Fast Paws

Implementation of Q&A bot using LLM application

made by Lukianytsia, Sachenko, Levchyk, Tovtyn, Krysanova, Kostianets
mentor: Ing. Martin Hasin

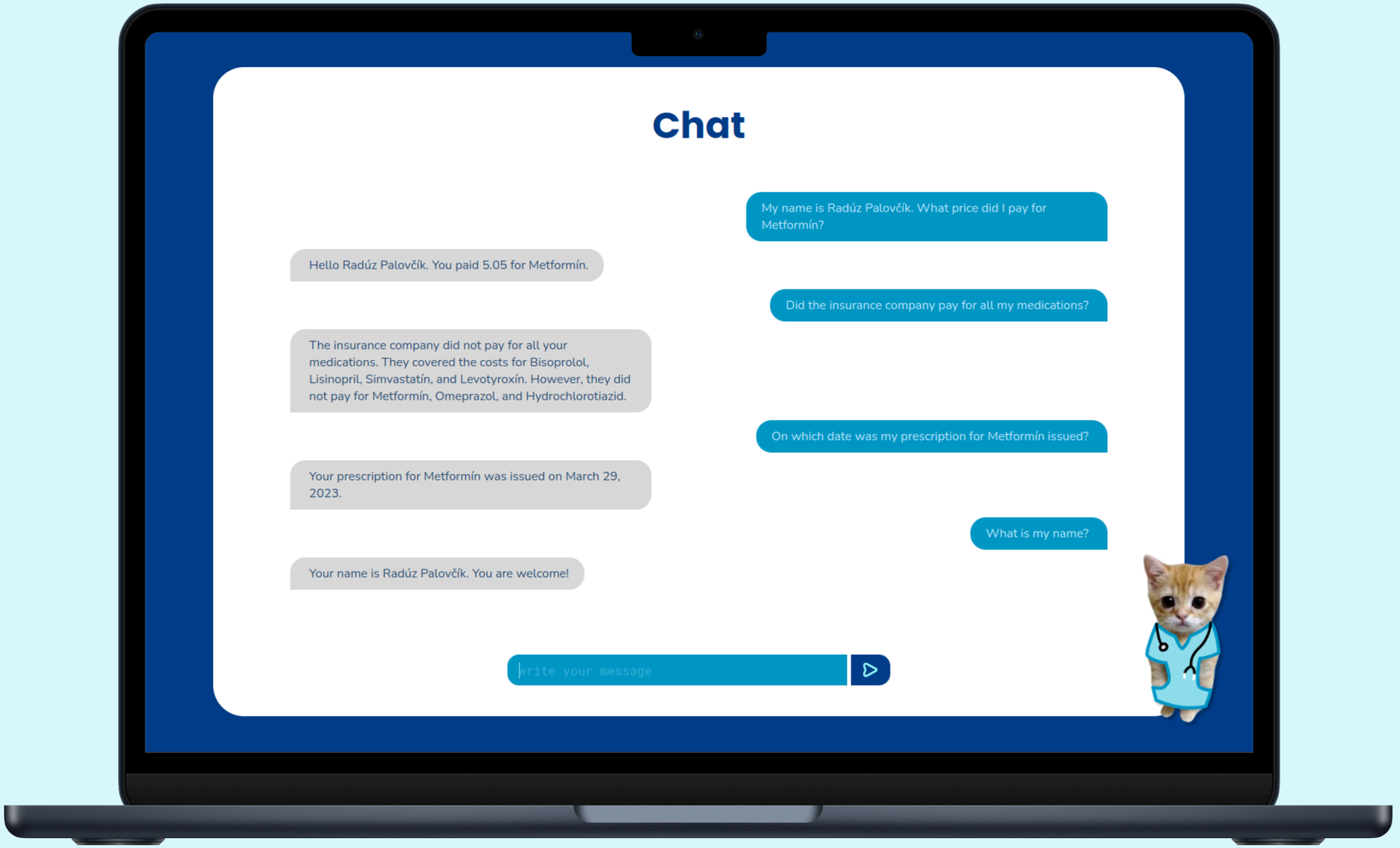
This is Tony

- he doesn't understand medicine
- he is so busy that he does not want to understand it
- he more easily perceives information from conversation

How can we **help** Tony?

