74.51	26.34	52.88
Conditional Entropy Rerank	**74.80**	**26.79**	**53.44**

Table 5: **Re-Ranking Strategies.** Evaluation on KDE (En→De) using `Llama-3.2-3B-Instruct` (Llama 3B) ($k = 5$).

Sensitivity to Candidate Pool Size. Finally, we analyzed the computational trade-off regarding the cardinality of the candidate pool $|\mathcal{C}| = N$. As detailed in Table 6, increasing the search space from $5 \times k$ to $20 \times k$ yields steady improvements. However, performance plateaus beyond $N = 100$ ($20 \times k$). Consequently, we selected a pool size of 100 for all experiments to balance maximum utility with inference latency.

	Dense Full Reranking		
N	**COMET**	**BLEU**	**ChrF**
1	74.04	24.95	52.42
5	74.43	26.01	52.80
10	74.42	26.20	53.08
20	**74.88**	26.80	**53.47**
30	74.83	**27.08**	53.38

Table 6: **Candidate Pool Sensitivity.** Increasing pool size (N) yields gains up to $N = 100$ (MULT=20), after which performance saturates.

7.3 Detailed examples used in Analysis

Table 7 shows full details for the examples shown in the analysis section.

7.4 Performance Across Tasks and Models

Table 8 details the results across individual datasets, while Table 9 provides a model-by-model breakdown. Tables 11 and 10 shows Gated (test-set calibrated) and Gated (dev-set calibrated) results per model and per dataset. The data indicate that our gating strategy generally yields higher performance than full reranking while simultaneously maintaining consistent computational efficiency. Although empirical performance occasionally trails that of full reranking, the efficiency gains persist across all configurations.

