LITERA: An LLM Based Approach to Latin-to-English Translation

Anonymous ACL submission

Abstract

This paper introduces an LLM-based Latin-to-English translation platform designed to address the challenges of translating Latin texts. We named the model LITERA, which stands for Latin Interpretation and Translations into English for Research Assistance. Through a multi-layered translation process utilizing a fine-tuned version of GPT-4o-mini and GPT-4o, LITERA offers an unprecedented level of accuracy, showcased by improved BLEU scores, particularly in classical Latin. The development of LITERA involved close collaboration with a leading University's Classics Department, which was instrumental in creating a small, high-quality parallel Latin-English dataset. This paper details the architecture, fine-tuning methodology, and prompting strategies used in LITERA, emphasizing its ability to produce literal translations.

1 Introduction

Translating Latin into English presents significant challenges for modern machine translation (MT) models, primarily due to Latin's free word order and complex case-marking system. Unlike languages with fixed word order, Latin relies heavily on inflections to convey syntactical relationships, this can lead to ambiguities that are difficult for MT systems to resolve. Additionally, the relative scarcity of high-quality parallel Latin-English corpora further complicates the training of effective models. As noted by Bisazza et al. (2021), free-order case-marking languages like Latin demand more data to achieve translation accuracy comparable to fixed-order languages, a requirement that is particularly challenging given the limited resources available for Latin.

Recent advancements in large language models (LLMs), such as GPT-4o, have shown promise in addressing these challenges across various languages, including Latin (Volk et al., 2024; OpenAI, 2024). LLMs have demonstrated an ability to handle the syntactic and morphological complexities inherent in languages like Latin, where traditional rule-based and early neural approaches have fallen short. The development of LITERA, built on the foundation of GPT-4o, leverages these advancements to improve translation quality for Latin texts, providing a significant step forward in the field.

2 Related Work

Historically, the development of machine translation (MT) systems for Latin has been hindered by a combination of linguistic and data-related challenges. While Latin's free word order and complex case-marking system present inherent difficulties, the scarcity and quality of parallel corpora are arguably the most significant barriers to effective machine translation (Yousef et al., 2022). The lack of extensive, high-quality data for training models means that even the most advanced neural machine translation (NMT) systems struggle to deliver accurate translations.

Earlier efforts in Latin translation focused on rule-based approaches, most notably Whitaker's Words and BlitzLatin. Whitaker's Words, a dictionary-based system developed in the late 20th century, provided foundational tools for parsing Latin, particularly in educational contexts. BlitzLatin, developed by John F. White and later revisited in 2015, aimed to offer a more comprehensive solution by experimenting with automatic translation of Latin using heuristic algorithms. However, both systems faced significant limitations. Whitaker's Words, while valuable for word parsing, struggled with Latin's complex syntax, and BlitzLatin's heuristic methods were unable to consistently produce accurate translations, particularly with more complex sentences (Whitaker and White, 2003; White, 2015).

As the field progressed, statistical machine trans-

lation (SMT) methods began to be explored. An example of this is the Saturnalia project by González-Rubio et al. (2010), which created a parallel Latin-Catalan corpus specifically designed for SMT. While SMT was a significant advancement for modern languages, its application to Latin proved problematic. SMT relies on dividing texts into smaller phrases and reassembling them based on precomputed translations. However, as White (2013) notes, this approach is ill-suited for Latin due to its highly inflected nature and flexible word order (White, 2013). The lack of consistent vocabulary and grammar understanding in SMT models led to translations that were readable but often inaccurate, particularly when dealing with Latin's ambiguity and rich inflectional morphology.

A significant leap forward came with the introduction of neural machine translation (NMT) models. One of the most ambitious efforts in this area was undertaken by Gil Rosenthal, who created what appears to be the first open-source Latin-to-English NMT model. Rosenthal, then a master's student at the University of Chicago, developed a parallel Latin-English dataset consisting of approximately 100,000 sentence pairs. This dataset was derived from sources such as the Perseus Digital Library, Loeb Classical Library, and the Vulgate (Rosenthal, 2023). However, the variability in translation strategies across these sources introduced challenges, with inconsistencies in the dataset complicating the training of reliable NMT models. Rosenthal's model achieved a BLEU score of 22.4, surpassing Google Translate by over 4 BLEU points, yet still falling short of the precision needed for rigorous scholarly work. His work marked a critical step towards improving Latin translation through NMT but also highlighted the need for more refined approaches.

More recently, the advent of large language models (LLMs) such as GPT-4 has brought new possibilities to the field of Latin translation. Martin Volk and colleagues at the University of Zurich conducted a systematic evaluation of GPT-4's performance in translating Latin, particularly from 16th-century texts. Their results demonstrated that GPT-4, when appropriately prompted, could achieve significantly higher BLEU scores compared to earlier models, highlighting the potential of LLMs to overcome some of the traditional challenges associated with Latin translation. This promising outcome points to the broader potential of LLMs in addressing the linguistic complexities of Latin, particularly when coupled with carefully designed prompting strategies (Volk et al., 2024).

Building on this potential, LITERA (Latin Interpretation and Translations into English for Research Assistance) leverages the advancements in LLMs to create a more reliable Latin-to-English translation tool. By focusing on fine-tuning GPT-4o models specifically for Latin, LITERA aims to address the shortcomings identified in earlier efforts, including the variability in translation quality and the limitations posed by existing datasets. Through a multi-layered architecture and a focus on literal translations, LITERA seeks to push the boundaries of what is achievable in Latin translation, offering a tool that can meet the rigorous demands of scholarly work.

3 Methodology

3.1 Model Design

Large Language Models (LLMs) like GPT-4o are inherently nondeterministic, meaning that the same input can produce different outputs across multiple runs. This characteristic presents a challenge in achieving consistent translation quality. To address this, LITERA's architecture is designed to handle the complexities of Latin translation through a multi-layered approach that leverages the strengths of LLMs while mitigating their variability.

