Balancing Translation Quality and Environmental Impact: Comparing Large and Small Language Models

Antonio Castaldo[1,2,†], Petra Giommarelli[1,2,†] and Johanna Monti[2]

[1]University of Pisa, Largo Bruno Pontecorvo, 3, 56127 Pisa, Italy
[2]University of Naples L'Orientale, Via Chiatamone, 61/62, 80121 Naples, Italy

Abstract

Large Language Models (LLMs) have demonstrated remarkable performance in machine translation (MT), specifically concerning high-resource European languages. However, their extensive computational requirements raise sustainability concerns. This paper investigates the potential of smaller, fine-tuned language models as a more sustainable alternative for MT tasks. We conduct a comparative analysis of model performance in terms of translation quality and CO_2eq emissions, and examine the key errors associated with using smaller models. Furthermore, we propose a novel metric that balances translation quality against environmental impact, aiming to inform more sustainable model selection in MT research and practice.

Keywords

machine translation, large language models, sustainability

1. Introduction

MT has been a core topic in natural language processing (NLP) for several decades, evolving from rule-based systems to statistical methods, and more recently to neural machine translation (NMT) and transformer-based models. The emergence of LLMs has significantly advanced the state-of-the-art in MT, demonstrating remarkable performance on various NLP tasks [1].

Their ability to generate fluent, context-aware translations in different domains has positioned LLMs at the forefront of MT research [2]. Their ability to model context, semantics, and discourse phenomena makes them highly attractive for both academic and industrial translation applications.

However, this performance comes at a significant environmental cost. Training and deploying LLMs consumes enormous computational resources, leading to considerable carbon emissions and infrastructure demands [3, 4]. These challenges have prompted the exploration of more sustainable alternatives.

This paper investigates whether smaller language models can serve as efficient and environmentally sustainable valid alternatives to LLMs in MT. Specifically, we will fine-tune the Gemma-3-4B[5] model on a parallel English-Italian (EN-IT) parallel corpus, and evaluate its performance, with human and automatic evaluation, against larger models. This setup allows us to assess the real-world viability of small models for machine translation when fine-tuned for specific language pairs and domains.

We conduct a comprehensive analysis of model performance, in terms of translation quality and CO_2eq emissions, validating our results with a human evaluation of the key errors associated with each model. Finally, we introduce a metric called Carbon-Adjusted Quality Score (CAQS), designed to facilitate sustainable model selection, that quantifies the trade-off between translation quality and sustainability.

2. Background

2.1. LLMs and Translation

LLMs have achieved state-of-the-art results in MT, by leveraging extensive pretraining on multilingual corpora, enabling them to deliver remarkable performance across a wide range of domains and language pairs [6]. In contrast to NMT systems, which rely primarily on parallel corpora, LLMs are pretrained on massive web-scale monolingual and multilingual datasets. This enables them to generate high-quality translations even in domains where parallel data is limited [7].

Notably, GPT-based models excel at producing contextually accurate translations, effectively capturing discourse relations and maintaining sentence-level coherence. They consistently outperform encoder-decoder architectures such as Transformer-big and M2M100, particularly in zero-shot and few-shot settings [8].

Moreover, LLMs support document-level translation by leveraging discourse-aware context windows, which enable the maintenance of lexical cohesion and consistent

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy

† The authors contributed equally.

antonio.castaldo@phd.unipi.it (A. Castaldo); petra.giommarelli@phd.unipi.it (P. Giommarelli); jmonti@unior.it (J. Monti)

0009-0008-3325-787X (A. Castaldo)

resolution of anaphoric references across sentences [9]. This capability results in more fluent translations, making LLMs increasingly favored in professional translation settings.

The adoption of LLMs, however, requires substantial computational resources and infrastructure, which may not be feasible for all organizations or languages. Beyond these practical limitations, the widespread adoption of LLMs also raises significant concerns about their environmental sustainability.

