Translate Smart, not Hard: Cascaded Translation Systems with Quality-Aware Deferral

Anonymous ACL submission

Abstract

Larger models often outperform smaller ones but come with high computational costs. Cascading offers a potential solution. By default, it uses smaller models and defers only some instances to larger, more powerful models. However, designing effective deferral rules remains a challenge. In this paper, we propose a simple yet effective approach for machine translation, using existing quality estimation (QE) metrics as deferral rules. We show that QE-based deferral allows a cascaded system to match the performance of a larger model while invoking it for a small fraction (30% to 50%) of the examples, significantly reducing computational costs. We validate this approach through both automatic and human evaluation.

1 Introduction

Larger models consistently outperform smaller ones in NLP tasks, but the trade-off is the increased computational cost. This raises the question:

> *How can we maintain high performance while reducing computational load?*

A promising solution is **model cascading**, where smaller models handle examples by default, and only a subset of hard instances is deferred to a larger model. However, this approach requires a robust deferral system that reliably determines when to defer. Common approaches often involve designing and training specialized deferral models, which determine when a large model is needed—*e.g.*, based on reliability or uncertainty estimates (Chen et al., 2023; Gupta et al., 2024). But do we really need to train new models for every task, or can existing resources speed up this process?

For **machine translation** (MT), extensive research on reference-free automatic evaluation offers an appealing alternative (Zerva et al., 2022, 2024; Blain et al., 2023). In this paper, we leverage recent **quality estimation** (QE) metrics to create straightforward and relatively lightweight deferral rules. This approach draws inspiration from professional translation workflows, where QE metrics help identify translations that should be deferred to expert post-editing (Castilho and O'Brien, 2017; Béchara et al., 2021). Our main contributions are:

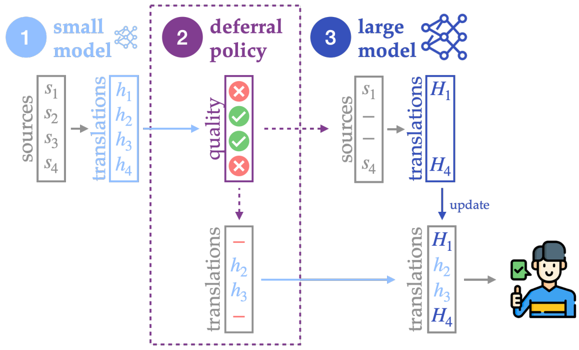

Figure 1: Cascaded translation system with QE-based deferral. A small model translates a batch of source sentences, and a relatively lightweight QE model scores the hypotheses. Sources with the lowest-scoring translations are deferred to a larger model. The extent of deferral is determined by a predefined compute budget.

- We introduce a **cascaded translation system** that uses pretrained QE metrics to determine whether to defer examples from a smaller model to a larger one, balancing efficiency and quality (§3). See Fig. 1 for an illustration.
- We confirm that the benefits of QE-based model cascading hold across different combinations of translation and QE models (§4).
- We perform **human evaluation**, further validating our approach on two language pairs (en-es and en-ja) in the WMT24 test set (§5).
- We release our code, all generated translations, and human quality assessments (when applicable) to encourage further research.[1]

[1]These resources will be made available upon acceptance.

2 Adaptive Inference in NLP

Adaptive inference techniques are increasingly being adopted in natural language processing tasks (Mamou et al., 2022; Varshney and Baral, 2022; Chen et al., 2023; Ong et al., 2024). These methods typically use models of different sizes and predictive power (often two, though most frameworks can easily accommodate more), with the primary goal of reducing the computational load by using the larger, more computationally expensive model only when necessary (*e.g.*, for more difficult examples or when a model is highly uncertain about its prediction). Current strategies include **routing**, where a decision rule determines which model to use, ensuring only one model is used to handle each input, and **cascading**, which starts with a smaller model and may invoke a larger one afterward based on the small model's output and a deferral rule. In this paper, we focus on the second approach.

The computational efficiency of model cascading comes at the cost of designing a **robust deferral system** that can reliably identify when to defer to the larger model. This is often handled using simple decision rules, such as nonparametric methods or other approaches based on uncertainty measures (Ramírez et al., 2024; Gupta et al., 2024). A recent alternative involves training external models specifically to predict when deferral is needed – for a given example, these models can be trained, *e.g.*, to assess if a given candidate is correct (Chen et al., 2023).[2] Here, we propose a simple and effective deferral rule for MT that is conceptually similar to this approach while offering a particularly straightforward solution for this task.

3 Quality-Aware Deferral for MT

Although human evaluations and reference-based metrics remain the standard for evaluating machine translations, reference-free/quality estimation (QE) metrics have shown strong correlations with human judgments (Zerva et al., 2024), holding promise in distinguishing between the quality of translations for the same source (Agrawal et al., 2024). Since QE models are typically much smaller than current translation models (Kocmi et al., 2024a), we propose to leverage them for an efficient deferral rule. Rather than training new bespoke decision models (§2), existing QE models can evaluate translations from a lightweight model and determine when to accept them or defer to a larger one.

How to choose which examples to defer? Setting a fixed threshold on QE scores is challenging—too high a threshold wastes computational resources, while too low a threshold risks compromising quality. Throughout this paper, we use a **budget-constrained computation** approach: we first translate all examples in a batch with the smaller model, then rank them based on QE scores, deferring only the lowest-scoring subset according to a predefined compute budget (the fraction of examples deferred to the larger model). This assumes parallel processing of entire batches rather than processing individual instances sequentially. We leave alternatives such as dynamic thresholding (Ramírez et al., 2024) for future work. See Fig. 1 for an illustration with 50% of deferral.

