



Why do LLMs Hallucinate?

AI conference, San Francisco

September 2023

About Me



Head of Developer Relations @ Vectara

Previously: Syntegra, Helix, LendUp, Hortonworks, Yahoo!
Leader in Machine Learning, Data Science and Big Data

Author of "Practical data science with Hadoop".

Worked on LLMs since late 2019 (my first LLM was **GPT-2**)



Hallucinations: Example

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.



Hallucinations in LLMs

One “definition”:

when the generated content is **incorrect** or **unfaithful** to the provided source content



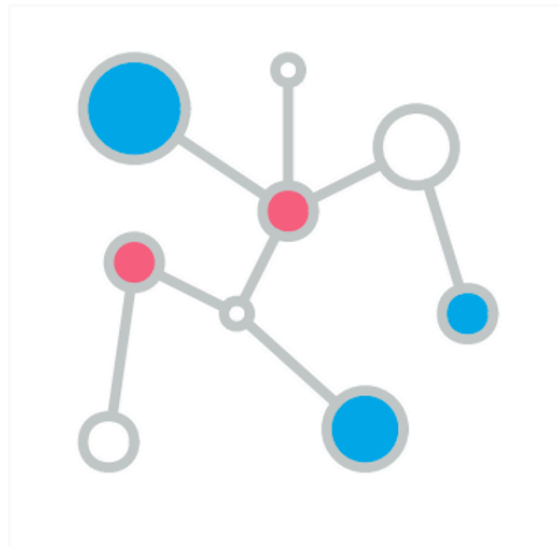
Why do LLMs Hallucinate?



LLM knowledge graph: conceptual paradigm

It is useful to imagine that an LLM has some kind of **knowledge graph** embedded in its weights.

When we ask the LLM to respond, it relies on its knowledge graph to generate a response.



It's not easy for us to understand what that knowledge really is, we can just observe it when we use the LLM.



Why do LLMs hallucinate? Reason #1

You ask a question to which the answer is NOT in the knowledge graph.

Most LLMs are not trained to say “I don’t know” (or caveat their answer) when they don’t have the answer in their knowledge graph.

LLMs generate **highly probable** responses; “highly probable” does not necessarily mean “true” or “correct”, and when this misalignment happens a hallucination is the end result.



Why do LLMs hallucinate? Reason #2

An answer to your question is in the knowledge graph, but:

- It might include fictional content
- It might include content that is subjective like opinions and beliefs.



Addressing Hallucinations



There's a ton of research going on...

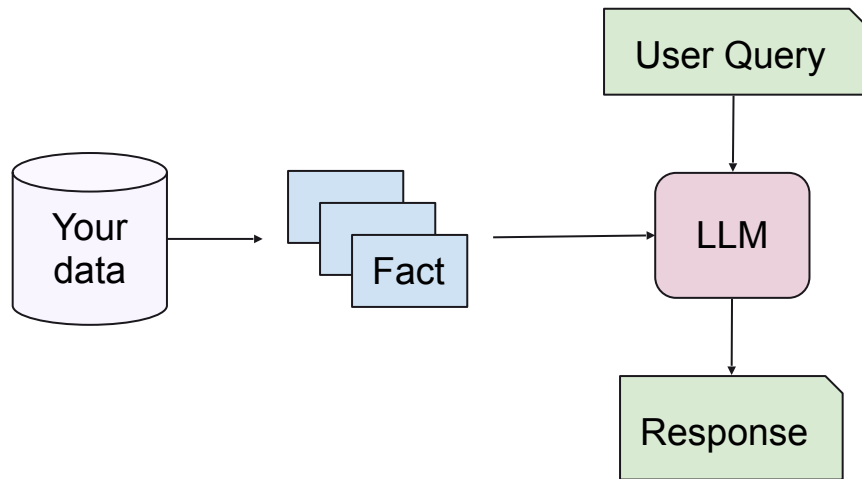
Most promising directions so far are:

