Research Proposal

Applicant's name:
Jinhu Qi

Proposed Topic/Title of Research:
Utilizing Large Language Models and Retrieval-Augmented Generation Techniques to Overcome Communication Barriers Across Different Languages and Cultural Backgrounds

Background:
Effective communication across different languages and cultural backgrounds remains a significant challenge in today's interconnected world. [1]Current translation tools often fail to accurately convey the original thoughts, leading to misunderstandings and inefficiencies. My research aims to address this gap by leveraging Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) techniques, combined with [2]knowledge graphs, to enhance social communication efficiency through a deeper understanding and transformation of thoughts.

The theoretical framework of this research is grounded in Natural Language Processing (NLP) and computational linguistics. By integrating language models with knowledge graphs, [3]the study aims to perform reasoning in a seemingly "white box" environment, thereby mitigating the hallucination problem of LLMs. The central research question is: How can LLMs, enhanced by RAG and knowledge graphs, accurately understand and convey the original thoughts behind human speech across different languages?

Methodology:
The research methodology involves several key steps:

1. Subject Selection and Data Collection:

    - Collect multilingual speech data from diverse cultural backgrounds.
    - Use NLP models such as BERT to process and analyze the speech data, focusing on understanding the semantic and syntactic structures.[6]

2. Data Mapping and Vectorization:

    - Map the processed information to a vector database using advanced vectorization techniques.[5]
    - Utilize knowledge graphs to enrich the contextual understanding and facilitate more accurate mapping.[4]

3. Retrieval and Query Techniques:

    - Implement retrieval and query techniques to fetch relevant data from the vector database.
    - Return the results to the LLM for data integration and generation.

4. Data Integration and Generation:

    - Integrate the retrieved data using LLMs to generate sentences in the target language that accurately convey the original meaning.[7]
    - Conduct iterative testing and refinement to ensure the accuracy and reliability of the

generated translations.

5. Statistical Analysis:

- Apply statistical methods to evaluate the performance of the proposed approach.
- Compare the results with existing translation tools to measure improvements in accuracy and understanding.

Outcomes and Value:
The expected outcomes of this research include:

- Development of a novel approach that combines LLMs, RAG techniques, and knowledge graphs to enhance multilingual communication.
- Creation of a more accurate and reliable translation system that captures the original thoughts behind human speech.
- Significant improvements in social communication efficiency and understanding across different languages and cultural backgrounds.

The anticipated impacts of the results are profound, with potential applications in social media, international business, and global collaboration. By reducing communication barriers, this research aims to foster social harmony and progress, ultimately contributing to a more connected and understanding world.

References:

[1]Li, Z. (2023). Research and Practical Exploration of English Translation from the Perspective of Cross-Cultural Communication. International Journal of Science and Engineering Applications.

[2]Bianchini, F., Calamo, M., Luzi, F.D., Macrì, M., & Mecella, M. (2024). Enhancing Complex Linguistic Tasks Resolution Through Fine-Tuning LLMs, RAG and Knowledge Graphs (Short Paper). CAiSE Workshops.

[3]Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2023). Unifying Large Language Models and Knowledge Graphs: A Roadmap. ArXiv. https://doi.org/10.1109/TKDE.2024.3352100

[4]Sanmartin, D. (2024). KG-RAG: Bridging the Gap Between Knowledge and Creativity. ArXiv. /abs/2405.12035

[5] Staš, J., Hládek, D., & Koctúr, T. (2023). Slovak Question Answering Dataset Based on the Machine Translation of the Squad V2.0. Journal of Linguistics/Jazykovedný casopis, 74, 381 - 390.

[6] Singh, S., & Mahmood, A. (2021). The NLP Cookbook: Modern Recipes for Transformer based Deep Learning Architectures. ArXiv. https://doi.org/10.1109/ACCESS.2021.3077350

[7] Wang, G., Li, Y., Liu, Y., Deng, G., Li, T., Xu, G., Liu, Y., Wang, H., & Wang, K. (2024). MeTMaP: Metamorphic Testing for Detecting False Vector Matching Problems in LLM Augmented Generation. ArXiv. /abs/2402.14480