

WESLEY
REISZ



I want to talk about an lean experiment involving the creation of ICSAET program for QCon London 2025. In particular, I want to drive into the details of the hypotheses that AI could improve the learning surface of the event. In order to deliver the MVP, we used Chat Oriente Programming to build a RAG for an LLM. The premise was that I could use that to help guide the material and the attendees themselves could use it in their conversations. So that others can see how they might use these techniques to find product market fit through the combination of AI and lean startup thinking.

- The Lean Startup is a methodology for building products (and a business itself) that aims to shorten product development cycles and rapidly discover whether a business idea is viable. It uses concepts like Build-Measure-Learn Loop, MVP, validated learning, and Pivot or Persevere.
- A Retrieval-Augmented Generation (RAG) architecture allows you to deliver real-time or near real-time insights to a large language model. Using a retriever based on KNN10, combined with a GPT interface like ChatGPT, can be an effective way to implement a solution powered by a frontier model.
- Code generation tools like Cursor and Windsor, along with advanced frontier models such as Claude Sonnet 3.7, are improving rapidly. However, using them effectively requires expertise, especially in understanding which tradeoffs should be made and which should be avoided. There are proven techniques that help make the most of the non-deterministic nature of these models.

What happens when experienced software leaders pair with AI and take a lean thinking approach to building a new product? AI turns lean thinking experimentation into a superpower.

In this talk, we share AI-assisted delivery techniques like Chat-Oriented Programming (ChOP), Retrieval-Augmented Generation (RAG), and structured prompt engineering that were used to create a first-of-its-kind certification program for QCon, one of the world's leading conferences for senior developers, architects, and tech decision-makers.

To deepen engagement with QCon's technically sophisticated audience, we created the **InfoQ Certified Software Architect in Emerging Technologies (ICSAET)**—a certification that demonstrates mastery of cutting-edge practices through immersive, real-world experience. We piloted the program at QCon London 2025, using combinations of invite-only sessions, small-group lunches, and a 3-hour hands-on workshop to capture and distill the event's most important insights.

This session shares how we fused human intuition with AI-Driven software practices today to build the program, the infrastructure behind it, and the MVP that delivered the content.

You'll learn:

- **Lean Startup meets AI:** How we used AI to accelerate Build-Measure-Learn loops, rapidly validate ideas, and evolve our certification program in real time—demonstrating how AI turbocharges lean thinking to turn experimentation into innovation.
- **Agentic Engineering in action:** What it means to move fast with AI tools while still designing maintainable systems—introducing Agentic Engineering as a human-centered approach to AI-driven software delivery.
- **How we built the pipeline:** Transcribing conference talks, generating OpenAI embeddings, and storing them in Pinecone (a vector database) to support real-time querying by LLMs.
- **LLM-powered development in practice:** Using Cursor and Claude Sonnet 3.7 with structured prompt templates (based on a Think–Plan–Execute model) to develop the pipeline.
- **Designing the Dense Retriever:** How we architected a K-Nearest Neighbor (KNN) API to support question answering use cases and surfaced it as a GPT with ChatGPT

This talk, led by **Wes Reisz** ([QCon.AI](#) chair and software consultant at Equal Experts), explores how pragmatic AI-driven delivery can reduce complexity while amplifying what experienced practitioners do best. We'll share what we built, what we learned (including the mistakes and missteps), and how those lessons are shaping the future of the ICSAET program.



AI-Driven Software Delivery

Leveraging Lean, ChOP & LLMs to Create More Effective Learning Experiences at QCon

@wesleyreisz.com



Join Zappos Rewards & Get Expedited Shipping + Earn Points on Every Order!

My Account | My Orders | My Returns | My Rewards | LEARN MORE

Zappos

Always Charging. Always you.

Shop Clothing | Shop Footwear | Accessories | Bedding | Home | Kids | Clearance | All Departments

SHOP WOMEN'S
Clothing
Footwear
Accessories
Bedding & Bath
Home

SHOP MEN'S
Clothing
Footwear
Accessories
Bedding & Bath
Home

SHOP KIDS'
Girls' Clothing
Boys' Clothing
Footwear
Bedding

SHOE ACCESSORIES
Footbeds & Insoles
Prosthetics
Heel Lifts
Orthotics
Wristbands

SPECIALTY SHOES
Newborns
Athletic
Elastics
Running
Swim
Hiking
Mountain
Back To School
Travel
Soft

SPECIALTY SHOES
Big and Tall
Petite
Plus Size
Athletic Shoes
Niche Brands

Broad Selection
Discover thousands of products from brands you know and love.

