DLUT and GTCOM's Large Language Model Based Translation System for WMT25

Hao Zong[1,2] Chao Bei[2] Conghu Yuan[2]
Wentao Chen[2] Huan Liu[2] Degen Huang[1]*
[1]Dalian University of Technology
[2]Global Tone Communication Technology Co., Ltd.
zonghao@mail.dlut.edu.cn
{beichao, yuanconghu, chenwentao and liuhuan}@gtcom.com.cn
huangdg@dlut.edu.cn

Abstract

This paper presents the submission from Dalian University of Technology (DLUT) and Global Tone Communication Technology Co., Ltd. (GTCOM) to the WMT25 General Machine Translation Task. Amidst the paradigm shift from specialized encoder-decoder models to general-purpose Large Language Models (LLMs), this work conducts a systematic comparison of both approaches across five language pairs. For traditional Neural Machine Translation (NMT), we build strong baselines using deep Transformer architectures enhanced with data augmentation. For the LLM paradigm, we explore zero-shot performance and two distinct supervised fine-tuning (SFT) strategies: *direct translation* and *translation refinement*. Our key findings reveal a significant discrepancy between lexical and semantic evaluation metrics: while strong NMT systems remain competitive in BLEU scores, fine-tuned LLMs demonstrate marked superiority in semantic fidelity as measured by COMET. Furthermore, we find that fine-tuning LLMs for direct translation is more effective than for refinement, suggesting that teaching the core task directly is preferable to correcting baseline outputs.

1 Introduction

The field of machine translation is undergoing a profound paradigm shift, marked by the ascent of general-purpose Large Language Models (LLMs) that challenge the dominance of specialized encoder-decoder Neural Machine Translation (NMT) architectures (Vaswani et al., 2017). For years, NMT systems, meticulously trained on vast parallel corpora, have been honed into highly effective, specialized tools for a single task: translation (Ott et al., 2019). In contrast, LLMs, pretrained on web-scale multilingual and multimodal data, emerge as powerful generalists, possessing not only cross-lingual capabilities but also extensive world knowledge and reasoning skills (Brown et al., 2020), which they can apply to translation with remarkable zero-shot proficiency. This dichotomy between the "specialized artisan" (NMT) and the "generalist polymath" (LLM) raises critical questions about the future trajectory of machine translation research.

This transition is further complicated by an evolution in evaluation philosophy. The community is increasingly moving away from lexical overlap metrics like BLEU (Papineni et al., 2002), which may unduly penalize valid, fluent translations that diverge stylistically from a single reference. The rise of semantic-aware metrics such as COMET (Rei et al., 2020) and its successor, XCOMET-XL (Guerreiro et al., 2023), reflects a demand for evaluations that prioritize meaning and fidelity. This shift is particularly pertinent when comparing NMT and LLMs, as LLMs often excel at producing semantically coherent and contextually appropriate outputs that might be lexically dissimilar to the reference. A core challenge, therefore, is to conduct a fair comparison that accounts for this evaluation dichotomy.

In this paper, we leverage our participation in the WMT25 General Machine Translation task as a standardized testbed to systematically investigate this ongoing paradigm shift. Our work is guided by two central research questions (RQs):

1. **(RQ1)** How do the performance characteristics of specialized NMT systems and general-purpose LLMs diverge, particularly under the contrasting lenses of lexical (BLEU) and semantic (COMET) evaluation metrics?

2. **(RQ2)** Among supervised fine-tuning (SFT) strategies for adapting LLMs to translation, which is more effective: direct instruction on source-to-target mapping (*direct translation*), or training the model to correct outputs from a baseline system (*translation refinement*)?

*Corresponding Author

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 285–292
November 8–9, 2025. ©2025 Association for Computational Linguistics

To address these questions, we developed a comprehensive suite of systems. Our NMT pipeline features deep Transformer models trained with the `fairseq` toolkit, enhanced by data augmentation. Our LLM pipeline is built upon the powerful Gemma3 model family (Team et al., 2025), which we adapt using the `LLaMa-Factory` framework (Zheng et al., 2024). Our main contributions are: (1) a robust empirical comparison of NMT and LLM systems across five language pairs, revealing a significant divergence between lexical and semantic evaluation scores; (2) a direct analysis of two distinct LLM fine-tuning strategies, demonstrating the superior efficacy of direct translation; and (3) insights into the qualitative differences between the outputs of these systems, highlighting the semantic strengths of modern LLMs.

