RAG agent for intuitive work with SQL databases

Marek Tenora
2025 Supervisor: Ing. Ondřej Klíma Ph.D.

System architecture

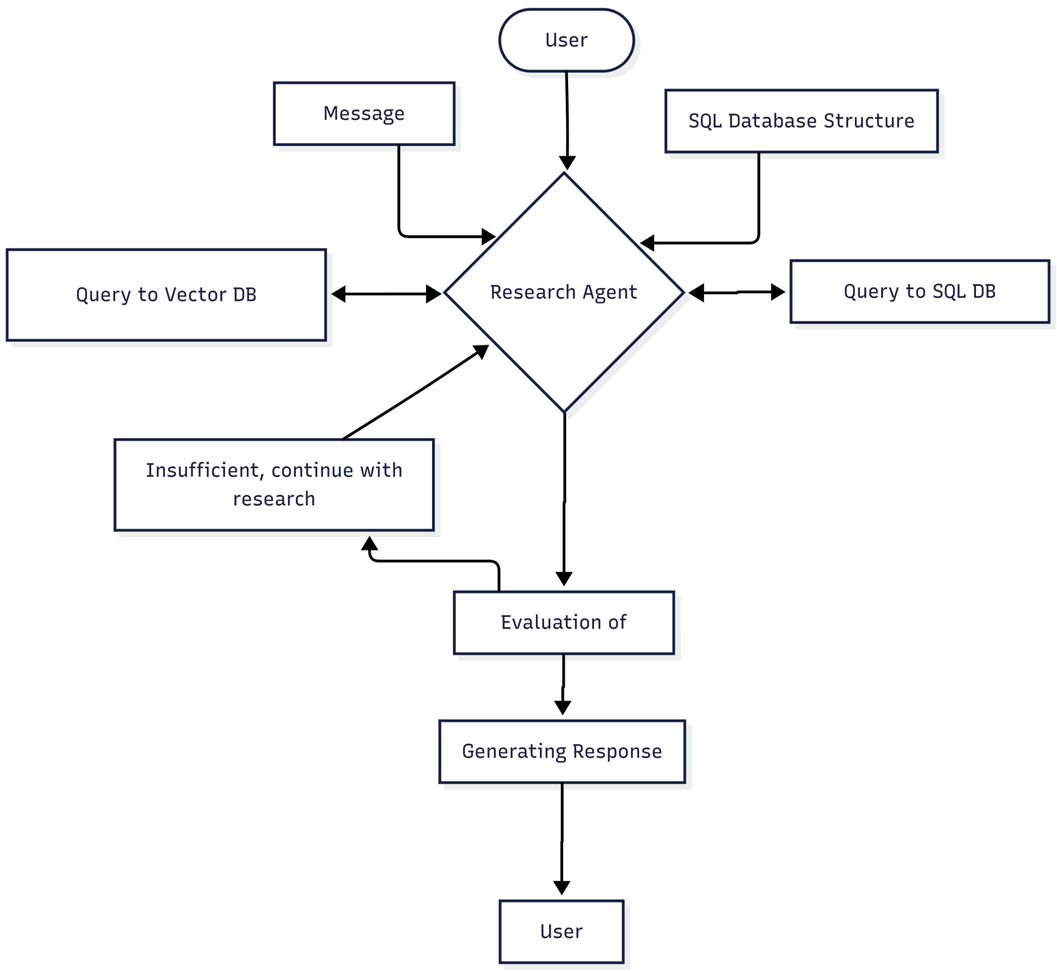

Figure 1

Information

The agent finds information by searching SQL and vector databases.

- The user is provided with the results of all the tools and reasoning.

Figure 4 - Screenshot of the results of the tools used

- The application allows you to set:
  - language model
  - embedding model
  - the target database on which the agent will work
- Technologies used:
  - The agent is built using the langchain and langgraph libraries
  - Backend: FastAPI
  - Frontend: Vue.js

Database

- The system uses 2 databases:
- SQL database - connected by the user, the main source of information
  - Vector database - created automatically by embedding the entire SQL database, used for semantic search
  - Benefits of semantic search:
- complex information that is difficult to search using SQL
  - multi-lingual search