Strategy	Retrieved In-Context Example (Shot)	Final Model Output
Case 1: Reranking Helped (Resolves Terminology Mismatch) **Input (PT):** As actualizações sobre o decurso do estudo e sobre o recrutamento serão apresentadas juntamente com os reavaliação anual/ RPS. **Reference (EN):** Progress and recruitment updates will be provided alongside the Annual Re-assessment/ **PSURs**.		
Baseline	*Src:* As actualizações do progresso e dos recrutamentos serão apresentadas juntamente com os PSURs normais. *Tgt:* Progress and recruitment updates will be provided alongside the standard PSURs.	Updates on the study course and recruitment will be presented together with the annual review/**RPS**.
Rerank	*Src:* O sumário da informação de segurança recolhida será apresentado juntamente com os reavaliação anual/ RPSs. *Tgt:* Summary safety information gathered will be provided alongside the Annual Re-assessment/ PSURs.	Updates regarding the course of the study and recruitment will be provided alongside the Annual Re-assessment/ **PSURs**.
Case 2: Baseline Was Better (Reranking Caused Hallucination) **Input (EN):** Like all medicines, Xenical can cause side effects, although not everybody gets them. **Reference (PT): Como todos os medicamentos, Xenical pode causar efeitos secundários, no entanto estes não se manifestam em todas as pessoas.**		
Baseline	*Src:* Like all medicines, Arixtra can cause side effects, although not everybody gets them. *Tgt:* Como todos os medicamentos, Arixtra pode causar efeitos secundários, no entanto, estes não se manifestam em todas as pessoas.	**Como todos os medicamentos, Xenical pode causar efeitos secundários, no entanto, estes não se manifestam em todas as pessoas.** *Eng: Like all medicines, Xenical can cause side effects although not everybody gets them.*
Rerank	*Src:* Take special care with XENICAL Weight loss may also affect the dose of medicines taken for other conditions... *Tgt:* Tome especial cuidado com XENICAL A perda de peso pode também obrigar a alterar a dose de medicamentos...	**Tome especial cuidado com XENICAL A perda de peso pode também obrigar a alterar a dose de medicamentos que esteja a tomar para outras doenças (p. ex. colesterol elevado ou diabetes).** *Eng: Take special care with XENICAL Weight loss may also affect the dose of medicines taken for other conditions (e.g. high cholesterol or diabetes).*
Case 3: Baseline Was Better (Reranking Caused Template Mismatch) **Input (EN):** Like all medicines, STOCRIN can cause side effects, although not everybody gets them. **Reference (PT):** Como todos os medicamentos, STOCRIN pode causar **efeitos secundários, no entanto estes não se manifestam em todas as pessoas.**		
Baseline	*Src:* Like all medicines, Cetrotide can cause side effects, although not everybody gets them. *Tgt:* Como os demais medicamentos, Cetrotide pode causar efeitos secundários, no entanto estes não se manifestam em todas as pessoas.	Como os demais medicamentos, STOCRIN pode causar **efeitos secundários, no entanto estes não se manifestam em todas as pessoas.** *Eng: Like the other medicines, STOCRIN can cause **secondary effects, however these do not manifest in all people.***
Rerank	*Src:* Always notify your doctor immediately if you have these symptoms or any side effects while taking STOCRIN. *Tgt:* Informe sempre o seu médico imediatamente se sentir qualquer um destes sintomas ou quaisquer efeitos indesejáveis...	Como todos os medicamentos, STOCRIN pode causar **efeitos colaterais, embora nem todo mundo os obtenha.** *Eng: Like all medicines, STOCRIN can cause **collateral effects, although not everyone obtains them.***
Case 4: Baseline Was Better (Reranking Lost Structural Template) **Input Task (NLU):** Some rooms have balconies. Can we know Some rooms have balconies off of them that overlook the ocean.? **Reference Label: Maybe.**		
Baseline	*Task:* There are sea views from some rooms and there is also a swimming pool. Can we know Half the rooms are oceanview.? *Label:* Maybe.	**Maybe.**
Rerank	*Task:* Rooms have private lanais, refrigerators, and coffeemakers. Can we know Rooms have refigerators.? *Label:* Yes.	No.
Case 5: Reranking Helped (Resolves Relational Mapping) **Input Task (NLU):** In what borough is there a neighborhood called Huguenot? Can we know A small group of Huguenots also settled on the south shore of Staten Island along the New York Harbor, for which the current neighborhood of Huguenot was named.? **Reference Label: Yes.**		
Baseline	*Task:* The borough of Staten Island is primarily located on what island? Can we know This situation of boroughs separated by water led to the development of an extensive infrastructure of bridges and tunnels.? *Label:* No.	No.
Rerank	*Task:* In what borough is the Douglaston neighborhood located? Can we know In contrast, New York City also has neighborhoods that are less densely populated and feature free-standing dwellings.? *Label:* No.	**Yes.**

Table 7: Comparison of Baseline vs. Reranking In-Context Retrieval on Translation and NLU Quality

Dataset	Baseline	Full Rerank	Gated (Test-Set Calibrated)	Savings	Gated (Dev-Set Calibrated)	Savings
Machine Translation (BLEU)						
Medical_EMEA	38.77	39.89	**40.46**†	26%	**40.12**†	29%
Medical_JRC-Acquis	39.39	40.83	**41.07**†	12%	**40.87**†	16%
Medical_KDE	33.11	34.26	**34.53**†	15%	**34.27**†	18%
Natural Language Understanding (Accuracy)						
AgNews	91.66	92.63	**92.67**†	47%	92.34†	58%
CR	93.76	93.62	**94.61**†	88%	**93.72**	80%
MNLI	65.90	67.50	**67.75**†	15%	67.13†	45%
QNLI	72.10	73.58	**74.04**†	34%	**73.63**	47%
SST-2	93.37	94.02	**94.34**†	67%	93.89	59%
SST-5	51.81	51.34	**53.16**†	62%	**52.25**	60%
Subj	90.56	92.46	**93.27**†	17%	**92.75**†	30%

Table 8: Performance breakdown by dataset using Uncertainty. Gated (test-set calibrated) represents the theoretical upper bound, while Gated (dev-set calibrated) represents practical levels. Gated values exceeding Full Reranking are in bold. † denotes statistical significance against the baseline ($p < 0.05$).

