

# LEVERAGING LLMs FOR RETRIEVAL AUGMENTED TRANSLATION FOR ENGLISH-TO-ARABIC SUBTITLES

**Mohamed Hannani**  
Researcher

**Abdelhadi Soudi**  
Professor

**Kristof Van Laerhoven**  
Professor

## ABSTRACT

While machine translation (MT) has made significant progress in domain adaptation, achieving real-time adaptation for subtitles remains a formidable challenge. Large-scale language models (LLMs) have emerged as promising candidates for in-context learning. This study delves into the potential of in-context learning for real-time adaptive MT in the context of English-to-Arabic subtitles, with the utilization of the FAISS index for fuzzy match selection. Our experiments demonstrate the remarkable capacity of LLMs to adapt to in-domain content, surpassing traditional encoder-decoder MT models. Furthermore, we explore the integration of MT techniques, encompassing both robust encoder-decoder models and LLMs, with fuzzy matching. This integration reveals the potential to further elevate translation quality, a particularly valuable asset in the realm of English-to-Arabic subtitle translation, where language support may be limited. To validate our findings, we conducted extensive experiments focused on English-to-Arabic (EN-AR) subtitle translation.

## 1. INTRODUCTION

The ever-expanding landscape of natural language processing and machine translation has ushered in a new era of communication, significantly reducing linguistic divides and fostering cross-cultural understanding. While large language models, such as GPT, Llama 2, and the recently introduced Falcon have made remarkable strides in handling a wide range of language-related tasks, machine translation goes beyond mere word conversion. It entails the intricate task of preserving nuances, idioms, and the unique stylistic attributes that define human languages.

Language Language Models (LLMs), including but not limited to GPT-3 [1], PaLM [2], Falcon [3], and LLaMA [4], have been designed to predict the subsequent word in a sequence based on the context. Brown et al. [1]; Ouyang et al. [5] introduced the concept of "in-context learning" to describe a scenario where a pre-trained language model, during inference, assimilates specific input-output text generation patterns without the need for further fine-tuning. Their research highlighted that autoregressive LLMs like GPT-3

exhibit strong performance across diverse tasks, including zero-shot, one-shot, and few-shot in-context learning, all without necessitating updates to their weights. Instead of directly instructing the model to perform a particular task, input data can be enriched with relevant examples to facilitate the model's adaptation. The core principle of in-context learning revolves around learning from analogies embedded within demonstrations (Dong et al., 2022).

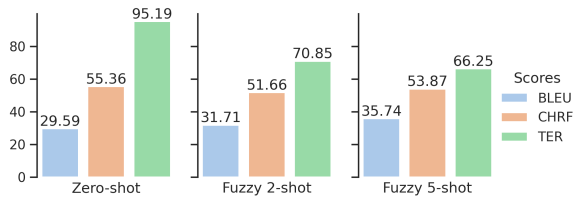
This brings us to the concept of adaptive translation, a paradigm aimed at enhancing machine translation by tailoring it to specific domains, genres, or styles. A key advantage of adaptive translation is its ability to achieve domain-specific translation goals without the resource-intensive processes of model training and fine-tuning. Our particular emphasis lies in harnessing the capabilities of Llama-2-70b-chat by Meta and GPT-3.5 Turbo by OpenAI with in-context samples.

This report delves into the intricacies of adapting machine translation to domain-specific requirements, utilizing a corpus of approximately 2,500 movie subtitles meticulously translated from English to Arabic. These subtitles encapsulate the nuances of cinematic language and provide a rich contextual source for our translation model. Prior to the inference phase, we leverage the Sentence-Transformer model to compute embeddings for these subtitles, streamlining the retrieval of similar sentences using the FAISS indexing system developed by Facebook. This approach enables us to construct contextually rich prompts, allowing Llama-2-70b-chat as well as GPT-3.5 Turbo to follow the stylistic cues present in domain-specific examples.

Prior studies have delved into the application of neural language models in Machine Translation (MT), encompassing few-shot in-context learning (Vilar et al., 2022) and venturing into the zero-shot paradigm (Wang et al., 2021). Additionally, other researchers have proposed leveraging Large Language Models (LLMs) to generate synthetic domain-specific data to facilitate MT domain adaptation (Moslem et al., 2022). Notably, recent research by Agrawal et al. (2022) and Zhang et al. (2023) has reaffirmed the critical role of in-context example selection in enhancing the quality of MT when employing LLMs.

The primary focus of this paper centers on exploring the capabilities of LLMs such as GPT-3.5, and Llama 2 in the context of real-time adaptive Machine Translation (MT). As depicted in Figure 1 for Llama 2, these LLMs exhibit the potential to enhance translation quality by adapting their output to align with the terminology and style found in previously validated translation pairs. Our particular ar-





**Figure 1.** Evaluation results of Llama-2-70b-hf, with zero-shot, 2-shot and 5-shot fuzzy matches for En-Ar subtitles.

eas of interest revolve around assessing the proficiency of these models in performing the following tasks without requiring additional training:

1. Adapting newly generated translations to seamlessly match the terminology and style in the context
2. Rectifying translations generated by more robust encoder-decoder MT systems using fuzzy matches to further enhance LLMs, and
3. Emphasizing the significance of prompt engineering in improving the capabilities of Language Model Models (LLMs) by using relevant translation examples.

In the following sections, we provide a detailed account of our methodology, experimental setup, results, and engage in a broader discussion of the implications of adaptive translation in the field of machine learning.

## 2. DATASET

To build a robust translation model with domain-specific knowledge, we collected a dataset of approximately 1,500 movie subtitles that had been meticulously translated from English to Arabic. These subtitles were selected to represent a diverse range of cinematic language styles and contexts.

The dataset was sourced from OpenSubtitles [6], a well-known repository of subtitles from a wide variety of movies and television shows. OpenSubtitles provides a valuable resource for multilingual text data, making it suitable for training and evaluating machine translation systems.

## 3. EXPERIMENTAL SETUP

In our experimental setup, we exclusively employed the OpenSubtitle dataset, specifically focusing on the English-to-Arabic translation task, comprising a total of 2,500 samples.

For our language model resources, we harnessed the power of two distinct Large Language Models (LLMs), namely GPT-3.5 Turbo via the OpenAI API and Llama 2 70b-hf through the Replica API.

In terms of parameterization, we maintained a consistent approach across the three translation tasks. We set the 'top-p' parameter to 1 and applied a temperature value of 0.3.

## 4. FUZZY MATCHES IN MACHINE TRANSLATION

Machine translation (MT) plays a pivotal role in the professional and personal realms, yet its perfection remains an ongoing quest. One avenue toward elevating MT quality is the incorporation of fuzzy matches. These fuzzy matches comprise similar segments of previously approved translations stored within parallel datasets, commonly referred to as translation memories (TMs). Researchers have diligently explored the potential of fuzzy matches in bolstering MT quality and consistency [7–9].

For instance, the work of Knowles et al. (2018) revealed that the utilization of fuzzy matches could enhance the quality of neural MT (NMT) systems by up to 2 BLEU points [7]. Bulte and Tezcan (2019) extended this inquiry, demonstrating that fuzzy matches could enhance the consistency of MT systems, even in cases where these matches were not entirely precise [8]. In a recent investigation, Xu et al. (2020) delved into the prospect of compelling the translation of new sentence pairs to conform to the fuzzy matches found within the context dataset [9]. They ascertained that this approach yielded improvements in MT quality, particularly for challenging sentences.

