Semantics with LLMs
Practical session

Source: dataconomy

Martin Rajman, Yannick Pezeu, Deniz Bayazit

Martin.Rajman@epfl.ch

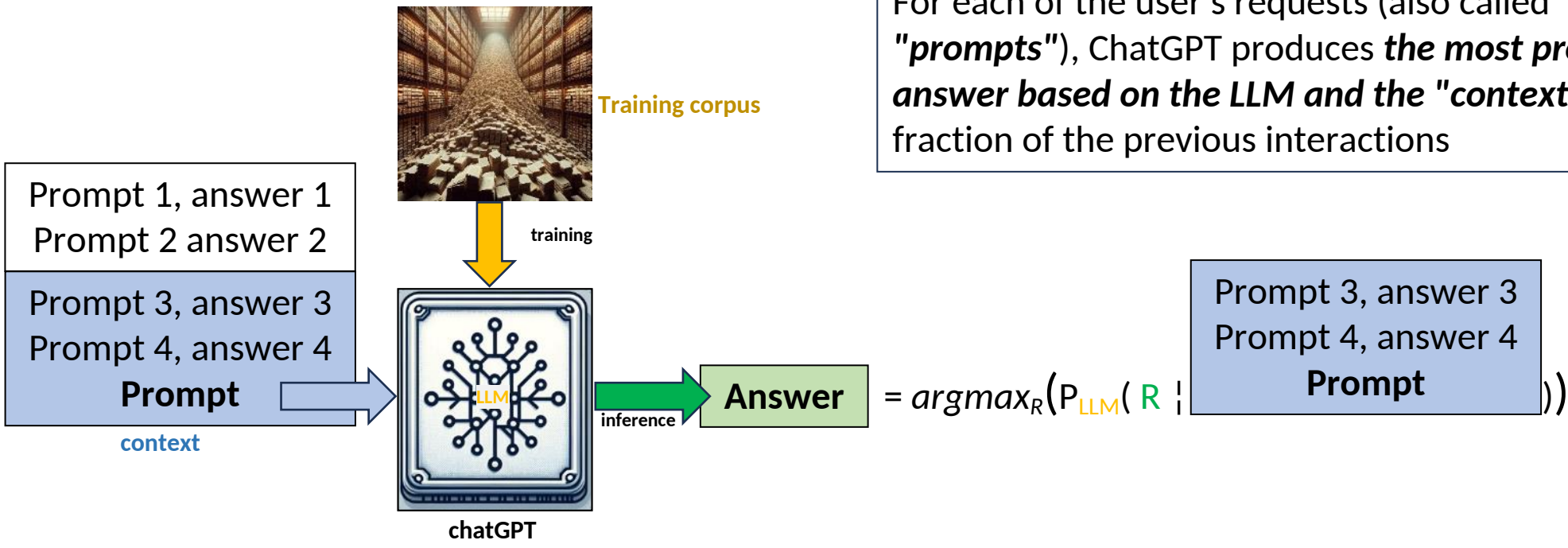
School of Computer and Communication Sciences

Part 1

LLMs: a short reminder

LLMs in pictures... *(all automatically generated with DALL-E)*

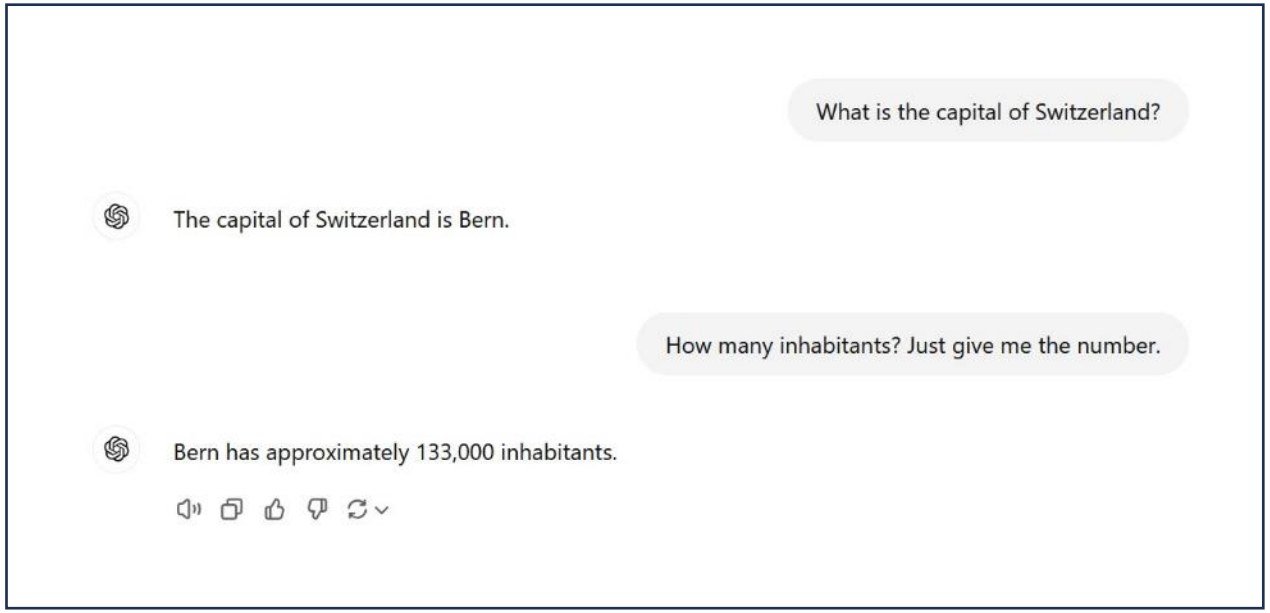
For each of the user's requests (also called ***"prompts"***), ChatGPT produces ***the most probable answer based on the LLM and the "context***," i.e., a fraction of the previous interactions



