From Nile Sands to Digital Hands: Machine Translation of Coptic Texts

 $^{\lambda}$ MBZUAI, UAE ξ University of British Columbia, Canada

 ϵ African Masters in Machine Intelligence, Senegal

^σUniversität des Saarlandes, Germany

{muhammed.yahia, mukhtar.mohamed, shady.shehata}@mbzuai.ac.ae

amohamed@aimsammi.org, muhammad.mageed@ubc.ca

Abstract

The Coptic language, rooted in the historical landscapes of Egypt, continues to serve as a vital liturgical medium for the Coptic Orthodox and Catholic Churches across Egypt, North Sudan, Libya, and the United States, with approximately ten million speakers worldwide. However, the scarcity of digital resources in Coptic has resulted in its exclusion from digital systems, thereby limiting its accessibility and preservation in modern technological contexts. Our research addresses this issue by developing the most extensive parallel Copticcentered corpus to date. This corpus comprises about 8,000 parallel sentences between Arabic and Coptic, and more than 24,000 parallel sentences between English and Coptic. We have also developed the first neural machine translation system between Coptic, English, and Arabic. Lastly, we evaluate the capability of leading proprietary Large Language Models (LLMs) to translate to and from Coptic using a few-shot learning approach (in-context learning). Our code and data are available at https: //github.com/UBC-NLP/copticmt.

1 Introduction

With written records dating back to 3,400 B.C., Egyptians trace their roots to one of the world's oldest civilizations. While modern Egyptians primarily speak the Egyptian dialect of Arabic, millions in Northeastern and Eastern Africa continue to use Coptic as their everyday language. This unique language combines elements of the ancient Egyptian languages Demotic, Hieroglyphic, and Hieratic. It was the language of the Ptolemaic rulers following the spread of Greek culture throughout much of the Near East. The Coptic script borrows from the Greek alphabet while retaining features from Egypt's writing traditions.

Large Language Models (LLMs) are known for their exceptional performance with languages extensively represented in their pre-training data (Wang et al., 2023; Zhang et al., 2023b). However, their performance significantly declines for languages with lower representation. Several methods have been proposed to address this issue. These methods include continual training incorporating new data representing the low-resource languages (Gupta et al., 2023; Zhao et al., 2024; Husain et al., 2024), instruction fine-tuning (Zhao et al., 2024; Azime et al., 2024), and vocabulary extension to cover new languages (Cui et al., 2024; Zhang et al., 2023b).

The primary challenge with these methods is their reliance on substantial amounts of new data. In the case of Coptic, as far as we know, no specialized dataset exists for pre-training or fine-tuning models. Furthermore, an examination of the vocabulary of several LLMs (Raffel et al., 2020; Xue et al., 2021; Touvron et al., 2023) reveals that Coptic characters are entirely absent. We hypothesize that the absence of Coptic language symbols will prevent these models from generating or understanding the Coptic language.

To address these challenges for Coptic, this study begins by compiling the available Coptic text from the internet into a unified parallel dataset. We then develop a compact translation model between Arabic/Coptic and English/Coptic language pairs. To develop the model, we initially establish a baseline using a simple Seq2Seq architecture (Vaswani et al., 2017). We train this model

^{*} These authors contributed equally.



Figure 1: Coptic translation

with English-Coptic and Arabic-Coptic pairs and then experiment with a many-to-many approach where the same model is trained to translate between all three languages. In the second phase, we perform full fine-tuning on more specialized models such as mT5 and M2M100 (Fan et al., 2021; Xue et al., 2021), using both pairwise and many-tomany translation methods. Moreover, as we have a limited dataset and unseen scripts, we hypothesize that the T5-Small version will generalize better to the dataset compared to the T5-Base since the latter is larger and can overfit the training data.

Furthermore, we assess the capabilities of the current top proprietary and open-source LLMs on Coptic-centered translation. We experiment with closed models such as GPT-4-Turbo-2024-04-09 (OpenAI, 2024), Claude 3 Opus (Anthropic, 2024), and Gemini 1.5 Pro-API-0409-Preview (Team et al., 2023). For open-source models, we investigate Llama-3-70b-Instruct (Touvron et al., 2023). We focus on translating between the various pairs without any training, utilizing an in-context (zero-shot and few-shot) learning approach.