2.2. LLMs Sustainability

While Large Language Models (LLMs) have enabled remarkable progress in NLP, their growing environmental footprint raises important sustainability concerns. Training large-scale models such as GPT-3, with hundreds of billions of parameters, can consume up to 1.3 GWh of electricity, comparable to the yearly energy usage of more than 100 US homes [10]. This results in hundreds of tons of CO_2 emissions, depending on the carbon intensity of the power grid.

In addition to training, the inference phase of LLMs also significantly contributes to their overall carbon footprint, particularly in large-scale deployments. While the energy cost of a single inference is lower than that of training, the cumulative emissions can become substantial depending on usage patterns. For example, serving a single ChatGPT prompt may emit over 4g of CO_2eq, more than 20 times the emissions of a typical web search [11].

The same study emphasizes that total environmental impact depends on a combination of factors: model size, batch size, and hardware type. The latter reflects the impact of producing high-performance GPUs, which involves substantial embodied carbon emissions. Although these emissions occur at production time, they contribute to the model's overall environmental cost throughout its operational lifetime.

2.3. Small Language Models

Recent research has emphasized the growing feasibility and importance of SLMs as efficient alternatives to LLMs in constrained environments [12, 13]. SLMs, typically ranging from hundreds of millions to a few billion parameters, are substantially more resource-efficient and accessible, especially when tailored to specific tasks.

SLMs benefit from architectural simplifications, such as compact tokenizers and reduced model width and depth, which are optimized to preserve key capabilities while minimizing parameter overhead [14]. Small models, like Gemma [15] and PanGu-π-1.5B Pro model with only a few billion parameters have recently outperformed much larger models on several benchmarks due to optimizations in model architecture and training strategy [14].

Moreover, recent studies show that even highly complex capabilities like multi-step reasoning, previously thought to emerge only in models over 100B parameters, can be acquired by SLMs through targeted fine-tuning and distillation. Distilling chain-of-thought reasoning abilities, for instance, from GPT-3.5 into FlanT5 variants (250M to 3B) resulted in significant performance improvements on math reasoning tasks without the need for full retraining of the model's weights [16].

A comprehensive survey of SLMs underscores the value of model compression techniques such as pruning, quantization, and knowledge distillation. These enable the deployment of efficient models on mobile and edge devices while maintaining competitive accuracy for many tasks [17]. The adoption of SLMs is particularly promising for democratizing NLP, enabling smaller institutions and low-resource languages to benefit from modern AI without the environmental or infrastructural burden of LLMs.

In this study, we evaluate whether SLMs, when combined with modern fine-tuning strategies and lightweight architectures, could offer a pragmatic and sustainable path forward for machine translation and other NLP applications.

3. Fine-tuning a SLM

To demonstrate the effectiveness of using SLMs as sustainable alternatives to larger, more resource-intensive models in machine translation, we compare two state-of-the-art models: GPT-4o-mini [18] and an open-source model, Gemma-3-4B [15], which is significantly smaller than its OpenAI counterpart.

We fine-tune Gemma-3-4B on a carefully curated subset of the OpenSubtitles corpus, obtained from the Opus Corpus [19]. We evaluate both models on a held-out test set of 400 segments for the English–Italian (EN-IT) language pair and present our findings.

3.1. Dataset Curation

For our experiments, we focused on the EN–IT subset of the OpenSubtitles corpus, made available through the Opus Corpus repository. While OpenSubtitles is a rich resource for dialogue-based translation data, it also contains a considerable amount of noise due to its automatic extraction and alignment process. Therefore, careful curation was necessary to ensure the quality and relevance of the dataset.

We began by removing duplicate entries and any empty lines. Following this, we applied the `langdetect` [20] tool to verify the language of each sentence. This

step was essential, as web-crawled corpora, although intended to be language-specific, occasionally contain segments in other languages. Sentences detected to be in languages outside our target pair, and that could not be classified with a high confidence score, were filtered out.

