Towards Improving Translation Ability of Large Language Models on Low Resource Languages

Amulya Ratna Dash and Yashvardhan Sharma
Birla Institute of Technology and Science, Pilani, Rajasthan, India
{*p20200105, yash*}*@pilani.bits-pilani.ac.in*

Keywords: Natural Language Processing, Machine Translation, Low Resource Languages, Large Language Model.

Abstract: With advancements in Natural Language Processing (NLP) and Large Language Models (LLMs), there is a growing need to understand their capabilities with low resource languages. This study focuses on benchmarking and improving the machine translation ability of LLMs for low resource Indic languages. We analyze the impact of training dataset sizes and overfitting by training for additional epochs on translation quality. We use LLaMA-3 as the base model and propose a simple resource efficient model finetuning approach which improves the zero-shot translation performance consistently across eight translation directions.

1 INTRODUCTION

In recent years, Large Language Models are becoming increasingly ubiquitous. LLM such as GPTs (Brown et al., 2020), Gemma (Team et al., 2024), LLaMA (Touvron et al., 2023) and others have shown impressive performance in various NLP tasks like text generation, question answering, summarization, etc. Machine Translation (MT) helps break down language barriers, provides access to global(foreign language) information, and at times simplifies some day-to-day tasks. As compared to other NLP tasks, machine translation involves the translation of text from one language to another, which makes it relatively complex. The accuracy of machine translation is critical as the user may not understand the foreign language and completely rely on the generated text.

Most of the LLMs are pretrained on huge amount of English language data along with very small amount of non-English data (Minaee et al., 2024), and this proportion is aligned with the available digital text content. In comparison to supervised sequence-to-sequence encoder-decoder Neural Machine Translation (NMT) models (Vaswani et al., 2017), the machine translation performance is lower for most of the LLMs other than very large and closed language models like GPTs. The MT performance further degrades for low resource languages. Given the growing popularity of LLMs, it is required to assess the translation ability of LLMs and explore ways to improve the translation performance.

In this paper, we try to understand how well LLMs perform machine translation for low resource Indic languages and explore whether finetuning open LLMs will improve the performance. Our findings show that fine-tuning LLMs with limited parallel data and computational resource can lead to significant improvement in translation ability. We perform an ablation study to evaluate how varying training dataset sizes (low and medium) and the introduction of overfitting through extended training epochs affect the model performance.

2 RELATED WORK

2.1 LLMs for Machine Translation

Previous works have explored the use of LLMs for MT, as these large models demonstrate language understanding skills during pretraining (Liu et al., 2024). In-context learning approach has been studied by Vilar et al., 2023, Agrawal et al., 2023 and Bawden and Yvon, 2023, show the importance of few shot examples, example quality and prompt template for better MT performance. Hendy et al., 2023 have performed a comprehensive evaluation of the translation quality of GPT models and found that the performance for high resource languages is comparable to dedicated translation systems. Singh et al., 2024 benchmarked LLMs on different generation tasks for Indic languages, and reported PaLM-2 (a large closed source LLM) leading in most tasks. Our work ex-

Dash, A. R. and Sharma, Y.
Towards Improving Translation Ability of Large Language Models on Low Resource Languages.
DOI: 10.5220/0013319000003905

In *Proceedings of the 14th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2025)*, pages 801-807
ISBN: 978-989-758-730-6; ISSN: 2184-4313

plores zero shot performance of LLMs for MT of low and very low resource Indic languages(and dialect).

2.2 Finetuning of LLMs

Limited prior work exists for adapting open LLMs to machine translation tasks, with results similar to closed models. Finetuning approach has been studied for relatively high resource languages and shows that MT performance can be improved via finetuning. Iyer et al., 2023 conclude that finetuning improves the ability of LLMs to translate ambiguous sentences. Xu et al., 2023 experiment on German, Czech, Russian, Chinese and Icelandic languages, adapt LLaMA for translation using two steps, first with continued pretraining on monolingual data and then finetuning on parallel data. Alves et al., 2024 adapt LLaMa-2 to multiple tasks in translation workflow, focusing on German, French, Dutch, Italian, Spanish, Portuguese, Korean, Russian and Chinese languages. In our work, we focus on improving the MT performance of LLaMA for low and very low resource Indic languages(and dialect) via single step finetuning on limited parallel data presented as instructions.

