Streamlining Tender Submissions: An AI Approach

Twan Lieuw A Soe[1], Sietske Tacoma[2], Steven Haveman[3], Devran Alper[4] and Stephan Nell[4]

[1] HU University of Applied Sciences Utrecht, Institute of Design & Engineering, Padualaan 99, 3584CH Utrecht, The Netherlands
[2] HU University of Applied Sciences Utrecht, Research Group Artificial Intelligence, Padualaan 99, 3584CH Utrecht, The Netherlands
[3] HU University of Applied Sciences Utrecht, Research Group Organizations in Digital Transitions & Institute of Design & Engineering, Padualaan 99, 3584CH Utrecht, The Netherlands
[4]Independent Researcher

Abstract. The research investigates integrating Azure OpenAI with multi-agent systems using the Retrieval-Augmented Generation (RAG) framework to propose a framework to optimize the RFX process. The study aims to adress the efficiency, precision, and relevancy in drafting proposals by leveraging specialized agents to handle different sections, thereby adressing drafting time and increasing response accuracy.

Keywords: Retrieval-Augmented Generation (RAG), Request for Proposal/Information (RFP, RFI, RFX), Generative Artificial Intelligence (GenAI), Large Language Models (LLMs), Multi-Agent Systems,

1 Introduction

Generative Artificial Intelligence (GenAI), through the use of Large Language Models (LLMs) offers a transformative approach to optimizing project acquisition processes, specifically in the creation of Requests for Proposal (RFP) and Requests for Information (RFI), collectively referred to as the RFX process. Traditional methods of responding to RFX documents, though structured, the RFX process is manual, time-consuming process, and prone to inefficiencies such as writer's block, redundant information gathering, and prolonged drafting times [1, 2, 12]. This research introduces an innovative system that integrates Azure OpenAI with a multi-agent system within the Retrieval-Augmented Generation (RAG) framework [13]. The multi-agent system allows for specialization, where each agent is tasked with handling different aspects of the RFX process. These agents operate independently but collaborate to generate tailored and precise responses for specific sections of the proposal, ultimately enhancing the efficiency, precision, and relevance of RFX documents [3, 4].

2 Methodology

This research investigates the potential of combining LLMs with the RAG framework to improve the efficiency of the RFX drafting process in consultancy firms. Specifically, showcasing how the use of GenAI can reduce drafting duration, increase response relevancy, and ultimately enhance productivity while reducing costs [1]. Additionally, this research aims to provide a system that can be used by organizations looking to optimize their proposal generation workflows, facilitating access to relevant data and minimizing errors or inconsistencies [5, 6, 7, 8, 9, 10, 11]. The research methodology follows three key stages.

1. **Selection of Methodology:** Fine-tuning LLMs, training domain-specific models, and integrating RAG were considered as possible solutions. RAG merges LLMs' generation capabilities with real-time retrieval of contextual data from an indexed document base, allowing for responses that are not only accurate but also contextually relevant to the client's specific needs [13, 14].
2. **Building a Proof-of-Concept (PoC):** The PoC system utilises Microsoft Azure services, leveraging the RAG framework to combine Azure OpenAI with specialized agents that perform distinct roles within the RFX process. For example, one agent may focus on compliance by ensuring that legal and regulatory requirements are met, while another agent handles keyword optimization, making sure that client-specific and industry-standard terminology are incorporated [3].
3. **Testing and User Feedback:** The PoC system was tested with five consultants who are directly involved in RFX drafting. Through semi-structured interviews and presentations, their feedback was gathered on key areas such as the system's usability, the quality of the generated content, and its ability to mitigate common challenges like writer's block. The consultants provided insights into the system's potential to improve their workflow, with a particular focus on reducing time spent retrieving and compiling information [1].

3 Discussion and conclusion

The PoC demonstrated promising results in the RFX process, with the use of multi-agents proving to be a key factor in the system's success. Specialized agents handled different sections of the proposal simultaneously, and feedback from the consultants confirmed that this approach could help alleviate common inefficiencies, such as prolonged drafting times and difficulty in accessing relevant information. The multi-agent framework showcased scalability, allowing the system to manage more complex proposals and larger volumes of RFX documents without sacrificing accuracy or speed. However, limitations like token truncation were identified, and consultants recommended increasing token limits and improving responses to detailed client inquiries. Overall, the multi-agent system showed potential for improving efficiency and relevance in the RFX process, enabling teams to focus on refining content. Future research could explore advanced models like GPT-4 and assess the scalability of the system for larger enterprise applications.

References

1. Candelon, F., Krayer, L., Rajendran, S., David, M. Z.: How People Can Create—and Destroy—Value with Generative AI. Retrieved from https://www.bcg.com/publications/2023/how-people-create-and-destroy-value-with-gen-ai, last accessed 2024/10/24
2. Dorn, W. R., Payne, J., Dorn, W. R., Ulrich, J.: The RFx Procs. In: Managing Indirect Spend Enhancing Profitability Through Strategic Sourcing, 45-62 (2011).
3. Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Wiest, O., Chawla, N. V.: Large Language Model based Multi-Agents: A Survey of Progress and Challenges. ArXiv, 1-10 (2024).
4. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kiela, D.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. ArXiv, 1-3 (2021).
5. Microsoft: Retrieval Augmented Generation (RAG) in Azure AI Search. Retrieved from https://learn.microsoft.com/en-us/azure/search/retrieval-augmented-generation-overview (2023, 11 20), last accessed 2024/10/21
6. Microsoft: What are Memories? Retrieved from https://learn.microsoft.com/en-us/semantic-kernel/memories/ (2023, 05 23), last accessed 2024/10/20
7. Microsoft: Personas: Giving your Agent a Role. Retrieved from https://learn.microsoft.com/en-us/semantic-kernel/concepts/personas?pivots=programming-language-csharp (2024, 6 25), last accessed 2024/10/20
8. Microsoft: Understanding the Kernel. Retrieved from https://learn.microsoft.com/en-us/semantic-kernel/concepts/kernel?pivots=programming-language-csharp, last accessed 2024/10/20
9. Microsoft: What are Agents? Retrieved from https://learn.microsoft.com/en-us/semantic-kernel/concepts/agents?pivots=programming-language-csharp, last accessed 2024/10/24
   Microsoft: What is a Planner? Retrieved from https://learn.microsoft.com/en-us/semantic-kernel/concepts/planning?pivots=programming-language-csharp, last accessed 2024/08/30
10. Microsoft: What is Semantic Kernel? Retrieved from https://learn.microsoft.com/en-us/semantic-kernel/overview/?tabs=Csharp, last accessed 2024/08/30
11. Microsoft: What's Azure AI Search? Retrieved from https://learn.microsoft.com/en-us/azure/search/search-what-is-azure-search, last accessed 2024/08/30
12. Nistala, P., Rajbhoj, A., Kulkarni, V., Noronha, S., Joshi, A.: An industrial experience report on model-based, AI-enabled proposal development for an RFP/RFI. Science of Computer Programming, 1-3 (2024).
13. Ovadia, O., Brief, M., Mishaeli, M., Elisha, O.: Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs. ArXiv, 1-14 (2024).
14. Soudani, H., Kanoulas, E., Hasibi, F.: Fine Tuning vs. Retrieval Augmented. ArXiv, 1-2, 11 (2024).