The rest of this paper is organized as follows: Section 2 is about related work. Section 3 offers a background about the Coptic language. In Section 4, we introduce our datasets. We present our methods in Section 5 and our results in Section 6. We conclude and provide an outlook about future work in Section in Section 7.

Consequently, we have endeavored to compile available parallel corpora and then assess the best methods to fine-tune LLMs for translating between these language pairs. Previous research, such as DIPMT (Ghazvininejad et al., 2023), has explored translation in low-resource languages through prompting, providing translations for uncommon words using existing dictionaries. This approach suits languages where LLMs exhibit basic syntactic understanding, potentially

learned through n-gram language modeling. However, traditional prompting methods are often inadequate when faced with a completely new language, from which LLMs must learn vocabulary and grammar from scratch. This led to the development of DIPMT++ and has inspired further research into machine translation for low-resource languages. And the Coptic not only most of the LLMs, hasn't been seen during the training moreover the Coptic characters are not present in the dictionary of most of the LLMs such as LLama2, LLama3 (Touvron et al., 2023), mT5 (Xue et al., 2021) With this in mind we have decided to handle the tasks of the multilingual machine translation between the Arabic, English and Coptic in the 5 setups the first one we are going to build the baseline using Seq2Seq (Vaswani et al., 2017) we are going to train in two fashions the first one in which we have pairs b (English \leftrightarrow Coptic) and (Arabic \leftrightarrow Coptic) alongside use of the many to many langauge in which the same model is trained to translate between three languages (Arabic \leftrightarrow Coptic \leftrightarrow) the goal is to test which approaches can best results as the available parallel corpus is limited and we aim to use whatever resource we have to the fullest, the second setup is which in we are going to perform full finetuning to (Fan et al., 2021; Xue et al., 2021) models and we are going to before also pairs and many translation, the third setups is using Zeroshot learning in which we are going to use open source models such as LLama3 and other closed source models such as GPT3.5, GPT4 and Gemini models, and as the closed source models will require us to access them via API for those closed model we are going to use portion of the test set that is sampled randomly. The fourth setup is fewshot learning, the fifth is the in-context learning for the Gemini Model, and finally, we are going to discuss how we can tune (Zhang et al., 2024) model to handle the case of the Coptic Arabic and English translation to handle the machine translation between the different languages. In this paper, we are taking our first steps to create such a machine translation system. It is like we are building a bridge between languages. To help our system learn, we are introducing it to 24,000 parallel Coptic and English sentences and 16,800 Arabic to Coptic Parallel data. We aim to build a multi-language machine translation system capable of translating between Arabic and Coptic, Coptic and Arabic, English and Coptic, and Coptic and English. We have collected

a corpus of parallel sentences and proposed transformers (Vaswani et al., 2017) pipeline for Coptic to English translation. Additionally, through transfer learning, we found that the mT5 (Xue et al., 2021) model can generalize and adapt to languages with unseen scripts, such as Coptic, even without prior exposure to Coptic data during training. Our data collection efforts and adaptation of large language models like mT5 (Xue et al., 2021) for unseen scripts also hold promise for other languages.

2 Related Work

This section reviews research on fine-tuning large language models (LLMs) for machine translation in low-resource languages. Most studies focus on languages encountered during LLM pre-training, excluding Coptic. We examine specific LLMs and pre-trained language models (PLMs) used in our study to determine their exposure to Coptic. We also review methods for fine-tuning LLMs for lowresource languages, establishing a baseline, and identifying challenges. Our goal is to adapt and enhance these techniques for Coptic, which is significantly underrepresented in machine translation research.

2.1 Large Language Models in Use

Since the introduction of transformers by Vaswani et al. (2017), large language models (LLMs) have significantly progressed, diversifying the landscape of machine translation technologies. Initially, these transformers utilized an encoderdecoder framework. Over time, however, pretrained language models (PLMs) and other LLMs have adopted various architectural frameworks. Some PLMs targeted at natural language understanding (NLU) employ an encoder-only model, while others focused on natural language generation (NLG) utilize solely the decoder component of the transformer model (Touvron et al., 2023; ?; Team et al., 2023; Anthropic, 2024; OpenAI, 2024). Meanwhile, specific models continue to implement the full encoder-decoder configuration, especially those designed for translation tasks, such as the model by Fan et al. (2021), which was trained on over a hundred languages, and text-totext models like MT5 (Xue et al., 2021), which can handle multiple tasks but require specific prompts.