- Improve the training data (cleaner / less bias)
- Incorporate human feedback using RLHF or similar techniques
- Improved prompting (esp for reasoning tasks), such as:
 - Chain-of-thought
 - Tree-of-thought
 - React
- Retrieval-augmented generation (aka Grounded Generation)



Addressing Hallucinations with RAG

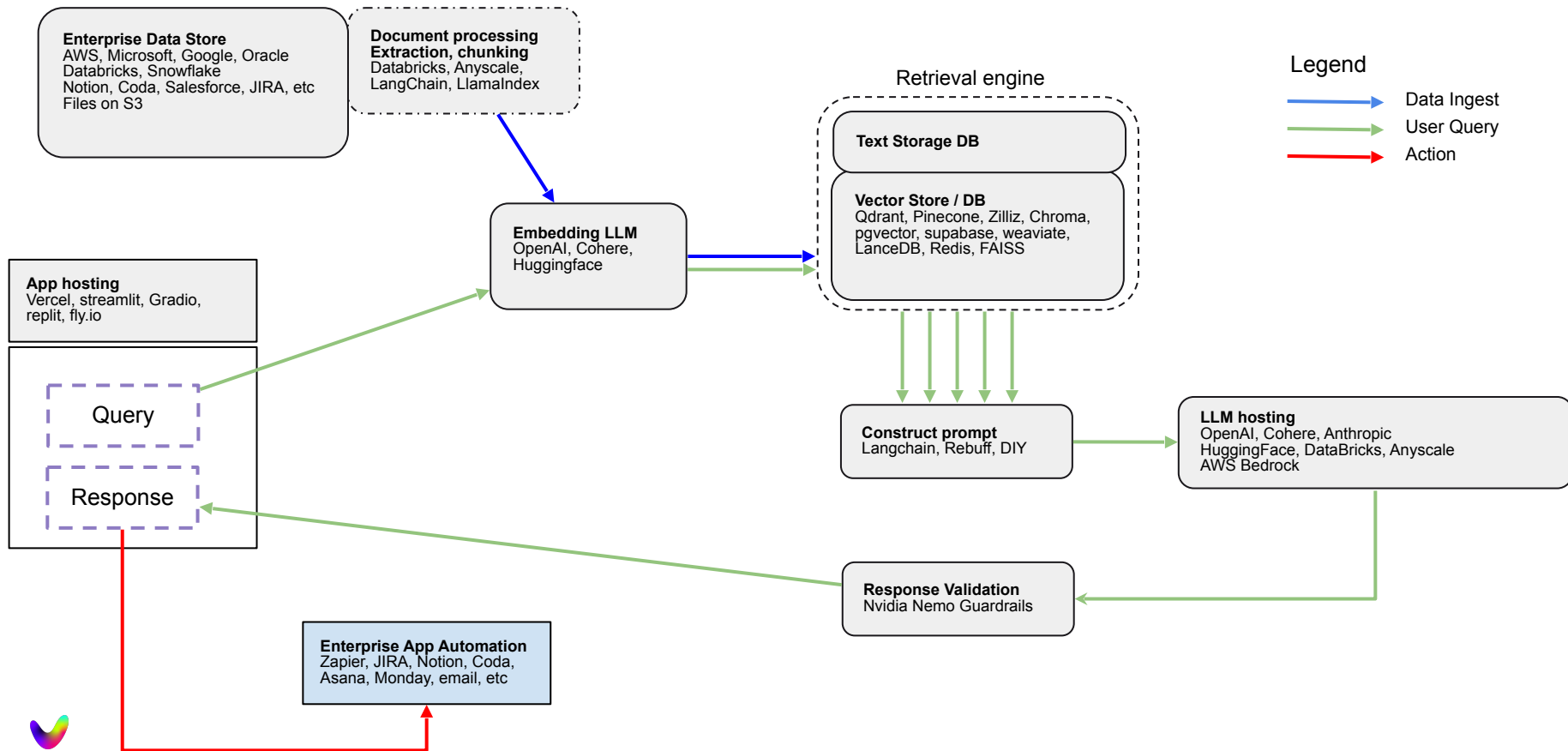
Using **retrieval augmented generation** helps the LLM focus on facts that matter and provide more accurate responses.



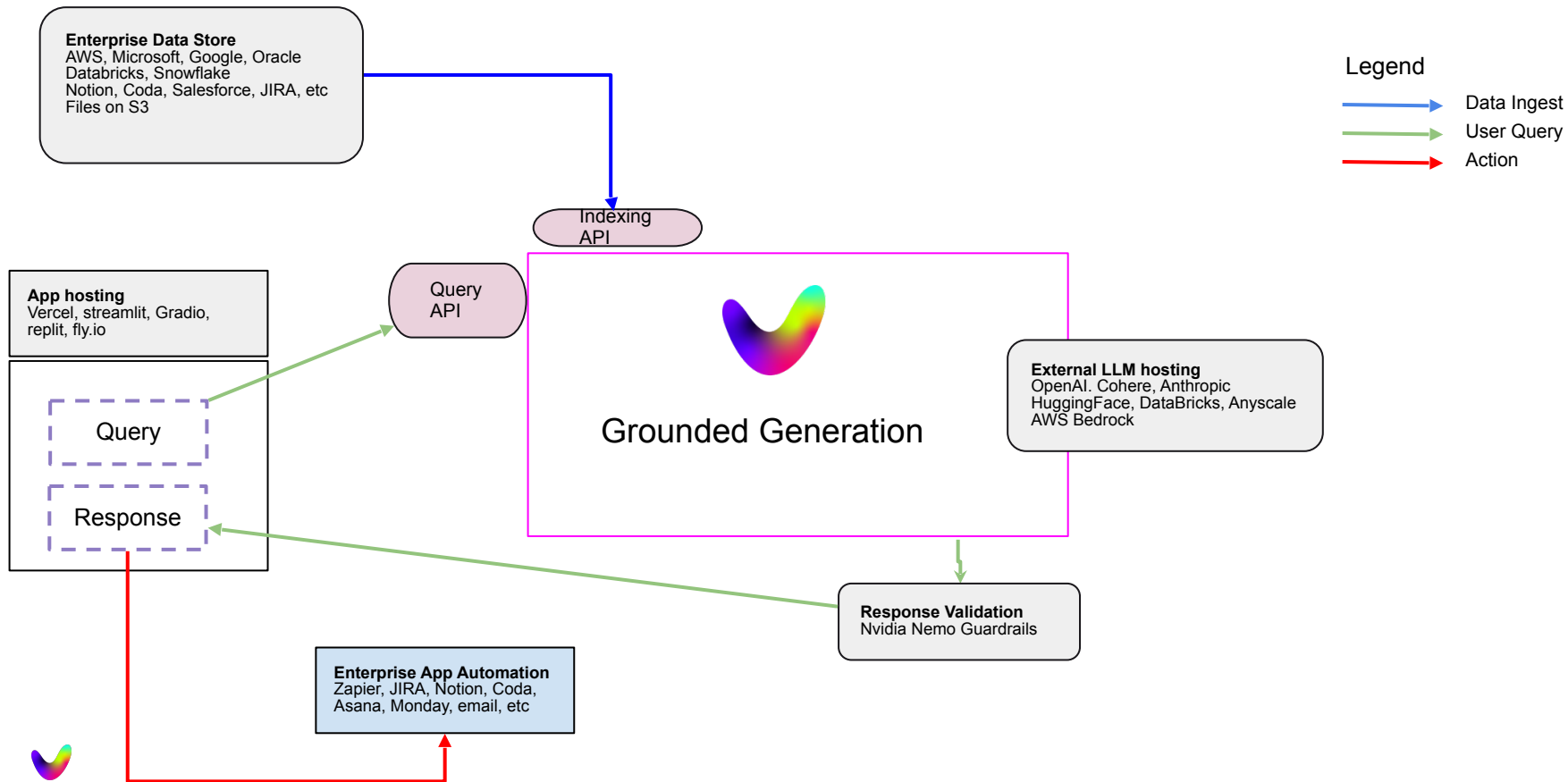
- LLM with “your data”
- AKA “**Grounded Generation**”
- Helps reduce hallucinations
- Use-cases
 - Question answering
 - Chatbot / Co-pilot