2 Task Description

The core of this task is bilingual text translation. The data, sourced using the 'mtdata' tool(Gowda et al., 2021) from the official WMT25 repository, consists of both parallel and monolingual corpora. Table 1 provides a detailed breakdown of the training data statistics. For our development and testing sets, we used newstest2019 for the Czech→German direction, wmttest2024 for Czech→Ukrainian, English→German, and English→Ukrainian, and flores200-devtest (NLLB Team, 2022) for English→Serbian.

3 Methodology

Our methodology is designed as a comparative study of two distinct translation paradigms. We first establish a strong baseline representing specialized NMT systems and then build upon a generalist LLM foundation, exploring different adaptation strategies.

3.1 Data Foundation: Preprocessing and Quality Filtering

A high-quality dataset is the bedrock of any translation system. Our preprocessing pipeline is standardized across all languages and includes punctuation normalization, tokenization, Truecasing, and Byte Pair Encoding (BPE) (Sennrich et al., 2015) to manage vocabulary size and handle rare words.

Beyond standard preprocessing, we implemented a rigorous quality filtering stage using the CometKiwi tool ('wmt23-cometkiwi-da-xl' model) (Rei et al., 2023). For our LLM fine-tuning, we adopted a nuanced data selection strategy. Rather than simply taking the top-N scoring sentence pairs, we extracted a 100,000-pair subset ranked between the 10,000th and 110,000th positions. This decision is based on the hypothesis that the absolute highest-scoring pairs often consist of overly simplistic, short, or formulaic sentences (e.g., from translation memories), which can lead to models that are fluent but lack complexity. By targeting a "high-quality but challenging" segment, we aim to create a more diverse and robust instruction dataset for fine-tuning.

Data Type	**Number of Sentences**
Parallel Data	
cs-de	120.39M
cs-uk	10.62M
en-uk	24.6M
en-ru	77.5M
en-sr	114.04M
Monolingual Data	
English	35M
Czech	42.6M
Ukrainian	14.8M
German	72.8M
Serbian	56.8M
Russian	56.2M
Development Sets	
cs-de	1997
cs-uk	2316
en-uk	997
en-ru	997
en-sr	1012

Table 1: Statistics for the training and development datasets.

3.2 Paradigm 1: Specialized NMT Systems

To represent the best in specialized NMT, we employed a deep Transformer architecture ('transformer_wmt_en_de' configuration in `fairseq`). These models, featuring 24 encoder and 24 decoder layers, serve as our high-performance baseline. To maximize the utility of available monolingual data—a cornerstone of competitive NMT—we incorporated iterative data augmentation:

1. **Back-Translation (BT):** We trained reverse-direction models (e.g., ru→en) to translate target-language monolingual data into the

source language, creating a large, synthetic parallel corpus to augment the primary training data.

2. **Forward-Translation (FT):** The improved models from the BT step were then used to translate source-language monolingual data, further enriching the training mixture in a subsequent iteration.

3.3 Paradigm 2: Generalist LLM-based Systems

Our exploration of the generalist paradigm centers on the Gemma3 model family, selected for its strong preliminary multilingual performance. Our approach systematically moves from zero-shot evaluation to targeted adaptation.

3.3.1 Foundation Model and Zero-Shot Baseline

We first established a zero-shot baseline by evaluating several prominent instruction-tuned LLMs (including the Qwen3 and Gemma3 series) on our development sets using a direct translation prompt. This step measures the intrinsic, out-of-the-box translation capabilities of these models without any task-specific training.

3.3.2 Supervised Fine-Tuning (SFT) Strategies

To adapt Gemma3 for high-quality translation, we investigated two distinct SFT strategies, each testing a different hypothesis about how LLMs best learn this complex task. **To achieve the most thorough adaptation possible, we performed full-parameter fine-tuning, allowing all weights of the base model to be updated during the training process.** This approach, while computationally intensive, ensures that the model can fully specialize its internal representations for the translation task.

