Building Chatbots that Know All About Your Business

Retrieval Augmented Generation (RAG)

Nikita Kozodoi, PhD

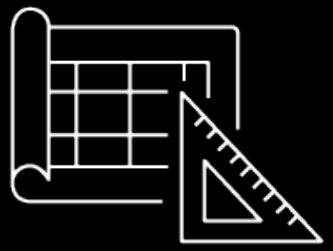
22.03.2024

About Me

https://kozodoi.me

- Applied Scientist at Amazon Web Services
- Building Generative AI solutions across industries
- Earned PhD in ML for Credit Risk Analytics
- Won 18 Kaggle competition medals

My Team: Generative AI Innovation Center



Design

Deploy

Drive

Design guidance:

- Select the GenAI use case with the highest business impact
- Design how to develop, train, and deploy it to production

Deploy recommended solutions:

- Develop and fine-tune a GenAI solution to meet your business objectives and demonstrate what's possible

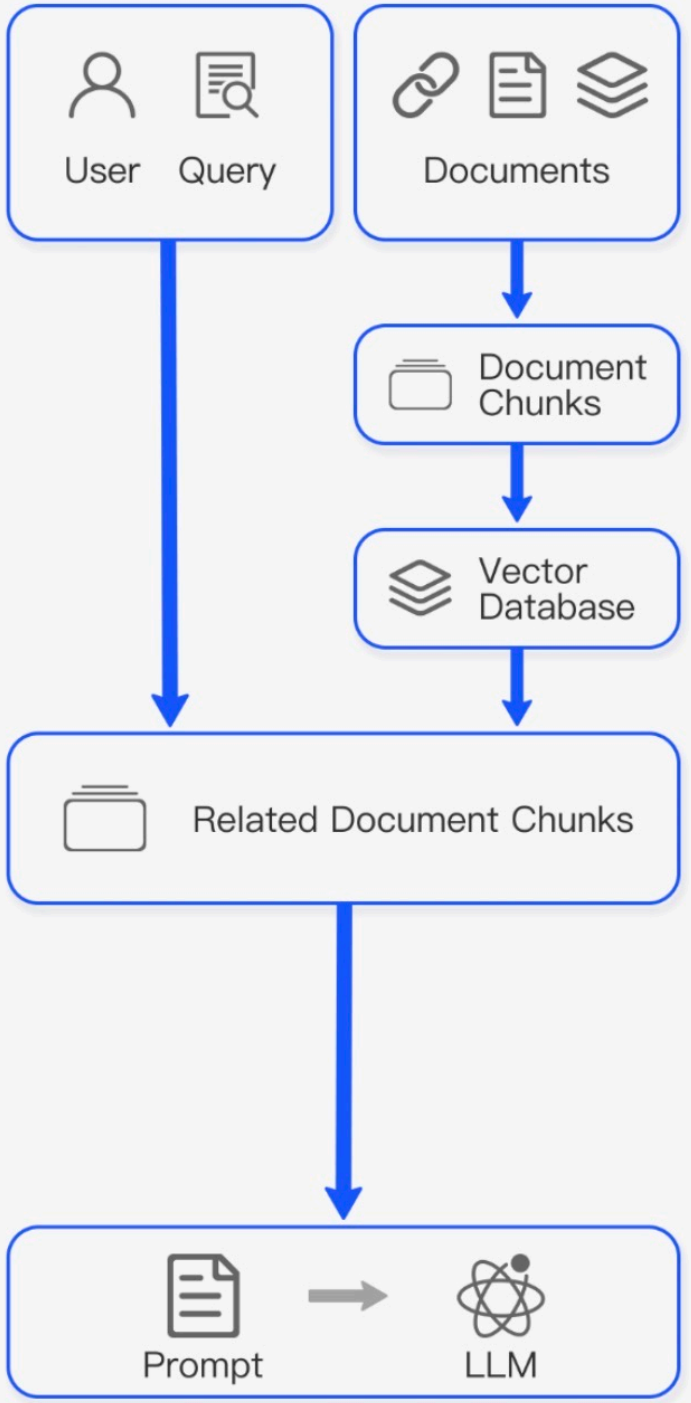
Drive adoption:

- Accelerate stickiness and adoption with a path to production for your GenAI solution integrated into your application.

https://aws.amazon.com/generative-ai/innovation-center/

Agenda

- Why Do We Need RAG Chatbots?
- How Do RAG Chatbots Work?
- Building RAG Systems
- Take Aways



https://arxiv.org/abs/2312.10997

Why Do We Need RAG Chatbots?

