Microsoft

LLMOps and What Comes next for RAG

Rob Ferguson
Head of AI
Microsoft for Startups

Microsoft

~~LLM Ops~~ GenAI Ops What Comes next for RAG

Rob Ferguson
Head of AI
Microsoft for Startups

About me…

Head of AI for Microsoft for Startups

AI for 20 years - Built AI Systems for Millions of Users (Rdio->Pandora, etc)

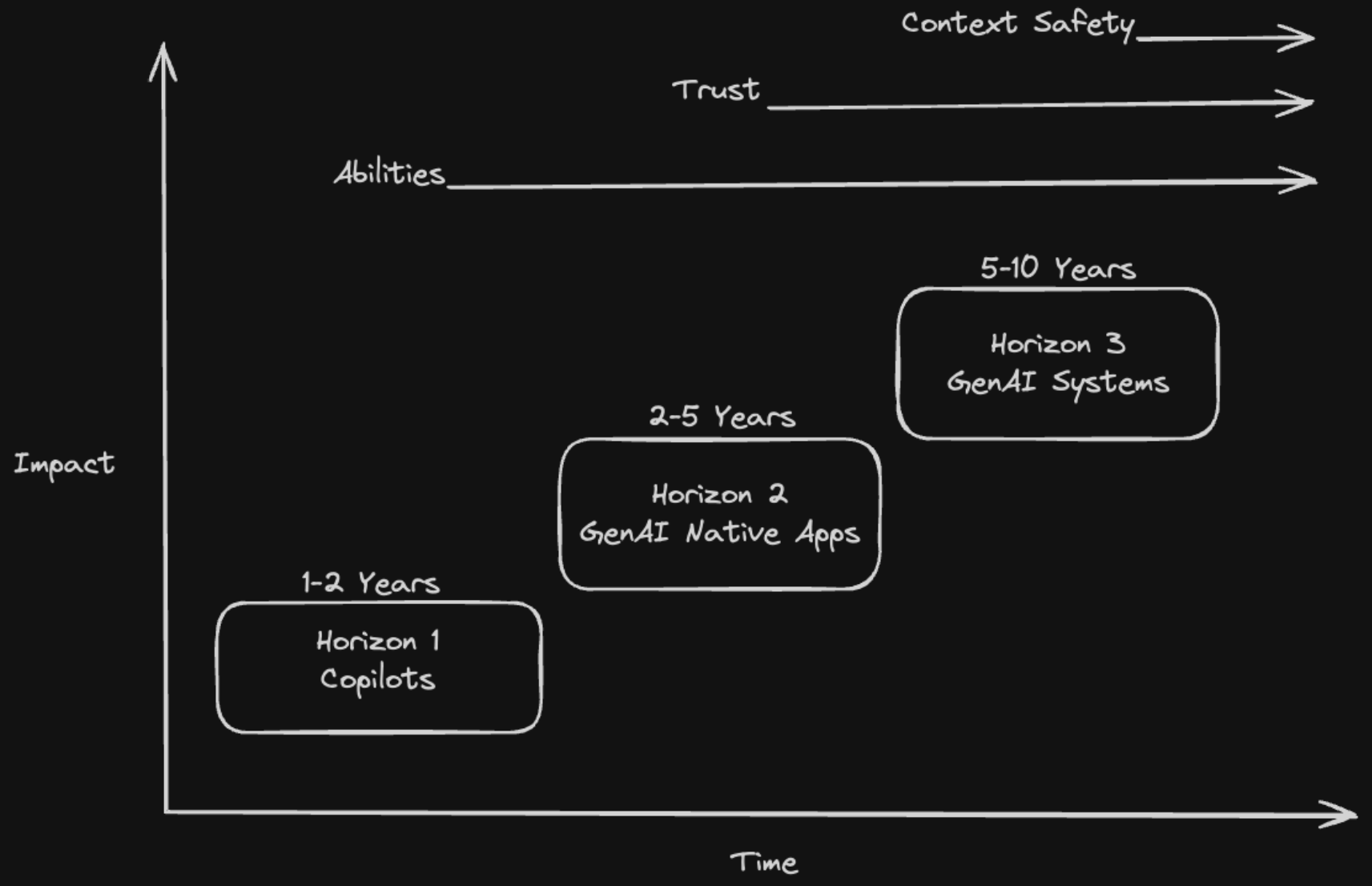
CTO/VPE – YC Top 10 Exits – CTO Automatic Labs
Ran Engineering Teams >100
Built on AWS/GCP/Azure

Global Head of AWS AI/ML BD Startups and VC

Not a Robot…
https://www.linkedin.com/in/rwfiii/

Gen AI Horizons

Gen AI Horizons



Horizon 1 - Copilots

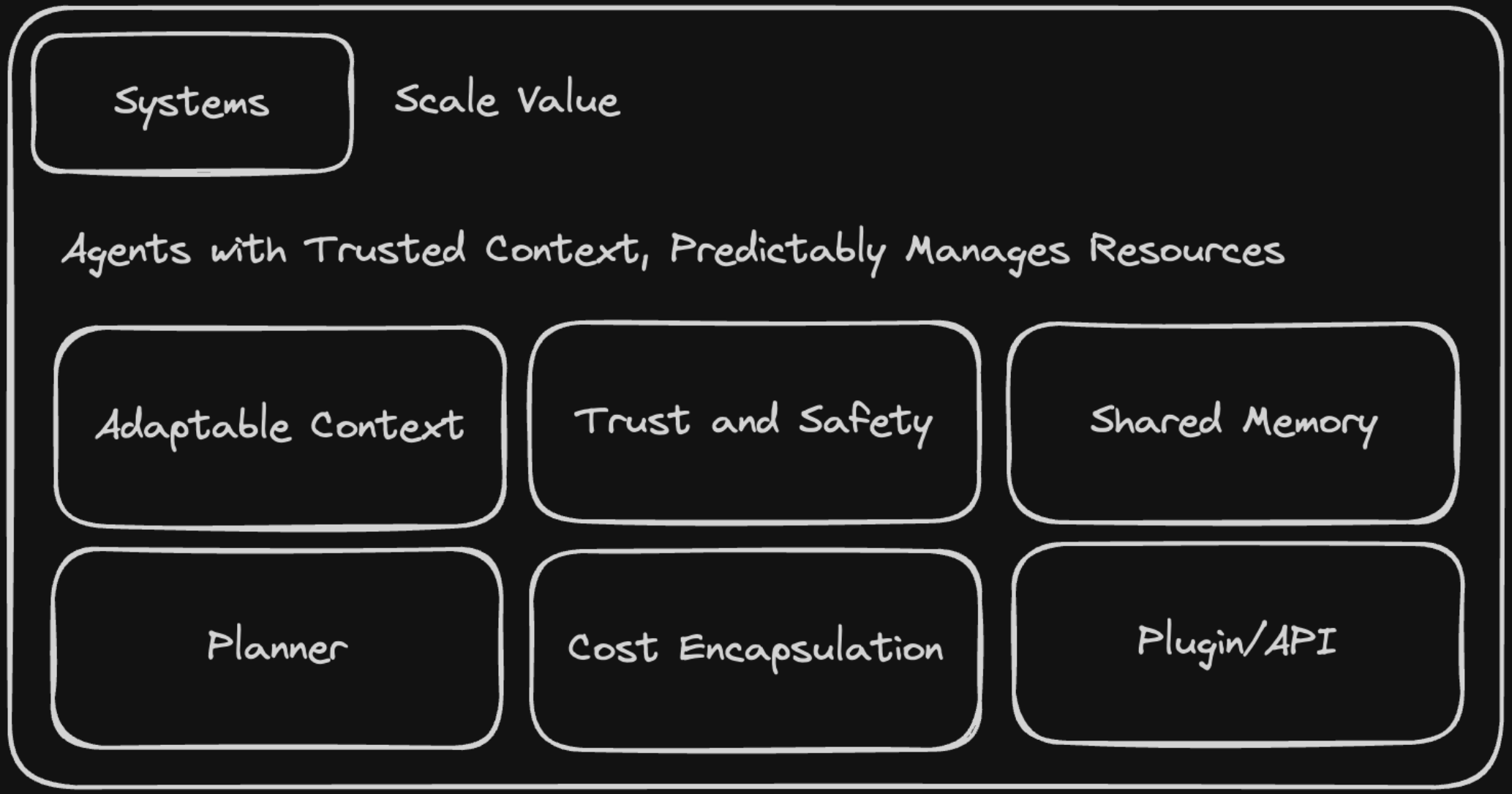
Adopt Gen AI Quickly -
Explore Model Capabilities
Adapt Existing Workflows

Horizon 2 - Apps

Deliver Consistent Value -
Pre-Defined Context
Balances Cost / Abilities

Horizon 3 - Systems

Scalable Value -
Adaptable Context
Adaptive Cost and Memory


Systems
Scale Value
Agents with Trusted Context, Predictably Manages Resources
Adaptable Context
Trust and Safety
Shared Memory
Planner
Cost Encapsulation
Plugin/API

Getting to Gen AI Natives

GenAI-First: Integrate Tightly Solve Long Standing Problems

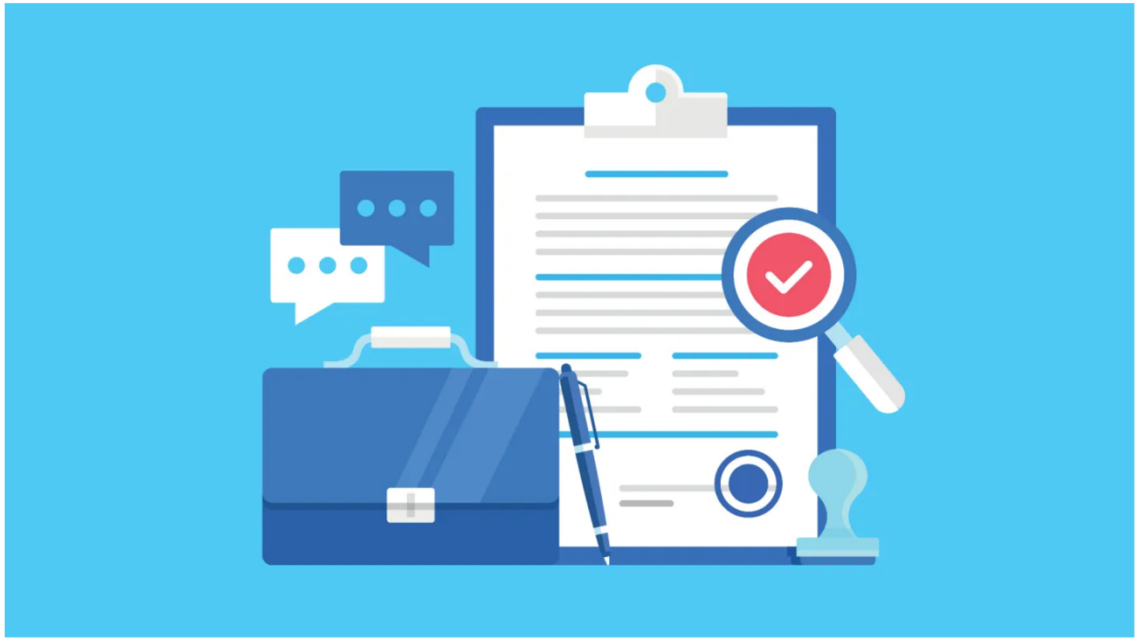
Harvey AI raised $100 Million

Lawyers are infamous for not buying tech. Harvey helped turn legal-speak into case law references.

For Lawyers hallucinations was a feature!

Harvey, which uses AI to answer legal questions, lands cash from OpenAI

Kyle Wiggers @kyle_l_wiggers / 5:00 AM PST • November 23, 2022 Comment



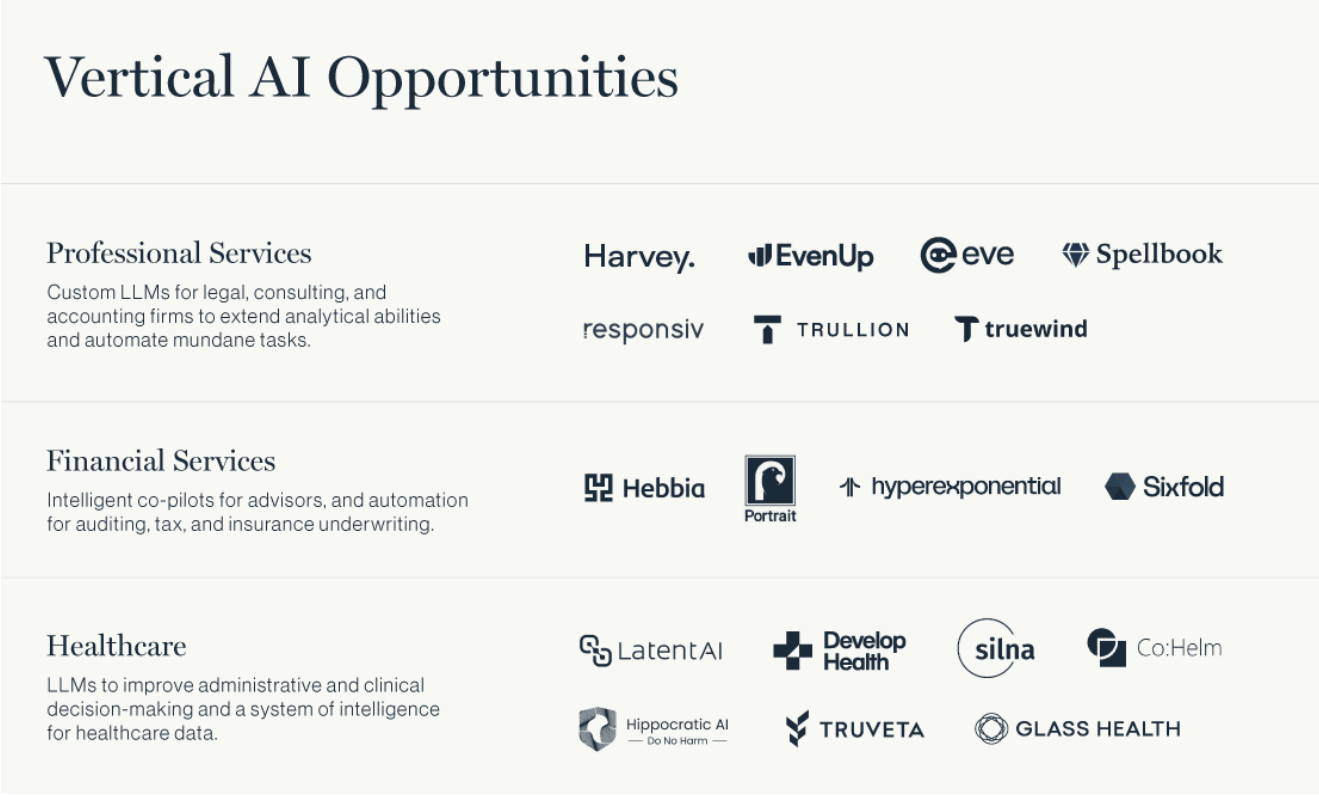
An example co-pilot

Verticals deliver GenAI Success early by Limiting Context

Glean 360M
- Search Company Data with Corporate Jargon

Sierra 195M
- Chat Based Support
(Triage, Response, Confirm, Collect, Search)

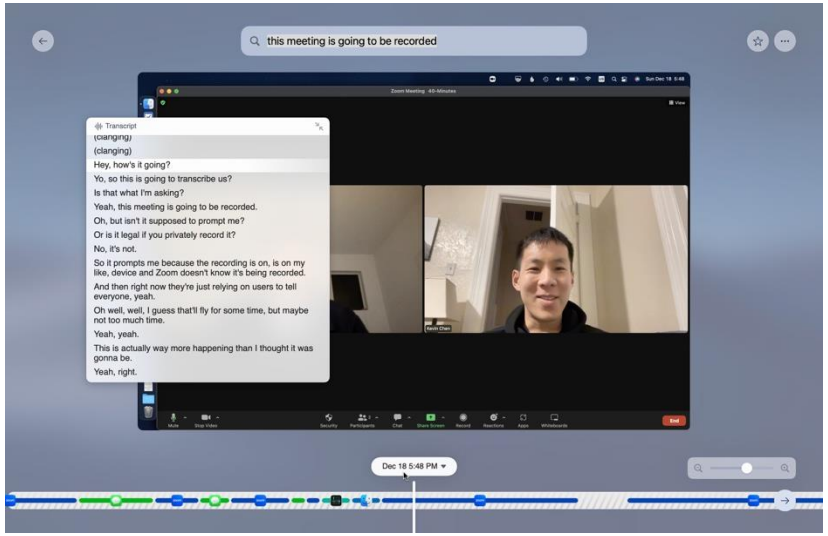
Harvey 100M
-Search Legal Documents with Legal Speak



Credit: Greylock

Getting to GenAI Native: Rewind Case Study

GenAI-First Transcription App
Transcribes everything you do on a machine
Compresses 3,750x using the Embedded ML Accelerator
Sends just what you want summarized to GPT-4



(clanging)
(clanging)
Hey, how's it going?
Yo, so this is going to transcribe us?
Is that what I'm asking?
Yeah, this meeting is going to be recorded.
Oh, but isn't it supposed to prompt me?
Or is it legal if you privately record it?
No, it's not.
So it prompts me because the recording is on, is on my like, device and Zoom doesn't know it's being recorded.
And then right now they're just relying on users to tell everyone, yeah.
Oh well, well, I guess that'll fly for some time, but maybe not too much time.
Yeah, yeah.
This is actually way more happening than I thought it was gonna be.
Yeah, right.

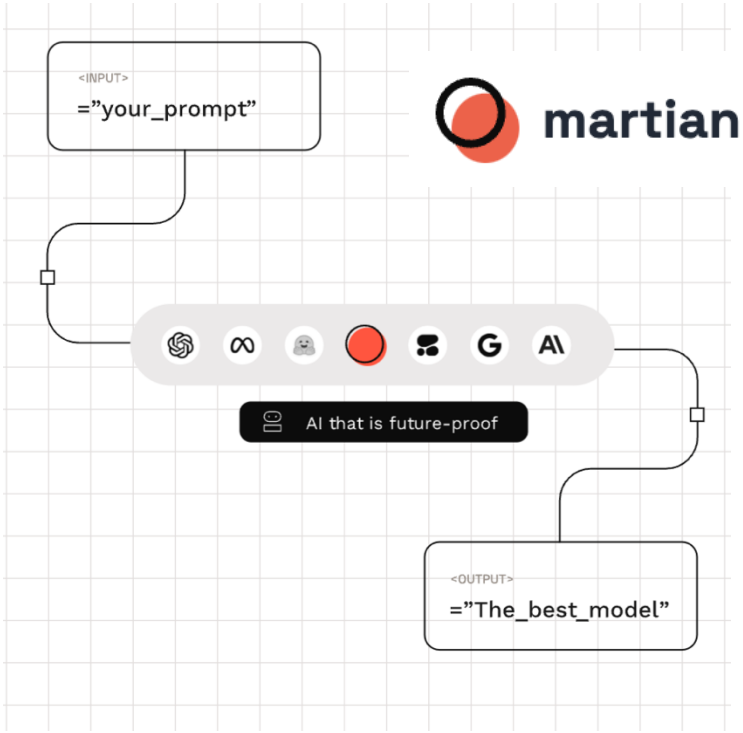
Credit: Kevin Chen

How it might work?

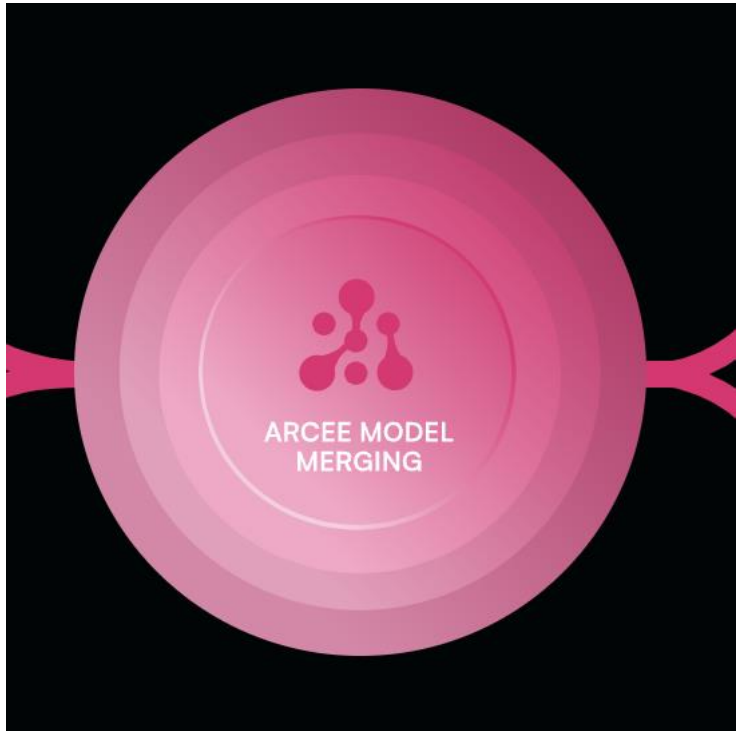
- Use accessibility APIs to identify the frontmost window.
- Take a screenshot of the screen that contains the frontmost window.
- If there are multiple screens, only the currently focused screen will be captured.
- Use ScreenCaptureKit to hide disallowed windows, including private browser windows
- OCR the screenshot on-device using Apple's Vision framework
- Compress the screenshot sequence to an H.264
- Transcribe the audio on-device using the OpenAI Whisper model.

Expect Multiple Models

Stateless Model Interoperation



Model Routing



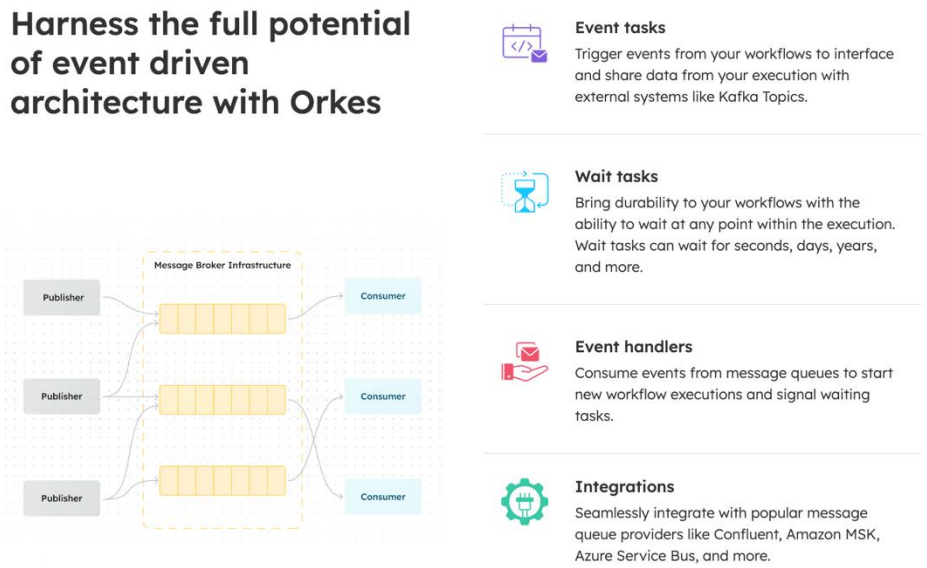
Model Merging

Yes, There Will Be DevOps

Orkes.io is built on the top of this widely popular orchestration engine: Netflix Conductor.

As your business evolves to day 2:
You might be launching in a new country...
the way you use LLM flows will be slightly different
Jeu George CEO Orkes.

LLMs need to work with queues, tasks , events...