Strategy 1: Direct Translation. In contrast, this strategy reframes the task from generation to a more complex process of critique and correction. The model is provided with a triplet: the source text, a potentially flawed translation from our NMT baseline, and the high-quality reference. The hypothesis is that by learning to identify and correct errors—essentially, learning the "delta" between a mediocre and an excellent translation—the model develops a more nuanced understanding of quality, error patterns, and stylistic appropriateness. This task is guided by a prompt that casts the LLM in the role of a professional "post-editor":

```
You are an expert in {Source
    language}-{Target language}
    translation, with a deep
    understanding of both languages'
    cultural nuances. Your
    translations are accurate, fluent
    , and elegant. Please translate
    the following {Source language}
    text into {Target language}. Only
     output the translation.

{Source language} text: {Source text
    }
{Target language} translation:
{Target text}
```

Strategy 2: Translation Refinement. This strategy reframes the task from generation to critique and correction. The model is provided with a source text, a potentially flawed translation from our NMT baseline, and the high-quality reference. It is then instructed to "polish" or "refine" the baseline translation. The hypothesis here is that learning to identify and correct errors is a more cognitively demanding task that could foster a deeper, more nuanced understanding of translation quality, error patterns, and stylistic appropriateness, potentially leading to a more robust translator. The prompt for translation refinement is as follows:

```
You are a professional {Source
    language}-{Target language}
    translation refinement expert who
     excels at making machine-
    translated content more natural
    and fluent, ensuring it aligns
    better with the target language's
     norms and contexts. Based on the
     provided source text and machine
     translation, refine and modify
    the translation to make it more
    accurate and natural.

{Source language} source: {Source
    text}
{Target language} translation: {
    Baseline translation text}
{Target language} corrected
    translation:
{Target text}
```

4 Experimental Setup

Our experiments were designed to ensure a fair and reproducible comparison between the NMT and LLM paradigms.

4.1 NMT System Configuration

We used the `fairseq-py` toolkit for all NMT experiments. Our deep Transformer models were trained with the following configuration:

- **Architecture:** 'transformer_wmt_en_de' (24 encoder/decoder layers, 16 attention heads, embedding size of 1024).
- **Optimizer:** Adam (Kingma and Ba, 2017) with $\beta_1 = 0.9$, $\beta_2 = 0.98$.
- **Learning Rate Schedule:** Inverse square root scheduler with a warm-up of 4,000 steps and a peak learning rate of 5×10^{-4}.
- **Regularization:** Dropout was set to 0.3 for the attention and activation functions, and label smoothing of 0.1 was applied.
- **Batching:** We used a maximum of 4096 tokens per batch per GPU. Models were trained for 100,000 steps or until convergence on the development set.

4.2 LLM System Configuration

All LLM experiments were conducted using the `LLaMa-Factory` framework.

- **Base Models:** We used the instruction-tuned versions of the Gemma3 family: 'Gemma3-12B-it' and 'Gemma3-27B-it'.
- **Fine-Tuning Method:** We employed **full-parameter supervised fine-tuning**. This involves updating all of the model's weights, rather than using a parameter-efficient method.
- **Hyperparameters:** Models were trained for 3 epochs over the 100k-pair instruction dataset. We used the AdamW optimizer with a learning rate of 2×10^{-5}, a cosine learning rate scheduler, and a warm-up ratio of 0.03. Training was performed with `bfloat16` mixed-precision to optimize memory usage and throughput.

4.3 Evaluation Metrics

To provide a multifaceted view of translation quality, we report scores from two distinct metrics:

- **sacreBLEU** (Post, 2018): A standardized implementation of BLEU that measures n-gram precision against a reference translation. It primarily reflects lexical similarity.
- **XCOMET-XL** (Guerreiro et al., 2023): A state-of-the-art semantic metric that uses a large pre-trained model to assess the meaning equivalence between the source, hypothesis, and reference. This metric aligns more closely with human judgments of translation quality.

5 Results and Discussion

In this section, we analyze our experimental results to answer the research questions posed in the introduction. We dissect the performance of each paradigm and discuss the implications of our findings. The comprehensive results for our NMT baselines and zero-shot LLM evaluations are consolidated in Table 2.

5.1 RQ1: NMT vs. LLMs and the BLEU-COMET Dichotomy

Our first research question explores the performance divergence between specialized NMT and generalist LLMs. The results in Table 2 reveal a fascinating and consistent trend that we term the BLEU-COMET dichotomy.

NMT systems remain formidable competitors on lexical metrics, often outperforming even large LLMs in BLEU score. This is most evident in the en→sr direction, where the NMT baseline achieves a BLEU score of 38.44, significantly higher than any other system. Similarly, for cs→de, the NMT baseline's BLEU of 29.67 is the highest in its category. Regarding data augmentation, back-translation shows a clear benefit for lower-resource pairs (e.g., providing a 1.73 BLEU point gain for cs→uk), but its impact diminishes or even slightly degrades BLEU on high-resource pairs like cs→de. Furthermore, forward-translation consistently proves detrimental to performance across most pairs, likely due to the introduction of unmitigated noise.

