



Measuring Hallucinations In RAG

Conf42 Large Language Models (LLMs)
April 2024

About Me



Head of Developer Relations @ Vectara

Previously: Syntegra, Helix, LendUp, Hortonworks, Yahoo!

My first LLM was **GPT-2**



The LLM revolution

In 5 years, every application we use (enterprise and/or consumer), every piece of knowledge we acquire, and even every device, will have a **GenAI interface** enabling us to **easily interact** with it.



Use cases with LLMs...

- Customer support chatbot
- Question answering applications
- Product recommendations
- Semantic search
- Workplace search
- And many others...



But... LLMs hallucinate

Prompt: did Will Smith ever hit anyone?

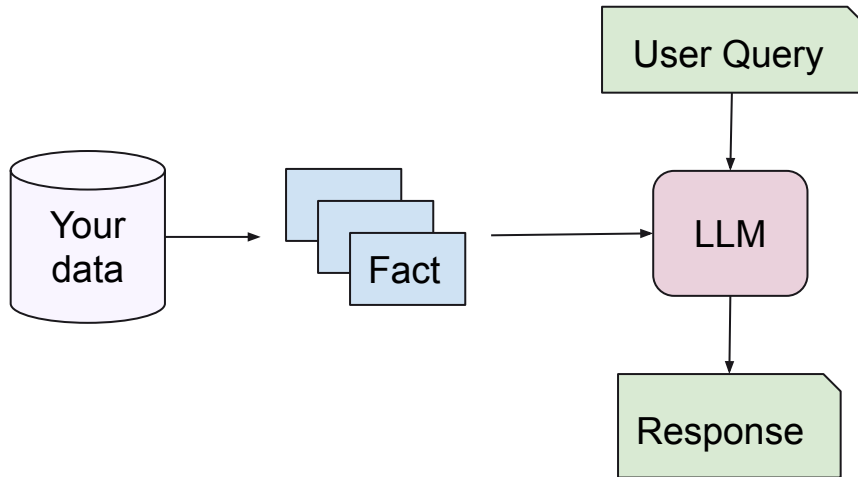
ChatGPT:

There is no public record or credible evidence to suggest that Will Smith has ever physically assaulted or hit anyone. It's important to be cautious when discussing or spreading unverified information about individuals, especially when it comes to serious allegations such as physical violence. As far as public knowledge and available information indicate, Will Smith has not been involved in any incidents of that nature.



