RAGtime in the Big Apple: Chat with a Decade of NYR Speakers



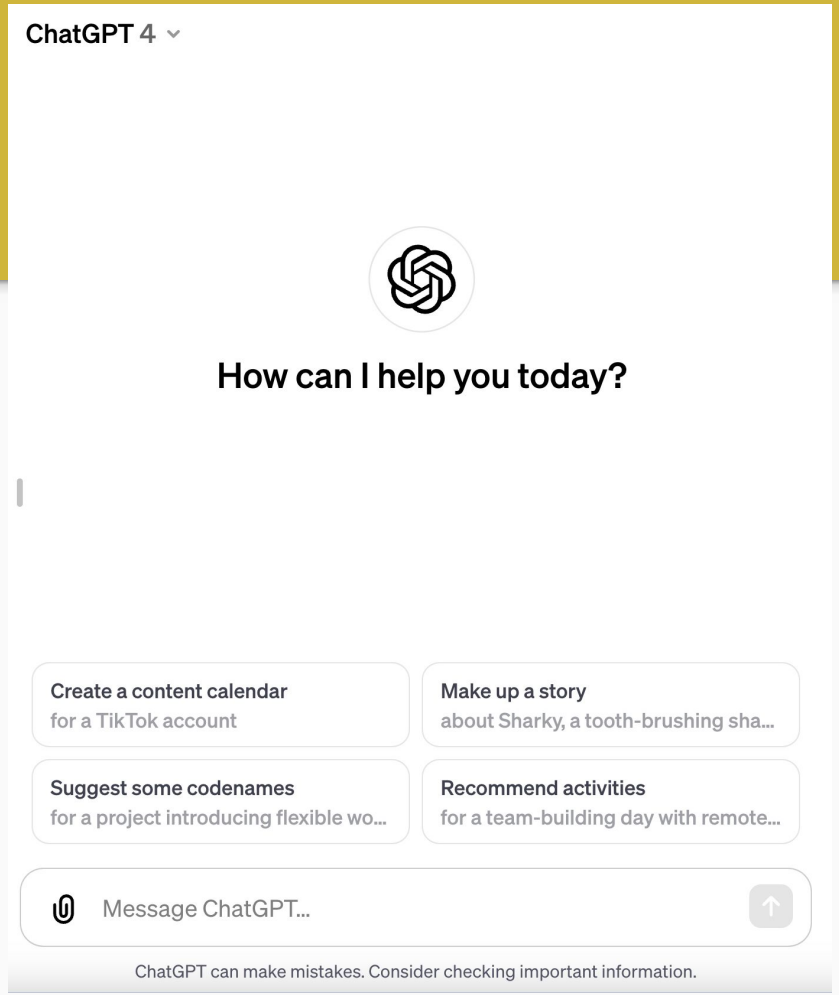
Alan Feder

Who am I?