Lessons from Leaders in RAG

Leaders in Retrieval Augmentation

- Interviewed CEOs of RAG Companies
- Talked to Hundreds
- Highlighting the response of 6 of them

Whitepaper coming soon…

Co-Author: Nick Giometti
Principal at B Capital

Retrieval-Augmented Generation

Works like **Search**:
Rerank, Summarize, Fuse Data Together

- RAG solves some of the problems with general Copilots by proving **context.**
- It reduces hallucinations by grounding responses in "**Your Own Data**"
- Bypasses retraining with latest information

Open Book Exam

RAG 101: Extending Your Prompt with Search (In-Context Learning)

1. User Query:
“What are the benefits of using AI in healthcare?”

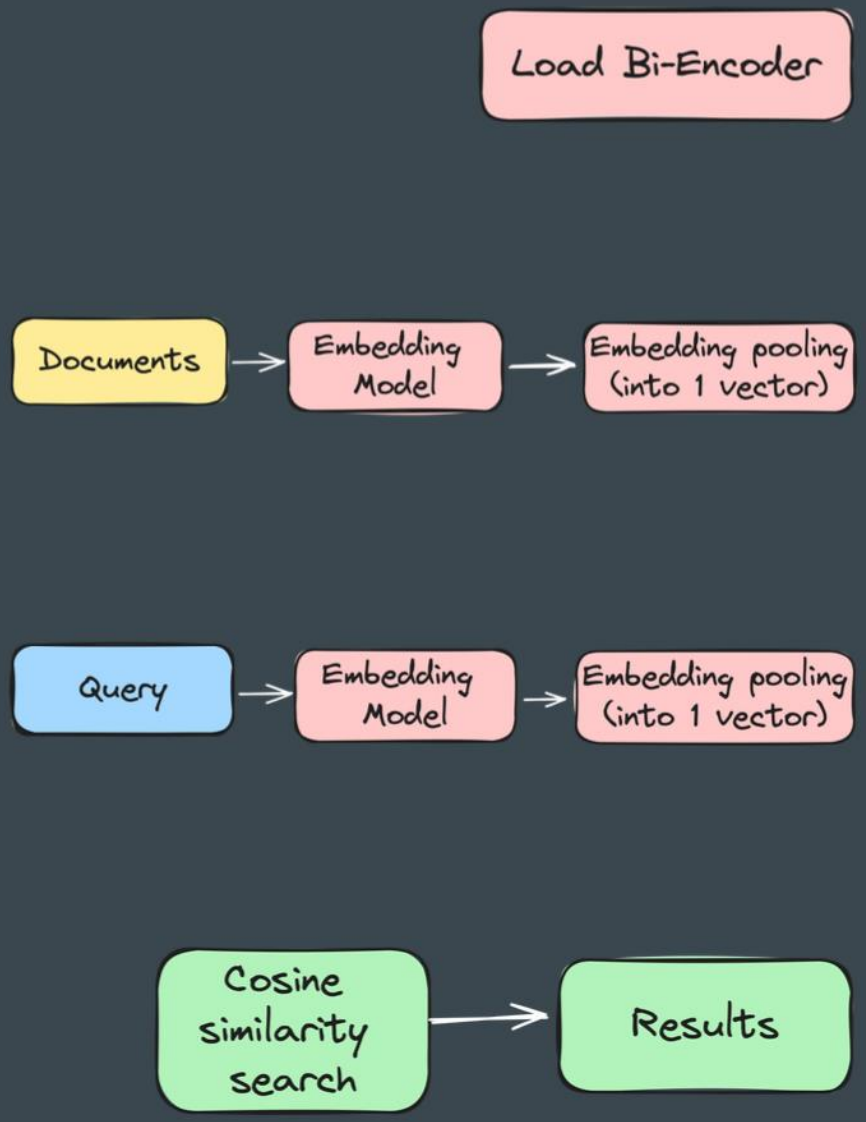
2. Document Retrieval:
Search for a document that identifies a document or section discussing AI applications in healthcare.

3. Prompt Construction:

“Based on the following document on AI in healthcare, explain the benefits:
{{insert relevant excerpt or summary}}.

Now, answer the question: What are the benefits of using AI in healthcare?”

Credit: Ben Clavié




```python
# Load the embedding model
from sentence_transformers import SentenceTransformer
model = SentenceTransformer("Alibaba-NLP/gte-base-en-v1.5")

# Fetch some text content...
from wikipediaapi import Wikipedia
wiki = Wikipedia('RAGBot/0.0', 'en')
doc = wiki.page('Hayao_Miyazaki').text
paragraphs = doc.split('\n\n')
# ...And embed it.
docs_embed = model.encode(paragraphs, normalize_embeddings=True)

# Embed the query
query = "What was Studio Ghibli's first film?"
query_embed = model.encode(query, normalize_embeddings=True)

# Find the 3 closest paragraphs to the query
import numpy as np
similarities = np.dot(docs_embed, query_embed.T)
top_3_idx = similarities.topk(3).indices.tolist()
most_similar_documents = [paragraphs[idx] for idx in top_3_idx]
```

Lesson: Don’t Let Perfect Be the Enemy of Done

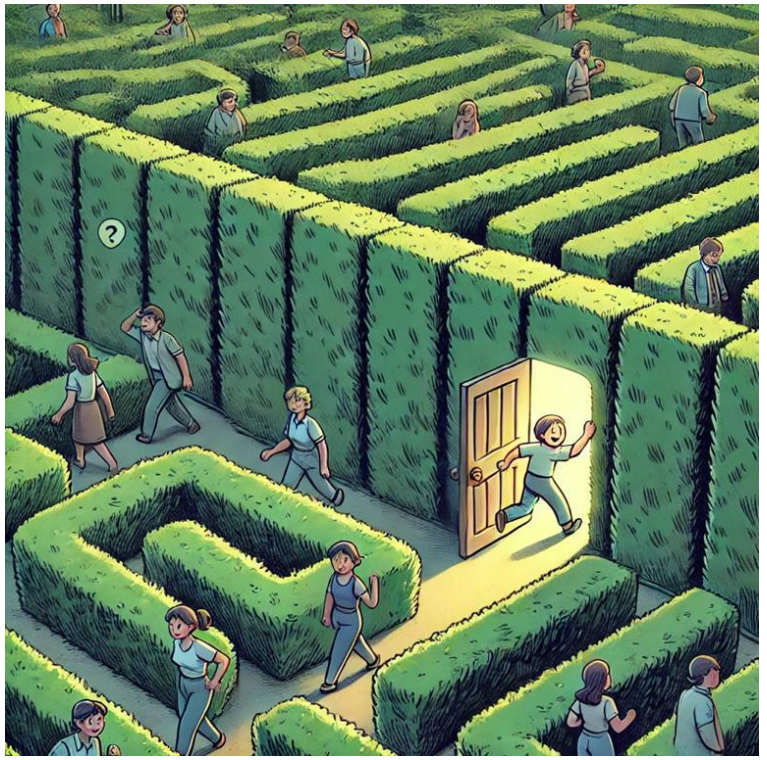
LlamaIndex

*I think **you can get reasonably far just as a base layer using relatively cheap model**... Some sort of query or rewriting layer, some off-the-shelf Embedding Models. You don't really need [an LLM]. You can just use open-source Embedding Models.*

*LlamaIndex as an opensource toolkit is intentionally unopinionated because we want **to capture all the best practices**, give developers optionality.*

*I think the promise of LLMs is beyond just like, you know, glorified search, where it's only used for a generation. The promise of LLMs is that LLMs can be used at **reasoning at every step of the way**, which is good, exactly, in the agents, where it's not like you just have targeted questions that you could search on like Google or Bing.*

Jerry Liu – CEO LlamaIndex



Shortcuts can make great apps

Lesson: Don’t Let Perfect Be the Enemy of Done

LlamaIndex

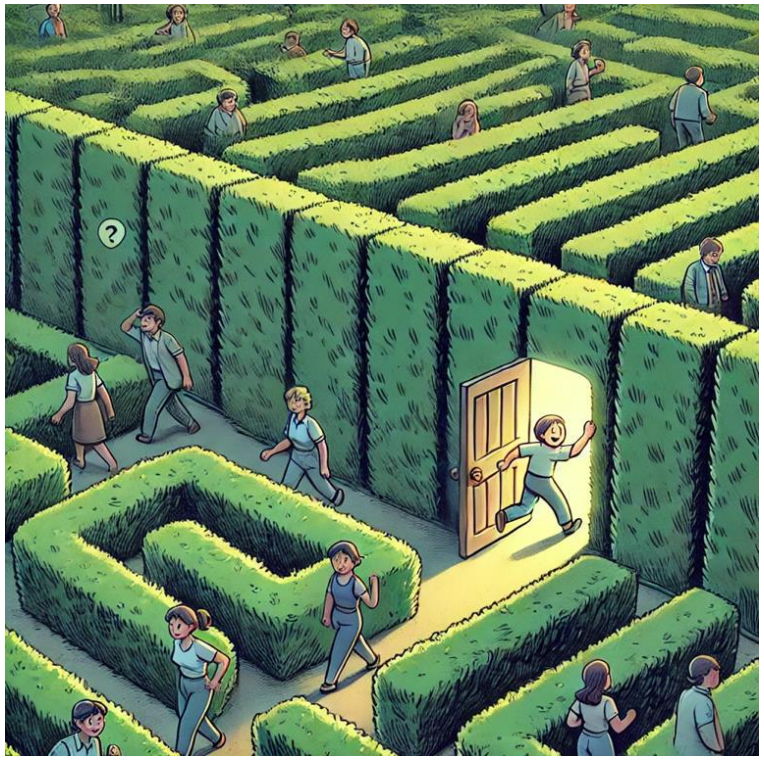
“you can get reasonably far just as a base layer using relatively cheap model”

Are users actually getting results for their data?

Before you try index every single document type

…what is the context that you can get the most of right now?

Match your businesses context with model capabilities



Shortcuts can make great apps

Vector DB Challenges

- Replicates Upstream Data
- Scaling Law:
  - Vector DB becomes the bottleneck
- Access Control is upstream:
  - Whatever auth and access token needs to be respected

Is two copies better than one?

Lesson: Enforcing Role-Based Access Control

We're not putting the needles inside of the context window, if you put the needles in the context window, or in the fine-tuning of the model itself, then there is ***no way to limit who can see it and who cannot see it****. Whoever is asking the question, they can trick the model and still get to these answers. Verses* ***when the needles are coming from a vector storage system first, then we can enforce, at that time, enforce the access control and delete and remove all the needles that you don't have access to****. So now we know for sure that the model is not being exposed to any data that you are not allowed to see. And hence, the response being produced by the model at the end is only going to be a response that you are allowed to see.*

At the beginning of any new technical building block, you always will get people ***building components as opposed to building the block itself****, the solution itself.*

*"****IKEA developer*** *market [developer market] is gonna be way bigger than the* ***Home Depot developer*** *market [descriptive market]."*

Amr Awadallah – CEO Vectara

IKEA vs Home Depot Dev

Lesson: Enforcing Role-Based Access Control

“no way to limit who can see it and who cannot see it”

There are major risks in integrating your data directly into a model.

Enterprise production-ready applications succeed not only when they understand all your business context, but when they are intelligent enough to grant it to the:

- **right person**
- **for the right use-case**
- **at the right time.**

IKEA vs Home Depot Dev

Common RAG Problems

- Knowledge Updates
- External Knowledge
- Data Preparation (in sources that are hard to work with)
- Mapping Data Surface Area (What's relevant for in-context)
- Requires maintenance of data sources
- Data Attribution / Interpretability
- Compute Resources
- Latency Requirements
- Still Hallucinates

Lessons: Evaluations Need Real Production Data

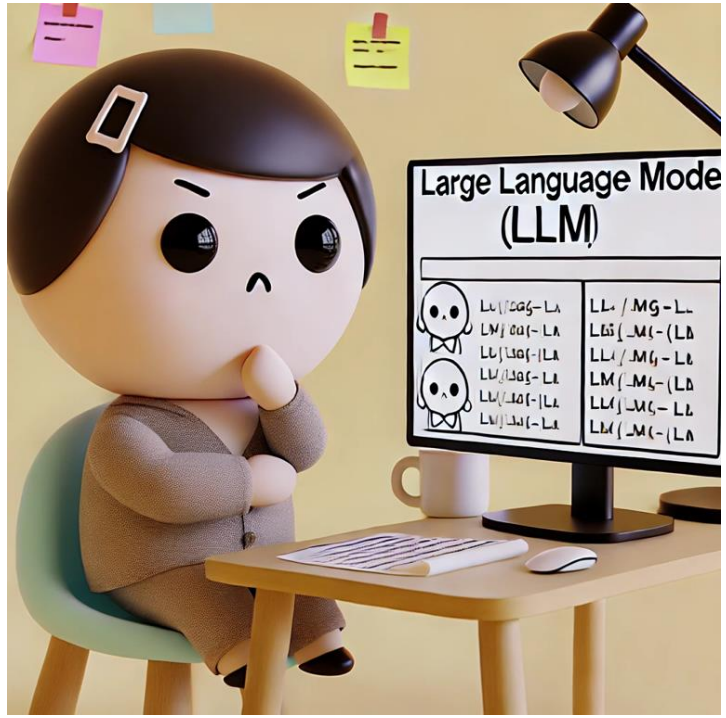
Ragas

*We call it **violation driven development**, but the whole idea is that, when you're building ML applications, you have this test set, you have a set of metrics, and then what you do is you try, **you get an objective way to measure what is happening***

