



**School of Science and Engineering**

**Exploring Data Set Labels and Text Types and their Impact on the  
Quality of the Performance of Machine Learning Models in  
English to Arabic Translation.**

**Capstone Design Interim Report**

March 2024

**Omar Hamaini**

Supervised by:

**Dr. Yousra Chtouki**

## **Student conduct and copyright**

The supervisor and the student (the Capstoner) agree that:

1. Permission has been obtained for any third party content (eg Data, illustrations, photographs, charts or maps).
2. The results described in this report have not previously been published

**Student's name and signature:**

Omar Hamaini



**Supervisor's name and signature:**

Yousra Chtouki

Title: Exploring Data Set Labels and Text Types and their Impact on the Quality of the Performance of Machine Learning Models in English to Arabic Translation.

## Capstone Report

### **Student Statement:**

I, Omar Hamaini, applied ethics to the design process and in the selection of the final proposed design. I affirm that I have held the safety of the public to be paramount and has addressed this in the presented design wherever may be applicable.

---

Omar Hamaini



Approved by the Supervisor

---

Dr. Yousra Chtouki

## **ACKNOWLEDGEMENTS**

I want to extend my sincere appreciation to my supervisor, Dr. Yousra Chtouki, for her invaluable guidance and support throughout this semester's project. Her mentorship has not only enabled me to navigate new tools but also encouraged me to push beyond my comfort zone. I would also like to thank both of my parents as well as my friends for being supportive and an invaluable source of motivation.

# CONTENT

ACKNOWLEDGEMENTS .....	2
ABSTRACT (ENGLISH) .....	4
RESUME (FRENCH) .....	5
INTRODUCTION .....	6
PROBLEM STATEMENT .....	7
PROJECT SPECIFICATION.....	7
1 STEEPLE Analysis.....	8
1.1. Social Impact.....	8
1.2. Technological Impact .....	8
1.3. Economic Impact.....	8
1.4. Environmental Impact .....	8
1.5. Political Impact.....	8
1.6. Legal Impact .....	8
1.5. Ethical Impact.....	8
2 Engineering Standards.....	9
3 Logic Model Framework .....	9
3.1. Target Population.....	9
3.2. Underlying Assumptions.....	9
3.3. Ressources/Challenges .....	9
3.4. Activities.....	10
3.5. Outputs of the Project .....	10
3.6. Outcomes.....	10
4 Literature Review .....	11
5 Methodology and Capstone Design.....	11
5.1. Capstone Design and Approaches/Model/Algorithm/Application Development.....	11
5.1.1 Large Language Models.....	11
5.1.2 Machine Learning Analysis .....	12
5.1.3 Evaluation Metrics .....	12
5.1.4 Software.....	12
6 Methodology and Capstone Design.....	13
7 References .....	14
APPENDIX B: Conference Proposal.....	14

## **ABSTRACT (ENGLISH)**

This capstone project investigates the optimization of machine translation systems to improve the quality of English to Arabic translations as well as evaluating the nuances in translation quality depending on the labels of data sets. The project aims to address the scarcity of high-quality Arabic content on digital platforms, particularly Wikipedia, by Large Language Models. It conducts a thorough investigation to assess the performance of various LLMs, including GPT, BERT, T5, and mT5, in translating diverse text types from English to Arabic. The methodology involves the integration of LLMs into a translation tool, implemented using Flask, a lightweight web framework in Python due to it being a scalable solution for facilitating translation services. Evaluation metrics such as BLEU score, Cosine Similarity, and Latent Semantic Analysis (LSA) are also utilized to assess translation accuracy and quality. Additionally, exploratory data analysis techniques, including Principal Component Analysis (PCA) and K-means clustering, are employed in order to analyze translation data and identify patterns. The project's findings demonstrate the effectiveness of LLMs in capturing linguistic nuances and producing quality translations across different text types. The results of the project have the potential to enhance the accessibility and quality of Arabic content on digital platforms and therefore addressing the linguistic disparities that are quite prevalent in information that is spread online.

