Zürcher Hochschule
für Angewandte Wissenschaften

School of
Management and Law

EQUIS
since 2023

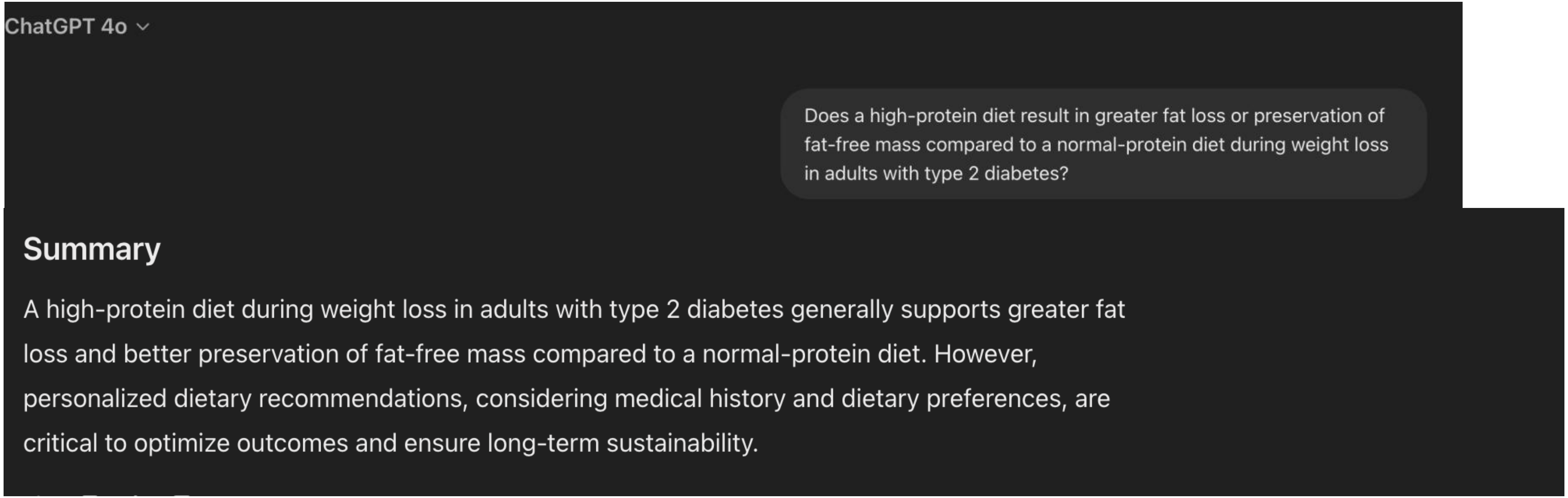
AACSB
since 2015

Retrieval Augmented Generation
Enhancing the Reliability of Generative AI

Building Competence. Crossing Borders.

Jasmin Heierli
jasmin.heierli@zhaw.ch / 11 March 2025

LMs are Great AND…



Does a high-protein diet result in greater fat loss or preservation of fat-free mass compared to a normal-protein diet during weight loss in adults with type 2 diabetes?

Summary

A high-protein diet during weight loss in adults with type 2 diabetes generally supports greater fat loss and better preservation of fat-free mass compared to a normal-protein diet. However, personalized dietary recommendations, considering medical history and dietary preferences, are critical to optimize outcomes and ensure long-term sustainability.