RAG: Reference Architecture



RAG as a service with Vectara



Why Retrieval Augmented Generation?

- Augment LLM with **your own data** to reduce hallucinations
- **Private:** data is not used for training the LLM
- **Inexpensive** to implement and use, since no re-training or human feedback collection is required.
- **Complementary** to model improvements like RLHF or better data curation: as LLMs continue to improve and evolve, RAG/GG becomes better as well.



Doing RAG Right: Retrieval Matters

It's critical to have the best retrieval engine possible:

- **Low recall:** if the right facts are missing, then LLM has nothing to work with
- **Low precision:** if wrong facts are included, we add noise

How?

- Optimal chunking of source documents
- SOTA embeddings model (English + other languages)



Doing RAG Right: Chunking Strategies

Madam Speaker, Madam Vice President, and our First Lady and Second Gentleman, members of Congress and the Cabinet, Justices of the Supreme Court, my fellow Americans: Last year, COVID-19 kept us apart. This year, we're finally together again.

Tonight — tonight we meet as Democrats, Republicans, and independents, but, most importantly, as Americans with a duty to one another, to America, to the American people, and to the Constitution, and an unwavering resolve that freedom will always triumph over tyranny.

Six — thank you. Six days ago, Russia's Vladimir Putin sought to shake the very foundations of the free world, thinking he could make it bend to his menacing ways. But he badly miscalculated. He thought he could roll into Ukraine and the world would roll over. Instead, he met with a wall of strength he never anticipated or imagined. He met the Ukrainian people.

...

Fixed Chunking

Madam Speaker, Madam Vice President, and our First Lady and Second Gentleman, members of Congress and the Cabinet, Justices of the Supreme Court, my fellow Americans: Last year, COVID-19 kept us apart. This year, we're finally together again.

Tonight — tonight we meet as Democrats, Republicans, and independents, but, most importantly, as Americans with a duty to one another, to America, to the American people, and to the Constitution, and an unwavering resolve that freedom will always triumph over tyranny.

Six — thank you. Six days ago, Russia's Vladimir Putin sought to shake the very foundations of the free world, thinking he could make it bend to his menacing ways. But he badly miscalculated. He thought he could roll into Ukraine and the world would roll over. Instead, he met with a wall of strength he never anticipated or imagined. He met the Ukrainian people.

...

NLP Chunking

Text from:
2022 State of the Union

Reference: <https://vectara.com/grounded-generation-done-right-chunking/>



Doing RAG Right: Feeding Chunks to LLM

- Create small chunks that have clean semantic meaning (e.g. sentences)
- Augment chunk with additional text around it when providing to LLM

Madam Speaker, Madam Vice President, and our First Lady and Second Gentleman, members of Congress and the Cabinet, Justices of the Supreme Court, my fellow Americans: Last year, COVID-19 kept us apart. This year, we're finally together again.

Tonight — tonight we meet as Democrats, Republicans, and independents, but, most importantly, as Americans with a duty to one another, to America, to the American people, and to the Constitution, and an unwavering resolve that freedom will always triumph over tyranny.

Six — thank you. Six days ago, Russia's Vladimir Putin sought to shake the very foundations of the free world, thinking he could make it bend to his menacing ways.

But he badly miscalculated. He thought he could roll into Ukraine and the world would roll over.

Instead, he met with a wall of strength he never anticipated or imagined. He met the Ukrainian people.

...