Addressing Hallucinations with RAG

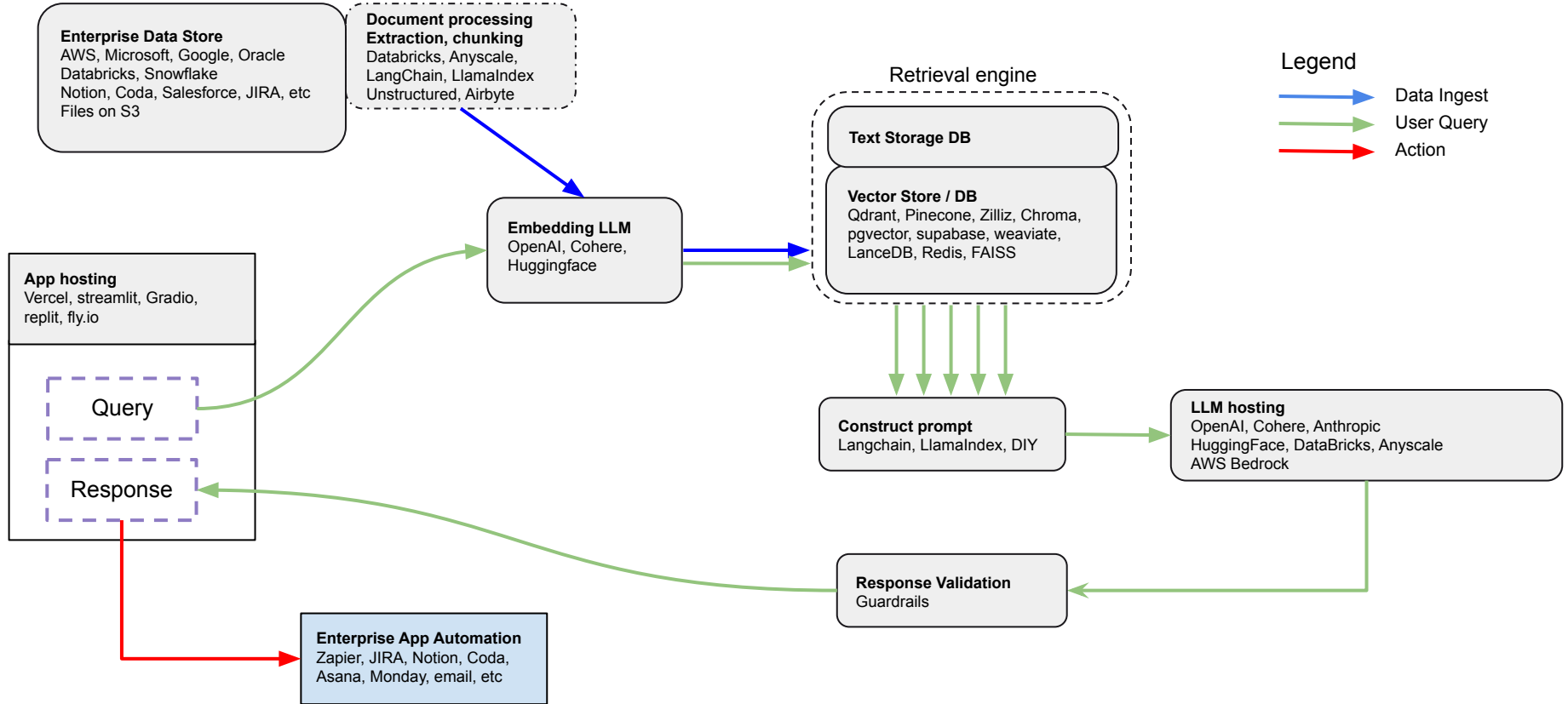
Using **retrieval augmented generation** helps the LLM focus on facts that matter and provide more accurate responses.