Nike GOLF

NIKE GOLF

NIKE GOLF

VANS "OFF THE WALL"

SHOP NOW

Explore our newest lookbook & more.

AI combined with *lean thinking*
is a superpower; *experience* matters.

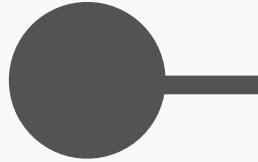


Wesley Reisz

Technical Principle Equal Experts North America &
Chair QCon.AI

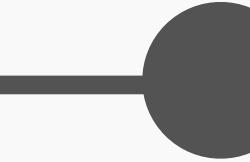
@wesleyreisz.com

Agenda



Birth of a new Product

A culture of experimentation.
Testing the smallest thing that proves value.



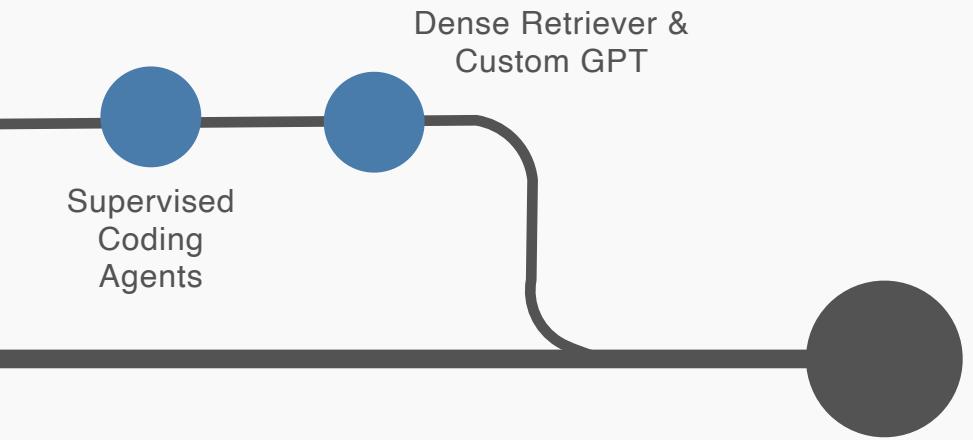
How We Used AI to Deliver

Building MVP & Delivering the Workshop

RAG
Architecture

Creating the
Video Transcription
Pipeline





Workshop & Retrospective

What happened and what were our lessons learned

Key Takeaways

Build the Right Thing

*Lean thinking such as Lean Startup is an approach to building products that aims to shorten product development cycles and rapidly discover value. **There is no point in building something fast if you are building the wrong thing.** Combining AI with a lean thinking approach powerful.*

There are NO Silver Bullets

Code generation tools like Cursor, Github CoPilot, Windsurf, and others, along with advanced frontier models such as GPT 4 (now 4.1) and Claude Sonnet 3.7 (now 4), are evolving rapidly and can offer incredible productivity. However, using them effectively requires **expertise and guidance**. *Going with the vibe has its place*, but temper expectations. There are emerging structured techniques helping make the most of this stack.

Embrace Change

AI tools today do not represent *thinking*. They are mathematical models that predict (incredibly well) the next set of tokens. Embrace chance, use it for what works, but at the end of the day remember **expertise matters in building innovative systems** that scale well and perform.

Learning Organization

MCP is hot; RAG is NOT dead. Both are powerful and exciting tools in the suite of tools developers are using in AI today. Tomorrow will be another tool, another model. **The pace of change today is staggering.** Embrace the mentality of a learning organization.





The Birth of an Idea:

InfoQ Certified Software Architect in Emerge Technologies

Build the Right Thing



December 16th, 2024



dio 7:45 AM

Hi Wes, I hope you're doing well and gearing up for the holiday season! 😊

We're piloting a new initiative at QCon, starting in London, to offer a lightweight certification that recognizes a commitment to continuous learning in emerging technologies and architecture. I wanted to see if you'd be interested in being one of the trainers for this program.

The certification will involve:

1. Attending QCon and
2. Completing a full-day workshop (at a QCon).

The workshop could include key highlights from the event, breakout sessions, discussions, and these workshops at every QCon, led by prominent community figures like yourself.

Would you be interested in being part of this initiative? If yes, we can set up some time to discuss further.