In our study, we explore the use of the aforementioned models (Vaswani et al., 2017; Fan et al., 2021; Xue et al., 2021; Touvron et al., 2023; Anthropic, 2024; Team et al., 2023; OpenAI, 2024) for machine translation between Coptic, Arabic, and English. Our initial step involves identifying which of these models have been trained in the Coptic language. Examination of the dictionaries from open-source models (Xue et al., 2021; Fan et al., 2021; Touvron et al., 2023) revealed an absence of Coptic vocabulary, indicating no direct training in this language. For closed-source models, zero-shot testing showed that while GPT-4 recognizes Coptic characters, its translation performance is poor, suggesting exposure to Coptic data but not to our specific parallel corpus (OpenAI, 2024). Conversely, Claude by Anthropic (2024) demonstrated some capability in zero-shot translation tasks involving Coptic, suggesting it may have been trained on a parallel corpus that includes Coptic, English, and Arabic (Anthropic, 2024).

2.2 Adapting Pretrained Language Models to Low-Resource Settings

Adapting large language models (LLMs) to lowresource languages is a multifaceted process. Commonly, this adaptation involves continual pretraining on monolingual texts to enhance model performance (Lin et al., 2023). However, extensive monolingual resources are not always available for the initial adaptation of the LLM embeddings to the target low-resource language before fine-tuning for translation tasks. In the field of neural machine translation, there is no prior work specifically addressing the Coptic language. Nonetheless, similar research can offer valuable insights. For example, Ahia and Ogueji (2020) developed the first supervised machine translation system between Nigerian Pidgin and English using Seq2Seq transformers (Vaswani et al., 2017), achieving BLEU scores of 24 and 21 for English to Pidgin and Pidgin to English translations, respectively. Notably, due to the limited corpus, their approach involved reducing the size of the transformer model without employing transfer learning from pre-trained language models. In contrast, Lin et al. (2023) enriched their translation efforts by assembling nearly 30K parallel texts from the Holy Bible and then applied transfer learning using the T5 English model and the mT5 model for the translation tasks, significantly enhancing the performance of transformerbased models. Similarly, Osman et al. (2023) enriched an Arabic-Swahili parallel corpus by sourcing texts from the Bible. They trained models (Fan

et al., 2021; Xue et al., 2021) for Arabic-Swahili machine translation, finding that the model from Fan et al. (2021) outperformed Xue et al. (2021). Their use of the T5 model explored state-of-the-art outcomes in bidirectional translation (Osman et al., 2023; Fan et al., 2021), underscoring the potential of LLMs in translating between English and Pidgin as well as other language pairs.

2.3 Teaching LLMS for unseen language

Adapting LLMs to Zhuang Zhang et al. (2024) introduced DIPMT++, a method designed to adapt large language models (LLMs), such as GPT-4 and LLaMA-2 to unseen language on the fly meaning without the need of collecting further monolingual corpus and pretrain the LLMs on them nor the need for Supervised Finetuing SFT. Zhang et al. (2024) has tested their approach with the Zhuang language. Although these models were not initially trained to handle Zhuang, they can tokenize Chinese characters, which are used in the Zhuang writing system. This capability allows the LLMs at least correct tokenization of the input. The DIPMT++ method utilizes a combination of a dictionary and a limited parallel corpus of 5,000 sentences to facilitate in-context learning (ICL). Central to this approach is a dictionary compiled from existing resources, which includes Zhuang words alongside their Chinese translations, thus providing critical lexical information. Within the DIPMT++ framework, this dictionary is directly integrated into translation prompts, pairing each word in the source sentence with its corresponding dictionary translation. This design uses sample translations as exemplars, enhancing the model's understanding and generative capabilities, significantly improving translation accuracy, and demonstrating the potential of ICL to adapt LLMs to new languages with scant data efficiently.

