Mitigating Translation Hallucinations in Large Language Models: A Chain of Thought and RAG-Based Approach

Jinhu Qi

2024–2025 The Chinese University of Hong Kong Ph.D. Application

Abstract

This research addresses the limitations of Large Language Models (LLMs) in translation, particularly in handling cultural nuances, specialized terms, and hallucinations. Despite their superior performance compared to traditional tools like Google Translate, LLMs often struggle with complex linguistic and cultural contexts due to their sequence-to-sequence architecture.

To overcome these challenges, this study integrates Retrieval-Augmented Generation (RAG) and Chain of Thought (CoT) reasoning into LLM-based translation. By combining RAG's context-aware retrieval with CoT's logical reasoning, the research aims to improve translation accuracy and reduce hallucinations. Three experimental approaches—RAG, CoT, and a hybrid model—will be systematically evaluated.

The study's outcomes will advance NLP and machine translation by enhancing the robustness and interpretability of LLMs. This project aligns with CUHK's strengths in AI and NLP, with practical applications in social media moderation and cross-cultural communication.

1 Introduction

1.1 Background

In today's globalized world, where social media and international communication have become essential aspects of daily life, individuals from different countries encounter significant challenges due to their distinct cultural and linguistic backgrounds. Although there are many translation tools available, such as Google Translate or Microsoft's Bing Translate, these tools often fail to provide accurate translations[Gao et al., 2013], and in many cases, they produce mistranslations or errors. When translation is inaccurate, it can reduce communication efficiency in both social media interactions and everyday conversations, potentially leading to misunderstandings[Singh et al., 2024].

This research aims to investigate the limitations of Large Language Models (LLMs) in addressing translation challenges, particularly related to linguistic and cultural diversity. While LLMs outperform traditional translation tools like Google Translate, their reliance on sequence-to-sequence architecture often leads to inaccuracies in context-sensitive translations. This study will explore how the integration of Retrieval-Augmented Generation (RAG) and Chain of Thought (CoT) mechanisms can improve LLMs' ability to provide accurate and culturally sensitive translations. LLMs, with their context-aware processing, offer advantages over traditional translation systems, yet still face limitations in precision. This study aims to systematically investigate and compare various translation strategies, with a focus on determining the most efficient approach. By leveraging techniques such as Retrieval-Augmented Generation (RAG) and the "Chain of Thought" (CoT) mechanism, this research will also explore how the integration of vector databases and knowledge graphs can mitigate hallucination issues in LLMs' translation tasks.

1.2 Analogous to Human Thinking Systems

Although LLMs outperform traditional translation software, such as Google Translate, due to their superior contextual understanding and handling of complex linguistic structures, there is still room for improvement in translation tasks. The architecture of LLMs, based on transformers, relies on a sequence-to-sequence method for predicting the next token. This process can be compared to human cognitive systems, specifically System 1 and System 2:

- **System 1 (Intuitive Thinking)**: Fast, automatic, and effortless, relying on intuition, heuristics, and pattern recognition. It uses memory and prior experiences to make quick judgments[Alter et al., 2007].
- **System 2 (Analytical Thinking)**: Slow, deliberate, and effortful, involving logical reasoning and step-by-step problem-solving. It is activated in unfamiliar or complex situations[Shankar, 2013].

Currently, LLMs operate more like System 1, generating translations based on pattern recognition and memory without deeper analysis. While this approach works well in many cases, it struggles when faced with complex sentences, misspellings, or culturally nuanced terms, leading to translation errors. To improve LLMs' translation performance, it is essential to integrate System 2-like logical reasoning, allowing LLMs to deeply analyze and reason through the meaning of a sentence before generating a translation.

For instance, when a sentence contains multiple typos, LLMs may have difficulty understanding the context, resulting in mistranslation. However, by training LLMs to employ System 2-style reasoning—using logic to infer the overall meaning of the sentence—the model could generate a more accurate translation. Research has shown that logical reasoning, such as step-by-step reasoning through Chain of Thought (CoT)[Lei and Deng, 2023], significantly improves LLM accuracy in handling complex tasks. Although the current transformer architecture is sequence-based, it lacks inherent logical reasoning capabilities[Thakkar and Jagdishbhai, 2023], highlighting the need for further improvements in this area to enhance translation performance[Pirozelli et al., 2023].