**Keywords:** LLM, Translation, Wikipedia, Artificial Intelligence, Flask

## **RESUME (FRENCH)**

Ce projet de fin d'études enquête sur l'optimisation des systèmes de traduction automatique pour améliorer la qualité des traductions de l'anglais vers l'arabe et évalue les nuances de qualité de traduction en fonction des libellés des ensembles de données. Le projet vise à remédier à la pénurie de contenu arabe de haute qualité sur les plateformes numériques, en particulier sur Wikipedia, grâce à des LLMs (Grand modèle de langage). Il mène une enquête approfondie pour évaluer les performances de divers LLMs, notamment GPT, BERT, T5 et mT5, dans la traduction de différents types de texte de l'anglais vers l'arabe. La méthodologie implique l'intégration des LLMs dans un outil de traduction, implémenté à l'aide de Flask, un Framework web léger en Python en raison de sa capacité à fournir des services de traduction évolutifs. Des métriques d'évaluation telles que BLEU Score, Cosine Similarity et LSA sont également utilisées pour évaluer l'exactitude et la qualité de la traduction. De plus, des techniques d'analyse exploratoire des données, notamment l'analyse PCA et K-means, sont employées pour analyser les données de traduction et identifier les motifs. Les résultats du projet démontrent l'efficacité des LLMs pour capturer les nuances linguistiques et produire des traductions de qualité dans différents types de texte. Les résultats du projet ont le potentiel d'améliorer l'accessibilité et la qualité du contenu arabe sur les plateformes numériques et donc de remédier aux disparités linguistiques qui sont assez prévalentes dans les informations diffusées en ligne.

**Mots clés:** LLM, Traduction, Wikipedia, AI, Flask

## **INTRODUCTION**

In a society driven by digital technology, the significance of language in bridging cultural divides cannot be overstated. This project, as an initiative, aims to address the language barriers between Arabic and English and leverages translation technologies. We first and foremost seek to empower Arabic speakers with greater access to information by enhancing the accessibility of knowledge currently available in English but not in Arabic. This way, mutual understanding among diverse linguistic communities such as Wikipedia for example is fostered. Therefore, addition of diverse content to the Arabic digital landscape is envisioned to promote clarity and inclusivity.



## **PROBLEM STATEMENT**

While English enjoys widespread availability of content and information online, Arabic still significantly lags behind in online presence. Traditional translation methods often fail to effectively handle the cultural nuances inherent in Arabic which exacerbates this linguistic gap and impacts many individuals. Previous research has highlighted the inadequacy of machine translation models such as Google Translate, Turjuman, and MarianMT in consistently providing accurate translations from English to Arabic. This inconsistency underscores the need for a more robust approach to translation. Therefore, the objective of this capstone project is to evaluate the accuracy and effectiveness of modern translation models, particularly Large Language Models (LLMs), in translating English content to Arabic. Additionally, the project seeks to investigate the impact of labeling data sets and categorizing text types on the performance quality of these models. The project aims to identify suitable translation solutions that can bridge the linguistic gap between English and Arabic online content and improve accessibility as well as quality especially for Arabic speaking users.

## **PROJECT SPECIFICATION**

This project focuses on investigating the influence of diverse text types and data set labels on the efficacy and performance quality of machine translation models, mainly LLMs. Our objectives include categorizing various text genres, assessing their impact on model performance and gaining insights into the adaptability of machine translation systems. The methodology used also involves data preprocessing and performance evaluation using established metrics.

# **1 STEEPLE Analysis**

## **1.1. Social Impact**

Any Arabic speaker would benefit from this project as we are providing them with content that is accurately translated from English to Arabic. The tool will also provide those who create content in English with the ability to translate their content to Arabic instead of manually rewriting everything from scratch, which can also further be provided to users.

## **1.2. Technological Impact**

The project goes through and shows the development of machine translation technology starting its conceptual phase and ending with the testing phase and offers insights on how the nature of the written text influences the performance of translation models. It also further demonstrates the strengths and weaknesses of the models.