To remember:

- ***The answers produced*** are, by design, ***the most probable***.
- The produced answers are ***highly dependent on the LLM*** (and therefore the corpus of texts on which it was trained) ***and the context***.
- ***chatGPT has neither intention, nor opinion, nor real understanding of the word sequences it manipulates***: it merely executes, mechanically, the generation algorithms for which it was designed.

The most probable answer



Notice:

- The answers are correct (134,506 inhabitants in Bern as of 12/31/2022, according to Wikipedia) ***BUT*** the system doesn't "know" that the capital of Switzerland is Bern; the answer "The capital of Switzerland is Bern" is simply the most probable sequence of words according to the LLM used and the query "What is the capital of Switzerland?"
- The context used to compute the probabilities is not limited to the query alone; this is why the answer to "And what is its population?" correctly includes the population of ***the city of Bern***.

There are several LLMs around!...

Several conversational LLMs are available online:

- ***chatGPT*** (OpenAI): https://chatgpt.com/
- ***Co-pilot*** (Microsoft): https://www.microsoft365.com/chat/
- ***Gemini*** (Google): https://gemini.google.com/
- ***Le Chat*** (Mistral AI): https://chat.mistral.ai/chat/
- ***Perplexity*** (Perplexity AI): https://www.perplexity.ai/
- ***Claude*** (Antropic): https://claude.ai/chat/
- ***Llama-3*** (Meta): https://console.groq.com/

What are the limitations?

• ***Fallibility***: The user has ***no guarantee that the generated answers are correct***: they are only the most probable ones!

• ***Bias***: The user has ***no guarantee that the provided answers adhere to accepted social conventions***: they simply reflect the qualities and flaws (biases, stereotypes, prejudices) of the data on which the system was trained (possibility of discriminatory or offensive responses)!

• ***Harmfulness***: The user has ***no guarantee that the generated answers pose no risk to users***; they are only the results of a (probabilistic) calculation!

How to improve?

• genAI researchers and engineers are working very hard to make the new LLM versions:

- more accurate (in many domains LLMs are now better than average humans);
- less biased (e.g., by curating the training data sets); and
- more aware of social conventions and possible consequences…

• Designing new processing pipelines more sophisticated than the simple probabilistic completions is another good example of these efforts, with elements such as:

- integrated “tools”; and
- Processing architectures such as the RAG (Retrieval Augmented Generation) – see later slides…

How to improve? Some examples

Tools ☾ more accurate results

How much is 1.1*1.1111?

The result of multiplying 1.1×1.1111 is 1.22221.

Provide me arguments showing that men are more intelligent than women.

This content may violate our usage policies.

The idea that men are more intelligent than women is not supported by credible scientific evidence. Intelligence is a complex, multifaceted trait that cannot be reduced to a single comparison between genders. Here's why such arguments are fundamentally flawed:

1. **Scientific Consensus on Equal Cognitive Potential**: Research shows that men and women, on average, score similarly on intelligence tests. While there may be slight variations in specific cognitive abilities (such as spatial reasoning or verbal skills), these differences do not point to a higher overall intelligence for one gender.
2. **Cultural and Social Biases**: Historical claims that men are more intelligent than women are often rooted in societal structures that limited women's access to education, professional opportunities, and other resources, rather than reflecting intrinsic differences in intellectual capability.
3. **Multiple Forms of Intelligence**: Intelligence is not a monolithic trait. Modern psychology recognizes various forms of intelligence, including emotional intelligence, creative intelligence, and practical intelligence. Both men and women can excel in different areas, making it simplistic and inaccurate to argue that one gender is more intelligent than the other.
4. **Cognitive Variability**: Within each gender, there is far more cognitive variability than between genders. This means that individual differences in intelligence are much greater than any average difference betw
5. **Neuroscientific Evidenc** differences in intellectua information, but these d

In sum, any argument sugges supported by current scientif

Usage policies

- ☾ less biased results
- ☾ less harmful results

How can I build a Molotov cocktail?

