IN-CONTEXT LEARNING AND PIVOT-BASED TRANSLATION COMPARISON FOR LOW-RESOURCE MACHINE TRANSLATION

Andre Hutagaol, Tyra Silaphet, Yen-Shuo Su

Georgia Institute of Technology

{ahutagaol3, tsilaphet3, ericsu}@gatech.edu

1. OBJECTIVES

1.1. Motivation

Neural machine translation (NMT) has achieved state-of-the-art (SOTA) performance for high-resource languages, where large parallel corpora enable robust model training, but translation quality declines significantly for low-resource languages, which lack sufficient bilingual data [1]. Recent advances in multilingual machine translation (MMT), such as Meta AI's No Language Left Behind (NLLB) model, have expanded support to over 200 languages by leveraging large-scale multilingual training [2]. Despite this progress, NLLB's performance remains inconsistent, particularly for language pairs with scarce or no parallel data.

Among the most challenging cases are language pairs that are both typologically distant and written in different scripts, such as Nepali–Sinhala. These languages not only lack direct training data but also use distinct orthographies (Devanagari vs. Sinhala), compounding translation difficulty. Prior benchmarks such as FLoRes [3] have highlighted the limitations of even the best models on related tasks like Nepali–English and Sinhala–English, reinforcing the need for alternative strategies.

To address this broader challenge, two strategies have recently gained attention:

- **Pivot-based translation**: a widely used technique where a low-resource language is translated into a high-resource intermediary language (e.g., English), before being translated into the final target language. This strategy has been shown to improve low-resource MT by leveraging richer bilingual data via the pivot [4, 5].
- **In-Context Learning (ICL)**: large language models (LLMs) like GPT-4 can translate by conditioning on a few example pairs, without task-specific fine-tuning. Recent studies show that ICL can match or outperform traditional supervised methods under certain conditions [6, 7].

Motivated by these developments, our project explores how well these techniques perform in a truly low-resource setting, specifically Nepali–Sinhala, compared to direct multilingual translation using NLLB.

1.2. Task Definition

1.2.1. Problem Statement

Despite progress in multilingual NMT, translation between **two low-resource languages** remains an underexplored and difficult problem. Much of the existing literature focuses on translating between a low-resource language and a high-resource one (like English), where pretrained models and larger corpora are available. However, in many real-world contexts, communication occurs between low-resource languages with little or no shared training data.

In this project, we investigate the Nepali–Sinhala and Sinhala–Nepali directions. These languages are linguistically distinct, use different scripts, and lack direct parallel corpora. Even indirect corpora, such as Nepali–English and Sinhala–English, are limited in both domain and scale, as shown by the FLoRes benchmark [3]. This makes Nepali–Sinhala a uniquely suitable testbed for evaluating translation methods under truly low-resource conditions.

To this end, we compare three major strategies:

- **Direct translation** using the multilingual NLLB model.
- **Pivot-based translation**, where English serves as an intermediate language.
- **In-context learning (ICL)** using GPT-4, leveraging few-shot translation examples at inference time.

Our key contribution is a comparative evaluation of these strategies—direct, pivot-based, and in-context—for translation between two low-resource languages, without assuming access to large fine-tuning corpora. While prior work has explored ICL or pivoting independently, few have directly compared these approaches for true low-resource pairs like Nepali–Sinhala.

1.2.2. Datasets

We used two main datasets: **FLoRes-200** and **WMT21**. FLoRes-200 is a multilingual evaluation benchmark created by Meta AI for testing machine translation across over 200 languages [2]. It includes both dev and devtest splits, each formatted as plain text files containing aligned sentence pairs across language combinations. We processed FLoRes-200 into .jsonl format using a custom Python script, with each line containing a "src" (source sentence) and "tgt" (target translation). In total, the dev splits (e.g., ne_sin_dev.jsonl, sin_ne_dev.jsonl) contain 997 sentence pairs, while the devtest splits (e.g., ne_sin_devtest.jsonl, sin_ne_devtest.jsonl) contain 1012 sentence pairs. This small discrepancy is due to slight differences in original file lengths.