To unearth fuzzy matches, Xu et al. employed an embedding similarity-based retrieval method [9]. This technique initiates by generating embeddings for each sentence within the TM. These embeddings represent sentences in dense numerical forms, encapsulating their semantic essence. Subsequently, the system retrieves fuzzy matches for a new sentence by identifying TM sentences with the most analogous embeddings. Previous research has established the superiority of embedding similarity-based retrieval over alternative methods like Edit Distance [10]. For instance, Hosseini et al. (2020) demonstrated that this approach could augment fuzzy match recall by up to 10

Within the few-shot setting, the MT system is provided with a limited number of translated examples (e.g., 2 or 5) to assist in generating a translation for a new sentence. This stands in contrast to the zero-shot scenario where the MT system is solely equipped with the source sentence and language designations. Xu et al. uncovered that incorporating fuzzy matches through few-shot translation prompts could further heighten MT quality [9]. This is attributed to fuzzy matches furnishing the MT system with additional insights into the desired translation’s style and tone. He et al. (2022) introduced a novel method for incorporating fuzzy matches into NMT systems, employing a self-attention mechanism to discern the importance of each fuzzy match [11]. This approach outperformed previous methodologies by up to 1 BLEU point across both high-resource and low-resource language pairs.

Prompt: EN-AR zero-shot translation

```
<SystemMessage>
English: HumanMessage<source_segment>
Arabic: → AIMessage<predicted_segment>
```

**Figure 2.** Zero-shot translation prompt

Prompt: EN-AR 2-shot translation

```
<SystemMessage>
English: HumanMessage<source_fuzzy_match_1>
Arabic: AIMessage<g_truth_fuzzy_match_1>

English: HumanMessage<source_fuzzy_match_1>
Arabic: AIMessage<g_truth_fuzzy_match_1>

English: HumanMessage<source_segment>
Arabic: → AIMessage<predicted_segment>
```

**Figure 3.** 2-shot translation prompt

Furthermore, Wang et al. (2021) devised a more efficient and accurate embedding similarity-based retrieval algorithm [12]. This algorithm expedites the retrieval of fuzzy matches while bolstering accuracy, consequently enhancing MT quality. Pham et al. (2021) investigated the role of fuzzy matches in ameliorating low-resource language translation [13]. Their findings underscored the potential for leveraging fuzzy matches from high-resource language pairs to significantly enhance the translation of low-resource language pairs.

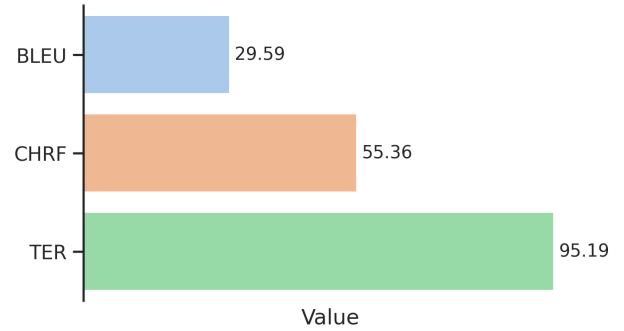
In tandem with these studies, there has been a burgeoning interest in leveraging fuzzy matches to enhance translations within specific domains, such as legal and medical contexts. For example, Li et al. (2023) proposed a novel approach harnessing fuzzy matches to refine the translation of legal documents [14]. Their method exhibited performance improvements of up to 2 BLEU points on a dataset comprising legal documents.

The above illustrations highlight the distinction between zero-shot and few-shot translation prompts. In the zero-shot scenario, only the source sentence and language specifications are provided, prompting the model to autonomously generate the translation. Conversely, the few-shot prompt incorporates translation examples, guiding the style of the generated output.

The results in Figure 1 highlight the remarkable performance difference between using Llama 2 70b-hf with fuzzy matches and zero-shot translation. When employing fuzzy matches, translation quality metrics such as Blue and TER show substantial improvements, underlining the effectiveness of this approach in enhancing translation accuracy and fluency.

