Introduction to Retrieval Augmented Generation (RAG)

Ph.D. A. Gemelli

UNIVERSITÀ DEGLI STUDI FIRENZE
DINFO
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE
letxbe·ai

Nice to meet you 🤗

- **Master Degree of Computer Science,** UniFi
- **European PhD**, CVC (Barcelona) and UniFi
- **AI Research Scientist** @ LetXbe, Paris

Contacts

email: andrea.gemelli@letxbe.ai

website: https://www.andreagemelli.me/

UNIVERSITÀ DEGLI STUDI FIRENZE | DINFO DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

letxbe·ai

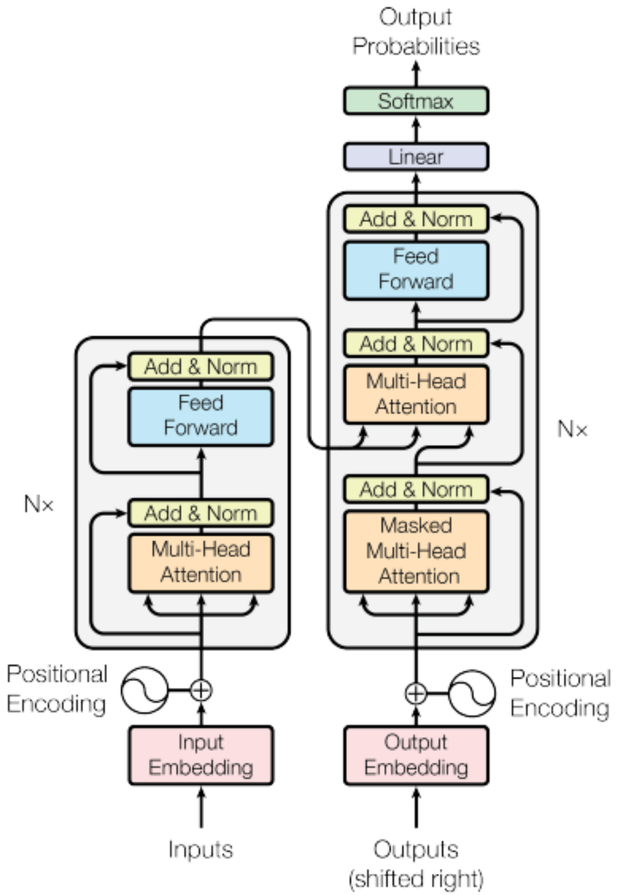
Lecture overview

- Introduction to LLMs and their limitations
- What is RAG?
- Build a RAG system from scratch
- Conclusions

Introduction to LLMs

What are large language models?

- **Transformer architecture** introduced in “Attention is all you need”
- Three type of decoders: **only-Encoder** (e.g. BERT), **only-Decoder** (e.g. GPT) and **Encoder-Decoder** Transformers (e.g. T5)
- The “LLMs” we usually refer nowadays, such as ChatGPT, Llama, Gemini, etc. are only-decoders generative models with **billions** of parameters



Introduction to LLMs

How they work?

I am at university as a teacher today and I am introducing my students the Large Language Models. Can you make a joke about AI and LLMs to help me break the ice?

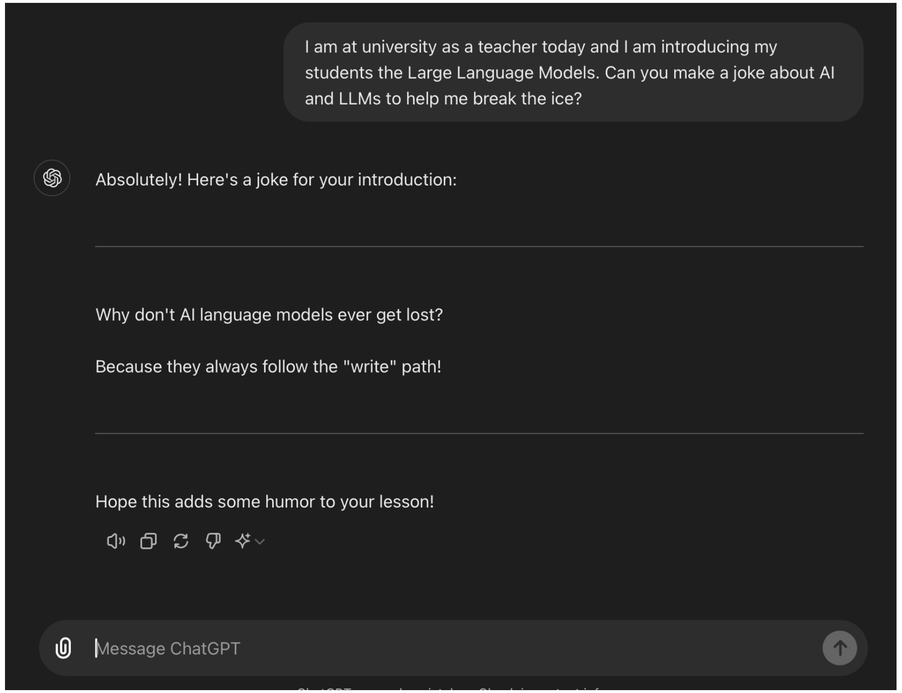
Prompt (Engineering)