3 The Coptic Language

3.1 The Evolution of the Coptic Language

Until the arrival of the Arabs in the seventh century, Coptic was the spoken language of ancient Egypt. It initially appeared in hieroglyphic script, then progressed to hieratic and demotic scripts (Suscopts, n.d.). Around the third century B.C., during the Ptolemaic era, Egyptians adopted the Greek alphabet (comprising 25 letters from Greek origins), creating what we now know as Coptic. However, due to the Greek alphabet's limitations in representing Egyptian sounds, seven demotic letters were integrated into the Coptic alphabet, resulting in a total of 32 letters in the Coptic alphabet (Suscopts, n.d.).

The Coptic language was the predominant spoken language in Egypt until the 11th Century, at which point it began to be supplanted by Arabic. Historical records suggest that in certain regions of Upper Egypt, Coptic continued to be spoken until the 17th century (Suscopts, n.d.). In contemporary times, Coptic is mainly used within the context of religious ceremonies and traditions within the Coptic Orthodox Church. Although it is still utilized in rituals like prayers, hymns, and reading sacred texts, its spoken usage in everyday conversations has greatly decreased.

3.2 The Coptic Diaspora

Coptic Oriental Orthodox Christianity is the predominant faith among the majority of ethnic Copts, making them the foremost Christian denomination in Egypt, as well as across the Middle East, Sudan, and Libya. In Egypt, Copts make up approximately 10-20% of the population (around 18 million followers). In Sudan and Libya, Copts represent just over 1% of the respective populations (contributors, 2024). Figure 2 provides a visual representation of countries where Coptic communities have indigenous roots.

The Coptic diaspora extends beyond Egypt, Sudan, and Libya, with significant communities in the United States, Canada, and Australia. The United States hosts the largest Coptic diaspora globally, with estimates ranging from 100K to 1M individuals, although official data remains scarce (contributors, 2024). Smaller Coptic communities, under 10K people, exist in countries including Britain, France, Germany, Austria, Switzerland, Italy, Lebanon, Jordan, and the United Arab Emirates. Recent missionary efforts have also led to the emergence of Coptic communities in several sub-Saharan African nations such as Zambia, Kenya, the Democratic Republic of the Congo, Zimbabwe, Namibia, and South Africa contributors (2024).

3.3 The Diverse Landscape of Coptic Dialects

Coptic encompasses six distinct dialects, with four spoken in Upper Egypt and two in Lower Egypt, primarily differing in their phonetic systems (The Editors of Encyclopaedia Britannica, 2017). The Fayyūmic dialect, spoken in Upper Egypt along the

Greek	α	β	γ	δ	3		ζ	η	θ	ι	к	λ	μ	ν	y	0	π	ρ	σ	τ	υ	φ	χ	ψ	ω
Coptic	۵	B	Г	Δ	е	٢	Ζ	н	θ	l	κ	λ	м	и	z	0	Π	Р	с	т	γ	ф	X	۲	w
Table 1: Adaptation of Greek letters into Coptic alphabet.																									

Coptic Letter	യ	q	þ	2	Х	6	t	
Sound	Shai	Fai	Khai	Hori	Ganga	Cheema	Tee	
Table 2: Contic letters derived from older Egyptian scripts								





Figure 2: The map illustrates the main regions where the Coptic community primarily resides in present-day Egypt, Libya, and Sudan.

Nile's west bank, persisted until the 8th century. Asyūțic, also known as Sub-Akhmīmic, thrived around Asyūț in the 4th century, while Akhmīmic was prevalent in and around the city of Akhmīm. Sahidic, initially spoken near Thebes, became the standard Coptic for Upper Egypt after the 5th century, known for its extensive documentation (The Editors of Encyclopaedia Britannica, 2017). In Lower Egypt, Bashmūric remains poorly understood, with limited surviving texts, while Bohairic, originating in the western part of Lower Egypt, including Alexandria and Memphis, has been the standard dialect for Coptic Christians since the 11th century (The Editors of Encyclopaedia Britannica, 2017).

4 Data Collection

During the early modern era, the Coptic language, once a medium of daily communication, became obsolete and was largely supplanted by Egyptian Arabic. Nevertheless, its significance persisted, primarily as the liturgical medium for Coptic Christians. Consequently, the predominant body of extant Coptic literature is constituted by ecclesiastical writings penned by illustrious figures of the Coptic Church and some seminal translations of the Bible into Coptic dialects. Notably, the Coptic rendition of the Bible spans both the Old and New Testaments and is available in multiple Coptic dialects, such as Sahidic or Bohairic.