Problems with out of the box LM applications

- Information can be outdated
  - Check the cut-off-date for training data
- Retraining and updating an LM is expensive and time-consuming
- LMs are inconsistent
  - They know how words statistically relate to each other, not what they mean
  - LMs can be trained on false facts
- LMs are fantastic generalists, but not necessarily experts
  - Fine-tuning becomes more viable, but still takes a lot of effort (and potentially money)
  - Fine-tuned models may still leak private data from the training corpus
- LMs don’t know about non-public/internal data
  - We also don’t want them to know
  - But we want them to use it for specific applications
- Our best bet is prompt engineering
  - Solid prompt engineering still solves many problems

Possible Solution

New Prompt:

Does a high-protein diet result in greater fat loss or preservation of fat-free mass compared to a normal-protein diet during weight loss in adults with type 2 diabetes? Answer this question using the following information only:

<info>

“Both the HP (High Protein) and NP (Normal Protein) groups reduced fat mass and increased fat-free mass percent, with no significant difference between groups." Additionally, the results indicate that "total mass was reduced by 10.2±1.6 kg (9.4%) in the HP group and 12.7±4.8 kg (11.8%) in the NP group," and "the preferential loss of fat mass in both diet groups in the present study lessens concerns related to potential weight loss-related adverse events.”

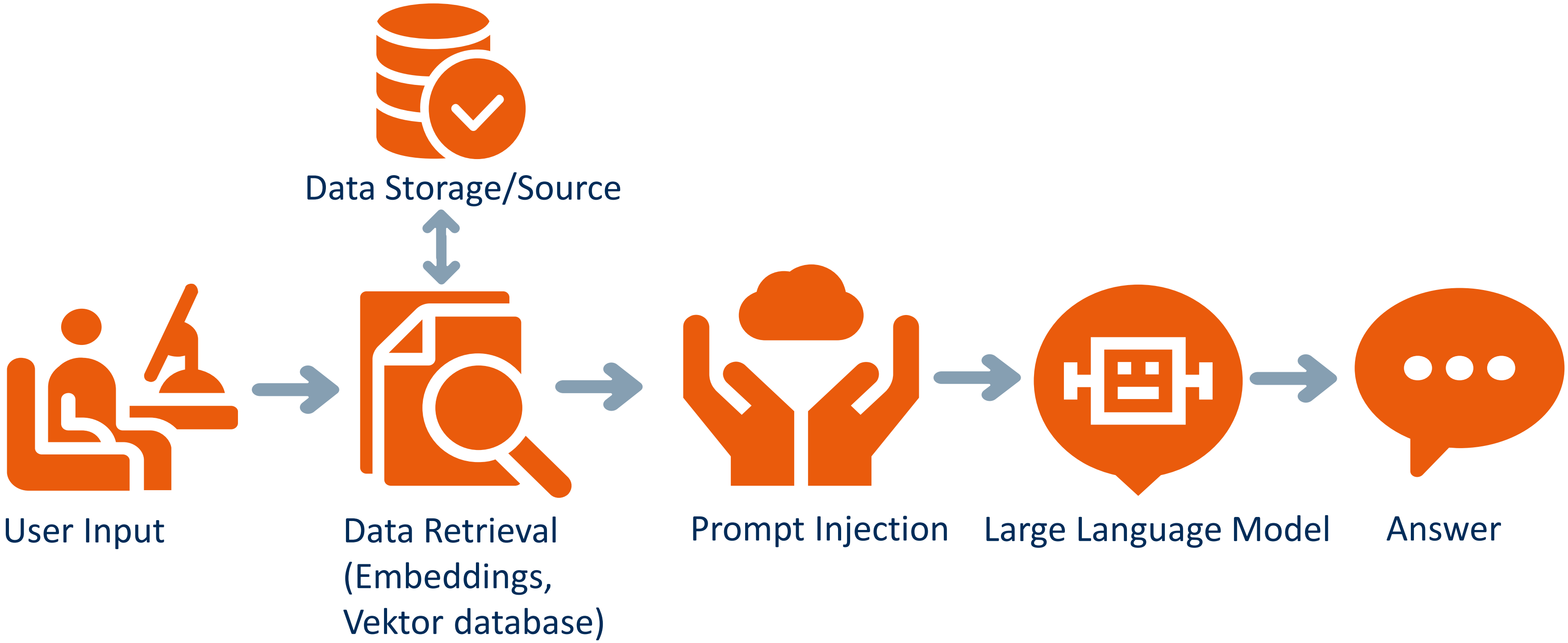
“Contrary to the hypothesis, the HP diet did not result in greater weight loss when compared to the NP diet. Instead, the groups had similar weight loss and body composition changes following the intervention. It was also hypothesized that the HP group would result in preferential loss of fat mass compared to fat-free mass, which was also not supported.”

