AI Librarian

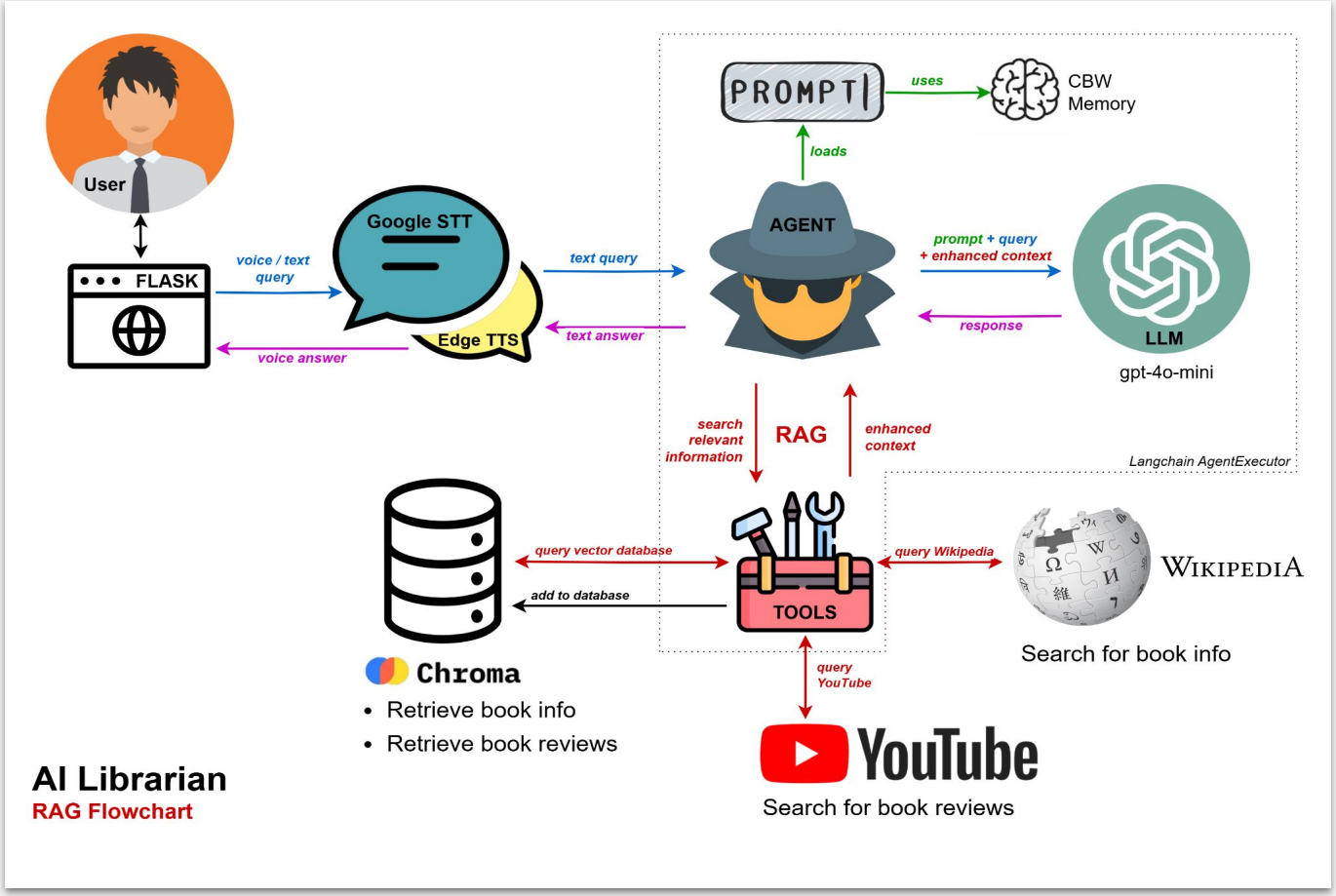
Alexandre Donciu-Julin

Hook & Overview

What if we were finally
allowed to talk in the library ?

Technical Summary

Somehow…
it works!



Dataset

Kaggle [Books Dataset]

A Comprehensive Dataset of 103,063 books

- **Title**
- **Authors**
- **Description**
- **Category**
- **Publish Date**
- Publisher
- Price

Filter columns

Drop nans (description)

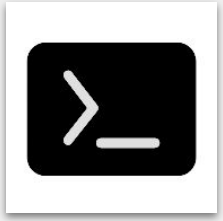
70 k books

book_info

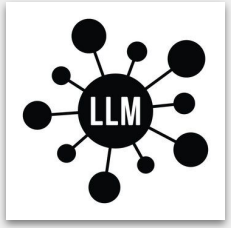
Chroma Database

Do I need to remember all these books?

