# Rag Intro

## Intro to RAG systems

Kristóf Osztopáni
kristof@componentsoft.ai

ComponentSoft

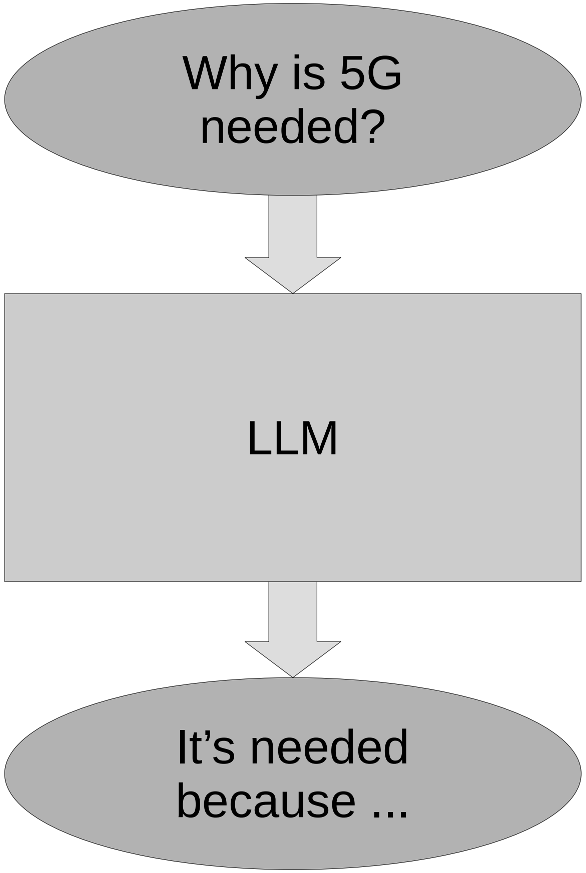
# Introduction

1. Researching LLM based applications for over a year
2. Few years of teaching problem solving
3. Doing competitive programming for the past 6 years

(c) 2024 Component Soft Ltd. - v1.0

# Background

- LLMs (since 2017)
  - **L**arge **L**anguage **M**odel
- Text in Text out
- Some notable ones
  - ChatGPT (gpt4-o)
  - Gemini (Gemini 1.5 Pro)
  - Llama 3.1


Why is 5G needed?

LLM

It’s needed because ...

# Example

Why is 5G needed?

5G is needed to provide faster internet speeds, lower latency, and greater capacity than previous generations of wireless technology. This enables better support for a growing number of connected devices, enhances mobile broadband experiences, facilitates advancements in technologies like the Internet of Things (IoT), smart cities, and autonomous vehicles, and improves overall network reliability for both consumers and businesses.

Is the answer true? Did the LLM hallucinate something? How can we be sure?

(c) 2024 Component Soft Ltd. - v1.0

# Problem

- Out of date information.
  - Trained with H100-80GB (700W TDP) for ~31M GPU Hours, That is 3.5 Years with 1000 GPUs
- Is it accurate? How can we be sure?
  - Read a somewhat filtered version of the Internet
- An algorithm with a Huge amount of data
  - ~ 812 GB
- No access to private data

The examples are about the training of Llama 3.1 405B, that is comparable in some benchmarks with gpt-4o (ChatGPT)

# What does an LLM “know”?

- What it learned during training.
  - Wikipedia, White Papers, etc…
  - Reddit, StackOverflow, Github public code, etc…
- It knows how to reason.
  - Can deduce non-trivial facts
- It knows its input text / question / prompt.

# Did it learn that it will have a knowledge cutoff date?

| Expected | Got |
|---|---|
| Who won yesterday's match? | Who won yesterday's match? |
| I'm not aware of any specific match you're referring to. Could you please provide more context or information about the match you're asking about, such as the teams or tournament involved? | I'm sorry, but I don't have access to real-time information or specific knowledge about yesterday's match. Could you please provide more context or details about the match you're referring to (e.g., which sport, league, or teams)? I'll do my best to help you find the answer! |

**Question**: What happened?

(c) 2024 Component Soft Ltd. - v1.0

# Context

- Use the reasoning ability of the LLM
  - For any incoming request, let's prepend a block with useful info
- What else could we inject into our prompt?
  - Lets solve our original problems with a similar approach

Today is 2024. November 7, you are a helpful assistant, that can answer questions which requires information before your cutoff date, which is 2023. December.