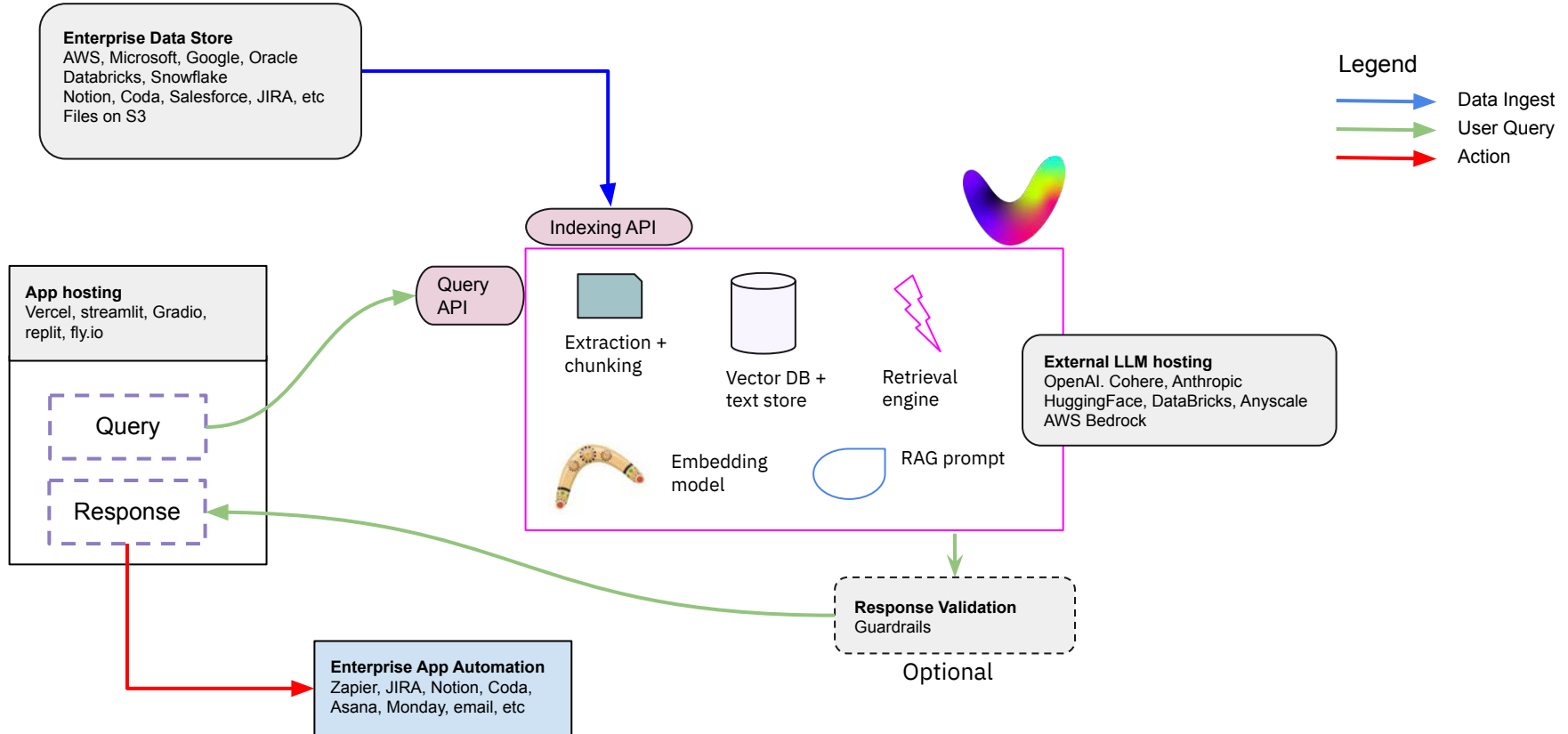
- LLM with “your data”
- Helps reduce hallucinations
- Use-cases
 - Question answering
 - Chatbot / Co-pilot



RAG: Do-It-Yourself Approach



RAG-as-a-Service with Vectara



Why Retrieval Augmented Generation?

- Augment LLM with **your own data** to **reduce hallucinations**.
- RAG outputs are **explainable** with citations, increasing user trust.
- **Private:** your data is not used for training the LLM.
- **Inexpensive** to implement and use, since no re-training or human feedback collection is required.
- Enables **per-person access control** so that generated output complies with regulations and data governance



Why Vectara?

- Building RAG is more complex than it seems:
 - Good retrieval is more than vector similarity
 - Cross-lingual retrieval is complex
 - DevOps, MLOps, promptOps: requires significant expertise
 - Chat-with-my-PDF (MVP) vs enterprise ready RAG
 - Handling citations
- Security, data privacy, permissioning, data governance
- Total cost of ownership