4.1 English Corpus Selection

To compile our dataset, we sourced translated iterations of Biblical texts and hagiographical narratives of Coptic saints from the Coptic SCRIPTORIUM. The SCRIPTORIUM, denoting <u>Sahidic Corpus Research</u>: Internet <u>Platform</u> for Interdisciplinary multilayer <u>Methods</u>, grants open access to a vast repository of Coptic manuscripts. These are available in multiple digital formats, among which are XML (.xml) and TreeTagger (.tt)¹.

Given its enriched metadata, we opted for the TreeTagger format, which was later transformed into HTML to streamline the process of accessing and parsing file details during the preparatory stage.

Our selected corpus comprises more than 120 Biblical texts, aggregating over 950 chapters. In contrast, the hagiographical narratives detailing the lives of saints encapsulate 36 books attributed to 12 distinct authors. Contrary to the modular structure of the Biblical texts, the hagiographical data is more aggregated, with each file amalgamating several chapters.

Scripts Genres	Books	Chapters	Verses
Old Testaments	73	728	14,609
New Testaments	66	259	7,907
Saints' Lives	36	-	3,428
Total Verses			25,944

Table 3: Distribution of genres in the collected Coptic corpus.

4.2 Arabic Corpus

For our study on the Arabic language, we gathered data from the Bible ². Initially, we prepared the Arabic texts by removing diacritics to ensure consistency. We then ensured our dataset was a parallel corpus by comparing books and chapters, using

¹For a comprehensive overview of the SCRIP-TORIUM initiative, readers are directed to https: //copticscriptorium.org ²https://ebible.org/

Coptic	English					
укнс граі ппршме еточаав пбючма	There is implanted in a holy man the					
NXICBO	earnest desire to learn,					
ЕЩАХЕ ЕММИТРОШПИОТ ·	I mean that not all have given thanks.					
ефхе генмеріт не некма нффпе пхоєіс	How amiable are thy tabernacles, O					
иибом	Lord of hosts!					

Table 4: Sample translations from Coptic to English from the Bible.

only those with matching verse counts. Chapters with differing numbers of verses were excluded from our study. In total, our dataset includes 39 unique books and 929 chapters, resulting in about 8,000 parallel verses between the Arabic and Coptic datasets.

5 Methods

This section outlines the models used for machine translation between Coptic, Arabic, and English. It includes three subsections: the first introduces the baseline model, the second covers the adaptation strategies, and the third discusses zero-shot and few-shot translation methods.

5.1 Baseline Model

We employ a Transformer-based model as our baseline, mirroring the architecture described by Vaswani et al. (2017). This model features six encoder and six decoder layers, eight attention heads, 512-dimensional embeddings, and 2,048 units in the feed-forward network. It is trained from scratch on Arabic-Coptic and Coptic-English parallel corpora. For tokenization, we use both a word-based method and Byte Pair Encoding (BPE). Ending with two models, one that supports translations between Arabic and Coptic and another that empowers translation between English and Coptic (see Table 6).

5.2 Fine-tuning

Building on studies by Lin et al. (2023) and Osman et al. (2023), we fine-tune the M2M100 and mT5 models in their base and small configurations. The mT5 Small model contains about 300 million parameters across eight encoder-decoder layers, while the mT5 Base model includes approximately 580 million parameters across 12 layers. We train with a batch size of 16 and max sequence length of 196 for 20 epochs. These configurations utilize the Transformer architecture effectively, with the Base model's larger size providing enhanced capabilities for handling complex multilingual tasks. We also experimented with Fan et al. (2021) model containing 418 million parameters using the same hyperparameters as the authors, but it did not produce legible results. As mentioned in Section 2, neither of the models was exposed to the Coptic dataset during their initial training. Therefore, following this initial phase, we plan to train the best-performing models using a unified translation model between the three language sets. This model facilitates many-to-many translations among Arabic, Coptic, and English.