Döner is All You Need GmbH

Magic Döner Menu

Magic Döner

Not So Magic Döner

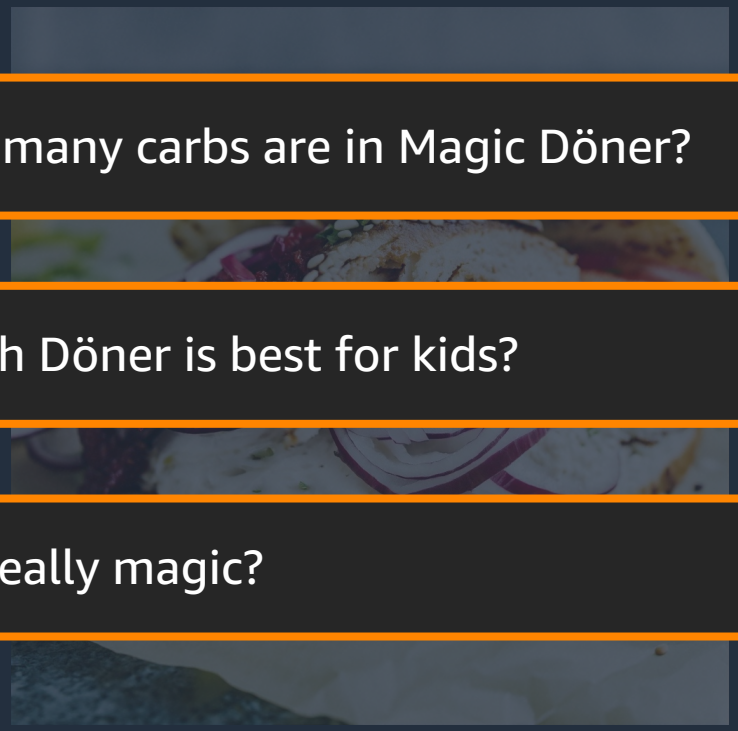
Döner is All You Need GmbH

Magic Döner Menu

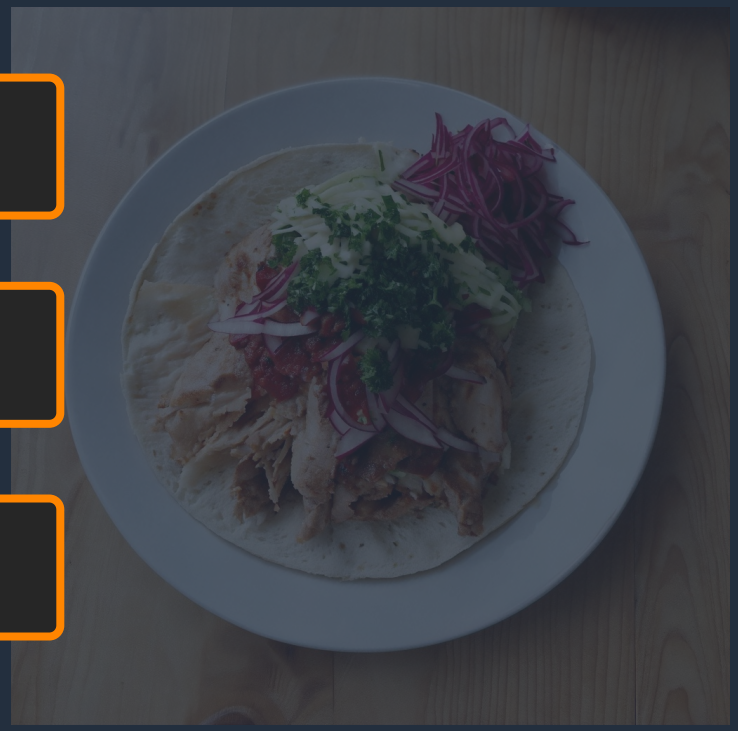
How many carbs are in Magic Döner?

Which Döner is best for kids?

Is it really magic?



Magic Döner



Not So Magic Döner

Search vs Q&A Chatbot

How many carbs are in Magic Döner? 

[1] Magic Döner Reviews

The secret to a great doner is in the meat mixture. Magic Döner combines lamb and beef with...

[2] 3 Ways to Ruin a Doner Kebab

Want to make an inedible doner? 1) Use low-quality meat; 2) Don't let it cook evenly; 3) Get the...

Search

How many carbs are in Magic Döner? 

