Exploring Translation Models: A Comparative Study of Seq2Seq and LLM Architectures for English-Italian Machine Translation

Pasquale Gravante - 896983
Angelo Giuseppe Limone - 903441
Antonio Mastroianni - 898723

January 5, 2025

Contents

1 Introduction

1.1 Problem Statement

Machine translation (MT) has become an indispensable tool for facilitating communication in a globalized world. Deep learning techniques, particularly the use of transformer-based models, have revolutionized the field by enabling the development of highly accurate and versatile translation systems. Despite these advancements, questions remain about the comparative advantages of different training paradigms. Specifically, it is unclear how models pre-trained on MT tasks differ in performance from those pre-trained on general tasks or trained entirely from scratch.

This study investigates these paradigms by comparing several approaches to English-to-Italian MT, including models pre-trained specifically for translation, models pre-trained on diverse tasks, and models trained without any pre-training. Additionally, we evaluate the impact of architectural modifications, such as increasing model dimensionality and vocabulary size, as well as the effectiveness of fine-tuning large language models (LLMs) with lightweight techniques like low-rank adaptation (LoRA).

1.2 Objectives and Research Questions

The primary objective of this project is to compare and contrast the performance of various model types and training strategies for MT. This involves evaluating pre-trained Seq2Seq models, general-purpose pre-trained models, and non-pre-trained models. Special attention is given to exploring whether increasing model complexity can offset the absence of pre-training and to understanding how LoRA fine-tuning affects LLM performance.

Key research questions addressed in this study include:

- How do models pre-trained for MT-specific tasks perform relative to general-purpose pre-trained models on translation tasks?
- Can training Seq2Seq models from scratch achieve competitive performance, particularly when architectural capacity is increased?
- Does fine-tuning pre-trained LLMs with LoRA improve MT performance significantly?
- How do non-pre-trained LLMs compare to their pre-trained counterparts in terms of translation quality?

1.3 Literature Review

Pre-trained models have proven highly effective in MT, particularly those trained on large-scale parallel corpora like the Helsinki-NLP OPUS-MT models. These systems excel in leveraging task-specific data to achieve robust performance. General-purpose models, such as T5-small, represent another avenue, demonstrating versatility across a range of downstream tasks, including MT, though typically requiring task-specific fine-tuning.

Training models from scratch remains relevant in cases where pre-trained resources are unavailable or insufficiently specialized. However, this approach often necessitates large datasets and prolonged training times to reach acceptable performance levels. Recent innovations in LLMs, such as Llama and Qwen, further complicate the landscape. These models, when fine-tuned using techniques like LoRA, offer efficient pathways to adapt large-scale architectures for specific tasks, reducing computational costs without sacrificing quality.

Finally, modifications to model architecture, including increased dimensionality and vocabulary size, have shown potential for improving capacity and translation quality. However, such enhancements come with significant computational trade-offs, underscoring the importance of balancing complexity and practicality.

1.4 Structure of the Report

This report is organized into several chapters. Chapter 2 presents the exploratory data analysis (EDA), offering insights into the datasets, preprocessing steps, and observed trends. Chapter 3 outlines the frameworks used, describing the architectures and training methodologies for pre-trained Seq2Seq models, non-pre-trained Seq2Seq models, and LLMs. In Chapter 4, the experiments and results are detailed, covering evaluation metrics, quantitative findings, and qualitative analyses of translation outputs. Chapter 5 discusses the results, highlighting key insights and addressing limitations. The final chapter, Chapter 6, concludes the study by summarizing the findings and proposing directions for future research.

2 Exploratory Data Analysis (EDA)

2.1 Dataset Overview: TED2013

The **TED2013 corpus** is a parallel dataset consisting of TED talk subtitles provided by **CASMACAT** (available here). The original data comes from WIT3 and is widely used for machine translation research and benchmarking tasks. The TED2013 dataset contains multilingual subtitles aligned across English and target languages such as Italian, French, German, Spanish, and many others.

> J. Tiedemann, 2012, *Parallel Data, Tools and Interfaces in OPUS.* In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012).

2.1.1 Dataset Language Pair (English-Italian)

The statistics below pertain specifically to the **English-Italian (en-it)** subset:

- **Total Sentences:** 159,391
- **Source Language Tokens (English):** 2,670,718
- **Target Language Tokens (Italian):** 2,508,941

These numbers demonstrate a robust bilingual alignment between English and Italian sentences, which can be leveraged for training machine translation models or conducting cross-linguistic analysis.

2.1.2 Dataset Characteristics

The TED2013 dataset provides multilingual parallel data with well-structured sentence alignments, typically used for tasks like machine translation, linguistic studies, and benchmarking pre-trained translation models. It consists of:

- **Bilingual Sentence Pairs:** The data consists of aligned English-Italian sentences extracted from TED talks and subtitles.
- **Domain-Specific Content:** The corpus focuses on technical and spoken communication topics related to TED talks, providing diverse content domains.

- **Preprocessing Options:** Tokenization and normalization are essential preprocessing steps to standardize sentence structures and align data properly for model training.

The dataset also includes sentence fragments, which can lead to domain-specific patterns and linguistic variability.

