Research Project 2025

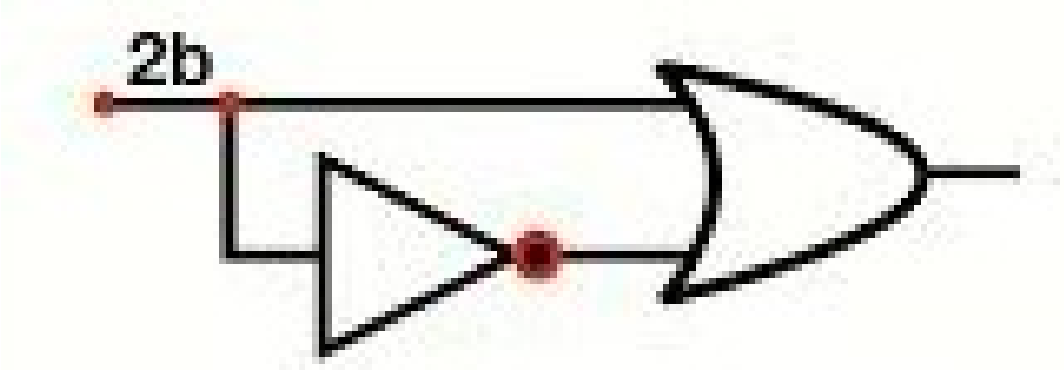
Talking with Tables

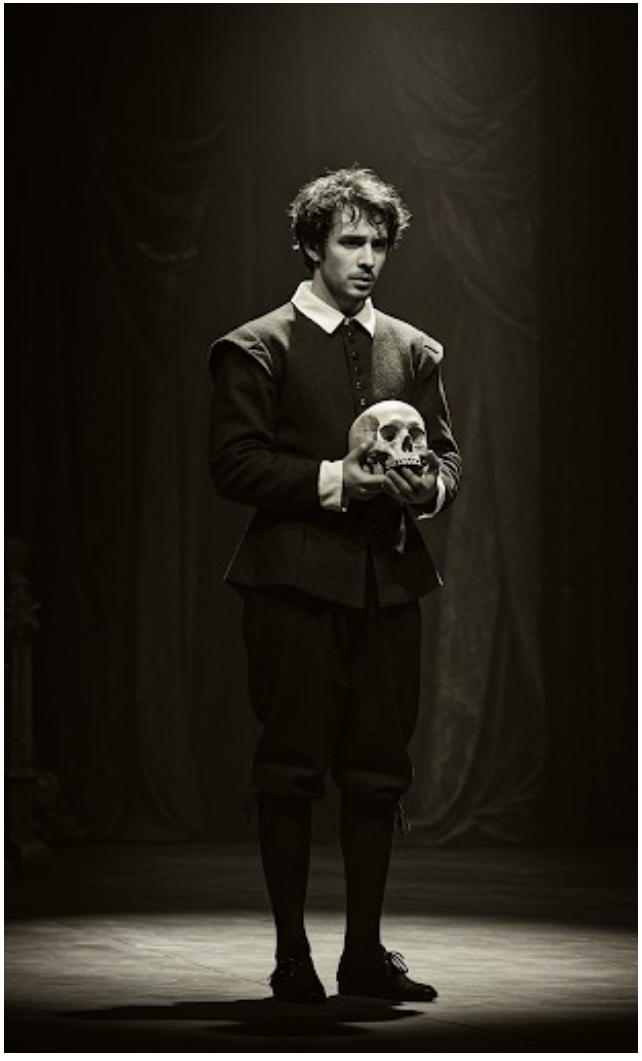
Introduction

- Prof. Dr. Jan-Torsten Milde
  - Head of the Digital Media program at Fulda UAS
  - I studied: **computational linguistics and text technology**
- Central research interest
  - Multimodal HCI (Speech and Gesture)
  - Robots in every day life
  - Cobots
  - NLP based information systems
  - Computer music :)

Talking to documents: RAG

KEEP CALM AND ...


2b



This is exactly the way LLMs work

no more, no less

RAG: Request Augmented Generation

- RAG for LLMs: RAG enhances LLMs by providing them with external knowledge, which helps to address issues like knowledge gaps, factuality, and hallucinations.
- It retrieves relevant documents and combines them with the original prompt to generate a more accurate and reliable output.

https://medium.com/intel-tech/tabular-data-rag-llms-improve-results-through-data-table-prompting-bcb42678914b

Talking to Tables

RAG: Talking to Tables

- Tabular Data Challenges: Handling tabular data in RAG systems presents unique challenges.
- Unlike text, tabular data needs to be **structured and formatted correctly** for LLMs to interpret it effectively.

- Problem space
  - Heterogeneity of Feature Types
  - Sparsity
  - Lack of Locality
  - Dependency on Preprocessing
  - Context-based Interconnection of Features
  - Order Invariance
  - Lack of Prior Knowledge about Structure
  - Sensitivity to Feature Engineering
  - Curse of Dimensionality with Categorical Data

https://arxiv.org/pdf/2402.17944

Why is this interesting ?

The Last Mile Problem

Data Analytics Marathon

Source: Brent Dykes | effectivedatastorytelling.com

Is there a (n easy) solution ?

- I don’t know
  - This is research
  - Let’s try to find out

What do you have to do ?

- Create a prototype of a system, that allows talking to a specific set of (table) data
  - Select and pre-process the data
  - Use python to implement the system
- Write a scientific paper about this prototype
- Attend the (five) sessions during lecture time
- Hand in the paper by the end of the semester

Semester plan and rules

Unfortunately, there are quite a few conflicting dates in the upcoming summer semester, so I will be offering a total of *6 in-person sessions* for the research seminar for up to *8 students*.

Datum	Wed (RP-DS)	Time	Room	Anmerkung
22.4.	+	46.009	17.00-18.00	Vorbesprechung
23.4.	+	51.203	15.30-17.00	Agentic AI, part I
7.5.	+	51.203	15.30-17.00	Agentic AI, part II
4.6.	+	51.203	15.30-17.00	Discussion Exposés
11.6.	+	51.203	15.30-17.00	Scientific Writing
2.7.	+	51.203	15.30-17.00	Presentation SoW
16.7.	+	51.203	15.30-17.00	(extra Meeting)
30.9.	+	51.203	15.30-17.00	Final Submission

2.1 Rules

Please register with me via email (milde@hs-fulda.de, Subject: RP DS) by tomorrow, *April 23, 2025, at the latest*. The first 8 students to register will receive the places. Please have a look at the schedule overview for the DS/AI research project.

Repository of the project

- Material for the course
  - Slides, links and examples

https://github.com/drmilde/research-project-SoSe-2025.git

References

- Tabular Data, RAG, & LLMs: Improve Results Through Data Table Prompting, by Eduardo Rojas Oviedo with Ezequiel Lanza
- Large Language Models (LLMs) on Tabular Data: Predic-tion, Generation, and Understanding - A Survey, by Xi Fang et. al.
- Machine Learning for Tabular Data, XGBoost, Deep Learning, and AI, by Mark Ryan and Luca Massaron
  - https://github.com/lmassaron/Machine-Learning-on-Tabular-Data.git