5.3 Zero-Shot and Few-shot

Recent advances in LLMs have showcased their ability to handle various tasks, even those not explicitly encountered during training (Brown et al., 2020). This is achieved by presenting a few examples of the desired task at the time of inference, eliminating the need to explicitly train the model on those tasks. This approach, known as *few*-shot learning or *in-context learning*, is proving to be particularly promising for the field of machine translation (Sia and Duh, 2022; Hendy et al., 2023; Zhu et al., 2023), especially when dealing with low-resource languages like Coptic, which have very limited parallel corpora.

In this context, LLMs offer two powerful approaches:

- Zero-shot prompting: By crafting carefully worded prompts that haven't been seen in the model's training data, we can guide LLMs to perform translations. These prompts often include explicit instructions alongside the source text, essentially asking the model to deduce how to translate based on its broader understanding of language.
- Few-shot learning: A more guided approach, few-shot learning involves providing the model with a small set of high-quality translation examples within the prompt itself. These examples act as context, demonstrating

Arabic	Coptic
	эен оунагт еөве инеөнафшпі ад`смоу
بِالإيمانِ، بارَكَ إسحاقُ يَعْقُوبَ وعيسو	`nxe`lcaak `єіакшв нем нсау
مَعَ أَنَّ اللهَ قالَ لهُ: «بِإسحاقَ سوفَ يكونُ لكَ نَسلٌ يَحمِلُ	фнетаусахі немац хе зен ^ Ісаак
اسمكَ!»	еу^евагем о^хрох нак
	гисте маном† `ппетеперноу эеп
لِذَلكَ عَرُّوا بَعضُكم بَعضًا بِهَذا الكَلامِ!	NAICAXI

Table 5: Sample translations from Arabic to Coptic.

the desired style, register, and translation patterns for the model to emulate.

Crucially, those approaches significantly depart from the data-intensive demands of supervised neural machine translation (NMT). In the context of Coptic (and low-resource languages in general), where parallel text data is limited, prompt-based translation offers a viable alternative for developing effective translation systems. This paper delves into applying zero-shot and few-shot learning to bridge the gap between modern LLMs and the unique challenges posed by Coptic translation.

The main goal of this experiment is to test the capabilities of large language models (LLMs) in understanding and translating to and from the Coptic language. To achieve this, we opted to use the top LLMs according to LLM Arena³, which include GPT-4-Turbo-2024-04-09 (OpenAI, 2024), Claude 3 Opus (Anthropic, 2024), Gemini 1.5 Pro API-0409-Preview (Team et al., 2023), and the best open-source model to date, Llama-3-70b-Instruct (Touvron et al., 2023).

In our initial zero-shot experiment, models are prompted to generate translations without prior examples, directly translating from the source to the target language. To carry out this experiment, we curated a dataset of 100 examples from each of the following language pairs: Arabic-Coptic, English-Coptic, and English-Arabic. In each pair, we tested translations in both directions. The English-Arabic set serves as a baseline for comparison, allowing us to gauge the models' performance on a more established language translation task.

In both zero-shot and few-shot experiments, we employed a random selection strategy due to the absence of a model (based on our testing) capable of generating sufficiently accurate embeddings for the Coptic language. Other strategies proposed in prior work (Zhang et al., 2023a; Chowdhery et al., 2022) were therefore not applicable in this context. For the template, we use the same instruction and format as recommended in the OpenAI playground for the default sentence-level translation task.

6 Results

In this section, we present the outcomes of our experiments. First, we offer the results of our Seq2Seq transformer-based models compared to mT5 (see Table 6), then our zero- and few-shot results (see Table 7). Our experiments involve several models: m2m100, mT5base, and mT5-small. As Table 6 shows, some of our Seq2Seq baseline models are able to produce translations between the various pairs (evaluated in BLEU scores). However, the models remain relatively weak.

Despite the fact that mT5 small and large models were not exposed to Coptic text during the pretraining stage, and their dictionaries do not include any Coptic text, the results demonstrate that the mT5 model can successfully adapt and generate Coptic text. However, it was observed that the m2m100 model struggles with encoding and decoding in this context, leading to frequent generation of "<unk>" tokens. Consequently, we have chosen not to report the results for the M2M100 model as the results highlight that M2M100 cannot adapt to unseen language with unseen scripts in the dictionary of the M2M100.