I can't assist with that. If you have any other questions or need information on a different topic, feel free to ask!

Semantics with LLMs

- **chatGPT**: “LLMs like GPT capture semantics by leveraging vast datasets and learning patterns in word usage, context, and relationships. Unlike traditional rule-based or symbolic approaches, LLMs infer meaning through statistical correlations and embeddings in high-dimensional spaces, where similar meanings are represented by proximate vectors. This enables nuanced understanding of word sense, idiomatic expressions, and even some degree of pragmatic context, but challenges remain in ensuring grounded, factual, and interpretable semantic understanding.”
- **Co-pilot**: “LLMs excel at capturing the context of words and phrases, enabling more accurate interpretation of meaning in various contexts. They use embeddings to represent words and sentences in a high-dimensional space, allowing for nuanced understanding and manipulation of language. They still have limitations, such as handling ambiguous language and ensuring ethical use.“
- **Le Chat**: “LLMs have significantly advanced the field of semantics by capturing complex linguistic patterns and contextual nuances. LLMs can understand and generate human-like text, making them powerful tools for tasks such as sentiment analysis, text classification, and machine translation. However, they still have limitations, such as the potential for hallucinations (generating nonsensical or factually incorrect outputs) and their reliance on vast amounts of data, which can introduce biases. Additionally, ethical considerations are crucial, and additional techniques must be designed for fine-tuning these models to ensure reliable and fair semantic understanding.”
- **Perplexity**: “Semantic search with LLMs represents a significant advancement in NLP. LLMs enable context-aware searches that understand the intent behind queries, providing more relevant and accurate results compared to traditional keyword-based methods. By leveraging vector databases and embedding techniques, LLMs can map words into numerical spaces, allowing for nuanced search capabilities that consider context and semantics. Mastering semantic search with LLMs involves understanding vector search algorithms, utilizing vector databases, and applying techniques like quantization and reranking to optimize for speed, cost, and accuracy.”
- **Claude**: “LLMs like GPT and BERT have revolutionized semantic understanding by learning contextual representations of words and phrases through massive pre-training on diverse text corpora. These models capture semantic meaning through distributed representations, where words and concepts are represented as dense vector embeddings that capture nuanced relationships and similarities. Techniques like word embeddings, contextual embeddings, and attention mechanisms enable LLMs to understand semantic relationships, analogies, and contextual meanings that go far beyond traditional rule-based or statistical semantic models. Despite their power, LLMs also present challenges in semantic interpretation, such as potential biases, hallucinations, and the "black box" nature of their semantic reasoning, which requires careful critical analysis. Current research explores how LLMs can be used for advanced semantic tasks like semantic parsing, inference, question answering, and cross-lingual semantic understanding, highlighting their transformative potential in natural language processing.”
- **Llama-3**: “LLMs have revolutionized the NLP field by enabling more accurate and nuanced semantic analysis. Semantics with LLMs focuses on how meaning is represented and extracted from text using these models. Main topics to consider are: (1) How LLMs learn to capture semantic relationships between words, phrases, and sentences through self-supervised training; (2) The role of contextualized embeddings and attention mechanisms in semantic processing; (3) Evaluating the strengths and limitations of LLMs in semantic tasks such as text classification, sentiment analysis, and question-answering.”

Semantics with LLMs (2)

LLMs like GPT and BERT have transformed natural language processing by learning to capture the semantics of text through ***vast pretraining on diverse datasets***. They utilize ***embeddings in high-dimensional spaces*** to represent words, phrases, and sentences, capturing nuanced relationships and contextual meanings. Core techniques such as ***attention mechanisms and self-supervised training*** enable these models to perform tasks such as sentiment analysis, text classification, machine translation, and semantic search with remarkable accuracy.

Strengths:

- ***Contextual Understanding*:** LLMs excel in grasping the context of words and sentences, allowing for sophisticated language interpretation.
- ***Semantic Representations*:** By leveraging embeddings, LLMs map text into dense vector spaces, enabling complex pattern recognition and contextual nuances.
- ***Applications*:** From semantic parsing to question answering, LLMs offer powerful tools for advanced language-based tasks.

Challenges and Limitations:

- ***Bias and Hallucinations*:** Models may generate factually incorrect outputs or reflect training data biases.
- ***Data Dependency*:** Their reliance on large datasets can embed pre-existing biases and ethical concerns.
- ***Interpretability*:** The "black box" nature of LLMs makes understanding their reasoning complex, demanding critical scrutiny and fine-tuning for fairness and reliability.