Finally, we applied COMET-QE [21], a quality estimation model, to score the remaining sentence pairs. Using these scores, we selected the top 100,000 highest-quality translations for use in our fine-tuning experiments. The strategy of mining large datasets and selecting top-k sentence pairs based on quality metrics for fine-tuning helps to further filter out noisy segments and ensures that the limited available data contribute maximally to model training [22]. This approach is consistent with our goal of reducing computational costs. By carefully curating a smaller but higher-quality dataset, we limit energy consumption and the associated environmental costs, while maximizing translation performance.

Table 1
Corpus size after each curation step.

Step	Pairs Remaining
Original Dataset	50,000,000
Training Set	100,000
Test Set	400

3.2. Training

The Gemma-3-4B model was fine-tuned for three epochs using Low-Rank Adaptation (LoRA) [23], a fine-tuning technique which injects small trainable matrices in the model's weights. The adoption of LoRA for fine-tuning has shown strong empirical results in machine translation [24, 25], enhancing efficiency, while reducing training time and computational costs. As demonstrated in experiments conducted by [26], fine-tuning with LoRA obtained the same improvements in terms of BLEU score [27], while drastically reducing training time and modifying only a small number of trainable parameters, with respect to supervised fine-tuning involving all parameters of the original network. In our case, we train effectively 0.42% of the trainable parameters, corresponding to the LoRA adapter matrices injected in Gemma-3-4B.

Our fine-tuning pipeline was implemented using the Hugging Face Transformers library [28], leveraging its integration with the PEFT library. For the LoRA configuration, we set the rank (`r`) to 16 and the scaling factor (`alpha`) to 16, with a dropout rate of 0.05 to improve generalization. The training was carried out on a single NVIDIA A100 GPU using mixed-precision (fp16) computation. We used the `CodeCarbon`[1] library to monitor the environmental impact of our training process.

CodeCarbon is a Python library that estimates carbon emissions by tracking the energy consumption of computing resources (CPU, GPU, RAM) during code execution and combining this data with the carbon intensity of the electricity grid based on geographic location.

The fine-tuning session consumed approximately **0.65 kWh**, resulting in an estimated **162 g CO_2eq** under an average EU grid intensity of 250 gCO_2/kWh.

3.3. Gemma-3 Evaluation

We conduct our evaluation on a held-out test set of 400 segments from the same corpus, ensuring no overlap with the training data. Table 2 reports the evaluation of EN–IT translation performance for Gemma-3-4B before and after LoRA fine-tuning, using BLEU [27], chrF [29], and COMET [30] as quality metrics. Our fine-tuned Gemma-3-4B model, with only 0.42% of additional trainable parameters, shows a notable improvement over the base version, achieving a +4 point gain in BLEU, a modest increase in chrF, and a +1 point gain in COMET. These results place our model on par with GPT-4o in COMET and above GPT-4o-mini in all three metrics.

In addition to performance, we also measure the environmental impact of inference using the CodeCarbon library. The estimated carbon emissions per inference for the fine-tuned model are approximately 0.028g CO_2eq, twice that of the base model, but significantly lower than GPT-4o models, each exceeding 0.42g per inference as estimated in a relevant study [31].

Our evaluation demonstrates that fine-tuning Gemma-3-4B with LoRA leads to competitive performance gains with low additional environmental cost.

4. Quality-Sustainability Trade-Off

In our second experiment, to further assess the viability of trading off quality for sustainability with the use of SLMs, we extend our evaluation on a set of multilingual LMs, of different parameter sizes. We select the models for our evaluation based on state-of-the-art performance and usage in the research community. We benchmark each model on the same held-out EN–IT test set, using BLEU, chrF and COMET, and log the CO_2eq emissions per inference using the CodeCarbon framework. Importantly, we emphasize in our approach that a sustainable model choice should not be based on its parameter size alone, but actual carbon emissions.