## **1.3. Economic Impact**

The project will give people in rural areas with low income and who do not have the financial means or the ability to study English to access the same kind of information that was originally written in English in Arabic instead, thus bridging that gap between economic classes.

## **1.4. Environmental Impact**

Considering the environmental impact of data collection methods can influence the project's choice of practices that are environmentally sustainable.

## **1.5. Political Impact**

This project has the potential to contribute to diplomatic efforts and cross-border collaborations.

## **1.6. Legal Impact**

The project complies with the legal requirements relevant to data usage as well as privacy and operates under full transparency with its users.

## **1.5. Ethical Impact**

Informing the users about the accuracy level of different translation models as well as how text types can affect their performance is a requirement.

## **2 Engineering Standards**

### **IEEE Standard 802.11:**

This standard is crucial as the project involves wireless communication. It ensures that the wireless data transfer, especially the kind that is involved in data acquisition, follows a widely accepted protocol.

### **ISO/IEC 27001:2022 Standard:**

This is an international standard for information security management systems. It is crucial for this project as it involves the need for securing sensitive information, data privacy and confidentiality.

### **ISO Standard 639-1/639-2:**

These are international standards that provide language codes. They ensure a standardized representation and identification of languages. In the context of my project, it helps maintain consistency in handling multilingual datasets.

## **3 Logic Model Framework**

### **3.1. Target Population**

First, users who want to exchange information efficiently and without worrying about language barriers as well as people from Arabophone countries who are not familiar with English. Second, academic researchers who prefer to rely on resources that are written in Arabic due to it being their mother tongue. Finally, linguists or other language enthusiasts whose professions consist of using English and/or Arabic in their works.

### **3.2. Underlying Assumptions**

- \* We expect the users to be familiar with using web-based applications.
- \* Categorizing text types and using LLMs instead of transformer models can provide more accurate and consistent translations.

### **3.3. Ressources/Challenges**

#### **Resources:**

- \* Available datasets in specific domains.
- \* Large Language Models.
- \* The Python framework Flask.

- \* PostgreSQL.
- \* Research articles and papers on LLMs and English to Arabic machine translation.
- \* The architecture of LLMs as well as their documentations.

### **Challenges:**

- \* Arabic has a complex structure which can create translation difficulties.
- \* Using LLMs can require additional computing resources.
- \* The performance of the models will rely on the quality of the datasets and the categorization of the data labels.
- \* There is a lack in translated featured articles on Wikipedia from English to Arabic.

### **3.4. Activities**

- \* Categorization of text types in chosen datasets.
- \* Integration of LLMs into a web-based Flask application.
- \* Routes configuration and testing.
- \* Implementation of BLEU score, LSA and Cosine Similarity.
- \* The scores being stored in an SQL table that connects to the Flask application.
- \* Descriptive analysis, PCA and K-means clustering.

### **3.5. Outputs of the Project**

A web-based Flask application that uses LLMs to perform translation of Wikipedia articles from English to Arabic, as well as storing the scores from the aforementioned metrics.

### **3.6. Outcomes**

- \* The results will indicate whether manually categorizing text types will affect the accuracy of the translations.
- \* The project will improve access to information by providing fully translated articles from English to Arabic.
- \* The project may provide future guidance and/or insights for research in the field of machine translation.

## 4 Literature Review

Machine translation, particularly from English to Arabic, presents significant challenges due to linguistic and cultural differences between the two languages. Transformer models have recently emerged as effective tools for machine translation tasks. Current transformer models, such as Turjuman and Google Translate, have shown limitations in producing accurate and high-quality translations. Some studies have identified several challenges with transformer-based translation models. One significant limitation is the inconsistency in translation accuracy across models. While some models, such as Turjuman, provide relatively consistent results, others, such as Google Translate, vary in quality. Another challenge is the inability of current transformer models to accurately capture the complexities of English to Arabic translation. Despite advances in neural machine translation, these models frequently struggle to maintain the semantic and syntactic structures of the original text and result in mistranslations and inaccuracies. To address the limitations of traditional transformer models, recent research has focused on the potential of LLMs to improve translation quality. One significant advantage of LLMs is their ability to capture complex linguistic patterns and contextual dependencies. They can effectively learn the nuances of English and Arabic by using large-scale pre-training on massive amounts of text data thus improving translation performance. By focusing on specific domains, such as medical or legal translation, researchers hope to create tailored solutions that address each domain's distinct linguistic and contextual requirements.

