BUILD A MULTI-AGENT APPROACH FOR A LARGE LANGUAGE MODEL (LLM) TO IMPROVE RESPONSES AND BY USING RETRIEVAL-AUGMENTED GENERATION (RAG) TO CONSIDER OWN DATA IN QUERIES

Bachelor Thesis by Anna Hansl

Supervisor: Dipl.-Ing. Dr.techn. Marian Lux

MOTIVATION

- Many tools on the market, chatbots are everywhere
- But no local, open-source, multi-agent tool that can easily be extended
- Use Case: Blog Post Generation
  - But: Can easily be changed and extended

BACKGROUND INFORMATION

LARGE LANGUAGE MODELS & TOKENIZATION

- LLM: AI Systems that are able to understand and use human language
- How? Always predicting the next word in a sentence based on context
  - Uses the Transformer architecture, which understands the relationships between words[1]
  - Many providers: OpenAI, Google, Meta, …
  - Used here: *llama3.1:8b-instruct-q8_0* – can be downloaded and run locally [2]
- Based on Tokenization: Data sets transformed into small sequences and are embedded[3]

RETRIEVAL-AUGMENTED GENERATION & EMBEDDING

- RAG: Two-phase AI Framework used for retrieving facts from other sources, like a website or a database[4]
  - Used to reduce LLM hallucinations and improve answer quality by providing a knowledge source

- Embedding: Vectorization of tokenized objects like text, picture and audios to continuous vector space[5]
  - Can then be used for vector calculations
  - Embedding model used in this implementation: *mxbai-embed-large*

MULTI-AGENT SYSTEMS & CREWAI

- Multi-Agent Systems: Multiple LLM Agents working together to perform a task[6]
  - Each agent powerful and can act on its own
  - Communication and task distribution most important aspects of framework

- CrewAI: Framework for multi-agent systems[7]
  - Enables users to assemble their „dream team“
  - Based on a LLM of choice

PROMPT ENGINEERING & OTHER SOFTWARE AND FRAMEWORKS

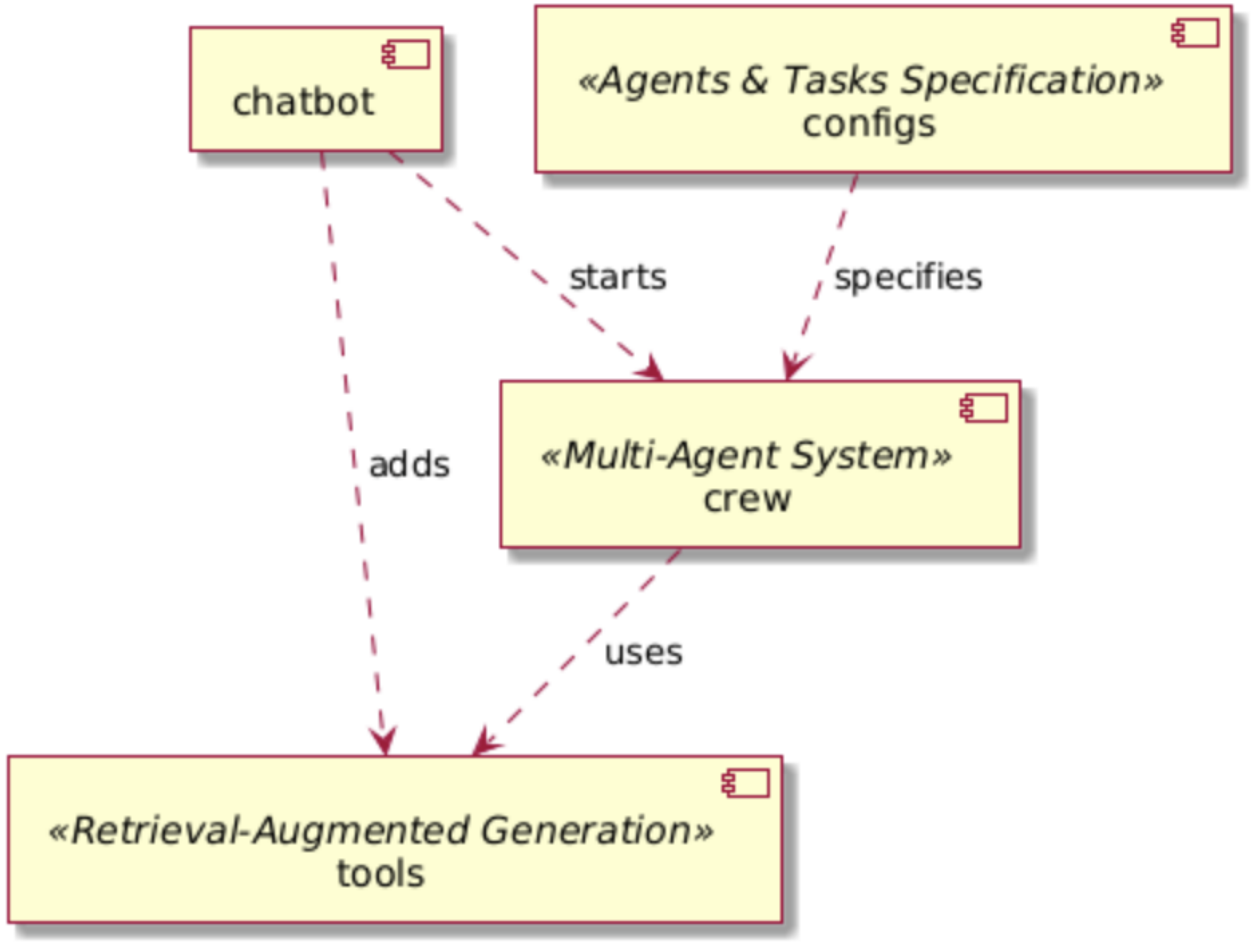
- Prompt Engineering is the making of queries to improve LLM responses by using thoughtfully crafted natural language
  - Many techniques around nowadays, eg. Chain-of-thought[8]