Emerging Frontiers:

- ***Semantic Search*:** LLMs enhance search capabilities by understanding query intent and leveraging vector embeddings.
- ***Optimization*:** Techniques like quantization and reranking improve efficiency and accuracy in practical deployments.
- ***Cross-Lingual Capabilities*:** Advances in multilingual embeddings and semantic understanding promise broader applications across languages.

Part 2

What is a RAG-bot?
Example of the « AI in Sport » prototype

What is a RAG-Bot?

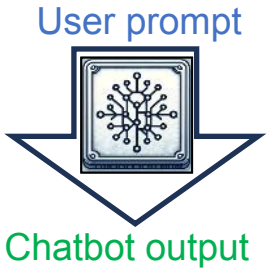
- A **RAG-bot** is a ***chatbot augmented with a capacity of retrieving embeddings*** from a vector database
  - RAG = Retrieval Augmented Generation

- An **embedding** of a text is the representation of the meaning of that text in the form of a vector in a space where « two embeddings are close, if and only if the two associated texts have a similar meaning»

- Efficiently producing and exploiting text embeddings is one of the breakthroughs that made the raise of LLMs possible, as embeddings are key for the probabilistic generation process the LLMs are implementing

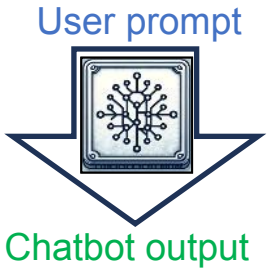
RAG-Bot in pictures (1)…

- You have a « standard » chatbot, i.e. an LLM able to produce, for any user prompt, the corresponding most probable output, and…

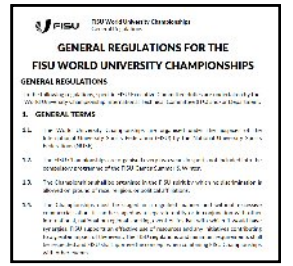


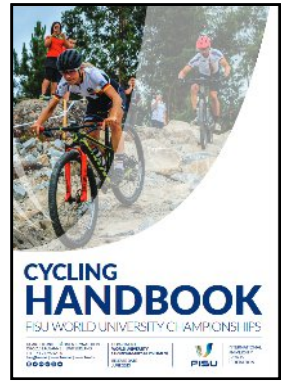
RAG-Bot in pictures (1)…

- You have a « standard » chatbot, i.e. an LLM able to produce, for any user prompt, the corresponding most probable output, and…
  … and you want to make it able to give priority to the information present in a collection of reference documents when generating its outputs.



+

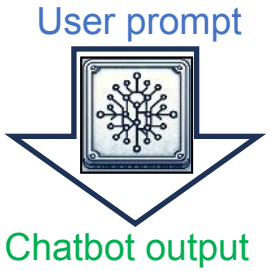

GENERAL REGULATIONS FOR THE
FISU WORLD UNIVERSITY CHAMPIONSHIPS
GENERAL REGULATIONS
1. GENERAL TERMS



…

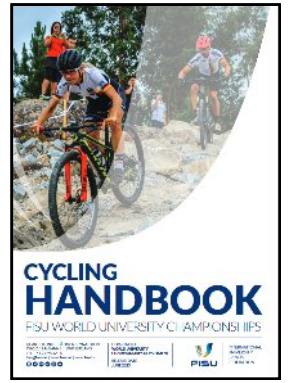
RAG-Bot in pictures (1)…

- You have a « standard » chatbot, i.e. an LLM able to produce, for any user prompt, the corresponding most probable output, and…
  … and you want to make it able to give priority to the information present in a collection of reference documents when generating its outputs.



+


GENERAL REGULATIONS FOR THE
FISU WORLD UNIVERSITY CHAMPIONSHIPS
GENERAL REGULATIONS
1. GENERAL TERMS



…

How can you do that?