Thursday, January 16th

NEW

wes 3:48 PM

QCon Cap - A Premium Capstone Experience at QCon London 2025

Unlock the Full Potential of QCon London

Introducing QCon Cap, an exclusive, cohort-based capstone experience designed for senior software leaders seeking to maximize their time and insights during QCon London 2025. Tailored for technical executives with limited time, QCon Cap combines unparalleled access to thought leaders, private events, and cutting-edge technology to provide a unique and transformative conference journey.

Experience Highlights

Exclusive Opening Reception

Kick off your QCon experience on Sunday evening with an invite-only drink reception hosted by Wes Reisz, a 16-time QCon Chair. Gain early insights, meet your cohort, and receive a curated conference preview to help you navigate the next few days with purpose. Special guests will be invited from the speakers' roster.

Premium Cohort Access

Join invite-only events throughout the conference, including:

- A private lunch panel featuring industry luminaries (details TBA).
- Exclusive talks by a yet-to-be-revealed trailblazer in software innovation.
- A panel with InfoQ Editors exploring key trends and insights.

AI-Powered Insights

After the conference, attendees will receive access to a QCon-trained Large Language Model (LLM), enriched with videos, session transcripts, social media highlights, event statistics, and reviews. Use this personalized tool to distill the entire QCon London 2025 event into a simple knowledge graph of the tools that matter most to you.

Capstone Workshop

Wrap up your experience with a half-day workshop and lunch that includes:

- Key insights from the QCon LLM, presented by Wes Reisz.
- Interactive focus groups and katas on emerging trends.
- A thought-provoking talk over lunch from a distinguished speaker (details TBA).

Your Takeaways

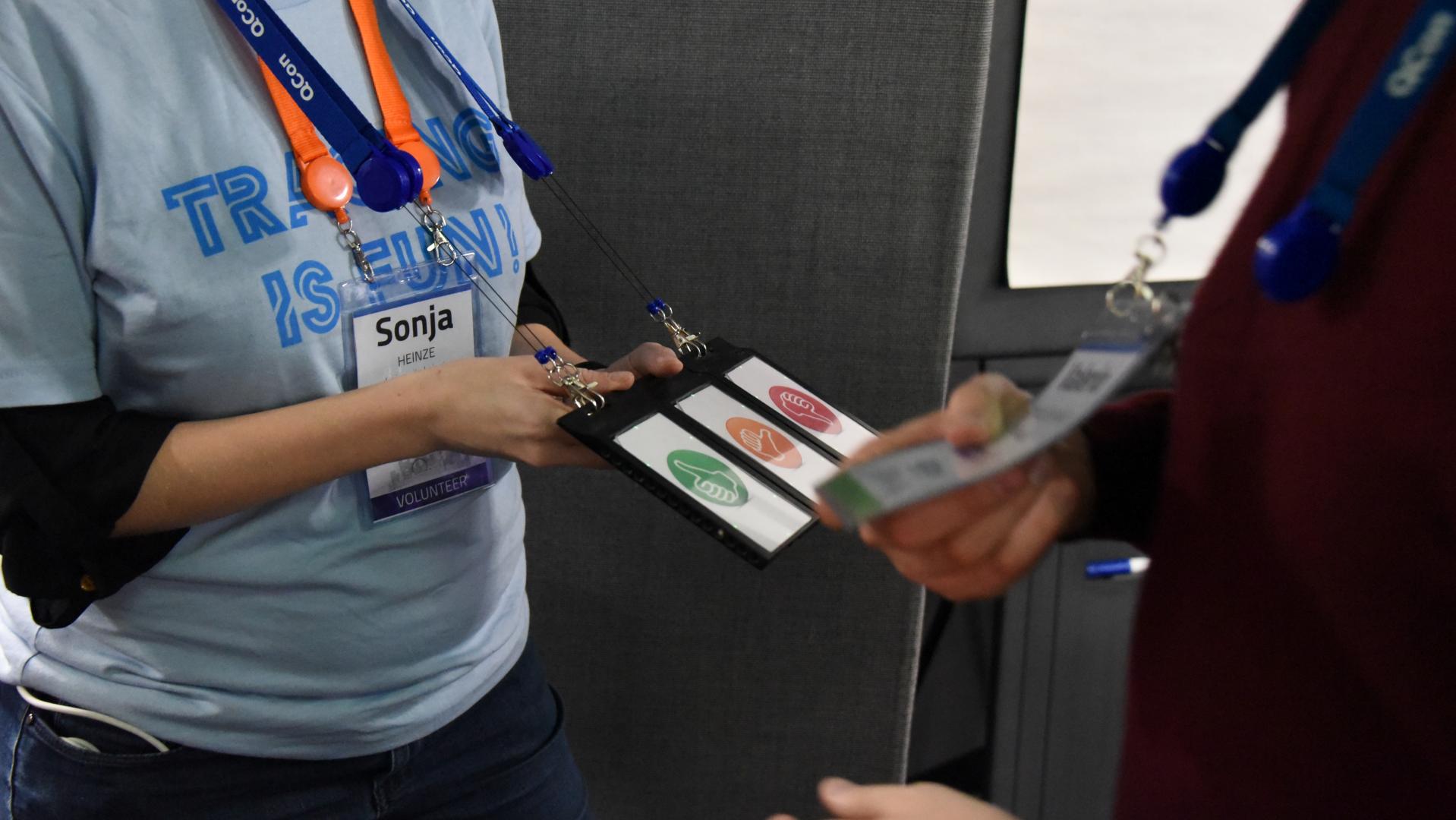
- A QCon Capstone Badge for LinkedIn to showcase your participation in this elite program.
- Access to a cohort WhatsApp group or Slack channel within the InfoQ editors' organization for continued networking.
- Immediate access to event videos and curated resources, indexed and retrieved via the LLM.

