# CUFE at NADI 2024 shared task: Fine-Tuning Llama-3 To Translate From Arabic Dialects To Modern Standard Arabic

**Michael Ibrahim**
Computer Engineering Department, Cairo University
1 Gamaa Street, 12613
Giza, Egypt
michael.nawar@eng.cu.edu.eg

## Abstract

LLMs such as GPT-4 and LLaMA excel in multiple natural language processing tasks, however, LLMs face challenges in delivering satisfactory performance on low-resource languages due to limited availability of training data. In this paper, LLaMA-3 with 8 Billion parameters is finetuned to translate among Egyptian, Emirati, Jordanian, Palestinian Arabic dialects, and Modern Standard Arabic (MSA). In the NADI 2024 Task on DA-MSA Machine Translation, the proposed method achieved a BLEU score of 21.44 when it was fine-tuned on the development dataset of the NADI 2024 Task on DA-MSA and a BLEU score of 16.09 when trained when it was fine-tuned using the OSACT dataset.

## 1 Introduction

The transformer architecture has revolutionized natural language processing, leading to the development of large self-supervised neural networks with tens or hundreds of billions of parameters trained on trillions of tokens. Recent models based on the transformer architecture include GPT series (Brown et al., 2020), LLaMA (Touvron et al., 2023), and others. As of late 2024, the release of LLaMA-3 sparked the growth of a thriving open-source community, resulting in the development of near-state-of-the-art models like Mistral (Jiang et al., 2023) and Mixtral (Jiang et al., 2024).

These LLMs can adapt to new tasks and instructions with minimal finetuning. To save on finetuning costs, the NLP community uses Parameter-Efficient Fine-Tuning (PEFT) techniques instead of full-parameter fine-tuning. These PEFT techniques are shown to significantly reduce a model's computational footprint (Houlsby et al., 2019). Low-Rank Adaptation (LoRA) (Hu et al., 2021), an innovation among these PEFT techniques, effectively balances performance and efficiency through low-rank matrix approximations during the fine-tuning phase. Other PEFT techniques, include LLM model weights compression through quantization techniques, cutting down storage and memory usage during both training and inference. By employing techniques such as NormalFloat Quantization (Chen et al., 2023), these methods aim to preserve model performance while considerably reducing LLMs' resource requirements, opening avenues for resource-efficient fine-tuning through their integration with PEFT methods like LoRA.

Large Language Models (LLMs) can be fine-tuned and used for translation of low-resource languages (Lankford et al., 2023). The adaptMLLM system outperformed baselines from the LoResMT2021 Shared Task (Ojha et al., 2021) in translating low-resource languages such as Irish. The improvements were 14% greater in the English to Irish direction (5.2 BLEU points) compared to the Irish to English direction (an increase of 117%, or 40.5 BLEU points).

In this paper, a LLAMA-3 model with an 8-Billions parameters model was finetuned to develop a sentence-level machine translation (MT) of four Arabic dialects into MSA. The rest of the paper is organized as follows. The related work is summarized in Section 2. The dataset used for training and validation was detailed in Section 3. In Section 4, the system is presented. Section 5 summarizes the study's key findings.

## 2 Related Work

### 2.1 Arabic Machine Translation

Arabic ranks fifth in the number of native and total speakers. In the early nineteenth century, Modern Standard Arabic (MSA) was developed from Classical Arabic as the standardized and academic variant of the language. Modern Standard Arabic, having a standard orthography and used in formal contexts, contrasts with Dialectal Arabic,

*Proceedings of The Second Arabic Natural Language Processing Conference*, pages 769–773
August 16, 2024 ©2024 Association for Computational Linguistics

commonly used unofficially and increasing in online usage. Complex words and certain expressions in Arabic dialects differ greatly from Modern Standard Arabic, evidenced by the use of letter concatenation, character repetition, and emoticons for emphasis (Baniata et al., 2018).

Different techniques were used to perform DA to MSA translation, the ELISSA system (Salloum and Habash, 2013), utilizes morphological analysis, transfer rules, dictionaries, and language models, adopts a rule-based approach for DA to MSA translation, and functions as a general preprocessor for DA when working with MSA NLP tools. 93% of ELISSA's MSA translations are correct. Improving MSA-pivoted DA-to-English MT with ELISSA for producing MSA versions of DA sentences increases BLEU score by 0.6% to 1.4% on various blind test sets.

In (Alnassan, 2023), the author distinguishes between rule-based and statistical machine translation methods, emphasizing the challenge of accurately translating dialects compared to standard or contemporary Arabic. They propose an "automatic standardization" solution employing machine translation techniques to generate standard Arabic text from dialect inputs. Instead of creating linguistic rules for each dialect, the authors employ statistical models. The goal is to develop integrated software for automatic standardization and translation of Arabic dialects. The authors propose that transforming dialectal text into standard Arabic could aid in the comprehension of various Arabic dialects.