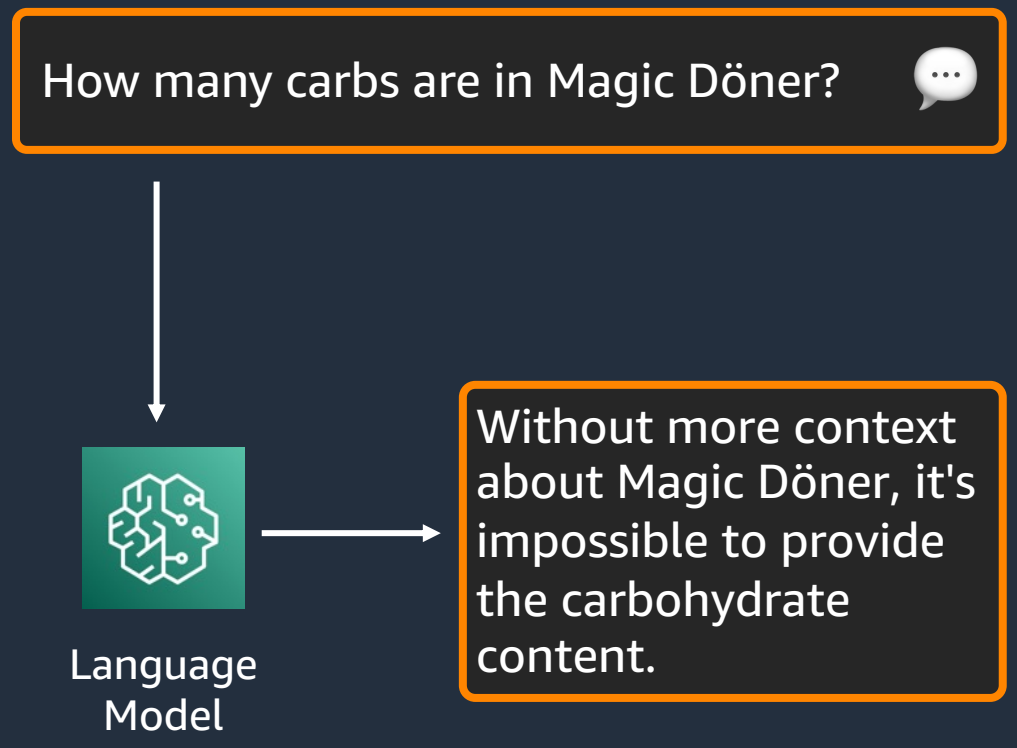
32g of carbs per serving.