As shown in Table 3, we highlight that the relationship between model size and emissions is **non-linear**. For instance, Qwen-3B [32], despite its relatively small size, exhibits disproportionately high emissions. This can be attributed to its reasoning behavior during inference,

[1] https://mlco2.github.io/codecarbon/index.html

Table 2
Evaluation of EN–IT translation performance for Gemma-3-4B before and after LoRA fine-tuning. Metrics include BLEU, chrF, and WMT22 COMET-DA. We also report estimated CO_2eq emissions per inference.

Model	BLEU	chrF	COMET	CO_2eq (g)
Gemma-3-4B (Base)	46.0	69.0	93.0	0.014
Gemma-3-4B (Ours)	50.0	72.0	**94.0**	0.028
GPT-4o-mini	49.0	71.0	92.0	*>0.42*
GPT-4o	52.0	73.0	**94.0**	*>0.42*

which results in extended reasoning outputs before generating a final answer. This behavior increases inference latency and environmental cost.

Similarly, the assumption that larger models necessarily produces more carbon emissions does not always hold. This is the case for models developed with a Mixture-of-Experts (MoE) architectures. In these models, only a subset of the total parameters is activated during inference. As a result, MoE models like Mixtral, although large in aggregate size, can have lower or comparable emissions to smaller, densely activated models. This decoupling of parameter size and runtime efficiency highlights the need for measuring more empirical results, such as CO_2eq emissions.

Therefore, we introduce a **Carbon-Adjusted Quality Score** (CAQS) metric as a measure of model cost-effectiveness, and we calculate it on each corpus translation generated by the models evaluated in our study. Our CAQS score penalizes each gram of carbon emissions exponentially, while ensuring that low-quality models are not rewarded more than high-quality ones, regardless of their efficiency. We define the CAQS metric as follows.

$$\text{CAQS} = \text{avg}(\text{METRICS}) \times \exp(-\lambda \times \text{CO}_2\text{eq}) \quad (1)$$

Here, λ is a sensitivity parameter that controls the strength of the carbon penalty and can be adjusted according to the user's desired trade-off between quality and sustainability. The exponential penalty function reflects the urgent need for sustainable AI, where a single increase in emissions becomes increasingly problematic. In our experiment, we use $\lambda = 2$ and provide ranking for interpretability.

Table 3 shows that Gemma-3-4B and Magistral-Small [33] rank first according to our metric, while larger and slighly superior models, like Llama-3.3-70B [34], are strongly penalised due to their high emissions. Similarly, we find that low-quality models, like Phi-2 [35] and Llama-3.2-1B are not exceedingly rewarded.

We emphasize the need for sustainable model choices in both industrial and academic settings, and recommend the adoption of a standardized approach: measuring CO_2eq emission using CodeCarbon or similar tools on a representative sample of the target corpus, then calculating a carbon-adjusted score that considers both translation quality and sustainability.

5. Error Analysis

To complement the quantitative results and better understand the practical implications of the quality-sustainability trade-off, we conduct a manual error analysis on the translations generated by four representative models: our fine-tuned version of Gemma-3-4B, and the baseline instruction-tuned Gemma-3-27B, Llama-3.2-3B and Llama-3.3-70B.

Annotation Process. We conducted our error analysis following the MQM framework [36], with two annotators who were native speakers of the target language, proficient in English, and with expertise in translation studies. The annotators applied a set of MQM categories: accuracy, fluency, style, locale conventions, and verity, along with their respective subcategories. Errors were rated using four severity levels: trivial, minor, major, and critical, corresponding to weights of 0, 1, 5, and 25, respectively.