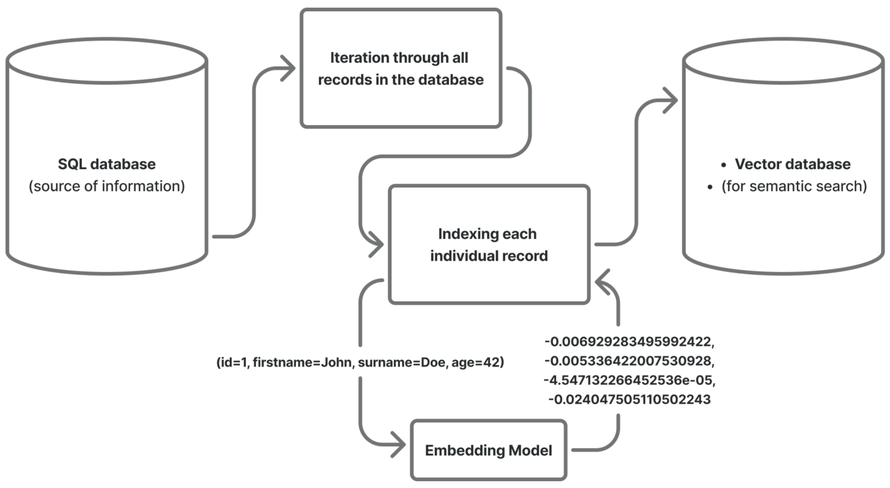

Figure 2 - database indexing

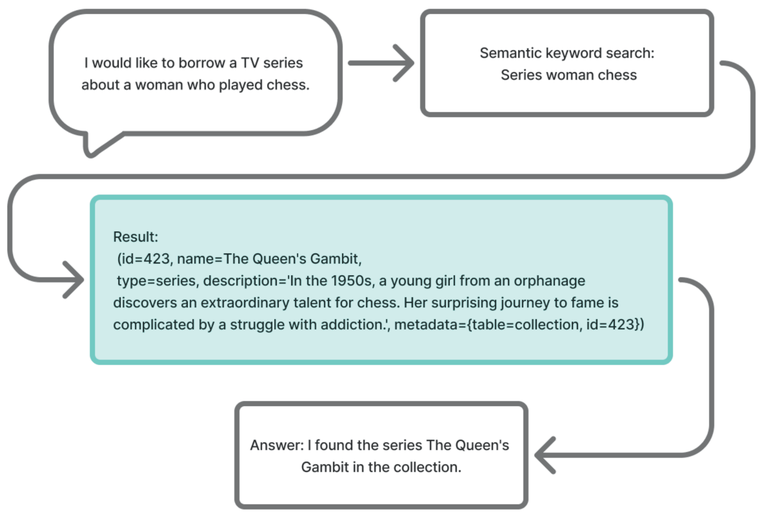

Figure 3 - example of semantic search

UI

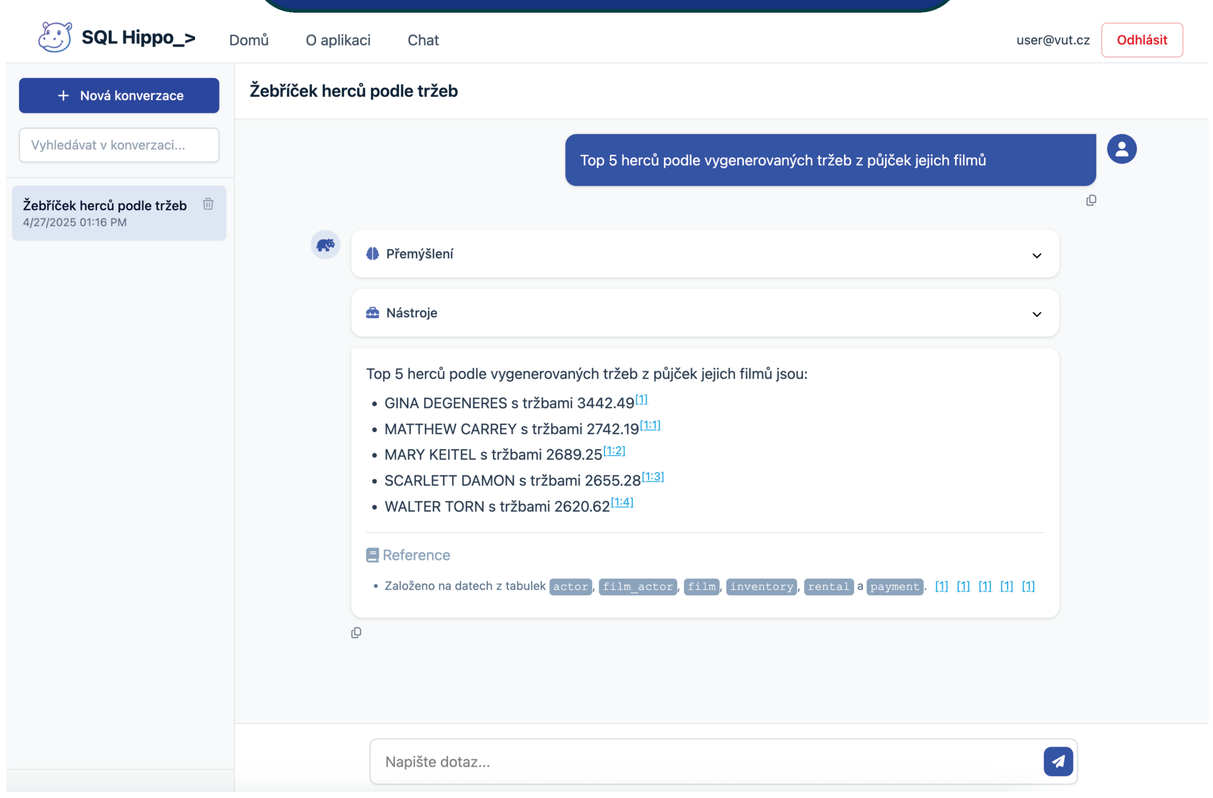

Figure 5 - Screenshot from the web application

Evaluation

Testing was conducted on manually scored questions that served as reference output. The scoring was performed by the o3-mini model.

latency	9.13s	11.17s	12.83s	17.41s	12.06s	10.85s	9.65s	12.72s	6.74s	8.27s	12.47s	18.47s	11.80s	11.51s	7.95s	17.78s	7.20s	3.76s	5.40s	12.52s	10.89s	8.70s	10.40s	8.08s	10.79s
tokens	27,367	35,067	35,307	49,292	28,587	22,463	33,821	23,715	39,049	32,854	37,079	75,206	29,695	42,055	28,697	56,874	26,856	12,809	20,065	38,190	28,315	28,809	30,555	27,828	28,456
correctness	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
correctness_judge	1	1	1	0	0	1	1	1	1	0	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1
relevance	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table 1 - evaluation

Excel@FIT 2025

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY