

# Why RAG?

## Instructors

Prashant Sahu

Manager - Data Science, Analytics Vidhya

Ravi Theja

Developer Advocate Engineer, LlamalIndex



# Context

- LLMs are a phenomenal piece of technology for knowledge generation and reasoning.
- They are pre-trained on large amounts of publicly available data.



# What can we do with LLMs?



# Blog Writing

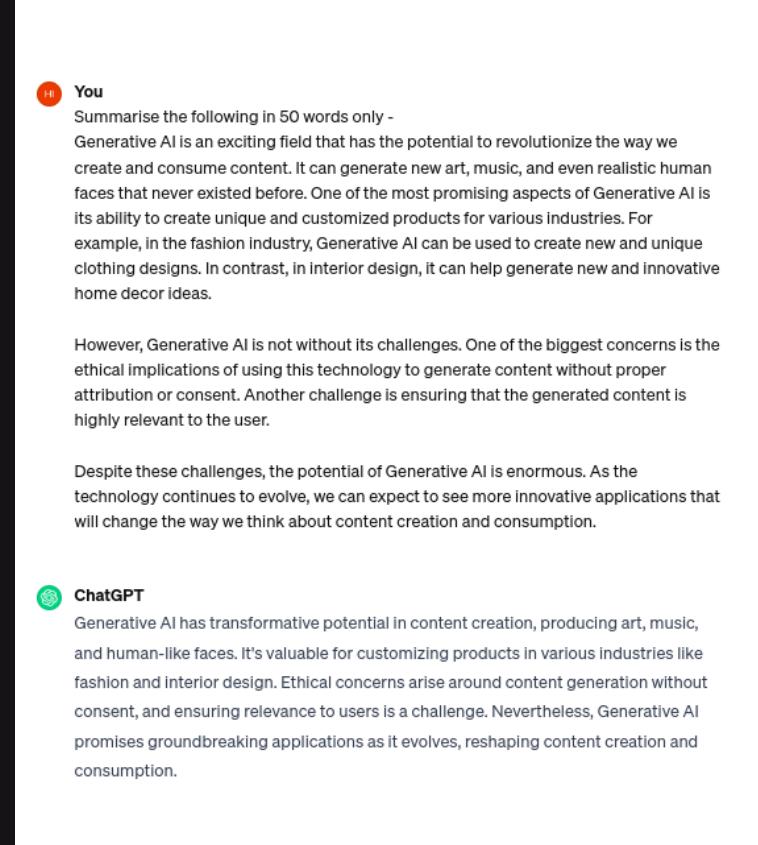
**You**  
Summarise the following in 50 words only -  
Generative AI is an exciting field that has the potential to revolutionize the way we create and consume content. It can generate new art, music, and even realistic human faces that never existed before. One of the most promising aspects of Generative AI is its ability to create unique and customized products for various industries. For example, in the fashion industry, Generative AI can be used to create new and unique clothing designs. In contrast, in interior design, it can help generate new and innovative home decor ideas.

However, Generative AI is not without its challenges. One of the biggest concerns is the ethical implications of using this technology to generate content without proper attribution or consent. Another challenge is ensuring that the generated content is highly relevant to the user.

Despite these challenges, the potential of Generative AI is enormous. As the technology continues to evolve, we can expect to see more innovative applications that will change the way we think about content creation and consumption.

 ChatGPT

Generative AI has transformative potential in content creation, producing art, music, and human-like faces. It's valuable for customizing products in various industries like fashion and interior design. Ethical concerns arise around content generation without consent, and ensuring relevance to users is a challenge. Nevertheless, Generative AI promises groundbreaking applications as it evolves, reshaping content creation and consumption.



# Summarization

# Code Assistance

# What can we do with LLMs?

A Generate a list of idea for writing a technical blog around Gen AI

View other drafts

Technical Blog Ideas for Gen AI:

Foundational Concepts:

- Demystifying Generative AI: An Introduction to the Core Concepts
- Unpacking the Hype: What is Gen AI Really Capable Of?
- The Building Blocks of Gen AI: Exploring Neural Networks and Machine Learning
- Deep Dive into Text Generation: How AI Can Craft Compelling Content
- Understanding the Ethics of Gen AI: Responsible Development and Deployment

Idea Generation

You who won the icc world cup 1983

ChatGPT

The ICC (International Cricket Council) Cricket World Cup in 1983 was won by the Indian cricket team. India defeated the West Indies in the final match to secure their first-ever Cricket World Cup victory. The final took place on June 25, 1983, at Lord's Cricket Ground in London, England. India's victory in the 1983 World Cup is considered a historic moment in Indian cricket history.