After annotating 10% of the dataset, inter-annotator agreement (IAA) was calculated to ensure the reliability of the annotations. The initial agreement, measured with Cohen's Kappa, was equal to $K = 0.28$, due to disagreements primarily on the severity levels to assign, rather than the identification of the error categories themselves. Following a collaborative resolution process, we refined the annotation guidelines and calculated agreement on the final annotations, reaching a Cohen's Kappa equal to $K = 0.53$. The annotators proceeded separately and annotated the translations generated by two models each.

Annotation Results. The results, displayed in Table 4 indicate that Gemma-3-27B, the largest model in the Gemma family, produced the fewest overall errors, with only one major error and 8 minor ones. In the context of our study, minor errors were defined as those that do not significantly alter the meaning expressed by the source text. Interestingly, we find that Gemma-3-4B demonstrates comparable performance to the much

Table 3
Comparison of translation quality and CO_2eq emissions per inference for various multilingual models on the EN–IT test set. Models are sorted and ranked by CAQS, where higher CAQS values indicate better effciency.

Model	Params (B)	BLEU	chrF	COMET	CO_2eq (g)	CAQS	Rank
Gemma-3-4B	4.0	50.0	72.0	94.0	0.028	68.08	**1**
Magistral-Small	7.0	48.6	70.2	92.7	0.053	63.41	2
Llama-3.2-3B	3.0	37.4	62.6	90.0	0.019	60.97	3
Gemma-3-27B	27.0	49.3	72.8	93.9	0.112	57.42	4
Llama-3.3-70B	70.0	49.5	71.3	93.6	0.115	56.78	5
Llama-3.2-1B	1.0	19.8	46.0	76.5	0.005	46.96	6
Phi-2	2.7	6.8	32.1	49.5	0.015	28.60	7
Qwen-3B	3.0	40.3	65.2	92.4	0.503	24.12	8

Table 4
Error severity distribution across models. The final score represents the weighted sum of all errors.

Model	Critical	Major	Minor	Score
Gemma-3-27B	0	1	8	13
Llama-3.3-70B	0	3	29	44
Gemma-3-4B	0	4	29	49
Llama-3.2-3B	1	28	56	221

larger and environmentally demanding model, Llama-3.3-70B. In terms of weighted scores, both models show similar results, with very few major errors and a comparable number of minor ones. The smallest Llama checkpoint presents a very high number of both major and minor errors, when compared to the Gemma-3-4B model. The findings may suggest that Llama-3's architecture is suboptimal for translation tasks across model sizes, given that Gemma-3-4B matches the performance of its largest checkpoint. However, the results should be interpreted with caution, as our evaluation was limited to a small test set and a single language pair.

In terms of error category distribution increasing parameter size leads to an overall performance improvement, as seen in Table 5. This trend is particularly evident within the Gemma models, where the jump from 4B to 27B parameters results in a significant drop in errors across all categories. In contrast, Llama-3.2 models exhibit a less linear improvement, suggesting diminishing returns from scaling model size. This observation, however, is limited by the fact that only the smallest Gemma model was LoRA-adapted, while the LLaMA models were evaluated in their original form. A more rigorous comparison, involving both original and adapted versions across model sizes, is left for future work.

When comparing Gemma-3-4B and Llama-3.3-70B, we find that most of the errors in the Gemma model are concentrated in surface-level issues, especially in spelling diacritics. These errors, however, do not compromise the overall understandability of the output. In contrast, Llama-3.3-70B displays fewer fluency issues but a higher number of style-related errors, including two rated as major. These style errors typically result in translations that sounds unnatural or awkward for a target-language speaker, thereby reducing the overall quality of the translation.

6. Conclusions

In this study, we investigated the potential of SLMs as sustainable alternatives to LLMs, for MT tasks focusing on the EN-IT language pair. Our results demonstrate that parameter-efficient fine-tuning of SLMs can achieve competitive translation quality while dramatically reducing environmental impact. The fine-tuned Gemma-3-4B model achieved performance comparable to GPT-4o and outperformed GPT-4o-mini across all metrics, while consuming approximately 15 times less energy per inference.