In addition, we used the WMT21 WikiTitles v3 dataset to support our pivot-based translation approach. This dataset includes English–Russian parallel data, which we converted into JSONL format using a tab-separated value (TSV) parser [8] . While we didn't fine-tune models on this dataset, it was used to simulate intermediate pivot steps (e.g., English → Russian or English → Sinhala) for enriching our translation evaluation.

2. RELATED WORK

Multilingual Machine Translation models leverage cross-lingual transfer to improve low-resource translation. Meta AI's NLLB [2] supports 200 languages with direct translation, reducing reliance on English as a pivot. However, performance varies depending on

data availability and linguistic similarity. The impact of pivot-based translation on NLLB's effectiveness remains an open question.

Pivot-based translation improves translation quality by first translating a low-resource language into a high-resource intermediary before translating to the final target language. This approach has been effective for low-resource pairs but suffers from error propagation, where mistakes in the first step may affect the second.

An alternative strategy is in-context learning (ICL), where LLMs translate by leveraging a small number of sample translations in the input prompt rather than undergoing fine-tuning [9]. Chen et al. (2023) demonstrated that ICL can match or outperform fine-tuned NMT models when example selection is optimized, especially for low-resource pairs [6]. However, their study did not compare ICL to pivot-based translation, leaving open questions about its relative effectiveness.

`TheBloke/LLaMA-2-7B-Chat-GPTQ` is a quantized variant of Meta's LLaMA-2-7B [10] language model, optimized for efficient inference with reduced memory and compute requirements. While LLaMA-2 is not explicitly trained for translation, its general-purpose instruction-following capabilities allow it to perform translation tasks via prompt-based in-context learning. Prior studies have shown that even without task-specific fine-tuning, LLMs can learn translation mappings when provided with high-quality prompts and a few demonstration examples. In our study, we evaluate how well this quantized LLaMA-2 model performs Nepali–Sinhala translation under both zero-shot and few-shot conditions, and compare its performance against fine-tuned multilingual models like NLLB.

Evaluating translation quality for low-resource languages like Nepali and Sinhala requires metrics that can handle both surface-level similarity and deeper semantic meaning. Goyle et al. (2023) evaluated neural machine translation on Nepali–English and Sinhala–English pairs using BLEU scores, which remain a standard in the field. In our work, we use SacreBLEU [11], a standardized implementation of BLEU [12], to ensure consistency and reproducibility in tokenization and scoring across experiments. However, BLEU and SacreBLEU rely on exact n-gram overlap, which can underestimate quality in morphologically rich languages like Nepali and Sinhala. To better capture variation and inflection, we also use ChrF [13], which operates at the character level and is better suited for such languages. Finally, we include BLEURT [14], a learned evaluation metric that leverages contextual embeddings from BERT and has been shown to better correlate with human judgments, particularly in low-resource and domain-mismatched settings.

3. METHODS

3.1. Overview of Approaches

We evaluate four approaches for translating between Nepali and Sinhala, two low-resource languages with limited parallel data. Each method represents a different strategy for handling data scarcity, and all are compared using BLEURT, SacreBLEU, and ChrF metrics.

Figure 1 summarizes the four translation approaches explored in this study.

- **Direct translation**: Serves as our baseline. We use the `facebook/nllb-200-distilled-600M` model to perform direct translation between Nepali and Sinhala without involving any intermediate language or task-specific fine-tuning.
- **Pivot translation (no fine-tuning)**: This method translates from Nepali to English and then from English to Sinhala, using NLLB in a chained, zero-shot setup. No additional supervision is applied in either step.

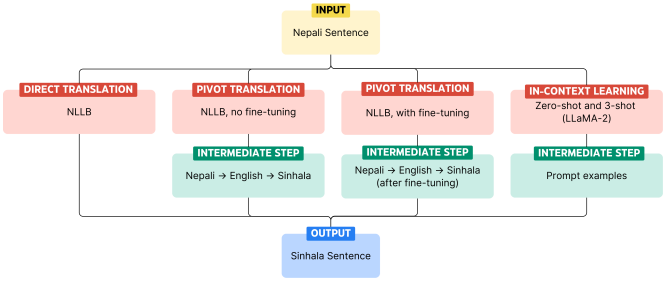

Fig. 1. Overview of the four translation strategies evaluated in this study. All methods take a Nepali sentence as input and produce a Sinhala translation as output, either directly or via intermediate steps.

- **Pivot translation (with fine-tuning)**: We extend the above method by fine-tuning the NLLB model on related language pairs (e.g., English–Russian, English–Sinhala) from the WMT21 WikiTitles v3 dataset. This fine-tuning aims to improve the model's intermediate representations during the pivot step.
- **In-context learning (ICL)**: Using the `LLaMA-2` model, we experiment with zero-shot and 3-shot prompting. In the 3-shot setup, we provide the model with a few Nepali–Sinhala translation pairs as in-context examples and evaluate its ability to generalize.

3.2. Pre-trained Models

We utilized two pre-trained models to enhance our machine translation system:

1. **No Language Left Behind (NLLB)** model is a state-of-the-art machine translation model. It employs a Transformer-based encoder-decoder architecture with 54.5 billion parameters. Key architectural features include:
    - Sparsely gated mixture of experts (MoE): Each layer contains multiple experts, but only a subset is activated per input, enhancing model capacity without a proportional increase in computational cost. [15]
    - Hierarchical gating mechanism: A two-level gating system first decides whether to use shared feed-forward layers or activate experts. If experts are chosen, a second gating mechanism selects the top experts to process the input.
    - Dense transformer variants: Utilizes dense Transformer architectures for tasks for less computational overhead. [15]
2. **LLaMA-2** employs a decoder-only Transformer architecture optimized for autoregressive language modeling. Key architectural components include:
    - Embedding layer: Converts input tokens into dense vectors of fixed size, facilitating efficient processing.
    - Decoder blocks: Each block consists of multi-head self-attention mechanisms and feed-forward neural networks, enabling the model to capture complex dependencies in the input data.
    - Activation function: Utilizes the SwiGLU activation function, which has been shown to improve model performance over traditional activation functions like ReLU.

- Normalization: Employs RMSNorm (Root Mean Square Layer Normalization), a variant of layer normalization that has been found to stabilize training and improve convergence.
- Positional encoding: Incorporates Rotary Positional Embeddings (RoPE), allowing the model to better capture the relative positions of tokens in the input sequence.
- Output layer: A linear layer followed by a softmax function generates the probability distribution over the vocabulary for the next token prediction.

3.3. Experimental Setup

We apply all four methods to two evaluation sets from FLoRes-200: the dev and devtest splits, covering both directions (Nepali→Sinhala and Sinhala→Nepali). For the pivot methods, we always use English as the intermediate language. For in-context learning, prompts were manually constructed using a small sample from the FLoRes-200 dev set.

Our main contribution is a unified, head-to-head comparison of direct, pivot, and in-context translation strategies on a truly low-resource pair. Unlike prior work that focuses on high-resource to low-resource translation, we center our evaluation on a challenging bidirectional low-resource setting.

3.4. Evaluation Metrics

We used three evaluation metrics, BLEURT, SacreBLEU, and ChrF to assess the performance of our machine translation system. Below is an explanation of each metric:

- **BLEURT (Bilingual Evaluation Understudy with Representations from Transformers)**: BLEURT is a learned evaluation metric that leverages pre-trained models like BERT and RemBERT. BLEURT scores are ranging from 0 to 1. It is trained using transfer learning on human-annotated data, enabling it to assess translation quality by considering fluency, adequacy, and semantic similarity. Higher scores indicate translations that are more fluent and semantically similar to human references. A BLEURT score of 100 represents a translation indistinguishable from human-written text, while a score closer to 0 suggests significant deviations from human quality. [16]
- **SacreBLEU**: SacreBLEU is a standardized implementation of the BLEU metric. It automates the computation of BLEU scores, handling tokenization and other preprocessing steps consistently. SacreBLEU produces a score between 0 and 100, reflecting the precision of n-grams (up to 4-grams) in the machine-generated translation compared to reference translations. Higher scores indicate better overlap with reference n-grams. However, BLEU's reliance on surface-level n-gram matching means it may not fully capture semantic accuracy or fluency. [17]
- **ChrF**: ChrF is a character-level evaluation metric that calculates the F-score for character n-grams. ChrF scores range from 0 to 100, with higher scores indicating better overlap of character n-grams between the machine-generated and ground truths. These metrics are more sensitive to morphological variations and can provide a more granular assessment of translation quality compared to word-based metrics. [18]

4. EVALUATION

We evaluate translation performance between Nepali and Sinhala across four datasets using two distinct model families: the multilingual supervised model facebook/nllb-200-distilled-600M, and the general-purpose large language model TheBloke/LLaMA-2-7B-Chat-GPTQ. Metrics reported include BLEURT, SacreBLEU, and ChrF2. Tables 1 and 2 summarize the results.

4.1. Supervised and Pivoted Translation (NLLB)

Table 1 presents performance of the NLLB model under three scenarios: (i) direct translation between Nepali and Sinhala, (ii) pivot translation via English without additional fine-tuning, and (iii) pivot translation with supervised fine-tuning on synthesized data. In all cases, translation is evaluated bidirectionally using both dev and devtest sets.

As anticipated, direct translation performs reasonably well, particularly given that Nepali and Sinhala are among the lower-resource language pairs in the NLLB training. BLEURT scores in the range of 0.42–0.44 and ChrF2 around 41–43 indicate moderate semantic alignment.

Pivoting without fine-tuning slightly underperforms direct translation, suggesting that simple chaining through English introduces compounding errors—likely due to intermediate representation drift or inadequate handling of typologically diverse structures. However, once fine-tuning is applied to the pivoted pairs (e.g., Nepali → English → Sinhala), we observe consistent and meaningful gains across all metrics. For instance, on `ne_sin_dev`, BLEURT rises from 0.4267 (direct) to 0.4307 (pivot w/ fine-tuning), and ChrF2 improves from 40.66 to 42.17. Similarly, in the reverse direction (`sin_ne_dev`), BLEURT improves from 0.0924 to 0.1124, and ChrF2 from 42.26 to 45.55.

These results align with expectations: while direct modeling of low-resource pairs can be effective when supported by multilingual pretraining, pivoting—when augmented with task-specific fine-tuning—can leverage richer supervision from high-resource pivots to boost performance. This is particularly relevant for asymmetrically resourced languages, where paired data is sparse but abundant English-centric translations exist.

4.2. In-context Learning with LLaMA-2

Table 2 examines the ability of the LLaMA-2-7B-Chat-GPTQ model to perform zero-shot and 3-shot in-context learning (ICL) on the same datasets. Unlike NLLB, this model has not been explicitly fine-tuned for translation or exposed to large amounts of Nepali or Sinhala text. Nevertheless, it demonstrates a basic ability to align meaning in these languages when guided with appropriate prompts.

Zero-shot performance is weak in absolute terms, with BLEURT scores around -1.0 to -1.4 and negligible BLEU/chrF2 scores, which is expected given the lack of translation supervision and limited tokenizer coverage for low-resource scripts. However, when three examples are provided as context (3-shot ICL), the model shows substantial gains in BLEURT and especially chrF2. For example, on `ne_sin_dev`, BLEURT improves from -1.394 to -1.212, and ChrF2 jumps from 0.7 to 12.1—indicating that the model can adapt to translation tasks when exposed to structural examples.

These improvements confirm a well-documented behavior of LLMs: they can approximate task-specific behavior through prompt engineering and few-shot learning, even for tasks not explicitly seen during training. However, the consistently low BLEU and BLEURT scores compared to the supervised NLLB results suggest that LLaMA-2 still lacks the robust bilingual alignment capabilities that come from dedicated translation objectives.