## 5 Methodology and Capstone Design

### 5.1. Capstone Design and Approaches/Model/Algorithm/Application Development

#### 5.1.1 Large Language Models

**-GPT (Generative Pre-trained Transformer):** It is known for its generative capabilities and has been successfully applied to a variety of NLP tasks including translation.

**-BERT (Bidirectional Encoder Representations from Transformers):** Its models use bidirectional attention mechanisms to capture contextual information from both the left and right contexts and makes them suitable for translation tasks.

**-T5 (Text-To-Text Transfer Transformer):** It represents NLP tasks as text-to-text transformations and allows them to handle translation tasks with ease.

**-mT5 (Multilingual T5):** An extension of T5, mT5 is trained on multilingual data and can translate between multiple languages such as English and Arabic.

### 5.1.2 Machine Learning Analysis

**-Dimensionality Reduction:** Techniques such Principal Component Analysis (PCA) can be used to reduce the dimensionality of LLM embeddings and makes them easier to visualize and analyze. It can help identify the most informative features or components in LLM embeddings and allow for more efficient downstream processing and interpretation of results.

**-Clustering:** Algorithms such as K-means clustering can be used to group similar translations or documents which facilitates the organization and analysis of translation results. These techniques aid in the identification of patterns and similarities in translation data and thereby provide insights into LLM effectiveness and translation quality.

### 5.1.3 Evaluation Metrics

**-The BLEU Score (Bilingual Evaluation Understudy):** a commonly used metric for comparing the quality of machine-translated text to reference translations.

**-Cosine Similarity:** Determines the similarity of two text vectors based on the cosine of their angle and reveals semantic similarity in translations.

**-Latent Semantic Analysis (LSA):** It is a technique for analyzing relationships between documents and the terms they contain by generating a set of related concepts.

### 5.1.4 Software

**-Flask:** This is a Python web framework that is both lightweight and versatile and makes it ideal for creating web applications and APIs. It is the translation service's backbone and can provide infrastructure for handling HTTP requests, processing translation requests and serving translated content to users.

**-PostgreSQL:** An open-source relational database management system that is renowned for its dependability and robustness. It is used to store translation scores, metadata, and other Flask-related information in this context.

**-HuggingFace's Transformers Library:** It provides an easy-to-use interface for working with NLP models including pre-trained LLMs. It loads the former, fine-tunes it on specific datasets and uses it

for various NLP tasks including text generation, sentiment analysis and translation among other tasks.

**-Python:** It is a flexible programming language for web development, data analysis, and machine learning. It is the primary programming language for developing the translation service which allows for integration with other tools and frameworks.

**-TensorFlow:** A deep learning framework for developing and training neural network models. It makes it easier to create, train, and deploy machine learning models for translation tasks. It is widely used in the industry and provides scalable solutions for model development and deployment.

## 6 Methodology and Capstone Design

**Wikipedia content in English:** The core dataset consists of a collection of English Wikipedia articles on various topics. To ensure accuracy, only articles with carefully translated versions were used in this project. This selection process sought to ensure that the translations retained the original meaning and quality of the English articles.

**Pre-trained LLMs:** The tools that will be used are pre-trained LLMs, namely GPT, BERT, T5, Mt5 and possibly Llama 2. They will be thoroughly imported into a web-based Flask application and connected to a PostgreSQL database.

**Storing records into Database:** The three implemented metrics for evaluating translation accuracy and quality of the LLMs are BLEUScore, Cosine Similarity and LSA. The scores will be saved into a PostgreSQL database mainly in order to make data retrieval and inspection easier.