We complement these results with a MQM human evaluation across a set of representative models, confirming that Gemma-3-4B performed comparably to the much larger Llama-3.3-70B, producing only minor fluency and spelling errors.

We also highlighted that the relationship between model size and carbon emissions is non-linear and highly dependent on architectural choices, emphasizing the need for accurate measurements of carbon emissions.

Given the non-linear relation between model size and environmental impact, we introduced the CAQS, a novel metric specifically designed to facilitate sustainable model selection by integrating translation quality and carbon emissions. CAQS includes a sensitivity parameter that allows users to adjust how strongly quality is penalized by the model's carbon footprint. According to this metric, Gemma-3-4B and Magistral-Small emerged as the most efficient models in our study, offering optimal trade-offs between sustainability and translation quality.

Table 5
MQM error category breakdown per 100 segments for each model.

Model	Accuracy	Fluency	Style	Others	Total
Gemma-3-4B	10	17	6	0	33
Gemma-3-27B	7	1	1	0	9
Llama-3.2-3B	40	16	29	0	85
Llama-3.3-70B	8	9	15	0	32

7. Limitations

In light of practical constraints related to time and resources, the main limitations of our study lie in the relatively small sample of segments and the domain-specific nature of the OpenSubtitles corpus, used for both training and inference. For this reason, we highlight that our evaluation results may not be reproducible in other domains.

As our evaluation focuses on a relatively high-resource language pair (EN-IT), our findings may not be applicable for distant or low-resource pairs. Finally, our carbon emission measurements are specific to the computational infrastructure used (NVIDIA A100 GPUs, EU electricity grid). Results may differ when deploying models on different hardware configurations, cloud providers, or geographical regions.

Acknowledgments

This work has been funded by the Italian National PhD programme in Artificial Intelligence, partnered by University of Pisa and University of Naples "L'Orientale", through a doctoral grant (ID 39-411-24-DOT23A27WJ-6603) established by Ex DM 318, of type 4.1, co-financed by the National Recovery and Resilience Plan.

References

[1] C. Lyu, Z. Du, J. Xu, Y. Duan, L. Wang, New trends in machine translation with large language models, 2023.

[2] W. Jiao, W. Wang, J.-t. Huang, X. Wang, S. Shi, Z. Tu, Is chatgpt a good translator? yes with gpt-4 as the engine, arXiv preprint arXiv:2301.08745 (2023).

[3] A. Singh, N. P. Patel, A. Ehtesham, S. Kumar, T. Talaei Khoei, A survey of sustainability in large language models: Applications, economics, and challenges, arXiv preprint arXiv:2412.04782 (2025).

[4] M. C. Rillig, M. Ågerstrand, M. Bi, K. A. Gould, U. Sauerland, Risks and benefits of large language models for the environment, Environmental Science & Technology 57 (2023) 3464–3466. doi:10.1021/acs.est.3c01106.

[5] Google DeepMind, Gemma: Open models for responsible ai, https://deepmind.google/models/gemma/, 2024. Accessed: 2025-05-27.

[6] A. Hendy, M. Abdelrehim, A. Sharaf, V. Raunak, M. Gabr, H. Matsushita, Y. J. Kim, M. Afify, H. H. Awadalla, How good are gpt models at machine translation? a comprehensive evaluation, 2023. URL: https://arxiv.org/abs/2302.09210. arXiv:2302.09210.

[7] Z. He, T. Liang, W. Jiao, Z. Zhang, Y. Yang, R. Wang, Z. Tu, S. Shi, X. Wang, Exploring human-like translation strategy with large language models, Transactions of the Association for Computational Linguistics 12 (2024) 229–246. doi:10.1162/tacl_a_00642.

[8] Y. Moslem, R. Haque, J. D. Kelleher, A. Way, Adaptive machine translation with large language models, arXiv preprint arXiv:2301.13294 (2023).

