CSCI-6509 - Evaluating RAG, Reflection Tuning, and Chain of Thought in LLM Translations: Individual and Combined Effects

Ajaykumar Premkumar Nair
B00968276
ajay.nair@dal.ca

Manish Shankar Jadhav
B00969328
mn649712@dal.ca

December 20, 2024

Abstract

This research investigates the effectiveness of two advanced techniques—Retrieval-Augmented Generation (RAG) and Chain of Thought—in enhancing Large Language Model (LLM) translation capabilities, specifically focusing on conversational French-to-English translation. The study particularly emphasizes challenging aspects of casual French conversation. While these techniques have shown promise individually in various natural language processing tasks, their specific impact on translation quality, both independently and in combination, remains unexplored, especially in preserving the cultural and contextual nuances of informal communication. Our study is to evaluate these techniques with everyday conversational French-to-English translation tasks. The evaluation framework combines automated metrics (BLEU) with human assessment criteria, including adequacy, fluency, and pragmatic accuracy, to provide a comprehensive quality assessment. Special attention will be paid to the preservation of humor and cultural context in the translations. We will assess each technique in isolation to establish individual performance metrics, examine their combination to identify potential effects, and analyze the impact of implementing both techniques simultaneously. The research aims to quantify improvements in translation accuracy and quality against a baseline of standard LLM translations, particularly focusing on how well each approach handles the inherent ambiguity and cultural specificity of conversational French [1]. By comprehensively evaluating these techniques in the context of conversational French-to-English translation, this study seeks to contribute valuable insights to the field of machine translation and provide practical recommendations for optimizing LLM translation systems. The findings will help inform the development of more effective translation architectures and demonstrate whether the computational overhead of implementing multiple enhancement techniques is justified by corresponding improvements in translation quality.

1 Problem Statement

This research addresses the critical challenge of accurately translating conversational French to English using Large Language Models (LLMs), with a specific focus on handling informal expressions, cultural nuances, and contextual understanding [2] [3]. Current LLM translations often fail to capture the subtle intricacies of everyday conversation and are not culturally aware of idiomatic expressions, leading to translations that appear artificial or lose their intended meaning [4]. This problem is significant because informal communication represents a substantial portion of real-world language use, and inaccurate translations can lead to miscommunication and cultural disconnects in increasingly global digital interactions. Additionally, as LLMs become more integrated into communication tools, the need for natural-sounding conversational translations becomes crucial for effective cross-cultural dialogue. The scope of our project encompasses the evaluation of two specific enhancement techniques: Retrieval-Augmented Generation (RAG) and Chain of Thought. We will assess these techniques both independently and in combination, using a comprehensive evaluation framework that includes both automated metric (BLEU) and human assessment criteria, specifically testing their effectiveness on a curated dataset of conversational French texts that include informal expressions, slang, and cultural references. Our goal is to quantifiably determine which technique or combination of techniques most effectively improves the quality of conversational French-to-English translations while considering the practical implications of implementation complexity and computational requirements.

2 Related Work

The field of machine translation (MT) has made significant strides with the advent of neural approaches, particularly with the introduction of Transformer-based architectures like BERT [5] and GPT [6]. However, conversational translation, which involves maintaining cultural nuances, humor, and pragmatic accuracy, remains a challenging subfield. Most research has focused on formal or technical text, with limited attention to informal conversational data, which often contains ambiguous or context-dependent expressions.

2.1 Conversational Machine Translation

Prior work on conversational machine translation has highlighted the importance of contextual understanding for accurate translations. Tiedemann and Scherrer [7] explored the integration of context into neural MT systems and demonstrated that including dialogue history improves translation fluency and adequacy. Similarly, recent research by Bawden et al. [8] evaluated discourse-level translation models and emphasized the challenge of capturing inter-sentential coherence in conversations. However, the translation of cultural and pragmatic elements in casual speech remains an underexplored area.