</info>

What is RAG?

- **R**etrieval **A**ugmented **G**eneration is a Natural Language Processing (NLP) technology that retrieves information from external sources (retrieval) and uses this information to generate answers or context (generation).
- In contrast to earlier NLP methods, it is not necessary to train or improve a *specific* machine-learning model in advance. RAG actively searches for relevant information and incorporates the result into the answer generation in real-time.
- The first RAG framework has been first introduced by Meta in 2020 (https://arxiv.org/abs/2005.11401v4).

RAG Architecture



Data Storage/Source

User Input

Data Retrieval (Embeddings, Vektor database)

Prompt Injection

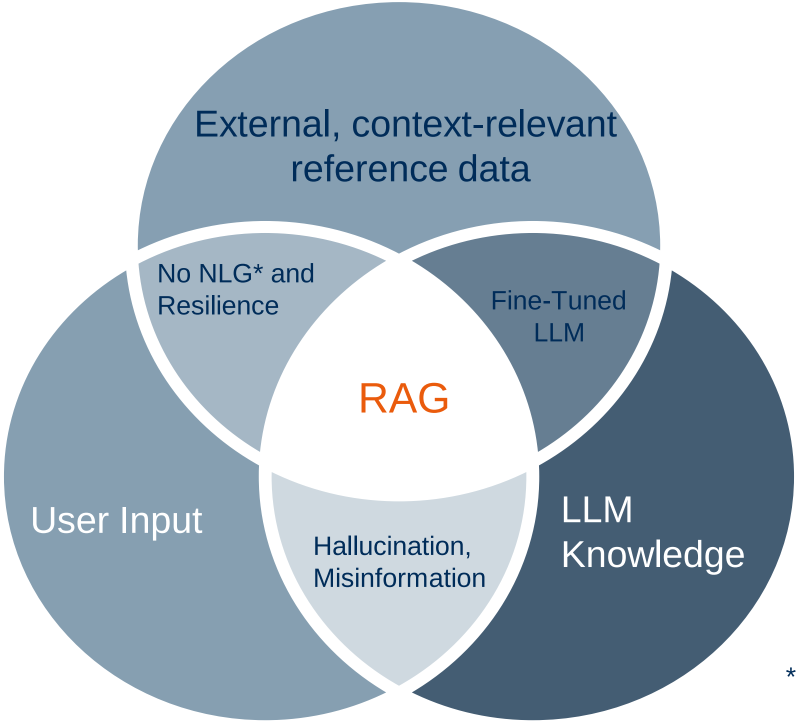
Large Language Model

Answer

RAG Capabilities

- LM „looks up the answer“ in the data you provide as opposed to searching in its memory
  - Open book vs closed book exam
  - Reduces Hallucinations
- Facts can be added in multiple ways
  - Adding browsing capabilities → useful to ensure that answers are up-to-date
  - Giving access to a curated database → useful to ensure that data is relevant and up to date
  - APIs
  - Knowledge Graphs etc.
- Low need for training data
- More likely disregards unnecessary information
- Respect context length limitations for prompts

Classification of Retrieval Augmented Generation



External, context-relevant reference data

No NLG* and Resilience

Fine-Tuned LLM

RAG

User Input

Hallucination, Misinformation

LLM Knowledge

*Natural Language Generation

Problem: How to pick the relevant information

- Given our user query: *Does a high-protein diet result in greater fat loss or preservation of fat-free mass compared to a normal-protein diet during weight loss in adults with type 2 diabetes?*
- Given a knowledge base (knowledge graph, database, collection of texts…)

What information from the knowledge base should be injected?
How would a human (you) solve this problem?