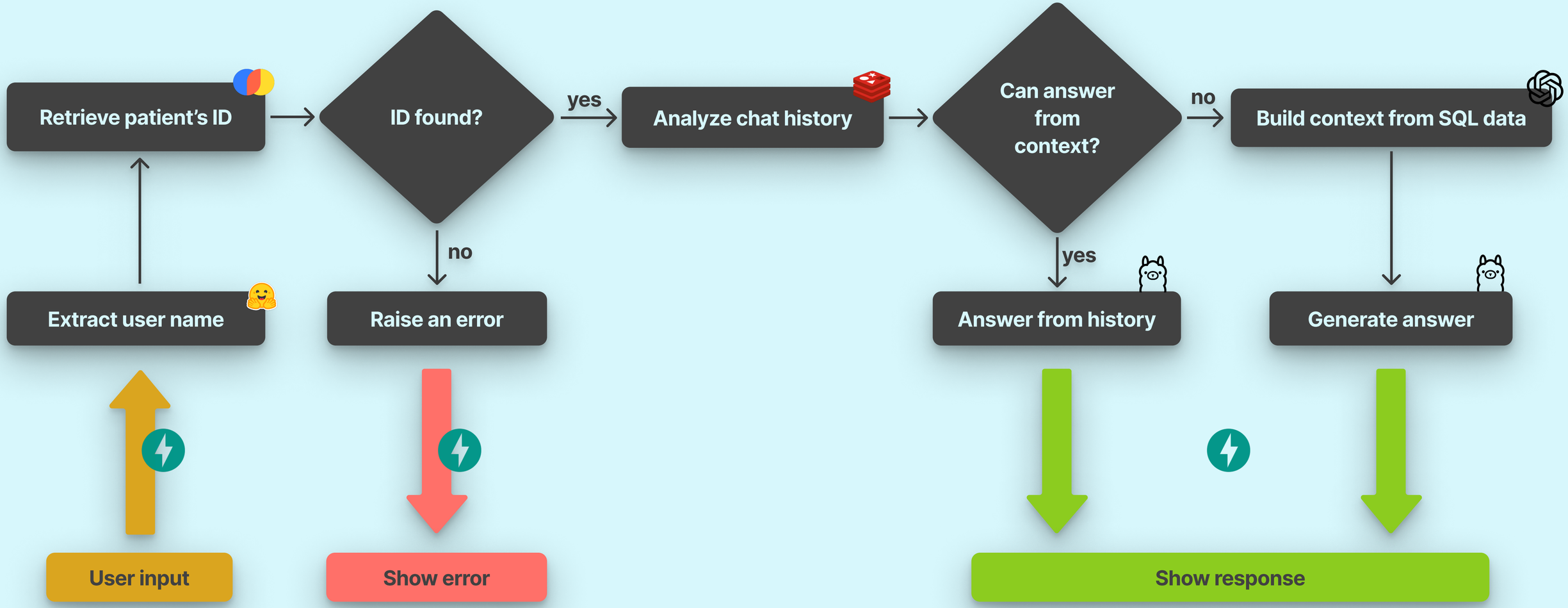
Expectations from the project

1. Create **workflow** for our patients
2. **RAG system** over static data
3. Compare **different LLMs**
4. Working with **tools**
5. Creating **an agent**