[9] L. Wang, C. Lyu, T. Ji, Z. Zhang, D. Yu, S. Shi, Z. Tu, Document-level machine translation with large language models, arXiv preprint arXiv:2304.02210 (2023).

[10] U.S. Energy Information Administration, Electricity use in homes, 2023. URL: https://www.eia.gov/energyexplained/use-of-energy/electricity-use-in-homes.php, accessed: 2025-06-16.

[11] S. Nguyen, B. Zhou, Y. Ding, S. Liu, Towards sustainable large language model serving, 2024. URL: https://arxiv.org/abs/2501.01990. arXiv:2501.01990.

[12] F. Wang, Z. Zhang, X. Zhang, Z. Wu, T. Mo, Q. Lu, W. Wang, R. Li, J. Xu, X. Tang, Q. He, Y. Ma, M. Huang, S. Wang, A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness, 2024. URL: https://arxiv.org/abs/2411.03350. arXiv:2411.03350.

[13] Y.-C. Lin, S. Sharma, H. Manikandan, J. Kumar, T. H. King, J. Zheng, Efficient multitask learning in small language models through upside-down reinforcement learning, 2025. URL: https://arxiv.org/abs/2502.09854. arXiv:2502.09854.

[14] Y. Tang, K. Han, F. Liu, Y. Ni, Y. Tian, et al., Rethink-

ing optimization and architecture for tiny language models, in: Proceedings of the 41st International Conference on Machine Learning, PMLR, 2024.

[15] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, et al, Gemma 3 technical report, 2025. URL: https://arxiv.org/abs/2503.19786. arXiv:2503.19786.

[16] Y. Fu, H. Peng, L. Ou, A. Sabharwal, T. Khot, Specializing smaller language models towards multi-step reasoning, in: Proceedings of the 40th International Conference on Machine Learning, PMLR, 2023.

[17] C. V. Nguyen, X. Shen, R. Aponte, Y. Xia, et al., A survey of small language models, 2024. arXiv:2410.20011.

[18] OpenAI, Gpt-4o, https://openai.com/gpt-4o, 2024. Accessed: 2025-07-23.

[19] J. Tiedemann, S. Thottingal, OPUS-MT – Building open translation services for the World, in: A. Martins, H. Moniz, S. Fumega, B. Martins, F. Batista, L. Coheur, C. Parra, I. Trancoso, M. Turchi, A. Bisazza, J. Moorkens, A. Guerberof, M. Nurminen, L. Marg, M. L. Forcada (Eds.), Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, European Association for Machine Translation, Lisboa, Portugal, 2020, pp. 479–480. URL: https://aclanthology.org/2020.eamt-1.61.

[20] S. Nakatani, Langdetect: Language detection library for python, https://pypi.org/project/langdetect/, 2014. Port of Google's language-detection library.

[21] R. Rei, J. G. C. de Souza, D. Alves, C. Zerva, A. C. Farinha, T. Glushkova, A. Lavie, L. Coheur, A. F. T. Martins, COMET-22: Unbabel-IST 2022 submission for the metrics shared task, in: P. Koehn, L. Barrault, O. Bojar, F. Bougares, R. Chatterjee, M. R. Costa-jussà, C. Federmann, M. Fishel, A. Fraser, M. Freitag, Y. Graham, R. Grundkiewicz, P. Guzman, B. Haddow, M. Huck, A. Jimeno Yepes, T. Kocmi, A. Martins, M. Morishita, C. Monz, M. Nagata, T. Nakazawa, M. Negri, A. Névéol, M. Neves, M. Popel, M. Turchi, M. Zampieri (Eds.), Proceedings of the Seventh Conference on Machine Translation (WMT), Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 2022, pp. 578–585. URL: https://aclanthology.org/2022.wmt-1.52/.

[22] E. A. Chimoto, B. A. Bassett, Comet-qe and active learning for low-resource machine translation, 2022. URL: https://arxiv.org/abs/2210.15696. arXiv:2210.15696.