Question Answering

# Challenges with LLMs

The image shows a side-by-side comparison of two LLM responses to the same query. On the left, under the heading "Recent Updates", is a screenshot of a ChatGPT interface. A user (AP) asks, "Who won the ICC World Cup 2023?". ChatGPT responds, "As of my last update in January 2022, I don't have information on events that occurred after that date, including the ICC World Cup 2023. You might want to check the latest news or sports updates for the most current information." Below the message are five small interaction icons. On the right, under the heading "Enterprise Data", is another screenshot of a similar interface. A user (AP) asks, "Give me the list of recent features released by llaindex". ChatGPT responds, "I'm sorry, but I can't provide real-time or updated information about specific companies or their recent releases. As of my last update in January 2022, LlamalIndex wasn't a recognized platform or service. If it's a new service that emerged after that time, I wouldn't have information on it. To find the latest features released by LlamalIndex or any other platform, I recommend checking their official website, blog, or social media channels for announcements and updates." Below this response are also five small interaction icons.

Recent Updates

Enterprise Data

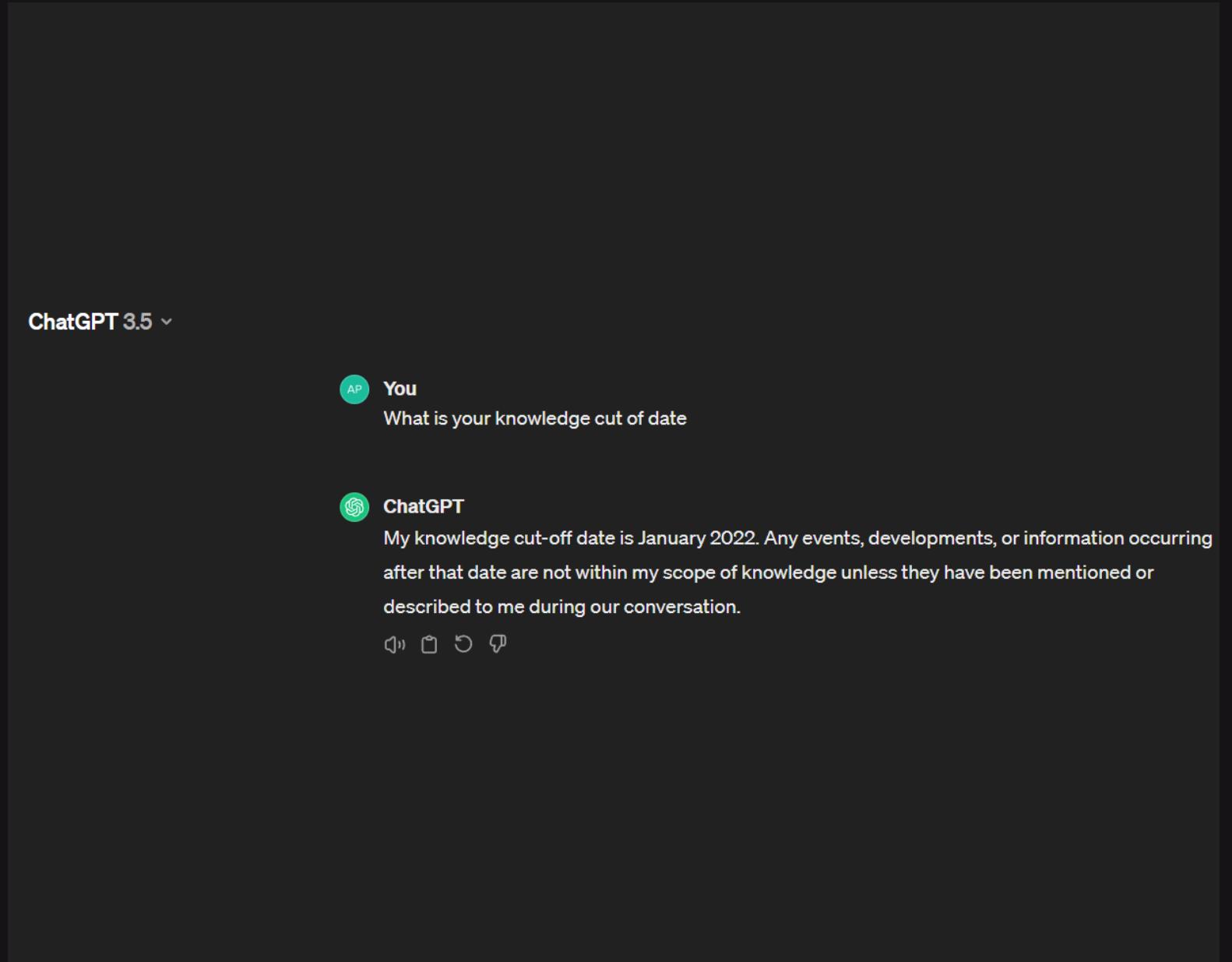
The knowledge cut-off  
date of ChatGPT(3.5)  
is **January 2022**