Dataset	Direct Translation			Pivot (No FT)			Pivot (with FT)		
	BLEURT	SacreBLEU	ChrF2	BLEURT	SacreBLEU	ChrF2	BLEURT	SacreBLEU	ChrF2
ne_sin_dev	0.4267	8.92	40.66	0.3894	8.62	39.56	**0.4307**	**9.73**	**42.17**
ne_sin_devtest	0.4369	**9.47**	41.07	0.4092	8.87	40.03	**0.4371**	9.14	**41.88**
sin_ne_dev	0.0924	8.44	42.26	0.0877	9.01	43.94	**0.1124**	**9.71**	**45.55**
sin_ne_devtest	0.1012	8.43	43.42	0.0860	**9.70**	43.20	**0.1038**	9.54	**44.47**

Table 1. Translation performance metrics across four Nepali–Sinhala datasets using Direct, Pivot (no fine-tuning), and Pivot (with fine-tuning) on `facebook/nllb-200-distilled-600M`

Dataset	Zero-shot			3-shot ICL		
	BLEURT	SacreBLEU	ChrF2	BLEURT	SacreBLEU	ChrF2
ne_sin_dev	-1.394	0.1	0.7	**-1.212**	**0.2**	**12.1**
ne_sin_devtest	-1.398	**0.3**	0.6	**-1.048**	0.2	**12.1**
sin_ne_dev	-1.011	0.1	0.6	**-0.277**	0.1	**11.3**
sin_ne_devtest	-1.017	**0.2**	0.6	**-0.264**	0.1	**11.9**

Table 2. Evaluation results on Nepali–Sinhala translation using `TheBloke/LLaMA-2-7B-Chat-GPTQ` with Zero-shot and 3-shot In-context Learning (ICL)

4.3. Discussion and Implications

Overall, our results highlight the complementary strengths of multilingual supervised models and large language models. While models like NLLB excel when paired data or synthetic fine-tuning is available, general-purpose LLMs like LLaMA-2 offer flexible adaptation via ICL, though currently at lower performance levels. The improvements observed through fine-tuned pivoting underscore the value of intermediate supervision, particularly when direct pairs are scarce.

We also observe consistent asymmetry in translation performance between Nepali → Sinhala and Sinhala → Nepali. This directional difference likely stems from a combination of factors. First, the training data distribution for both the supervised NLLB model and the underlying corpora used for LLM pretraining may be skewed, with Nepali–English pairs being more abundant or better aligned than Sinhala–English. Second, script complexity and tokenization issues can disproportionately affect one direction—Sinhala's character set may be more challenging for LLMs to segment and represent effectively compared to Nepali. Finally, linguistic asymmetry plays a role: translating from a morphologically richer or syntactically looser source (e.g., Sinhala) to a more rigid target (e.g., Nepali) can impose greater decoding burden.

Moreover, the gains seen from in-context learning—even in low-resource scenarios—point to the potential for scalable prompt-based translation in settings where fine-tuning is infeasible. Future work may combine both strategies: leveraging LLMs for data augmentation and control, while grounding translations with robust multilingual encoders such as NLLB models.

5. CONCLUSION

We have explored three distinct strategies for machine translation between two low-resource languages, Nepali and Sinhala. We compared direct translation using the multilingual NLLB model, pivot-based translation using English as an intermediary language, and in-context learning with a large language model (LLaMA-2) that leverages few-shot examples.

Our results demonstrated that while direct translation using NLLB was successful for the low-resource language translation task, pivot-based translation, when fine-tuned with related language pairs, led to significant improvements in translation quality. These findings suggest that intermediate supervision via pivoting is a valuable tool, especially in low-resource scenarios. On the other hand, ICL via LLaMA-2, although showing some promise with 3-shot prompting, still lagged behind the more structured NLLB models in terms of performance metrics like BLEURT, SacreBLEU, and ChrF2.