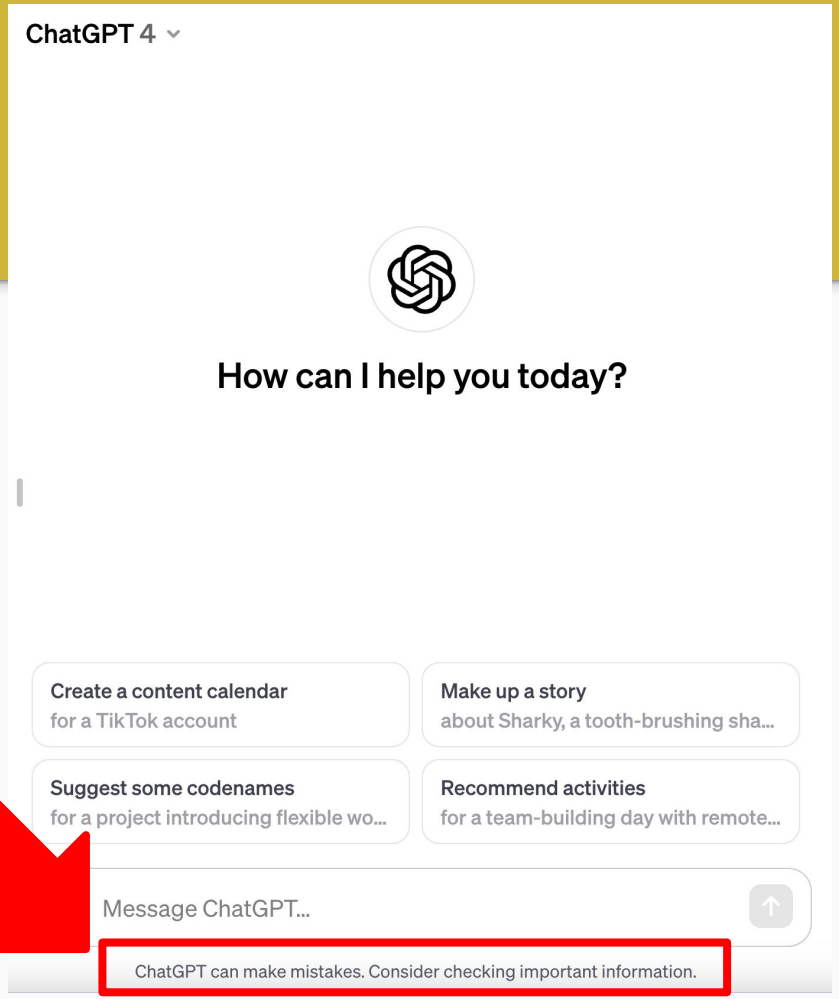
- Alan Feder
- Data Scientist / Machine Learning Engineer
- Freelance Consultant
- NYC Expat
- Excited by ML / Statistics / NLP



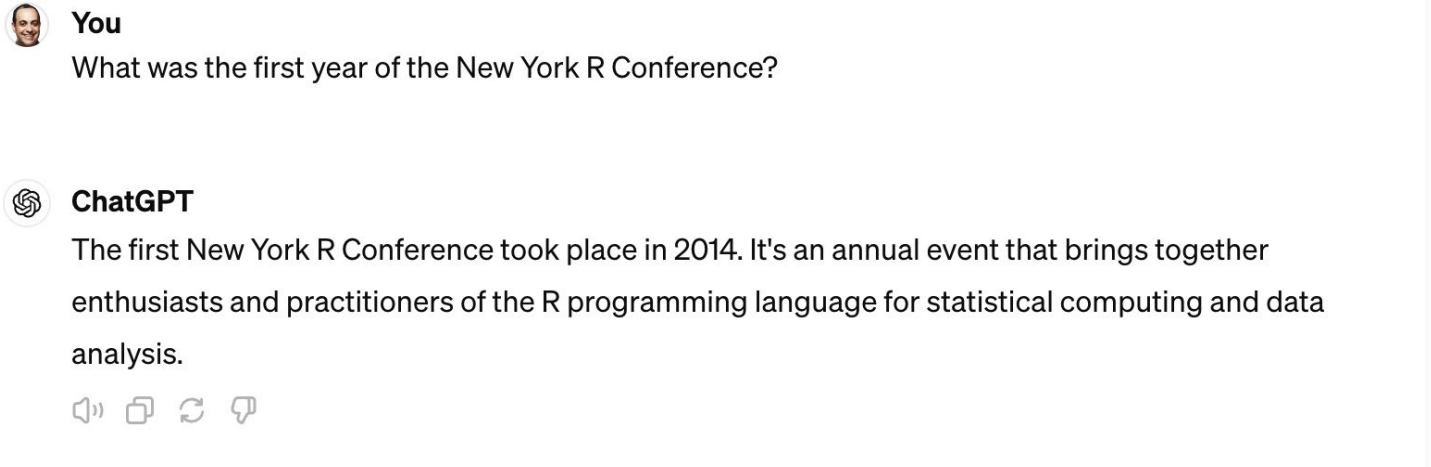
ChatGPT


ChatGPT 4
How can I help you today?
Create a content calendar
for a TikTok account
Make up a story
about Sharky, a tooth-brushing sha...
Suggest some codenames
for a project introducing flexible wo...
Recommend activities
for a team-building day with remote...
Message ChatGPT...
ChatGPT can make mistakes. Consider checking important information.