3 METHODOLOGY

3.1 Models

To investigate the MT ability of LLMs, we use three foundation language models: GPT 3.5 Turbo[1], Gemma 1.1[2] and LLaMA 3[3]. GPT 3.5 is a closed model, where as Gemma and LLaMA are open models. Gemma 1.1 and LLaMA 3 offer two model sizes each: 2 billion and 7 billion parameters for Gemma, and 2 billion and 8 billion parameters for LLaMA. We use the 7 billion model of Gemma and the 8 billion model of LLaMA, henceforth referred to as Gemma 1.1 7B and LLaMA 3 8B.

3.2 Languages

To evaluate MT performance on low resource languages, we consider 4 Indic languages, i.e. Hindi, Bengali, Odia and Chhattisgarhi. All four languages belong to Indo-Aryan branch of the Indo-European language family.

Hindi is one of the main languages of India along with English, and uses Devanagari script. Hindi is spoken by 345 million speakers around the world.

Bengali (or Bangla) is the second most spoken (237 million speakers) language in India and official language of Bangladesh. It uses Bengali script and is the fifth most spoken native language in the world.

Odia (or Oriya) language is spoken (50 million speakers) predominantly in the Indian state of Odisha, and uses Odia script.

Chhattisgarhi language is spoken (16 million speakers) in the Indian state of Chhattisgarh and parts of neighbouring states. It is considered as a dialect of Hindi, at the same time due to its distinct linguistic features, also considered as a separate language.

3.3 Evaluation Data

FLORES-200 (Costa-jussà et al., 2022) is a human-translated benchmark where the same English sentences are translated into 200 languages. FLORES-200 is split into *dev(977)*, *devtest(1012)* and *test(992)* categories. We use the publicly available *devtest* split having 1012 sentences for evaluating MT performance of LLMs, henceforth referred to as the test dataset.

3.4 Evaluation Metrics

We use Character n-gram F-score(ChrF) (Popović, 2015) and COMET-22 (Rei et al., 2022) scores as evaluation metrics. ChrF evaluates translations at character level rather than word or token level, making it more robust for rich morphological and low resource languages as compared to token-level metrics. Crosslingual Optimized Metric for Evaluation of Translation 2022 (COMET 22) is a neural framework to evaluate translations, which considers context and semantics of the text and has better alignment with human judgement. COMET 22 is based on the XLM-R model which doesn't include Chhattisgarhi language, so we evaluate Chhattisgarhi tasks using ChrF only.

3.5 Finetuning of Model

Based on the baseline MT performance of Gemma and LLaMA models on Hindi and Odia languages, we select LLaMA as the base model for finetuning. We finetune the LLaMA 3 8B model on parallel data presented as instructions using LoRA (Hu et al., 2021) technique. LoRA is a technique that yields better performance, improves sample efficiency, and reduces

[1]http://platform.openai.com/docs/models

[2]https://huggingface.co/google/gemma-1.1-7b-it

[3]https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

Table 1: Composition of training dataset.

Language Pair	Count(E2)	Count(E3)
English - Hindi	1493	24846
Hindi - English	1504	24854
English - Bengali	1498	28724
Bengali - English	1499	28725
English - Odia	1489	18950
Odia - English	1508	18969
English - Chhattisgarhi	504	504
Chhattisgarhi - English	493	493
Total	9988	146065

reserved memory for finetuning LLMs to particular tasks.

4 EXPERIMENTS

4.1 Tasks

We consider eight translation tasks grouped into two categories. First category is Translation of test dataset from English into Hindi, Bengali, Odia and Chhattisgarhi, collectively referred to as $en \rightarrow xx$. The second category is translation of the test dataset from Hindi, Bengali, Odia and Chhattisgarhi into English, collectively referred to as $xx \rightarrow en$.