Exemplary Knowledge base: https://www.khou.com/article/news/local/houston-we-have-a-myth-what-was-really-the-first-word-spoken-from-the-moon/285-412522095

Requirements?

Given a knowledge base, e.g. the news article above and our user query: What do we need?

- The application shall have a complete prompt: role, task, chain of thought…
- The prompts shall be injected with paragraphs from our knowledge base
- The best matching paragraphs shall be identified with a user query
- The system shall store the data in a ChromaDB
- The data shall be ingested by the ChromaDB

offline

- The system shall use the prompt, including the injected paragraph, to generate a response with an LM

online

Solution 1: Langchain & Chroma

- Currently one of the state-of-the-art solutions for RAG
- Langchain: Framework to develop LM-powered applications with(out) 3rd party integrations
- Chroma: Open-source vector store (embedding database)

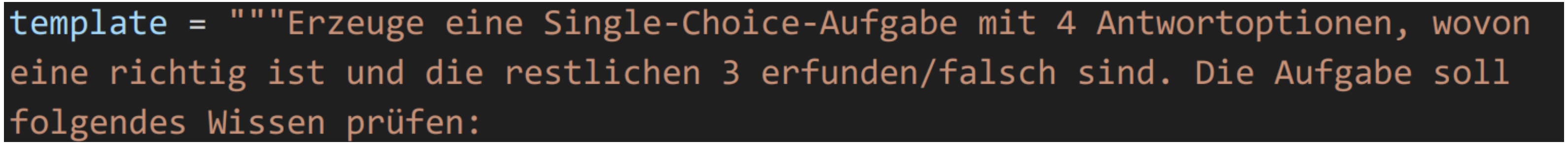
🦜🔗

LangChain + FAISS

https://medium.com/@onkarmishra/using-langchain-for-question-answering-on-own-data-3af0a82789ed

Langchain RetrievalQA Chain

- Chain: sequence of calls to a LM, a tool, or a data processing step
- Performs a retrieval step to find relevant documents in a database
- Inject relevant documents into LM to generate response
- Q(uestion) A(nswering) chain requires:
  - LM
  - Retriever (database)
  - Prompt Template



```
template = """Erzeuge eine Single-Choice-Aufgabe mit 4 Antwortoptionen, wovon
eine richtig ist und die restlichen 3 erfunden/falsch sind. Die Aufgabe soll
folgendes Wissen prüfen:
```

Exercise

- Offline: From PDF to ChromaDB, from User Query to relevant documents with prompt injection
- Online: Response Generation with injected prompt

https://github.com/zhaw-iwi/RAG-with-vector

Langchain Limitations

- If someone has done it before, you can often do it too
  - If you use the same version of Langchain and related packages
- Bleeding edge framework with lots of capabilities
  - Risk: new versions rename or (re)move desired functionality
- The current algorithm (February ‘24) does not include deduplication of retrieved documents
  - Reason: A document may contain several relevant parts for the query

🫤

Other Solutions: Assistant API

- Interface fo GPT-3.5 and GPT-4 provided by Openai
- Allows setup of custom assistants
- Has text processing capabilities
- Separate upload of documents to Openai
- Call to Assistant API with file ID generated after upload

Other Solutions: Assistant API

- Simple setup with just 2 steps
- Lots of under-the-hood processes
  - No influence or information on how the data is preprocessed
  - No influence or information on how the data is retrieved
  - Limitation on number of documents (max. 20) and size (max. 512 MB)
- Only available with Openai
- See: https://platform.openai.com/docs/assistants/overview

Other Preferred Solution: Direct Interface between Chroma and LM

- Data is loaded with an independent loader
- Text is split with customer or Langchain text splitter
- Free choice of embedding model
- Use any other vector store or database

Thank you.