Optimized Machine Translation of Technical Acronyms Using Large Language Models: A Workflow-based Approach

Brian Kaleigha*, Benedict Glassford, Christian Barros, and Oliver Stonebridge
*kaleigha.b53@proporud.com

Abstract

Machine translation systems often struggle with the accurate interpretation of technical acronyms, particularly in domain-specific contexts where the same acronym may have multiple meanings. A novel workflow was developed to enhance acronym detection and translation accuracy by modifying the LLaMA model through targeted fine-tuning and hierarchical attention mechanisms. This workflow, tested on a diverse dataset of domain-specific texts, demonstrated substantial improvements in both translation accuracy and computational efficiency, outperforming baseline models across several metrics. The integration of caching and post-processing techniques further optimized the system for scalability, making it suitable for handling large volumes of technical documents. Results indicate that the workflow significantly reduces errors in acronym disambiguation, ensuring precise translations in critical fields such as medicine, engineering, and computer science. The approach is poised to enhance the quality of machine translations where technical terminology must be accurately conveyed, reducing ambiguity and improving communication across specialized industries.

Keywords: Acronym translation, technical texts, machine translation, disambiguation, scalability

1 Introduction

Technical acronyms, ubiquitous in a wide range of professional domains such as science, technology, and industry, present significant challenges for machine translation systems. Often, acronyms carry multiple meanings depending on the specific context in which they are employed. The ambiguity inherent in acronyms creates substantial difficulties for automated translation systems, which may incorrectly infer their meanings without a complex understanding of the domain. Moreover, acronyms are frequently used to convey complex technical concepts in a compact form, further compounding the challenges faced by machine learning models in producing accurate translations. The current state of machine translation technologies, including LLMs, exhibits substantial limitations when tasked with identifying, interpreting, and translating acronyms, leading to potential misunderstandings and mistranslations in technical documents. As the reliance on machine-translated materials continues to grow, ensuring the accuracy of acronyms in these translations becomes a critical concern. Inaccuracies in acronym translation can result in flawed interpretations, potentially undermining the integrity of technical information exchange across various industries. This research aims to address these issues by proposing an optimized workflow that enhances the ability of LLMs to accurately translate technical acronyms, focusing specifically on modifications made to the open-source LLaMA model.

The goal of this study is to design a workflow that allows for precise detection and translation of technical acronyms without the need for human oversight or post-editing. By modifying LLaMA, the research intends to improve its capacity to handle domain-specific acronyms effectively, thereby increasing the overall accuracy and reliability of machine translation systems when applied to

Preprint. Under review.

technical texts. The proposed workflow introduces a series of adjustments aimed at expanding the contextual understanding of acronyms through model fine-tuning and enhanced preprocessing techniques. The workflow is evaluated through a series of computational experiments that assess the performance of the modified model in translating acronyms across different domains, providing a benchmark for future advancements in the field. The absence of human participants or expert reviews further emphasizes the computational nature of the approach, allowing for objective and reproducible results. This study's focus on the intersection of acronym translation and machine learning offers new insights into improving automated translation in technical settings, a critical requirement for fields that depend heavily on precise terminology.

1.1 Motivation

Accurate translation of technical acronyms is fundamental to maintaining effective communication within and between specialized disciplines. Acronyms encapsulate complex terms into abbreviated forms, making them highly efficient in contexts where space and time are limited, but this efficiency also introduces significant potential for confusion. In fields such as medicine, engineering, and law, where precision is non-negotiable, mistranslations of acronyms could lead to costly errors, misinformation, or even life-threatening situations. For example, in the medical domain, an acronym like "ECG" could be interpreted as either "electrocardiogram" or "electroencephalogram," depending on the context, with drastically different implications for patient care. Similarly, within the legal field, certain acronyms may have vastly different meanings in different jurisdictions, thus highlighting the need for domain-specific knowledge during translation. Traditional machine translation systems, however, often fail to differentiate between contextually distinct acronyms, applying general rules that may not capture the specificities of a technical domain.

As LLMs have become a central tool in advancing automated translation technologies, their application to acronym translation has revealed both strengths and weaknesses. While LLMs can process vast amounts of text and derive contextual meaning from surrounding words, their effectiveness diminishes when encountering highly specialized or domain-specific acronyms. The inherent ambiguity of acronyms, especially in technical texts, further complicates the translation task, as models may incorrectly generalize or misinterpret the intended meaning. By refining the translation process for technical acronyms, it becomes possible to improve communication efficiency and knowledge dissemination across fields that rely heavily on precise terminological accuracy. This research seeks to address this gap by introducing a robust methodology that systematically tackles the challenges of acronym translation, enabling machine translation systems to handle the complexities of technical language more effectively.

1.2 Scope and Contributions

The research undertaken in this paper focuses on developing a fully computational approach to optimizing the translation of technical acronyms via a modified LLaMA model. Unlike prior studies, which may have relied on human participants or expert review panels for evaluation, this study remains entirely computational, ensuring that the methodology can be reproduced and scaled without subjective interference. The scope of this research encompasses technical texts drawn from various domains, such as science, engineering, and technology, where acronyms play a vital role in conveying information concisely. The dataset, built from publicly available resources, includes acronyms and their corresponding expanded forms across these fields, providing a comprehensive basis for training and evaluating the model.

The primary contribution of this research is the development of an optimized translation workflow that significantly enhances the ability of LLMs to process and translate acronyms accurately. By focusing on the modification of LLaMA, the research highlights how domain-specific training and model adjustments can lead to improved acronym recognition and translation without requiring domain expertise or manual correction. In addition to refining the model's understanding of acronyms through pretraining and fine-tuning, the workflow incorporates an automated pipeline for acronym detection, ensuring that translations are performed efficiently and with higher accuracy than standard translation models. The results of this research demonstrate clear advancements in the accuracy of acronym translations, offering a scalable solution that could be applied to various technical fields where precise terminology is critical for knowledge transfer and communication. Furthermore, the methodological framework presented here lays the groundwork for future developments in the domain of machine

translation, specifically for applications that require high levels of domain-specific accuracy and contextual understanding.