Model	Baseline	Full Rerank	Gated (Test-Set Calibrated)	Savings	Gated (Dev-Set Calibrated)	Savings
Machine Translation (BLEU)						
Llama-3.2-3B-Instruct	34.22	35.69	**36.00**†	19%	**35.73**†	20%
Llama-3.3-70B-Instruct	43.16	44.16	**44.38**†	18%	44.16†	24%
Meta-Llama-3.1-8B-Instruct	37.28	38.57	**39.03**†	22%	**38.73**†	26%
Mistral-7B-Instruct-v0.3	35.00	36.23	**36.53**†	11%	**36.31**†	15%
Qwen2.5-32B-Instruct	41.38	42.53	**42.74**†	16%	42.52†	21%
Qwen2.5-3B-Instruct	32.53	33.86	**34.38**†	19%	**34.16**†	21%
aya-expanse-8b	37.64	38.59	**38.84**†	14%	**38.60**†	19%
c4ai-command-r7b-12-2024	35.52	36.96	**37.58**†	21%	**37.16**†	21%
Natural Language Understanding (Accuracy)						
Llama-3.2-3B-Instruct	76.32	77.99	**78.47**†	36%	77.70†	46%
Llama-3.3-70B-Instruct	84.12	85.05	**85.55**†	36%	84.93	48%
Meta-Llama-3.1-8B-Instruct	79.06	79.74	**80.21**†	47%	79.60	55%
Mistral-7B-Instruct-v0.3	78.31	78.55	**79.76**†	62%	**79.30**	68%
Qwen2.5-32B-Instruct	85.91	86.67	**87.13**†	49%	86.59†	54%
Qwen2.5-3B-Instruct	78.01	78.91	**80.14**†	57%	**79.49**	55%
aya-expanse-8b	80.06	80.74	**81.24**†	47%	**80.78**	49%
c4ai-command-r7b-12-2024	77.24	78.22	**78.76**†	43%	78.14	60%

Table 9: Performance breakdown by model using Uncertainty. Gated (Test-Set Calibrated) represents the theoretical upper bound, while Gated (Dev-Set Calibrated) represents practical levels. Gated values exceeding Full Reranking are in bold. † denotes statistical significance against the baseline ($p < 0.05$).