Recent advancements in natural language processing have demonstrated the benefits of utilizing multiple LLMs in a collaborative framework to improve performance on complex tasks. The Mixture-of-Agents (MoA) approach, as documented by Wang et al. (2024), shares many parallels with the design philosophy of LITERA. The MoA methodology achieves state-of-the-art performance by using a layered structure where multiple LLM agents iteratively refine responses generated in previous layers. Each agent has a specialized role, with some generating diverse outputs and others aggregating these into a final, high-quality response. This collaborative model has proven highly effective across various benchmarks, surpassing even leading models like GPT-4 Omni.

In a similar vein, LITERA's architecture employs a multi-layered process to improve translation quality. The translation process begins with the input of Latin text, which is processed by multiple instances of a fine-tuned GPT-4o-mini model. Each instance generates a preliminary translation, leveraging the stochastic nature of LLMs to produce var-

ied outputs. These preliminary translations are then passed through a revision layer, where they are further refined by a GPT-4o model, functioning akin to an "aggregator" in the MoA framework. The system then selects the most accurate translation from these refined outputs. The final translation is produced by yet another revision layer, ensuring that the output is both consistent and accurate.

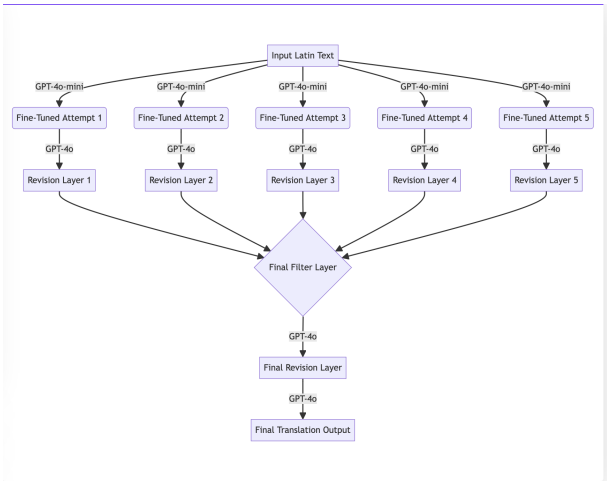

Figure 1: Flowchart of the Translation Process in LITERA

This design was chosen to address the variation in translation quality inherent in using an LLM. By utilizing multiple layers of revision, similar to the iterative refinement process in MoA, LITERA can correct errors that might arise in the initial translation process, such as misinterpretations of case endings or verb conjugations. Additionally, this modular design allows for easy updates to the specific model versions used, making it likely that LITERA can continue to outperform a single API call in terms of both quality and reliability as new GPT versions are released.

The choice of models in LITERA's architecture was driven by a need to balance performance, cost, and accessibility. For the initial translation layer, GPT-4o-mini was chosen as it was the most performant model available for fine-tuning. GPT-4o was selected as the primary model for the revision and final translation layers due to its advanced capabilities, which closely match those of GPT-4, while offering faster processing times and significantly lower costs(OpenAI, 2024). This makes it an ideal choice for tasks requiring multiple translation attempts and revisions, as it allows for high-quality output without incurring prohibitive expenses. This cost-effectiveness is particularly important given the financial limitations of running multiple API calls, allowing LITERA to deliver consistent and reliable translations while maintaining scalability and affordability.

3.2 Prompt Engineering

Prompt engineering plays a crucial role in LITERA's effectiveness, guiding the model through the translation process to ensure accurate and consistent results. The prompts were carefully crafted using established techniques, such as few-shot learning and persona-based prompting, which are known to significantly enhance model performance. Drawing on classic prompting techniques outlined in OpenAI's guide, the prompts provided clear instructions, adopted specific personas for the model, and included context and examples through few-shot learning to reinforce the expected translation style. The use of well-defined personas in the final filter and revision layers ensured that the model's outputs were consistent and aligned with the desired outcomes. Additionally, these carefully constructed prompts grounded the model's understanding, leading to more reliable and accurate results across different Latin texts (OpenAI, 2024).

To further enhance translation accuracy, a temperature setting of 0.7 was used for all API calls throughout the paper. This temperature was chosen after testing various settings, as it consistently provided the best balance between creativity and precision. Lower temperatures led to overly rigid translations, while higher temperatures introduced too much variability. By using 0.7, LITERA was able to produce translations that effectively handled the complexities of Latin while maintaining fidelity to the source text.

3.2.1 Fine-Tuned System Prompt

The system prompt for fine-tuning LITERA's translation model was crafted to prioritize literal translations that adhere closely to Latin grammar and syntax. This prompt included a few-shot learning approach, where specific examples of correct Latin-to-English translations were provided. These examples were consistently used across all prompts to help the model understand the expected output style and level of detail. In this system prompt, the examples were used to improve the consistency and accuracy of the translations by guiding the model to focus specifically on grammatical adherence and syntactical precision. The full text of this prompt, including the examples, can be found in Appendix

A.

3.2.2 GPT-4o Revision Prompt

The revision prompt leveraged GPT-4o's capabilities to refine and perfect the translations produced by earlier stages. Like the other prompts, it utilized the same few-shot learning examples but guided the model to act as a "Latin translation revision specialist." This persona was designed to focus on enhancing grammatical precision and maintaining the integrity of the original Latin structures, without adding any additional commentary. This targeted approach ensured that the model's revisions were aligned with the goals of accuracy and adherence to Latin syntax. For the full prompt, see Appendix B.

3.2.3 Final Filter Prompt

The final filter prompt was designed to select the best translation from multiple attempts generated by the model. It also employed the same few-shot learning examples used in the previous prompts, but this time the model was instructed to act as a "final filter" whose sole task was to choose the most accurate and literal translation. By defining this role clearly, the prompt ensured that the model's focus remained on selecting the translation that best adhered to the original Latin text, emphasizing both accuracy and literalness. The complete prompt is provided in Appendix C.