4.2 Data

We collect our parallel training data from the Bharat Parallel Corpus Collection (BPCC Gala et al., 2023) and FLORES-200 dataset. BPCC-H Wiki and BPCC-H Daily datasets which are manually translated multi-domain Indic parallel corpus are used for Hindi, Bengali, and Odia languages. The *dev* split of FLORES-200 dataset is used for Hindi, Bengali, Odia and Chhattisgarhi languages. The above individual datasets are split into both translation directions for uniformity across all 8 translation tasks. Data processing includes formatting the data as per LLaMA 3 chat template. The composition of the processed dataset is available in Table 1. The processed dataset is split into *train(0.9)* and *validation(0.1)* dataset. The template used for training data is available in Figure 1. A sample record from processed dataset is available in Figure 2.

4.3 Experiment Design

Experiment 1 (E1): Evaluate GPT3.5, Gemma1.1 7B and LLaMA3 8B models on the translation tasks.
Experiment 2 (E2): Finetune LLM on 10K parallel data sampled from the processed dataset.

Table 2: Hyperparameters and Training size.

Parameters	Values
Learning rate	0.0003
Train batch size	8
Eval batch size	4
Seed	3407
Gradient accumulation	8
Total train batch size	64
Optimizer	Adam(Kingma, 2014)
Adam β_1	0.9
Adam β_2	0.999
Adam ε	1e-08
LR scheduler type	cosine
LR warmup steps	0.01
Num of Epochs	4
Num of GPU	1 (A100 40GB)
Train examples	E2: 8989 E3: 131458

Experiment 3 (E3): Finetune LLM on the full(100K+) dataset.

Both the experiments E2 and E3, include the step to evaluate the finetuned model on the translation tasks.

4.4 Training Setup

The model was finetuned in a multilingual many-to-many approach, as is evident from the composition of the training dataset. Among the open models, we shortlisted LLaMA 3 8B for finetuning experiments **E2** and **E3** based on the outcome of **E1** experiment (baseline). The hyperparameters and data size used for experiments **E2** and **E3** are detailed in Table 2.

5 RESULTS

5.1 Baseline Experiment Results (E1)

In comparison to Gemma 1.1 model, GPT 3.5 and LLaMa 3 perform better in the eight translation tasks. Gemma 1.1 model performed poorly for the Odia translation tasks. The scores for $xx \rightarrow en$ tasks are better as compared to $en \rightarrow xx$, as all three LLMs are pretrained to generate fluent and coherent English text. GPT 3.5 model performs better for Hindi tasks and Hindi's dialect Chhattisgarhi tasks. The detailed scores from experiment **E1** are available in Table 3.

Role	Instruction and Response Template
Human	Don't provide any justification or extra output, just the translated text. Translate the following sentence from <*Source language*> to <*Target language*>: <*Text in Source language*>
Assistant	<*Text in Target language*>

Figure 1: Template of training dataset.

Role	Text
Human	Don't provide any justification or extra output, just the translated text. Translate the following sentence from English to Odia: what's the directions to the nearest pizza hut<|eot_id|>
Assistant	ନିକଟତମ 'ପିଜ୍ଜା କୁଟୀରକୁ ଯିବା ପାଇଁ ଦିଗ ସୂଚନା କ'ଣ?<|eot_id|>

Figure 2: Sample training record (and Prompt) for *English→Odia* language.

Table 3: Translation performance for GPT 3.5, Gemma 1.1, and LLaMA 3 **(E1)**.

	GPT 3.5		**Gemma 1.1**		**LLaMA 3**	
Language Pair	ChrF	COMET	ChrF	COMET	ChrF	COMET
English - Hindi	**47.80**	**77.78**	40.08	74.51	43.77	76.57
English - Bengali	39.36	81.73	36.21	81.31	**39.94**	**83.20**
English - Odia	27.94	72.47	2.07	39.18	**31.35**	**74.94**
English - Chhattisgarhi	**35.79**	-	31.43	-	32.75	-
Hindi - English	**60.37**	**87.94**	55.95	86.21	60.10	87.86
Bengali - English	53.28	85.88	49.36	83.91	**54.72**	**86.28**
Odia - English	45.81	82.27	17.89	51.25	**47.30**	**82.92**
Chhattisgarhi - English	**52.34**	-	43.80	-	51.36	-