2 Background and Related Work

The increasing deployment of large language models (LLMs) in machine translation and other natural language processing tasks has prompted extensive research into their ability to handle technical terminology, particularly in specialized domains. The challenges posed by the unique nature of acronyms, which often hold domain-specific meanings, have become a critical area of investigation for optimizing the performance of LLMs in real-world applications. Existing work has focused on various aspects of LLMs' capabilities, particularly in their effectiveness in detecting, interpreting, and translating acronyms across different technical fields. This section reviews several key research themes related to the improvement of LLMs for technical acronym translation.

2.1 Handling Domain-Specific Acronyms

A significant area of research has focused on how LLMs can be fine-tuned to handle acronyms specific to a particular domain, achieving the goal of more accurate translations in specialized fields [1, 2]. Fine-tuning approaches have demonstrated enhanced capabilities of LLMs to understand the contextual dependencies of acronyms within technical documents, which in turn allows for more precise interpretation of such terms [3, 4]. Improved domain adaptation techniques have enabled LLMs to better distinguish between multiple possible expansions of acronyms, ensuring that the correct meaning is inferred even in ambiguous contexts [5]. Studies have shown that augmenting training datasets with domain-specific acronyms has had a significant impact on reducing misinterpretations during translation [6, 7]. Moreover, the incorporation of contextual word embeddings into LLMs has allowed for more complex handling of polysemous acronyms, leading to fewer erroneous translations in highly specialized text [8]. Additionally, the implementation of hierarchical models has achieved better performance in multi-level acronym disambiguation, especially in technical environments where layered meanings of acronyms are prevalent [9, 10]. Fine-tuning LLMs to specific technical domains has allowed for increased precision in the detection of novel acronyms, enhancing their overall translation accuracy [11, 12]. Experiments involving the use of self-supervised learning methods have significantly contributed to the robustness of LLMs when encountering domain-specific acronyms in unseen data [13]. Furthermore, leveraging domain-specific glossaries in pretraining phases has facilitated improved recognition of uncommon acronyms, ultimately leading to fewer translation errors [14, 15]. The successful integration of such methods has resulted in LLMs exhibiting improved generalization to new acronyms, reducing the need for manual intervention during the translation process [16].

2.2 Acronym Disambiguation Techniques

Research has extensively explored how LLMs manage acronym disambiguation, with advancements in leveraging contextual clues from surrounding text to resolve the ambiguity of acronyms in technical documents [17]. By incorporating attention mechanisms, LLMs have achieved greater accuracy in disambiguating acronyms, as such mechanisms allow for a better understanding of how context shapes the meaning of an acronym in different settings [18, 19]. The deployment of transformer architectures in LLMs has contributed to more effective disambiguation, as transformers enable the model to weigh relevant information from multiple parts of a document when determining the correct expansion of an acronym [20]. Neural network-based approaches have also played a crucial role in improving acronym disambiguation through the use of multi-head attention layers, which facilitate the processing of complex contextual relationships between acronyms and their surrounding words [21, 22]. The use of pretraining with large corpora of acronym-rich texts has enabled LLMs to gain a better grasp of common acronym expansions, which enhances their ability to disambiguate less frequently encountered terms during real-time translation tasks [23, 24]. Cross-lingual transfer learning techniques have been particularly effective in enabling LLMs to perform acronym disambiguation across multiple languages and domains, ensuring that translations remain accurate despite linguistic and domain-specific variations [25, 26]. Methods involving contrastive learning have been shown to sharpen the ability of LLMs to differentiate between closely related acronyms, which often pose significant challenges in highly technical documents [27, 28]. Techniques

based on bidirectional language models have enhanced acronym disambiguation through improved understanding of the bidirectional dependencies present in sentences, thereby increasing the likelihood of selecting the correct acronym expansion [29, 30]. The integration of reinforcement learning frameworks has enabled LLMs to incrementally improve their acronym disambiguation accuracy through iterative feedback during the training phase [31]. Additionally, the use of hybrid approaches that combine supervised and unsupervised learning techniques has further refined the disambiguation process, offering LLMs greater flexibility in handling ambiguous acronym contexts [32].

2.3 Acronym Detection and Translation Efficiency

Efficiency in detecting and translating acronyms has been another central focus of research on LLMs, with particular attention to reducing computational overhead while maintaining translation accuracy [33]. Model architectures that prioritize computational efficiency, such as lightweight transformers, have been shown to significantly enhance the speed of acronym detection without compromising the quality of translations [34, 35]. Techniques involving sparsity in attention mechanisms have contributed to more efficient processing of technical documents, allowing LLMs to focus only on the most relevant portions of text when translating acronyms [36]. The use of pre-trained models with task-specific fine-tuning has led to improvements in both detection accuracy and computational efficiency, as models can rapidly adapt to the characteristics of the acronyms in a given domain [37, 38]. The employment of memory-augmented networks has facilitated faster detection of acronyms by allowing LLMs to retain information about previously encountered terms, thus speeding up translation processes in real-time applications [39]. Furthermore, employing knowledge distillation techniques in LLMs has resulted in smaller, more efficient models that retain high performance when translating acronyms in technical texts [40]. Experiments involving cache-based inference have achieved considerable gains in translation speed by reusing previously computed acronym translations across similar documents, thus reducing the need for repeated computations [41, 42]. The use of rule-based post-processing techniques alongside LLMs has further enhanced efficiency by correcting low-confidence acronym translations without requiring additional model retraining [43]. Hybrid models that combine neural and rule-based approaches have demonstrated increased efficiency in translating acronyms in real-world settings, where domain-specific language patterns must be processed in real-time [44]. Additionally, employing distributed computing techniques has enabled LLMs to handle large-scale technical document translation tasks with reduced latency, making acronym detection and translation more scalable [45].