Computational efficiency. The standard approximation for the number of floating point operations (FLOPs) required for inference with a transformer model is $2ND$, where N represents the number of model parameters and D is the number of tokens generated at inference time (Sardana et al., 2024; Snell et al., 2024). For a cascaded approach with superscripts S and L denoting the smaller and larger models, respectively, this becomes:

$$2BD_S(N_S + N_{QE}) + 2\eta BD_L N_L, \qquad (1)$$

where B is the batch size and η is the proportion of instances the larger model processes. Assuming $D_S \approx D_L$, this approach achieves computational parity with the larger model (*i.e.*, $2BDN_L$) when:

$$\eta^\star = 1 - \frac{N_S + N_{QE}}{N_L}. \qquad (2)$$

This expression provides a simple rule of thumb: to maintain computational efficiency, the larger model should handle at most $\eta^\star$ of the examples. For instance, if it is $10\times$ larger than the smaller model and the QE model is negligible ($N_{QE} \ll N_S$), then $\eta^\star \approx 0.9$. This means the cascading is more efficient than always using the larger model as long as fewer than 90% of the examples are deferred.

4 Experiments and Analysis

We consider Tower-v2 models (Rei et al., 2024) of different size and predictive power: **Tower-v2 70B**, an improved iteration of Tower (Alves et al.,

[2] Likewise, routing typically involves training external models to *(i)* predict the performance of the small model (Šakota et al., 2024), or *(ii)* determine if the small model is likely to outperform the large one (Ding et al., 2024).

2024), obtained by continued pretraining Llama-3 (AI@Meta, 2024) on a multilingual dataset with 25 billions of tokens, followed by supervised fine-tuning for translation-related tasks;[3] and **Tower-v2 7B**, a more lightweight version using Mistral (Jiang et al., 2023). Check App. A for more details.

Deferral. We use two versions of COMETKIWI: `wmt22-cometkiwi-da` (Rei et al., 2022), which with only 0.5B parameters achieves a strong correlation with human judgments (Zerva et al., 2022); and `wmt23-cometkiwi-da-xxl` (Rei et al., 2023), a scaled version with 10.5B parameters. As baselines, we consider `random` selection; deferral rules based on source length computed using Tower-v2's tokenizer, *i.e.*, deferring either the shortest (`length`) or the longest (`-length`) sources;[4] and a confidence measure based on the smaller model's normalized log-probability (`logprobs`), *i.e.*, deferring texts with the lowest likelihoods. We also compare our approach with quality-aware decoding (Fernandes et al., 2022) in App. B.

Evaluation. We use the WMT24 test sets (Kocmi et al., 2024a), which span multiple domains (news, social, speech, and literary) and 11 language pairs (en-cs, en-de, en-es, en-hi, en-is, en-ja, en-ru, en-uk, en-zh, cs-uk, and ja-zh). For each language pair, we treat the full test set as a single batch for computing QE thresholds (§3).[5] We evaluate systems with METRICX (Juraska et al., 2023) to reduce the risk of "reward hacking" (Fernandes et al., 2022) and better reflect real quality improvements. Since biases may still exist when using a different evaluation metric than the reward model (Kovacs et al., 2024), we also conduct human evaluation (§5).

4.1 Larger is not necessarily better

Although Tower-v2 70B outperforms Tower-v2 7B across all language pairs (Table 1 shows aggregated results), a closer look at its win rates shows it only outperforms the smaller model in 43% of individual examples. This confirms that larger models do not consistently do better on every example, opening the possibility of using smaller models for a subset of examples without compromising overall performance, thus improving efficiency.

[3]Combined with quality-aware decoding (Fernandes et al., 2022), this is the winning submission of the WMT24 general translation shared task (Kocmi et al., 2024a).

[4]Source length is often used to assess translation difficulty (Kocmi and Bojar, 2017; Wan et al., 2022; Wang et al., 2023).

[5]Results are then averaged across language pairs for better visualization unless otherwise stated.

	M ↑	C ↑	Win rate
Tower-v2 7B	-3.01	83.94	43% / 32%
Tower-v2 70B	**-2.79**	**84.71**	NA

Table 1: Translation quality measured with METRICX (M) and COMET (C) on the WMT24 test set. Win rates against Tower-v2 70B, according to M. The bars represent the proportions of losses, ties, and wins. Following Kocmi et al. (2024b), translations with differences in M below 0.122 are considered ties (90% human accuracy).

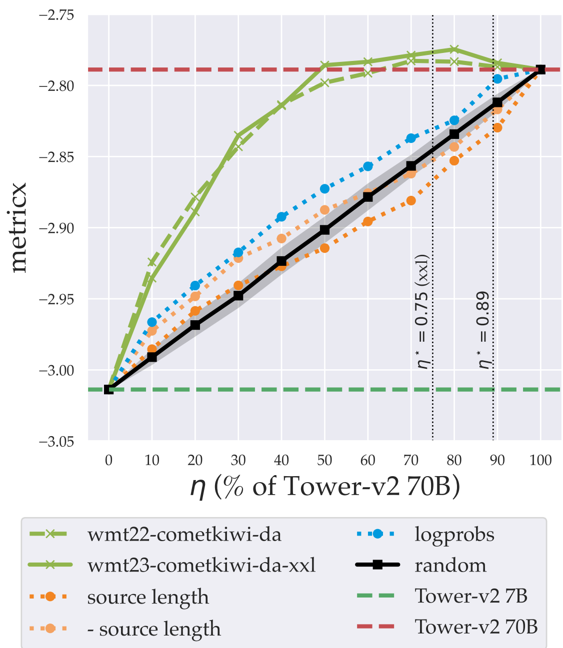

Figure 2: Translation quality of cascading combining Tower-v2 7B and Tower-v2 70B according to METRICX, as the inference computation budget varies. Horizontal lines show the performance of each model alone.