In contrast, LLMs exhibit a clear and striking superiority in semantic fidelity, even in a zero-shot setting. The Gemma3-27B-it model achieves the highest zero-shot COMET score in four out of five language pairs. The most dramatic example is en-uk, where the Gemma3-27B-it's COMET score of 80.31 massively surpasses the NMT system's 67.55, despite having only a marginal advantage in BLEU. This pattern holds for cs→de (94.24 vs. 93.71), cs→uk (91.47 vs. 88.65), and en→ru (91.73 vs. 91.55). The only exception is en→sr, where the NMT baseline's COMET score

	cs→de		cs→uk		en→uk		en→ru		en→sr	
Model / System	**BLEU**	**COMET**	**BLEU**	**COMET**	**BLEU**	**COMET**	**BLEU**	**COMET**	**BLEU**	**COMET**
Specialized NMT Systems										
NMT Baseline	**29.67**	93.71	26.59	85.71	25.97	66.02	28.21	90.11	**38.44**	**90.45**
+ Back-translation	29.59	92.80	28.32	88.65	26.22	67.55	**28.86**	91.55	36.05	86.21
+ Forward-translation	26.37	82.50	24.50	82.50	25.25	66.10	28.55	90.81	31.66	83.54
Generalist LLMs (Zero-shot)										
Qwen3-8B	12.72	90.70	13.52	84.25	17.42	67.53	19.79	86.21	10.41	69.25
Qwen3-14B	22.24	92.63	26.57	87.65	23.44	71.52	27.59	87.69	10.58	78.27
Gemma3-12B-it	24.29	93.62	29.24	91.01	25.65	79.08	26.43	91.01	11.42	86.27
Gemma3-27B-it	25.53	**94.24**	**30.50**	**91.47**	**27.22**	**80.31**	28.02	**91.73**	26.62	90.25

Table 2: Comprehensive results comparing our specialized NMT systems against zero-shot performance of generalist LLMs across all five language pairs. While NMT+BT often leads in BLEU, the Gemma3-27B-it model consistently achieves the highest COMET scores, highlighting the BLEU-COMET dichotomy.

	cs→de		cs→uk		en→uk		en→ru		en→sr	
Model and SFT Strategy	**BLEU**	**COMET**	**BLEU**	**COMET**	**BLEU**	**COMET**	**BLEU**	**COMET**	**BLEU**	**COMET**
Gemma3-12B-it Fine-tuned										
Direct Translation SFT	25.45	94.01	27.01	91.23	28.25	80.23	28.31	90.11	31.28	87.16
Refinement SFT	25.71	94.17	26.60	91.57	24.98	79.88	28.01	87.66	30.49	86.59
Gemma3-27B-it Fine-tuned										
Direct Translation SFT	**26.69**	**94.50**	**30.50**	**92.10**	**29.57**	**81.56**	**29.53**	**91.50**	**31.90**	**90.97**
Refinement SFT	26.32	94.32	29.94	91.01	27.83	80.61	29.26	90.50	31.33	88.97

Table 3: Results of supervised fine-tuning on Gemma3 models. The Direct Translation strategy consistently outperforms the Refinement strategy across nearly all models and language pairs. The fine-tuned Gemma3-27B-it with Direct SFT emerged as our best overall system.

is competitive (90.45 vs. 90.25). This powerful trend suggests that LLMs' vast world knowledge allows them to generate more fluent and semantically equivalent translations, a quality that is rewarded by COMET but can be unfairly penalized by BLEU's rigid lexical matching.

5.2 RQ2: Efficacy of SFT Strategies

Our second research question investigates the more effective SFT strategy for adapting LLMs to translation. The results from our fine-tuning experiments, presented in Table 3, provide a decisive answer.

Direct Translation consistently and significantly outperforms Translation Refinement. For both the 12B and 27B model sizes and across all five language pairs, the models fine-tuned with the direct translation task achieved superior scores on both BLEU and COMET. For instance, in the en-uk direction, the Gemma3-27B-it model fine-tuned for direct translation achieved a COMET score of 81.56, while the refinement-tuned model scored only 80.61. Similarly, for en-sr, the direct translation model achieved a COMET of 90.97, a full two points higher than the refinement model's 88.97. The Gemma3-27B-it with Direct Translation SFT emerged as our best overall system, achieving the highest COMET score across the board.