2.2 Data Preprocessing

Preprocessing is a critical step in preparing datasets for machine translation tasks. For this work, the TED2013 English-Italian dataset underwent several preprocessing steps to ensure data consistency and compatibility with the models. These steps are essential to standardize the input data, reduce noise, and improve training performance.

2.2.1 Tokenization

Tokenization was performed to split the text into subword units. Different tokenizers were employed based on the specific model architecture:

- **Marian Tokenizer:** Used for the Marian MT Helsinki model, this tokenizer is designed for efficient subword segmentation and is tailored to work seamlessly with Marian models.
- **T5 Tokenizer:** The T5 model utilizes a SentencePiece-based tokenizer, which encodes text into subwords based on statistical patterns in the dataset. This tokenizer was applied when training and evaluating the T5 models.
- **Custom Tokenizer:** For the models trained from scratch, a custom SentencePiece tokenizer was used. It was trained on the TED2013 dataset to create a vocabulary specific to the training corpus, ensuring optimal subword segmentation.
- **Tokenizer for LLama Instruct:** The tokenizer for the `LLaMA Instruct` model is based on the SentencePiece library and employs a subword tokenization approach, which splits text into smaller units to handle rare or unknown words effectively. This tokenizer is trained on the same multilingual dataset as the model, ensuring alignment with the linguistic structures found in the data and enhancing performance across diverse languages and domains. The preprocessing pipeline of the tokenizer normalizes input text, applies byte-pair encoding (BPE), and

outputs tokenized sequences tailored for the LLaMA Instruct model, ensuring that input data is consistently formatted for both inference and fine-tuning. This robust design makes the tokenizer an essential component in bridging raw textual input and the model's capabilities for nuanced language understanding.

2.2.2 Normalization

Normalization was conducted to standardize the text. This included:

- Lowercasing all text to ensure consistency.
- Removing extraneous whitespace, punctuation, and special characters that could introduce noise.
- Standardizing formatting for common patterns, such as dates and numbers.

2.2.3 Sentence Filtering

The dataset contains both full sentences and fragments. To improve model training:

- Sentences that were excessively short (e.g., fewer than three tokens) or overly long (e.g., more than 96 tokens) were removed to avoid noisy or problematic data.
- Pairs with significant length mismatches between the source and target sentences were discarded to improve alignment quality.

2.3 Visual Insights

In this section, we provide a detailed exploration of the dataset using some visualizations. This includes exploring the most common words and sentences as well as their distributions in the source (English) and target (Italian) datasets.

2.3.1 Most Common Words

(a) Most common words in English

(b) Most common words in Italian.

Figure 1: Comparison of the most common words in the English and Italian datasets.

The word cloud in Figure 1a highlights the most frequent words in the English dataset. The prominence of words like *people*, *one*, *know*, and *how* indicates a focus on abstract concepts and interpersonal discussions. This distribution suggests that TED talks often involve explanations and explorations of human-centric ideas.
For Italian, Words like *e* (and), *che* (that), *questo* (this), and *ma* (but) dominate, reflecting the structure of Italian discourse, which often relies on conjunctions and demonstratives. The frequent use of these words is indicative of descriptive and explanatory sentence structures in the translated texts.

2.3.2 Sentence Length Distribution

In Figure 2, we observe the sentence length distributions for both the source (English) and target (Italian) datasets. The majority of sentences fall within a length of 10 to 20 words, with a slight skew towards shorter sentences in the Italian dataset. This reflects the nature of translations, where Italian translations tend to be more concise than their English counterparts due to linguistic differences. Longer sentences in both datasets are less frequent but often correspond to more complex or compound sentences.

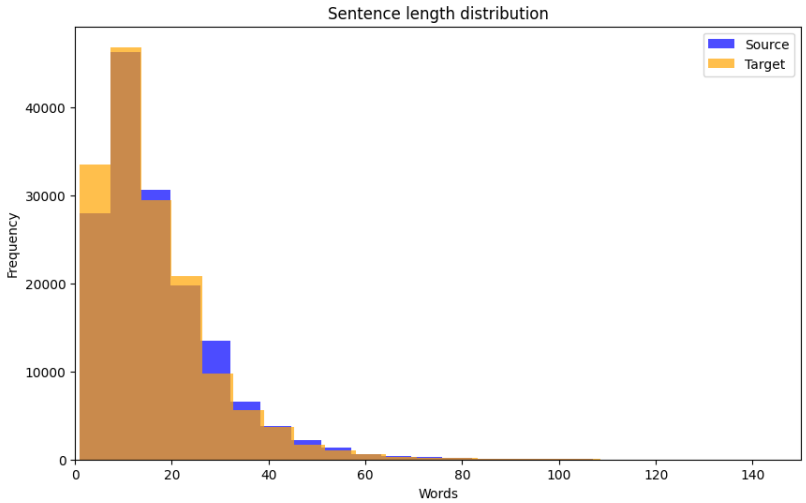

Figure 2: Sentence length distribution for English and Italian datasets.