**Exploratory Data Analysis:** As mentioned before, we will be use PCA and K-means clustering:

**-PCA** is a popular data analysis method that simplifies datasets. By reducing the number of parameters, PCA helps to preserve the essence of the data while providing a more concise representation. This technique is especially useful in tasks where the original dataset contains a large number of variables and allows analysts to identify and focus on the most important components.

**-K-Means Clustering** is an unsupervised learning algorithm that groups items into clusters based on shared characteristics. K-means works by iteratively assigning data points to clusters and optimizing cluster centroids to group similar items together while minimizing differences between clusters. This approach is known for making it easier to find inherent patterns and structures in data.

## 7 References

- Alkabi, M. N., Hailat, T. M., Alshawakfa, E. M., & Alsmadi, I. M. (2013b, November 1). *Evaluating English to Arabic Machine Translation Using BLEU*. Citeseerx. <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=84ff4f0e3b53503490950f29fa4497c8aba7f158#page=77>
- Arxiv. (n.d.-a). <https://arxiv.org/pdf/2302.14520.pdf>
- Introduction. NLP Course. (n.d.). <https://huggingface.co/learn/nlp-course/chapter1/1>
- Introduction to tensorflow. TensorFlow. (n.d.). <https://www.tensorflow.org/learn?hl=en>
- ISO/IEC 27001:2022. ISO. (2022, October 25). <https://www.iso.org/standard/27001>
- Language Codes. Codes for the representation of names of languages (Library of Congress). (n.d.). [https://www.loc.gov/standards/iso639-2/php/langcodes\\_name.php?iso\\_639\\_1=en](https://www.loc.gov/standards/iso639-2/php/langcodes_name.php?iso_639_1=en)

## APPENDIX B: Conference Proposal

**Title:** Bridging Language Gaps: Optimizing Machine Learning Models for Enhanced Arabic Wikipedia Content Generation.

**Selected topics:** Natural Language Processing, Artificial Intelligence, Machine Learning. Chosen








**Keywords:** LLM, Machine Learning, Artificial Intelligence, Wikipedia, English to Arabic Translation, Flask.

**Abstract:** With the growing demand for accurate and fluent translation between English and Arabic due to the lack of content in Arabic on Wikipedia that is equivalent in quality to its English counterpart, the development of effective machine translation systems has become increasingly essential. In this capstone project, we present a comprehensive investigation into the performance of machine learning models for English to Arabic translation and focus on the integration of Large Language Models in order to improve translation quality. The idea of the project began with a critical analysis of existing transformer-based translation models and revealed significant limitations in translation accuracy and fluency. Drawing from these insights, the utilization of state-of-the-art LLMs, such as GPT and Llama 2 as well as the categorization of data set labels was proposed to overcome these challenges and enhance translation performance. Experiments are conducted to evaluate the efficacy of LLMs in English to Arabic translation tasks by considering various metrics including BLEU score, Cosine Similarity, and Latent Semantic Analysis. We demonstrate the superior capabilities of LLMs in capturing linguistic nuances and producing high-quality translations. We also discuss strategies for integrating LLMs into Flask-



based applications, configuring routes, and implementing evaluation metrics for assessing translation quality. This project contributes to advancing machine translation for English to Arabic and offering insights into the efficacy of LLMs and domain specific modeling approaches. Findings from the project can also pave the way for the development of reliable translation systems capable of meeting diverse cultural needs.

#### Receipt of Submission:

 Edas Help <help@edas.info>

To: Omar Hamaini <77018> Wed 2/28/2024 9:09 PM

Dear Mr. Omar Hamaini:

Thank you for registering your paper 1571002773 (*Bridging Language Gaps: Optimizing Machine Learning Models for Enhanced Arabic Wikipedia Content Generation*) to **2024 IEEE International Conference on Industry 4.0, Artificial Intelligence and Communications Technology (IAICT)**. You still have to upload your manuscript at [1571002773](https://edas.info/index.php?c=31880). Your manuscript can be .

You can see all your submissions and their status at <https://edas.info/index.php?c=31880> using your EDAS user id **O.hamaini@au.ma**.

Once you upload your manuscript, you will receive another email confirmation.

Regards,  
The conference chairs