1.3 Problem Statement

- **Cultural Nuances and Communication Gaps:** Current LLMs face challenges in accurately representing cultural nuances[Fung et al., 2024], limiting their ability to bridge communication gaps caused by linguistic and cultural diversity.

- **Sequence-to-Sequence Limitations:** Due to their transformer-based sequence-to-sequence architecture (predicting the next token), LLMs often rely on intuition and memory rather than logical reasoning[Lu et al., 2023]. This leads to inaccuracies in translating sentence meanings and specialized terms, especially in cross-cultural contexts. While LLMs outperform tools like Google Translate, they still struggle with precision.

- **Lack of Transparent Reasoning:** LLMs lack transparent reasoning chains, which results in trust issues and the risk of misinformation[Brown and Hawe, 2024]. Users are unable to fully understand or verify the logic behind LLM outputs, leading to reduced trust in cross-cultural communication scenarios.

1.4 Research Objectives

- To integrate knowledge graphs and vector databases into LLMs via Retrieval-Augmented Generation (RAG) techniques, creating transparent (white-box) reasoning mechanisms for translation tasks.

- To apply the "Chain of Thought" mechanism, allowing models to reason through multiple steps of translation by leveraging prompt engineering. This will enable LLMs to approach translation tasks in a step-by-step manner, rather than relying solely on intuition and memory to generate output.

- To conduct comparative experiments between various translation approaches—RAG-based, Chain of Thought-based, and combined methods—evaluating their translation performance using well-defined metrics and standards.

- To reduce hallucinations in LLM-generated translations and enhance their ability to accurately represent and convey cross-cultural information.

- To improve the accuracy and cultural sensitivity of LLMs, specifically in the contexts of social media and real-world applications, where misunderstandings caused by linguistic diversity are prevalent.

1.5 Significance of the Study

This research will provide a robust framework for using LLMs, enhanced by the Chain of Thought and RAG techniques, to address translation challenges in social media and real-world communication. By focusing on the transparency of the reasoning process and optimizing translation accuracy, the study will offer practical solutions to the communication barriers that hinder social cohesion and cross-cultural understanding. Furthermore, the outcomes of this research will contribute to the field of natural language processing, particularly in advancing the application of LLMs in cross-cultural and multilingual settings.

2 Literature Review

2.1 Large Language Models (LLMs)

2.1.1 Summary of Existing Research

Recent advancements in Large Language Models (LLMs) have demonstrated their ability to handle complex linguistic structures. A 2024 study utilizing a fine-tuned ChatGLM-6B model for sentiment analysis of classical Song Dynasty poetry achieved a notable 0.840 F1 score, highlighting the model's proficiency in navigating intricate linguistic nuances[Ihnaini et al., 2024]. Furthermore, LLMs such as GPT-4 have shown promise in bridging cross-cultural gaps, as evidenced by a study analyzing over 80,000 Wikipedia articles to compare ideological differences between Russian and English authors[Panasyuk et al., 2024].

2.1.2 Critical Analysis

While LLMs are adept at handling large datasets and complex sentence structures, challenges remain. For instance, LLMs often struggle with the accurate translation of specialized technical content and culture-specific elements[Bielykh, 2024]. The performance of these models can be inconsistent when handling region-specific or culturally embedded concepts[ind, 2021]. Additionally, although LLMs have improved contextual understanding, their reliance on pre-trained data can lead to inaccuracies, especially when dealing with dynamic or evolving terminology in specialized fields.

2.1.3 Research Gap

The limitations of LLMs in accurately capturing culture-specific nuances and translating domain-specific terms reveal a significant gap in current research[Esfandiari and Khatibi, 2022]. Existing models often fail to address these subtleties, which can result in misunderstandings or distorted translations[Zhou and Sun, 2022]. This study aims to integrate knowledge graphs and vector databases into LLMs using RAG techniques, improving the transparency and accuracy of translations in cross-cultural contexts.