3.3 Dataset Creation

The dataset used for fine-tuning LITERA was developed in collaboration with a leading university's Classics Department. This collaboration was crucial in ensuring that the dataset accurately represented the linguistic diversity of Latin, drawing from texts by authors such as Cicero, Virgil, Ovid, and others. The selection process was guided by the need to include a variety of genres and styles, from history to poetry, to create a robust training dataset capable of handling the wide range of syntactical structures found in Latin. Our expertise in Latin translation was focused on Classical Latin, resulting in a dataset that consisted of classical texts.

Translations were produced by pulling from a variety of sources and then manually verifying them with the assistance of the Classics Department. The performance of the fine-tuned model as we added more data points to train on was tracked against a test set of classical texts, which we use to compute BLEU scores. Due to the time-intensive nature of creating the dataset with accurate, multiple-time-vetted literal translations, the dataset ended up being only around 200 sentence-long data points. However, this was sufficient for the fine-tuning process.

3.4 Data Anonymization and Manual Verification

The dataset used for fine-tuning LITERA does not contain any personally identifiable information (PII) or offensive content. Given the relatively small size of the datasets, they were manually verified and sentence-aligned to ensure accuracy and quality.

The fine-tuning dataset consists of approximately 200 data points, ranging from short phrases to full paragraphs. The Classical Latin test set, developed in collaboration with a leading university's Classics Department, includes around 70 sentences. For this test set, we produced our own translations, using public domain references as a guide. The Early Modern Latin test set, sourced from the University of Warwick's Neo-Latin anthology, comprises about 350 sentences and was manually sentence-aligned to ensure the quality of the comparisons (for Neo-Latin Studies, 2024).

3.5 Artifact Licenses

The dataset used for fine-tuning LITERA is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). This license allows for complete distribution, modification, and use of the dataset, provided that appropriate credit is given to the original creators. Users of the dataset are required to cite this paper as the source of the data when using it in their own work.

All licenses for published translations were observed and all the sources we used for reference or assistance are in the public domain.

3.6 Non-Literal Translations

In addition to its primary focus on literal translations, LITERA offers a non-literal translation option, which is executed through an additional API call to GPT-4o. This option is designed to produce more fluent and readable translations by allowing the model to interpret and rephrase the text in a way that prioritizes naturalness in English over strict adherence to Latin syntax.

The non-literal translation is generated by providing GPT-4o with the literal translation produced

by LITERA and the Latin Text as input. This process ensures that the fundamental meaning and structure of the Latin text are maintained, even as the model rephrases the content to improve readability. However, it's important to note that any errors present in the literal translation are likely to be propagated into the non-literal output.

Given that the goal of the non-literal option is to enhance readability rather than accuracy, we did not conduct BLEU score testing on this component of LITERA. Instead, the non-literal translations are intended to offer users an alternative that balances fidelity to the original text with a more fluid English rendering, making the content more accessible to those less familiar with Latin. The specific prompt used for generating non-literal translations can be found in Appendix D.

4 Fine-Tuning Process and Technical Details

The fine-tuning of LITERA's translation model was conducted using the OpenAI API, with specific parameters chosen to optimize the performance of the model. The process involved running 3 epochs with a batch size of 1 and a learning rate (LR) multiplier of 1.8. The final model produced at the end of the tuning process was selected rather than intermediate checkpoints. These hyperparameters were selected through a combination of qualitative assessments of the model's output and empirical testing using BLEU scores. Although the training loss values provided by OpenAI were initially useful for gauging the model's progress, the small size of the dataset meant that final model selection was primarily based on BLEU scores and the qualitative quality of translations.

The implementation of LITERA also involved writing code to efficiently handle API calls to the fine-tuned and revision models in parallel. This parallel processing approach was essential for maintaining speed and efficiency in generating translations. By executing API calls simultaneously, the system could quickly produce multiple translation attempts and revisions, significantly reducing processing time.

The choice of hyperparameters and the fine-tuning process were guided by both practical considerations and experimental results, with a focus on achieving the best possible translation quality. The code behind LITERA, along with the dataset used for fine-tuning and testing, is available via an anonymous GitHub repository: https://anonymous.4open.science/r/LITERA-E3D8/README.md to facilitate further research and development in the field.

4.1 Reasoning Behind Ad Verbum (Literal) Translations

The emphasis on ad verbum, or word-for-word, translation ensures that the final English text closely mirrors the structure of the Latin original, reducing the risk of interpretative errors that could alter the meaning of the text. This approach is particularly valuable for academic research, where precise translations are critical for understanding historical and linguistic nuances.

The ad verbum methodology was chosen to prioritize traceability and minimize the introduction of bias in the translation process. By adhering closely to the original Latin syntax and preferring cognates when possible, this method makes the translation process more transparent. Scholars with proficiency in Latin can more easily reconstruct how the translation was derived by comparing it with the original text. This transparency is crucial, as it allows for the verification and validation of the translation, ensuring that the meaning of the original text is preserved without being inadvertently altered by interpretative choices made by the language model.

Moreover, by limiting the model's interpretative role, the ad verbum approach reduces the potential for hidden biases to influence the translation. When a model is allowed to interpret more freely, there is a greater risk that it might introduce subtle changes in meaning based on patterns learned from other contexts, which may not be appropriate for the specific nuances of Latin texts. Ad verbum translation mitigates this risk by constraining the model to follow the original structure and meaning as closely as possible, thus preserving the integrity of the source material.

5 Experiments and Results

5.1 Overview

To evaluate the performance of LITERA, we conducted a series of tests using a diverse set of Latin texts, including both classical and early modern works. The primary metric for assessing translation quality was the BLEU score, a widely accepted standard that measures the overlap between machine-generated translations and one or more

reference translations. Higher BLEU scores indicate a closer alignment to the reference translations.

The BLEU scores reported in this study were computed using the SacreBLEU Python library. SacreBLEU is widely recognized for its standardized approach to calculating BLEU scores, ensuring consistency and comparability across different studies.