Application design

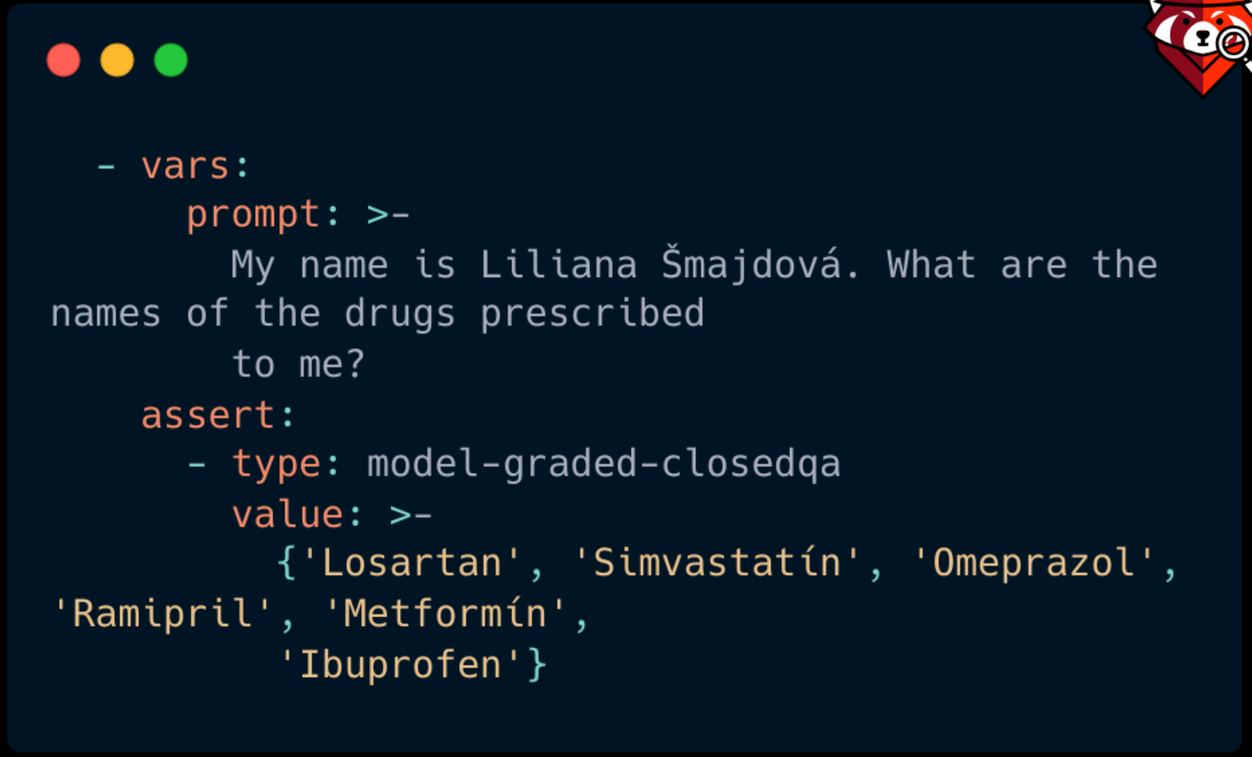
Application flow



Model evaluation

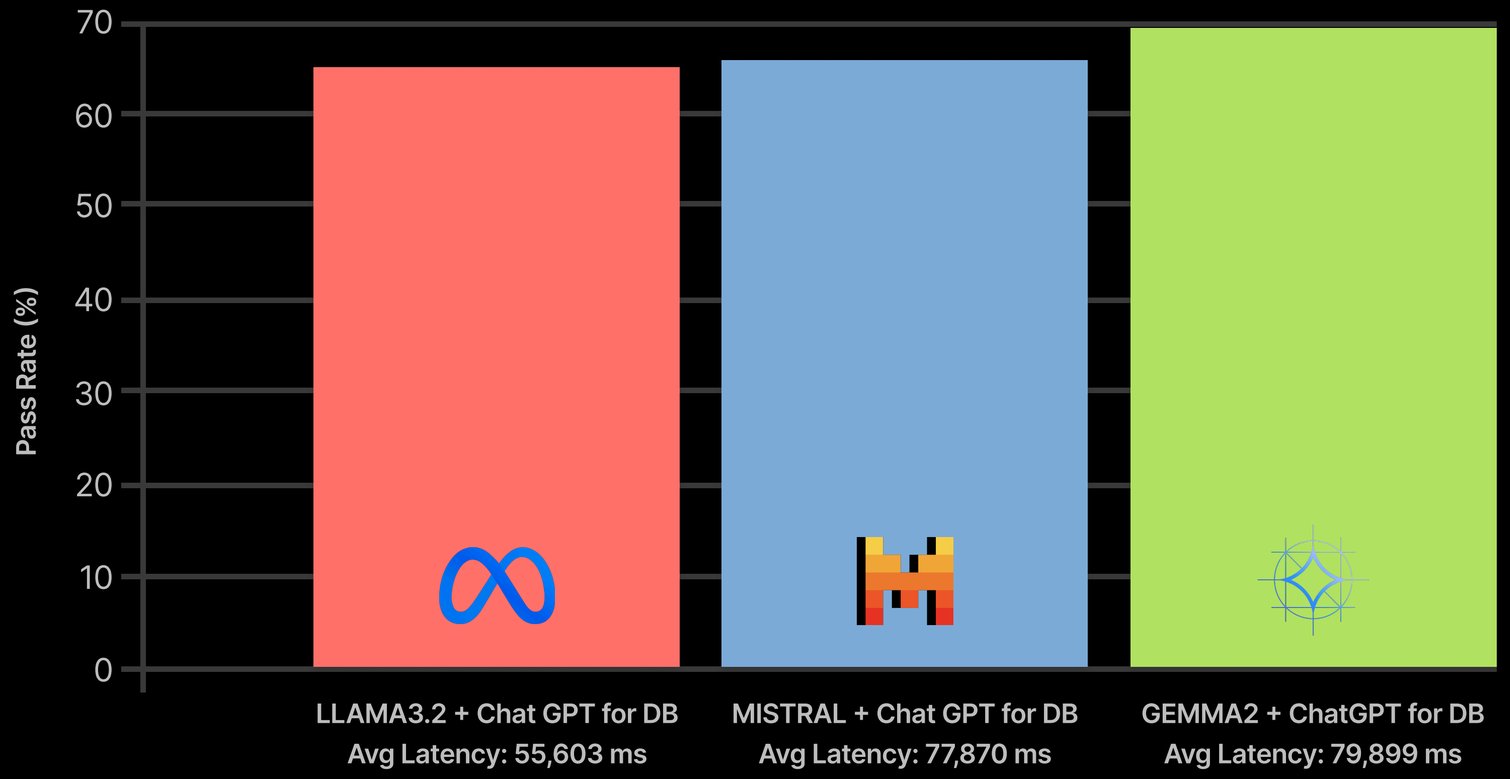
Evaluation methods

- Used **PromptFoo** to evaluate models.
- **100** randomly generated **tests** for diverse and robust **evaluation**, using **Pandas** on our **dataset**.
- Graded **model responses** using another **LLM** for **consistency**.
- Compared **models** based on **accuracy** and **latency metrics**.



```
  - vars:
      prompt: >-
        My name is Liliana Šmajdová. What are the
names of the drugs prescribed
        to me?
    assert:
      - type: model-graded-closedqa
        value: >-
          {'Losartan', 'Simvastatín', 'Omeprazol',
'Ramipril', 'Metformín',
          'Ibuprofen'}
```