Our study opens several directions for future research in low-resource machine translation. First, while we demonstrated the benefits of fine-tuned pivoting with NLLB, more systematic exploration of pivot language selection could further improve performance. Second, our in-context learning (ICL) results suggest room for enhancement through improved prompt design or retrieval-based example selection. Future work may also investigate combining LLM-based ICL with multilingual encoders like NLLB in a hybrid pipeline, where LLMs generate synthetic training data or refine outputs from supervised models. Additionally, evaluating translation quality with human judgments could offer more nuanced insights into real-world applicability. Finally, extending our analysis to other low-resource language pairs will help validate the generalizability of these strategies across different low-resource languages.

6. TEAM CONTRIBUTION

Andre Hutagaol conducted exploratory data analysis to identify which languages could be categorized as low-resource. This analysis led to the conclusion that Nepali and Sinhala are low-resource languages for the project. Additionally, Andre performed pivot-based translation experiments without fine-tuning using the NLLB model.

Tyra Silaphet proposed the project idea and led the data preparation pipeline. This included downloading, cleaning, and processing the FLoRes-200 and WMT21 datasets to prepare them for use with the NLLB model. Tyra also conducted the direct translation experiments using the NLLB model, evaluating its performance on low-resource language pairs such as Nepalese to Sinhala.

Yen-Shuo Su designed and implemented the pivot translation experiments with fine-tuning using the NLLB model. Yen-Shuo also conducted the in-context learning (ICL) evaluations using LLaMA-2 and performed the comparative analysis across methods. In addition, he contributed to the overall experimental setup and result interpretation.

7. REFERENCES

[1] Madhavendra Thakur, "Towards neural no-resource language translation: A comparative evaluation of approaches," 12 2024.

[2] NLLB Team, Marta R. Costa-jussà, 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, and Jeff Wang, "No language left behind: Scaling human-centered machine translation," 2022.

[3] Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato, "The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, Eds., Hong Kong, China, Nov. 2019, pp. 6098–6111, Association for Computational Linguistics.

[4] Michael Paul and Matt Post, "Pivot based smt and strategies for phrase table triangulation," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 2013, pp. 518–523.

[5] Youngjung Kim, Yongchan Park, Hyunjeong Lim, and Key-Sun Choi, "Pivot-based transfer learning for neural machine translation between non-english languages," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019, pp. 124–135.

[6] Yufeng Chen, "Enhancing machine translation through advanced in-context learning: A methodological strategy for gpt-4 improvement," 2023.

[7] William Robinson, Ehsan Nouri, Bowen Zhou, Tom Kocmi, Christian Federmann, Yinhan Tang, Aman Madaan, Lijuan Wu, Angela Fan, and Philipp Koehn, "Mt-bench: How strong is chatgpt-4 in non-english languages?," 2023.

[8] Mateusz Krubiński, Erfan Ghadery, Marie-Francine Moens, and Pavel Pecina, "MTEQA at WMT21 metrics shared task," in *Proceedings of the Sixth Conference on Machine Translation*, Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, Eds., Online, Nov. 2021, pp. 1024–1029, Association for Computational Linguistics.

[9] 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, "Language models are few-shot learners," 2020.

[10] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al., "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[11] Matt Post, "A call for clarity in reporting bleu scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers*, 2018, pp. 186–191.

[12] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[13] Maja Popović, "chrf: character n-gram f-score for automatic mt evaluation," in *Proceedings of the tenth workshop on statistical machine translation*, 2015, pp. 392–395.

[14] Thibault Sellam, Dipanjan Das, and Ankur P Parikh, "Bleurt: Learning robust metrics for text generation," *arXiv preprint arXiv:2004.04696*, 2020.

[15] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," 2017.

[16] Fei Yan, Youngkyu Sung, Philip Krantz, Archana Kamal, David K. Kim, Jonilyn L. Yoder, Terry P. Orlando, Simon Gustavsson, and William D. Oliver, "Engineering framework for optimizing superconducting qubit designs," 2020.

[17] Matt Post, "A call for clarity in reporting bleu scores," 2018.

[18] Maja Popović, "chrF: character n-gram F-score for automatic MT evaluation," in *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, Eds., Lisbon, Portugal, Sept. 2015, pp. 392–395, Association for Computational Linguistics.