6.1 Zero/Few shot learning

Table 7 displays the BLEU scores for zero-shot and five-shot pairwise translation tasks across various LLMs: GPT-4-Turbo-2024-04-09 (OpenAI, 2024), Claude 3 Opus (Anthropic, 2024), Gemini 1.5 Pro API-0409-Preview (Team et al., 2023), and Llama-3-70b-Instruct (Touvron et al., 2023). Translation tasks were conducted for the following language pairs: English-Coptic, Arabic-Coptic, and Arabic-English, where sentences from one language were translated into the other for each pair. The Arabic-English language pair serves as our

³https://chat.lmsys.org/?leaderboard

Model	Coptic +	→ Arabic	$\textbf{Coptic}\leftrightarrow \textbf{English}$			
	AR to Coptic	Coptic to AR	EN to Coptic	Coptic to EN		
Baseline (WT)	0.33	3.37	3.02	7.41		
Transformers (BPE)	1.86	0	0	6.65		
mT5 (Arabic, Coptic)	2.89	0.873	0.023	0.003		
mT5 (English, Coptic)	0.013	0.009	6.84	14.9		
mT5 (Many to Many)	1.974	0.571	2.12	4.466		

Table 6: Results for the baseline model, a Seq2Seq transformer-based model featuring a 6-6 encoder-decoder configuration and 8 attention heads. This model utilizes two types of tokenization: WT stands for word-based, and BPE (Byte Pair Encoding), where "AR" stands for Arabic, "EN" stands for English. For the mT5, we have three models: the first is trained on Arabic-Coptic pairs, the second on English-Coptic pairs, and the third is a many-to-many model trained on translations between Arabic, Coptic, and English. We are evaluating these using an mT5-based framework.

Experiment	Model	Arabic	↔ Coptic	English	↔ Coptic	Arabic	⇔English
Zero Shot	GPT-4	4.62	2.55	11.87	3.58	8.19	32.72
	Gemini	0.0	3.46	17.75	0.0	14.61	52.5
	Claude3	6.05	6.44	15.34	4.62	8.23	45.48
	Llama3	1.69	1.57	20.82	1.09	4.16	24.14
Five Shot	GPT-4	10.59	3.46	11.25	4.07	4.99	23.59
	Gemini	6.83	4.44	26.89	5.3	8.91	39.2
	Claude3	12.19	53.04	56.45	9.98	27.78	83.05
	Llama3	8.03	2.43	14.13	2.72	4.99	28.37

Table 7: Comparison between the top LLMs (GPT-4-Turbo-2024-04-09 (OpenAI, 2024), Claude 3 Opus (Anthropic, 2024), Gemini 1.5 Pro API-0409-Preview (Team et al., 2023), and Llama-3-70b-Instruct (Touvron et al., 2023)) in zero-shot and five shot translation between Coptic, Arabic, and English language pairs.

baseline to demonstrate the models' performance on languages encountered during their pretraining.

For the zero-shot experiment, we observe that most results are very low (compared to the Arabic-English baseline), with translations from Coptic to other languages performing better than translations to Coptic. This indicates that the models understand Coptic better than they generate it. Claude 3 Opus produces the best results across most language pairs, with GPT-4 and Gemini following closely behind.

In the five-shot experiment, performance significantly improves across all language pairs, illustrating that in-context learning and providing LLMs with a few examples can greatly enhance their performance in low-resource languages like Coptic. Claude 3 Opus again produces the best results, with GPT-4 and Gemini closely following.

7 Conclusion

We presented our efforts to develop a machine translation system tailored for the Coptic language, a language characterized by its limited resources. Our endeavors led to the assembly of the most extensive parallel corpus involving Arabic, English, and Coptic to date. Building on this foundation, we introduced a novel framework emphasizing crosslingual adaptive training for the mT5 model, particularly beneficial for languages with rare scripts. The empirical results from our study underscore the potential of cross-lingual adaptive training as an instrumental approach for enhancing language processing with unseen scripts. As we look ahead, our research is geared towards refining adaptation techniques, with a particular focus on harnessing the capabilities of English-centric as well as non-English-centeric pretrained LLMs. In addition, we plan to amass an even more extensive Coptic dataset. With this enriched dataset, we plan to continue pretraining several models such as mT5 as well as larger models such as LLama3.