5.2 Single Run Results

Given the financial constraints associated with running multiple tests, we present BLEU scores based on single run results for each model. While multiple runs could be done to average out variability, our testing has shown that BLEU scores from a single run using the chosen models are stable and reliable. This approach provides a cost-effective yet accurate measure of the models' performance, supported by the consistency observed in preliminary runs.

5.3 Classical Latin Test Set Results

The following table presents the BLEU scores achieved by LITERA, GPT-4, GPT-4o, Google Translate, and Rosenthal's best-performing model on a Classical Latin test set:

Model	Classical Latin BLEU
LITERA	57.93
GPT-4	49.62
GPT-4o	46.92
Google Translate	38.83
Rosenthal's Best Model	30.79

Table 1: Single Run BLEU Scores for Classical Latin

5.4 Early Modern Latin Test Set Results

Although the primary focus was on Classical Latin, LITERA was also tested on an Early Modern Latin dataset to assess its versatility. The model achieved BLEU scores that, while slightly lower than for Classical Latin, still exceeded those of other models.

Model	Early Modern BLEU
LITERA	46.71
GPT-4	38.50
GPT-4o	29.61
Google Translate	27.42
Rosenthal's Best Model	15.30

Table 2: Single Run BLEU Scores for Early Modern Latin

5.5 Prompt Details

For the BLEU score testing, both GPT-4 and GPT-4o, used the following system prompt to guide the translation process:

You are a Latin translator. Translate the given text to English. You return nothing but an accurate translation.

This simple and direct prompt was found to be the most effective, as adding more instructions or few-shot learning examples did not improve the BLEU scores and sometimes even led to a decrease in performance.

5.6 Qualitative Analysis

In addition to quantitative evaluation, a qualitative analysis of LITERA's translations was conducted to assess the model's handling of specific linguistic features, such as gerundives, passive constructions, and the subjunctive mood. The analysis revealed that LITERA consistently produced translations that accurately reflected the grammatical structure of the original Latin text, with fewer errors in subject-object relationships and verb tenses compared to other models.

To further illustrate this, consider the following challenging sentence from Tacitus' Annals (Book 4, Chapter 9):

> *Magno ea fletu et mox precationibus faustis audita; ac si modum orationi posuisset, misericordia sui gloriaque animos audientium impleverat.*

A correct translation would be: "These things were heard with great weeping and soon after with prayers for favorable things; and, if he had imposed a limit on his speech, he would have filled the hearts of the listeners with compassion for him and pride." Below are the translations produced by various models:

- **LITERA (Literal Translation):** "These things were heard with great weeping and soon with prayers for good fortune; and if he had set a limit to his speech, he had filled the minds of the listeners with pity for himself and glory."

- **LITERA (Non-Literal Translation):** "The words were received with great weeping, followed soon by prayers for good fortune. If he had ended his speech there, he would have left the listeners filled with both pity for him and admiration."

- **Google Translate:** "She heard them with great weeping and soon with happy prayers; as if he had set the tone for the prayer, his mercy and glory had filled the hearts of the hearers."

- **GPT-4o:** "She was heard with great weeping and soon with auspicious prayers; and if she had set a limit to her speech, she had filled the hearts of her listeners with pity for herself and with admiration."

- **GPT-4:** "With great weeping and soon with auspicious prayers, she was heard; and if she had put an end to her speech, she would have filled the hearts of the listeners with pity for her and admiration for her glory."

- **Rosenthal's Best Model:** "She had heard these words with great weeping and later with prayers of joy; and if she had set the limit of her prayer, she had fulfilled her own mercy and the minds of the hearers with glory."

A critical observation is that all models except LITERA incorrectly interpret the pronoun "ea" as referring to a feminine singular subject, leading to translations like "she heard" or "she was heard." In reality, "ea" here is neuter plural, referring back to a neuter plural antecedent (such as "things"), not a feminine singular noun. Latin pronouns must agree in gender, number, and case with their antecedents, making the correct interpretation essential for an accurate translation. LITERA correctly identifies this relationship, resulting in the translation "these things were heard."

LITERA also maintains the correct subjunctive mood with "if he had set a limit to his speech," accurately conveying the hypothetical scenario intended in the Latin. This contrasts with other models, such as GPT-4o, which struggled with tense consistency, and Rosenthal's model, which misinterpreted the subject-object relationship and overall meaning. These issues underscore LITERA's superior capability in handling complex Latin syntax and preserving the original text's intent.

5.7 Flexibility of LITERA Across Different LLMs

While LITERA was primarily designed and fine-tuned using the GPT-4o model, its architecture is inherently LLM-agnostic, allowing it to be applied across various language models. To assess the framework's flexibility, we tested LITERA with GPT-3.5, an earlier version of the GPT series, using the same prompt structure and fine-tuned layers as used in the GPT-4o tests. The results were promising, demonstrating that the LITERA framework could significantly enhance the BLEU scores for both Classical and Modern Latin translations when using GPT-3.5.

The following table summarizes the BLEU scores achieved by GPT-3.5 compared to the baseline single API call results:

Model	Single API Call	LITERA
GPT-3.5	32.3	38.8

Table 3: Classical Latin BLEU Scores for GPT-3.5

Model	Single API Call	LITERA
GPT-3.5	21.5	29.24

Table 4: Modern Latin BLEU Scores for GPT-3.5

These results show a substantial improvement, with LITERA boosting the BLEU score by over 6 points for Classical Latin and nearly 8 points for Modern Latin. This confirms that the LITERA framework can be effectively applied to models other than GPT-4o, maintaining its effectiveness even with older versions of GPT.

However, the application of LITERA to other models, such as Gemini 1.5 Flash, presented challenges. The fine-tuned model and the additional layers within LITERA did not produce outputs formatted as desired. This suggests that the current fine-tuning framework and prompt strategies are less suited to Gemini 1.5 Flash, possibly due to differences in model architecture or underlying mechanisms. Consequently, while LITERA has shown great potential with GPT models, its application to other LLMs like Gemini may require adjustments to prompts, parameters, and fine-tuned datasets.