ChatGPT


ChatGPT 4
How can I help you today?
Create a content calendar
for a TikTok account
Make up a story
about Sharky, a tooth-brushing sha...
Suggest some codenames
for a project introducing flexible wo...
Recommend activities
for a team-building day with remote...
Message ChatGPT...
ChatGPT can make mistakes. Consider checking important information.

ChatGPT “hallucinates”



ChatGPT “hallucinates”



ChatGPT sometimes doesn’t know

AO How do they use R at AirBNB?

Airbnb, like many tech companies, uses R for various data science and analytics tasks. While I don't have insider knowledge of their specific practices, here are some common ways R is typically used at companies like Airbnb:

1. Data Analysis: R is widely used for exploratory data analysis, allowing data scientists to clean, manipulate, and visualize large datasets to uncover insights and patterns.
2. Statistical Modeling: R provides a wide range of statistical modeling techniques, such as linear regression, logistic regression, and time series analysis. These can be used to predict user behavior, forecast demand, and optimize pricing.
3. Machine Learning: R has extensive libraries for machine learning, like caret and mlr, which enable data scientists to build and deploy predictive models. This could be used for tasks like recommending listings to users or detecting fraudulent activity.
4. Data Visualization: R has powerful data visualization capabilities with libraries like ggplot2 and plotly. These can be used to create interactive dashboards and

What is a RAG?

- ***Retrieves*** a list of documents to identify which ones are most relevant
- The LLM ***generates*** a final answer using only the items found above

- The final output can link back to the document(s) found in the first step for verification

How could you use a RAG?

- Tech Support
- Benefits Policies
- Internal Research Documents

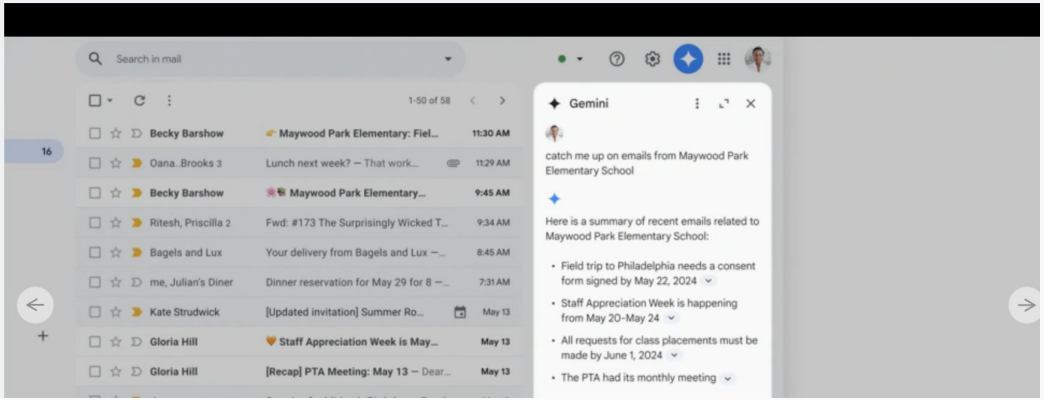
How could you use a RAG?

ACTUALLY USEFUL AI —

AI in Gmail will sift through emails, provide search summaries, send emails

Gmail will soon be able to summarize recent emails from a contact.