2.2 Retrieval-Augmented Generation (RAG)

2.2.1 Summary of Existing Research

Retrieval-Augmented Generation (RAG) has been developed to address the issue of hallucinated content in LLM outputs by grounding responses in factual, externally retrieved data[Gao et al., 2023]. The use of RAG allows LLMs to generate more reliable and transparent responses by retrieving information from external, up-to-date sources, which is particularly valuable in rapidly evolving fields such as medicine and technology[Lyu et al., 2024].

2.2.2 Critical Analysis

While RAG significantly improves the factual accuracy of LLM outputs, its implementation is not without challenges. The dependence on external data sources can sometimes slow down the response time, and the accuracy of the information retrieved is dependent on the quality of the external source[Yu et al., 2024]. Furthermore, RAG systems have yet to fully explore their potential in handling deeply embedded cultural content, where nuanced understanding is crucial for accurate translation.

2.2.3 Research Gap

Although RAG enhances LLM performance by retrieving contextual information, there is limited research on its application in culturally nuanced translation tasks[Yu et al., 2024][Radeva et al., 2024]. This study proposes to fill this gap by combining RAG with knowledge graphs to improve the accuracy of LLMs in cross-cultural communication, particularly for domain-specific content.

2.3 Chain of Thought (CoT)

2.3.1 Summary of Existing Research

The Chain of Thought (CoT) approach plays a critical role in enhancing LLMs' reasoning capabilities. CoT prompts LLMs to generate a sequence of intermediate reasoning steps, enabling the models to tackle complex tasks more systematically and logically[Lei and Deng, 2023]. This progressive reasoning approach has proven especially effective in multi-step reasoning tasks such as arithmetic, symbolic reasoning, and common-sense reasoning[Liu et al., 2023].

2.3.2 Critical Analysis

Despite its success in structured reasoning tasks, CoT's application in translation tasks remains underexplored. Most research focuses on CoT's efficacy in tasks requiring logical steps, like arithmetic, but few studies have investigated how CoT could be used to enhance translation tasks[Lei and Deng, 2023], where cultural and linguistic context plays a pivotal role.

2.3.3 Research Gap

There is a lack of research on how CoT could improve LLMs' handling of cross-cultural translations, especially in tasks that require more than factual retrieval, but instead demand nuanced understanding and multi-step reasoning[Ranaldi et al., 2023]. One challenge with CoT methods is that LLMs can be unreliable for evaluating intermediate ideas. Researchers are exploring methods such as pairwise comparisons to address this issue[Zhang et al., 2024]. This study aims to investigate how CoT, integrated with RAG, can enhance the step-by-step translation of culturally complex content.

2.4 Knowledge Graphs and Vector Databases

2.4.1 Summary of Existing Research

Knowledge graphs provide structured information about entities and their relationships, improving the reasoning capabilities of LLMs by enabling them to perform more complex queries[Tao et al., 2024]. Vector databases, on the other hand, allow efficient semantic search, retrieving information based on meaning rather than keywords[Ma et al., 2024].

2.4.2 Critical Analysis

While knowledge graphs help in structuring domain-specific information, they still face challenges in capturing rapidly evolving or culturally specific knowledge. Additionally, vector databases, though effective in semantic search, are limited by the quality of embeddings and the granularity of information they retrieve[Wang et al., 2024]. The scalability of these systems in real-time translation tasks also remains an open challenge.

2.4.3 Research Gap

The combination of knowledge graphs and vector databases in the context of translation tasks is still underexplored, particularly for cross-cultural and domain-specific content[Tao et al., 2024]. This study seeks to address this gap by integrating these tools with RAG techniques to improve translation accuracy in real-time applications.

2.5 Existing Translation Tools and Technologies

2.5.1 Summary of Existing Research

Existing translation tools, such as Google Translate and Bing Translator, often face challenges in accurately translating colloquial and culturally nuanced content[Aldawsar, 2024]. Research has shown that LLM-based translation tools, such as Bing AI Chat, perform better in translating informal language, such as Arabic colloquial expressions, when compared to traditional translation systems[Son and Kim, 2023].

