NTNU
Kunnskap for en bedre verden

Department of Computer Science and Informatics

TDT39 - Empirical Studies in ICT

Research Plan

Title:
Semantic Routers: A Novel Approach to Enhancing Multi-Agent Systems

Author:
Adam Sioud

Word Count:
1194

27 Nov 2024

1 Purpose

Recent advancements in Large Language Models (LLMs) have transformed artificial intelligence by enabling agents capable of reasoning, decision-making, and task execution in complex environments [1, 2]. However, LLMs face practical limitations such as restricted context length, delayed knowledge updates, and inability to directly interact with external tools (e.g., calculators, databases), which impede their effectiveness in dynamic, multi-step workflows [3]. These constraints hinder the seamless operation of LLMs when handling complex and multi-step tasks.

A promising alternative is Multi-Agent Systems (MAS), which distribute tasks among specialized agents to enhance modularity, scalability, and collaborative intelligence [4, 5]. MAS enable distributed problem-solving and dynamic task allocation, mitigating LLMs' limitations and empowering agents to handle more complex processes. However, MAS face significant challenges that hinder their efficiency and scalability. Key issues include communication overload, coordination complexity, context management, and dynamic capability acquisition, which collectively reduce system performance and adaptability [4, 6]. Furthermore, MAS often lack mechanisms to effectively route external inputs to the most appropriate agents or tools, leading to inefficiencies and suboptimal task execution. Agents may also produce outputs that are unstructured or not immediately actionable, necessitating additional processing to make them useful for decision-making or further tasks.

Current MAS frameworks often rely on monolithic architectures, creating bottlenecks in scalability and task execution, particularly in domains like software engineering and AI-driven web navigation [4, 6]. For example, Guo et al. [6] highlight that existing MAS inadequately address effective communication and information exchange between agents, leading to suboptimal collaboration. Han et al. [4] emphasize difficulties in scaling MAS due to inefficient agent coordination. The Magnetic-One system [7], while demonstrating multi-agent collaboration, suffers from inefficient message routing and lacks adaptability in dynamic settings. These limitations underscore the need for more scalable and flexible MAS designs that can efficiently manage communication and coordination among agents.

To address these challenges, this research proposes a novel approach: **semantic routers**. Semantic routers act as intelligent decision-making layers within MAS, leveraging embedding-based routing to semantically interpret agent interactions and dynamically allocate messages based on contextual relevance. They steer external inputs to the most suitable agents or tools and refine agent outputs into more actionable or readable formats. Unlike traditional MAS communication protocols that rely on static and inflexible pathways, semantic routers adapt dynamically to contextual changes, reducing redundant communication and improving coordination efficiency [4, 6]. This adaptability enhances scalability, efficiency, and reliability in dynamic environments.

Recent studies have demonstrated the potential of semantic routing in enhancing AI system performance. Manias et al. [8] showed that semantic routers could improve the accuracy and efficiency of LLM-assisted intent-based network management in 5G core networks. Similarly, Janakiram [9] discussed how semantic routers enable agents to select appropriate language models for specific tasks, reducing dependency on large models and improving overall workflow efficiency. The work by ClearPeaks [10] demonstrates how semantic routers can interpret user inputs to select the most appropriate function or agent, improving the relevance and efficiency of responses in AI systems. Eskili [11] illustrates how semantic routing in chatbots prevents unwanted or irrelevant outputs, enhancing the user experience by providing more actionable and readable information. Inspired by Buckley's work on semantic routing using Azure AI Search [12], this research integrates embedding models and vector databases to optimize communication pathways in MAS.

Despite preliminary work on semantic routing in AI systems [9–12], there is a lack of research on applying semantic routers within MAS to address the specific challenges of communication overload, coordination complexity, input steering, and output refinement. By integrating semantic routers into MAS, this study seeks to enhance scalability, efficiency, reliability, and adaptability in multi-agent workflows, contributing to the fields of information systems, software engineering, and artificial intelligence.

1.1 Objectives

- Develop a framework that integrates semantic routers into MAS.
- Evaluate the effectiveness of semantic embedding techniques for dynamic message routing, input steering, and output refinement.
- Assess the impact of semantic routing on system performance, resource utilization, and scalability.

1.2 Research Questions

RQ1: How do semantic routers enhance the scalability, efficiency, reliability, and adaptability of multi-agent systems?

RQ1.1: Which semantic embedding techniques are most effective for dynamic message prioritization, input routing, and output refinement in MAS?

RQ1.2: How does embedding-based routing impact resource utilization, task execution efficiency, and output usability in MAS?

RQ1.3: What are the scalability limitations of semantic routing in large-scale, heterogeneous MAS environments?

By addressing these questions, this research will fill a gap in knowledge by providing a novel methodology for enhancing MAS communication and coordination through semantic routers.

2 Contributions

This research will deliver a novel methodology by integrating semantic routers into Multi-Agent Systems to enhance communication, coordination, input steering, and output refinement. By designing and implementing a semantic routing framework utilizing embedding-based techniques, the study addresses the identified gap in MAS communication efficiency. The novelty lies in creating a dynamic, context-aware communication mechanism that surpasses static protocols, improving scalability, efficiency, and reliability in MAS operations.