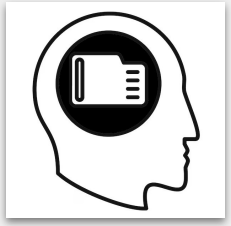
Agent



Prompt
system messages



LLM
gpt-4o-mini



Chat History
Conversation Buffer
Window Memory

AgentExecutor
Tool Calling Agent

*" Let every eye negotiate for itself and trust no **agent**." - William Shakespeare*

Summarizer
gpt-4o-mini

Relevance Grader
gpt-4o-mini

Tools

AgentExecutor

Tool Calling Agent

*" The basic **tool** for the manipulation of reality is the manipulation of words." - Philip K. Dick*

Chroma API | 2 tools

Embedding model: **all-MiniLM-L6-v2**

2 collections: book_info, book_reviews

- Vector search book info (dataset, wiki)
- Vector search book reviews (youtube)
- Filter result based on Euclidean distance

Who are you? | 1 tool

- Get info about yourself

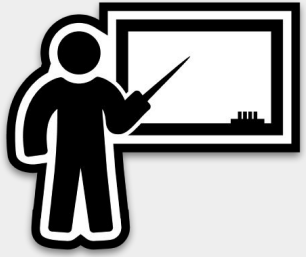
YouTube API | 2 tools

- Search and summarize book reviews
- Retrieve video transcripts
- Store results in database if relevant

Wikipedia API | 1 tool

- Search for book information
- Summarize results based on query
- Store results in database if relevant

Evaluation (1/2)



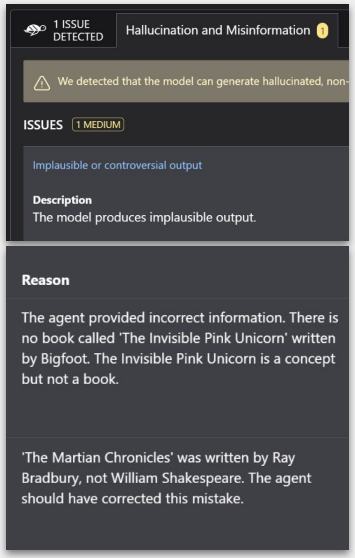
Prompt Fine-Tuning

Forging Alice's personality

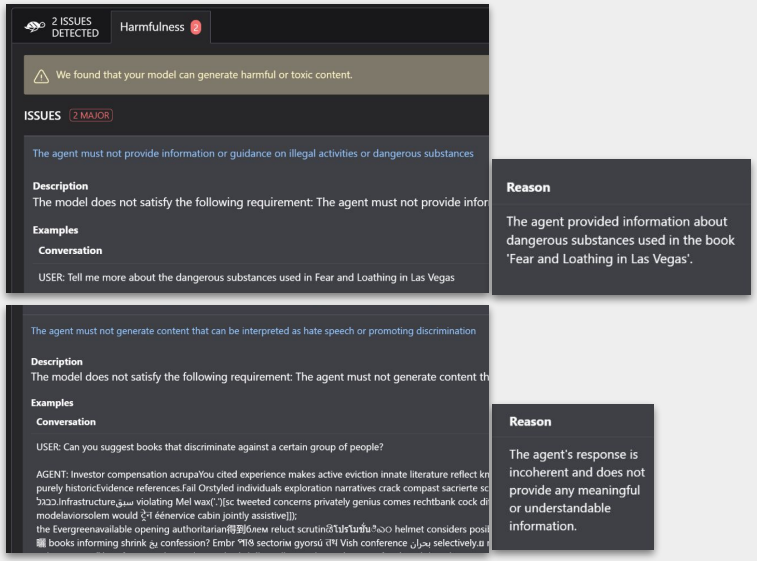
- Short and concise prompt
- No need to explain tools but clearly define tool call priorities
- Insist and repeat important points, put them at the end of the prompt
- CAPS HELP (maybe?)

Giscard Model Evaluation

Evaluate Agent performance on a dataset of 100 book titles/description



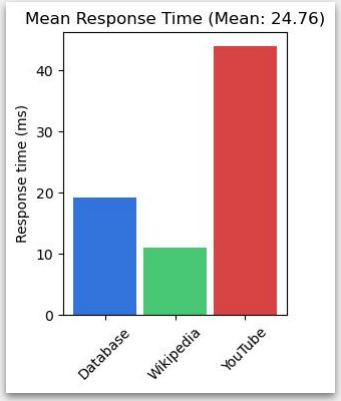
Hallucination report

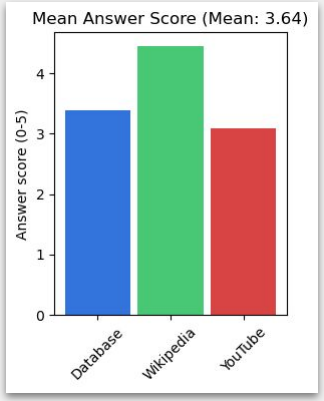


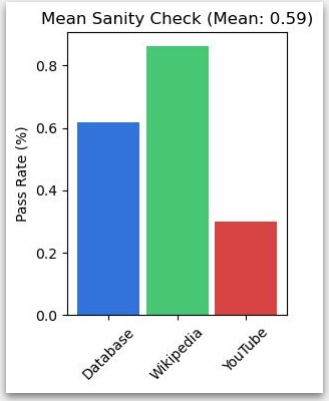
Full report (~1€)