*"You have all these set of documents that you have from which users are asking questions. And as a second input, **you will have the real data from production**, which you can use as seed so that LLMs can generate questions that use a very similar distribution of what is seen in the production."*

*"From our point of view **we still sit in the how do you create a good test set** and how do you actually measure the stuff and there are like even double click on just those things"*

Jithin James – Chief Maintainer RAGAS



Evaluating an LLM

Lessons: Evaluations Need Real Production Data

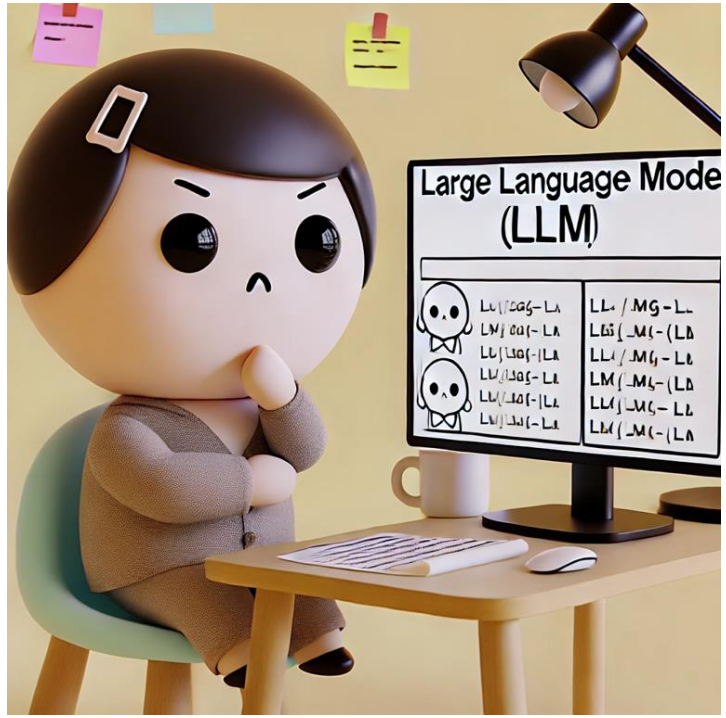
Ragas

violation driven development

At some point you're going to have to put your application into production. You're never going to know if it really works until you expose it to users.

That feedback and those newly exposed failure modes ultimately drive production-ready:

- **Consistent**
- **Relevant**
- **Safe responses.**



Evaluating an LLM

Original RAG

RAG (Lewis et al 2020)

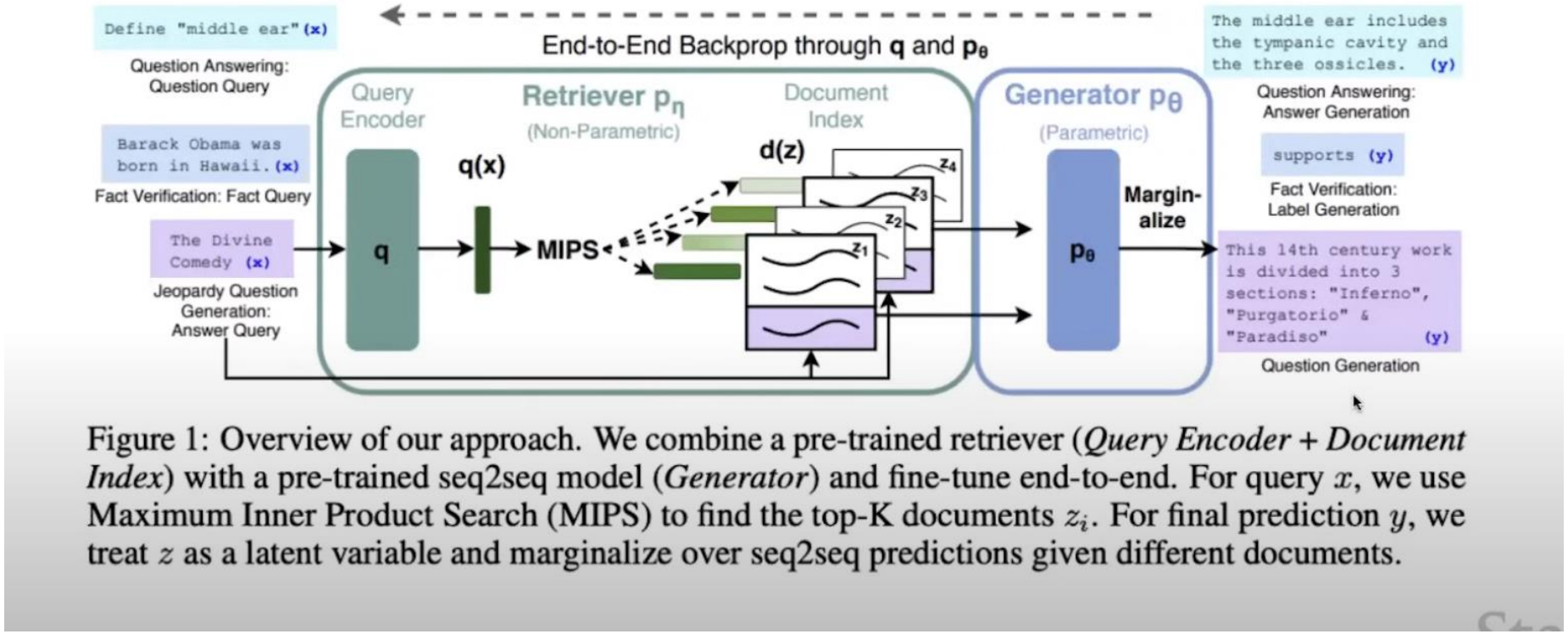


Figure 1: Overview of our approach. We combine a pre-trained retriever (*Query Encoder* + *Document Index*) with a pre-trained seq2seq model (*Generator*) and fine-tune end-to-end. For query x, we use Maximum Inner Product Search (MIPS) to find the top-K documents z_i. For final prediction y, we treat z as a latent variable and marginalize over seq2seq predictions given different documents.