Dataset	Qwen 3B	Llama 3B	Command	Mistral	Aya	Llama 8B	Qwen 32B	Llama 70B
EMEA (de→en)	+0.44 BLEU (13.4%)	+0.18 BLEU (19.9%)	+0.76 BLEU (17.0%)	+0.21 BLEU (17.3%)	-0.06 BLEU (37.2%)	+0.25 BLEU (32.4%)	+0.12 BLEU (38.2%)	-0.27 BLEU (27.7%)
EMEA (en→de)	+0.41 BLEU (26.5%)	+0.41 BLEU (19.1%)	-3.47 BLEU (18.2%)	+0.26 BLEU (45.4%)	+0.53 BLEU (31.1%)	+0.70 BLEU (27.8%)	-0.11 BLEU (34.5%)	-0.34 BLEU (35.9%)
EMEA (en→es)	+0.38 BLEU (40.4%)	+0.12 BLEU (37.1%)	+0.23 BLEU (17.7%)	+0.08 BLEU (15.5%)	+0.42 BLEU (45.7%)	+0.70 BLEU (55.8%)	-0.12 BLEU (35.9%)	-0.03 BLEU (26.6%)
EMEA (es→en)	+0.50 BLEU (29.5%)	+0.36 BLEU (29.9%)	+1.17 BLEU (38.5%)	-0.20 BLEU (18.7%)	+0.36 BLEU (12.1%)	+0.28 BLEU (29.5%)	+0.03 BLEU (32.1%)	-0.27 BLEU (23.1%)
EMEA (en→pt)	+0.60 BLEU (24.2%)	-0.03 BLEU (20.1%)	+2.40 BLEU (30.9%)	+1.77 BLEU (72.1%)	+0.23 BLEU (52.7%)	+1.28 BLEU (41.9%)	+0.42 BLEU (24.8%)	+1.18 BLEU (36.1%)
EMEA (pt→en)	-0.05 BLEU (4.1%)	-0.20 BLEU (15.5%)	+0.15 BLEU (20.0%)	-0.06 BLEU (5.2%)	-0.15 BLEU (13.6%)	-0.13 BLEU (19.4%)	+0.11 BLEU (24.4%)	-0.17 BLEU (48.2%)
JRC-Acquis (de→en)	+0.50 BLEU (17.8%)	-0.12 BLEU (29.1%)	+0.41 BLEU (23.3%)	+0.11 BLEU (29.0%)	+0.01 BLEU (10.3%)	-0.12 BLEU (28.0%)	-0.15 BLEU (6.7%)	-0.06 BLEU (15.3%)
JRC-Acquis (en→de)	+0.34 BLEU (20.2%)	+0.05 BLEU (24.4%)	+0.57 BLEU (32.2%)	+0.05 BLEU (13.5%)	-0.02 BLEU (20.7%)	+0.32 BLEU (35.9%)	+0.09 BLEU (19.5%)	-0.10 BLEU (27.7%)
JRC-Acquis (en→es)	+0.82 BLEU (35.0%)	+0.37 BLEU (26.4%)	+0.09 BLEU (13.7%)	+0.41 BLEU (20.9%)	-0.05 BLEU (8.2%)	-0.10 BLEU (22.0%)	-0.03 BLEU (28.8%)	+0.10 BLEU (33.9%)
JRC-Acquis (es→en)	+0.35 BLEU (8.1%)	+0.03 BLEU (16.2%)	+0.41 BLEU (28.7%)	-0.09 BLEU (-2.7%)	-0.10 BLEU (20.0%)	+0.06 BLEU (28.5%)	+0.04 BLEU (15.8%)	+0.01 BLEU (18.8%)
JRC-Acquis (en→pt)	-0.04 BLEU (18.7%)	-0.18 BLEU (4.4%)	-0.07 BLEU (10.1%)	-0.24 BLEU (-4.9%)	-0.33 BLEU (8.5%)	-0.22 BLEU (11.6%)	-0.03 BLEU (7.6%)	+0.03 BLEU (13.3%)
JRC-Acquis (pt→en)	-0.02 BLEU (-5.3%)	-0.01 BLEU (-5.4%)	-0.08 BLEU (2.2%)	-0.22 BLEU (-1.8%)	-0.05 BLEU (-2.6%)	-0.41 BLEU (12.9%)	-0.04 BLEU (8.1%)	-0.20 BLEU (14.7%)