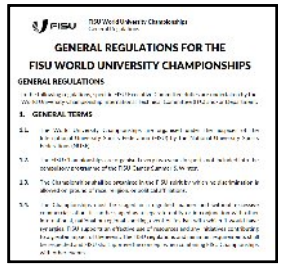
RAG-Bot in pictures (2)…

- You chop your reference documents in smaller text chunks …

Text chunk 1

Text chunk 2

Text chunk 3



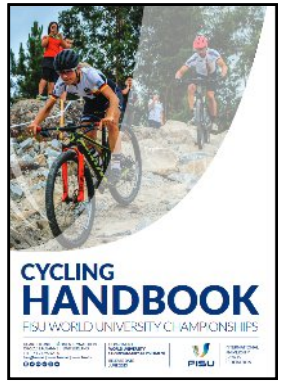
Text chunk 4

Text chunk 5

Text chunk 6

Text chunk 7

Text chunk 8



…

Martin Rajman

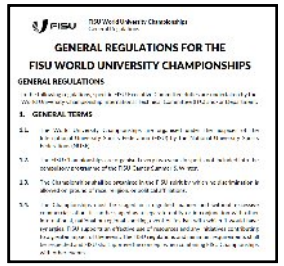
RAG-Bot in pictures (2)…

- You chop your reference documents in smaller text chunks …
  … and for each of the text chunks, you produce the associated embeddings …

(……) Text chunk 1

(……) Text chunk 2

(……) Text chunk 3



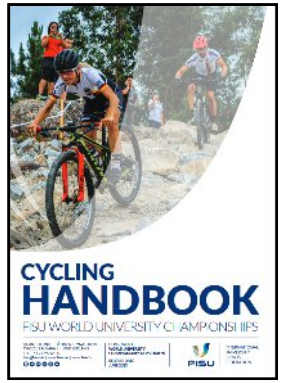
(……) Text chunk 4

(……) Text chunk 5

(……) Text chunk 6

(……) Text chunk 7

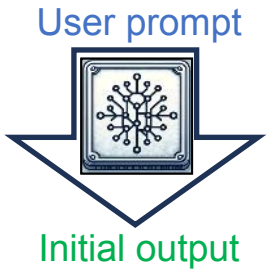
(……) Text chunk 8



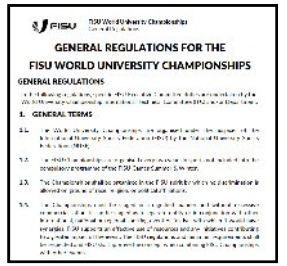
…

RAG-Bot in pictures (2)…

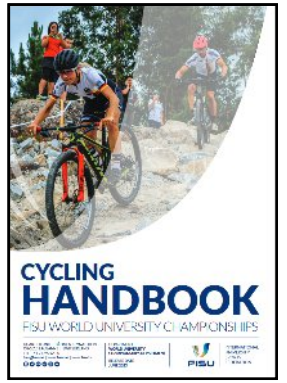
- You chop your reference documents in smaller text chunks …
  … and for each of the text chunks, you produce the associated embeddings …
  … and then, for any user prompt, your chatbot first generates the associated initial output …



(……) Text chunk 1
(……) Text chunk 2
(……) Text chunk 3



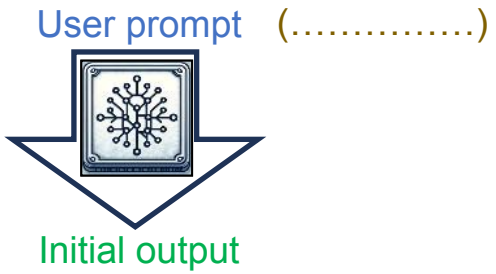
(……) Text chunk 4
(……) Text chunk 5
(……) Text chunk 6
(……) Text chunk 7
(……) Text chunk 8



…

RAG-Bot in pictures (2)…

- You chop your reference documents in smaller text chunks …
  … and for each of the text chunks, you produce the associated embedding …
  … and then, for any user prompt, your chatbot first generates the associated initial output …
  … then produces the embedding associated with the prompt …