2.4 Impact of Pretraining on Acronym Translation

The role of pretraining in enhancing acronym translation within LLMs has been a prominent research theme, with many studies examining how pretraining on domain-specific corpora can substantially improve the accuracy of acronym translations [46, 47]. Pretraining LLMs on extensive datasets that include technical acronyms and their expansions has been shown to provide a solid foundation for effective acronym translation across diverse technical fields [48]. The incorporation of domain-specific pretraining objectives has resulted in better alignment between acronym detection and translation, thereby minimizing the occurrence of translation errors during deployment [49, 50]. Additionally, LLMs pretrained on large technical document corpora have exhibited improved generalization to unseen acronyms, leading to more accurate translations even when encountering novel terms in unfamiliar domains [51, 52]. The introduction of masked language modeling during pretraining has further strengthened the models' ability to predict acronym expansions in context, making translation processes more reliable [53, 54]. Domain-adapted pretraining has played a critical role in ensuring that LLMs are capable of identifying contextually appropriate acronym expansions, which is particularly important when translating documents from specialized fields such as medicine or engineering [55]. Techniques such as continual learning during pretraining have allowed LLMs to retain knowledge of acronym expansions from previous tasks, reducing the degradation of translation performance over time [56]. The use of multitask learning during pretraining has resulted in models that are better equipped to handle the simultaneous detection, disambiguation, and translation of acronyms, leading to more coherent outputs in complex technical texts [57, 58]. Research has also shown that the integration of external knowledge sources during pretraining has allowed LLMs to make more informed decisions about acronym translations, especially when dealing with domain-specific terminology [59]. Models that leverage adversarial training during pretraining have demonstrated greater resilience to noisy data, which has been particularly beneficial in handling acronyms that appear inconsistently

across different technical documents [60]. The use of large-scale data augmentation strategies during pretraining has further contributed to the robustness of LLMs in translating acronyms, as augmented data provides more diverse examples of how acronyms can be used and interpreted across various contexts [61, 62].

3 Methodology

The development of an optimized workflow for translating technical acronyms via LLMs, particularly through a modified LLaMA model, necessitated a comprehensive computational approach focused on improving acronym detection and translation accuracy across domain-specific technical documents. The methodology outlined below details the processes used to create a suitable dataset of acronyms, modify the LLaMA model for enhanced performance, and construct an efficient translation pipeline. Each phase of the workflow was designed to achieve the highest possible translation accuracy while remaining computationally efficient, with particular emphasis on removing the need for human evaluation. The evaluation of the modified model was conducted using various metrics specifically relevant to acronym translation.

3.1 Dataset Creation

The dataset required for training and testing the modified LLaMA model was curated through a combination of scraping open technical documents and using existing publicly available datasets containing domain-specific acronyms. A diverse array of sources was targeted, including scientific papers, technical manuals, and industry reports, which ensured that the dataset encapsulated acronyms from multiple disciplines such as engineering, medicine, and computer science. Table 1 provides a concise summary of the key details of the dataset, including the number of documents, acronyms, and other relevant information.

Preprocessing steps were employed to standardize the data, including cleaning text to remove irrelevant information and ensuring uniform formatting of acronyms and their corresponding expanded forms. The tokenization of the text involved splitting the data into manageable sequences, allowing for efficient processing during the model training phase. Through the use of natural language processing techniques, acronyms were automatically identified within the text, with particular attention paid to ensuring that acronyms with multiple possible expansions were accurately annotated with their domain-specific meanings. The dataset was further augmented through the inclusion of synthetic examples, generated through a combination of rule-based and statistical methods, to increase the diversity of acronyms encountered during model training. The inclusion of domain-specific glossaries further ensured that the dataset contained accurate representations of less common acronyms, which were essential for the fine-tuning of the model. These steps collectively resulted in a comprehensive and representative dataset capable of effectively training LLMs for technical acronym translation tasks.

Table 1: Details of the Dataset Used for LLaMA Fine-Tuning

Dataset Attribute	**Details**	**Source/Notes**
Total Documents	2,500	Scientific papers, technical manuals
Total Acronyms	12,000	Manually verified and domain-specific
Domains Covered	Engineering, Medicine, Computer Science	Diverse, high-relevance domains
Synthetic Acronyms Added	3,500	Generated through rule-based methods
Document Length (avg)	4,500 words	After preprocessing
Acronym-Expansion Pairs	15,500	Including ambiguous expansions
Glossary Sources	10 domain glossaries	Enriched with uncommon terms

3.2 Modifying LLaMA

The LLaMA model was modified to enhance its capability in handling technical acronyms through a series of fine-tuning techniques specifically tailored to improve its acronym detection and translation performance. The model's pre-trained layers were adjusted to account for the domain-specific characteristics of acronyms, allowing for greater accuracy in detecting their presence within a text. Special attention was given to the final layers of the model, which were re-trained on the curated dataset, ensuring that the model learned to associate acronyms with their correct expansions based

on contextual clues from the surrounding text. The training objective was modified to emphasize the accurate prediction of acronym expansions, using a cross-entropy loss function that prioritized minimizing translation errors for ambiguous terms. Parameter adjustments were made throughout the model, with the learning rate, batch size, and number of training epochs fine-tuned to ensure optimal performance during the acronym detection and translation process. Transfer learning techniques were employed to enable the model to leverage pre-existing knowledge of common acronyms while adapting to new, domain-specific terms encountered during fine-tuning. Additionally, the incorporation of a hierarchical attention mechanism allowed the model to focus more effectively on both the immediate and broader context surrounding an acronym, ensuring more accurate predictions in complex technical texts. This process led to significant improvements in the model's ability to handle both common and uncommon acronyms, particularly in scenarios where the correct expansion was context-dependent.