Who won yesterday's match?

I'm sorry, but I don't have any information on live events or real-time data beyond my cutoff date in 2023. Therefore, I cannot provide details about any events or matches that took place after that date.

# How could we solve these problems?

- Problems
  - Out of date information.
  - Is it accurate?
  - How can we be sure of an answer?
  - No access to private data

# The original problems

- Problems
  - Out of date information.
    - Solution: Inject new information
  - Is it accurate?
    - Solution: Inject relevant information to the question
  - How can we be sure?
    - Solution: Reference existing documents
  - No access to private data
    - Solution: Inject private data

# Context length

- The context length is the length full history of a conversation
- LLMs have a limited context length
  - ~ 128 000 Words for Llama 3.1 405B
- LLMs have limited attention
  - If we fill up the context, then it may not find an answer that is in the provided context
- We can’t just inject everything

(c) 2024 Component Soft Ltd. - v1.0

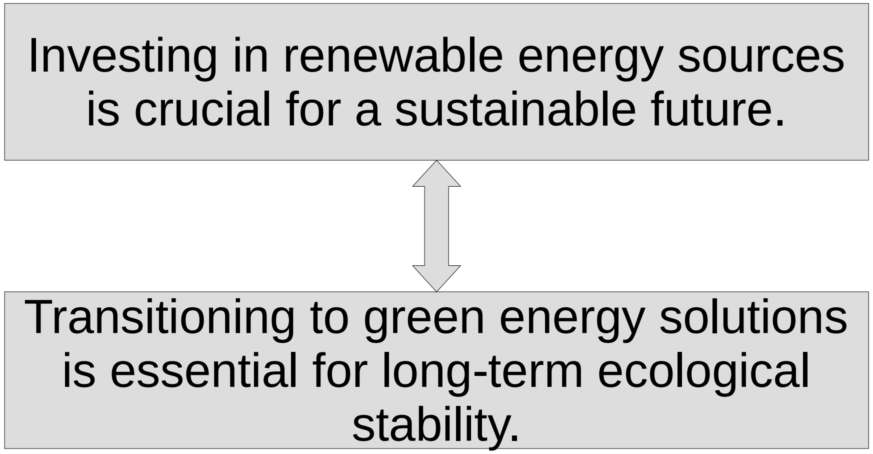
# Relevance

- We should only inject the **most relevant** information that we know about.
- What is relevance?
  - Two sentences are relevant to each other if
    - They are Similarly worded: Using mostly the same words
    - They mean the same or they are talking about the same concept: Semantic Similarity

# Semantic Similarity

They are similar because

- About green energy
- About the future
- About being important
- Not about software
- Not about school



Investing in renewable energy sources is crucial for a sustainable future.

Transitioning to green energy solutions is essential for long-term ecological stability.

(c) 2024 Component Soft Ltd. - v1.0

# Semantic Similarity

Investing in renewable energy sources is crucial for a sustainable future.

With just one sentence

- (0.8) About green energy
- (0.7) About the future
- (0.6) About being important
- (0.2) Not about software
- (0.0) Not about school

# Semantic Similarity

- **Question:** If we can score a sentence, with how much is it about green energy, then when is it similar to our original sentence
  - Our sentence, had a 0.8 score for being about green energy

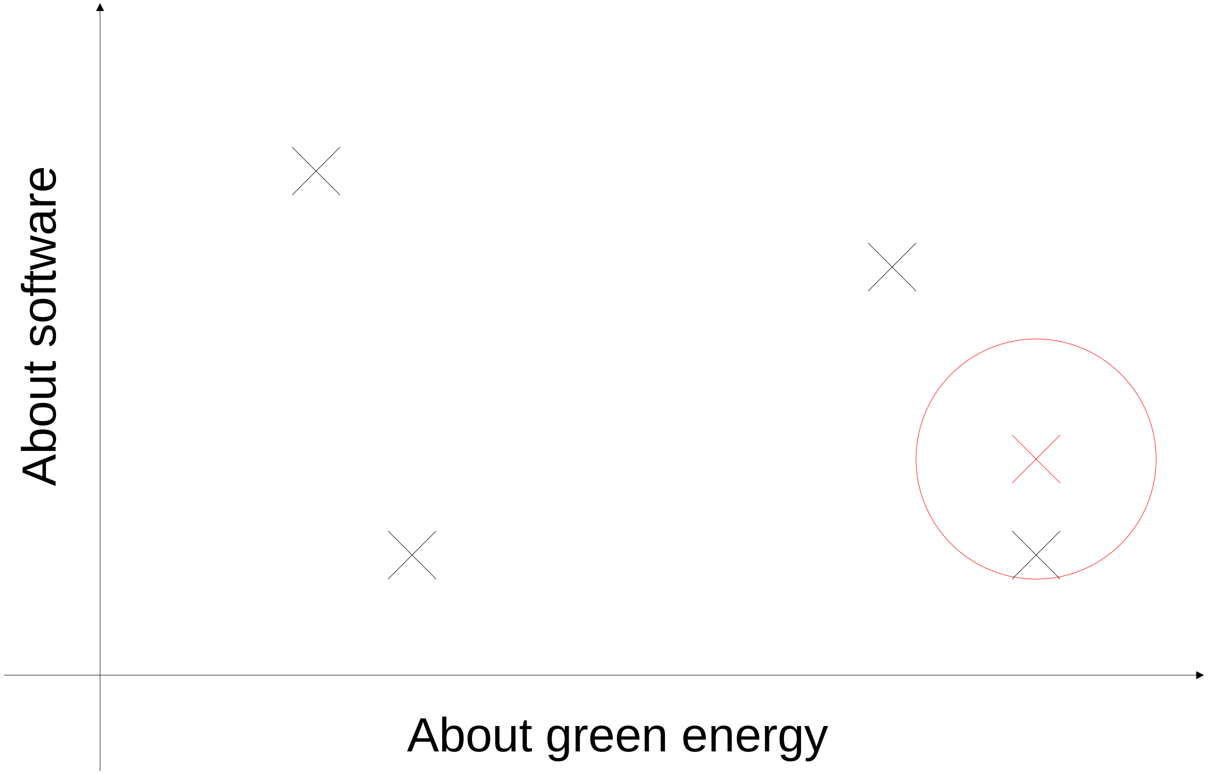
# Embeddings

- For a given set of topics, we can score a sentence for every topic.