ChatGPT 3.5 ▾

You AP What is your knowledge cut off date

ChatGPT My knowledge cut-off date is January 2022. Any events, developments, or information occurring after that date are not within my scope of knowledge unless they have been mentioned or described to me during our conversation.

🔉 🗃 ⏴ ⏵



# Limitations of LLMs

- Factuality errors
- Limited context
- Difficulties with evolving information

How do we best augment LLMs  
with our own private data?



# Paradigms for Inserting Knowledge into LLMs

✓ Fine Tuning

✓ In Context  
Learning

✓ RAG

# Paradigms for Inserting Knowledge into LLMs

✓ Fine Tuning

# Fine Tuning

Fine-tuning is a process that takes a model that has already been trained and then tunes or tweaks the model to make it perform a similar task.

## Limitations OF Fine Tuning

- High Computation
- Data Preparation Effort
- High Cost
- Need ML expertise

# Paradigms for Inserting Knowledge into LLMs

- ✓ In Context Learning

# In-Context Learning - DHSGPT

Fix the model, put context into the prompt

DHS stands for DataHack Summit. DataHack Summit (DHS or dhs) 2023 is India's most futuristic AI and data science conference, bringing together the best and brightest in the field of AI to share their knowledge, experiences, and passion for AI. The theme of the conference is Infinite Possibilities: Exploring the future with Generative AI. DHS 2023 totally has 70+ AI Talks, 30+ Hack Sessions and 8+ Workshops. It's the 4th edition of DHS.

What is the theme of DHS?



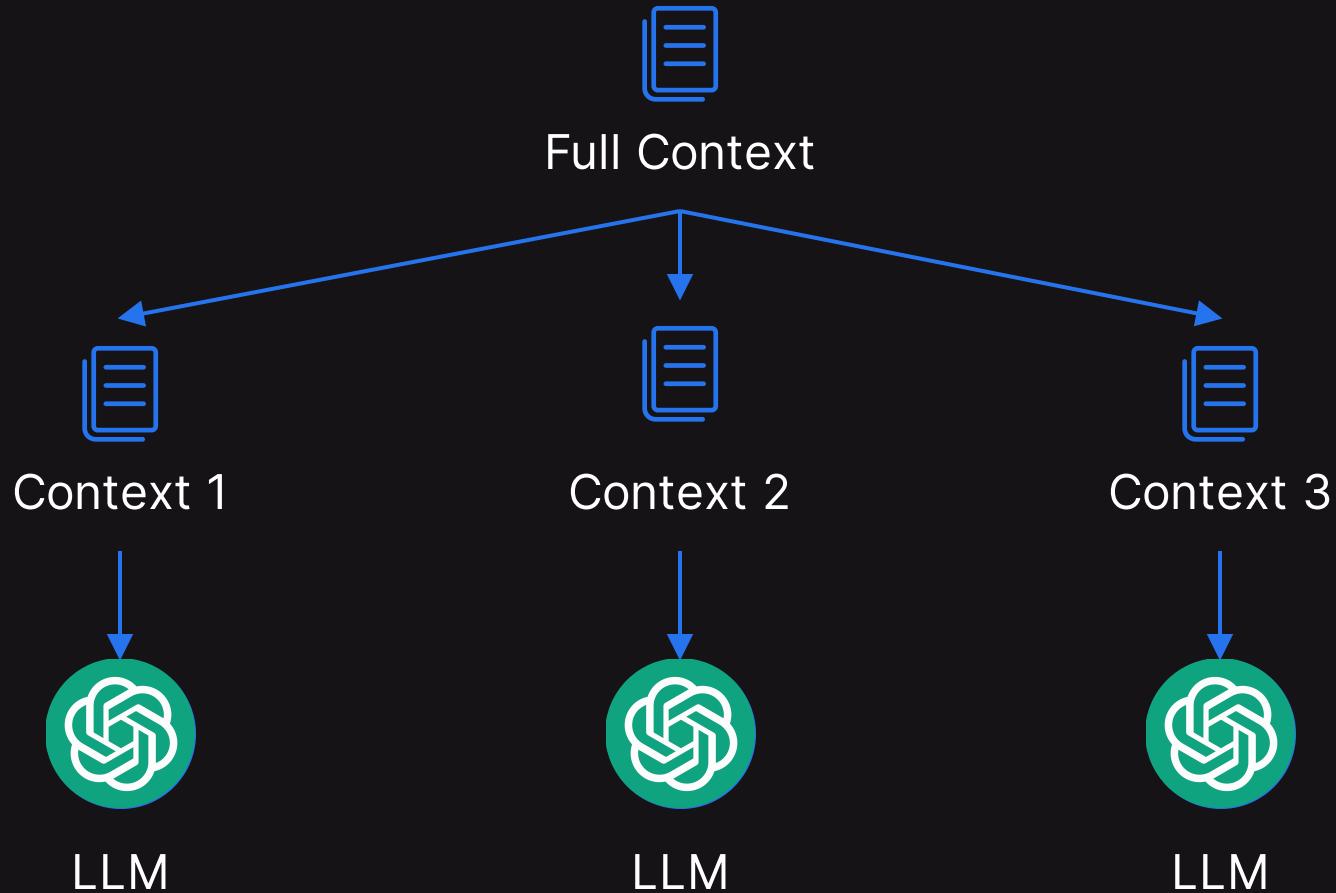
Infinite possibilities:  
Exploring the future  
with Generative AI

# In-Context Learning - DHSGPT

Error on the maximum context length

“This model's maximum context length is 4097 tokens. However, your messages resulted in 5060 tokens. Please reduce the length of the messages.”

# Dealing with Long Context Data



# Challenges with In-Context Learning

- How to retrieve the right context for the query?
- How to deal with source data that is potentially very large (GBs, TBs)?
- Performance vs Latency vs Cost

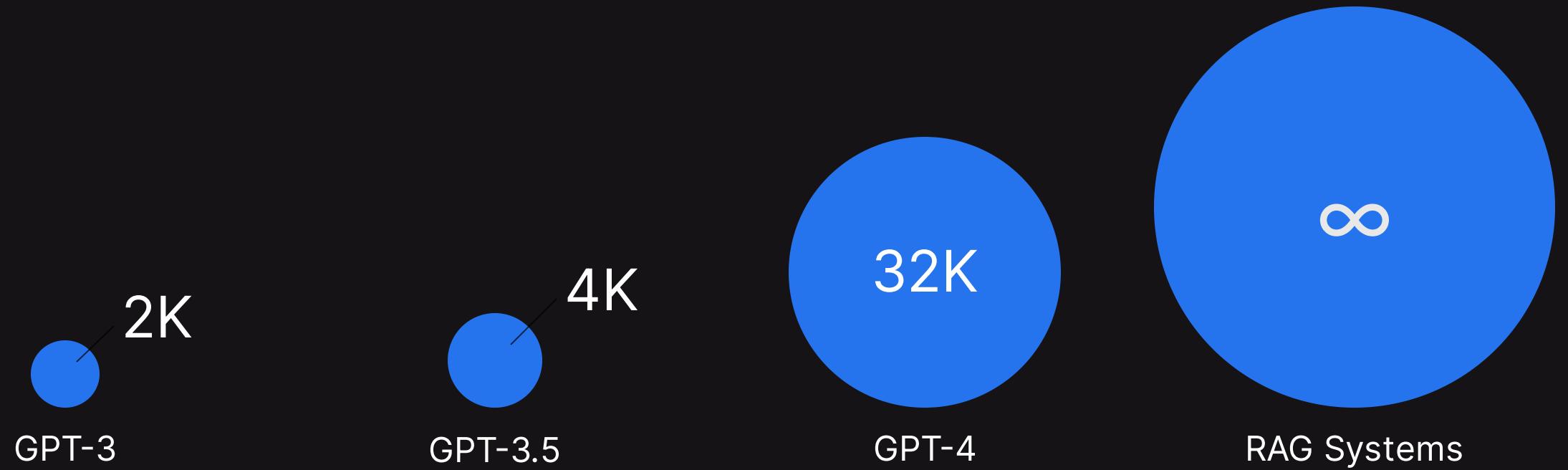
# Paradigms for Inserting Knowledge into LLMs



RAG

# Retrieval Augmented Generation

# Why RAG?



# Why RAG?

- Augmenting the LLM with external knowledge sources
- Improving context-awareness
- Facilitating easier knowledge updates

# Thank You

---