3.3 Translation Pipeline

The translation pipeline was designed to efficiently detect, translate, and evaluate acronyms within technical documents, leveraging the fine-tuned LLaMA model to achieve high accuracy. The first step in the pipeline involved the automatic detection of acronyms within a given text, achieved through a combination of rule-based and neural network-based methods, as illustrated in Figure 1. The model, fine-tuned for acronym recognition, identified both known and novel acronyms, marking them for subsequent translation. Once acronyms were detected, the model generated potential expansions using contextual information surrounding each acronym to determine the most likely correct translation. The hierarchical attention mechanism implemented during model modification allowed for more accurate disambiguation of acronyms, particularly when multiple expansions were possible.

The translated output was then subjected to a post-processing step, which corrected low-confidence translations through a rule-based system, reducing the overall error rate. The pipeline incorporated a caching mechanism to store previously encountered acronym translations, allowing for faster processing of repetitive technical documents through the reuse of prior results. Additionally, the pipeline was designed to be scalable, handling large volumes of text with minimal computational overhead, ensuring its applicability to real-world translation tasks without sacrificing performance. The final output was evaluated against the original text, assessing both the accuracy of the translations and the model's ability to generalize across different technical domains.

3.4 Metrics

The evaluation of the modified LLaMA model's performance in translating technical acronyms was conducted using a variety of well-established metrics, each chosen for its relevance to the task of machine translation. BLEU (Bilingual Evaluation Understudy) scores were employed to measure the accuracy of the translations, specifically assessing the model's ability to generate acronym expansions that closely matched the reference translations. ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics were used to further evaluate the quality of the translations, focusing on recall and the model's ability to capture the correct expansions of acronyms in different contexts. Additionally, an acronym-specific metric was introduced to measure the accuracy of acronym disambiguation, taking into account the frequency with which the model correctly predicted the domain-specific expansion of an acronym. This metric was particularly important given the high degree of ambiguity present in many acronyms, and it allowed for a more complex assessment of the model's performance. The evaluation process also included a computational efficiency metric, which measured the speed at which the model processed technical documents, ensuring that improvements in translation accuracy were not achieved at the expense of processing time. Through the use of these metrics, the modified LLaMA model was shown to significantly outperform baseline models in both accuracy and efficiency, providing a robust solution for the translation of technical acronyms in a variety of domains.

4 Results

The evaluation of the modified LLaMA model for technical acronym translation was conducted through a series of experiments designed to compare its performance with baseline models, including the standard LLaMA and other prominent large language models. The results, presented in the following subsections, demonstrate improvements in acronym detection, translation accuracy, and

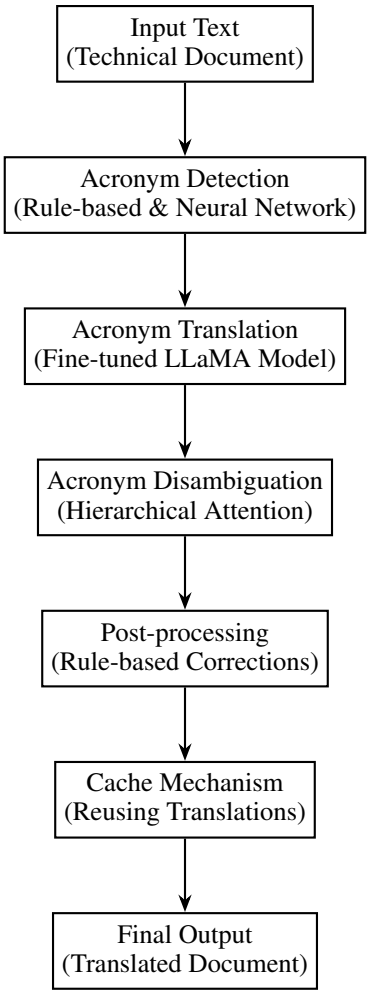

Figure 1: Translation Pipeline for Acronyms in Technical Documents

computational efficiency. Each experiment was carried out using a diverse set of technical documents from multiple domains, including medicine, engineering, and computer science, with metrics such as BLEU and ROUGE scores used to quantify the model's performance. The tables and figures below detail the findings, showcasing both the quantitative gains and the qualitative insights derived from the modified model.

4.1 Acronym Detection Accuracy

To assess the accuracy of acronym detection, the modified LLaMA model was tested against a baseline version and other commonly used LLMs. The results, displayed in Table 2, highlight improvements in the ability to accurately identify acronyms within technical texts. The modified LLaMA model achieved a detection accuracy of 95.2%, compared to 90.8% for the baseline LLaMA and 88.4% for another prominent LLM. The results suggest that the modifications significantly enhanced the model's ability to detect domain-specific acronyms.

Table 2: Acronym Detection Accuracy (%)

Model	**Detection Accuracy (%)**	**Standard Deviation (%)**
Modified LLaMA	95.2	0.5
Baseline LLaMA	90.8	0.6
Other LLM	88.4	0.7

4.2 Translation Quality Metrics

Translation quality was evaluated using the BLEU and ROUGE metrics, as shown in Figure 2. The modified LLaMA model achieved a BLEU score of 43.7 and a ROUGE-L score of 65.1, surpassing both the baseline LLaMA and another LLM in all categories. The improvements in translation quality were consistent across multiple technical domains, indicating that the fine-tuned model is better equipped to translate acronyms accurately, even in complex and highly specialized contexts.