[1] Magic Döner Reviews

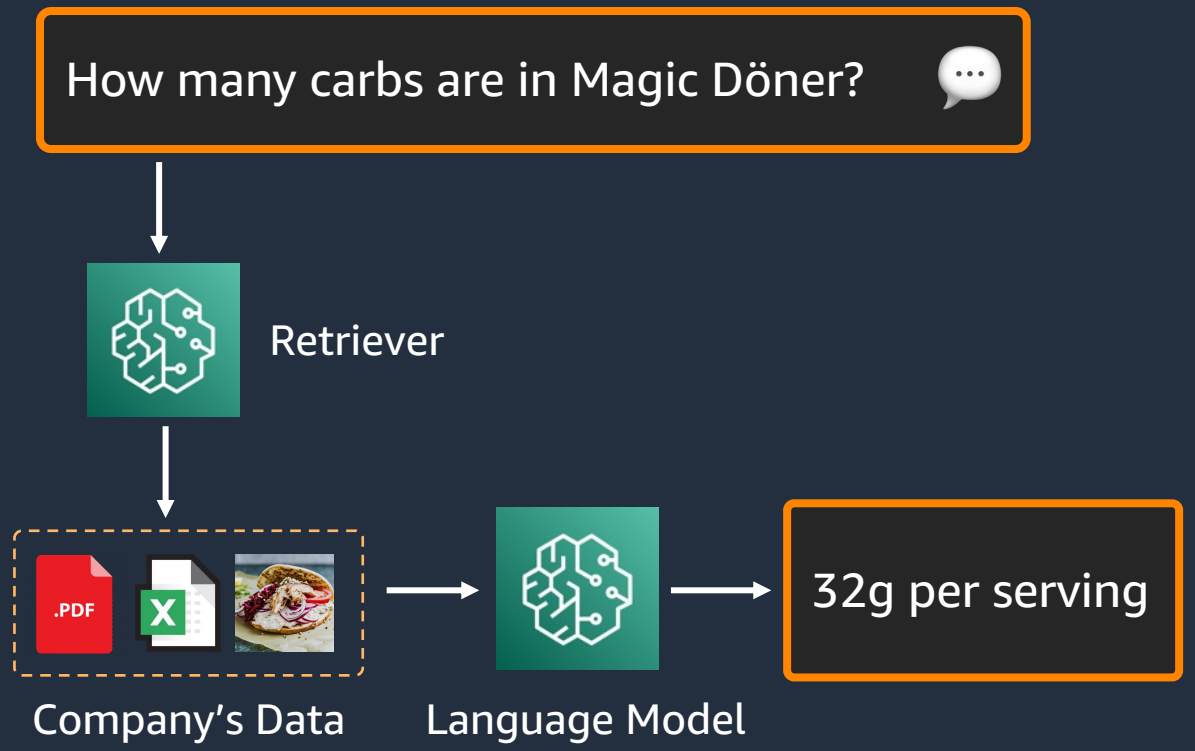
[2] Nutrition Cards

Q&A Chatot

Why Do We Need RAG?



Chatbot with no RAG

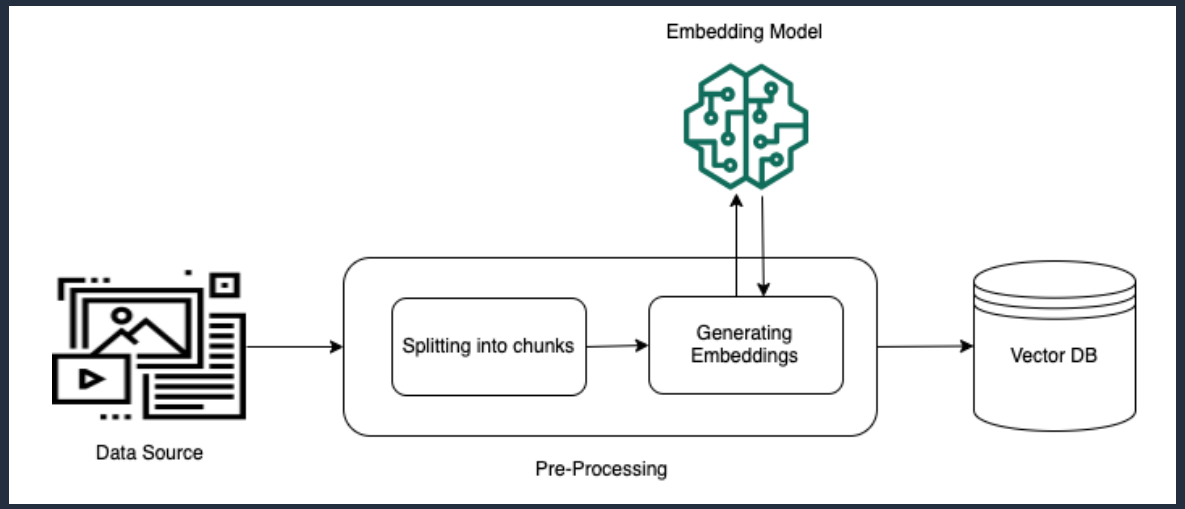


Chatbot with RAG

How Do RAG Chatbots Work?

RAG Chatbot Architecture: Stage 1/2

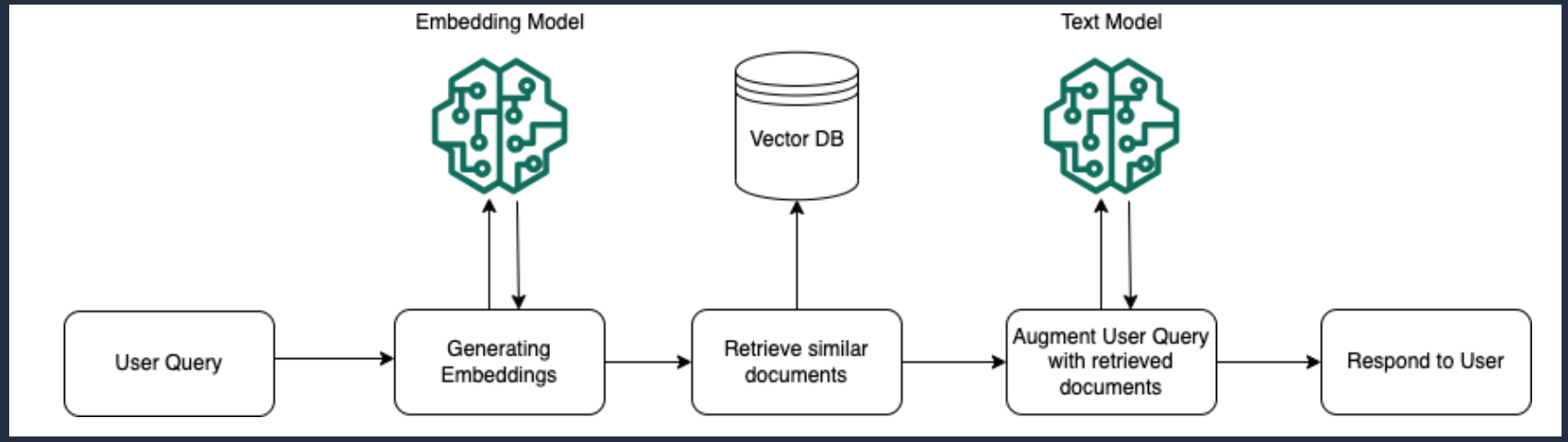
- Each document is parsed to extract text
- Texts are split into chunks (e.g. paragraphs)
- Each chunk is embedded into a vector
- Vectors and text chunks are stored in a database