2.3.3 Source - Target Words Ratio

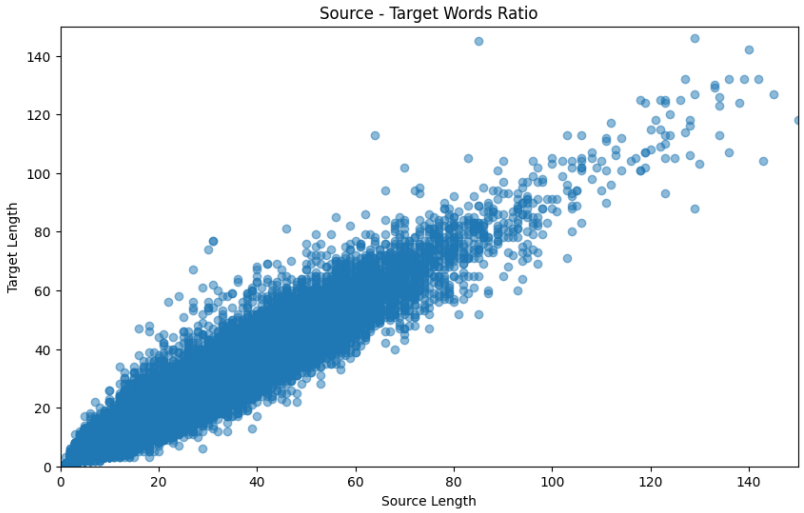

Figure 3: Source - Target Words Ratio for English and Italian datasets.

In Figure 3 we can observe that, except for some rare cases, the majority of the data exhibits approximatively the same length in words in both the English and Italian versions.

3 Frameworks and Models

This chapter details the different machine translation frameworks and models employed in this study, categorized into three main groups based on their pre-training and architectural configurations.

3.1 Seq2seq

3.1.1 Seq2Seq Pre-trained on Machine Translation Tasks

For this category, we utilized the Helsinki-NLP OPUS-MT English-to-Italian model. This model is part of the OPUS-MT project, which offers a suite of transformer-based machine translation systems pre-trained on extensive parallel corpora. The Helsinki-NLP model is a classic encoder-decoder architecture specifically pre-trained for translation tasks.

Model Configuration: The model employs a transformer architecture with 6 encoder and decoder layers, a hidden dimensionality of 512, and 8 attention heads. The vocabulary size is tailored for machine translation, incorporating subword tokens derived from joint byte pair encoding (BPE). This configuration provides the model with sufficient capacity to learn translation-specific nuances.

Training and Fine-tuning: We fine-tuned the model on an English-to-Italian dataset to adapt its pre-trained knowledge to our specific task domain. Training hyperparameters included a learning rate of 3×10^{-4}, batch size of 64, and 10 epochs with early stopping based on validation performance. The model achieved robust results out-of-the-box due to its pre-trained MT-specific representations.

3.1.2 Seq2Seq Pre-trained on General Tasks

The second framework utilized a general-purpose Seq2Seq model, T5-small, which is not pre-trained specifically for MT but rather for diverse text-to-text tasks. This choice allows us to evaluate the adaptability of such models to translation tasks through fine-tuning.

Model Configuration: T5-small employs the same encoder-decoder structure but with a focus on general linguistic knowledge rather than task-specific training. The model consists of 6 layers each for the encoder and decoder, a model dimensionality of 512, and 8 attention heads. The vocabulary size is

approximately 32,000 subword tokens, covering a wide range of general-use cases.

Training and Fine-tuning: The model was fine-tuned using the same English-to-Italian dataset as the OPUS-MT model. Hyperparameters included a learning rate of 3×10^{-4}, a batch size of 64, and 10 epochs. Unlike the OPUS-MT model, T5-small required more training iterations to converge due to its general-purpose pre-training. However, the results highlight its adaptability to translation when provided with sufficient task-specific data.

3.1.3 Seq2Seq Non-Pre-trained Model

To understand the impact of pre-training, we trained a Seq2Seq model entirely from scratch. This model matched the T5-small architecture in terms of dimensionality and layer configuration, allowing for a direct comparison.

Model Configuration: The scratch model shared the same architecture as T5-small but was initialized with random weights. It included 6 encoder and decoder layers, a hidden dimensionality of 512, and 8 attention heads. The vocabulary was based on a similar BPE tokenization scheme, yielding approximately 32,000 tokens. Additionally, an experimental variation with increased dimensionality (1024 hidden units) and an expanded vocabulary size of 64,000 tokens was tested to explore the potential of larger architectures.

Training: Training the non-pre-trained model was significantly more resource-intensive. The model required 15 epochs and a batch size of 64 with gradient accumulation steps set to 2 to simulate larger effective batches. Despite these efforts, the lack of pre-trained knowledge led to slower convergence and higher susceptibility to overfitting, particularly for the expanded architecture. While the increased model size showed marginal improvements, it also demanded significantly more computational resources.

3.2 Large Language Models

The second architecture taken into account is the Large Language Model (LLM). LLMs have demonstrated impressive performances in handling complex linguistic tasks, such as machine translation. What makes these results possible is their transformer-based architecture, which allows us to analyze

entire sentences and understand their context. Training them on diverse multilingual corpora enhances their ability to identify cross-linguistic patterns, making them highly effective for machine translation tasks. These advantages, combined with the potential for continuous improvement through updated training data and methods, position LLMs as a pivotal tool in modern machine translation, despite challenges such as computational demands and training time. The model we have selected is the `Llama-3.2-1B-Instruct` stands out for its optimized architecture, which balances computational efficiency and generalization capabilities. This makes it particularly well-suited for training on limited hardware, a key consideration in our case. It is an autoregressive model and it is comprised of the decoding component only. Hence, it will require a distinct approach to training compared to traditional machine translation models.