3 Research Method

A design and creation research strategy will be employed to develop the semantic routing framework, aligning with the objective of producing a new methodology. The framework will be implemented within an existing MAS platform. Controlled simulations will generate data, creating environments with diverse inputs and agent interactions. Quantitative analysis will measure performance metrics such as routing accuracy, message latency, and resource utilization. Statistical methods will assess improvements over traditional MAS protocols, ensuring that the research questions are effectively addressed. Potential threats to validity, like simulation biases, will be mitigated through careful experimental design and multiple scenario testing. Alternative methods, such as purely rule-based or heuristic-driven routing mechanisms, were considered but ruled out due to their lack of adaptability to dynamic environments and limited scalability.

4 Participants

The study involves virtual agents within the MAS platform, simulating real-world behaviors, and expert reviewers—faculty members or industry professionals—who will provide feedback on the

framework's design and usability. Non-researcher participants (expert reviewers) are essential to validate the practical relevance of the framework. They will be recruited via professional networks, with informed consent obtained. Ethical considerations are minimal since no personal or sensitive data is collected; feedback will be anonymized to protect privacy. As the researcher, I will design, implement, and evaluate the framework, maintaining objectivity to ensure the validity of the results.

5 Research Paradigm

The research adopts a positivist paradigm, focusing on objective measurement and hypothesis testing. By utilizing quantitative methods and statistical analysis within simulations, the study seeks to establish empirical evidence on the effectiveness of semantic routers in MAS. This paradigm is appropriate as it supports developing generalizable findings and understanding cause-and-effect relationships, aligning with the goal of enhancing MAS performance through a new methodology.

6 Dissemination

The results will be disseminated through a comprehensive Master's thesis report detailing the research process, findings, and implications. A functional prototype will demonstrate dynamic message routing within MAS, showcasing its effectiveness in controlled simulations. This research is ambitious for a half-year Master's thesis, with broader objectives like input steering and output refinement potentially serving as a foundation for future PhD work. Findings will be presented at academic conferences, published in journals, and shared on social media to engage a broader audience, particularly those interested in demos and practical applications in AI and MAS.

References

[1] T. B. Brown *et al.*, 'Language models are few-shot learners', *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020. DOI: 10.48550/arXiv.2005.14165.

[2] L. Wang *et al.*, 'A survey on large language model based autonomous agents', *Frontiers of Computer Science*, vol. 18, no. 1, pp. 1–35, 2024. DOI: 10.48550/arXiv.2308.11432.

[3] Y. Cheng *et al.*, 'Exploring large language model based intelligent agents: Definitions, methods, and prospects', *arXiv preprint*, vol. 2401, p. 03 428, 2024. DOI: 10.48550/arXiv.2401.03428.

[4] S. Han *et al.*, 'Llm multi-agent systems: Challenges and open problems', *arXiv preprint*, vol. 2402, p. 03 578, 2024. DOI: 10.48550/arXiv.2402.03578.

[5] N. Crawford *et al.*, 'Bmw agents – a framework for task automation through multi-agent collaboration', *arXiv preprint*, vol. 2406, p. 20 041, 2024. DOI: 10.48550/arXiv.2406.20041.

[6] T. Guo *et al.*, 'Large language model based multi-agents: A survey of progress and challenges', *arXiv preprint*, vol. 2402, p. 01 680, 2024. DOI: 10.48550/arXiv.2402.01680.

[7] A. Fourney *et al.*, *Magentic-one: A generalist multi-agent system for solving complex tasks*, Nov. 2024.

[8] D. M. Manias, A. Chouman and A. Shami, 'Semantic routing for enhanced performance of llm-assisted intent-based 5g core network management and orchestration', in *IEEE Globecom 2024 Proceedings*, 2024. DOI: 10.48550/arXiv.2404.15869.

[9] J. M. S. V. Janakiram, *Semantic router and its role in designing agentic workflows*, The New Stack, Accessed: 2024-11-27, Sep. 2024. [Online]. Available: https://thenewstack.io/semantic-router-and-its-role-in-designing-agentic-workflows/.

[10] ClearPeaks, *The semantic router: Ai's pathway to understanding user input*, ClearPeaks Blog, Accessed: 2024-11-27, Nov. 2024. [Online]. Available: https://www.clearpeaks.com/the-semantic-router-ais-pathway-to-understanding-user-input/.

[11] B. T. Eskili, *Smarter chatbots: How semantic routing prevents the unwanted*, Marvelous MLOps, Accessed: 2024-11-27, Mar. 2024. [Online]. Available: https://medium.com/marvelous-mlops/smarter-chatbots-how-semantic-routing-prevents-the-unwanted-be31f34a7df6.

[12] C. Buckley, *Semantic router using azure ai search*, ISE Developer Blog, Accessed: 2024-11-27, Aug. 2024. [Online]. Available: https://devblogs.microsoft.com/ise/semantic-routing-using-azure-ai-search/.