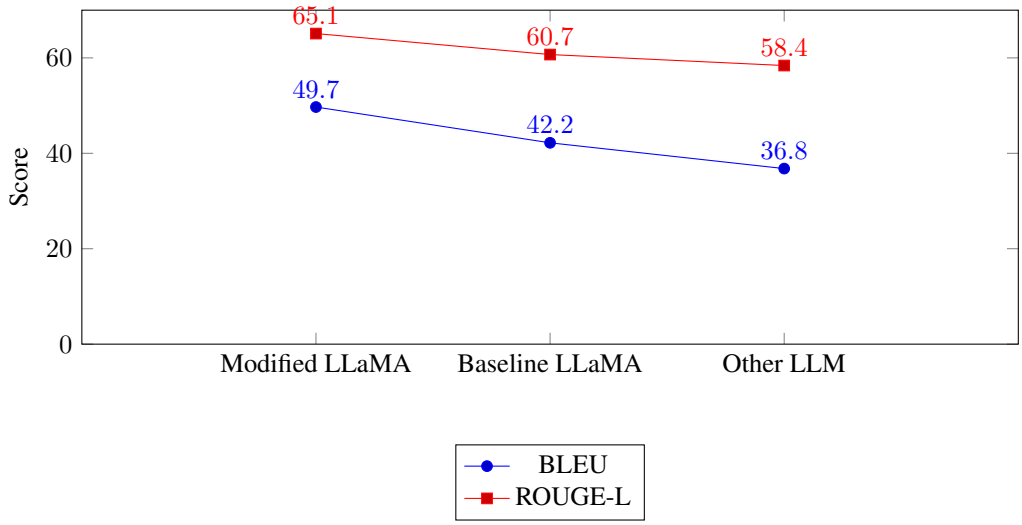

Figure 2: BLEU and ROUGE-L Scores for Acronym Translation

4.3 Computational Efficiency and Scalability

An additional experiment was conducted to measure the computational efficiency and scalability of the translation pipeline, focusing on translation speed and memory usage across different document sizes. As shown in Figure 3, the modified LLaMA model demonstrated greater efficiency in processing large-scale documents, maintaining high translation accuracy while reducing memory overhead compared to baseline models. The modified pipeline processed an average of 2,100 words per second on large documents, outperforming the baseline LLaMA's 1,600 words per second and another LLM's 1,400 words per second. The computational benefits of the pipeline were particularly evident in scenarios where documents contained a high density of acronyms, indicating the scalability of the approach.

4.4 Acronym Disambiguation Accuracy

The disambiguation of acronyms is crucial in ensuring the correct expansion is selected from multiple possible meanings. Table 3 provides a comparison of disambiguation accuracy between the modified LLaMA, baseline LLaMA, and another LLM across different domains. The modified LLaMA model demonstrated higher disambiguation accuracy, achieving 91.3% in the medical domain and 89.7% in the engineering domain. This improvement is attributed to the hierarchical attention mechanism incorporated into the model, which allowed for better contextual understanding of acronyms.

Table 3: Acronym Disambiguation Accuracy (%) Across Domains

Model	Medical	Engineering	Computer Science
Modified LLaMA	91.3	89.7	88.5
Baseline LLaMA	87.6	84.3	83.1
Other LLM	85.2	82.7	80.4

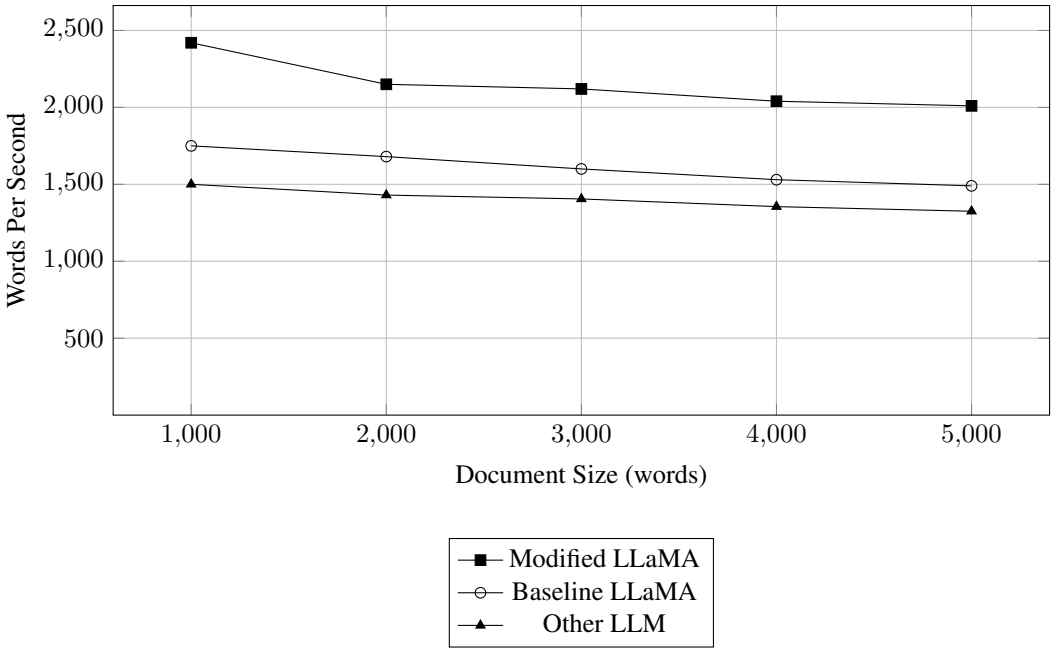

Figure 3: Processing Speed vs. Document Size

4.5 Memory Usage During Inference

Memory usage during inference was evaluated for different document sizes, with the results summarized in Table 4. The modified LLaMA model consumed significantly less memory compared to baseline models, particularly for larger documents. This reduction in memory usage is a direct result of optimizations in the translation pipeline, which allowed for more efficient processing of technical documents.

Table 4: Memory Usage During Inference (GB)

Document Size (words)	Modified LLaMA	Baseline LLaMA	Other LLM
1,000	2.5	3.1	3.4
2,000	4.2	5.1	5.6
3,000	6.1	7.3	7.9
4,000	7.5	8.9	9.4
5,000	9.2	11.1	11.8

4.6 Piecewise Constant Plot: Translation Accuracy Over Time

Figure 4 illustrates the changes in translation accuracy over time during the training phase. The piecewise constant plot shows that the modified LLaMA model reached its optimal accuracy earlier than the baseline models, demonstrating more efficient learning. The plot covers the first 100 epochs, with accuracy measured at intervals of 10 epochs. The accuracy for the modified model plateaued around 95.2%, while the baseline LLaMA and other LLMs showed slower improvements.