2.5.2 Critical Analysis

While LLM-based tools outperform traditional machine translation systems in terms of cultural sensitivity and context understanding, they still struggle with lexical and structural ambiguity[Alzain et al., 2024]. Furthermore, these tools often fail to provide accurate translations for domain-specific content that requires deep understanding of both linguistic and cultural context.

2.5.3 Research Gap

Current translation tools lack the ability to consistently handle complex cultural nuances and specialized terms[Son and Kim, 2023], particularly in informal or technical settings[Ahmed and Lenchuk, 2024]. This study aims to fill this gap by developing a framework that integrates RAG, knowledge graphs, and CoT to enhance LLMs' translation accuracy and cultural sensitivity in real-world applications.

3 Research Methodology

3.1 Research Design

This research will follow a pipeline-based approach, akin to Langchain, where different components are integrated to achieve the research objectives. The first step involves reviewing reports from well-known open-source LLMs (such as Meta, Qwen, or Mistral) to determine which model is most suitable for translation tasks. After identifying the model, extensive testing will be conducted to integrate vector-RAG and graph-RAG with the chosen LLMs, evaluating the translation quality after knowledge integration. In a parallel branch, Chain-of-Thought (CoT) techniques will be adapted to translation logic, followed by comparative experiments to analyze performance differences.

The study will consist of three main experimental approaches:

- **RAG-based approach:** The LLM will first extract unfamiliar terms, ambiguous words, technical terms, and culturally specific vocabulary from the user query. These terms will be fed into pre-stored vector databases and knowledge graphs for retrieval. The output will then be compared between retrieving from the vector database, the knowledge graph, and combining both, to evaluate translation quality.

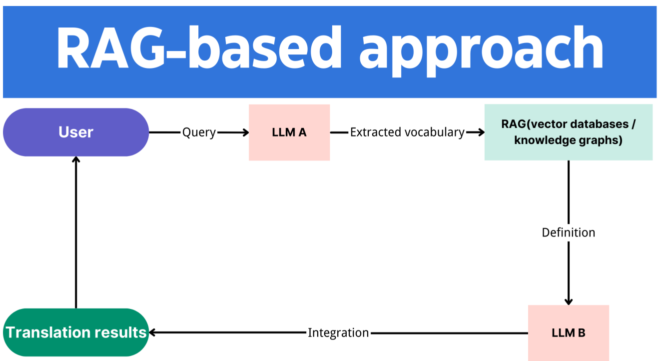

Figure 1: RAG-based approach

- **CoT-based approach:** Various CoT models will be designed through prompt engineering, teaching the LLMs how to reason through translation step-by-step. User queries will be passed to CoT-trained LLMs, and comparative experiments will assess which CoT approach yields the best translation results.

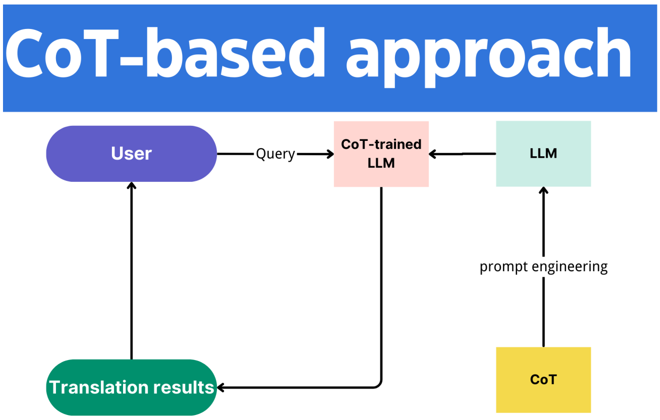

Figure 2: CoT-based approach

- **Combined RAG and CoT approach:** In this approach, CoT-trained LLMs will first extract unfamiliar, ambiguous, or culturally specific terms, feeding them into either a vector database, a knowledge graph, or a combination of both. A separate CoT-trained LLM with a self-critic mechanism will then process the output to produce the final translation, with experiments comparing the translation quality across various setups.