4.2 QE is an effective deferral rule

Fig. 2 shows the performance of a cascaded system combining Tower-v2 7B and Tower-v2 70B according to METRICX under varying inference budgets (results are averaged across language pairs). Each curve represents a different deferral rule. As expected, the random baseline fails to identify examples that benefit from larger models, resulting in suboptimal performance. Source length-based decision rules or using the small model's `logprobs` perform slightly better or worse than random, suggesting that simple heuristics are inefficient for deferral. In contrast, QE-based deferral (our proposal) achieves the best overall performance, enabling the cascaded system to match the performance of the large model while invoking it for only 50% to 60% of the examples. From Eq. (2),

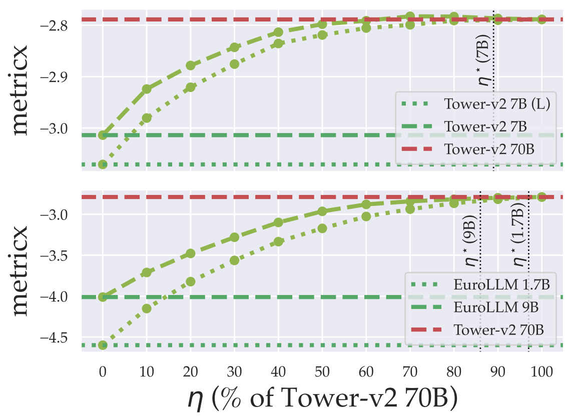

Figure 3: Translation quality of cascaded systems with deferral based on `wmt22-cometkiwi-da`. Large model: Tower-v2 70B. Small models: Tower-v2 7B (L), Tower-v2 7B (top); EuroLLM 1.7B, EuroLLM 9B (bottom).

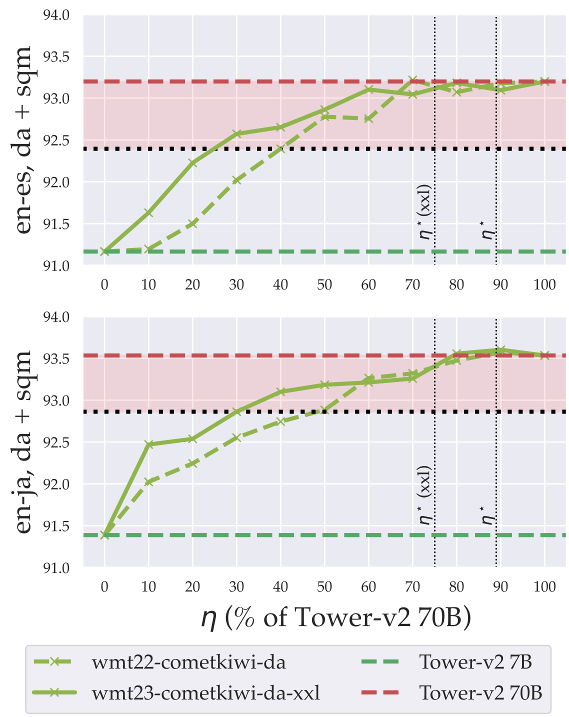

Figure 4: Translation quality of a cascaded system combining Tower-v2 7B and Tower-v2 70B according to human scores (in a scale from 0 to 100), as the inference computation budget varies. Systems in the shaded area are not significantly different from Tower-v2 70B according to the paired-permutation test with $p = 0.01$.

computational parity is reached at $\eta^\star = 89\%$ when using `wmt22-cometkiwi-da` ($N_Q = 0.5B$) and $\eta^\star = 75\%$ with `wmt23-cometkiwi-da-xxl` ($N_Q = 10.5B$). Matching Tower 70B's performance at such a small η shows that our approach effectively balances efficiency and quality.

4.3 Cascading works across different setups

We have shown that QE-based cascading works well across QE models of different sizes (Fig. 2). Here, we study whether it still provides gains when the smaller model is weaker. We train another version of Tower 7B using Llama-3 instead of Mistral, referred to as **Tower 7B (L)**, and use two versions of **EuroLLM** (Martins et al., 2024) with 1.7B ($\eta^\star = 0.97$) and 9B parameters ($\eta^\star = 0.86$). Fig. 3 shows that while these models underperform Tower-v2 7B, cascading with Tower 70B remains competitive. This indicates that QE-based cascading is robust across different generation models, even when both belong to the same family (top) or when the small model is much smaller (bottom).

5 Human Evaluation

Since using QE metrics during inference can bias automatic evaluations, we conduct a human study to validate our approach. We randomly sample 500 source instances and ask human annotators to rate translations from Tower-v2 7B and Tower-v2 70B on a continuous scale from 1 (no overlap in meaning) to 100 (perfect translation). This is done for en-es and en-ja. Further details are in App. C.

Fig. 4 shows the performance of cascaded systems using QE-based deferral. We use a paired-permutation test (Good, 2000; Zmigrod et al., 2022) to compare the performance of Tower-v2 70B with our systems under varying budgets. The shaded region shows that our approach achieves performance comparable to Tower-v2 70B while invoking it for only 30% to 50% of the examples,[6] confirming that it substantially reduces computational costs without compromising translation quality. App. C.1 provides further evidence using other QE models.

