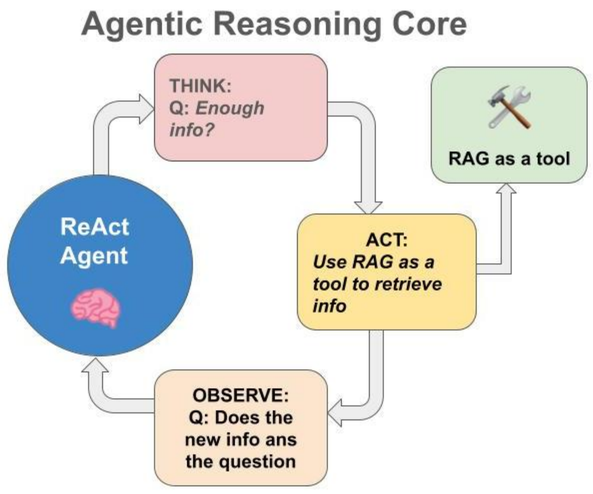
# An Open-Source Agentic Reasoning Framework for German Nuclear Documents

Rohil Rao, Prof. Dr. Thomas Kopinski

Fachhochschule Südwestfalen
University of Applied Sciences

## Abstract

We present an agentic framework that performs iterative **reasoning over German-language documents** in the nuclear decommissioning domain. The system orchestrates specialized language models for answering complex questions that require multi-step reasoning. To assess performance, we created a dataset: the **German-Nuclear-Decom-QA**, a domain-specific evaluation set inspired by GAIA [1], comprising highly objective questions with varying difficulty. In our evaluation, the workflow consistently classifies user intent, retrieves relevant, document-grounded evidence, and generates reasoned, accurate answers. These results indicate that agentic workflows can support reliable question answering in safety-critical settings without overreliance on a single model.




## Dataset

To robustly evaluate the system, we created the **German-Nuclear-Decom-QA** dataset of highly objective QA pairs derived from German nuclear decommissioning documents. The dataset is structured to include three levels of complexity with objective QA pairs, inspired by the GAIA benchmark [1]:

- **SIMPLE**: **Answerable by a single, direct fact from one document.**
- **MEDIUM**: **Require synthesis of information or simple calculations from a single document.**
- **HARD**: **Questions requiring cross-document synthesis or multi-step reasoning**.

| QA PAIR TYPES | QUESTION | ANSWER |
|---|---|---|
| **FACTUAL - TEXT** | Welcher Reaktortyp ist das Kernkraftwerk Grafenrheinfeld (KKG)? | Druckwasserreaktor |
| **NUMERICAL** | Wie viele Jahre liegen zwischen der ersten nuklearen Kettenreaktion im Block B des Gundremmingen und dem Beginn des Betriebs von KKG-BELLA? | 22 |
| **YES-NO** | Liegt der nächste Ort zum KRB II-Standort weiter als 5 km entfernt? | Nein (Mehrere Dörfer befinden sich in 1–4 km Entfernung.) |

## Approach

Our system is a multi-agent workflow. The process is broken down into four key stages:

**Intent Classification:** An Intent Agent first classifies the user query as either "corpus-relevant" or "general." This initial step prevents irrelevant queries from consuming computational resources and directs the workflow down the most appropriate path.

**Document Retrieval:** For relevant queries, a Retriever Agent performs a hybrid RAG search (vector and BM25) on the pre-processed document corpus. It retrieves and scores the top-k most relevant document chunks.

**Routing:** A Router Agent determines if the retrieved documents are sufficiently relevant to answer the query, using a pre-defined relevance threshold. If the documents meet the threshold, the query proceeds to the Reasoning Agent, it is routed to a general-purpose Large Language Model (LLM) to provide a response.

**Reasoning Core:** At the core is our **Reasoning Agent** (inspired from ReAct agents [4]) which utilizes the existing context to reason over the context and can **dynamically and autonomously** fetch additional information using the RAG retriever as a tool.

**Answer Generation:** The final stage involves two agents. A **Summarizer Agent** and for evaluation purposes, a **Final Answer Agent** then distills this summary into a succinct, single-word or number response.