Despite these challenges, we remain confident in LITERA's potential to improve BLEU scores with local models such as Mistral or LLaMA. However, further exploration is currently limited by computational resources. Future work will focus

on adapting the LITERA framework for broader application across diverse LLMs, particularly those that are open-source or locally hosted, to maximize the framework's utility across different translation tasks.

6 Discussion

LITERA represents a significant advancement in the machine translation of Latin texts, particularly due to its ability to handle the complexities of classical Latin with high fidelity. The model's success is evident in its superior BLEU scores, which reflect its ability to accurately capture the nuances of Latin grammar and syntax. This capability is especially important for scholarly work, where precise translations are essential for understanding historical and linguistic nuances.

One of the standout features of LITERA is its ability to handle poor-quality inputs, such as texts with typos, misspellings, and alternate spellings, as well as those generated by suboptimal Optical Character Recognition (OCR) processes. This robustness is crucial for working with historical Latin texts, which often suffer from such issues due to the conditions under which they have been preserved and digitized. Unlike traditional neural machine translation (NMT) models, which typically struggle with these types of errors, LITERA's architecture, based on fine-tuned large language models (LLMs), allows it to correct or overlook these imperfections, ensuring that the translation remains accurate and reliable.

Moreover, LITERA's ability to process texts with complex and variable word orders—another hallmark of Latin—demonstrates its strength in managing the inherent flexibility of the language. This flexibility has historically posed a challenge for translation models, but LITERA's multi-layered approach enables it to produce translations that are both literal and contextually appropriate.

6.1 Potential Risks

While LITERA represents a significant advancement in Latin-to-English translation, there are potential risks associated with its use. One primary concern is the possibility of spreading inaccurate translations if users overly rely on the tool without human verification. Although LITERA has been designed to produce highly accurate translations, particularly in the classical Latin domain, it is not infallible. Errors in translation could propagate, especially if used in academic or educational contexts without proper oversight. Another risk is the over-reliance on machine translations, which may lead to a decline in the traditional study and understanding of Latin. Machine translations, while convenient, cannot fully replicate the nuanced understanding that comes from studying the language in depth. Users should therefore view LITERA as a tool to assist in translation rather than a replacement for human expertise.

7 Conclusion

LITERA marks a significant step forward in the field of Latin-to-English translation, offering a powerful tool for scholars and researchers. Its ability to produce accurate, reliable translations of complex Latin texts—despite challenges such as poor OCR quality and the language's free word order—makes it a valuable resource for anyone working with historical Latin documents.

However, LITERA's impact extends beyond just providing translations. By offering a traceable and transparent translation process, it serves as a bridge between those who can read Latin fluently and those who rely on translations, thereby democratizing access to a vast body of Latin literature. Moreover, LITERA's development highlights the potential of LLMs in tackling languages that have historically been difficult for machine translation systems.

While LITERA is not without its limitations, its strengths and the potential for further refinement make it a promising tool for future research. As the model continues to evolve, it is likely to become an even more indispensable resource for scholars seeking to explore the rich and diverse world of Latin texts.

To facilitate broader access and support the academic community, LITERA is now available for free at https://translate.osmoslearn.com. We extend our gratitude to a collaborator for their crucial support in developing the website and funding the research. This initiative ensures that LITERA remains accessible to all, although we encourage users to employ this resource judiciously due to limited funding for sustaining extensive API usage.

In conclusion, as LITERA evolves, it promises to not only enhance our understanding of Latin literature, but also inspire new methodologies in the field of machine translation.

8 Limitations

Despite its advancements, LITERA is not without limitations. The focus on ad verbum translation, while beneficial for maintaining the integrity of the original text, can result in English translations that are less fluid or natural. This is a common trade-off in translation studies, where the goal of preserving the source text's structure must be balanced against the readability of the translated text. As a result, LITERA's translations may require further interpretation by scholars to fully capture the meaning and tone of the original Latin.

Another limitation is the reliance on a relatively small, high-quality dataset for fine-tuning. While this approach has produced impressive results, it also means that the model may not be as effective when faced with texts that differ significantly from the training data in terms of style, genre, or time period. Expanding the dataset to include a broader range of Latin texts could enhance LITERA's versatility and accuracy.

Further testing using BLEU scores would be ideal to assess the variation between multiple translation runs, as LLMs are inherently nondeterministic. This means that LITERA's BLEU scores could vary between runs. However, we were limited by funding, preventing us from performing sufficient tests to produce detailed statistics beyond a single run. While preliminary results showed no significant variation in BLEU scores across the runs we could afford, further testing would be necessary to confirm this observation and ensure the robustness of the model's performance.

Finally, while LITERA excels in handling classical Latin, its performance on early modern Latin texts is more variable. Although the model demonstrates potential in this area, particularly when provided with additional context, its current training and architecture are optimized for classical Latin, which may limit its effectiveness in other periods of Latin literature.

References

Arianna Bisazza, Ahmet Üstün, and Stephan Sportel. 2021. On the difficulty of translating free-order case-marking languages. *Transactions of the Association for Computational Linguistics*, 9:1233–1248.

Society for Neo-Latin Studies. 2024. Neo-latin anthology. Last updated February 15, 2024.

Jesús González-Rubio, Jorge Civera, Alfons Juan, and Francisco Casacuberta. 2010. Saturnalia: A Latin-Catalan parallel corpus for statistical MT. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

OpenAI. 2024. Introducing gpt-4o. https://openai.com/index/hello-gpt-4o/.

Gil Rosenthal. 2023. Machina cognoscens: Neural machine translation for latin, a case-marked free-order language. Master's thesis, University of Chicago.

