

RWTHgpt latest developments

Initial situation

- Problem: Unregulated use of generative AI
 - Confidential information could be shared with not GDPR compliant service providers
- Solution: Provide GDPR compliant access to generative AI with RWTH terms of use
- Requirements:
 - Data remain on trusted servers in europe
 - Prompts and completions are not used to improve models
 - Budget control
 - Choice of different LLMs

RWTHgpt architecture

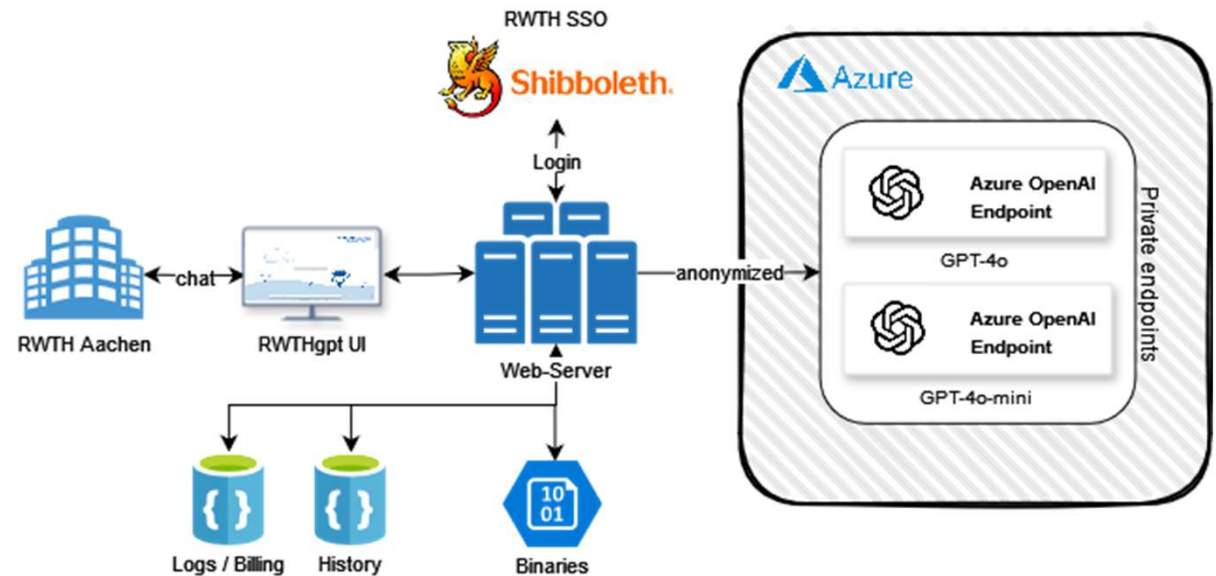
On premises

- Authentication via RWTH SSO (SAML)
 - Anonymized against Azure
- Chat history
- Customizable budgets
- Customizable limits (Token / Prompt)
- Individual budgets, limits, system prompt per role / group / project / ...
- RWTH UI

Azure

- Multiple LLMs hosted in Europe (Azure)
- Abuse Monitoring
 - Prompts stored for 30 days
- Prompts not used to improve LLMs or any 3rd party services
- Encrypted communication via SSL

RWTHgpt - simplified system architecture



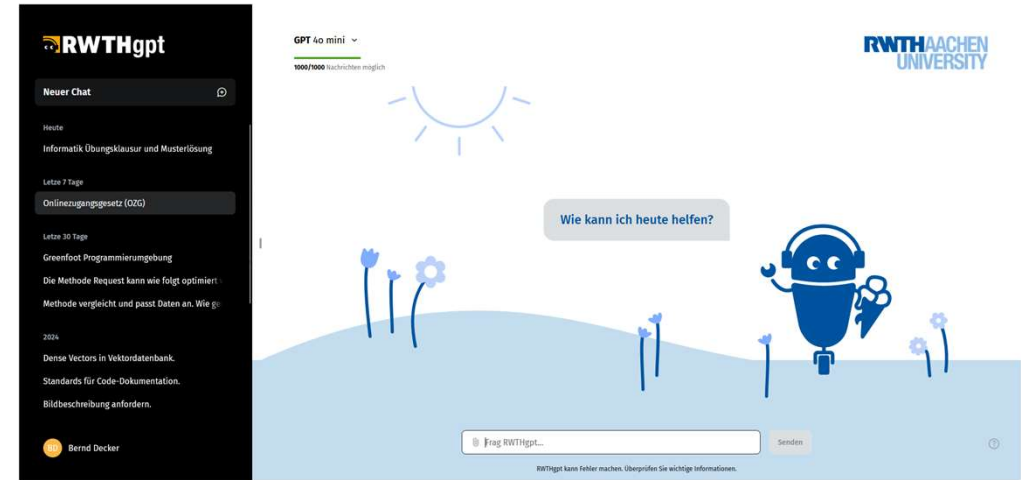
Go Live

Go Live (July 2024)

- RWTH Employees (approx. 10000)
- Staged roll out over 2 weeks
- Limits:
 - GPT-3.5t: 1000 prompts per hour
 - Menatime replaced by GPT-4o-mini
 - GPT-4o: 100 prompts per hour
 - Max context length 128000 tokens (4096 completion)

Infrastructure

- 4 Application Server (VM - Windows Server 2019)
 - 16 GB RAM
 - 4 vCPUs
 - 100 GB Storage
- MongoDB Cluster
- FileServer



Application features

- Textgeneration, summarization, translation
- Question and answer
- Chat with text documents
- Image processing
- Prompt editing
- Chat history