3.2.1 Pre-trained LLM

Model Configuration: We split the sampled dataset (50% of the original one) in three different datasets: train, test and validation. The first one has more than 50000+ examples (80%), test has around 15000 examples (15%), while validation has more or less 7000 examples (the 5% of sampled dataset). The `LLaMA-3.2-1B-Instruct` model is designed for instruction-following tasks. With 1 billion parameters, the model balances computational efficiency and performance. The architecture uses transformer blocks with multi-head attention mechanisms and feed-forward networks, optimized for causal language modeling tasks.

The key configurations are a hidden size of 2048, 16 attention heads, and 24 transformer layers. Positional embeddings are incorporated to maintain sequence ordering, and the activation function employed is GELU for better gradient flow and convergence. The model uses pre-normalization with layer normalization applied before attention and feed-forward layers, promoting stable training.

Training and Fine-tuning (LoRA): For the training of our pre-trained LLM, we need to talk about LoRA (Low-Rank Adaptation). It has emerged as an efficient and scalable method for adapting pretrained models to specific tasks. By allowing selective parameter adjustments, LoRA minimizes computational overhead while maintaining high performance. Through the optimization of a small subset of parameters, LoRA significantly reduces memory usage and computational demand. The low-rank adaptations can be saved also as light modules, so it is easy to switch between different con-

figurations already fine-tuned.

LoRA offers several adjustable parameters that determine its performance and resource efficiency. We have different parameters to adjust. The rank determines the dimensionality of the low-rank matrices added to the model, while the alpha scales the impact of the low-rank adaptations. It balances the influence of new parameters against the pretrained model. The dropout can be introduced during fine-tuning to prevent overfitting. At the same time, LoRA allows users to specify which layers of the model to adapt. Fewer layers adapted mean faster fine-tuning but may reduce task-specific performance. Clearly the size and quality of the training dataset influence the effectiveness of LoRA fine-tuning. Smaller datasets may require lower ranks and careful regularization. Finally the learning reate controls the speed at which the low-rank matrices are optimized. Task-specific tuning is required for best results.

The advantages of LoRA are reduction in computational and storage requirements, an easy integration with existing pretrained models and a reusable adaptation. The limitations are that may not fully leverage the potential of large datasets if the rank is too low and It requires careful tuning of parameters to achieve optimal performance. We have established that r and alpha have a value of 16, we set the dropout to 0.1, while the other parameters have not been set.

Considerations: The model shows a decrease in both training and validation loss during the steps, so we can tell that we have a good learning for our model. The validation loss is lightly higher than the training loss, so we don't have problems about overfitting and underfitting.

3.2.2 Non-pre-trained LLMs

Since we decided to train from scratch a Causal Language Model for a machine translation task, a specific task had to be designed in order for the model to translate sentences. The chosen task structure is the following:

```
<START_SYMBOL_source> source sentence <END_SYMBOL_SOURCE>
<START_SYMBOL_TARGET> target sentence <END_SYMBOL_TARGET>
```

Training a LLM completely from scratch would be a challenging and expensive task, both from a computational and a time point of view. Thus, in order to achieve this goal, we had to resort to some strategies in order to make the task feasible. Hence, we firstly reduced the size of the input data

to a fraction of the initial size. Moreover, the configuration of the model without pretrained weights wasn't considered in its entirety as, due to the high amount of trainable parameters, part of it has been pruned.

Model Configuration: `Llama-3.2-1B-Instruct` is designed as an instruction-tuned language model, optimized for conversational and task-specific guidance. It was pruned to only consider the first six layers and slightly over 600M trainable parameters.
Two attempts of training were made in order to understand how the results would change according to different amounts of data given as input. The first attempt considered only 15% of the data and lasted 2:30 hours wile the latter considered 50% of the data and lasted 12:00 hours.

Step	Training Loss	Validation Loss
100	10.428000	8.675980
200	7.374600	6.499942
300	5.742600	5.239958

Table 1: First model training with 15% of the data

Step	Training Loss	Validation Loss
500	6.515500	4.321095
1000	3.616100	3.922055
1500	2.464800	4.158695
2000	1.561500	4.387391

Table 2: Second model training of 50% of the data

Considerations: The first model shows a steady decrease in both training and validation loss over 300 steps, indicating effective learning. However, the validation loss remains higher than the training loss, suggesting possible underfitting due to the limited training data. Model 2 instead starts with higher loss values than first model but quickly converges to lower training losses by step 2000. However, the validation loss plateaus and slight increases, suggest overfitting as the model may have memorized the training data.
Model 2 likely experiences overfitting due to the mismatch between its large number of parameters (600 million) and the relatively small training dataset (76,768 observations). This imbalance allows the model to memorize the training data rather than generalize effectively to unseen examples, as indicated by the divergence between training and validation loss, where the

latter plateaus or slightly increases despite continued reductions in training loss. Training for 10 epochs amplifies this issue, providing high opportunity for memorization. High-capacity models like Model B typically require larger datasets or pretraining on extensive corpora to generalize effectively. Without these, the model's performance is constrained by its tendency to overfit.