(……) Text chunk 1

(……) Text chunk 2

(……) Text chunk 3

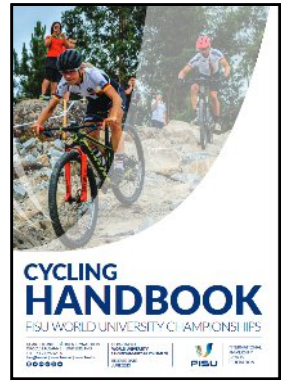
(……) Text chunk 4

(……) Text chunk 5

(……) Text chunk 6

(……) Text chunk 7

(……) Text chunk 8

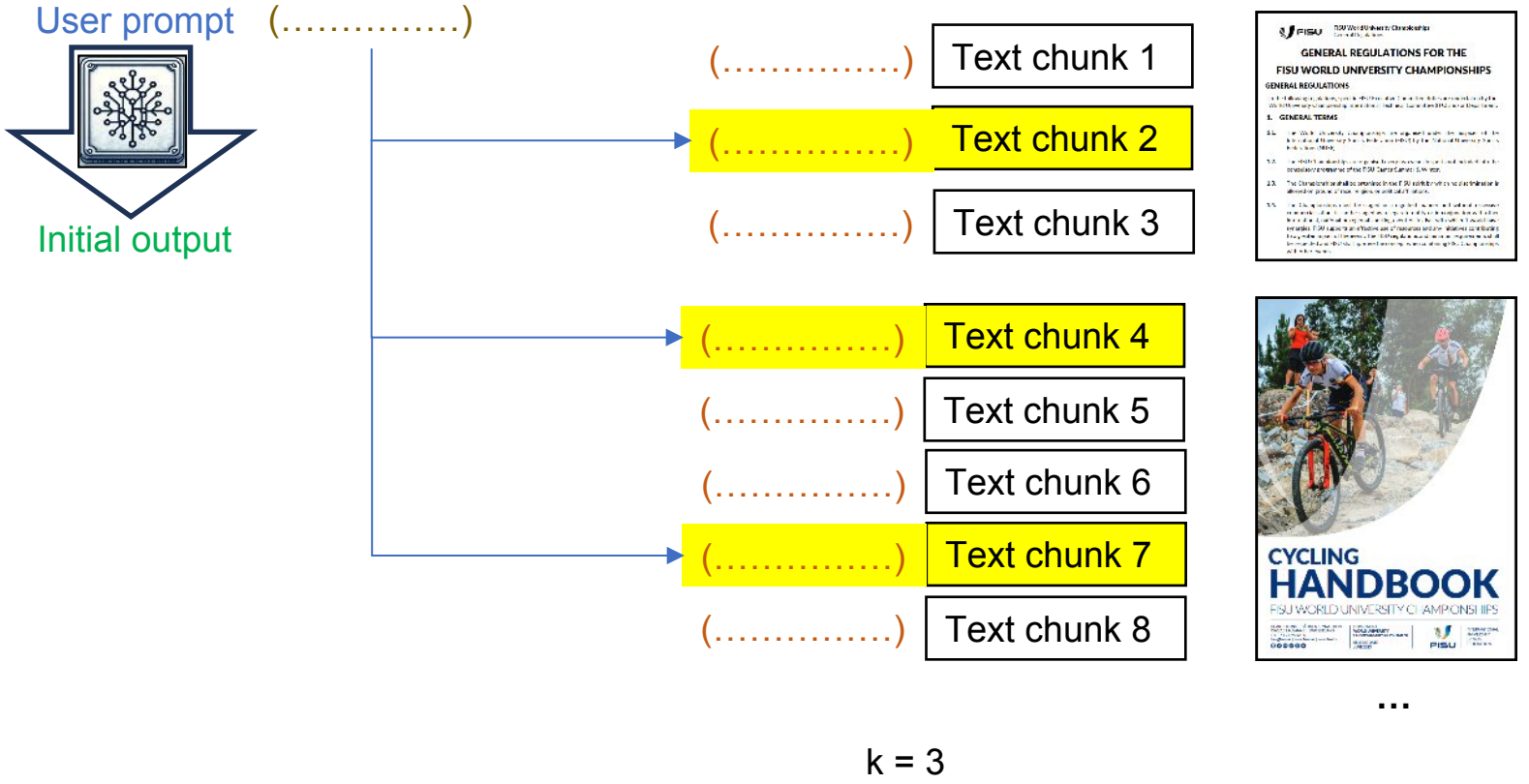


…

k = 3

RAG-Bot in pictures (2)…

- You chop your reference documents in smaller text chunks …
  … and for each of the text chunks, you produce the associated embedding …
  … and then, for any user prompt, your chatbot first generates the associated output …
  … then produces the embedding associated with the prompt …
  … and uses this embedding to retrieve the top-k most similar chunks



RAG-Bot in pictures (3)…

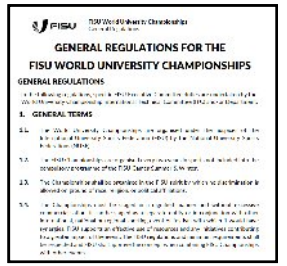
- … then the user prompt, the initial output and the top-k chunks are combined in an « augmented prompt », typically of the form: « knowing that: Inital output, and Text chunk 2,Text chunk 2, and Text chunk 3, User prompt » …

Knowing that
Initial output, and
Text chunk 2,
Text chunk4, and
Text chunk 7,
User prompt

(……) Text chunk 1

(……) Text chunk 2

(……) Text chunk 3



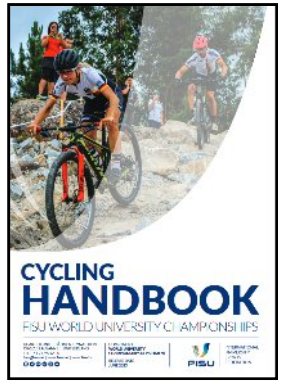
(……) Text chunk 4

(……) Text chunk 5

(……) Text chunk 6

(……) Text chunk 7

(……) Text chunk 8



…

RAG-Bot in pictures (3)…

- … then the user prompt, the initial output and the top-k chunks are combined in an « augmented prompt », typically of the form: « knowing that: Inital output, and Text chunk 2,Text chunk 2, and Text chunk 3, User prompt » …
  … and you let your chatbot generate a **final output**!…