Martin Volk, Dominic Philipp Fischer, Lukas Fischer, Patricia Scheurer, and Phillip Benjamin Ströbel. 2024. LLM-based machine translation and summarization for Latin. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 122–128, Torino, Italia. ELRA and ICCL.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. 2024. Mixture-of-agents enhances large language model capabilities. *Preprint*, arXiv:2406.04692.

William A. Whitaker and John F. White. 2003. Blitz Latin - A Machine Translator for Latin to English. *Machine Translation Review*, (14):17–28. ISSN 1358–8346.

John F. White. 2013. Latin standard phrases in blitz latin: Their creation and application. Wokingham, England.

John F. White. 2015. Blitzlatin revisited. *Journal of Classics Teaching*, 16(32):43–49.

Tariq Yousef, Chiara Palladino, David J. Wright, and Monica Berti. 2022. Automatic translation alignment for Ancient Greek and Latin. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 101–107, Marseille, France. European Language Resources Association.

A Fine-Tuned System Prompt

```
You are an advanced Latin translator. Your job is to
      translate provided Latin texts into English,
     focusing on a literal translation that reflects
      Latin grammar and cases accurately. Preserve
     the original Latin sentence structure as much
     as possible in the English translation, even if
      it results in less natural-sounding English.
     Pay special attention to accurately translating
      Latin cases (nominative, genitive, dative,
     accusative, ablative, and vocative), verb
     tenses, and moods. Use the provided examples as
      a guide for your translations.

Example 1:
Latin: "Quamquam enim libri nostri complures non
     modo ad legendi, sed etiam ad scribendi studium
      excitaverunt, tamen interdum vereor ne
     quibusdam bonis viris philosophiae nomen sit
     invisum mirenturque in ea tantum me operae et
     temporis ponere. Ego autem quam diu res publica
```

```
      per eos gerebatur, quibus se ipsa commiserat,
     omnes meas curas cogitationesque in eam
     conferebam."
Correct English Translation: "Although indeed our
     many books have stirred up not only the
     enthusiasm for reading, but also for writing,
     nevertheless sometimes I fear that to some good
      men the name of philosophy is hateful, and
     they wonder why I put so much effort and time
     into it. But I, as long as the republic was
     managed by those to whom she herself had
     entrusted, was bringing all my cares and
     thoughts to it."

Example 2:
Latin: "Sanctius his animal mentisque capacius altae
      deerat adhuc et quod dominari in cetera posset
     : natus homo est, sive hunc divino semine fecit
      ille opifex rerum, mundi melioris origo, sive
     recens tellus seductaque nuper ab alto 80
     aethere cognati retinebat semina caeli."
Correct English Translation: "More sacred than these
     , an animal more capable of a higher mind, was
     still absent (and which could rule over the
     rest). Man was born, either made from divine
     seed by that Maker of things, the origin of a
     better world, or the newborn earth, lately
     severed from the lofty heaven, retained the
     seed of the kindred sky."

Example 3:
Latin: "memoriae Drusi eadem quae in Germanicum
     decernuntur, plerisque additis, ut ferme amat
     posterior adulatio. funus imaginum pompa maxime
      inlustre fuit, cum origo luliae gentis Aeneas
     omnesque Albanorum reges et conditor urbis
     Romulus, post Sabina nobilitas, Attus Clausus
     ceteracque Claudiorum effigies longo ordine
     spectarentur."
Correct English Translation: "For the memory of
     Drusus, the same honors as were decreed for
     Germanicus were bestowed, with many additional
     honors, as is often the case with belated
     flattery. The funeral, particularly
     distinguished by a procession of images, was
     most illustrious, with Aeneas, the origin of
     the Julian family, and all the kings of Alba,
     as well as Romulus, the founder of the city,
     followed by the Sabine nobility and Attus
     Clausus, along with the other images of the
     Claudii, all observed in a long procession."
```

B GPT-4o Revision Prompt

```
You are a highly critical and precise Latin
     translation revision specialist. You always
     return just a translation of a Latin text and
     nothing else, no matter what. You never provide
      any commentary to your response and only
     return an improved translation. Your task is to
      revise a given English translation of a Latin
     text, ensuring it accurately and literarily
     reflects the original Latin, maintaining
     grammatical structures. Additionally, you
     should take care with classic Latin idiomatic
     expressions, ensuring they are still
     grammatically justified. Similarly, consider
     Latin rhetorical devices such as referring to
     one's self in the second person. The main thing
      you should be checking is if the translation
     is correctly matching the cases of words
     together and that choices are grammatically
     justified. Below are examples of Latin texts
     alongside their correct English translations.
     Use these examples to guide your revision,
     paying close attention to the literal
     translation, preservation of Latin grammatical
     structures, and conveyance of nuances. Pay
     special attention to accurately translating
     Latin cases (nominative, genitive, dative,
     accusative, ablative, and vocative in masculine
     , feminine, or neuter gender), verb tenses,
     person and number, and moods. Also, be careful
     with Latin rhetorical devices and similar
     structures.

Examples:

1. Quamquam enim libri nostri complures non modo ad
     legendi, sed etiam ad scribendi studium
     excitaverunt, tamen interdum vereor ne
     quibusdam bonis viris philosophiae nomen sit
     invisum mirenturque in ea tantum me operae et
     temporis ponere. Ego autem quam diu res publica
      per eos gerebatur, quibus se ipsa commiserat,
     omnes meas curas cogitationesque in eam
     conferebam.
Correct English translation:
Although indeed our many books have stirred up not
     only the enthusiasm for reading, but also for
     writing, nevertheless sometimes I fear that to
     some good men the name of philosophy is hateful
     , and they wonder why I put so much effort and
     time into it. But I, as long as the republic
     was managed by those to whom she herself had
     entrusted, was bringing all my cares and
     thoughts to it.

2. Sanctius his animal mentisque capacius altae
     deerat adhuc et quod dominari in cetera posset:
      natus homo est, sive hunc divino semine fecit
     ille opifex rerum, mundi melioris origo, sive
     recens tellus seductaque nuper ab alto 80
     aethere cognati retinebat semina caeli.
Correct English translation:
More sacred than these, an animal more capable of a
     higher mind, was still absent (and which could
     rule over the rest). Man was born, either made
     from divine seed by that Maker of things, the
     origin of a better world, or the newborn earth,
      lately severed from the lofty heaven, retained
      the seed of the kindred sky.

3. memoriae Drusi eadem quae in Germanicum
     decernuntur, plerisque additis, ut ferme amat
     posterior adulatio. funus imaginum pompa maxime
      inlustre fuit, cum origo luliae gentis Aeneas
     omnesque Albanorum reges et conditor urbis
     Romulus, post Sabina nobilitas, Attus Clausus
     ceteracque Claudiorum effigies longo ordine
     spectarentur.
Correct English translation:
For the memory of Drusus, the same honors as were
     decreed for Germanicus were bestowed, with many
      additional honors, as is often the case with
     belated flattery. The funeral, particularly
     distinguished by a procession of images, was
     most illustrious, with Aeneas, the origin of
     the Julian family, and all the kings of Alba,
     as well as Romulus, the founder of the city,
     followed by the Sabine nobility and Attus
     Clausus, along with the other images of the
     Claudii, all observed in a long procession.
```