Stage 1. Indexing

RAG Chatbot Architecture: Stage 2/2

- User question is embedded into a vector
- Top-K similar chunks are fetched from the database
- User question and chunks are send to LLM
- LLM provides response to the user



Stage 2. Runtime

Building RAG Systems

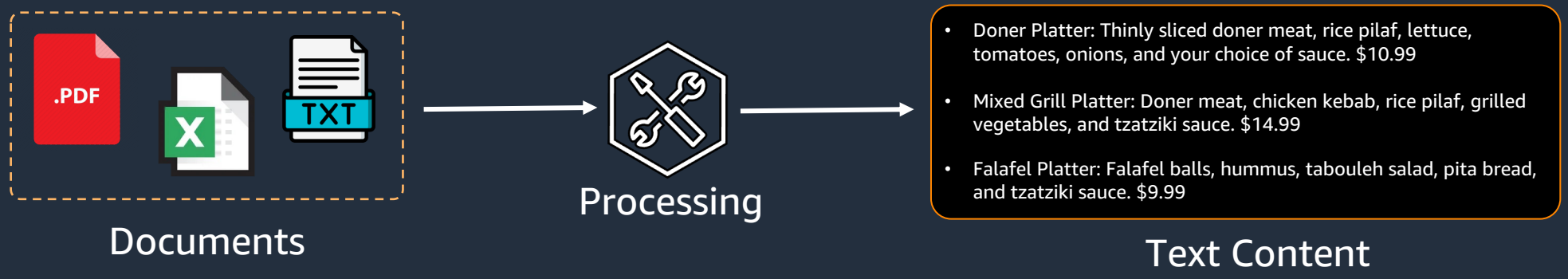
#1. Data Processing

Element 1: File Format & Layout

Embedding models require each document to be parsed into text

Challenges:

- How to deal with file formats? PDF, MS Office, Audio, …
- How to deal with tables and diagrams?

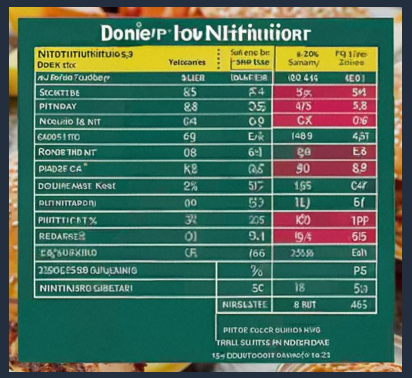


Element 1: File Format & Layout

Embedding models require each document to be parsed into text

Challenges:

- How to deal with file formats? PDF, MS Office, Audio, …
- How to deal with tables and diagrams?



Processing

Nutrition Facts	Per Serving	% Daily Value
Calories	450	23%
Total Fat	22g	34%
Saturated Fat	8g	40%
Trans Fat	0g	-
Cholesterol	75mg	25%
Sodium	980mg	41%
Total Carbs	40g	13%
Dietary Fiber	3g	12%

Text Content

Element 2: Chunking Strategy

Important to carefully select chunking strategy: size, splits, rules, etc

Why:

- Embedding very large chunks => **too much noise** => poor recall
- Embedding very small chunks => **limited context** => incomplete answers

Instructions:

1. In a large bowl, mix together meat, grated onion, garlic, cumin, paprika, cayenne, oregano, salt, and black pepper.

2. On a work surface, shape the meat mixture into one long, tightly packed loaf about 8-10 inches long and 6 inches wide.