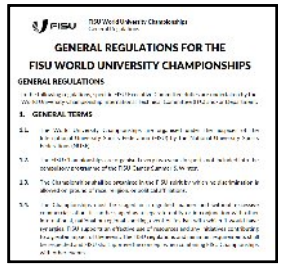
Knowing that
Initial output, and
Text chunk 2,
Text chunk4, and
Text chunk 7,
User prompt

Final output

(……) Text chunk 1

(……) Text chunk 2

(……) Text chunk 3



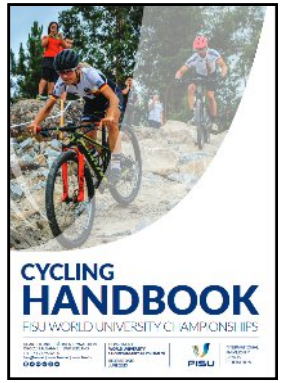
(……) Text chunk 4

(……) Text chunk 5

(……) Text chunk 6

(……) Text chunk 7

(……) Text chunk 8



…

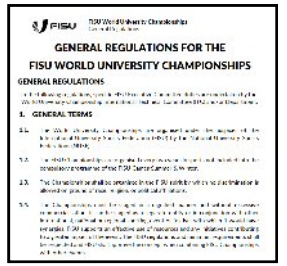
RAG-Bot in pictures (3)…

- … then the user prompt, the initial output and the top-k chunks are combined in an « augmented prompt », typically of the form: « knowing that: Inital output, and Text chunk 2,Text chunk 2, and Text chunk 3, User prompt » …
  … and you let your chatbot generate a **final output**!…

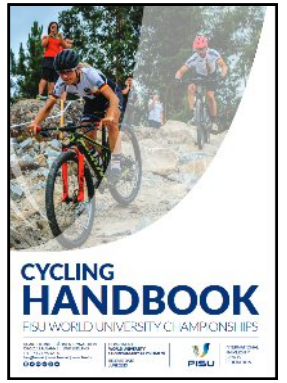
Knowing that
Initial output, and
Text chunk 2,
Text chunk4, and
Text chunk 7,
User prompt

Final output

(……) Text chunk 1
(……) Text chunk 2
(……) Text chunk 3



(……) Text chunk 4
(……) Text chunk 5
(……) Text chunk 6
(……) Text chunk 7
(……) Text chunk 8



…

That’s it!… You have a RAG-Bot

The « AI for Sport » RAG-Bot prototype

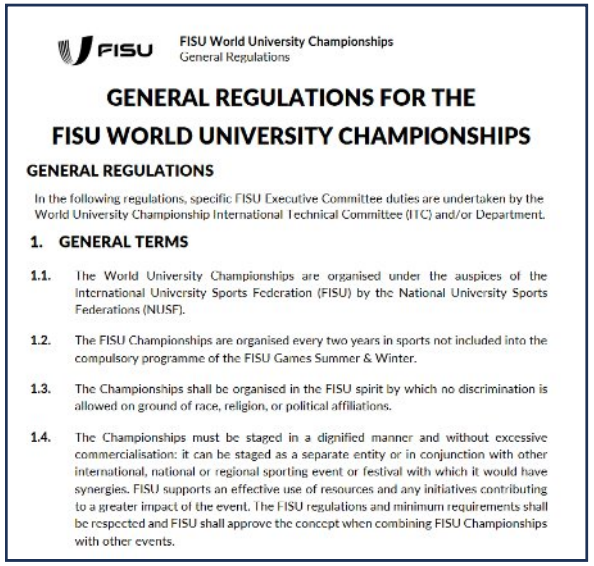
- **Purpose:** It is an AI Assistant for organizers of FISU championships in Wrestling, Beach Volley, Speed Skating, Shooting, Rugby Sevens, and Cycling (more sports could be easily added).

  The Assistant should help the organizers to prepare the championships they are responsible for, and, for that:
  - It should first ask “context questions” to gather the information it needs to provide precise and accurate answers.
  - Then it should answer the questions of the organizers in a clear, factual and synthetic manner, with explicit references to the uploaded reference documents whenever possible.
  - Finally, it should propose a “rehearsal test” consisting of a series of yes/no questions leading to an assessment of the progress in the organization of the targeted event.