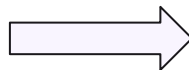
Additional
Text

Matching
chunk



Doing RAG Right: Embedding Models

“Madam Speaker, Madam Vice President, and our First Lady and Second Gentleman, members of Congress and the Cabinet, Justices of the Supreme Court, my fellow Americans: Last year, COVID-19 kept us apart. This year, we’re finally together again.”



0.7	0.24	0.3	0.9	...		
-----	------	-----	-----	-----	--	--

Chunk

Embedding vector

Not all embedding models are created equal

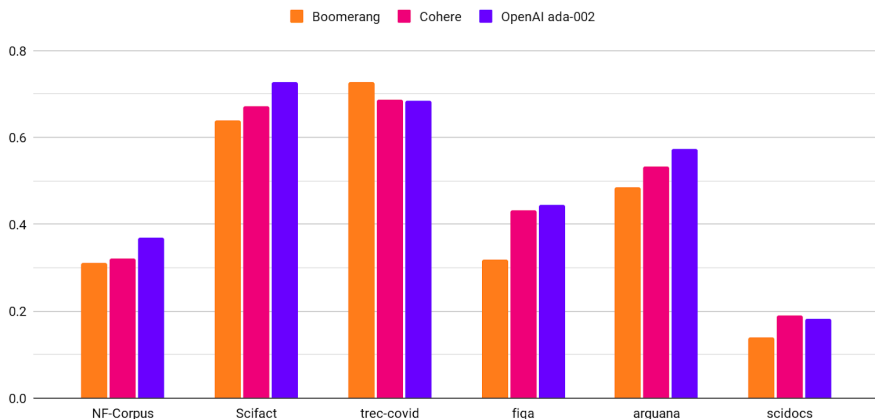


Vectara's Boomerang 🪃 Embedding Model

English

Figure 1: BEIR Benchmark

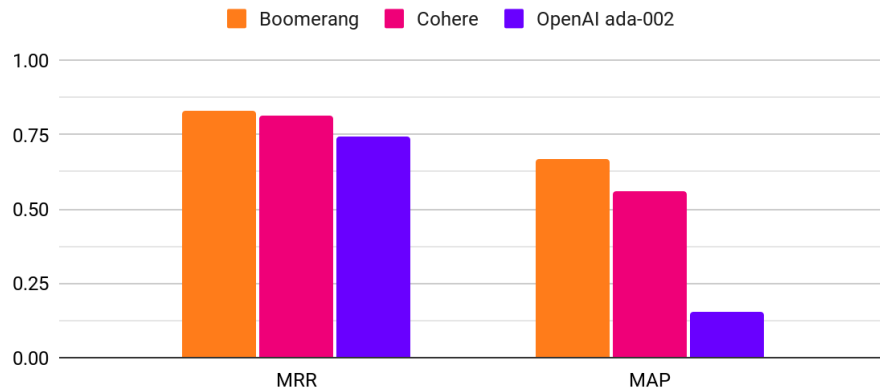
nDCG@10



Cross-Lingual

Figure 5: XQuAD-R Cross-Lingual Benchmark

Multilingual queries with multiple results in different languages



Source: <https://vectara.com/introducing-boomerang-vectaras-new-and-improved-retrieval-model>



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?

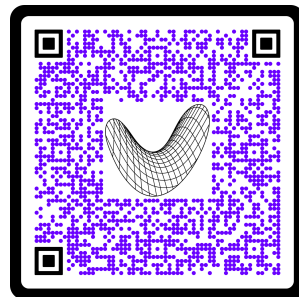
URL: <https://asknews.demo.vectara.com>



Thank you!

Any questions?

Try Boomerang Now!



Sign-up: <https://console.vectara.com/signup>
Docs: <https://docs.vectara.com/docs>
Discord: <https://discord.gg/GFb8gMz6UH>
Github: <https://github.com/vectara/>
Startup program: <https://vectara.com/startups>



[Integration](#)



[Integration](#)