3. Carefully skewer the meat loaf lengthwise onto a rotisserie spit or strong metal skewer, packing it firmly together.

VS

Instructions:

1. In a large bowl, mix together…

… meat, grated onion, garlic, cumin, paprika, cayenne, oregano…

Building RAG Systems

#2. LLM Selection & Prompting

Element 1: Embedding Model

English | Chinese | French | Polish

Overall MTEB English leaderboard 🔮

- **Metric:** Various, refer to task tabs
- **Languages:** English

Rank	Model	Model Size (GB)	Embedding Dimensions	Max Tokens	Average (56 datasets)	Classification Average (12 datasets)	Clustering Average (11 datasets)	Pair Classification Average (3 datasets)	Reranking Average (4 datasets)	Retrieval Average (15 datasets)
1	SFR-Embedding-Mistral	14.22	4096	32768	67.56	78.33	51.67	88.54	60.64	59
2	voyage-lite-02-instruct		1024	4000	67.13	79.25	52.42	86.87	58.24	56.6
3	GritLM-7B	14.48	4096	32768	66.76	79.46	50.61	87.16	60.49	57.41
4	e5-mistral-7b-instruct	14.22	4096	32768	66.63	78.47	50.26	88.34	60.21	56.89
5	GritLM-8x7B	93.41	4096	32768	65.66	78.53	50.14	84.97	59.8	55.09
6	echo-mistral-7b-instruct-last	14.22	4096	32768	64.68	77.43	46.32	87.34	58.14	55.52
7	mxbai-embed-large-v1	0.67	1024	512	64.68	75.64	46.71	87.2	60.11	54.39
8	UAE-Large-V1	1.34	1024	512	64.64	75.58	46.73	87.25	59.88	54.66
9	text-embedding-3-large		3072	8191	64.59	75.45	49.01	85.72	59.16	55.44
10	voyage-lite-01-instruct		1024	4000	64.49	74.79	47.4	86.57	59.74	55.58

https://huggingface.co/spaces/mteb/leaderboard

Element 2: Language Model

Total #models: **73**. Total #votes: **408144**. Last updated: March 13, 2024.

Contribute your vote 🗳️ at chat.lmsys.org! Find more analysis in the notebook.

Rank	🤖 Model	⭐ Arena Elo	📊 95% CI	🗳️ Votes	Organization	License	Knowledge Cutoff
1	GPT-4-1106-preview	1251	+5/-4	48226	OpenAI	Proprietary	2023/4
1	GPT-4-0125-preview	1249	+5/-6	22282	OpenAI	Proprietary	2023/12
1	Claude 3 Opus	1247	+6/-6	14854	Anthropic	Proprietary	2023/8
4	Bard (Gemini Pro)	1202	+6/-7	12623	Google	Proprietary	Online
4	Claude 3 Sonnet	1190	+6/-6	14845	Anthropic	Proprietary	2023/8
5	GPT-4-0314	1185	+4/-6	27245	OpenAI	Proprietary	2021/9
7	GPT-4-0613	1159	+4/-5	43783	OpenAI	Proprietary	2021/9
7	Mistral-Large-2402	1155	+5/-6	18959	Mistral	Proprietary	Unknown
8	Qwen1.5-72B-Chat	1146	+4/-5	16729	Alibaba	Qianwen LICENSE	2024/2
8	Claude-1	1145	+5/-6	21929	Anthropic	Proprietary	Unknown
8	Mistral Medium	1145	+5/-4	23931	Mistral	Proprietary	Unknown

https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard

Selecting a Language Model

Things to watch out for:

- 🎯 **Benchmark performance:** how good are the answers?
- 🏎️ **Inference latency:** how long does one answer take?
- 💰 **Generation cost:** how expensive is an average answer?
- 🗣️ **Supported languages:** is my target language supported?
- 📄 **Usage license:** can I use it for commercial purposes?
- 🚀 **Deployment type:** do I need to deploy it myself?

Prompt Engineering

You are a helpful AI assistant.

…

Carefully read the context and answer the human question.

{context}

Human: {question}

Answer:

Prompt Template

1. In a large bowl, mix together meat, grated onion, garlic, cumin, paprika, cayenne, oregano, salt, and black pepper.

2. On a work surface, shape the meat mixture into one long, tightly packed loaf about 8-10 inches long and 6 inches wide.

3. Carefully skewer the meat loaf lengthwise onto a rotisserie spit or strong metal skewer, packing it firmly together.

Retrieved Text Chunks

How many carbs are in Magic Döner?