4 Experiments and Results

4.1 Evaluation Metrics

This chapter discusses the metrics used to evaluate the quality of the translations generated by the different models. Each metric provides a unique perspective on translation performance, ranging from n-gram overlaps to semantic adequacy.

4.1.1 ROUGE-L

ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation) is a metric originally developed for summarization tasks but also applicable to translation. It calculates the longest common subsequence (LCS) between the generated translation and the reference translation. By focusing on LCS, ROUGE-L captures sentence-level fluency and coherence.

Key Characteristics:

- **Precision and Recall**: ROUGE-L considers both precision (how much of the generated text matches the reference) and recall (how much of the reference is covered by the generated text).
- **Flexibility**: The LCS approach ensures the metric is not overly strict about word order, tolerating minor deviations in phrasing.

Limitations: While useful for capturing general sentence structure, ROUGE-L does not account for semantic meaning, which limits its ability to measure adequacy in translations.

4.1.2 BLEU

BLEU (Bilingual Evaluation Understudy) is one of the most widely used metrics for evaluating machine translation systems. It measures the n-gram overlap between the generated translation and one or more reference translations.

Key Characteristics:

- **N-Gram Matching**: BLEU evaluates n-gram precision (unigrams, bigrams, etc.), emphasizing exact matches in short sequences.

- **Brevity Penalty**: To avoid favoring overly short translations, BLEU incorporates a brevity penalty when the generated translation is much shorter than the reference.
- **Multi-Reference Capability**: BLEU can compare the generated translation against multiple references, increasing its robustness to valid paraphrases.

Limitations: BLEU struggles to evaluate semantic equivalence and often penalizes creative or valid translations that do not closely match the references. It is also sensitive to minor variations in tokenization.

4.1.3 COMET

COMET (Cross-lingual Optimized Metric for Evaluation of Translation) is a modern metric that leverages neural networks to evaluate translation quality. Unlike traditional metrics, COMET goes beyond surface-level token matching and focuses on semantic alignment.

Key Characteristics:

- **Semantic Representation**: COMET uses pre-trained language models to encode both the source and reference sentences, allowing it to assess the semantic adequacy of the generated translation.
- **Human-Like Scoring**: By correlating strongly with human judgments, COMET provides a more accurate evaluation of translation quality compared to n-gram-based metrics.
- **Multi-Dimensional Analysis**: COMET evaluates adequacy (semantic similarity to the source) and fluency (grammatical correctness in the target language).

Limitations: COMET requires significant computational resources and can be sensitive to the quality of the language model used for evaluation. Additionally, its reliance on training data introduces potential biases.

Why COMET is Chosen: Given its strong correlation with human judgments, COMET serves as a robust metric to complement n-gram-based methods like BLEU and ROUGE-L, addressing their limitations in capturing semantic nuances.

4.2 Quantitative Results

This section presents the quantitative evaluation of the six models described earlier. Each model's performance is assessed using three widely accepted metrics for machine translation: ROUGE-L, BLEU, and COMET, as introduced in Section 4.1. These metrics evaluate the accuracy and quality of the generated translations compared to reference translations.

The results are summarized in Table 3, highlighting differences in performance between pre-trained, non-pre-trained, and augmented models.

Table 3: Quantitative Results: Evaluation Metrics for Each Model

Model	**ROUGE-L**	**BLEU**	**COMET**
Pre-trained Seq2Seq on MT	0.582	0.283	0.82
Pre-trained Seq2Seq (T5-small)	0.524	0.244	0.02
Non-Pre-trained Seq2Seq	0.3	0.064	-1.14
Non-Pre-trained Seq2Seq (Incr. Config.)	0.256	0.2	-1.36
Pre-trained LLM (LoRA Fine-Tuned)	0.552	0.297	0.79
Non-Pre-trained LLM (Decr. Config.)	0.18	0.0279	-1.24

The evaluation of the models across BLEU, ROUGE-L, and COMET metrics reveals distinct performance trends. Pre-trained Seq2Seq models, especially those fine-tuned for machine translation tasks like OPUS-MT, demonstrated the highest scores across all metrics, underscoring the value of task-specific pre-training. In contrast, general-purpose pre-trained models like T5-small showed moderate performance, reflecting their adaptability but highlighting limitations in task-specific optimization. Non-pre-trained models consistently underperformed, with significant gaps in BLEU and COMET scores, highlighting their reliance on extensive training data and computational resources for generalization.
The results observed of pre-trained LLM tell us that the final scores are similar to the Pre-trained Seq2Seq on MT, with slightly higher BLEU and slightly lower ROUGE-L and COMET.
These results emphasize the importance of pre-training and fine-tuning strategies in achieving robust machine translation performance, with metrics like COMET providing deeper insights into semantic adequacy and fluency compared to n-gram-based evaluations like BLEU and ROUGE-L.