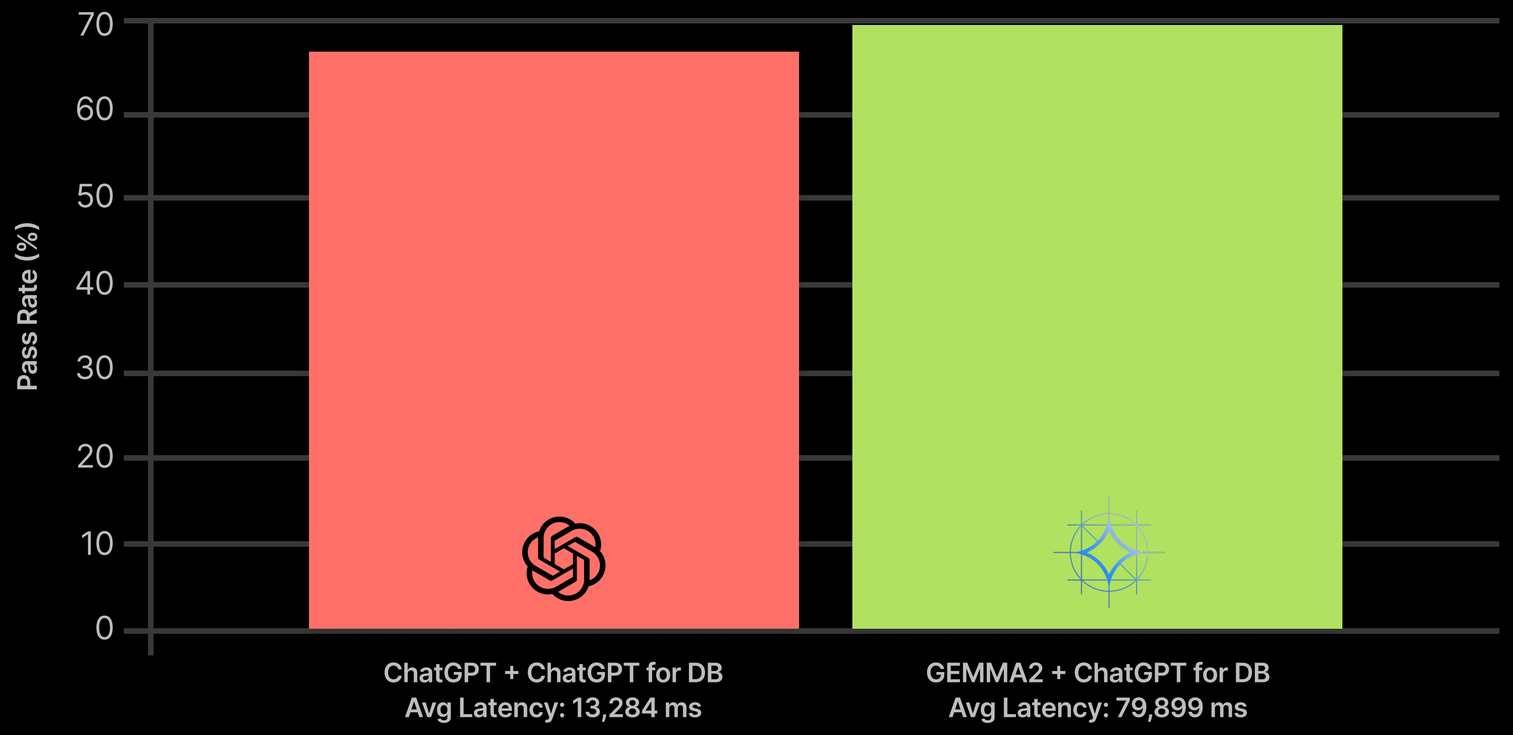
Hosted models



Evaluation results

- **GEMMA 2** - superior in **precision** and **relevance**, best for **complex**, **multi-step** queries.
- **MISTRAL** - balanced **performance**, ideal for **general-purpose** use.
- **OLLAMA 3.2** - **fastest** model, best for **real-time applications**, slight drop in **accuracy** for **complex** queries.

Cloud and Hosted models



Evaluation results

- **Hosted GEMMA2** achieves the highest **accuracy** but at the cost of **higher latency**.
- **Cloud ChatGPT** model have slightly lower **accuracy** but significantly **better latency**.

Challenges

- Abstraction level
- SQL queries quality
- Reasoning capabilities
- Tool calling
- Resources limitations

Future work

- Dynamic dataset integration
- Enhanced multilingual support
- Broader query capabilities
- Improving system security

Conclusions

- Developed a healthcare Q&A system using advanced LLMs.
- Used hybrid data architecture for better results.
- React-based interface ensured seamless user experience.
- Evaluated and selected optimal LLM for performance.

Demo

You can use this names for testing

- Servác Ďurčo
- Dominik Bobuľa
- Alexia Barbora
- Ladislava Palko
- Lucia Selecká
- Judita Pavlíčeková

*works only from TUKE network