6 Conclusions and Future Work

We propose a simple yet effective approach to model cascading for MT using QE metrics for deferral. Our method matches the quality of larger models while requiring them to handle only a subset of examples, significantly reducing computational costs. This is shown through automatic and human evaluations. The effectiveness of our framework depends on the quality of existing QE models, and improving them can further strengthen our approach (App. C.2).

[6]Systems within the shaded region are also significantly better than Tower-v2 7B according to the same statistical test.

7 Limitations

We highlight three main limitations of our work. First, we focus on a two-stage cascade, where examples are handled by a small model or deferred to a larger one. Extending this to a multistage setup with more than two models could further improve efficiency but also add complexity. Second, our study is limited to machine translation. QE-based deferral works particularly well in MT due to the availability of high-quality human-labeled data for training QE models. Extending this approach to other tasks where such data is scarce is not straightforward. Finally, our method assumes the smaller model is reasonably competitive with the larger one, which is a fair assumption for MT, as shown in our experiments. If the gap in win rates is too large, cascading offers little benefit, as most examples would require deferral.

References

Sweta Agrawal, António Farinhas, Ricardo Rei, and Andre Martins. 2024. Can automatic metrics assess high-quality translations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14491–14502, Miami, Florida, USA. Association for Computational Linguistics.

AI@Meta. 2024. Llama 3 model card.

Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. In *First Conference on Language Modeling*.

Frederic Blain, Chrysoula Zerva, Ricardo Rei, Nuno M. Guerreiro, Diptesh Kanojia, José G. C. de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, Constantin Orasan, and André Martins. 2023. Findings of the WMT 2023 shared task on quality estimation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653, Singapore. Association for Computational Linguistics.

Hannah Béchara, Constantin Orăsan, Carla Parra Escartín, Marcos Zampieri, and William Lowe. 2021. The role of machine translation quality estimation in the post-editing workflow. *Informatics*, 8(3).

Sheila Castilho and Sharon O'Brien. 2017. Acceptability of machine-translated content: A multi-language evaluation by translators and end-users. *Linguistica Antverpiensia, New Series–Themes in Translation Studies*, 16.

Lingjiao Chen, Matei Zaharia, and James Zou. 2023. Frugalgpt: How to use large language models while reducing cost and improving performance. *Preprint*, arXiv:2305.05176.

Dujian Ding, Ankur Mallick, Chi Wang, Robert Sim, Subhabrata Mukherjee, Victor Rühle, Laks V. S. Lakshmanan, and Ahmed Hassan Awadallah. 2024. Hybrid LLM: Cost-efficient and quality-aware query routing. In *The Twelfth International Conference on Learning Representations*.

António Farinhas, José de Souza, and Andre Martins. 2023. An empirical study of translation hypothesis ensembling with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11956–11970, Singapore. Association for Computational Linguistics.

António Farinhas, Haau-Sing Li, and Andre Martins. 2024. Reranking laws for language generation: A communication-theoretic perspective. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Christian Federmann. 2018. Appraise evaluation framework for machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 86–88, Santa Fe, New Mexico. 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.

Markus Freitag, Behrooz Ghorbani, and Patrick Fernandes. 2023. Epsilon sampling rocks: Investigating sampling strategies for minimum Bayes risk decoding for machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9198–9209, Singapore. Association for Computational Linguistics.

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.

Phillip Good. 2000. *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer New York, NY.

Nuno M. Guerreiro, Elena Voita, and André Martins. 2023. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference* of the European Chapter of the Association for Computational Linguistics, pages 1059–1075, Dubrovnik, Croatia. Association for Computational Linguistics.

Neha Gupta, Harikrishna Narasimhan, Wittawat Jitkrittum, Ankit Singh Rawat, Aditya Krishna Menon, and Sanjiv Kumar. 2024. Language model cascades: Token-level uncertainty and beyond. In *The Twelfth International Conference on Learning Representations*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

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.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. 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, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024a. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

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

Tom Kocmi and Ondřej Bojar. 2017. Curriculum learning and minibatch bucketing in neural machine translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 379–386, Varna, Bulgaria. INCOMA Ltd.

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024b. Navigating the metrics maze: Reconciling score magnitudes and accuracies. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1999–2014, Bangkok, Thailand. Association for Computational Linguistics.

Geza Kovacs, Daniel Deutsch, and Markus Freitag. 2024. Mitigating metric bias in minimum Bayes risk decoding. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1063–1094, Miami, Florida, USA. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Jonathan Mamou, Oren Pereg, Moshe Wasserblat, and Roy Schwartz. 2022. Tangobert: Reducing inference cost by using cascaded architecture. *Preprint*, arXiv:2204.06271.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M Guerreiro, Ricardo Rei, Duarte M Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, et al. 2024. Eurollm: Multilingual language models for europe. *arXiv preprint arXiv:2409.16235*.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M Waleed Kadous, and Ion Stoica. 2024. Routellm: Learning to route llms with preference data. *Preprint*, arXiv:2406.18665.

Guillem Ramírez, Alexandra Birch, and Ivan Titov. 2024. Optimising calls to large language models with uncertainty-based two-tier selection. *Preprint*, arXiv:2405.02134.

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

Ricardo Rei, Jose Pombal, Nuno M. Guerreiro, João Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tania Vaz, Duarte Alves, Amin Farajian, Sweta Agrawal, Antonio Farinhas, José G. C. De Souza, and André Martins. 2024. Tower v2: Unbabel-IST 2024 submission for the general MT shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 185–204, Miami, Florida, USA. 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. 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.