4.3 Qualitative Results

While quantitative metrics provide valuable insights into the overall performance of translation models, qualitative analysis offers a detailed understanding of their strengths and limitations. By analyzing specific translations, we can evaluate how well each model handles syntax, semantics, idiomatic expressions, and overall fluency. In this section, we present example sentences for each model, comparing the generated translations against the reference translations. Observations are provided to highlight notable patterns and model-specific behaviors.

4.3.1 Pre-trained Seq2Seq on MT (Helsinki-opus-mt-en-it)

The pre-trained Seq2Seq model fine-tuned on machine translation (Helsinki-opus-mt-en-it) demonstrates strong syntactic and semantic accuracy. Its outputs align closely with the reference translations, particularly for straightforward and literal sentences.

Table 4: Example Translations for Pre-trained Seq2Seq on MT (Helsinki-opus-mt-en-it)

Input (EN)	**Reference (IT)**	**Generated (IT)**
We're here to celebrate compassion	Siamo qui per celebrare la compassione	Siamo qui per celebrare la compassione
Well, we went to the forests of Singapore and Southeast Asia	Bene, siamo andati nelle foreste di Singapore e del sud-est asiatico	Beh, siamo andati nelle foreste di Singapore e del sudest asiatico
For example, they charged Saudi foreign fighters substantially more than Libyans, money that would have otherwise gone to al Qaeda	Per esempio, ai combattenti stranieri sauditi addebitavano più che ai libici, soldi che altrimenti sarebbero andati ad al Qaeda	Per esempio, hanno caricato combattenti sauditi stranieri sostanzialmente più di soldi libici che altrimenti sarebbero andati ad al Qaeda
So, when I look at creativity, I also think that it is this sense or this inability to repress my looking at associations in practically anything in life	Allora, quando penso alla creatività, penso anche che sia questo senso o questa incapacità di reprimere il mio volere vedere associazioni praticamente in tutto quello che succede	Quindi, quando guardo alla creatività, penso anche che sia questo senso o questa incapacità di repressere il mio osservare le associazioni in pratica qualunque cosa nella vita
I'm on this till the whole thing spreads with chat rooms and copycats and moms maybe tucking kids into bed singing Hush little baby don't say a word	la notizia non si diffonderà con chat room, strilli di giornale e mamme che mandano a letto i figli cantando "Zitto bimbo mio non dire una parola"	Sono qui finché tutto non si è diffuso con chat rooms e copiatori e mamme che magari infilano i bambini nel letto cantando "Zitto bambino non dire una parola"

Observations: The model performs well with syntactic alignment and semantic accuracy. However, minor errors in word choice or idiomatic phrases ("chat rooms e copiatori") highlight challenges with cultural adaptation. Literal translations are generally accurate.

4.3.2 Pre-trained Seq2Seq not on MT (T5-small)

The T5-small model, pre-trained on general tasks but not machine translation, exhibits mixed results. It handles simple sentences reasonably well but struggles with idiomatic expressions and complex sentence structures.

Table 5: Example Translations for Pre-trained Seq2Seq not on MT (T5-small)

Input (EN)	**Reference (IT)**	**Generated (IT)**
In an asexual species, if you get two different mutations in different creatures, a green one and a red one, then one has to be better than the other	In una specie asessuata, due mutazioni in creature diverse, una rossa ed una verde, devono essere una migliore dell'altra	In una specie assue, se otteniamo due mutazioni diverse in creature diverse, una verde e una rossa, allora una deve essere meglio dell'altra
He's sort of a Homer Simpson with fins	È una sorta di Homer Simpson con le pinne	È una specie di Homer Simpson con polvere
so if algorithms are going to curate the world for us if they re going to decide what we get to see and what we don t get to see then we need to make sure that they re not just keyed to relevance	dunque se gli algoritmi dovranno prendersi cura del mondo in nostra vece se ci diranno cosa vedere e cosa non vedere allora dobbiamo assicurarci che non siano guidati unicamente dalla rilevanza	quindi se gli algoritmi saranno curati il mondo per noi se decideranno cosa vediamo e cosa non vediamo allora dobbiamo assicurarci che non sono solo chiamati alla pertinenza
They're not even autonomous	Non sono nemmeno autonomi	Non sono nemmeno autonomi
There's a marker line called the trim line above our little red illustration there	C'è una linea di demarcazione chiamata linea di rifinitura sopra la nostra piccola illustrazione rossa.	C'è una linea di marcatori chiamata linea di taglio sopra la nostra piccola immagine rossa

Observations: T5-small demonstrates significant errors in terminology and struggles with fluency. While some outputs are syntactically correct, others fail to capture key semantic nuances or idiomatic expressions, indicating limitations in handling linguistic context.

4.3.3 Non Pre-trained Seq2Seq

The non-pre-trained Seq2Seq model shows poor performance. Generated translations often contain repetitive phrases and lack semantic coherence.