2.2 Retrieval-Augmented Generation (RAG)

The use of Retrieval-Augmented Generation (RAG) has gained traction as a means to improve language model performance by incorporating external knowledge. Lewis et al. [9] introduced RAG for knowledge-intensive tasks, demonstrating improvements in factual accuracy and context comprehension.

2.3 Chain of Thought Reasoning

Chain of Thought (CoT) prompting has been explored as a method to enhance LLM reasoning capabilities, particularly in tasks requiring multi-step logical deductions [10]. While CoT has not been extensively applied to MT, its ability to improve complex reasoning aligns with the needs of conversational translation, where understanding context, idioms, and nuances often requires sequential reasoning.

2.4 Evaluation Metrics in Machine Translation

Translation quality evaluation traditionally relies on automated metrics such as BLEU [11], which measures n-gram overlap between the hypothesis and reference translations. However, the limitations of BLEU in capturing semantic nuances and cultural context have been noted [12]. Recent works advocate combining automated metrics with human assessments, particularly for tasks like conversational translation where subjective factors such as humor and cultural adaptation play a critical role [13].

2.5 Challenges in Informal Language Translation

Translating informal language poses unique challenges due to the presence of idiomatic expressions, slang, and humor. These issues highlight the need for techniques that go beyond standard word-level translation, making the integration of advanced methods like RAG and CoT particularly relevant.

3 Methodology

Our research methodology employs a systematic approach to evaluate the effectiveness of Retrieval-Augmented Generation (RAG) [9] and Chain of Thought [10] in improving French-to-English conversational translations. The steps are as follows:

- **Establish Baseline Performance:** To set a reference point for comparison, we evaluate the translation quality of a standard Large Language Model (LLM) without any enhancement techniques. This baseline serves as the benchmark to quantify improvements achieved by RAG and Chain of Thought.
- **Data Curation:** The dataset used in this research was sourced from a publicly available Kaggle dataset containing approximately 12 million rows of French-English translation pairs. To ensure efficient processing and focus on conversational contexts, we selected a representative subset of 10,000 rows. This subset includes:
  - Informal expressions and slang commonly used in conversational French.
  - Idiomatic phrases that require contextual and cultural understanding.
  - Complex sentence structures to challenge model reasoning capabilities.
- **Technique Implementation and Testing:** Each enhancement technique was implemented and tested independently to evaluate its individual contributions:

- **RAG:** The Retrieval-Augmented Generation method integrates external knowledge retrieval into the translation process. It retrieves contextual information from the dataset to assist in translating slang, idiomatic expressions, and culturally specific phrases.
- **Chain of Thought:** This reasoning-based approach enables the model to break down and translate complex or nuanced expressions step by step, improving semantic accuracy in challenging cases.

3.1 LLM Selection

The selection of Large Language Models (LLMs) for this study was guided by the following criteria:

- **Compact yet effective for translation:** We sought models that are relatively small in size but have been specifically trained or fine-tuned for high-quality translation tasks. These models typically have fewer parameters but are optimized for specific translation tasks, making them suitable for evaluating the effect of Retrieval-Augmented Generation (RAG) on models with limited capacity.
- **Large and general-purpose:** We also included larger, general-purpose LLMs that were not explicitly trained for translation tasks. These models have a greater number of parameters, which allows them to handle a broader range of tasks, including more complex reasoning. While not fine-tuned for translation, they benefit from their large-scale training on diverse data, which we hypothesize may enable them to handle conversational translation with greater flexibility.
- **Efficient and runnable on basic hardware:** A key consideration was ensuring that the selected models could be executed on standard hardware setups without requiring specialized equipment like high-performance GPUs. This was important to make the models accessible for researchers with limited computational resources and to ensure that the evaluation could be conducted on widely available infrastructure.
- **Accessible via an API:** For practical and scalable experimentation, we prioritized models that are available through an API, which simplifies the integration process and allows for easy access to the models without requiring deep expertise in deployment or infrastructure management. This ensures reproducibility and accessibility for future work in this domain.