Other Software and Frameworks:

- LangGraph by LangChain[9]
- OpenAI Swarm, Autogen, Magentic-One, …
- Many LLMs and embedding models, Ollama is just one of them
- And there are chatbots a dime a dozen

DEMO

ARCHITECTURE



IMPLEMENTATION

RESULTS & EVALUATION

RESULTS & EVALUATION

- What were the primary objectives of the project and were they achieved?
  - Create a local, open-source, multi-agent tool that can easily be extended - Yes
- What methods and technologies were used and how effective were they?
  - CrewAI, Ollama, Telegram Chatbot - Partially effective (Except for Telegram Chatbot)
- What are the limitations of the implementation?
  - Very slow, and telegram chatbot can be hard to adapt
- Based on the project's outcomes, what future research do you recommend?
  - Try to improve speed by using a different framework and LLM
  - Try to make a user interface (eg. website) that adapts the whole tool for a specified use case for the user

LITERATURE

- [1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, u. & Polosukhin, I. (2017). Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett (eds.), *Advances in Neural Information Processing Systems*, : Curran Associates, Inc..
- [2] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. ArXiv, abs/2302.13971.
- [3] Webster, Jonathan & Kit, Chunyu. (1992). Tokenization as the initial phase in NLP. 1106-1110. 10.3115/992424.992434.
- [4] Gupta, Shailja & Ranjan, Rajesh & Singh, Surya. (2024). A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions. 10.48550/arXiv.2410.12837.
- [5] Peters, Matthew & Neumann, Mark & Iyyer, Mohit & Gardner, Matt & Clark, Christopher & Lee, Kenton & Zettlemoyer, Luke. (2018). Deep contextualized word representations. 10.48550/arXiv.1802.05365.
- [6] Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., & Mordatch, I. (2019). Emergent Tool Use From Multi-Agent Autocurricula. ArXiv, abs/1909.07528.
- [7] Introduction. CrewAI. (n.d.). https://docs.crewai.com/introduction
- [8] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22). Curran Associates Inc., Red Hook, NY, USA, Article 1800, 24824–24837.
- [9] Gupta, M. (2024, November 20). Magentic-one, autogen, langgraph, Crewai, or openai swarm: Which Multi-AI Agent Framework is best?. Medium. https://medium.com/data-science-in-your-pocket/magentic-one-autogen-langgraph-crewai-or-openai-swarm-which-multi-ai-agent-framework-is-best-6629d8bd9509

THANK YOU FOR YOUR ATTENTION!