RON AMADEO - UPDATED 5/14/2024, 1:44 PM



Context Windows

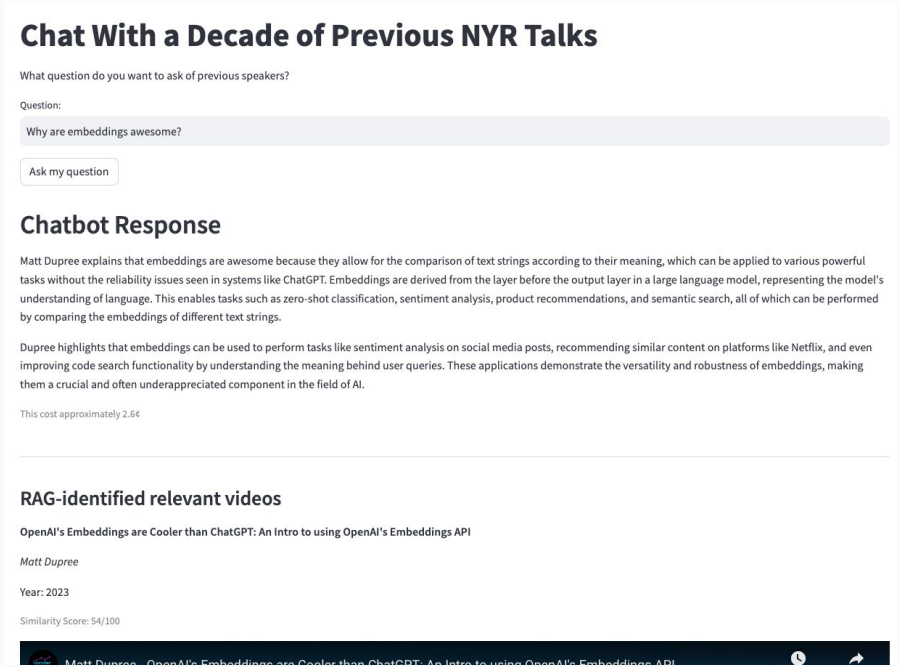
- LLMs can only remember a limited number of tokens

	Training Data	Params	Context length
Llama 3	A new mix of publicly available online data.	8B	8k
		70B	8k

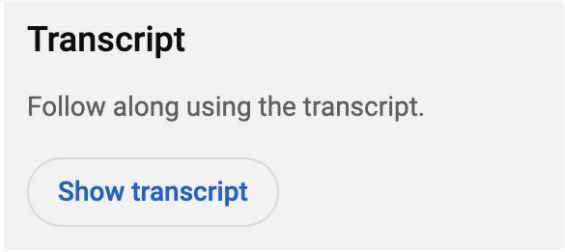
MODEL	DESCRIPTION	CONTEXT WINDOW
gpt-4o	New **GPT-4o** Our most advanced, multimodal flagship model that's cheaper and faster than GPT-4 Turbo. Currently points to gpt-4o-2024-05-13.	128,000 tokens

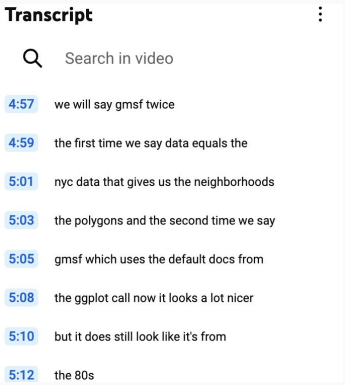
Embeddings

- LLMs and all other modern ML NLP models work on pre-trained embeddings (vectors)



My step 0 - YouTube transcription











Whisper

Split & Store your document

Split document

- This is one of the hardest parts
- Smaller chunks are more specific, but less context
- Overlaps are often good
- Use-case dependent

My method

- Video Embedding based on abstract
- Within videos: break transcript up into chunks of 40 seconds, with a 5 second overlap between each chunk

Embed your chunk

- Any embedding model will do
- OpenAI: `text-embedding-3-small`, `text-embedding-3-large`
- SBERT: Sentence Transformers
- MTEB Leaderboard

My method

- OpenAI embedding

Embed your question

- MUST BE SAME EMBEDDING MODEL!