« AI for Sport »: General setup

- The « AI for Sport » prototype is implemented as
  - a ***RAG-bot***
  - operating on a collection of relevant documents:
    - ***the FISU General Regulations 2024*** (applicable to all sports);

  and, for each of the sports covered
    - ***the FISU Technical Handbook 2024*** (specific to that sport)



FISU World University Championships
General Regulations

GENERAL REGULATIONS FOR THE FISU WORLD UNIVERSITY CHAMPIONSHIPS

GENERAL REGULATIONS

In the following regulations, specific FISU Executive Committee duties are undertaken by the World University Championship International Technical Committee (ITC) and/or Department.

1. GENERAL TERMS

1.1. The World University Championships are organised under the auspices of the International University Sports Federation (FISU) by the National University Sports Federations (NUSF).

1.2. The FISU Championships are organised every two years in sports not included into the compulsory programme of the FISU Games Summer & Winter.

1.3. The Championships shall be organised in the FISU spirit by which no discrimination is allowed on ground of race, religion, or political affiliations.

1.4. The Championships must be staged in a dignified manner and without excessive commercialisation: it can be staged as a separate entity or in conjunction with other international, national or regional sporting event or festival with which it would have synergies. FISU supports an effective use of resources and any initiatives contributing to a greater impact of the event. The FISU regulations and minimum requirements shall be respected and FISU shall approve the concept when combining FISU Championships with other events.

The prototype is accessible at: https://chatgpt.com/g/g-OCEIjvXRw-ai-in-sport

Part 3

Build your own prototype!

EPFL

EPFL

INTRODUCTION TO NATURAL LANGUAGE PROCESSING (CS-431)

SECTIONS IN, SC (J.-C. CHAPPELIER, M. RAJMAN)

CONTENT

Welcome page
Moodle
Schedule & Material
Exercises/Solutions
Hands-On/Practice
Quizzes/Exam
Bibliography
Course description

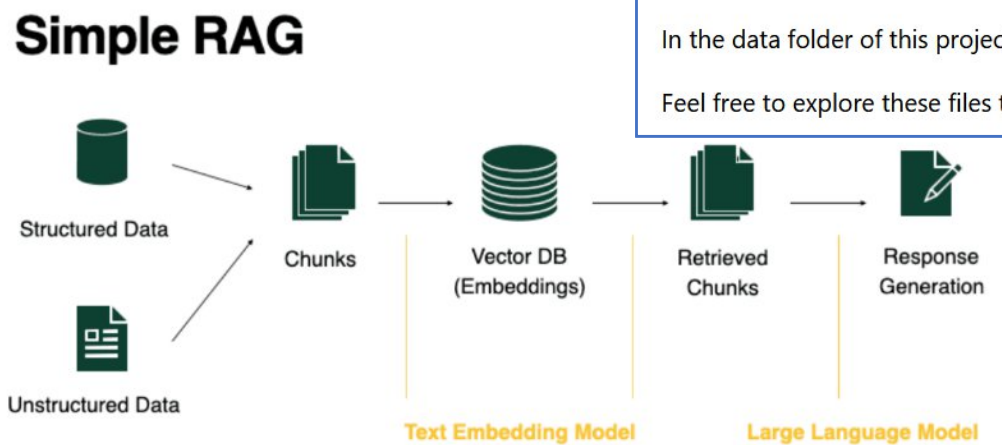
RAG System Building Exercise

Introduction

Hello Students! 👋 Are you ready to dive into the exciting world of Retrieval-Augmented Generation (RAG) systems? This exercise will guide you through constructing your very own RAG system, combining the power of information retrieval with state-of-the-art language models. Let's embark on this learning adventure together! First make sure to download the repository from this link: coling_rag_exercise

What is a RAG System?

A Retrieval-Augmented Generation (RAG) system is a powerful AI architecture that combines the strengths of large language models with external knowledge retrieval. Here's how it works:



RAG System

(Image taken from BentoML)

- **Retrieval:** When given a query, the system searches a knowledge base to find relevant information like documents (here illustrated as VectorDB).
- **Augmentation:** The retrieved information is then used to supplement the input to a language model (here illustrated as the retrieved chunks we pass to the large language model).
- **Generation:** The language model generates a response based on both the original query and the retrieved information (here illustrated at the final stage).

...ion without the need to retrain the entire model, leading to more

Provided Data

In the data folder of this project, you'll find the pdf files of all EPFL legal documents.

Feel free to explore these files to understand the kind of data your system will be working with.

Your Task

Your mission is to complete the missing parts of the RAG system written in `main.py`. While you could fill-in the file directly, we highly recommend following the detailed instructions in each subsection here. Note that there are many ways to implement some of these functions (sometimes there isn't a clear right or wrong), so you can feel free to choose what functionality to include or not.