Nikhil Sardana, Jacob Portes, Sasha Doubov, and Jonathan Frankle. 2024. Beyond chinchilla-optimal: Accounting for inference in language model scaling laws. In *Forty-first International Conference on Machine Learning*.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *Preprint*, arXiv:2408.03314.

Neeraj Varshney and Chitta Baral. 2022. Model cascading: Towards jointly improving efficiency and accuracy of NLP systems. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11007–11021, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yu Wan, Baosong Yang, Derek Fai Wong, Lidia Sam Chao, Liang Yao, Haibo Zhang, and Boxing Chen. 2022. Challenges of neural machine translation for short texts. *Computational Linguistics*, 48(2):321–342.

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

Chrysoula Zerva, Frederic Blain, José G. C. De Souza, Diptesh Kanojia, Sourabh Deoghare, Nuno M. Guerreiro, Giuseppe Attanasio, Ricardo Rei, Constantin Orasan, Matteo Negri, Marco Turchi, Rajen Chatterjee, Pushpak Bhattacharyya, Markus Freitag, and André Martins. 2024. Findings of the quality estimation shared task at WMT 2024: Are LLMs closing the gap in QE? In *Proceedings of the Ninth Conference on Machine Translation*, pages 82–109, Miami, Florida, USA. Association for Computational Linguistics.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. 2022. Findings of the WMT 2022 shared task on quality estimation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ran Zmigrod, Tim Vieira, and Ryan Cotterell. 2022. Exact paired-permutation testing for structured test statistics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4894–4902, Seattle, United States. Association for Computational Linguistics.

Marija Šakota, Maxime Peyrard, and Robert West. 2024. Fly-swat or cannon? cost-effective language model choice via meta-modeling. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, page 606–615, New York, NY, USA. Association for Computing Machinery.

A Experimental Details

Through the paper, we experiment with the following generation models:

- **Tower-v2 70B** (Rei et al., 2024): An improved iteration of Tower (Alves et al., 2024), obtained by continued pertaining Llama-3 (AI@Meta, 2024) on a multilingual dataset with billions of tokens, followed by supervised finentuning for translation-related tasks. It has 70B parameters. Compared to the first iteration of Tower, this model is better at paragraph and document-level translation and supports more language (15, instead of 10), including all the languages in the WMT24 test sets. Combined with quality-aware decoding (Fernandes et al., 2022), this is the winning submission of the WMT24 general translation shared task (Kocmi et al., 2024a).
- **Tower-v2 7B** (Rei et al., 2024): A smaller version of Tower-v2 70B based on Mistral (Jiang et al., 2023).
- **Tower-v2 7B (Llama-3)**: We follow the recipe described above to train a smaller version of Tower-v2 70B based on LLama-3. This model slightly underperforms its Mistral counterpart.
- **EuroLLM Instruct (9B and 1.7B)** (Martins et al., 2024): EuroLLM models are open-weight multilingual models trained on 4 trillion tokens covering all European Union and many other relevant languages across several data sources: web data, parallel data (en-xx and xx-en), and high-quality datasets. The instruction-tuned models are obtained after finetuning the base models on the EuroBlocks dataset, which includes general instruction-following and machine translation tasks.

We generate all translations with greedy decoding using vLLM (Kwon et al., 2023) for faster inference. Table 2 shows the performance of these models on the WMT24 test sets (Kocmi et al., 2024a),[7] according to METRICX and COMET (results are averaged across all language pairs), along with their win rates against Tower-v2 70B.[8] Our use of datasets and models aligns with their intended purposes as defined by the licenses.

[7] Publicly available for research purposes at `https://www2.statmt.org/wmt24/translation-task.html`.

[8] Following Kocmi et al. (2024b), translations with differences in METRICX below 0.122 are considered ties when comparing two systems (90% human accuracy). We use the same threshold for detecting ties at the segment level.

	M ↑	C ↑	Win rate	
Tower-v2 7B	-3.01	83.94	43%	32%
Tower-v2 7B (L)	-3.07	83.73	45%	32%
EuroLLM 9B	-4.01	80.56	52%	28%
EuroLLM 1.7B	-4.60	77.42	66%	20%
Tower-v2 70B	**-2.79**	**84.71**	NA	

Table 2: Translation quality measured with METRICX (M) and COMET (C) on the WMT24 test set. Win rates against Tower-v2 70B, according to M. The bars represent the proportions of losses, ties, and wins.

B Quality-Aware Decoding

There is a large body of work on **reranking for language generation**, where we start by generating multiple hypotheses with a language model, and then use a reranker to select the best one (Farinhas et al., 2024). For machine translation, an example is quality-aware decoding (Fernandes et al., 2022; Freitag et al., 2022). The simplest/cheapest approach is QE reranking, where we first generate multiple translation hypotheses and then rerank them using a quality estimation model. This strategy is often used to reduce the propensity of language models to hallucinate or generate critical errors (Guerreiro et al., 2023; Farinhas et al., 2023). While our approach is conceptually different—designed with efficiency in mind, whereas QE reranking is often computationally expensive—it is nonetheless valuable to compare its performance against QE reranking based on hypotheses generated by the small model.

Computational efficiency. Following the discussion in §3, the number of FLOPS required for inference with a large model on a batch of B examples is given by:

$$2BDN_L, \tag{3}$$

where N_L represents the number of model parameters and D is the number of generated tokens. In this section, we assume that our goal is to **reduce the computational cost by** $(1-X)\%$, meaning that we operate under a computational budget of:

$$X \cdot 2BDN_L. \tag{4}$$

The number of FLOPs required to run inference with our cascaded approach is given by:

$$2BD(N_S + N_{QE} + \eta N_L), \tag{5}$$

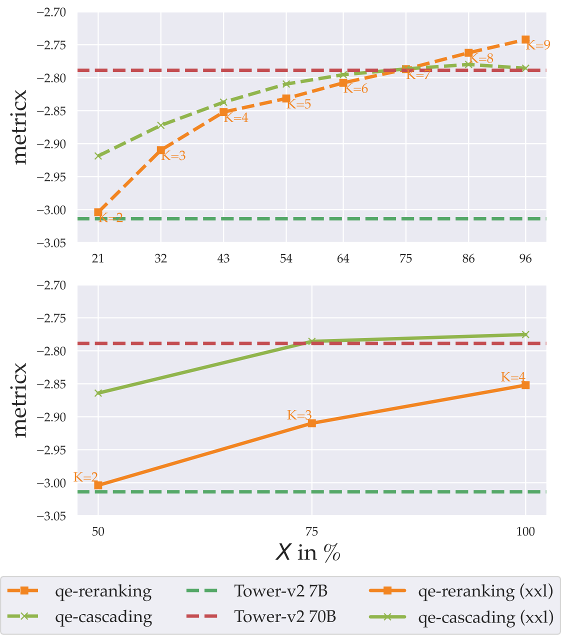

Figure 5: Translation quality of a cascaded system combining Tower-v2 7B and Tower-v2 70B (in green) *v.s.* QE reranking with hypotheses generated by Tower-v2 7B (in orange), measured with METRICX, as X varies. Horizontal lines show the performance of the smaller and larger models alone.

which leads to the following expression for X:

$$X = \eta + \frac{N_S + N_{QE}}{N_L}. \quad (6)$$

For QE reranking, the computational cost is:

$$2BDK(N_S + N_{QE}), \quad (7)$$

where K is the number of generated hypotheses. This yields:

$$X = K \cdot \left(\frac{N_S + N_{QE}}{N_L}\right). \quad (8)$$

These expressions allow us to obtain the values of η for which our cascaded approach incurs the same computational cost as QE reranking with K hypotheses:

$$\eta = (K-1) \cdot \left(\frac{N_S + N_{QE}}{N_L}\right). \quad (9)$$

Experiments and discussion. We generate up to 9 hypotheses with Tower-v2 7B using ϵ-sampling with $\epsilon = 0.02$ (Freitag et al., 2023).[9] Fig. 5 illustrates the trade-off between computational efficiency and translation quality (measured with METRICX) for a cascaded approach with QE-based deferral *versus* QE reranking. As expected, quality improves as the computational budget increases for both methods. While QE reranking is an effective way to improve translation quality when generating multiple hypotheses is feasible, our cascaded approach achieves better quality at lower computation costs, making it a more efficient alternative when computational efficiency is a priority.

[9]For our setup, according to Eq. (8), the number of FLOPs required for QE reranking with more than 9 hypotheses already exceeds the budget of $2BDN_L$ if we use `wmt22-cometkiwi-da`. When using `wmt23-cometkiwi-da-xxl`, computational parity is achieved with $K = 4$.

C Human Evaluation

In order to perform human evaluation, we recruited professional translators who were native speakers of the target language on the freelancing site Upwork.[10] We followed a DA+SQM (direct assessment + scalar quality metric) source contrastive evaluation (Kocmi et al., 2022) using Appraise (Federmann, 2018). We sampled 500 source instances from the WMT24 test set for en-ja and en-es and asked one translator per language pair to read two alternative translations for each source and evaluate them on a continuous scale from 0 to 100. The scale featured seven labeled tick marks (from 0 to 6) indicating different quality labels combining *accuracy* and *grammatical correctness*. Translators could further adjust their scores to reflect preferences or assign the same score to translations of similar quality. They were paid a market rate of around 20 USD per hour, and completing the task took approximately 12 to 14 hours for each language pair.

C.1 Deferral based on other QE metrics

We have seen that QE-based cascading works well with COMETKIWI models of different sizes (Fig. 4). Here, we show that this is also the case when using two reference-free versions of METRICX (Juraska et al., 2024): `metricx-24-hybrid-large-v2p6` and `metricx-24-hybrid-xl-v2p6` (Fig. 6, orange curves).

C.2 Oracle selection

The effectiveness of our framework depends on the quality of existing QE models, and improving them can further strengthen our approach. To access the performance ceiling of cascading, we report results with oracle deferral, *i.e.*, a deferral strategy that maximizes translation quality according to humans

[10]https://upwork.com

(Fig. 6, black curves).[11] The high oracle values indicate significant potential for improvement, suggesting that having better QE models could directly boost the effectiveness of our cascaded approach.

C.3 Annotation guidelines

We share below the annotation guidelines shared with the freelancers.

Task overview. This task involves evaluating two alternative translations of a source text and assigning a rating to each translation based on its overall quality and adherence to the source content. You should consider accuracy, fluency, and overall quality when assessing the different translations.

Annotation scale. Each translation should be evaluated on a continuous scale from 0 to 6 with the quality levels described below:

- **6 (perfect meaning and grammar):** The meaning of the translation is completely consistent with the source and the surrounding context, if applicable. The grammar is also correct.

- **4 (most meaning preserved and few grammar mistakes):** The translation retains most of the meaning of the source. It may have some grammar mistakes or minor contextual inconsistencies.

- **2 (some meaning preserved):** The translation preserves some of the meaning of the source but misses significant parts. The narrative is hard to follow due to fundamental errors. Grammar may be poor.

- **0 (nonsense/no meaning preserved):** Nearly all information is lost between the translation and source. Grammar is irrelevant.