Token count
37

I am at university as a teacher today and I am introducing my students the Large Language Models. Can you make a joke about AI and LLMs to help me break the ice?

40, 939, 540, 16490, 472, 261, 14044, 4044, 326, 357, 939, 49659, 922, 4501, 290, 27976, 20333, 50258, 13, 4101, 481, 1520, 261, 41751, 1078, 20837, 326, 451, 19641, 82, 316, 1652, 668, 2338, 290, 14821, 30

Tokenization



Model + **Generation**

Introduction to LLMs

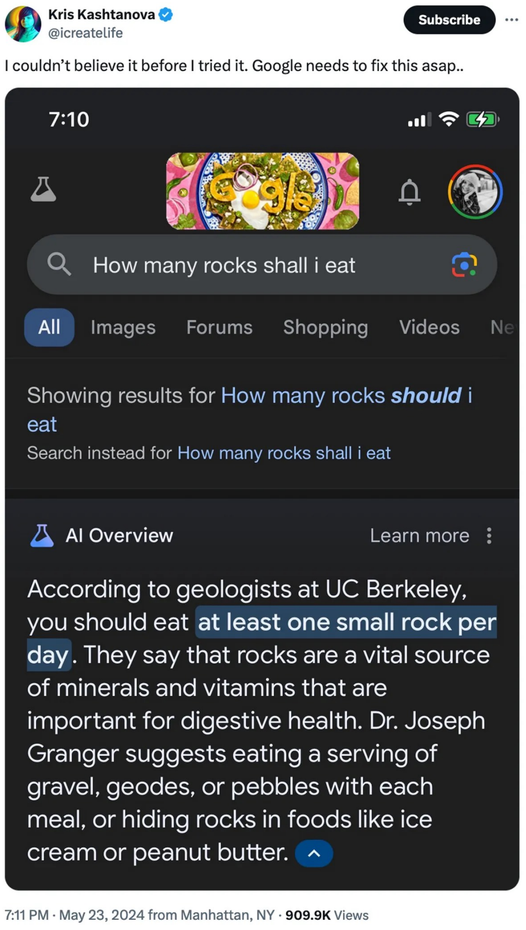
Main limitations

LLMs may have problems of **inconsistency** due to:

- **Outdated information:** trained on data up to X months/years ago
- **Hallucinations:** answering questions with different and unrelated answers seen at "some point" during training
- **Limited knowledge:** data vary in time and things can change

Introduction to LLMs

Main limitations: google suggesting to eat rocks



Introduction to LLMs

Main solutions

To cope with these limitations, among other solutions, the most used and applied are:

- **Supervised Fine Tuning and Alignment:** e.g. RLHF, DPO, PPO etc.
- **Retrieval Augmented Generation**, augmenting the accessible knowledge accessible by the LLM and force it “to stick” with it!

What is RAG?

