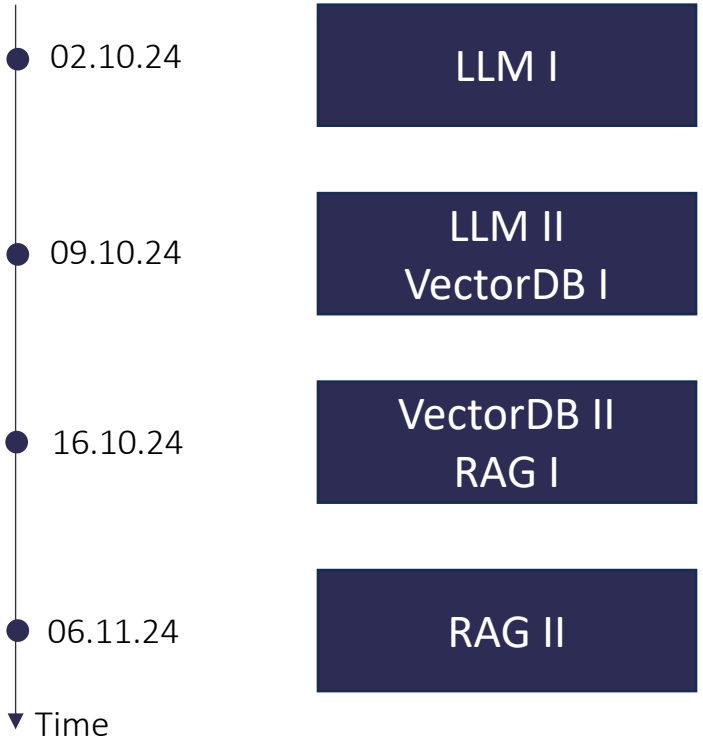
Course Schedule

gollnickdata.de

Schedule

Overview



Schedule

LLM Part I

02.10.2024

- What are Large Language Models?
- How are they trained?
- Which providers can be used?
- System Setup (VS Code, Python 3.12, API keys, etc.)
- How can you interact with an LLM via Python?
- What are system-user-AI messages?
- How to use prompt templates?
- How to use chains?

Schedule

LLM Part II / Vector DB Part I

09.10.2024

- What is a Vector Database?
- What is the difference between a classical and a vector DB?
- How does the data ingestion process work?
- How can data be loaded?
- How are chunks created?
- What are embeddings?
- How do you create chunk embeddings?
- Which database providers can be chosen? How do you select a specific provider?
- How do you store data in a vector database?

Schedule

Vector DB Part II / RAG Part I

16.10.2024

- What is RAG?
- How does it work?
- Implementing RAG in our custom project.
- If time allows, how to make it part of a web application.

Schedule

RAG Part II

06.11.2024

- Technical questions of workshop participants related to details of their own RAG system implementation.
- Basic techniques for tuning and managing the quality of RAG systems in the context of selected domains.
- Initiating discussions about possible further joint research projects.

Schedule

Buffer for technical workshops

13.11.2024

- Technical questions of workshop participants based on identified issues related to details of their own RAG system implementation.
- Security, regulatory compliance, and scalability of RAG systems.
- Short demos of participants' own RAG systems.
- Buffer for technical questions and workshops based on identified issues.

Schedule

Agentic Systems

20.11.2024

Technical Questions

Agentic Systems

Schedule

20.11.2024: Technical Questions – Practical (1_166324_P)

What problems can happen while using RAG?