Table 6: Example Translations for Non Pre-trained Seq2Seq

Input (EN)	**Reference (IT)**	**Generated (IT)**
in an asexual species if you get two different mutations in different creatures a green one and a red one then one has to be better than the other	in una specie asessuata due mutazioni in creature diverse una rossa ed una verde devono essere una migliore dell altra	in una specie di riferimento che avrete a che fare in due specie diversi e poi il rosso ha una rossa rossa che è meglio di un altro
He's sort of a Homer Simpson with fins	e una sorta di homer simpson con le pinne	e una specie di raq con sahariana
so if algorithms are going to curate the world for us if they re going to decide what we get to see and what we don t get to see then we need to make sure that they re not just keyed to relevance	dunque se gli algoritmi dovranno prendersi cura del mondo in nostra vece se ci diranno cosa vedere e cosa non vedere allora dobbiamo assicurarci che non siano guidati unicamente dalla rilevanza	quindi se i tetti sono i rifiuti per noi e il mondo ci stanno avvenendo per avere il mondo che vedremo e non vedremo che ci che vediamo a chiave che vediamo
They're not even autonomous	Non sono nemmeno autonomi	non sono nemmeno i risultati
There's a marker line called the trim line above our little red illustration there	C'è una linea di demarcazione chiamata linea di rifinitura sopra la nostra piccola illustrazione rossa.	c è una linea chiamata chiamata rossa rossa il nostro piccolo rosso c è il nostro piccolo rosso

Observations: The non-pre-trained Seq2Seq model demonstrates significant limitations in generating coherent translations. The outputs frequently include repetitive phrases, where the model generates redundant or nonsensical text. Additionally, semantic fidelity is often compromised, with translations deviating significantly from the intended meaning. While short and simple inputs yield somewhat accurate translations, the model struggles with

longer and more complex sentences, underscoring the challenges of training a Seq2Seq model from scratch without leveraging prior linguistic knowledge or pre-training. These results emphasize the necessity of pre-training for achieving semantic and syntactic coherence in translation tasks.

4.3.4 Non Pre-trained Seq2Seq (Increased Dimensionality)

The non-pre-trained Seq2Seq model with increased dimensionality shows poor performance. Generated translations often contain repetitive phrases and lack semantic coherence.

Table 7: Example Translations for Non Pre-trained Seq2Seq (Increased Dimensionality)

Input (EN)	**Reference (IT)**	**Generated (IT)**
in an asexual species if you get two different mutations in different creatures a green one and a red one then one has to be better than the other	in una specie asessuata due mutazioni in creature diverse una rossa ed una verde devono essere una migliore dell altra	se il rosso è una specie di rosso e il rosso è il rosso è una rossa rossa rossa rossa rossa e l altra e l altra è una rossa e poi è una rossa rossa
He's sort of a Homer Simpson with fins	È una sorta di Homer Simpson con le pinne	e una scatola di sabbia sana a sana a sana
So, if algorithms are going to curate the world for us if they re going to decide what we get to see and what we don t get to see then we need to make sure that they re not just keyed to relevance	dunque se gli algoritmi dovranno prendersi cura del mondo in nostra vece se ci diranno cosa vedere e cosa non vedere allora dobbiamo assicurarci che non siano guidati unicamente dalla rilevanza	quindi se non vedremo che ci aiutano a vedere che ci siano aiuti a vedere che ci ad accadere a spiegare il mondo e ci che vediamo che ci avranno bisogno di
They're not even autonomous	Non sono nemmeno autonomi	non sono senza risparmiare
there s a marker line called the trim line above our little red illustration there	C'è una linea di demarcazione chiamata linea di rifinitura sopra la nostra piccola illustrazione rossa.	c è una piccola chiamata rosso rosso chiamato rosso rosso

Observations: Despite the increased dimensionality, this model performs worse than the smaller, non-pre-trained counterpart. One possible explanation lies in the data and training process: the lack of pre-training means that the larger model's additional parameters do not provide an advantage but instead make optimization more challenging. The increased complexity without prior knowledge results in a higher risk of overfitting or converging to suboptimal patterns during training. This indicates that dimensionality alone cannot compensate for the lack of a robust initial representation, underscoring the critical role of pre-training in leveraging larger models effectively.

4.3.5 Pre-trained LLM

In the following table we will show the input, the reference and the actual translation of the pre-trained LLM. Generally we can see good performance.

Table 8: Translation for pre-trained LLM

Input (EN)	**Reference (IT)**	**Generated (IT)**
in an asexual species if you get two different mutations in different creatures a green one and a red one then one has to be better than the other	in una specie asessuata due mutazioni in creature diverse una rossa ed una verde devono essere una migliore dell altra	in una specie asexuale se ottenete due mutazioni diverse in creature diverse una verde e una rossa allora una deve essere meglio dell altra
He's sort of a Homer Simpson with fins	È una sorta di Homer Simpson con le pinne	È un po come un Homer Simpson con le pinne
So, if algorithms are going to curate the world for us if they re going to decide what we get to see and what we don t get to see then we need to make sure that they re not just keyed to relevance	dunque se gli algoritmi dovranno prendersi cura del mondo in nostra vece se ci diranno cosa vedere e cosa non vedere allora dobbiamo assicurarci che non siano guidati unicamente dalla rilevanza	Così, se gli algoritmi saranno in grado di curare il mondo per noi se saranno in grado di decidere ciò che vedremo e ciò che non vedremo allora dobbiamo assicurarci che non ci siano solo algoritmi legati alla rilevanza
They're not even autonomous	Non sono nemmeno autonomi	Non sono nemmeno autonome
there s a marker line called the trim line above our little red illustration there	C'è una linea di demarcazione chiamata linea di rifinitura sopra la nostra piccola illustrazione rossa.	c è una linea di taglio chiamata linea di taglio sopra la nostra piccola illustrazione

Observations:

We can tell that, seeing the table, the model is able to capture both syntactic and semantic informations of sentences. We can observe that the generated translations are not so much different from the reference of original dataset, underlying the ability of the model as we can expect from a pre-trained model.