User Question

Building RAG Systems

#3. Evaluation Pipeline

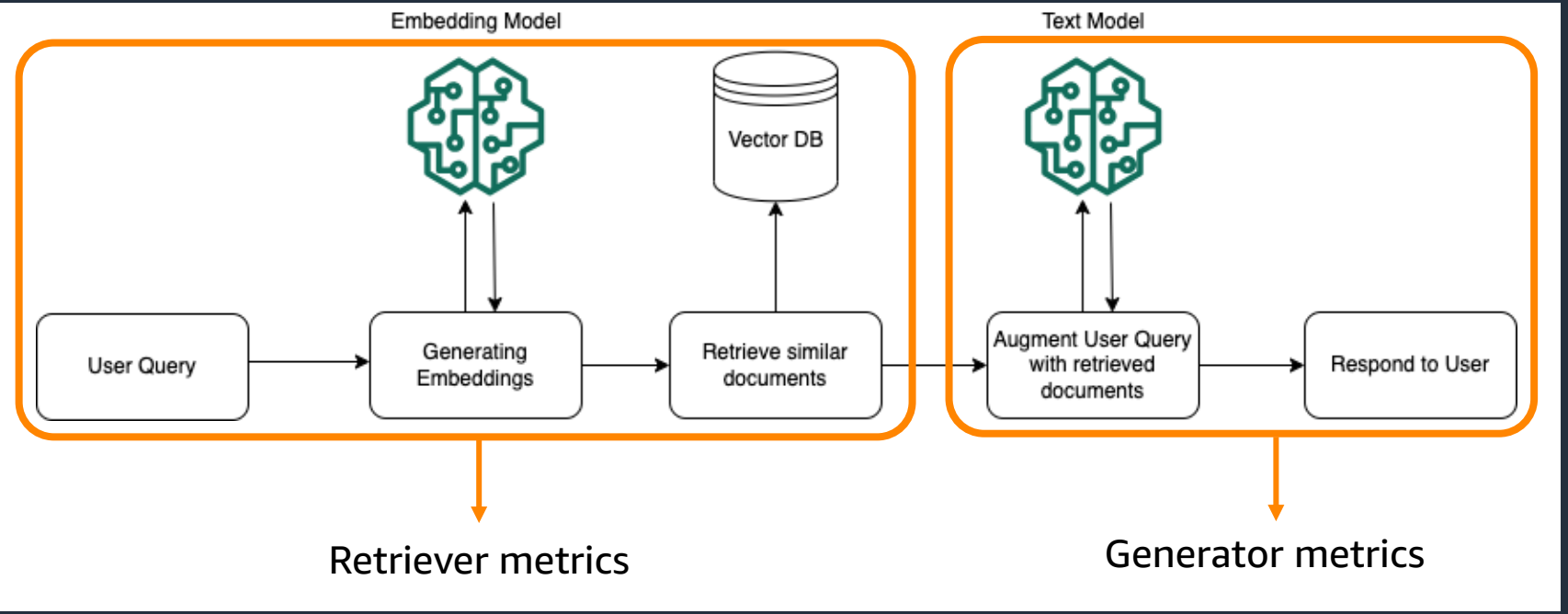
Building Evaluation Pipeline

Collect a test set with human-labelled ground truth answers

Question	Sources	Correct answer
How many carbs are in Magic Döner?	[doc1, doc2, doc3]	32 grams per serving.
Which Döner is good for kids?	[doc2, doc4, doc5]	Go for Magic Chicken Döner. It is more mild and familiar in flavor for children.
…	…	…
If döners could talk, what would they say while spinning and cooking on the rotisserie?	[doc1, doc3, doc4]	Keep on spinning, baby! A few more turns and I'll have the perfect crispy outer layer!