- How can I sensibly assign weights to embeddings to make them more accurate?
- scripts\TechnicalQuestions\1_166324_P\main.py
- [embedding model leaderboard]

```
Similarity Scores:
Paragraph: Dogs like to go for walks, Similarity Score: 0.6106970310211182

Distances:
Paragraph: Dogs like to go for walks, Distance: 48.289947509765625

Sorted Similarity Scores:
Paragraph: Dogs like to go for walks, Similarity Score: 0.6106970310211182
Paragraph: They like to run around the park and play with other dogs, Similarity Score: 0.5818507075309753
Paragraph: Iwo likes pancakes, Similarity Score: 0.3023950457572937
Paragraph: , Similarity Score: 0.22406500577926636

Sorted Distances:
Paragraph: They like to run around the park and play with other dogs, Distance: 43.6199951171875
Paragraph: Dogs like to go for walks, Distance: 48.289947509765625
Paragraph: Iwo likes pancakes, Distance: 57.65611267089844
Paragraph: , Distance: 93.8907470703125
```

Schedule

20.11.2024: Technical Questions – Practical (1_166324_P)

What problems can happen while using RAG?

- At the same time, I am not sure whether the cosine similarity result is acceptable and accurate.
- cosine similarity most used, but maximum margin relevance another one

```
Similarity Scores:
Paragraph: Dogs like to go for walks, Similarity Score: 0.6106970310211182



Distances:
Paragraph: Dogs like to go for walks, Distance: 48.289947509765625



Sorted Similarity Scores:
Paragraph: Dogs like to go for walks, Similarity Score: 0.6106970310211182
Paragraph: They like to run around the park and play with other dogs, Similarity Score: 0.5818507075309753
Paragraph: Iwo likes pancakes, Similarity Score: 0.3023950457572937
Paragraph: , Similarity Score: 0.22406500577926636



Sorted Distances:
Paragraph: They like to run around the park and play with other dogs, Distance: 43.6199951171875
Paragraph: Dogs like to go for walks, Distance: 48.289947509765625
Paragraph: Iwo likes pancakes, Distance: 57.65611267089844
Paragraph: , Distance: 93.8907470703125
```

Schedule

20.11.2024: Technical Questions – Practical (1_166311_P-t)

Theoretical Questions

How can a RAG model improve the accuracy of recommendations based on specific product attributes like dimensions and descriptions?

- long-tail recommendations (not just relying on popularity)
- use LLM to create descriptions (dynamic content creation)

What are the challenges of using RAG models in product recommendation systems for retail, particularly in inventory management?

- data quality and consistency
- frequent changes of items
- scalability (large inventory)
- cold-start problem (new or niche products have few interactions)

Schedule

20.11.2024: Technical Questions – Practical (1_166311_P-t)

Theoretical Questions

How does RAG address limitations in standard LLMs for handling frequently updated data, such as product inventory?

- RAGs separate language generation from data retrieval → maintain stable LLM while regularly updating external data source

What retrieval strategies are best suited for RAG models in contexts where product data is heterogeneous (e.g., varied formats of dimensions, descriptions, and categories)?

- semantic retrieval
- multi-modal retrieval (CLIP) can process visual and textual inputs
- hybrid retrieval systems (text + traditional databases)

Schedule

20.11.2024: Technical Questions – Practical (1_166311_P-t)

Theoretical Questions

How can current trends (from sources like online magazines, social media, and trending topics) be incorporated into RAG-based recommendations to enhance relevance and optimize inventory management?

- include tools
  - web search
  - web scraping
  - → Agentic RAG

Schedule

20.11.2024: Technical Questions – Practical (1_166311_P-t)

Practical Questions

How should product data be structured in a database to enable efficient RAG-based recommendations based on dimensions and descriptions?

- unified data model (model that reflects structured and unstructured data)
- consistent schema (same schema for all products)
- implement hierarchical category structure
- use tags and keywords
- use metadata like ratings, popularity, or historical sales

Schedule

20.11.2024: Technical Questions – Practical (1_166311_P-t)

Practical Questions

What steps are necessary to convert a standard relational product database into a vector database for effective semantic retrieval?

- data extraction – text AND numerical information
- include title, description, attributes
- data preprocessing (clean data, normalize numerical data)
- feature engineering
- fusion
  - combine embeddings and numerical data into a single vector representation

Schedule

20.11.2024: Technical Questions – Practical (1_166311_P-t)

Practical Questions

Which embedding models are most suitable for converting product descriptions and dimensions into vectors for similarity matching in RAG?

