vectara

How to build a GenAI application with Vectara

A step-by-step guide

Data Council Conference
Austin, Texas
March 2024

About Me

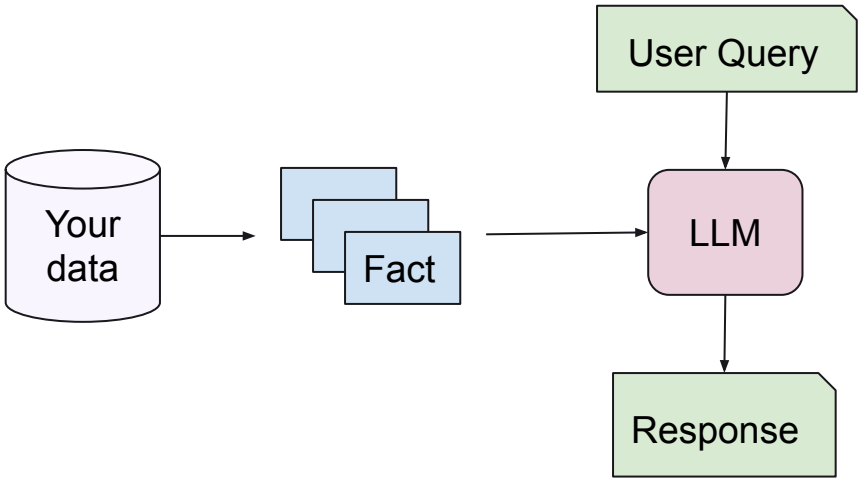
Head of Developer Relations @ Vectara

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

My first LLM was **GPT-2**

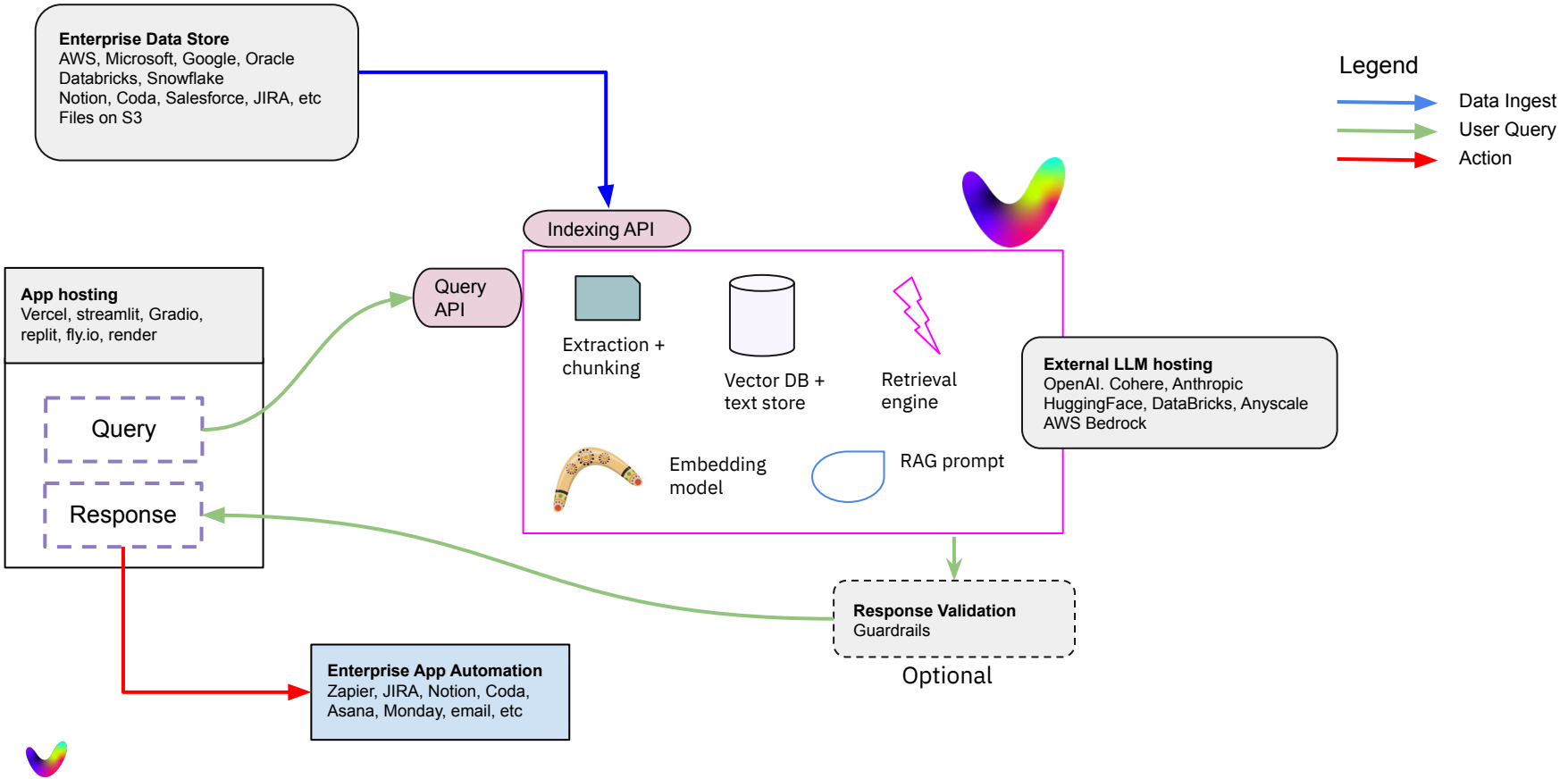
What is RAG?

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

Vectara: RAG-as-a-Service



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

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.

HHEM: Hughes Hallucination Evaluation Model

Model	Hallucination Rate	Factual Consistency Rate	Answer Rate	Average Summary Length (Words)
GPT 4	3.0 %	97.0 %	100.0 %	81.1
GPT 4 Turbo	3.0 %	97.0 %	100.0 %	94.3
GPT 3.5 Turbo	3.5 %	96.5 %	99.6 %	84.1
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
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 (beta)	8.6 %	91.4 %	99.8 %	86.6
Mixtral 8x7B	9.3 %	90.7 %	99.9 %	90.7
Amazon Titan Express	9.4 %	90.6 %	99.5 %	98.4
Mistral 7B	9.4 %	90.6 %	98.7 %	96.1
Google Palm 2 Chat (beta)	10.0 %	90.0 %	100.0 %	66.2
Google Palm 2	12.1 %	87.9 %	92.4 %	36.2
Google Palm 2 Chat	27.2 %	72.8 %	88.8 %	221.1

What is HHEM?

- Open source model
- Detects hallucinations

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

Demo Time

- Sign up for a free account: https://console.vectara.com/signup
- Create a Corpus + API Key
- Upload data (SOTU 2024 text)
- Run some queries
- Create chatbot using Create UI
- Run it!
- Done!

Thank you!

Try Vectara Now!

SCAN ME

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