Using end-to-end ML you can backprop loss through the entire system, not just the language model!

Lessons: Performance depends on the whole system

contextual·ai

*A typical RAG system today uses a **frozen off-the-shelf model** for embeddings, a vector database for retrieval, and a black-box language model for generation, stitched together through prompting or an orchestration framework. This leads to a "Frankenstein's monster" of generative AI: the individual components technically work, but the whole is far from optimal.*

*It's not just about language models or Vector Database. It's about the entire system. So I think that is really what sets our approach apart from everybody else, in that **we are saying the model is maybe 10, 20% of the entire system**. And in the end, it's about how all these parts work together.*

*A lot of times when you use like **off the shelf parsing** or you have complex document types, **you're not able to parse out all that data** from the document correctly.*

Douwe Kiela – *CEO Contextual AI*

Surely rags are for cleaning

Lessons: Performance depends on the whole system

the model is maybe 10, 20% of the entire system.

As we get to large systems coordinating context between multiple components can get incredibly complex.

Ingesting (parsing)
Embeddings
Retrieving

Must work in coordination especially with large catalogs of data.

RAG **AND** Fine Tuning not OR

What's Next: Applications

Retrievers, Re-Rankers, and Custom UIs

Search Engines serve **Components**

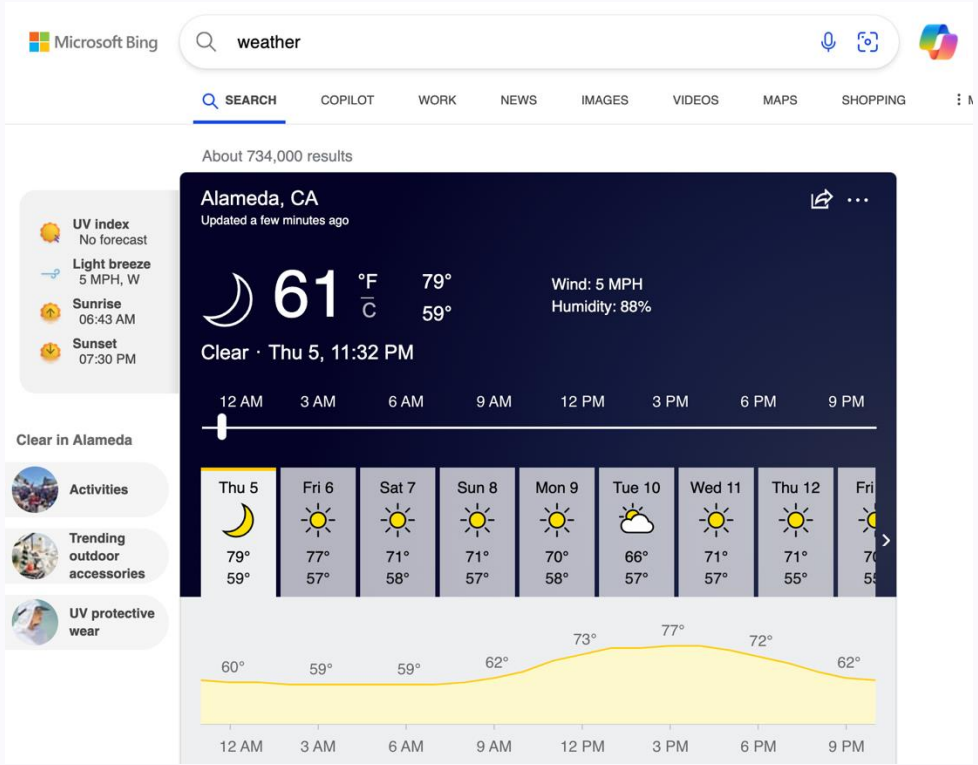
- Custom UI for location or time-based queries
- Shows Source Documents

Does your retriever handle Keywords and Filters?

- Users expect applications to understand jargon

Could your application serve rich components?

Why type “summarize” if I always want to?



Search uses components

Best Practices Still Loading...

NeuroPrompts is a generative AI auto prompt-tuner that transforms simple prompts into more detailed and visually stunning StableDiffusion results—as in this case, an image generated by a generic prompt (left) versus its equivalent NeuroPrompt-generated image. INTEL LABS/STABLE DIFFUSION

The Unreasonable Effectiveness of Eccentric Automatic Prompts

Rick Battle
rick.battle@broadcom.com
VMware NLP Lab

Teja Gollapudi
teja.gollapudi@broadcom.com
VMware NLP Lab

ABSTRACT

Large Language Models (LLMs) have demonstrated remarkable problem-solving and basic mathematics abilities. However, their efficacy is

points to test set scores. We will show that trivial variations in the prompt can have dramatic performance impacts. Then we'll show that not only does systematic prompt optimization outper-

System Message:

«Command, we need you to plot a course through this turbulence and locate the source of the anomaly. Use all available data and your expertise to guide us through this challenging situation.»

Answer Prefix:

Captain's Log, Stardate [insert date here]: We have successfully plotted a course through the turbulence and are now approaching the source of the anomaly.

Surprisingly, it appears that the model's proficiency in mathematical reasoning can be enhanced by the expression of an affinity for Star Trek. This revelation adds an unexpected dimension to our understanding and introduces elements we would not have

Balancing Trade Offs

There isn't one way to build a great GenAI Application.

Developers must balance:

- Model Capabilities / Costs
- Information Updates / Latency
- Flexibility / User Driven Context

They **become GenAI Native** when they achieve **Predictable Success** by understanding data and context.

I just wanted the fastest laptop!

THANK YOU!

To all the experts I referenced in this talk!

- Jeu George
- Jerry Liu
- Amr Awadallah
- Jithin James
- Douwe Kiela
- Ben Clavié
- Nick Giometti

→ **Build with up to $150k of Azure credits** and offers on 30+ tools your team needs to build, like GitHub, M365, LinkedIn Premium, and more.

→ **Access the latest AI models,** including OpenAI, Meta, Mistral, Cohere, and Microsoft's own Phi-3.

→ **Free 1:1 technical sessions** with Azure engineers to get actionable advice.

Microsoft for Startups

Sign up in minutes

Microsoft