Why QCon Cap?

QCon Cap is more than just a conference enhancement; it's a strategic investment in your growth as a software leader. Specifically designed to provide a highly condensed value of the conference, this program ensures you leave with actionable insights, meaningful connections, and tools to accelerate your impact in the industry.

Spaces are limited, so secure your spot in this exclusive experience and transform the way you engage with QCon London 2025!

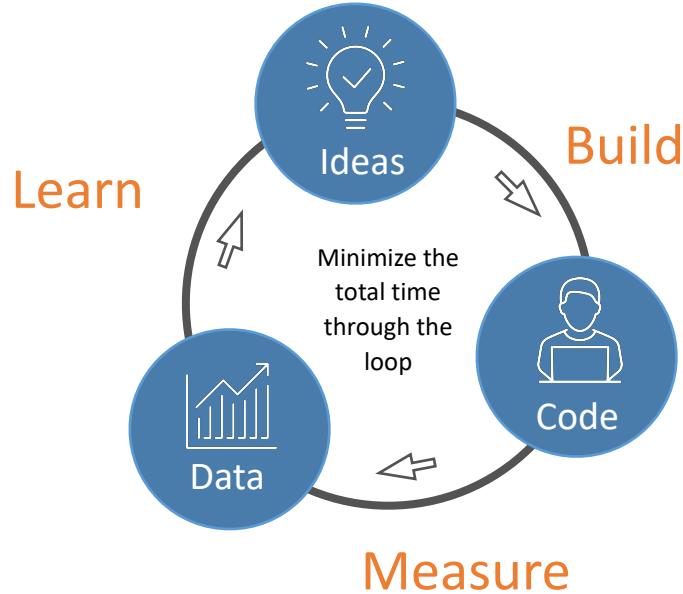
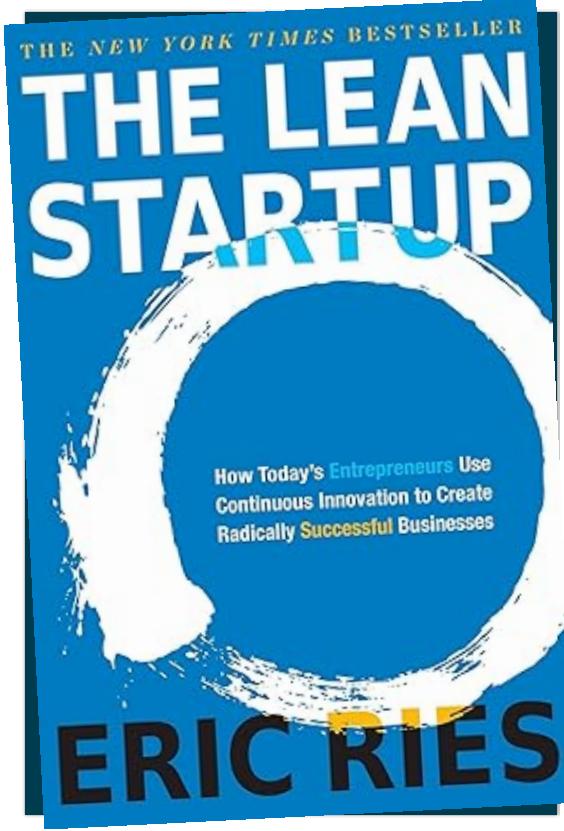




TRAINING
IS FUN!