The reasoning behind these selection criteria was to provide a balanced evaluation across different types of models. By comparing a small, translation-focused LLM with a larger, general-purpose LLM, we aim to assess how RAG influences models of varying capacities. Specifically, the smaller model offers insights into the effectiveness of RAG on specialized, task-oriented architectures, while the larger, general-purpose LLM allows us to explore whether the additional parameters and broader training data improve conversational translation performance, even in the absence of domain-specific training. Ultimately, this comparison will help us understand how RAG can enhance LLMs across different contexts and model architectures.

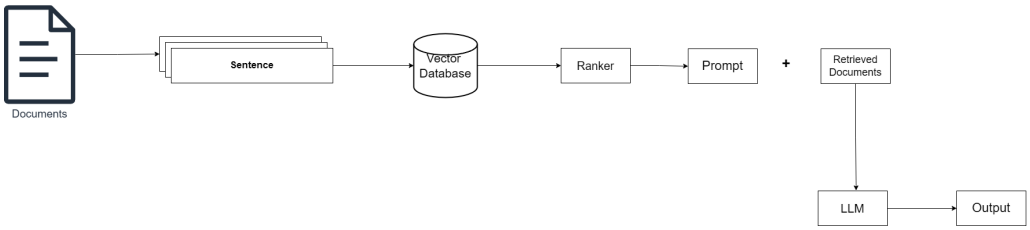

Figure 1: RAG Architecture

3.1.1 Final LLMs selected

- facebook/nllb-200-distilled-1.3B
- ChatGPT 4o
- ChatGPT o1-preview

3.2 Embedding Model Selection

- **Evaluation Metrics:** The performance of the baseline and each enhancement technique was assessed using the BLEU metric, which measures the lexical overlap and alignment between generated translations and reference translations. BLEU provided a quantitative evaluation of the improvements achieved by RAG and Chain of Thought.
- **Comparison and Analysis:** Finally, the results were compared to determine the relative effectiveness of RAG and Chain of Thought, both independently and in combination. The analysis focused on identifying the approach that yielded the most significant improvement in conversational translation quality while considering practical implementation factors such as computational efficiency and ease of integration.

4 RAG Architecture

The proposed Retrieval-Augmented Generation (RAG) architecture as shown in Figure 1 comprises a pipeline of components designed to enhance language model outputs through document-grounded responses. The system begins with document preprocessing, where input documents undergo sentence-level segmentation. These segments are then transformed into vector embeddings and stored in a vector database, enabling efficient similarity-based retrieval operations. When processing queries, the ranking component evaluates semantic similarity between the query and stored vectors, identifying the most relevant document segments. The system then constructs a prompt that combines the user query with these retrieved contexts. This composite prompt, along with the relevant document segments, is fed into a Large Language Model (LLM) for final output generation. The architecture's key advantage lies in its ability to ground LLM responses in specific document content, significantly reducing hallucination while maintaining the model's natural language generation capabilities. The vector database approach ensures scalable retrieval performance, while the ranking mechanism maintains output relevance by selecting only the most pertinent document contexts for inclusion in the generation process.