KDE (de→en)	+0.17 BLEU (18.5%)	+0.11 BLEU (41.7%)	+0.63 BLEU (29.7%)	-0.12 BLEU (10.3%)	-0.16 BLEU (13.6%)	+0.56 BLEU (30.9%)	+0.06 BLEU (14.5%)	+0.07 BLEU (17.3%)
KDE (en→de)	+0.15 BLEU (14.2%)	-0.09 BLEU (-0.9%)	-0.01 BLEU (16.2%)	-0.02 BLEU (6.2%)	-0.14 BLEU (5.9%)	-0.04 BLEU (6.6%)	-0.09 BLEU (11.9%)	-0.05 BLEU (13.6%)
KDE (en→es)	+0.28 BLEU (18.1%)	+0.03 BLEU (12.3%)	+0.01 BLEU (14.8%)	-0.23 BLEU (6.8%)	-0.24 BLEU (22.4%)	-0.01 BLEU (14.1%)	-0.16 BLEU (16.5%)	-0.04 BLEU (12.4%)
KDE (es→en)	+0.40 BLEU (23.0%)	-0.09 BLEU (19.2%)	+0.19 BLEU (21.6%)	-0.18 BLEU (7.7%)	+0.06 BLEU (14.5%)	-0.16 BLEU (28.5%)	-0.33 BLEU (36.4%)	+0.01 BLEU (18.1%)
KDE (en→pt)	+0.25 BLEU (25.3%)	+0.16 BLEU (25.3%)	+0.10 BLEU (3.1%)	-0.09 BLEU (0.3%)	-0.15 BLEU (1.1%)	-0.02 BLEU (15.9%)	-0.07 BLEU (11.9%)	+0.00 BLEU (29.2%)
KDE (pt→en)	+0.02 BLEU (38.9%)	-0.31 BLEU (33.6%)	+0.08 BLEU (34.0%)	-0.05 BLEU (4.0%)	+0.08 BLEU (18.1%)	-0.14 BLEU (32.8%)	-0.01 BLEU (19.0%)	-0.03 BLEU (23.8%)
NLU_AgNews	+0.01 Acc (72.9%)	-0.39 Acc (66.2%)	-0.14 Acc (48.7%)	-0.00 Acc (80.7%)	-0.11 Acc (43.7%)	-0.26 Acc (58.6%)	-0.36 Acc (55.7%)	-0.30 Acc (40.5%)
NLU_CR	+0.11 Acc (66.2%)	+1.20 Acc (71.9%)	+0.43 Acc (91.2%)	-0.32 Acc (78.0%)	+0.13 Acc (89.6%)	-0.61 Acc (79.0%)	+0.21 Acc (83.1%)	+0.05 Acc (84.0%)
NLU_MNLI	+0.29 Acc (35.0%)	+0.10 Acc (44.3%)	-0.61 Acc (54.4%)	-0.23 Acc (42.3%)	-0.50 Acc (46.0%)	-0.23 Acc (62.0%)	-0.19 Acc (32.3%)	-1.06 Acc (42.5%)
NLU_QNLI	+0.93 Acc (35.3%)	-0.80 Acc (18.8%)	-0.91 Acc (67.6%)	+1.35 Acc (74.3%)	-0.05 Acc (27.5%)	-0.28 Acc (30.4%)	+0.01 Acc (67.7%)	+0.84 Acc (51.0%)
NLU_SST-2	-0.36 Acc (34.6%)	-0.48 Acc (39.8%)	+0.09 Acc (92.5%)	+0.05 Acc (64.3%)	-0.30 Acc (59.2%)	-0.22 Acc (74.0%)	+0.34 Acc (64.7%)	-0.28 Acc (43.7%)
NLU_SST-5	+1.95 Acc (70.4%)	-0.71 Acc (52.2%)	+0.87 Acc (57.9%)	+0.65 Acc (59.0%)	+1.68 Acc (57.3%)	+1.51 Acc (70.4%)	+0.47 Acc (61.1%)	+0.87 Acc (55.0%)
NLU_Subj	+1.25 Acc (67.3%)	-0.74 Acc (27.4%)	-0.43 Acc (10.2%)	+4.02 Acc (76.6%)	-0.40 Acc (18.3%)	-0.21 Acc (11.5%)	-0.12 Acc (10.9%)	-0.28 Acc (19.7%)

