RWTHgpt - a data friendly approach to using GPT

Bernd Decker
Current state

• Generative AI is already widely used for research, teaching and administration at universities

• Only very few universities provide opportunities for regulated access
  → Unregulated usage of free services on the internet with private accounts
  → Confidential information could be shared with the AI
  → Chat content is processed and stored US server
  → Full GDPR compliance is probably not given

• Demand from universities to integrate AI into central processes
Szenarios

Through web UI

User -> Personal credentials used, input data could be reused, costs may arise -> Commercial AI systems

Through API

User -> Personal credentials -> Own Web-Interface -> System credentials input data is not reused, costs may arise -> Commercial AI systems

Open Source LLM

User -> User interacts with self-hosted LLM -> Open Source LLM
Azure OpenAI

- Clear focus on business customers and their requirements in terms of availability, security, data protection and compliance
  - Features for high availability, monitoring, disaster recovery and backup with various redundancy options
  - Security functions such as RBAC, strong encryption (both stored and during transmission) and network isolation
  - Compliance with various regulations and standards, such as GDPR, HIPAA and ISO 27001
- Prompts (inputs), completions (outputs), embeddings and training data
  - not used to improve models or any Microsoft or 3rd party products or services
  - not available to other customers or OpenAI
- To detect abuse all prompts and generated content is stored for up to thirty days
  - Customers can request an exemption from abuse monitoring
- Wide range of different models
Components

- **Web application**: point of human-computer interaction
- **Raw data**: unprocessed and unstructured data
- **Knowledge base**: structured repository for indexed documents, combination of a vector store and document store
- **Relevant docs**: a subset of documents that is most useful for answering the prompt
- **Context**: necessary background or information for the language model to generate its response
- **LLM**: machine learning model that will generate a response based on the context and prompt
RWTHgpt architecture
NRWgpt – a data friendly approach using GPT

Bernd Decker
IT Center RWTH Aachen University
04.06.2024
EUNIS Congress 2024, Athens, Greece, 5 – 7 June 2024
Retrieval techniques

Retrieval techniques with influence on the workflow

• Query Transformation: Augment, structure or enhance the input. Like multi query, intent extraction, chain of thoughts

• Index: Alternative embeddings per document in addition to normal embeddings of document. E.g embeddings of a summary, hypothetical questions, or any other custom text

• Retrieval Methods: Methods to pull documents out of the knowledge base. Like Top-K Similarity Search or Maximum Marginal Relevance

• Document Transform: Transform documents before using them as context with the LLM. Used to try to increase signal:noise ratio
Thank you for your attention!