C Final Filter Prompt

```
You are the final filter of an AI Latin translator.
     Your job is to choose the best out of several
     translation attempts for a Latin text. You
     always simply return the best translation (copy
      the full text). You judge best by accuracy and
      adherence to the Latin text. The goal
     translation is literally accurate and doesn't
     aim for sounding nice in English but rather
     reflecting the Latin grammatical structures.
     Below are a few examples of what a good literal
      translation is.

Examples:

1. Quamquam enim libri nostri complures non modo ad
     legendi, sed etiam ad scribendi studium
     excitaverunt, tamen interdum vereor ne
     quibusdam bonis viris philosophiae nomen sit
     invisum mirenturque in ea tantum me operae et
     temporis ponere. Ego autem quam diu res publica
      per eos gerebatur, quibus se ipsa commiserat,
```

```
     omnes meas curas cogitationesque in eam
     conferebam.
Correct English translation:
Although indeed our many books have stirred up not
     only the enthusiasm for reading, but also for
     writing, nevertheless sometimes I fear that to
     some good men the name of philosophy is hateful
     , and they wonder why I put so much effort and
     time into it. But I, as long as the republic
     was managed by those to whom she herself had
     entrusted, was bringing all my cares and
     thoughts to it.

2. Sanctius his animal mentisque capacius altae
     deerat adhuc et quod dominari in cetera posset:
      natus homo est, sive hunc divino semine fecit
     ille opifex rerum, mundi melioris origo, sive
     recens tellus seductaque nuper ab alto 80
     aethere cognati retinebat semina caeli.
Correct English translation:
More sacred than these, an animal more capable of a
     higher mind, was still absent (and which could
     rule over the rest). Man was born, either made
     from divine seed by that Maker of things, the
     origin of a better world, or the newborn earth,
      lately severed from the lofty heaven, retained
      the seed of the kindred sky.

3. memoriae Drusi eadem quae in Germanicum
     decernuntur, plerisque additis, ut ferme amat
     posterior adulatio. funus imaginum pompa maxime
      inlustre fuit, cum origo luliae gentis Aeneas
     omnesque Albanorum reges et conditor urbis
     Romulus, post Sabina nobilitas, Attus Clausus
     ceteracque Claudiorum effigies longo ordine
     spectarentur.
Correct English translation:
For the memory of Drusus, the same honors as were
     decreed for Germanicus were bestowed, with many
      additional honors, as is often the case with
     belated flattery. The funeral, particularly
     distinguished by a procession of images, was
     most illustrious, with Aeneas, the origin of
     the Julian family, and all the kings of Alba,
     as well as Romulus, the founder of the city,
     followed by the Sabine nobility and Attus
     Clausus, along with the other images of the
     Claudii, all observed in a long procession.
```

D Non-Literal Prompt

```
When provided with Latin text and a literal English
     translation, you are to produce a non-literal,
     lightly interpreted translation. This
     translation should capture the essence,
     cultural context, and nuances of the original
     Latin, offering clarity and insight into its
     meanings, implications, and subtleties that a
     direct translation might miss.

TASK:
1. Read the Latin text and its literal English
     translation.
2. Consider the historical, cultural, and contextual
      background of the Latin text.
3. Provide a non-literal, interpreted English
     translation that conveys the essence and
     meanings of the Latin text while being faithful
      to the text.

INPUT FORMAT:
Latin Text: [Insert Latin text here]
Literal English Translation: [Insert literal English
      translation here]

OUTPUT FORMAT:
Translation: [Insert Latin text here]

EXAMPLE:
Latin Text: "Obstipu re  omn s  nec  t  lia  dicta
     prob runt, ante omnesque Lelex animo maturus
     et aevo, sic ait:   inmensa   est finemque
     potentia caeli non habet, et quicquid superi
     voluere, peractum est, qu que minus dubites,
     tiliae contermina quercus collibus est Phrygiis
      modico circumdata muro; ipse locum vidi; nam
     me Pelopeia Pittheus misit in arva suo quondam
     regnata parenti. haud procul hinc stagnum est,
     tellus habitabilis olim, nunc celebres mergis
     fulicisque palustribus undae; Iuppiter huc
     specie mortali cumque parente venit Atlantiades
      positis caducifer alis. mille domos adiere
     locum requiemque petentes, mille domos clausere
      serae;"
Literal English Translation: "All were awestruck nor
      did they approve of such words, Before
     everyone Lelex, experienced in mind and age, So
      said:   The   power of the sky is great and has
      no end, And whatever the gods have wished is
     accomplished, So that you may doubt less, an
     oak tree Next to a linden tree on Phrygian
     hills is surrounded by a modest wall; I myself
     saw the place; for Pittheus sent me into
     Pelopeian fields that are ruled by his parent.
     By no means far from here is a swamp, once
     habitable land, Now waves, frequented by
     seagulls and water fowl; Jupiter came here in
     mortal disguise, and With his parent came
     Mercury the caduceus bearer after his wings had
      been set aside They approached a thousand
     homes, seeking a place for rest, But locks
     closed a thousand homes;"

NON-LITERAL TRANSLATION:
Translation: The others were startled, and
     disapproved of his words, Lelex above all,
     experienced in mind and years, who said:   The
      power of the gods is great and knows no limit,
      and whatever heaven decrees comes to pass. To
     help convince you, in the hills of Phrygia, an
     oak and a lime tree stand side by side,
     surrounded by a low wall. I have seen the place
     , since Pittheus, king of Troezen, sent me into
      that country, where his father Pelops once
     ruled. There is a swamp not far from there,
     once habitable land but now the haunt of diving
     -birds and marsh-loving coots. Jupiter went
     there, disguised as a mortal, and Mercury, the
     descendant of Atlas, setting aside his wings,
     went with his father, carrying the caduceus. A
     thousand houses they approached, looking for a
     place to rest: a thousand houses were locked
     and bolted.

INSTRUCTIONS:
- Use the literal English translation as a base, but
      enrich it with interpretations that convey the
      text's deeper meanings and implications.
- Aim for clarity and insightfulness, ensuring that
     the translation is accessible to those
     unfamiliar with Latin or its cultural context.
- If uncertain about specific details, make informed
      assumptions based on historical and cultural
     knowledge.
- Return only a translation.
```

