Introduction to Retrieval Augmented Generation (RAG)Ph.D. A. Gemelli











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

Contacts

email: <u>andrea.gemelli@letxbe.ai</u>

website: <u>https://www.andreagemelli.me/</u>













letxbe-di







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







A. Gemelli



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?

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? I am at university as a teacher today and I am introd ucing my students the Large Language Models. Can you Absolutely! Here's a joke for your introduction: \$ make a joke about AI and LLMs to help me break the ic e? Why don't AI language models ever get lost? Because they always follow the "write" path! 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, 1 Hope this adds some humor to your lesson! 9641, 82, 316, 1652, 668, 2338, 290, 14821, 30 ቀ ይ ይ ቀ Message ChatGPT

Prompt (Engineering)



Introduction to RAG

Tokenization

Model + Generation

A. Gemelli





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- Outdated information: trained on data up to X months/years ago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





Google AI overview suggests adding glue to get cheese to stick to pizza, and it turns out the source is an 11 year old Reddit comment from user F*cksmith 🤮





Introduction to RAG

Follow





I couldn't believe it before I tried it. Google needs to fix this asap..

7:10 穼 💋 囚 \bigcirc Q How many rocks shall i eat All Images Forums Shopping Videos Showing results for How many rocks *should* i eat Search instead for How many rocks shall i eat 👗 Al Overview Learn more According to geologists at UC Berkeley, you should eat at least one small rock per day. They say that rocks are a vital source of minerals and vitamins that are important for digestive health. Dr. Joseph Granger suggests eating a serving of gravel, geodes, or pebbles with each meal, or hiding rocks in foods like ice cream or peanut butter. 🔨

7:11 PM · May 23, 2024 from Manhattan, NY · 909.9K Views





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





A. Gemelli

10

What is RAG?

Main components

- (called chunks)
 - Transformers (HF), Mistral
- **Retrieval**: use similarity measures to find the closest matches with the query in the DB
 - e.g. FAISS, Pinecone



• Indexing: extraction o raw data, conversion in full text format and segmentation into smaller parts

• This part comprehends also the crucial choice of **embeddings** and **vectorDB.** e.g. Sentence

• 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!

A. Gemelli



11

Rag from scratch! Let's code!

 Colab code: <u>https://colab.research.google.com/drive/</u> <u>1f7CZKe2KM8kD9jSTSDIu5j7tY0N0JbDZ?usp=sharing</u>







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:

- <u>Attention is all you need</u>
- <u>Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks</u>
- <u>Retrieval-Augmented Generation for Large Language Models: A Survey</u> **Technology**
- Sentence Transformers
- <u>HuggingFace</u>
- <u>Faiss</u>
- <u>Colab link</u>
- **Blogs** Mistral RAG



<u>wledge-Intensive NLP Tasks</u> <u>ge Language Models: A Survey</u>





Questions?



Introduction to RAG