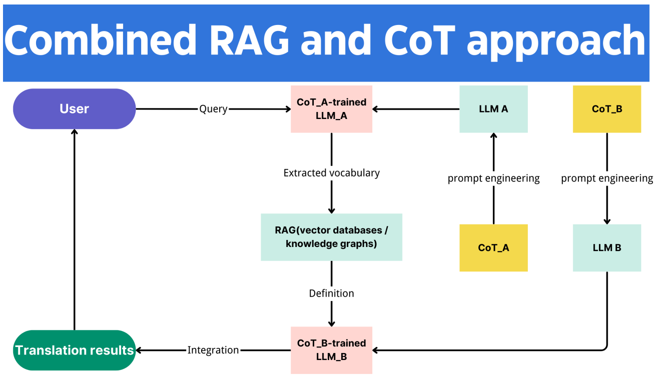

Figure 3: Combined RAG and CoT approach

- **LLM Enhanced MT Translation approach:** The approach begins by sending the user's language query to a machine translation system (MT), such as Google Translate or Microsoft's Bing Translator[Vashee, 2024]. The translation output from the MT is then fed to a large language model (LLM) trained by CoT for further refinement and improvement. The result is returned to the user after the CoT-trained LLM enhances the translation quality. The final product is a high-quality translation in the preferred language, evaluated by a specially designed evaluation system.

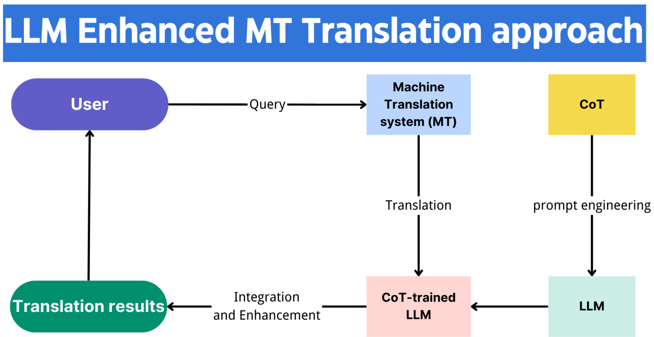

Figure 4: LLM Enhanced MT Translation approach

3.2 Data Collection Methods

The data for this study will be sourced from two categories:

- **Publicly available translation evaluation datasets:** These datasets will include pre-labeled inputs and outputs for assessing translation quality.
- **Large-scale multilingual and multicultural social media datasets:** These will be used for training and testing the models. To supplement the data, large-scale surveys may be conducted online to gather user feedback on translation accuracy, using majority vote to determine ground truth.

The datasets will be carefully curated to ensure diverse cultural contexts, enabling the testing of cross-cultural applicability of the models.

3.3 Analysis Techniques

The performance of the translations will be evaluated based on:

- **Translation accuracy:** How accurately the LLMs translate content from one language to another.
- **Cultural sensitivity:** The model's ability to understand and reflect cultural nuances in the translation.
- **Hallucination reduction:** The effectiveness of RAG and CoT in reducing hallucinations in the generated translations.

Comparative analysis will be conducted between:

- Microsoft's and Google's standard translation systems
- Standard LLMs
- LLMs integrated with different RAG types (vector databases, knowledge graphs)
- CoT-trained LLMs
- Combined RAG and CoT approaches.

3.4 Tools and Technologies

The study will use a variety of tools, including:

- Open-source LLMs of various sizes (with different parameter scales)
- Tools for vector databases and knowledge graphs
- RAG integration techniques
- LLM invocation tools such as *llama-factory*
- CoT design schemes and strategies for translation reasoning.

4 Expected Contributions

4.1 Anticipated Results

- This research aims to enhance LLMs' accuracy in translation tasks, including improvements in translation correctness, cultural sensitivity, and the accurate handling of specialized terms.
- By integrating vector-RAG, graph-RAG, and CoT mechanisms, this study will reduce hallucinations in LLM outputs, leading to more reliable and context-aware translations.