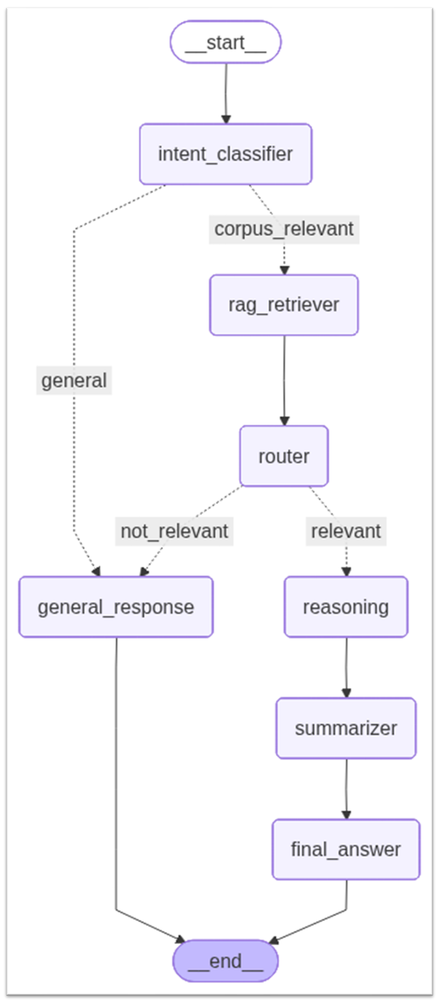


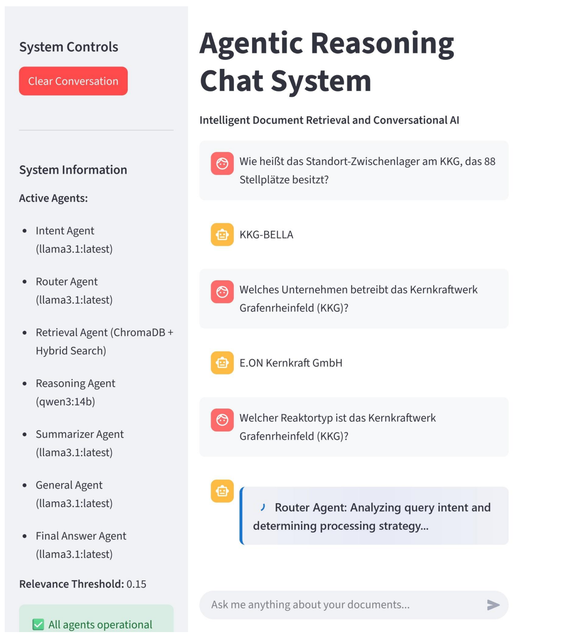

Fig: Workflow design of our agentic reasoning system for QA desinged using LangGraph

## Results

Evaluation performed on our **German-Nuclear-Decom-QA** dataset, shows that the framework successfully processes queries and retrieves information from the document corpus. Apart from evaluating the framework on the dataset, we also tested it via the UI shown on the right.

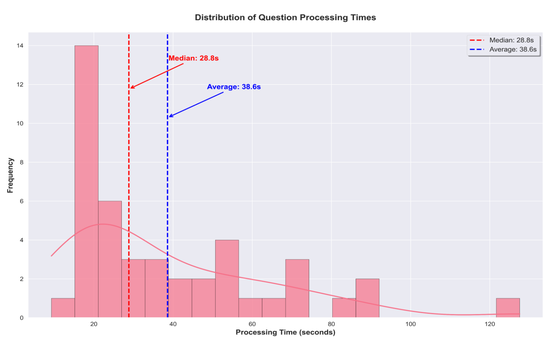
Adaptive Reasoning: The system's logic scales with complexity; average tool calls increase from 0 for simple to 1.80 for hard questions, with answer latency times rising from 25.14s to 114.71s.

The framework achieves an average 77.3% for semantic match accuracy, 66.70% for near-match accuracy, which drops to 46.7% for strict exact match evaluations, highlighting the challenge of generating precise responses and evaluation.




| Model | Questions | Exact Match Accuracy | Near Match Accuracy |
|---|---|---|---|
| GPT-4O | 45 | 76.33% | 82.22% |
| GEMINI 2.5 PRO | 45 | 81.21% | 91.11% |
| Ours | 45 | 46.7% | 66.7% |


| Difficulty Level | Number of Questions | Average Tool Calls per Question | Average Time to Answer (seconds) | Average Time per Tool Call (seconds) |
|---|---|---|---|---|
| Simple | 15 | 0.00 | 23.2 | 0.00 |
| Medium | 15 | 0.87 | 40.4 | 5.82 |
| Hard | 15 | 1.80 | 98.6 | 7.91 |




## Analysis & Conclusion

- This work demonstrates that an open-source agentic framework is a powerful approach for reasoning on domain-specific German language documents.
- Our evaluation shows that the core reasoning logic is sound; the system correctly identifies complex queries that require multi-step, autonomous tool-assisted information retrieval, as seen by the scaling of tool calls with question difficulty.
- Our framework's value is its modular, extendable and completely **open source** design, which enables targeted improvements - a capability unavailable in SOTA proprietary systems.
- **Future work** will focus on enhancing the multi-step reasoning capabilities of the framework:
  - **Improve Reasoning Quality:** We plan to integrate a supervised fine-tuned model into the workflow to evaluate if it improves performance.
  - **Enhance Retrieval** by adding question reformulator agents to improve the quality of inputs to the RAG retriever tool.
  - **Expand Evaluation:** We will continue to extend the German-Nuclear-Decom-QA dataset to cover more complex, multi-hop scenarios.

Check out the code and the detailed dataset and results here!

## References

1. Mialon, G., Fourrier, C., Swift, C., Wolf, T., LeCun, Y., & Scialom, T. (2023). GAIA: A benchmark for general AI assistants. arXiv. https://arxiv.org/abs/2311.12983
2. Anthropic. (2024, December 19). Building effective agents. Engineering at Anthropic. Retrieved from https://www.anthropic.com/engineering/building-effective-agents
3. LangChain. (2025, July 16). Open Deep Research [Blog post]. LangChain Blog. Retrieved from https://blog.langchain.com/open-deep-research/
4. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2022). ReAct: Synergizing Reasoning and Acting in Language Models. arXiv. Retrieved from https://arxiv.org/abs/2210.03629