5.2 Finetuning Experiment Results (E2 and E3)

The objective of experiment **E2** was to determine the possibility of improving the translation performance of LLaMa 3 by finetuning. Based on Validation loss, we select *Epoch 2* checkpoint as the generalized model, and consider the same as the final model of experiment **E2**. Using **10K** parallel data and **1 hour** of single GPU training time, we could see a significant improvement in zero-shot translation performance of the finetuned model. The average ChrF score increased by **+2.42** for *xx→en* and **+6.14** for *en→xx* over vanilla LLaMA 3 model. The detailed scores from experiment **E2** are available in Table 4 and Table 5.

As part of experiment **E3**, with around 12 hours of single GPU training time and using **140K** parallel data spread across all 8 language pairs, we could see further improvement in zero-shot translation performance as compared to vanilla LLaMA 3 model. To achieve the best generalization performance, we selected the checkpoint from *Epoch 2* based on the validation loss. This checkpoint serves as our final finetuned model. The average ChrF score increased by **+3.93** for *xx→en* and **+11.37** for *en→xx* over vanilla

Table 4: Translation performance on *en* → *xx* for model trained with 10K examples **(E2)**.

Language Pair	**ChrF**	**COMET**
English - Hindi	48.07	77.64
English - Bengali	43.74	83.34
English - Odia	38.21	77.22
English - Chhattisgarhi	42.36	-
Average	43.09	79.40

Table 5: Translation performance on *xx* → *en* for model trained with 10K examples **(E2)**.

Language Pair	**ChrF**	**COMET**
Hindi - English	59.71	87.98
Bengali - English	54.68	86.72
Odia - English	50.80	85.20
Chhattisgarhi - English	57.98	-
Average	55.79	86.63

LLaMA 3 model. The detailed scores from experiment **E3** are available in Table 6 and Table 7.

Table 6: Translation performance on *en* → *xx* for model trained with full dataset **(E3)**.

Language Pair	ChrF	COMET
English - Hindi	51.94	78.91
English - Bengali	48.75	85.15
English - Odia	46.58	81.82
English - Chhattisgarhi	46.02	-
Average	48.32	81.96

Table 7: Translation performance on *xx* → *en* for model trained with full dataset **(E3)**.

Language Pair	ChrF	COMET
Hindi - English	61.06	88.03
Bengali - English	55.52	86.33
Odia - English	54.15	86.37
Chhattisgarhi - English	58.46	-
Average	57.30	86.91

6 ANALYSIS

As evident in Figure 3 and 4, the translation performance of our finetuned model *'IndicMT-Llama-3-8B'* is significantly better than 8B vanilla LLaMa 3 and extremely large(more than 10x) closed model GPT 3.5 for both *en*→*xx* and *xx*→*en* translation directions.

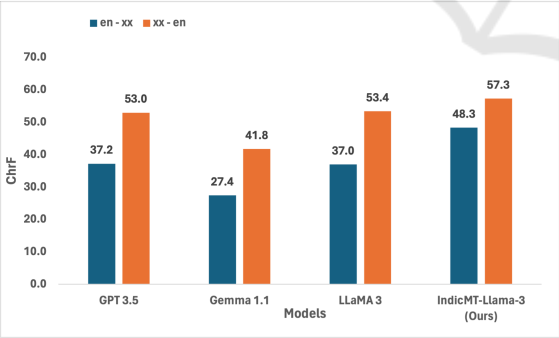

Figure 3: Average zero-shot translation performance *(ChrF)* on FLORES-200 for *English* ↔ {*Hindi, Bengali, Odia, Chhattisgarhi*}.