Stats

Avg daily costs per user

- 0,12 €

Avg daily max per user

- 4,15 €

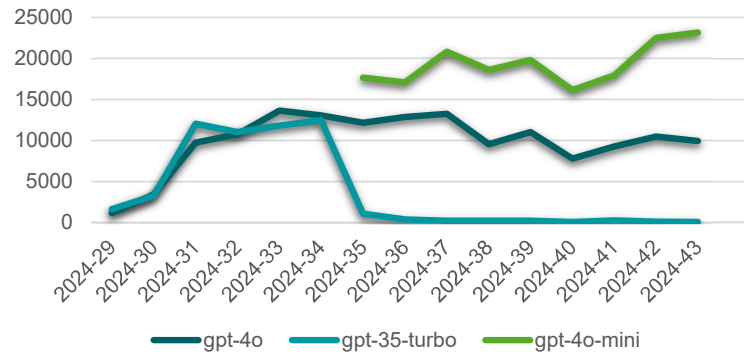
Avg prompt length

- 2256 tokens

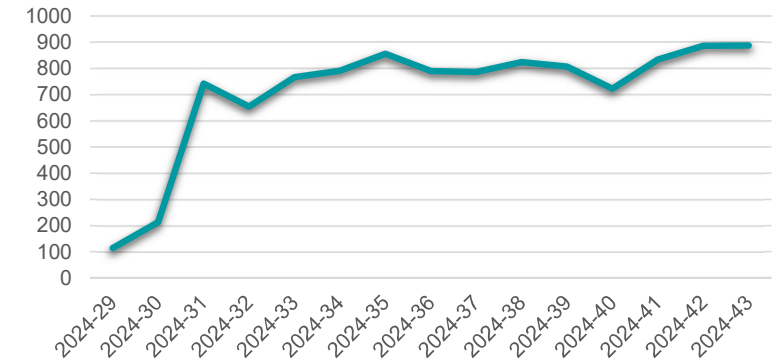
Avg max prompt length

- 43364 tokens

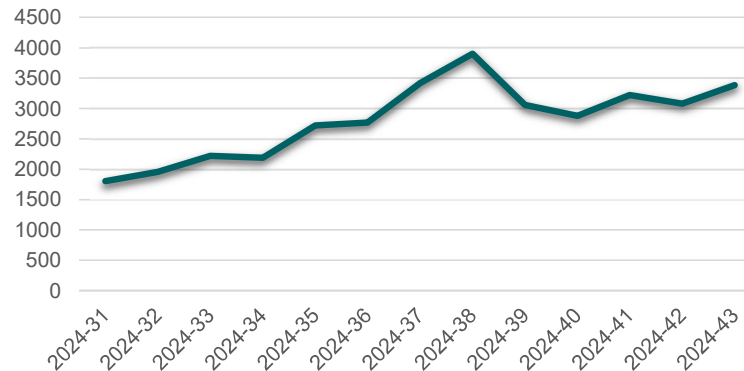
Prompts per week and model



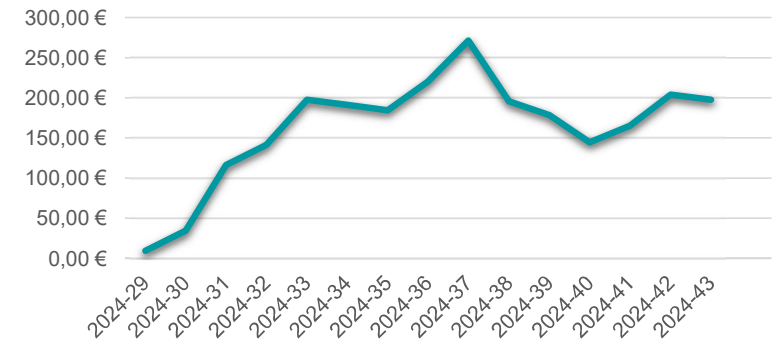
Unique users per week



Avg prompt length per week



Costs per week



Next steps

Onboarding students (November 2024)

- Seperate budget and accounting
- Limits on weekly token usage

Additional Models

- GPT-o1
- OpenSource Models (Llama, Mixtral,...)

KI:connect.nrw

- Statewide service for universities
- Additional features:
 - Login through DFN-AAI (SAML)
 - Themeable UI
 - Retrieval Augmented Generation (RAG)
 - Integration of open source LLMs
 - APIs for applications
 - OpenSource Repository (local installations)

Lessons learned

- Users still don't understand how generative AI works and what they can and cannot expect
 - Optimize the system-prompt
 - Offer prompt training courses

The screenshot shows a Google search interface with the query "Bernd Decker EUNIS president". The search results are displayed under the "Web" tab. The first result is from Eunis - European University Information Systems, titled "PRESENTATIONS – EUNIS2019", with a snippet mentioning Bernd Decker's impact on university networking. The second result is a PDF titled "EUNIS 2019 Congress Proceedings" from 2019, with a snippet identifying Bernd Decker as the deputy head of the IT process support team at RWTH Aachen University. The third result is from RWTH Aachen University, titled "Device specific credentials to protect from identity theft in Eduroam", with a snippet listing Bernd Decker as a corresponding author.

Bernd Decker EUNIS president. Don't use the internet for the answer.

The screenshot shows a ChatGPT response to the query "Bernd Decker EUNIS president". The response states: "Bernd Decker is indeed the president of EUNIS (European University Information Systems Organization). EUNIS is an association focused on advancing digital transformation and information technology within European higher education institutions. Decker, as president, plays a key role in driving initiatives that support IT development, collaboration, and innovation across universities in Europe. He likely works with a team to implement strategies that enhance the educational experience and administrative efficiency through technology and data integration." Below the text are several small icons for interaction.

ChatGPT answer

**Thank you very much
for your attention**