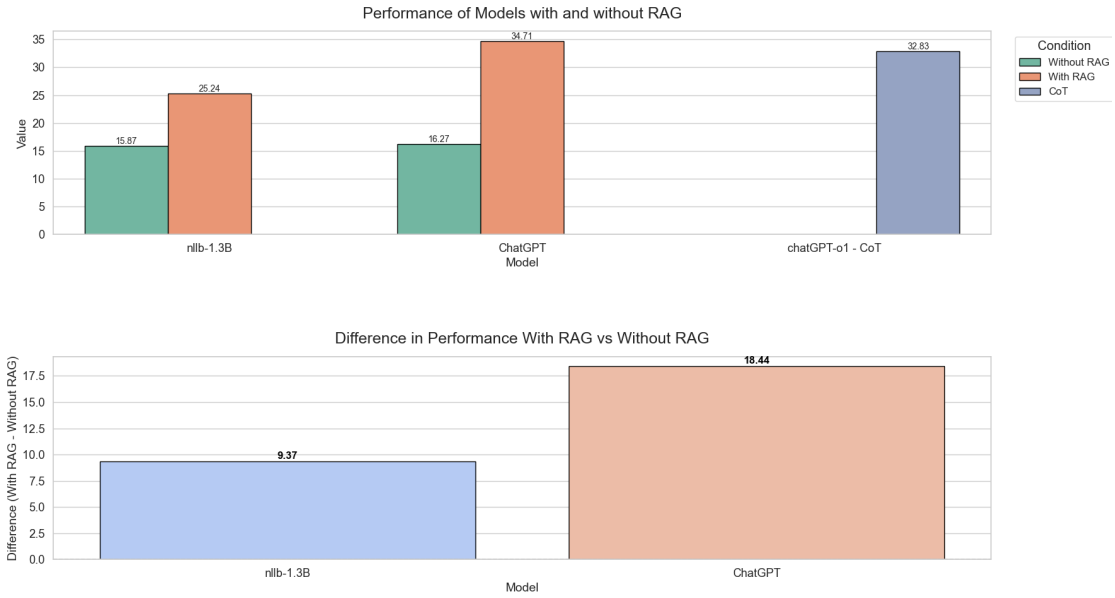

Figure 2: Performance Metrics

5 Results

The performance of the evaluated models with and without the Retrieval-Augmented Generation (RAG) enhancement is summarized in Figure 2. The results highlight the impact of RAG on translation quality, as measured by the BLEU score.

5.1 Performance Analysis

Figure 2 (top) compares the BLEU scores of three models: NLLB-1.3B, the ChatGPT base model, and the ChatGPT-o1 model with Chain of Thought (CoT). The following observations were made:

The NLLB-1.3B model achieved a BLEU score of 15.87 without RAG and improved to 25.24 with RAG, representing a significant gain of 9.37 points. The ChatGPT model demonstrated a substantial improvement from 16.27 without RAG to 34.71 with RAG, resulting in an 18.44-point increase. The ChatGPT-o1 CoT model achieved the highest BLEU score of 32.83, but the CoT technique's contributions are evaluated independently of RAG.

5.2 Comparison of Improvement Across Models

Figure 1 (bottom) quantifies the improvement in BLEU scores due to RAG.

The NLLB-1.3B model showed a modest improvement relative to its baseline performance, highlighting the utility of RAG in enhancing smaller LLMs. The ChatGPT model displayed the largest performance gain, underlining the synergy between RAG and larger, context-aware models.

5.3 Interpretation of Results

The results indicate that RAG significantly enhances the ability of LLMs to handle conversational French-to-English translations. The larger performance gains observed in the ChatGPT models suggest that the combination of RAG and advanced reasoning capabilities amplifies translation quality. However, CoT's BLEU score of 32.83 emphasizes the role of logical reasoning in translating complex, nuanced expressions.

We also observe that Chain of Thought (CoT) is a highly effective method for improving the model's performance, as the CoT score alone surpasses that of ChatGPT with Retrieval-Augmented Generation (RAG).

Overall, the combination of RAG and Chain of Thought proves to be highly effective in addressing the challenges posed by informal expressions and idiomatic phrases in conversational translations.

6 Conclusion

The results demonstrate the impact of both Retrieval-Augmented Generation (RAG) and Chain of Thought (CoT) techniques on improving French-to-English conversational translations. The performance analysis reveals several key insights:

The implementation of RAG substantially enhanced translation quality across all tested models, with particularly notable improvements in the ChatGPT model, which achieved an 18.44-point increase in BLEU score (from 16.27 to 34.71). This suggests that RAG's external knowledge integration capabilities are especially effective when combined with more sophisticated language models. The NLLB-1.3B model showed a more modest but still significant improvement of 9.37 points (from 15.87 to 25.24), indicating that even smaller, specialized models can benefit from RAG enhancement. Chain of Thought reasoning proved to be highly effective, with the ChatGPT-o1 CoT model achieving a BLEU score of 32.83. This strong performance highlights the importance of structured reasoning approaches in handling the complexities of conversational translation, particularly when dealing with cultural nuances and idiomatic expressions.