[23] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-Rank Adaptation of Large Language Models, 2021. URL: http://arxiv.org/abs/2106.09685. doi:10.48550/arXiv.2106.09685, arXiv:2106.09685 [cs].

[24] J. Zheng, H. Hong, F. Liu, X. Wang, J. Su, Y. Liang, S. Wu, Fine-tuning large language models for domain-specific machine translation, 2024. URL: https://arxiv.org/abs/2402.15061. arXiv:2402.15061.

[25] D. M. Alves, N. M. Guerreiro, J. Alves, J. Pombal, R. Rei, J. G. C. de Souza, P. Colombo, A. F. T. Martins, Steering large language models for machine translation with finetuning and in-context learning, 2023. URL: https://arxiv.org/abs/2310.13448. arXiv:2310.13448.

[26] X. Zhang, N. Rajabi, K. Duh, P. Koehn, Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA, in: P. Koehn, B. Haddow, T. Kocmi, C. Monz (Eds.), Proceedings of the Eighth Conference on Machine Translation, Association for Computational Linguistics, Singapore, 2023, pp. 468–481. URL: https://aclanthology.org/2023.wmt-1.43/. doi:10.18653/v1/2023.wmt-1.43.

[27] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: P. Isabelle, E. Charniak, D. Lin (Eds.), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040/. doi:10.3115/1073083.1073135.

[28] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, HuggingFace's Transformers: State-of-the-art Natural Language Processing, 2020. URL: http://arxiv.org/abs/1910.03771. doi:10.48550/arXiv.1910.03771, arXiv:1910.03771 [cs].

[29] M. Popović, chrF: character n-gram F-score for automatic MT evaluation, in: O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, C. Hokamp, M. Huck, V. Logacheva, P. Pecina (Eds.), Proceedings of the Tenth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Lisbon, Portugal, 2015, pp. 392–395. URL: https://aclanthology.org/W15-3049/. doi:10.18653/v1/W15-3049.

[30] R. Rei, C. Stewart, A. C. Farinha, A. Lavie, COMET: A neural framework for MT evaluation, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online,

2020, pp. 2685–2702. URL: https://aclanthology.org/2020.emnlp-main.213/. doi:10.18653/v1/2020.emnlp-main.213.

[31] N. Jegham, M. Abdelatti, L. Elmoubarki, A. Hendawi, How hungry is ai? benchmarking energy, water, and carbon footprint of llm inference, 2025. URL: https://arxiv.org/abs/2505.09598. arXiv:2505.09598.

[32] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu, F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R. Gao, S. Liu, S. Luo, T. Li, T. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou, Z. Qiu, Qwen3 technical report, 2025. URL: https://arxiv.org/abs/2505.09388. arXiv:2505.09388.

[33] Mistral-AI, :, A. Rastogi, A. Q. Jiang, A. Lo, G. Berrada, G. Lample, J. Rute, J. Barmentlo, K. Yadav, e. a. Kartik Khandelwal, Magistral, 2025. URL: https://arxiv.org/abs/2506.10910. arXiv:2506.10910.

[34] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, e. a. Alex Vaughan, The llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

[35] M. Abdin, J. Aneja, H. Awadalla, A. Awadallah, A. A. Awan, N. Bach, A. Bahree, A. Bakhtiari, J. Bao, e. a. Harkirat Behl, Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL: https://arxiv.org/abs/2404.14219. arXiv:2404.14219.

[36] A. Lommel, S. Gladkoff, A. Melby, S. E. Wright, I. Strandvik, K. Gasova, A. Vaasa, A. Benzo, R. M. Sparano, M. Foresi, J. Innis, L. Han, G. Nenadic, The multi-range theory of translation quality measurement: Mqm scoring models and statistical quality control, 2024. URL: https://arxiv.org/abs/2405.16969. arXiv:2405.16969.