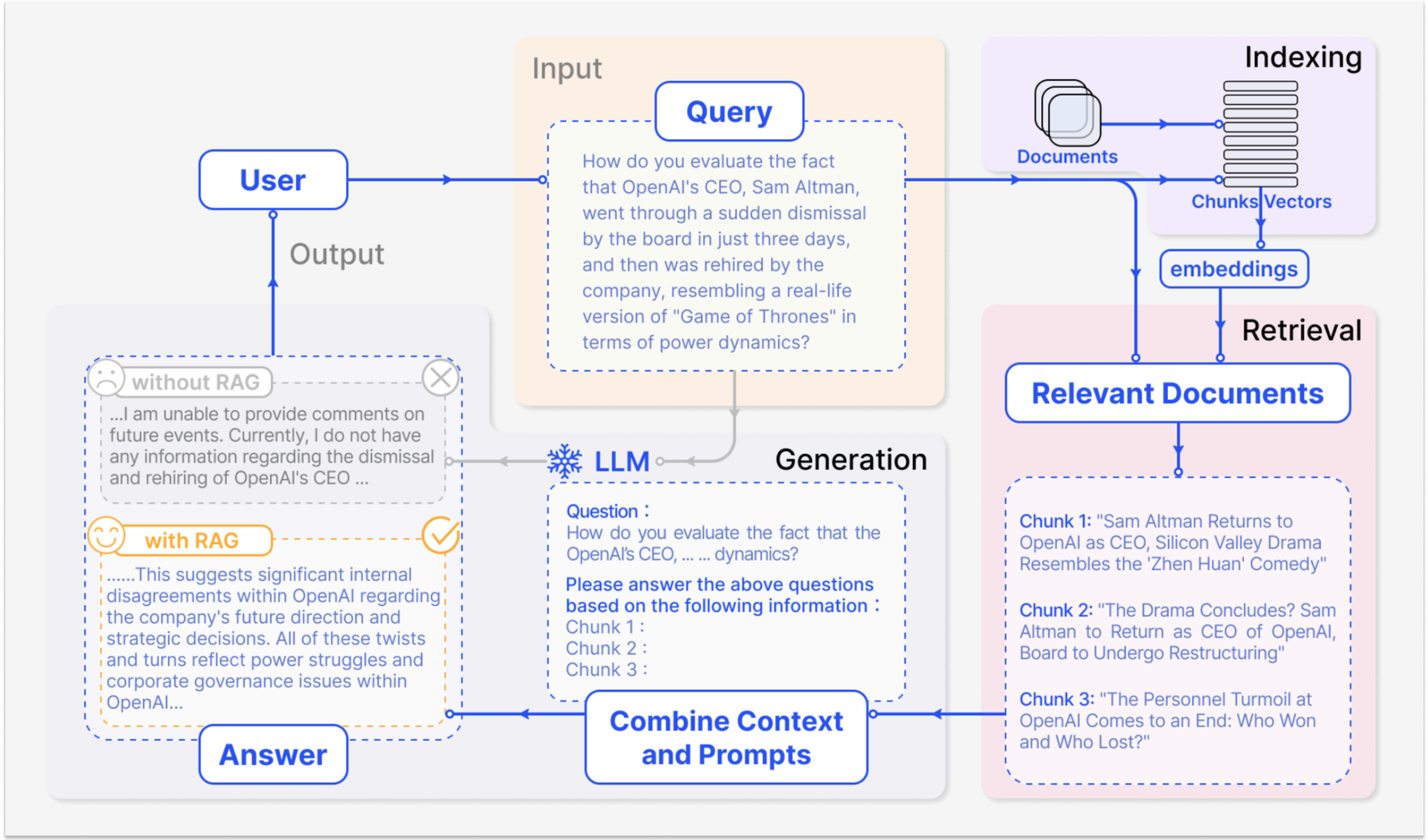
Example: how many moons Jupyter has?

HuggingChat 🤗 example:
https://huggingface.co/chat/conversation/665f64d84f6689afa29b1b60

Wikipedia page:
https://en.wikipedia.org/wiki/Moons_of_Jupiter

What is RAG?

Main components



What is RAG?

Main components

- **Indexing**: extraction o raw data, conversion in full text format and segmentation into smaller parts (called chunks)
  - This part comprehends also the crucial choice of **embeddings** and **vectorDB.** e.g. Sentence Transformers (HF), Mistral
- **Retrieval**: use similarity measures to find the closest matches with the query in the DB
  - e.g. FAISS, Pinecone
- **Generation**: augment the LLM query prompt with the retrieved context, enhancing its capabilities and reducing aforementioned limitations! Plus: *you don't need to "have access" to the LLM in use!*

Rag from scratch!

Let’s code!

- **Colab code: https://colab.research.google.com/drive/1f7CZKe2KM8kD9jSTSDIu5j7tY0N0JbDZ?usp=sharing**

Conclusions

- RAG helps **reducing LLMs known limitations**
- RAG **improves LLMs outputs** without the need to access and fine-tune di final model (tradeoff with fine-tuning and alignment)
- Lots of research “in the middle”: which embeddings to use? What about the chunk choice?

Conclusions

Papers:

- Attention is all you need
- Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks
- Retrieval-Augmented Generation for Large Language Models: A Survey

Technology

Sentence Transformers

HuggingFace

Faiss

Colab link

Blogs

Mistral RAG

Questions?