Limitations

While our research has yielded promising results, there are inherent limitations that need to be addressed. One of the primary challenges we faced was the adaptation of beyond-centric translation models, such as m2m100 Fan et al. (2021), to scripts they have never encountered during training. The vocabulary of these models is not inherently designed to accommodate unseen scripts, which poses a challenge for languages like Coptic. To overcome this, there is a pressing need to explore factorization techniques that can modify the weights of these models, enabling them to adapt their vocabulary to new scripts effectively.

In addition, we were not able to perform experiments under more extensive settings due to computing limits.

Acknowledgments

We acknowledge support from Canada Research Chairs (CRC), the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-2018-04267), the Social Sciences and Humanities Research Council of Canada (SSHRC; 435-2018-0576; 895-2020-1004; 895-2021-1008), Canadian Foundation for Innovation (CFI; 37771), Digital Research Alliance of Canada,⁴ and UBC ARC-Sockeye.

References

- Orevaoghene Ahia and Kelechi Ogueji. 2020. Towards supervised and unsupervised neural machine translation baselines for nigerian pidgin. *arXiv preprint arXiv:2003.12660*.
- Anthropic. 2024. Claude 3 haiku: Our fastest model yet. https://www.anthropic.com/ news/claude-3-haiku.
- Israel Abebe Azime, Atnafu Lambebo Tonja, Tadesse Destaw Belay, Mitiku Yohannes Fuge, Aman Kassahun Wassie, Eyasu Shiferaw Jada, Yonas Chanie, Walelign Tewabe Sewunetie, and Seid Muhie Yimam. 2024. Walia-Ilm: Enhancing amharic-Ilama by integrating task-specific and generative datasets.
- 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.

- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.
- Wikipedia contributors. 2024. Coptic diaspora. *Wikipedia, The Free Encyclopedia*. Accessed: 2024-06-21.
- Yiming Cui, Ziqing Yang, and Xin Yao. 2024. Efficient and effective text encoding for chinese llama and alpaca.
- Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. J. Mach. Learn. Res., 22(1).
- Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phraselevel prompting of large language models for machine translation. *arXiv preprint arXiv:2302.07856*.

⁴https://alliancecan.ca

- Kshitij Gupta, Benjamin Thérien, Adam Ibrahim, Mats L. Richter, Quentin Anthony, Eugene Belilovsky, Irina Rish, and Timothée Lesort. 2023. Continual pre-training of large language models: How to (re)warm your model?
- 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.
- Jaavid Aktar Husain, Raj Dabre, Aswanth Kumar, Jay Gala, Thanmay Jayakumar, Ratish Puduppully, and Anoop Kunchukuttan. 2024. Romansetu: Efficiently unlocking multilingual capabilities of large language models models via romanization.
- Pin-Jie Lin, Muhammed Saeed, Ernie Chang, and Merel Scholman. 2023. Low-resource crosslingual adaptive training for nigerian pidgin. In *Proceedings of the 24th INTERSPEECH conference.*
- Inc OpenAI. 2024. Gpt4. https: //platform.openai.com/docs/models/ gpt-4-turbo-and-gpt-4 [Accessed: (Use the date of access)].
- Asim Awad Osman, Ahmed Emadeldin Almahady, Muhammed Saeed, and Hiba Hassan Sayed. 2023. Machine translation baselines for arabicswahili.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.
- Suzanna Sia and Kevin Duh. 2022. Prefix embeddings for in-context machine translation. In Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track), pages 45–57.
- Suscopts. n.d. Coptic alphabet. PDF document, *Suscopts*. Accessed: 2024-06-21.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu,

Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

- The Editors of Encyclopaedia Britannica. 2017. Suhaj governorate, egypt. *Encyclopedia Britannica*. Accessed: 2024-06-21.
- 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*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. Advances in neural information processing systems, 30.
- Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen tse Huang, Wenxiang Jiao, and Michael R. Lyu. 2023. All languages matter: On the multilingual safety of large language models.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 483–498, Online. Association for Computational Linguistics.
- Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: A case study.
- Chen Zhang, Xiao Liu, Jiuheng Lin, and Yansong Feng. 2024. Teaching large language models an unseen language on the fly. *arXiv preprint arXiv:2402.19167*.
- Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023b. Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 7915–7927, Singapore. Association for Computational Linguistics.

- Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llama beyond english: An empirical study on language capability transfer.
- 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.