We attribute this clear victory to two primary factors. First, the refinement task introduces a higher cognitive load: the model must simultaneously comprehend the source, analyze the errors in a flawed translation, and generate a correction. This may represent a less direct and noisier learning signal. Second, the provided baseline translation may act as a negative anchor, implicitly constraining the model's output space and preventing it from generating a truly novel and superior translation from scratch. It learns to "edit" rather than to "create."

5.3 Overall Performance and Future Outlook

Our best-performing systems for all language pairs were the Gemma3-27B-it models fine-tuned using the direct translation strategy. As shown by our final official scores in Table 3, these systems achieved competitive results. However, a gap remains when compared to the top-ranking teams in the official evaluation.

Our analysis suggests that while full-parameter SFT on a high-quality 100k dataset is effective, it represents only the initial stage of true model alignment. To reach the highest echelons of translation quality, future work should focus on more advanced alignment techniques that have proven successful in general-domain LLMs. Promising directions include:

- **Continual Pre-training:** Further adapting the base LLM on large-scale, in-domain monolingual and bilingual data before the SFT stage.
- **Preference Optimization:** Moving beyond standard SFT to methods like Direct Preference Optimization (DPO) (Rafailov et al., 2024), which learns from human or AI-judged preferences between translation candidates, thereby optimizing directly for perceived quality.

This work confirms that while the era of LLMs is here, achieving state-of-the-art translation performance requires more than just scale; it demands sophisticated and targeted adaptation strategies.

6 Conclusion

In this paper, we presented a systematic comparison between specialized Neural Machine Translation (NMT) systems and general-purpose Large Language Models (LLMs) within the framework of the WMT25 General MT Task. Our work was designed to investigate the ongoing paradigm shift in the field, focusing on the divergence in performance characteristics and the efficacy of different LLM adaptation strategies.

Our investigation yielded clear answers to our initial research questions. **First (RQ1)**, we identified a significant and consistent "BLEU-COMET dichotomy." While our highly optimized NMT systems remained competitive, and occasionally superior, in terms of lexical similarity (BLEU), LLMs demonstrated a marked advantage in semantic fidelity (COMET), even in a zero-shot setting. This finding underscores the limitations of traditional metrics in the age of LLMs and highlights the unique ability of these large models to produce fluent and semantically equivalent translations.

Second (RQ2), our experiments on supervised fine-tuning strategies provided a decisive result: direct translation proved to be a more effective adaptation method than translation refinement. We hypothesize that teaching the model the core source-to-target mapping task directly provides a cleaner and more potent learning signal than asking it to perform the more complex, multi-step task of identifying and correcting errors from a baseline system.

Our best systems, based on full-parameter fine-tuning of the Gemma3-27B-it model, achieved highly competitive results. However, our analysis suggests that the next frontier for LLM-based translation lies beyond standard SFT. The path to state-of-the-art performance will require more sophisticated alignment techniques that can better bridge the gap between the LLMs' vast generative capabilities and the nuanced preferences of human evaluation. Future work should therefore prioritize the exploration of methods such as continual pre-training on in-domain corpora and, most promisingly, preference optimization techniques like DPO.

Acknowledgments

This work also receives substantial support from the 2030 Artificial Intelligence Research Institute of Global Tone Communication Technology Co., Ltd. This paper is also supported by National Key Research and Development Program of China (2022ZD0116100).

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Thamme Gowda, Zhao Zhang, Chris Mattmann, and Jonathan May. 2021. Many-to-English machine translation tools, data, and pretrained models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 306–316, Online. Association for Computational Linguistics.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2023. xcomet: Transparent machine translation evaluation through fine-grained error detection.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

James Cross Onur Çelebi Maha Elbayad Kenneth Heafield Kevin Heffernan Elahe Kalbassi Janice Lam Daniel Licht Jean Maillard Anna Sun Skyler Wang Guillaume Wenzek Al Youngblood Bapi Akula

Loic Barrault Gabriel Mejia Gonzalez Prangthip Hansanti John Hoffman Semarley Jarrett Kaushik Ram Sadagopan Dirk Rowe Shannon Spruit Chau Tran Pierre Andrews Necip Fazil Ayan Shruti Bhosale Sergey Edunov Angela Fan Cynthia Gao Vedanuj Goswami Francisco Guzmán Philipp Koehn Alexandre Mourachko Christophe Ropers Safiyyah Saleem Holger Schwenk Jeff Wang NLLB Team, Marta R. Costa-jussà. 2022. No language left behind: Scaling human-centered machine translation.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

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.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André F. T. Martins. 2023. Scaling up cometkiwi: Unbabel-ist 2023 submission for the quality estimation shared task.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 technical report.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma.

2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.