E Python Translation Script

```python

from openai import OpenAI
import os
from concurrent.futures import ThreadPoolExecutor,
     as_completed

# Initialize the OpenAI client with your API key
clientGPT = OpenAI(api_key="Insert␣your␣API␣key␣here
     ")

# Function to read prompts from files
def read_prompt(file_path):
 with open(file_path, 'r', encoding='utf-8') as file
      :
 return file.read()

# Reading the prompts
filter_prompt = read_prompt('FinalFilterPrompt.txt')
revision_prompt = read_prompt('RevisionPrompt.txt')
fine_tuned_system_prompt = read_prompt('
     FineTunedSystemPrompt.txt')
```

```python
non_literal_prompt = read_prompt('NonLiteralPrompt.txt')

def generate_translation(text):
 """
 This function generates an initial translation using a fine-tuned GPT model
 and subsequently refines it using a GPT-4 revision model.
 """
 # Initial translation using the fine-tuned model
 initial_response = clientGPT.chat.completions.create(
 model="Insert your fine-tuned model ID here", # Replace with your fine-tuned model ID
 messages=[{
 "role": "system",
 "content": fine_tuned_system_prompt
 }, {
 "role": "user",
 "content": text
 }],
 temperature=0.7,
 top_p=1,
 frequency_penalty=0,
 presence_penalty=0
 )

 translated_text = initial_response.choices[0].message.content

 # Revision of the initial translation
 revised_response = clientGPT.chat.completions.create(
 model="gpt-4o",
 messages=[{
 "role": "system",
 "content": revision_prompt
 }, {
 "role": "user",
 "content": f"Return a corrected translation or the same if it is accurate:\nLatin text: {text}\nTranslation:\n{translated_text}"
 }],
 temperature=0.7,
 top_p=1,
 frequency_penalty=0,
 presence_penalty=0
 )

 return revised_response.choices[0].message.content

def generate_non_literal_translation(text):
 """
 This function expects the input to be structured with both the original Latin text
 and a literal English translation attempt. The input should be provided in the
 following format:

 Latin Text: [Insert Latin text here]
 Literal English Translation: [Insert literal English translation here]

 The function uses these inputs to generate a non-literal, interpreted English
 translation that focuses on conveying the essence, meanings, and contextual
 nuances of the Latin text. The non-literal translation aims to enhance readability
 and accurately reflect the historical and cultural context of the original text.
 """
 response = clientGPT.chat.completions.create(
 model="gpt-4o",
 messages=[{
 "role": "system",
 "content": non_literal_prompt
 }, {
 "role": "user",
 "content": text
 }],
 temperature=0.7,
 top_p=1,
 frequency_penalty=0,
 presence_penalty=0
 )

 non_literal_translation = response.choices[0].message.content

 return non_literal_translation

def perform_latin_translation_workflow(text):
 """
 This function orchestrates the translation workflow, generating multiple candidate translations,
 selecting the best one, and refining it to produce a final output.
 """
 translations = []
 with ThreadPoolExecutor() as executor:
 futures = [executor.submit(generate_translation, text) for _ in range(5)]
 for future in as_completed(futures):
 translations.append(future.result())

 # Prepare the comparison prompt
 comparison_prompt = f"Given these five translations, select the best one based on this Latin provided text: \n{text}\n1. {translations[0]}\n2. {translations[1]}\n3. {translations[2]}\n4. {translations[3]}\n5. {translations[4]}"

 # Make the comparison call to gpt-4o
 best_choice_response = clientGPT.chat.completions.create(
 model="gpt-4o",
 messages=[{
 "role": "system",
 "content": filter_prompt
 }, {
 "role": "user",
 "content": comparison_prompt
 }],
 temperature=0.7,
 top_p=1,
 frequency_penalty=0,
 presence_penalty=0
 )

 # Revise the final output using the same revision model
 final_revised_response = clientGPT.chat.completions.create(
 model="gpt-4o",
 messages=[{
 "role": "system",
 "content": revision_prompt
 }, {
 "role": "user",
 "content": f"Return a corrected translation or the same if it is accurate: \nLatin text: {text}\nTranslation:\n{best_choice_response.choices[0].message.content}"
 }],
 temperature=0.7,
 top_p=1,
 frequency_penalty=0,
 presence_penalty=0
 )

 return final_revised_response.choices[0].message.content
```