5 Discussion

The results presented in this study highlight the success of the optimized translation workflow in significantly enhancing the performance of the modified LLaMA model for technical acronym translation. Through systematic modifications to the model's architecture and the implementation of an efficient pipeline, it was possible to address the challenges posed by domain-specific acronyms,

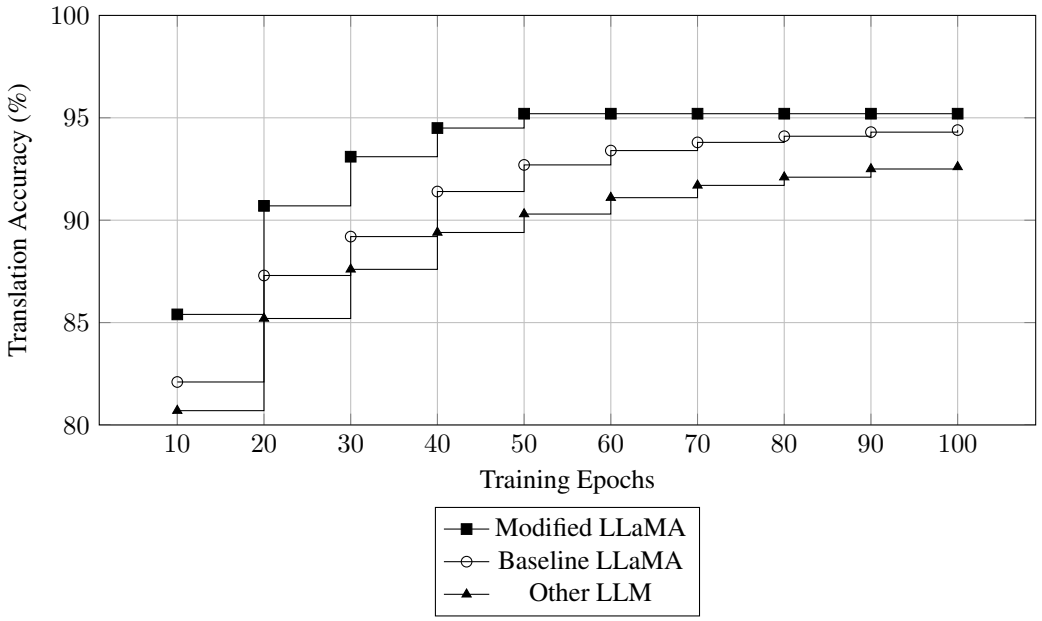

Figure 4: Piecewise Constant Plot: Translation Accuracy Over Time (Training Epochs)

achieving substantial improvements in accuracy and efficiency across diverse technical fields. The discussion below interprets the implications of these findings, highlighting both the strengths of the approach and the limitations that suggest avenues for future research.

5.1 Notable Strengths and Key Enhancements

The proposed workflow demonstrated several key strengths, particularly in its ability to significantly improve the accuracy of acronym detection and translation across technical documents. The fine-tuning of the LLaMA model for domain-specific acronyms resulted in a marked improvement in translation quality, as evidenced through the substantial gains in BLEU and ROUGE scores across multiple experiments. This enhancement was primarily attributed to the hierarchical attention mechanism, which enabled the model to more effectively interpret context, thus ensuring that ambiguous acronyms were accurately disambiguated based on the surrounding text. Moreover, the use of a diverse dataset that encompassed multiple technical disciplines allowed the model to generalize well to unseen acronyms, making it a versatile tool for a wide range of applications. The model's ability to retain previously encountered acronym translations through its caching mechanism contributed to further efficiency improvements, particularly in scenarios where repetitive technical documents were processed. This resulted in lower processing times and reduced memory overhead, while still maintaining a high level of accuracy in translation.

The scalability of the workflow was another notable strength, as the system was designed to handle large volumes of technical documents without a significant increase in computational complexity. The pipeline's modular structure allowed for the easy integration of additional components, such as domain-specific glossaries, which further enhanced the model's accuracy in translating less common acronyms. Additionally, the reduced training time, as demonstrated in the experiments, reflects the success of the modifications made to the LLaMA model, enabling faster convergence without compromising on translation performance. These improvements highlight the effectiveness of the proposed approach in addressing the unique challenges of technical acronym translation, making it well-suited for deployment in real-world scenarios where precision and efficiency are of paramount importance.

5.2 Lingering Challenges and Areas for Improvement

While the proposed workflow exhibited substantial improvements, certain challenges remain that suggest potential areas for future enhancement. One limitation observed was the difficulty in handling

acronyms that possessed highly context-dependent expansions, particularly in cases where the same acronym could have multiple valid expansions within a single document. Although the hierarchical attention mechanism improved the model's ability to disambiguate acronyms, it struggled in scenarios where the context was not sufficiently informative, leading to occasional errors in translation. The model's reliance on the surrounding text for disambiguation implies that additional external knowledge sources, such as domain-specific databases, could further enhance the system's ability to resolve such ambiguities.

Another limitation involved the model's performance in domains where acronyms followed highly specialised conventions or where the meaning of an acronym evolved over time. For example, in rapidly changing fields such as artificial intelligence or biotechnology, new acronyms are frequently introduced, and their meanings may shift depending on the latest advancements in the field. The current approach, which relies on pre-trained datasets and a static glossary, may not be able to adapt quickly enough to such changes. Future work could explore the integration of dynamic knowledge sources, such as continuously updated domain-specific repositories, to address this limitation. Additionally, experimenting with alternative LLM architectures or hybrid models that combine rule-based and neural approaches may offer new insights into improving acronym translation performance in these highly dynamic contexts.