Here close sentences are similar

In practice we have more than 2 topics, we have thousands

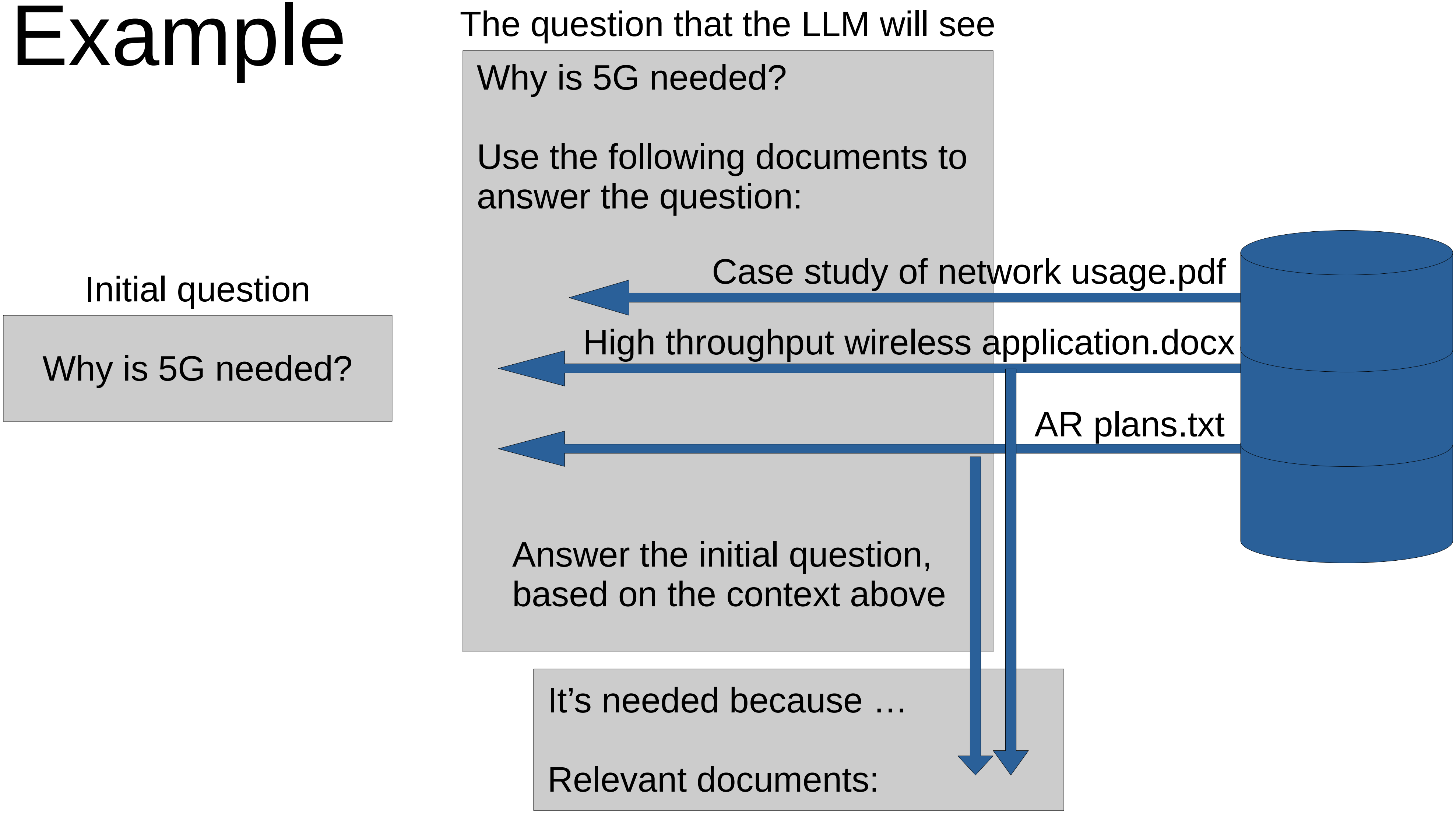
If we do this scoring, for every piece of information we have, and also for the question, then the most relevant information for the question will be the closest points to the question



# Problems

- Out of date information.
  - Solution: Inject new relevant information
- Is it accurate?
  - Solution: Inject relevant information to the question
- How can we be sure?
  - Solution: We can add a reference to the closest documents
- No access to private data
  - Solution: Inject relevant private data

# Example



The question that the LLM will see

Why is 5G needed?

Use the following documents to answer the question:

Case study of network usage.pdf

High throughput wireless application.docx

AR plans.txt

Answer the initial question, based on the context above

Initial question

Why is 5G needed?

It's needed because …

Relevant documents:

(c) 2024 Component Soft Ltd. - v1.0

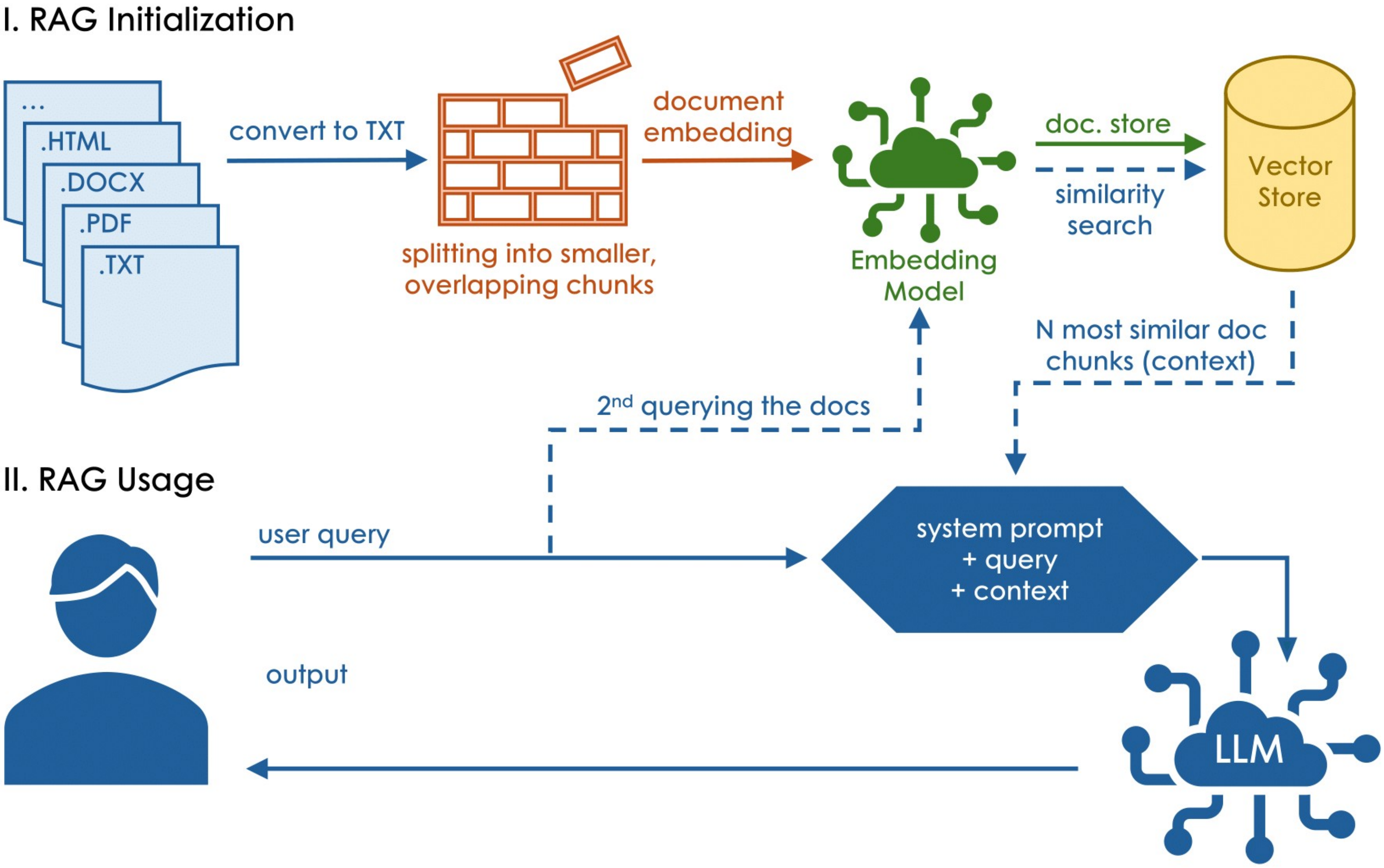
# “Upload” documents

- Individual documents are still too big
    - Split them into smaller chunks
    - **Question:** how should we split?
        - Split every 2000 words or 100000 characters

# “Upload” documents

- Individual documents are still too big
  - Split them into smaller chunks
  - How should we split?
    - Split every 2000 words or 100000 characters
    - By Document structure (Pages, Headers, etc...)
    - By Semantic groups
      - Somehow group regions of text by topics
    - By a combination of the above

# RAG Architecture



# Problems and enhancements

- What models to choose, for communication, and for embedding?
- How to split the documents?
  - Semantic splitting, Hierarchical chunking
- How to find the relevant documents?
  - Advanced search methods (Reranking, HyDE, etc..)

(c) 2024 Component Soft Ltd. - v1.0