6.1 Impact of Additional Parallel Training Data

The final model trained with additional 130K parallel training data had a roughly 2x improvement in ChrF score. The improvement for *en*→*xx* was higher as compared to *xx*→*en*. The additional training data

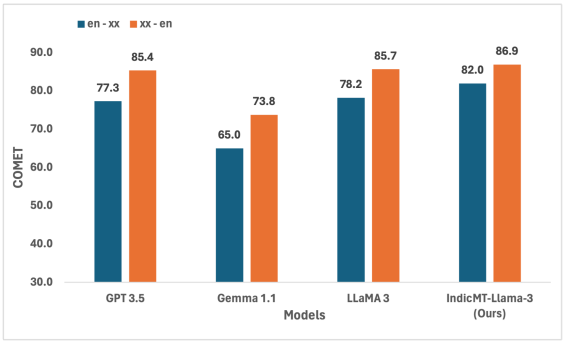

Figure 4: Average zero-shot translation performance *(COMET)* on FLORES-200 for *English* ↔ {*Hindi, Bengali, Odia, Chhattisgarhi*}.

helped the model better learn the Odia and Bengali translation tasks, as seen by an improvement of up to **+8.37** ChrF score.

6.2 Impact of Increasing the Number of Epochs

We analyzed the checkpoints of epoch 2 and epoch 4 for the model finetuned with full training dataset(140K). The ChrF score increased by **+0.28** for *en*→*xx*, but decreased by **-5.37** for *xx*→*en* translation direction. This mixed performance suggests that the model overfits the training dataset when finetuned beyond epoch 2, forgetting its English text generation ability.

6.3 Observations from Human Evaluation

- Over-generation and repetition of words were observed more when translating from English.
- Omission of words and generation of words from the source (non-English) language appeared in translated text in Roman script when translating to English.
- The overall fluency of translated text was better in the fine-tuned model compared to the base model for both directions(to and from English).
- Table 8 shows few random translations from base and finetuned model.

7 CONCLUSION AND FUTURE WORK

As the field of NLP and LLMs progresses rapidly, it is important to explore and improve their perfor-

Table 8: Sample translation of FLORES-200 evaluation dataset using Base (B) and Finetuned (F) model.

English (en)	Boating is a national pastime in Finland, with a boat to every seven or eight people.
B: Hindi → en	Rowing is a national pastime in Finland, one boat for every seven or eight people.
F: Hindi → en	Canoeing is a national pastime in Finland, with a boat accommodating one, seven or eight people.
B: Bengali → en	A boat ride with seven or eight people is a national pastime in Finland.
F: Bengali → en	Boat races with seven or eight persons in a boat is a national pastime in Finland.
B: Odia → en	It takes about 7 or 8 people to write a letter to Finland with a ferry.
F: Odia → en	In Finland, canoeing is a national pastime, with a canoe for every 7 or 8 people.
B: Chhattisgarhi → en	Finnland me donga me ghuma na ekthan rashtriya shagel ho, ek donga ke sang har saat ya aath mankhem bar donga haway.
F: Chhattisgarhi → en	In Finland, going by canoe is a national pastime, a canoe being the mode of transport for one or two persons in a dongo.

mance on low resource languages. To address this gap, this paper evaluates how well LLMs perform machine translation for low resource Indic languages and explores ways to improve the translation ability of LLMs.

Our study concludes that open LLMs can be finetuned to improve their multilingual translation ability with very limited per language parallel data and minimal computational resources. Our model *'IndicMT-Llama-3-8B'* achieves superior translation performance for four Indic languages, compared to open LLMs, in both English-to-Indic and Indic-to-English tasks. The methodology and findings can be generalized to other languages in improving the translation performance of any LLM.

We have not included recent proprietary closed models like GPT 4 (Achiam et al., 2023) and Gemini (Team et al., 2023) in our study due to exceptionally large model size as compared to less than 10B models like Gemma 1.1 7B & LLaMA 3 8B. The general translation ability of our model beyond the four Indic languages should have improved, but we have not validated.

Future work will evaluate the general translation ability of our finetuned model, and extend the study to additional languages(Indic and other language families) and evaluation datasets. Research on methods to further improve *xx→en* translation performance will be helpful.