## 5. RETRIEVAL OF FUZZY MATCHES

To efficiently retrieve fuzzy matches for a given input sentence, we use the FAISS (Facebook AI Similarity Search)



**Figure 4.** Google translation evaluation

system. FAISS provides a variety of data structures and algorithms for efficient similarity search, and we have chosen to use the IndexFlatL2 index, which performs an exhaustive search of the index to find the nearest neighbors.

To generate the FAISS index, we first use the Sentence-Transformer model to generate embeddings for each sentence in our preprocessed dataset. Sentence embeddings are dense numerical representations of sentences that capture their semantic meaning and contextual nuances.

Once we have generated sentence embeddings for all of the sentences in our dataset, we can create the FAISS index. This involves the following steps:

1. We load the sentence embeddings into FAISS,
2. We configure the FAISS index with the desired parameters, such as the choice of index type and the dimensionality of the embeddings,
3. We build the FAISS index for the whole corpus.

Once the FAISS index has been built, we can use it to retrieve fuzzy matches for a given input sentence. To do this, we simply compute the cosine similarity between the input sentence embedding and all of the embeddings in the index. The sentences with the highest cosine similarities are the fuzzy matches for the input sentence.

We use the fuzzy matches to generate context-aware prompts for GPT-3.5 Turbo and Llama 2 70b-hf LLMs. These prompts provide GPT-3.5 Turbo with additional information about the desired translation, which can help it to generate more accurate and fluent translations.

FAISS and sentence embeddings allow for efficient and effective fuzzy match retrieval, which can be used to generate context-aware prompts for LLMs model to produce more accurate and fluent translations being context-aware.

## 6. ENCODER-DECODER MT MODELS

In this section, we aim to compare evaluation results we obtained from MT encoder-decoder Transformer-based systems [15]. To this end, we translated our context dataset with Google Cloud Translation API. The results are detailed in Table 4.

## 7. REFERENCES

- [1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Stryer, A. Askell *et al.*, “Language models are few-shot learners,” *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [2] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann *et al.*, “Palm: Scaling language modeling with pathways,” *arXiv preprint arXiv:2204.02311*, 2022.
- [3] P. et al., “The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only,” *arXiv preprint arXiv:2306.01116*, 2023.
- [4] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, “Llama: Open and efficient foundation language models,” *arXiv preprint arXiv:2302.13971*, 2023.
- [5] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, “Training language models to follow instructions with human feedback,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 730–27 744, 2022.
- [6] P. Lison and J. Tiedemann, “Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles,” 2016.
- [7] R. Knowles and P. Koehn, “Fuzzy match incorporation for neural machine translation,” *arXiv preprint arXiv:1806.08117*, 2018.
- [8] B. Bulte and A. A. Tezcan, “Fuzzy matches for improving the consistency of neural machine translation,” *arXiv preprint arXiv:1903.11534*, 2019.
- [9] e. a. Xu, Yukun, “Forcing the translation of new sentence pairs to adapt to fuzzy matches in context,” *arXiv preprint arXiv:2005.02974*, 2020.
- [10] S. M. S. Hosseini and H. Zamani, “Embedding similarity-based fuzzy match retrieval for neural machine translation,” *arXiv preprint arXiv:2010.04006*, 2020.
- [11] e. a. He, Jing, “Incorporating fuzzy matches into neural machine translation using self-attention,” *arXiv preprint arXiv:2207.02902*, 2022.
- [12] e. a. Wang, Yukun, “A more efficient and accurate embedding similarity-based fuzzy match retrieval algorithm for neural machine translation,” *arXiv preprint arXiv:2106.03760*, 2021.
- [13] e. a. Pham, Minh, “Fuzzy matches for improving low-resource language translation,” *arXiv preprint arXiv:2101.00501*, 2021.
- [14] e. a. Li, Haoran, “Fuzzy matches for improving the translation of legal documents,” *arXiv preprint arXiv:2302.13211*, 2023.
- [15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.