4.2 Potential Impacts on the Field of Computer Science and Translation in the Linguistic Field

- This research is expected to significantly advance the fields of NLP and machine translation by developing LLMs that are not only more robust and interpretable but also better equipped to handle cross-cultural content. By integrating RAG and CoT mechanisms, this study will contribute to the development of LLMs that can provide more accurate, culturally aware translations, reduce hallucinations, and offer practical solutions for applications such as social media moderation, multilingual communication, and real-time translation services.
- Practical applications include social media moderation, automated translation, and multilingual communication systems, making significant advancements in real-world NLP applications.

4.3 Alignment with CUHK's Research Strengths

- CUHK's strong research focus on AI, NLP, and machine learning makes it an ideal institution to conduct this interdisciplinary research, aligning with the university's strengths in cutting-edge technological advancements.

5 Expected Timeline

- **First Month (4 weeks)**: Comprehensive literature review, dataset collection, and initial model design.

- **Second and Third Months (6 weeks)**: LLM development, testing, and integration with RAG and CoT mechanisms.
- **Fourth and Fifth Months (6 weeks)**: Conduct experiments, analyze results, and refine the model based on the findings.
- **Sixth Month (4 weeks)**: Final testing, evaluation of the model, and preparation of results for publication.

6 Resources Required

6.1 Computational Resources

- Access to high-performance computing clusters for training large-scale models.
- NVIDIA RTX 6000 Ada or other high-end GPUs to support efficient model training and testing.

6.2 Datasets

- Publicly available translation evaluation datasets.
- Large-scale multilingual and multicultural social media datasets for training and evaluating the models.

6.3 Software Requirements

- Tools like Neo4j for knowledge graph management.
- Tools like FAISS for vector database management.
- LLM invocation tools such as *llama-factory* for developing LLMs, CoT, and RAG models.

References

Indigenous knowledge systems and environmental management doc file, 2021.

Amer Ahmed and Iryna Lenchuk. The interaction between morphosyntactic features and the performance of machine translation tools: The case of google translate, systran, and microsoft bing in english-arabic translation. *Theory and Practice in Language Studies*, page n. pag, 2024.

A. H. Aldawsar. Evaluating translation tools: Google translate, bing translator, and bing ai on arabic colloquialisms. *Arab World English Journal*, 2024.

A. L. Alter, D. M. Oppenheimer, N. Epley, and R. N. Eyre. Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General*, 136(4):569–576, 2007. doi: 10.1037/0096-3445.136.4.569.

E. Alzain, K. A. Nagi, and F. Algobaei. The quality of google translate and chatgpt english to arabic translation: The case of scientific text translation. *Forum for Linguistic Studies*, 2024.

O. Bielykh. Stumbling blocks in the translation of technical documentation. *Forum for Education Studies*, 2024.

D. Brown and G. Hawe. Exploring trade-offs in explainable ai. *ACM SIGAda Ada Letters*, 43:36–42, 2024.

R. Esfandiari and Z. Khatibi. A cross-cultural, cross-contextual study of interactional metadiscourse in academic research articles: An interpersonal approach. *Southern African Linguistics and Applied Language Studies*, 40:227–242, 2022.

Y. Fung, R. Zhao, J. Doo, C. Sun, and H. Ji. Massively multi-cultural knowledge acquisition & lm benchmarking. *ArXiv*, 2024. URL `https://arxiv.org/abs/2402.09369`.

G. Gao, H. Wang, D. Cosley, and S. R. Fussell. Same translation but different experience: the effects of highlighting on machine-translated conversations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 449–458, 2013.

Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, et al. Retrieval-augmented generation for large language models: A survey. *ArXiv*, 2023. URL `https://arxiv.org/abs/2312.10997`.

B. Ihnaini, W. Sun, Y. Cai, Z. Xu, and R. Sangi. Sentiment analysis of song dynasty classical poetry using fine-tuned large language models: A study with llms. In *2024 7th International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pages 590–597, 2024. doi: 10.1109/ICAIBD62003.2024.10604440.

I. Lei and Z. Deng. Hint of thought prompting: An explainable and zero-shot approach to reasoning tasks with llms. *ArXiv*, 2023. URL `https://arxiv.org/abs/2305.11461`.