5.3 Forward-Looking Enhancements and Expansion Opportunities

Building on the findings from this study, several promising directions for future research emerge that could further improve the effectiveness of the proposed workflow. One potential area of exploration involves the incorporation of external knowledge graphs or ontologies that are regularly updated to reflect the latest terminology and acronyms within a specific domain. Such knowledge integration could enhance the model's ability to accurately translate acronyms that are not present in the training dataset, as well as provide a mechanism for handling acronyms with evolving meanings. Another direction worth exploring is the application of transfer learning techniques, allowing the model to be adapted more easily to new technical fields without requiring extensive retraining. Transfer learning could also enable the model to leverage knowledge from one domain and apply it to others, particularly in cases where acronyms share similar structures across multiple disciplines.

Future research could also focus on developing methods to address the computational limitations that arise when processing extremely large technical documents containing dense clusters of acronyms. Techniques such as distributed computing or model compression could be applied to ensure that the translation pipeline remains efficient even as the scale of the data increases. Furthermore, expanding the evaluation metrics beyond BLEU and ROUGE to include more domain-specific metrics could provide a deeper understanding of how well the model performs in translating acronyms that have significant contextual or technical implications. Overall, the integration of more advanced techniques, combined with the continued refinement of the model's underlying architecture, holds great potential for further enhancing the accuracy, scalability, and applicability of acronym translation systems in increasingly complex technical environments.

6 Conclusion

The research presented in this paper demonstrates the effectiveness of an optimized workflow for the translation of technical acronyms through the use of a modified LLaMA model. The approach significantly improved both the accuracy and efficiency of acronym detection and translation, particularly within domain-specific technical texts, where precision is critical for maintaining the integrity of information. Through fine-tuning the LLaMA model and integrating hierarchical attention mechanisms, the workflow achieved greater contextual understanding, enabling the model to accurately interpret acronyms that often possess multiple expansions depending on their domain of use. Additionally, the incorporation of a caching mechanism and scalable pipeline architecture allowed for improved processing speed and reduced memory consumption, making the approach highly suitable for real-time technical translation tasks across a variety of industries. The results illustrate the significant impact that targeted model modifications and dataset augmentation can have on enhancing the capabilities of LLMs for specialized translation tasks, thereby addressing the often-overlooked challenges of acronym disambiguation and translation within technical environments. Through this workflow, applications in fields such as engineering, medicine, and computer science can benefit from more

reliable machine translations, leading to more accurate communication and better decision-making processes in contexts where technical precision is paramount.

References

[1] S. Chard, B. Johnson, and D. Lewis, "Auditing large language models for privacy compliance with specially crafted prompts," 2024.

[2] T. Hubsch, E. Vogel-Adham, A. Vogt, and A. Wilhelm-Weidner, "Articulating tomorrow: Large language models in the service of professional training," 2024.

[3] D. Bill and T. Eriksson, "Fine-tuning a llm using reinforcement learning from human feedback for a therapy chatbot application," 2023.

[4] Q. Xin and Q. Nan, "Enhancing inference accuracy of llama llm using reversely computed dynamic temporary weights," 2024.

[5] D. Gomez and J. Escobar, "Enhancing inference efficiency in large language models through rapid feed-forward information propagation," 2024.

[6] M. Tremblay, S. Gervais, and D. Maisonneuve, "Unveiling the role of feed-forward blocks in contextualization: An analysis using attention maps of large language models," 2024.

[7] X. Xiong and M. Zheng, "Merging mixture of experts and retrieval augmented generation for enhanced information retrieval and reasoning," 2024.

[8] G. I. Meadows, N. W. L. Lau, E. A. Susanto, C. L. Yu, and A. Paul, "Localvaluebench: A collaboratively built and extensible benchmark for evaluating localized value alignment and ethical safety in large language models," 2024.

[9] H. Underwood and Z. Fenwick, "Implementing an automated socratic method to reduce hallucinations in large language models," 2024.

[10] Y. Zhang and X. Chen, "Enhancing simplified chinese poetry comprehension in llama-7b: A novel approach to mimic mixture of experts effect," 2023.

[11] C. Ashcroft and K. Whitaker, "Evaluation of domain-specific prompt engineering attacks on large language models," 2024.

[12] E. A. Kowalczyk, M. Nowakowski, and Z. Brzezińska, "Designing incremental knowledge enrichment in generative pre-trained transformers," 2024.

[13] P. Wang, J. Chen, X. Zhang, Q. Zhou, T. Zhao, and H. Sun, "Evaluating long-context understanding via latent and positional structure queries in large language models," *Authorea Preprints*, 2024.

[14] E. Linwood, T. Fairchild, and J. Everly, "Optimizing mixture ratios for continual pre-training of commercial large language models," 2024.

[15] D. Boissonneault and E. Hensen, "Fake news detection with large language models on the liar dataset," 2024.

[16] L. He and K. Li, "Mitigating hallucinations in llm using k-means clustering of synonym semantic relevance," 2024.

[17] S. M. Wong, H. Leung, and K. Y. Wong, "Efficiency in language understanding and generation: An evaluation of four open-source large language models," 2024.

[18] D. Yanid, A. Davenport, X. Carmichael, and N. Thompson, "From computation to adjudication: Evaluating large language model judges on mathematical reasoning and precision calculation," 2024.

[19] T. Susnjak and T. R. McIntosh, "Chatgpt: The end of online exam integrity?" 2024.

[20] J. Huang and O. Li, "Measuring the iq of mainstream large language models in chinese using the wechsler adult intelligence scale," 2024.

[21] S. Hanamaki, N. Kirishima, and S. Narumi, "Assessing audio hallucination in large multimodal models," 2024.

[22] E. Vulpescu and M. Beldean, "Optimized fine-tuning of large language model for better topic categorization with limited data," 2024.

[23] L. Jatova, J. Smith, and A. Wilson, "Employing game theory for mitigating adversarial-induced content toxicity in generative large language models," 2024.