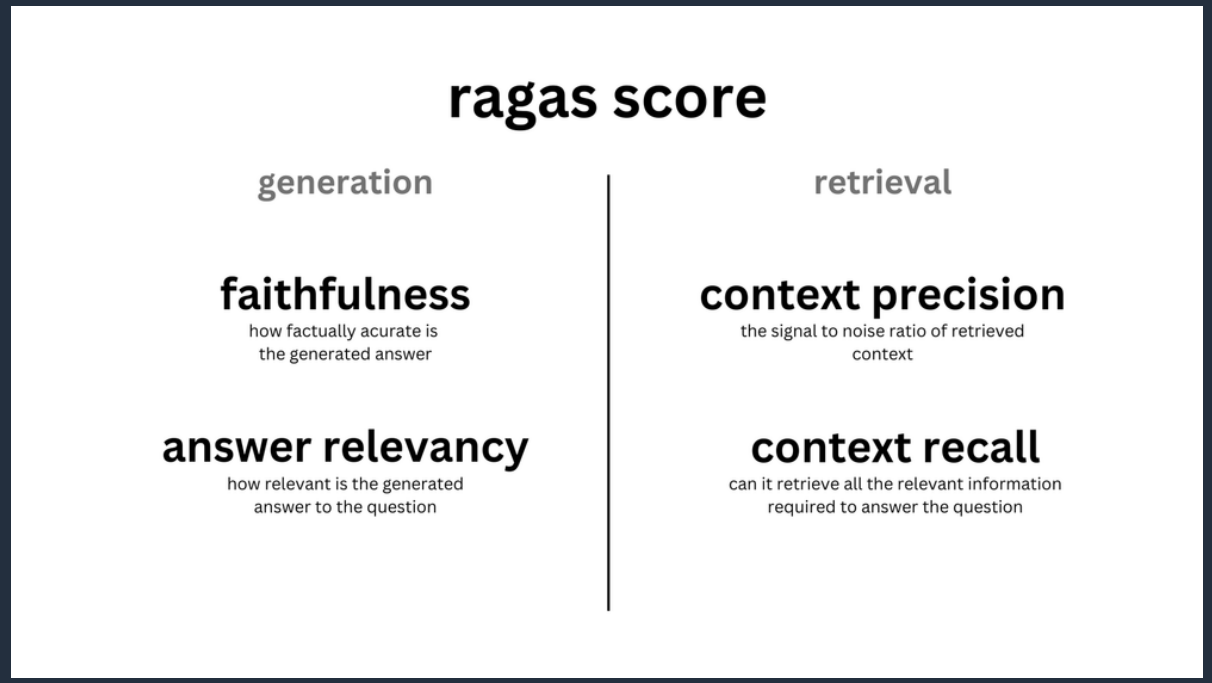
Building Evaluation Pipeline

Evaluate retriever & generator separately

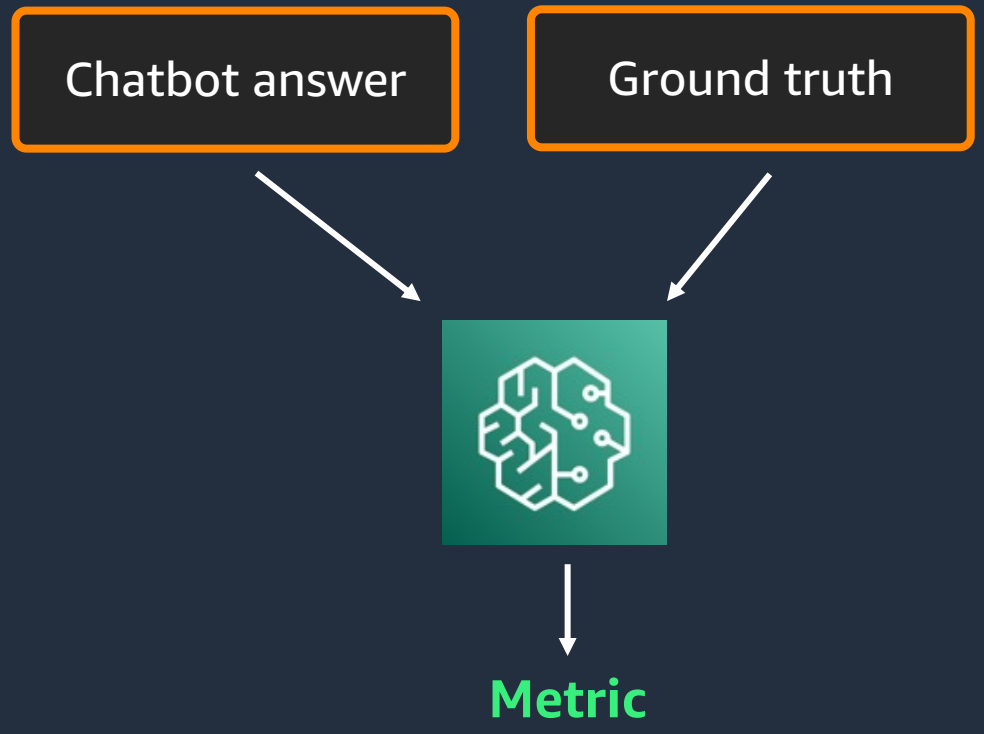


Building Evaluation Pipeline

Automated LLM-based evaluation



https://docs.ragas.io/en/stable/concepts/metrics/index.html



Take Aways

- **Data processing** is crucial to make sure LLM has good “notes” when answering questions
- **Selecting LLMs** requires thinking about many dimensions, including cost, latency, and others
- Building an **evaluation pipeline** is very important to keep the RAG system robust

https://kozodoi.me

Nikita Kozodoi, PhD