REFERENCES

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Agrawal, S., Zhou, C., Lewis, M., Zettlemoyer, L., and Ghazvininejad, M. (2023). In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873.

Alves, D. M., Pombal, J., Guerreiro, N. M., Martins, P. H., Alves, J., Farajian, A., Peters, B., Rei, R., Fernandes, P., Agrawal, S., et al. (2024). Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

Bawden, R. and Yvon, F. (2023). Investigating the translation performance of a large multilingual language model: the case of bloom. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 157–170.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., et al. (2022). No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Gala, J., Chitale, P. A., AK, R., Gumma, V., Doddapaneni, S., Kumar, A., Nawale, J., Sujatha, A., Puduppully, R., Raghavan, V., et al. (2023). Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv:2305.16307*.

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., and Awadalla, H. H. (2023). How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Iyer, V., Chen, P., and Birch, A. (2023). Towards effective disambiguation for machine translation with large language models. In *Proceedings of the Eighth Conference on Machine Translation*, pages 482–495.

Kingma, D. P. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Liu, Y., He, H., Han, T., Zhang, X., Liu, M., Tian, J., Zhang,

Y., Wang, J., Gao, X., Zhong, T., et al. (2024). Understanding llms: A comprehensive overview from training to inference. *arXiv preprint arXiv:2401.02038*.

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., and Gao, J. (2024). Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Popović, M. (2015). chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Rei, R., De Souza, J. G., Alves, D., Zerva, C., Farinha, A. C., Glushkova, T., Lavie, A., Coheur, L., and Martins, A. F. (2022). Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

Singh, H., Gupta, N., Bharadwaj, S., Tewari, D., and Talukdar, P. (2024). IndicGenBench: A multilingual benchmark to evaluate generation capabilities of LLMs on Indic languages. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11047–11073, Bangkok, Thailand. Association for Computational Linguistics.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., Tafti, P., Hussenot, L., Sessa, P. G., Chowdhery, A., Roberts, A., Barua, A., Botev, A., Castro-Ros, A., Slone, A., Héliou, A., Tacchetti, A., Bulanova, A., Paterson, A., Tsai, B., Shahriari, B., Lan, C. L., Choquette-Choo, C. A., Crepy, C., Cer, D., Ippolito, D., Reid, D., Buchatskaya, E., Ni, E., Noland, E., Yan, G., Tucker, G., Muraru, G.-C., Rozhdestvenskiy, G., Michalewski, H., Tenney, I., Grishchenko, I., Austin, J., Keeling, J., Labanowski, J., Lespiau, J.-B., Stanway, J., Brennan, J., Chen, J., Ferret, J., Chiu, J., Mao-Jones, J., Lee, K., Yu, K., Millican, K., Sjoesund, L. L., Lee, L., Dixon, L., Reid, M., Mikuła, M., Wirth, M., Sharman, M., Chinaev, N., Thain, N., Bachem, O., Chang, O., Wahltinez, O., Bailey, P., Michel, P., Yotov, P., Chaabouni, R., Comanescu, R., Jana, R., Anil, R., McIlroy, R., Liu, R., Mullins, R., Smith, S. L., Borgeaud, S., Girgin, S., Douglas, S., Pandya, S., Shakeri, S., De, S., Klimenko, T., Hennigan, T., Feinberg, V., Stokowiec, W., hui Chen, Y., Ahmed, Z., Gong, Z., Warkentin, T., Peran, L., Giang, M., Farabet, C., Vinyals, O., Dean, J., Kavukcuoglu, K., Hassabis, D., Ghahramani, Z., Eck, D., Barral, J., Pereira, F., Collins, E., Joulin, A., Fiedel, N., Senter, E., Andreev, A., and Kenealy, K. (2024). Gemma: Open models based on gemini research and technology.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Vilar, D., Freitag, M., Cherry, C., Luo, J., Ratnakar, V., and Foster, G. (2023). Prompting palm for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427.

Xu, H., Kim, Y. J., Sharaf, A., and Awadalla, H. H. (2023). A paradigm shift in machine translation: Boosting translation performance of large language models. *arXiv preprint arXiv:2309.11674*.