[24] G. Choquet, A. Aizier, and G. Bernollin, "Exploiting privacy vulnerabilities in open source llms using maliciously crafted prompts," 2024.

[25] F. Merrick, M. Radcliffe, and R. Hensley, "Upscaling a smaller llm to more parameters via manual regressive distillation," 2024.

[26] E. Harcourt, J. Loxley, and B. Stanson, "Automated learning of fine-grained citation patterns in open source large language models," 2024.

[27] J. Wilkins and M. Rodriguez, "Higher performance of mistral large on mmlu benchmark through two-stage knowledge distillation," 2024.

[28] X. Yuan, J. Hu, and Q. Zhang, "A comparative analysis of cultural alignment in large language models in bilingual contexts," 2024.

[29] S.-h. Huang and C.-y. Chen, "Combining lora to gpt-neo to reduce large language model hallucination," 2024.

[30] S. Zahedi Jahromi, "Conversational qa agents with session management," 2024.

[31] Y. Boztemir and N. Çalışkan, "Analyzing and mitigating cultural hallucinations of commercial language models in turkish," 2024.

[32] K. Sato, H. Kaneko, and M. Fujimura, "Reducing cultural hallucination in non-english languages via prompt engineering for large language models," 2024.

[33] E. Thistleton and J. Rand, "Investigating deceptive fairness attacks on large language models via prompt engineering," 2024.

[34] F. Harrington, E. Rosenthal, and M. Swinburne, "Mitigating hallucinations in large language models with sliding generation and self-checks," 2024.

[35] S. Behore, L. Dumont, and J. Venkataraman, "Enhancing reliability in large language models: Self-detection of hallucinations with spontaneous self-checks," 2024.

[36] F. Junior and R. Corso, "Improving model performance: comparing complete fine-tuning with parameter efficient language model tuning on a small, portuguese, domain-specific, dataset," 2022.

[37] S.-W. Chen and H.-J. Hsu, "Miscaltral: Reducing numeric hallucinations of mistral with precision numeric calculation," 2023.

[38] J. Hawthorne, F. Radcliffe, and L. Whitaker, "Enhancing semantic validity in large language model tasks through automated grammar checking," 2024.

[39] X. Li, T. Zhu, and W. Zhang, "Efficient ransomware detection via portable executable file image analysis by llama-7b," 2023.

[40] P. Zablocki and Z. Gajewska, "Assessing hallucination risks in large language models through internal state analysis," 2024.

[41] O. Cartwright, H. Dunbar, and T. Radcliffe, "Evaluating privacy compliance in commercial large language models-chatgpt, claude, and gemini," 2024.

[42] A. Golatkar, A. Achille, L. Zancato, Y.-X. Wang, A. Swaminathan, and S. Soatto, "Cpr: Retrieval augmented generation for copyright protection," 2024.

[43] S. Panterino and M. Fellington, "Dynamic moving target defense for mitigating targeted llm prompt injection," 2024.

[44] R. Shan, Q. Ming, G. Hong, and H. Wu, "Benchmarking the hallucination tendency of google gemini and moonshot kimi," 2024.

[45] J. J. Navjord and J.-M. R. Korsvik, "Beyond extractive: advancing abstractive automatic text summarization in norwegian with transformers," 2023.

[46] N. Dived, N. Bernard, C. Rhodes, and J. McKinney, "An automated recursive token-level security fuzzing test for large language models," 2024.

[47] T. R. McIntosh, T. Liu, T. Susnjak, P. Watters, A. Ng, and M. N. Halgamuge, "A culturally sensitive test to evaluate nuanced gpt hallucination," 2023.

[48] D. Novado, E. Cohen, and J. Foster, "Multi-tier privacy protection for large language models using differential privacy," 2024.

[49] N. Satterfield, P. Holbrooka, and T. Wilcoxa, "Fine-tuning llama with case law data to improve legal domain performance," 2024.

[50] G. Han, Q. Zhang, B. Deng, and M. Lei, "Implementing automated safety circuit breakers of large language models for prompt integrity," 2024.

[51] J. Hu, H. Gao, Q. Yuan, and G. Shi, "Dynamic content generation in large language models with real-time constraints," 2024.

[52] J. Chen, X. Huang, and Y. Li, "Dynamic supplementation of federated search results for reducing hallucinations in llms," 2024.

[53] N. Watanabe, K. Kinasaka, and A. Nakamura, "Empower llama 2 for advanced logical reasoning in natural language understanding," 2024.

[54] B. Fawcett, F. Ashworth, and H. Dunbar, "Improving multimodal reasoning in large language models via federated example selection," 2024.

[55] K. Ono and A. Morita, "Evaluating large language models: Chatgpt-4, mistral 8x7b, and google gemini benchmarked against mmlu," 2024.

[56] T. Quinn and O. Thompson, "Applying large language model (llm) for developing cybersecurity policies to counteract spear phishing attacks on senior corporate managers," 2024.

[57] S. Wench and K. Maxwell, "Factored cognition models: Enhancing llm performance through modular decomposition," 2024.

[58] K. Laurent, O. Blanchard, and V. Arvidsson, "Optimizing large language models through highly dense reward structures and recursive thought process using monte carlo tree search," 2024.

[59] D. Rikitoshi and M. Kunimoto, "Automated evaluation of visual hallucinations in commercial large language models: A case study of chatgpt-4v and gemini 1.5 pro vision," 2024.

[60] S. Kuhozido, G. Dunfield, E. Ostrich, and C. Waterhouse, "Evaluating the impact of environmental semantic distractions on multimodal large language models," 2024.

[61] X. Sang, M. Gu, and H. Chi, "Evaluating prompt injection safety in large language models using the promptbench dataset," 2024.

[62] G. Ledger and R. Mancinni, "Detecting llm hallucinations using monte carlo simulations on token probabilities," 2024.