Annotation interface. Figs. 7 and 8 show the annotation interface. If two candidates were the same or of the same quality, the annotators were asked to use "**match sliders**" to give them the exact same score. And, they could also use the absolute scale range to show preference between the translations.

[11] Oracle performance goes down after reaching a *plateau* due to our budget-constrained approach, which enforces deferral for a fixed percentage of examples.

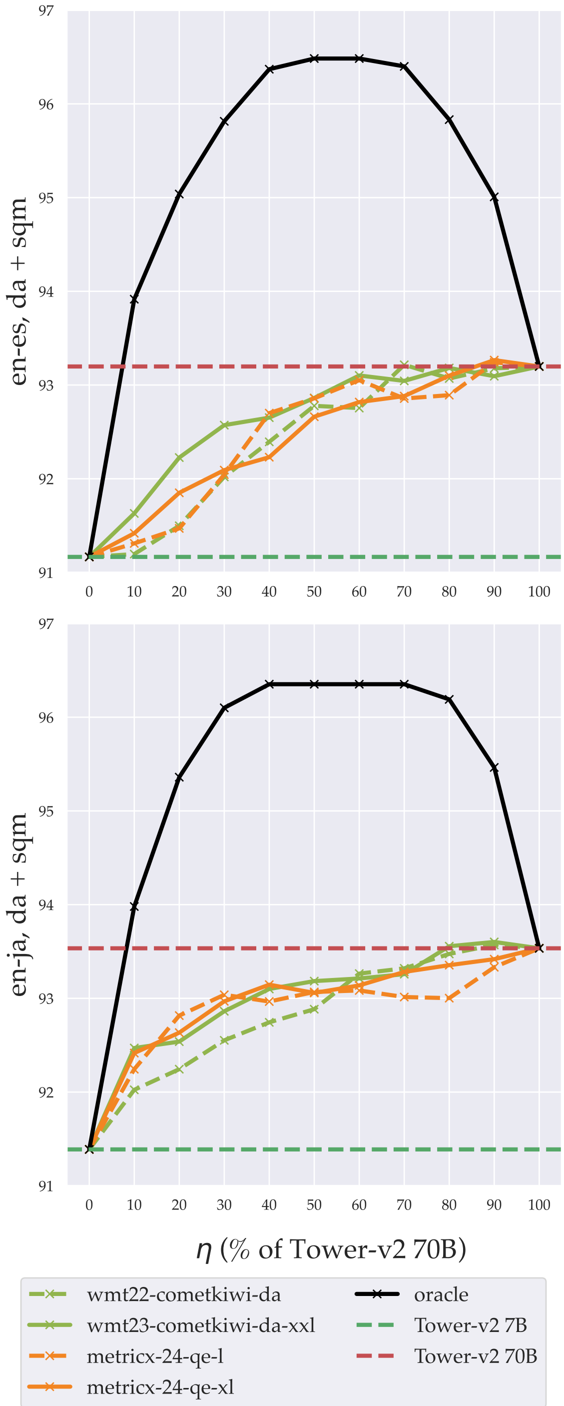

Figure 6: Translation quality of a cascaded system combining Tower-v2 7B and Tower-v2 70B according to human scores (in a scale from 0 to 100), as the inference computation budget varies. Deferral is based on different QE models (green and orange curves). The black curve shows the oracle selection.