4.3.6 Non Pre-trained LLM

The non pre-trained LLM with reduced dimensionality clearly shows poor performances. While some words or couple of words that are generally common are translated well, it is evident that the model lacks of semantic coherence.

Input (EN)	**Reference (IT)**	**Generated (IT)**
in an asexual species if you get two different mutations in different creatures a green one and a red one then one has to be better than the other	in una specie asessuata due mutazioni in creature diverse una rossa ed una verde devono essere una migliore dell altra	In una specie aspettativa se una specie una specie diversa è diversa una grande diversa e poi si mette in un un altra all altra all altra cosa all altra i le altre in
He's sort of a Homer Simpson with fins	È una sorta di Homer Simpson con le pinne	Lui ha fatta con una specie di nome
So, if algorithms are going to curate the world for us if they re going to decide what we get to see and what we don t get to see then we need to make sure that they re not just keyed to relevance	dunque se gli algoritmi dovranno prendersi cura del mondo in nostra vece se ci diranno cosa vedere e cosa non vedere allora dobbiamo assicurarci che non siano guidati unicamente dalla rilevanza	Quindi se stiamo in modo il mondo ci stiamo cercando di spiegare ciò che stiamo cercando di capire cosa stiamo cercando di capire cosa stiamo facendo per raccogliere ciò che vogliamo fare per capire non lo vediamo questo
They're not even autonomous	Non sono nemmeno autonomi	Hanno raggiungono la peggiora ete
there s a marker line called the trim line above our little red illustration there	C'è una linea di demarcazione chiamata linea di rifinitura sopra la nostra piccola illustrazione rossa.	C è una linea di sostenitore chiamata la linea che c è chiamata lontana del nostro piccolo tridimensionale insieme questi questo

Observations:

The outputs clearly show poor translation skills. Semantic fidelity is of-

ten compromised, with translations deviating significantly from the intended meaning. While The beginning of the inputs yield somewhat accurate translations, the model struggles with longer and more complex sentences, demonstrating the harshness of training a LLM model from scratch. These results emphasize the necessity of pre-training for achieving semantic and syntactic coherence in translation tasks.

4.3.7 Summary and Insights

The qualitative analysis highlights clear differences between pre-trained and non-pre-trained models. Pre-trained models excel in fluency and accuracy, while non-pre-trained models struggle with basic linguistic constructs. Augmented dimensionality and vocabulary do not mitigate issues arising from the absence of pre-training. These findings align with the quantitative results, reinforcing the importance of pre-trained architectures in neural machine translation tasks.

5 Conclusion and Future Work

5.1 Summary of Findings - Key insights

Summarize the key findings of the project. We have performed the training of several models based on different architectures with the purpose of comparing their performances for a machine translation task.

Among the approaches tested, the Seq2Seq model pretrained specifically on machine translation tasks demonstrated the best performance, significantly outperforming the others due to its task-specific pretraining. General Seq2Seq pretrained models, while showing acceptable results, were less effective compared to those optimized for translation. On the other hand, Seq2Seq models trained from scratch exhibited substantial limitations, generating outputs that lacked coherence and grammatical accuracy.

Pretrained large language models (LLMs) proved to be surprisingly effective in translation tasks, even without task-specific fine-tuning. This highlights their potential for versatility and adaptability across a range of applications. However, LLMs trained from scratch showed similar deficiencies to the non-pretrained Seq2Seq models. From the experiments, it is evident that fine-tuning plays a pivotal role in achieving high-quality translations. Models that leverage it can harness the linguistic patterns and syntactic structures learned during training on extensive datasets, resulting in improved accuracy and coherence. The advantages of pretrained LLMs also became clear, as they managed to balance flexibility and accuracy, making them suitable for diverse tasks even in the absence of fine-tuning. Conversely, non-pretrained models struggled with repetition, incoherence, and a lack of semantic understanding, reaffirming the value of knowledge transfer from general tasks to specific ones.

While increasing the dimensionality of non-pretrained models did not lead to significant improvements, this result highlights the need to balance computational efficiency with effective initialization strategies. Pretraining not only reduces the computational burden but also accelerates convergence during task-specific training.

5.2 Future works

For future work, combining pretrained LLMs with targeted fine-tuning approaches, such as LoRA, could further enhance their performance for machine translation tasks. Additionally, expanding the dataset with more diverse and representative examples could improve generalization capabilities. Finally, hybrid approaches that integrate the robustness of LLMs with the specificity of Seq2Seq models pretrained on machine translation may offer a promising direction for further exploration. In addition, expanding the work with the inclusion of the inverse translation task from Italian to English could further enrich this study.

In conclusion, this study underscores the importance of fine-tuning and task-specific optimization in machine translation. While LLM-based approaches show considerable promise for flexibility and adaptability, achieving high-quality results still necessitates tailoring models to the task at hand.