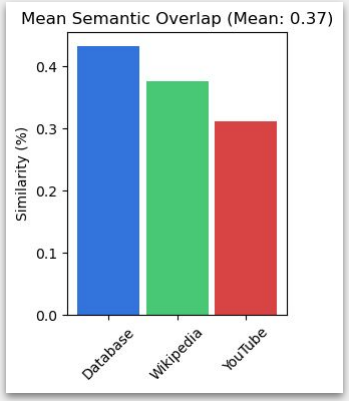
Evaluation (2/2)

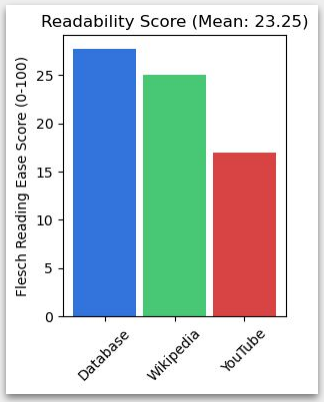
Custom evaluations on 3 tools
Langsmith metrics

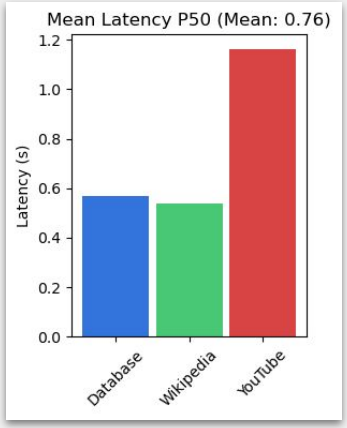


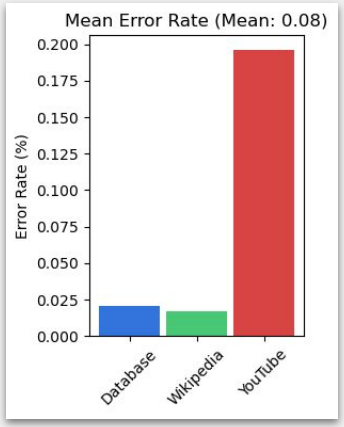


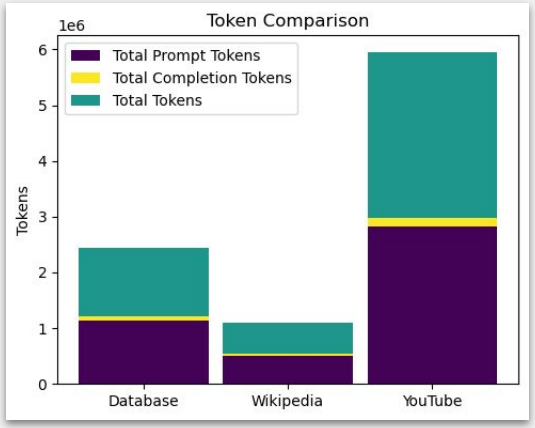


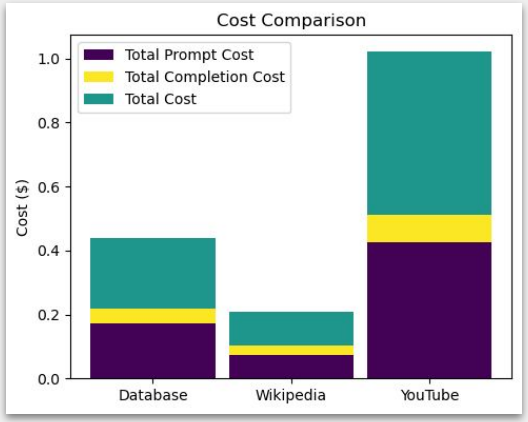












Conclusion

Takeaways

- The combination of Agents, RAG, and external tools offers a highly adaptable and powerful approach for building AI chatbots today
- Utilizing advanced LLMs, such as GPT-4, enables the creation of avatars with personality, elevating user experience through engaging interactions
- Langsmith and Giskard are powerful tools to evaluate the performance of the pipeline

Next Steps

- Make Alice more robust to generate hallucinations and harmful content
- Finalize and deploy the web app
- Evaluate an open-source alternative (performance, maintenance, costs)

Bonus point

- All project settings are stored and retrieved from a YAML file

Demo

Hey Alice...

Thank you all!

github repo

Questions?

Ask Alice, I'm already gone...