- capture semantic meaning of text description / product description, as well as numerical attributes
- setup: vector embeddings for description, keep all other information as meta data
- perform pre-filter based on user selection

Schedule

20.11.2024: Technical Questions – Practical (1_166334_T)

Theoretical Questions

What are the main challenges in integrating image analysis within RAG systems?

- need multimodal model
- image analysis demands significant computational power
- ensure relevant image-text pairs
- RAG response time will be slower
- image quality might impact result

Schedule

20.11.2024: Technical Questions – Practical (1_166334_T)

Theoretical Questions

How does RAG handle missing or incomplete data?

- define fallback strategy like „no relevant information“ for similarity < threshold
- rephrase query
- augment with alternative data sources

Schedule

20.11.2024: Technical Questions – Practical (1_166334_T)

Theoretical Questions

Are there any control techniques to ensure that RAG always generates truthful responses, especially in critical areas such as medicine or law ?

- verify data sources
- post processing to verify generated responses
- human oversight
- restrict retrieval to specific trusted datasets
- provide confidence score with each response – indicator for retrieval evidence

Schedule

20.11.2024: Technical Questions – Practical (1_166334_T)

Theoretical Questions

What are the best practices for selecting data sources for the RAG system to ensure their reliability and relevance?

- use data from reputable sources (peer-reviewed, official reports)
- verify data consistency across multiple sources
- ensure sources are up-to-date
- ensure data quality
- verify legal compliance
- continuous evaluation

Schedule

20.11.2024: Technical Questions – Practical (1_166334_T)

Theoretical Questions

How can we address privacy concerns when using sensitive data in RAG systems, especially in healthcare or finance sectors?

- main idea: critical information should not leave the company network → avoid closed source providers, run LLM locally
- show Ollama
- alternatives
  - anonymization
  - data encryption
  - access control

Schedule

20.11.2024: Technical Questions – Practical (1_166334_T)

Theoretical Questions

What role does reinforcement learning play in the ongoing adaptation and improvement of RAG systems?

- not much yet, but
- RL can be used in tasks like retrieval, generation, or balancing the components
- idea: use user-feedback on answers as rewards to define the system behavior

Schedule

20.11.2024: Technical Questions – Practical (1_166334_T)

Theoretical Questions

How can RAG systems incorporate user feedback dynamically without requiring complete retraining?

- fetch user response, store it in DB
- retrieve data from vector DB (similarity)
- also retrieve user ranking of documents
- apply re-ranking algorithm

Schedule

20.11.2024: Technical Questions – Practical (1_166334_T)

Theoretical Questions

How could I calculate the profitability of using RAG in a given company?

- assumption „customer support“
- define metrics
- calculate cost of operating RAG system
- calculate improved efficiency
- calculate savings
- consider customer satisfaction

Schedule

20.11.2024: Technical Questions – Practical (1_166334_T)

Theoretical Questions

Do you think that the small companies will go bankrupt because of their bigger competitors using AI?

- yes, if a large provider can offer it directly, they will die
- https://www.threads.net/@codeforreal/post/C2pQzuRvuGz?from_lookaside=1
- https://www.perplexity.ai/

Schedule

20.11.2024: Technical Questions – Practical (166334_P)

Theoretical Questions

What features or aspects of the email are most important for correctly assessing emotions? Should they be keywords, phrases, sentence structure, or maybe other characteristics?

- some words are strongly associated with emotions, e.g. „happy“, „angry“
- sentiment dictionaries exist
- phrases „this makes me upset“ – explicit emotional context
- sentence structure: exclamation mark ! or !!! indicate intensity
- contextual sentiment – sentiment can shift within a sentence
- ML trained on labeled emotion datasets

Schedule

Buffer for technical workshops

27.11.2024

- Short demos of participants’ RAG systems. Buffer for technical questions about building RAG systems. (Discussion)

Schedule

Buffer for technical workshops

04.12.2024

- Short demos of participants’ RAG systems. Buffer for technical questions about building RAG systems. (Discussion)