B. Liu, C. Lyu, Z. Min, Z. Wang, J. Su, and L. Wang. Retrieval-augmented multi-modal chain-of-thoughts reasoning for large language models. *ArXiv*, 2023. URL `https://arxiv.org/abs/2312.01714`.

P. Lu, B. Peng, H. Cheng, M. Galley, K. Chang, Y. N. Wu, S. Zhu, and J. Gao. Chameleon: Plug-and-play compositional reasoning with large language models. *ArXiv*, 2023. URL `https://arxiv.org/abs/2304.09842`.

Y. Lyu, Z. Li, S. Niu, et al. Crud-rag: A comprehensive chinese benchmark for retrieval-augmented generation of large language models. *ArXiv*, 2024. URL `https://arxiv.org/abs/2401.17043`.

J. Ma, Z. Gao, Q. Chai, W. Sun, P. Wang, H. Pei, et al. Debate on graph: A flexible and reliable reasoning framework for large language models. *ArXiv*, 2024. URL `https://arxiv.org/abs/2409.03155`.

Aleksey Panasyuk et al. Uncovering deep-rooted cultural differences (uncover). In *Defense + Commercial Sensing*, 2024.

P. Pirozelli, M. M. José, P. D. Filho, A. A. Brandão, and F. G. Cozman. Assessing logical reasoning capabilities of encoder-only transformer models. *ArXiv*, 2023. URL /abs/2312.11720.

I. Radeva, I. P. Popchev, L. Doukovska, and M. Dimitrova. Web application for retrieval-augmented generation: Implementation and testing. *Electronics*, 2024.

L. Ranaldi, G. Pucci, F. Ranaldi, E. S. Ruzzetti, and F. M. Zanzotto. Empowering multi-step reasoning across languages via tree-of-thoughts. *ArXiv*, 2023. URL https://arxiv.org/abs/2311.08097.

N. Shankar. The situation has provided a cue; this cue has given the expert access to information stored in memory, and the information provides the answer. intuition is nothing more and nothing less than recognition. In *Intuition in Judgment and Decision Making*, 2013.

S. Singh, D. S. Chaubey, R. Raj, V. Kumar, M. Paliwal, and S. Mahlawat. Social media communication, consumer attitude and purchase intention in lifestyle category products: a pls-sem modeling. *Marketing Intelligence & Planning*, 2024.

J. Son and B. Kim. Translation performance from the user's perspective of large language models and neural machine translation systems. *Inf.*, 14:574, 2023.

D. Tao, F. Huang, C. Wang, Y. Huang, and M. Jiang. Clue-guided path exploration: Optimizing knowledge graph retrieval with large language models to address the information black box challenge. *ArXiv*, 2024. URL https://arxiv.org/abs/2401.13444.

K. Y. Thakkar and N. Jagdishbhai. Exploring the capabilities and limitations of gpt and chat gpt in natural language processing. *Journal of Management Research and Analysis*, 2023.

Kirti Vashee. The path to llm-based machine translation. *Imminent*, 2024. URL https://imminent.translated.com/llm-based-machine-translation.

G. Wang, Y. Li, Y. Liu, G. Deng, T. Li, G. Xu, et al. Metmap: Metamorphic testing for detecting false vector matching problems in llm augmented generation. *ArXiv*, 2024. URL https://arxiv.org/abs/2402.14480.

S. Yu, M. Cheng, J. Yang, and J. Ouyang. A knowledge-centric benchmarking framework and empirical study for retrieval-augmented generation. *ArXiv*, 2024. URL https://arxiv.org/abs/2409.13694.

Z. Zhang, S. Han, H. Yao, G. Niu, and M. Sugiyama. Generating chain-of-thoughts with a pairwise-comparison approach to searching for the most promising intermediate thought. *ArXiv*, 2024. URL https://arxiv.org/abs/2402.06918.

M. Zhou and Z. Sun. Source work in china-based and western ssci journal articles: Preferences of reporting structures of english academic articles published in china and in english-speaking countries. *Chinese Journal of Applied Linguistics*, 45:596–611, 2022.