These findings have several important implications for the field of machine translation:

1. The significant improvements achieved through RAG and CoT suggest that future translation systems should consider incorporating these techniques, particularly for conversational and culturally-nuanced content.

2. The varying degrees of improvement across different model sizes indicate that the choice of enhancement technique should be considered in relation to the underlying model architecture and available computational resources.

3. The success of both RAG and CoT in improving translation quality suggests that hybrid approaches combining multiple enhancement techniques may offer the most promising path forward for addressing the challenges of conversational translation.

In conclusion, both RAG and Chain of Thought are valuable techniques for improving conversational French-to-English translation, with the potential to significantly enhance the quality of machine translation systems. The results provide a foundation for future work in developing more sophisticated and culturally aware translation systems.

References

[1] P. Chen, J. Tang, and A. Birch, "Evaluating the translation performance of large language models based on euas-20," *arXiv preprint arXiv:2408.03119*, 2024.

[2] H. Namukwaya, "Beyond translating french into english: Experiences of a non-native translator," *TranscUlturAl: A Journal of Translation and Cultural Studies*, vol. 5, p. 61, 03 2014.

[3] Z. Wang, *Study on the Importance of Cultural Context Analysis in Machine Translation*, 05 2013, vol. 212, pp. 29–35.

[4] P. Singh, M. Patidar, and L. Vig, "Translating across cultures: Llms for intralingual cultural adaptation," 2024. [Online]. Available: https://arxiv.org/abs/2406.14504

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019. [Online]. Available: https://arxiv.org/abs/1810.04805

[6] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020. [Online]. Available: https://arxiv.org/abs/2005.14165

[7] J. Tiedemann and Y. Scherrer, "Neural machine translation with extended context," in *Proceedings of the Third Workshop on Discourse in Machine Translation*, B. Webber, A. Popescu-Belis, and J. Tiedemann, Eds. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 82–92. [Online]. Available: https://aclanthology.org/W17-4811

[8] R. Bawden, R. Sennrich, A. Birch, and B. Haddow, "Evaluating discourse phenomena in neural machine translation," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, M. Walker, H. Ji, and A. Stent, Eds. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 1304–1313. [Online]. Available: https://aclanthology.org/N18-1118

[9] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," 2021. [Online]. Available: https://arxiv.org/abs/2005.11401

[10] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," 2023. [Online]. Available: https://arxiv.org/abs/2201.11903

[11] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, P. Isabelle, E. Charniak, and D. Lin, Eds.

Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. [Online]. Available: https://aclanthology.org/P02-1040

[12] C. Callison-Burch, M. Osborne, and P. Koehn, "Re-evaluating the role of Bleu in machine translation research," in *11th Conference of the European Chapter of the Association for Computational Linguistics*, D. McCarthy and S. Wintner, Eds. Trento, Italy: Association for Computational Linguistics, Apr. 2006, pp. 249–256. [Online]. Available: https://aclanthology.org/E06-1032

[13] L. Fischer and S. Läubli, "What's the difference between professional human and machine translation? a blind multi-language study on domain-specific MT," in *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, A. Martins, H. Moniz, S. Fumega, B. Martins, F. Batista, L. Coheur, C. Parra, I. Trancoso, M. Turchi, A. Bisazza, J. Moorkens, A. Guerberof, M. Nurminen, L. Marg, and M. L. Forcada, Eds. Lisboa, Portugal: European Association for Machine Translation, Nov. 2020, pp. 215–224. [Online]. Available: https://aclanthology.org/2020.eamt-1.23

7 Appendix

7.1 Useful Links

In this section, we provide a few useful links related to the project:

- Hugging Face - NLLB 200 Distilled 1.3B Model
- OpenAI Chat
- Kaggle - English to French Translation Dataset