Cosine similarity

- Run the cosine similarity between your question embedding and *each of the other document/chunk embeddings*
    - Since the embedding represents the topic of the text, those texts that are about the same topic as the question will have the highest similarities
- Vector databases speed this up

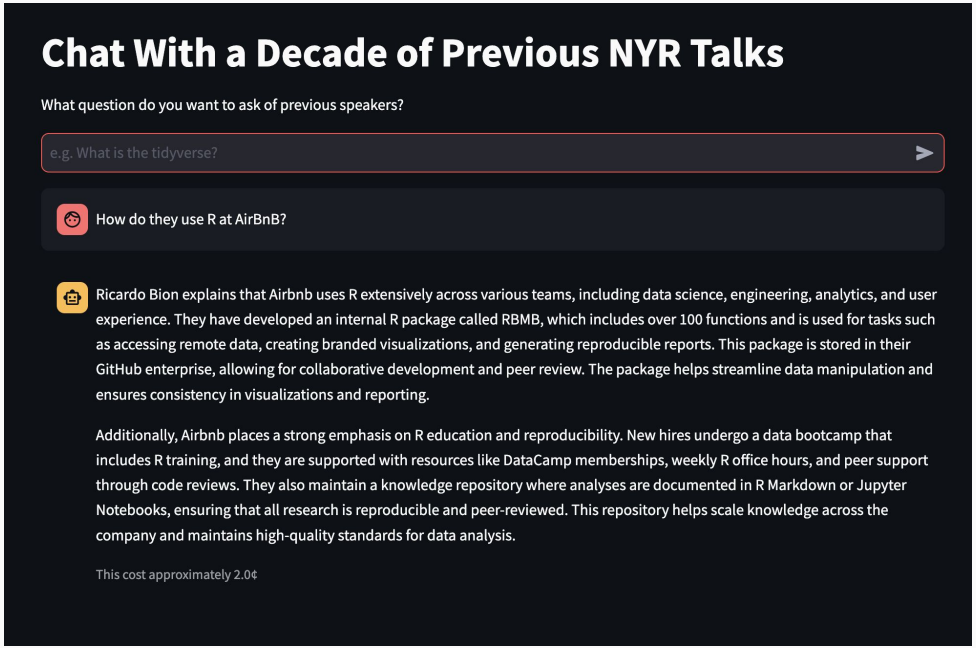
Retrieval

- **Retrieve** the chunks that are most similar to the question
- How many chunks to retrieve?
    - ¯_(ツ)_/¯
    - I did 1-4 videos, and then looked inside the videos

System Prompt

```
You are an AI assistant that helps answer questions by searching through video
transcripts. I have retrieved the transcripts most likely to answer the user's
question.  Carefully read through the transcripts to find information that
helps answer the question.
Be brief - your response should not be more than two paragraphs.
Only use information directly stated in the provided transcripts to answer the
question. Do not add any information or make any claims that are not explicitly
supported by the transcripts. If the transcripts do not contain enough
information to answer the question, state that you do not have enough
information to provide a complete answer. If only one of the transcripts
answers the question, don't reference the other and don't explain why its
content is irrelevant.
Format the response clearly.
```

Generate the output



Show the proof

- You could link back to the source, or even display the text

---

- I show the video starting from the minute with the highest cosine similarity

DEMO

Improvements to make

- Smarter chunking
- More complex retrieval
- Prompt engineering
- Chat - back and forth
- Better embeddings
- Better models

Alan Feder



AlanFeder@gmail.com
https://www.alanfeder.com/
alanfeder
AlanFeder

RAG is a great way to allow ChatGPT to answer questions about your documents!

App: https://nyr.alanfeder.com/
Code: https://github.com/AlanFeder/nyr-rag-app/
Slides: https://bit.ly/nyr-rag