Table 10: Performance gains (BLEU/Accuracy) and token savings (%) evaluated using the Gating (dev-set calibrated) configuration. Cell colors represent the percentage of tokens saved.

Dataset	Qwen 3B	Llama 3B	Command	Mistral	Aya	Llama 8B	Qwen 32B	Llama 70B
EMEA (de→en)	+0.65 BLEU (15.3%)	+0.54 BLEU (23.9%)	+0.96 BLEU (14.1%)	+0.48 BLEU (9.1%)	+0.26 BLEU (22.1%)	+0.51 BLEU (35.2%)	+0.49 BLEU (26.3%)	+0.03 BLEU (12.9%)
EMEA (en→de)	+0.69 BLEU (27.3%)	+0.64 BLEU (22.3%)	+0.22 BLEU (6.2%)	+0.48 BLEU (56.7%)	+0.71 BLEU (35.4%)	+0.94 BLEU (24.1%)	+0.16 BLEU (17.6%)	+0.07 BLEU (5.6%)
EMEA (en→es)	+0.61 BLEU (25.7%)	+0.40 BLEU (45.2%)	+0.55 BLEU (17.2%)	+0.33 BLEU (21.0%)	+0.64 BLEU (48.6%)	+1.01 BLEU (38.0%)	+0.19 BLEU (22.9%)	+0.22 BLEU (17.6%)
EMEA (es→en)	+0.69 BLEU (33.0%)	+0.69 BLEU (31.8%)	+1.42 BLEU (38.0%)	+0.04 BLEU (4.1%)	+0.52 BLEU (4.2%)	+0.56 BLEU (33.0%)	+0.20 BLEU (26.8%)	-0.04 BLEU (20.8%)
EMEA (en→pt)	+0.94 BLEU (24.0%)	+0.25 BLEU (19.0%)	+2.53 BLEU (26.3%)	+1.85 BLEU (69.1%)	+0.64 BLEU (71.5%)	+1.74 BLEU (51.6%)	+0.66 BLEU (27.2%)	+1.44 BLEU (32.4%)
EMEA (pt→en)	+0.08 BLEU (5.8%)	+0.14 BLEU (11.3%)	+0.52 BLEU (18.3%)	+0.17 BLEU (-0.5%)	+0.16 BLEU (-1.1%)	+0.15 BLEU (17.5%)	+0.27 BLEU (22.6%)	+0.12 BLEU (51.3%)
JRC-Acquis (de→en)	+0.60 BLEU (11.8%)	+0.23 BLEU (23.2%)	+0.62 BLEU (19.2%)	+0.32 BLEU (36.1%)	+0.20 BLEU (3.1%)	+0.14 BLEU (19.5%)	-0.02 BLEU (0.1%)	+0.03 BLEU (10.8%)
JRC-Acquis (en→de)	+0.54 BLEU (21.2%)	+0.29 BLEU (22.5%)	+0.74 BLEU (34.8%)	+0.32 BLEU (4.7%)	+0.21 BLEU (19.6%)	+0.59 BLEU (30.9%)	+0.27 BLEU (21.9%)	+0.15 BLEU (11.4%)
JRC-Acquis (en→es)	+1.05 BLEU (33.9%)	+0.58 BLEU (18.0%)	+0.31 BLEU (8.9%)	+0.59 BLEU (18.2%)	+0.17 BLEU (2.3%)	+0.28 BLEU (13.3%)	+0.27 BLEU (20.8%)	+0.32 BLEU (21.7%)
JRC-Acquis (es→en)	+0.54 BLEU (8.0%)	+0.28 BLEU (16.6%)	+0.55 BLEU (27.2%)	+0.00 BLEU (-5.4%)	+0.07 BLEU (9.3%)	+0.27 BLEU (36.4%)	+0.19 BLEU (17.4%)	+0.10 BLEU (10.4%)
JRC-Acquis (en→pt)	+0.21 BLEU (17.1%)	+0.15 BLEU (-2.0%)	+0.15 BLEU (14.8%)	-0.15 BLEU (-8.1%)	-0.03 BLEU (-4.7%)	+0.05 BLEU (1.6%)	+0.15 BLEU (6.5%)	+0.18 BLEU (7.6%)
JRC-Acquis (pt→en)	-0.00 BLEU (-5.5%)	+0.00 BLEU (-5.7%)	+0.03 BLEU (2.5%)	+0.00 BLEU (-4.9%)	+0.00 BLEU (-5.7%)	+0.00 BLEU (-5.6%)	+0.10 BLEU (3.4%)	-0.03 BLEU (4.4%)