For low-resource languages, different approaches were employed, (Lample et al., 2017) explored the possibility of translating without utilizing any parallel data. This model maps sentences from bilingual corpora into the same latent space. By mastering the ability to rebuild in two languages from a common feature space, the model can translate accurately without relying on labeled data. This model attained BLEU scores of 32.8 on Multi30k and 15.1 on WMT English-French without being trained on any parallel data.

### 2.2 LLM Applications

Utilizing LLMs for a wide range of NLP tasks is currently popular. The most effective and efficient utilization of these models is yet to be determined. Three primary methods exist for constructing applications utilizing LLMs.

- **Zero-shot prompting** It is querying prompts that aren't in LLM's training data to elicit responses. These prompts often include detailed instructions and a primary question. Crafting precise prompts is essential for maximizing the effectiveness of large language models.

- **Few-shot learning** Few-shot learning involves giving LLMs a limited number of examples to produce appropriate responses. Zero-shot prompting refers to a method without providing any examples. In few-shot learning, examples are incorporated into the prompt template to guide the model's response.

- **Fine-tuning**. The two methods above enable task adaptation without the requirement for additional training on the LLMs, while fine-tuning necessitates further training of the LLMs with task-specific data. When tailored datasets are available, this method is especially advantageous.

According to (Yang et al., 2024), LLMs perform effectively in most NLP tasks based on their examination of 'use cases' and 'no use cases' for particular downstream tasks using the mentioned methods.

### 2.3 LLMs for Machine Translation

LLMs have started to gain prominence in machine translation research. (Hendy et al., 2023) emphasized ChatGPT's advantage in machine translation through prompting. These models may not always surpass the performance of SOTA MT systems and commercial translators. According to (Zhu et al., 2023), decoder-only LLMs, though competitive, fall behind encoder-decoder-based multilingual NLLB (Costa-jussà et al., 2022).

(Bawden and Yvon, 2023) reveal challenges in machine translation through prompting, including copying issues, mistranslation of entities, and hallucinations. However, in a few-shot learning setting, these limitations can be mitigated.

## 3 Data

The development dataset provided by OSACT 2024 shared task (Atwany et al., 2024) on DA-MSA translation was employed for fine-tuning the Llama-3 8-Billions parameters model. The development

| **Dialect** | **Development** | **Test** |
|---|---|---|
| Egyptian | 100 | 500 |
| Emirati | 100 | 500 |
| Jordanian | 100 | 500 |
| Palestinian | 100 | 500 |

Table 1: NADI DA-MSA data split statistics

| **Dialect** | **DEV BLEU** | **TEST BLEU** |
|---|---|---|
| Egyptian | 14.16 | 17.19 |
| Emirati | 17.26 | 22.10 |
| Jordanian | 13.47 | 22.57 |
| Palestinian | 17.03 | 23.89 |

Table 2: Results of the development and test sets when the model is fine-tuned with the OSACT dataset

| **Dialect** | **BLEU** |
|---|---|
| Egyptian | 14.86 |
| Emirati | 17.35 |
| Jordanian | 15.98 |
| Palestinian | 16.20 |

Table 3: Results of the test sets when the model is fine-tuned with the NADI development dataset

dataset includes comprised a total of 1001 source-to-target examples, evenly distributed among dialects as follows: 200 Egyptian, 200 Maghrebi, 200 Levantine, 201 Gulf, and 200 Iraqi examples.

The performance of the developed method was assessed using the development and test dataset provided by NADI 2024 shared task (Abdul-Mageed et al., 2024) on DA-MSA translation. The development dataset includes 100 Egyptian, 100 Emirati, 100 Jordanian, and 100 Palestinian sentences, totaling 400 sentences, and the test dataset includes 500 Egyptian, 500 Emirati, 500 Jordanian, and 500 Palestinian sentences, totaling 2000 sentences. The dataset statistics are summarized in table 3.