Vectara's RAG-as-a-service provides an end-to-end solution that just works



HHEM: Hallucination Evaluation Model

Model	Hallucination Rate	Factual Consistency Rate	Answer Rate	Average Summary Length (Words)
Intel Neural Chat 7B	2.8 %	97.2 %	89.5 %	57.6
GPT 4	3.0 %	97.0 %	100.0 %	81.1
GPT 4 Turbo	3.0 %	97.0 %	100.0 %	94.3
Microsoft Orca-2-13b	3.2 %	96.8 %	100.0 %	66.2
GPT 3.5 Turbo	3.5 %	96.5 %	99.6 %	84.1
Cohere Command R	3.9 %	96.1 %	99.9 %	51.2
Google Gemini Pro	4.8 %	95.2 %	98.4 %	89.5
Llama 2 70B	5.1 %	94.9 %	99.9 %	84.9
Llama 2 7B	5.6 %	94.4 %	99.6 %	119.9
Llama 2 13B	5.9 %	94.1 %	99.8 %	82.1
Anthropic Claude 3 Sonnet	6.0 %	94.0 %	100.0 %	108.5
Anthropic Claude 3 Opus	7.4 %	92.6 %	95.5 %	92.1
Google Gemma-7b-it	7.5 %	92.5 %	100.0 %	113.0
Cohere-Chat	7.5 %	92.5 %	98.0 %	74.4
Cohere	8.5 %	91.5 %	99.8 %	59.8
Anthropic Claude 2	8.5 %	91.5 %	99.3 %	87.5
Microsoft Phi 2	8.5 %	91.5 %	91.5 %	80.8
Google Palm 2	8.6 %	91.4 %	99.8 %	86.6
Mixtral 8x7B	9.3 %	90.7 %	99.9 %	90.7

Chatbots May 'Hallucinate' More Often Than Many Realize

When summarizing facts, ChatGPT technology makes things up about 3 percent of the time, according to research from a new start-up. A Google system's rate was 27 percent.

<https://www.nytimes.com/2023/11/06/technology/chatbots-hallucination-rates.html>

Leaderboard: <https://huggingface.co/spaces/vectara/leaderboard>
Model: https://huggingface.co/vectara/hallucination_evaluation_model



Building Applications with Vectara

- [Sign up](#) for a free account
- Create a Vectara corpus and index your data
 - Using [Standard Indexing API](#) or [File Upload API](#)
 - Upload files from the Vectara console
 - [Vectara-ingest](#) - an open source project created by Vectara
 - [Airbyte](#) or [Unstructured](#) integration
- Build your UI and point it at your Vectara corpus
 - Build your own with the Query API
 - Use [Vectara-answer](#) to build question-answer apps
 - Try [Create-UI](#)
 - [React-search](#) and [React-chatbot](#)



Sample App: AskNews

AskNews

Sample news aggregator built using Vectara

Language: English

History

Filter by source

All sources

BBC

NPR

FOX

CNBC

CNN

Describe your interest above or try one of these topics.

Should AI be regulated?

What is Aisha Abdel Gawad's debut novel about?

What was introduced in Apple's latest WWDC?

what happened to Harry and Megan in NYC?



Sample App: AskNews with HHEM

AskNews

Sample news aggregator built using Vectara

Language: English

History

Filter by source

All sources

BBC

NPR

FOX

CNBC

CNN

Describe your interest above or try one of these topics.

Should AI be regulated?

What is Aisha Abdel Gawad's debut novel about?

What was introduced in Apple's latest WWDC?


what happened to Harry and Megan in NYC?





Sample App: Tax chat

Vectara chat demo: Tax Chat

Ask about taxes

 How may I help you?

 Is college tuition a deductible expense?

 College tuition is generally not considered a deductible expense for tax purposes. However, there are specific tax benefits available for education expenses, such as the American opportunity tax credit and the lifetime learning credit. These credits can help offset the cost of tuition and related expenses. It's essential to review the rules for these credits to determine eligibility and how to claim them on your tax return. While tuition itself may not be deductible, there are tax benefits designed to assist with educational costs, providing some relief for individuals pursuing higher education.

URL: <https://huggingface.co/spaces/vectara/IRS-chat>



Thank you!



Try Vectara Now!



- Sign-up:** <https://console.vectara.com/signup> (Free to get started: 50MB text, 15K queries/month)
- Docs:** <https://docs.vectara.com/docs>
- Discord:** <https://discord.gg/GFb8gMz6UH>
- Github:** <https://github.com/vectara/> - OS projects: vectara-answer, create-UI, react-search, vectata-ingest
- Demo apps:** <https://vectara.com/demos>
- Example Notebook** <https://github.com/vectara/example-notebooks/blob/main/notebooks/using-vectara-with-llamaindex.ipynb>
- Startup program:** <https://vectara.com/startups> - gain access to the full power of Vectara, financial incentive, success journey