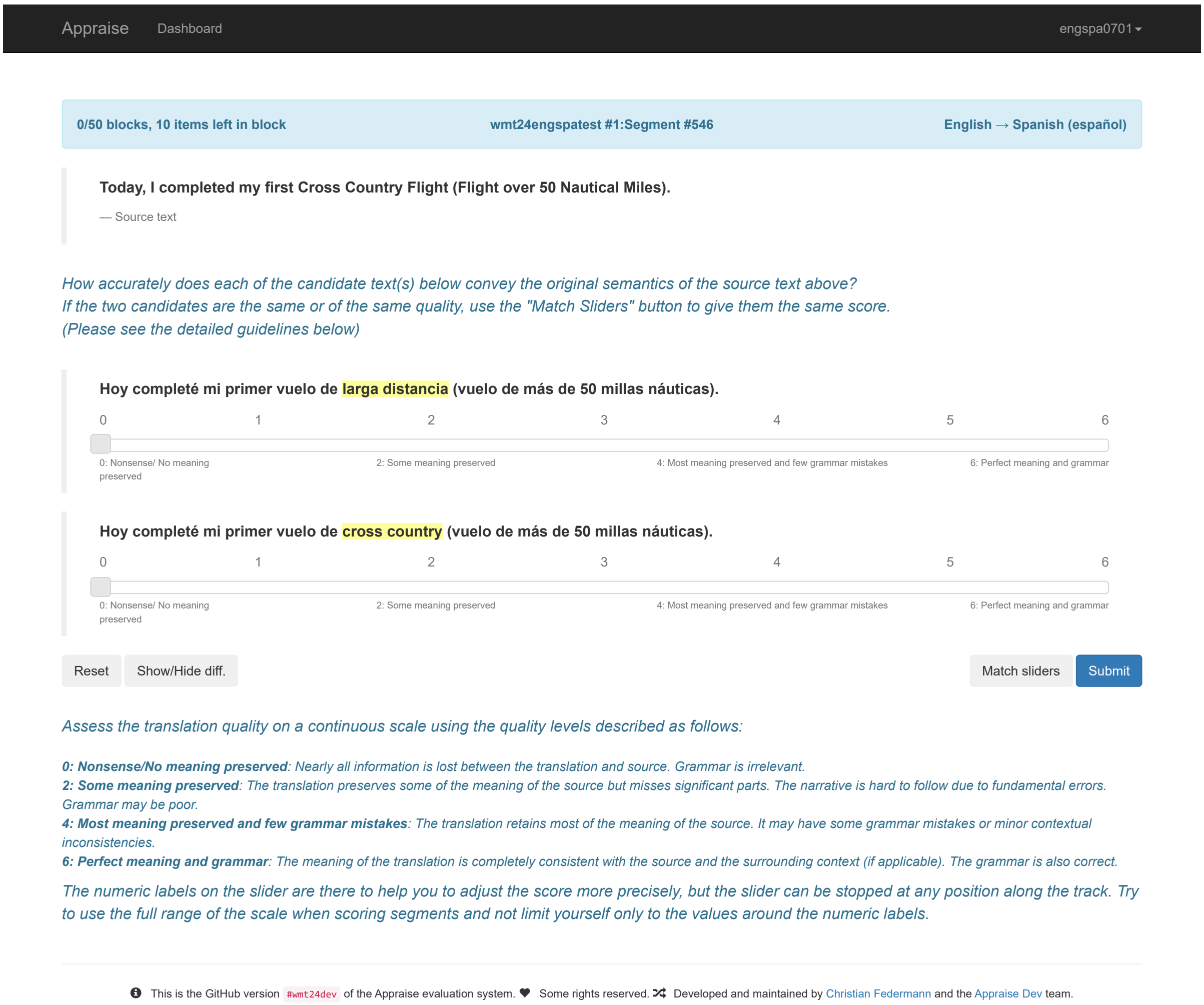

Figure 7: Annotation interface for en-es.

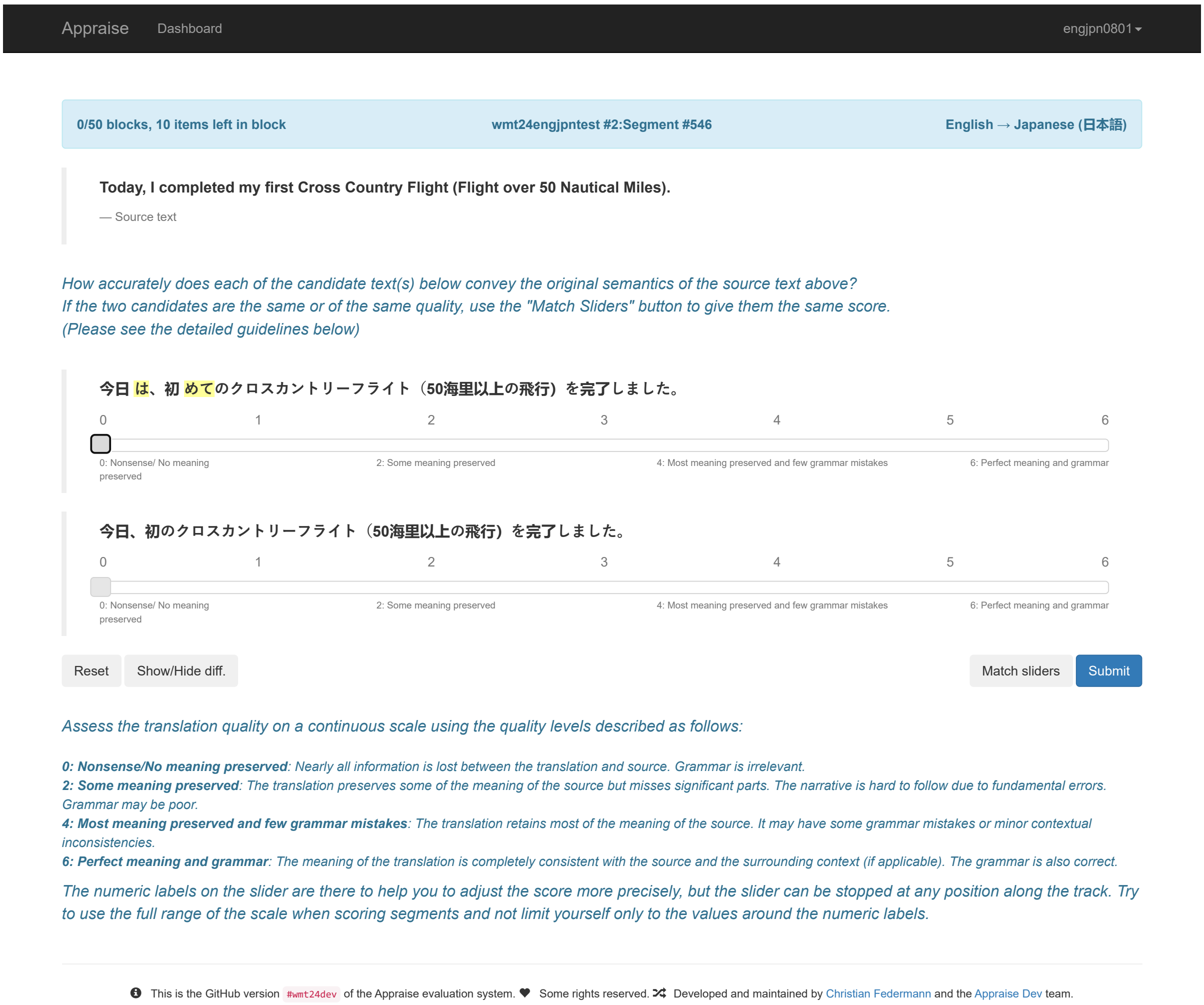

Figure 8: Annotation interface for en-ja.