## 4 Methodology

The Llama-3 model was fine-tuned with LoRA keeping the Llama-3 weight frozen updating only the LoRA matrices using the following formatted prompt-response triplets: Dialect [dialect], Arabic Dialect Sentence [source] and Modern Standard Arabic Sentence [target]. The dev set, like the training set, is formatted similarly. This prompt customizes the model for the DA-MSA machine translation assignment. The prompt is formatted as follows:

> *<s> [INST] «SYS» You are an expert Arabic proofreader! «/SYS» Translate the following text from the [dialect] dialect to Modern Standard Arabic. [source] [/INST] [target] </s>*

For example, the first annotation in the OSACT DA-MSA dataset will have the following prompt

> *<s> [INST] «SYS» You are an expert Arabic proofreader! «/SYS» Translate the following text from the Egyptian dialect to Modern Standard Arabic.* تتعلم ازاي *[/INST]* كيف تتعلم *</s>*

During the finetuning of the Llama-3 model, the learning rate of the Adam optimizer was set to 1e-4. The evaluation is conducted every 50 steps with a batch size of 1. The low-rank approximation rank is set to 64, and its scaling factor for adaptation is set to 16 in the LoRA configurations. The model trainable parameters are all linear layers: “q_proj”, “up_proj”, “o_proj”, “k_proj, ‘down_proj”, “gate_proj”, and “v_proj”. A 0.05 dropout is applied in the LoRA layer. The model’s weights are quantized to 4-bit precision and mixed-precision training with float16 and float32 is enabled to reduce memory requirements and accelerate the training process. This model was trained on Google Colab with a single NVIDIA A100 GPU with 40GB of memory. The finetuning of the model and the generation of the test results took almost one hour. The code used for training and generating the results can be found on the following GitHub repository. [1]

When the model was fine-tuned using the OSACT dataset, the system achieved a BLUE score of 15.52 on the NADI development dataset, and a BLUE score of 16.09 on the NADI testing dataset. The system results are summarized in table 2. When the model was fine-tuned using the NADI development dataset, the system achieved a BLUE score of 21.44 on the NADI testing dataset. The system results are summarized in table **??**

One final comment about the model, when the model is fine-tuned using the OSACT dataset, the OSACT dataset doesn’t have any “Emirati”, “Jordonain” or “Palestinian” sentences, so, during the generation of the translated sentences, the “Emirati” sentence were considered as “Gulf” dialect, and

[1] https://github.com/MichaelIbrahim-GaTech/NADI2024-subtask-3.git

the "Jordanian" and "Palestinian" sentences were considered as "Levantine" dialect.

## 5 Conclusions and Future Work

In this paper, PEFT methods like quantization and LoRA were used to finetune a LLama-3 model with 8 billion parameters, this model was adapted to translated between 4 different Arabic dialects and Modern Standard Arabic. The proposed method is efficient and it achieved a BLEU score of 21.44 on the NADI 2024 Task on DA-MSA Machine Translation when finetuned using the development dataset of the NADI shared task, and a BLEU score of 16.09 when finetuned on the OSACT dataset.

In the future, three interesting topics can be further explored: (1) We can further explore the effect of changing the model parameters on the results (i.e. learning rate, batch size, LoRA rank, etc.), (2) We can try different LLMs other than Llama-3, and (3) We can assess the proposed method on other Arabic dialects.

## References

Muhammad Abdul-Mageed, Amr Keleg, AbdelRahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. 2024. NADI 2024: The Fifth Nuanced Arabic Dialect Identification Shared Task. In *Proceedings of The Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*.

Abidrabbo Alnassan. 2023. Automatic standardization of arabic dialects for machine translation. *arXiv preprint arXiv:2301.03447*.

Hanin Atwany, Nour Rabih, Ibrahim Mohammed, Abdul Waheed, and Bhiksha Raj. 2024. Osact 2024 task 2: Arabic dialect to msa translation. In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation@ LREC-COLING 2024*, pages 98–103.

Laith H Baniata, Seyoung Park, and Seong-Bae Park. 2018. A neural machine translation model for arabic dialects that utilizes multitask learning (mtl). *Computational Intelligence & Neuroscience*.

Rachel Bawden and François Yvon. 2023. Investigating the translation performance of a large multilingual language model: the case of bloom. *arXiv preprint arXiv:2303.01911*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jiaao Chen, Aston Zhang, Xingjian Shi, Mu Li, Alex Smola, and Diyi Yang. 2023. Parameter-efficient fine-tuning design spaces. *arXiv preprint arXiv:2301.01821*.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Séamus Lankford, Haithem Afli, and Andy Way. 2023. adaptmllm: Fine-tuning multilingual language models on low-resource languages with integrated llm playgrounds. *Information*, 14(12):638.

Atul Kr Ojha, Chao-Hong Liu, Katharina Kann, John Ortega, Sheetal Shatam, and Theodorus Fransen. 2021. Findings of the loresmt 2021 shared task on covid and sign language for low-resource languages. *arXiv preprint arXiv:2108.06598*.

Wael Salloum and Nizar Habash. 2013. Dialectal arabic to english machine translation: Pivoting through modern standard arabic. In *Proceedings of the 2013*

*Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 348–358.

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

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.