KDE (de→en)	+0.39 BLEU (26.6%)	+0.46 BLEU (54.7%)	+0.89 BLEU (25.3%)	+0.16 BLEU (-1.0%)	+0.07 BLEU (6.7%)	+0.89 BLEU (22.7%)	+0.17 BLEU (14.9%)	+0.24 BLEU (18.3%)
KDE (en→de)	+0.46 BLEU (8.9%)	+0.06 BLEU (1.7%)	+0.30 BLEU (25.6%)	+0.27 BLEU (-3.8%)	+0.16 BLEU (-0.3%)	+0.21 BLEU (4.0%)	+0.16 BLEU (8.0%)	+0.11 BLEU (26.6%)
KDE (en→es)	+0.57 BLEU (25.4%)	+0.23 BLEU (15.7%)	+0.24 BLEU (13.6%)	-0.01 BLEU (-1.7%)	+0.12 BLEU (15.9%)	+0.16 BLEU (17.4%)	+0.04 BLEU (12.2%)	+0.10 BLEU (4.9%)
KDE (es→en)	+0.58 BLEU (21.8%)	+0.15 BLEU (14.7%)	+0.42 BLEU (28.7%)	+0.04 BLEU (-0.3%)	+0.19 BLEU (14.4%)	+0.28 BLEU (26.6%)	+0.09 BLEU (13.9%)	+0.32 BLEU (14.2%)
KDE (en→pt)	+0.52 BLEU (19.8%)	+0.43 BLEU (24.3%)	+0.27 BLEU (8.9%)	+0.10 BLEU (-4.5%)	+0.06 BLEU (1.7%)	+0.23 BLEU (3.4%)	+0.23 BLEU (6.4%)	+0.35 BLEU (32.2%)
KDE (pt→en)	+0.40 BLEU (27.3%)	+0.05 BLEU (7.2%)	+0.47 BLEU (39.3%)	+0.28 BLEU (3.4%)	+0.43 BLEU (5.9%)	+0.21 BLEU (27.6%)	+0.19 BLEU (17.3%)	+0.18 BLEU (20.0%)
NLU_AgNews	+0.35 Acc (84.7%)	+0.00 Acc (57.7%)	+0.00 Acc (48.9%)	+0.00 Acc (77.0%)	+0.00 Acc (67.6%)	+0.00 Acc (45.0%)	+0.00 Acc (-3.4%)	+0.00 Acc (-3.5%)
NLU_CR	+0.90 Acc (85.9%)	+2.10 Acc (76.5%)	+1.25 Acc (93.2%)	+0.85 Acc (81.0%)	+0.80 Acc (94.9%)	+0.55 Acc (91.1%)	+0.85 Acc (93.6%)	+0.60 Acc (85.5%)
NLU_MNLI	+1.00 Acc (37.3%)	+1.00 Acc (37.8%)	+0.00 Acc (16.0%)	+0.00 Acc (-3.4%)	+0.00 Acc (-3.5%)	+0.00 Acc (45.3%)	+0.00 Acc (-3.5%)	+0.00 Acc (-3.7%)
NLU_QNLI	+1.10 Acc (9.9%)	+0.10 Acc (-0.3%)	+0.00 Acc (2.0%)	+1.40 Acc (87.2%)	+0.05 Acc (19.3%)	+0.00 Acc (8.4%)	+0.05 Acc (84.8%)	+1.00 Acc (62.2%)
NLU_SST-2	+0.10 Acc (53.1%)	+0.10 Acc (44.4%)	+0.60 Acc (91.5%)	+0.30 Acc (53.9%)	+0.10 Acc (77.4%)	+0.45 Acc (70.1%)	+0.80 Acc (94.8%)	+0.10 Acc (50.8%)
NLU_SST-5	+3.10 Acc (65.2%)	+0.05 Acc (40.1%)	+1.90 Acc (50.9%)	+1.40 Acc (59.6%)	+2.55 Acc (61.0%)	+2.25 Acc (72.5%)	+1.50 Acc (78.9%)	+1.80 Acc (65.8%)
NLU_Subj	+2.00 Acc (65.3%)	+0.00 Acc (-3.8%)	+0.00 Acc (-3.2%)	+4.55 Acc (80.5%)	+0.00 Acc (12.1%)	+0.00 Acc (-3.5%)	+0.00 Acc (-3.7%)	+0.00 Acc (-3.9%)

Table 11: Performance gains (BLEU/Accuracy) and token savings (%) evaluated using the Gating (test-set calibrated) configuration. Cell colors represent the percentage of tokens saved.