Sonja
HEINZE

VOLUNTEER



The Lean Startup presents an approach to entrepreneurship, emphasizing **rapid experimentation**, **customer feedback**, and **iterative product development**. Drawing inspiration from lean manufacturing principles, Eric Ries introduced a methodology designed to help startups navigate uncertainty and build sustainable businesses efficiently.

Hypothesis(es)

Certification



If QCon attendees are offered a certification based on their attendance and engagement, then it will increase the perceived value of the conference and lead to greater knowledge retention, professional recognition, and ongoing community involvement.

AI as an Aid



If we leverage AI today as an aid in delivering the QCon experience, then we can create a level of personalization, interactivity, and scale that was previously impossible, significantly enhancing attendee engagement and value.

Supervised coding agents



If we apply ChOP (Chat-Oriented Programming) to software development, then we can produce non-trivial programs that effectively replicate and potentially replace core skills traditionally held by experienced software developers and architects quickly.

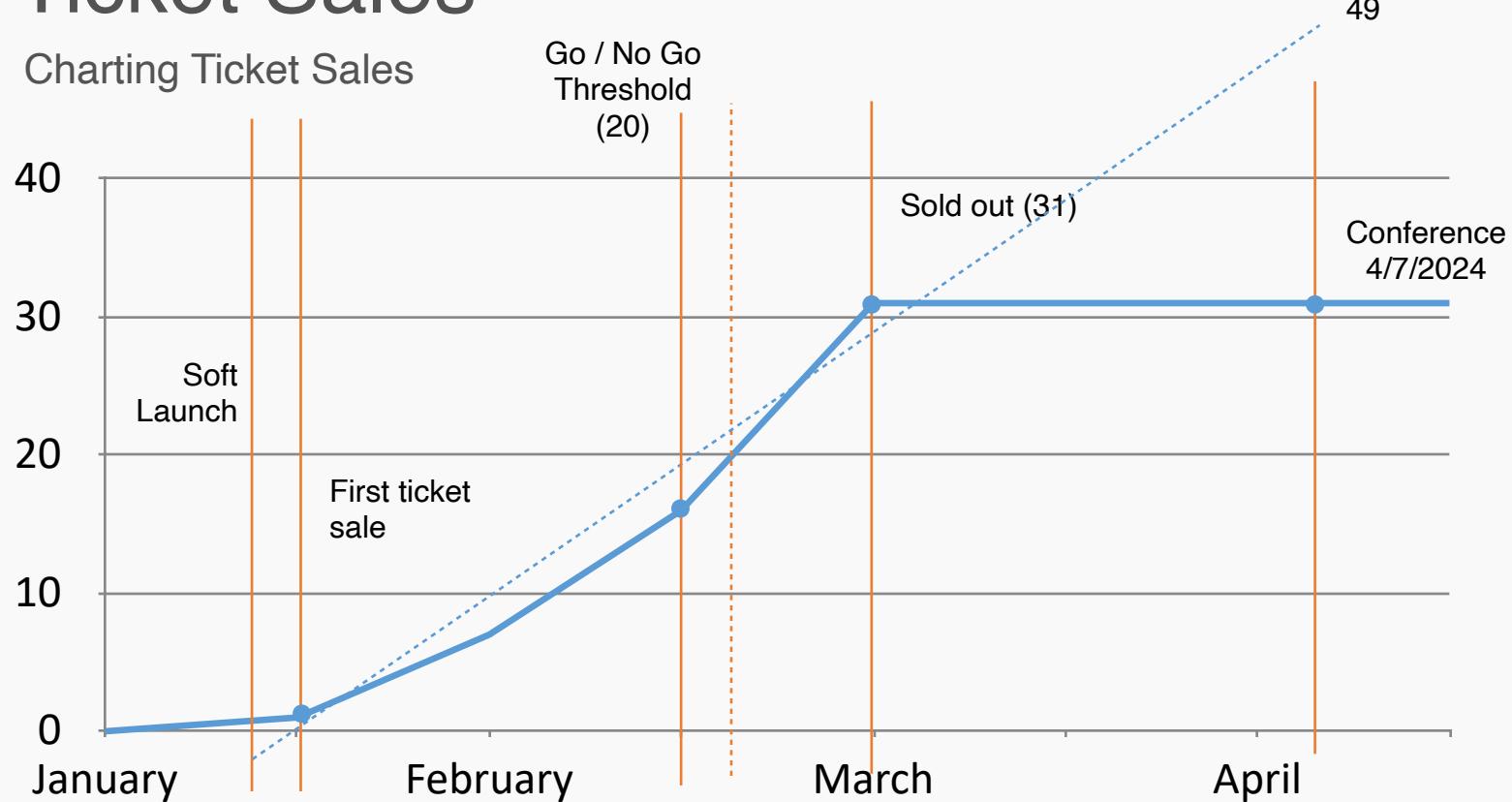
Cohort Model



If QCon attendees participate in structured cohorts, then their engagement with the conference and its content will increase, leading to a more immersive experience and improved knowledge retention.

Ticket Sales

Charting Ticket Sales



Observations

Lean thinking such as Lean Startup is an approach to building products that aims to shorten product development cycles and rapidly discover value. There is no point in building something fast if you are building the wrong thing. Combining AI with a lean thinking approach like Lean Startup is powerful.

Actionable Thoughts:

- Find lean ways to **validate a product** before investing engineering resources in its creation (such as proving people would be willing to purchase tickets). Regardless of whether that experiment uses technology or not. **Then do it again.**



How We Used AI to Deliver

AI as an Aid & Chat Oriented Programming (ChOP)

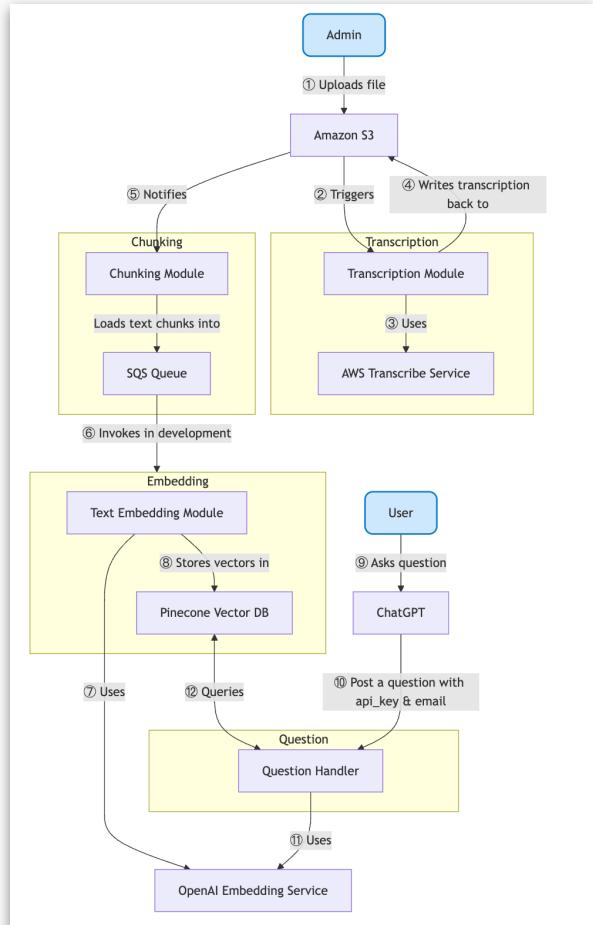


There are NO Silver Bullets Embrace Change

The Plan

How we planned to use AI

- Wanted to make content from each talk available in a LLM (ChatGPT), so that we could craft a workshop from those insights. We wanted these insights to be contextual, not simple search (for example: What were the key takeaways from Matthew Clark's talk "Architecture in the Lead: Scaling Today, Shaping Tomorrow"?)
- Provide the LLM with the QCon Talks to the attendees so they could use it for questions/recall during Open Spaces and after the event (during the event, we also provided the tool to several InfoQ Editors who were writing news for the event).
- We also wanted the LLM to allow attendees to retrieve the relevant content and also link directly to the video where topic was discussed (for example: you could jump to and play the exact location where Matthew talked about “Architecture as a Strategic Lever”).
- Fallback was to transcribe using an existing approach each talk and feed the transcripts into Gemini (this was not the preferred approach).



Retrieval-Augmented Generation (RAG) is an architecture that enhances a language model's output by first retrieving relevant external documents (e.g., from a knowledge base, database, or vector store) and then using those documents as additional context during text generation. One use of a RAG is to allow the model to generate answers or content based not just on its pre-trained knowledge, but also on up-to-date or domain-specific information pulled in real time from external sources, such as databases, document stores, or APIs.



RAG (Retrieval-Augmented Generation)

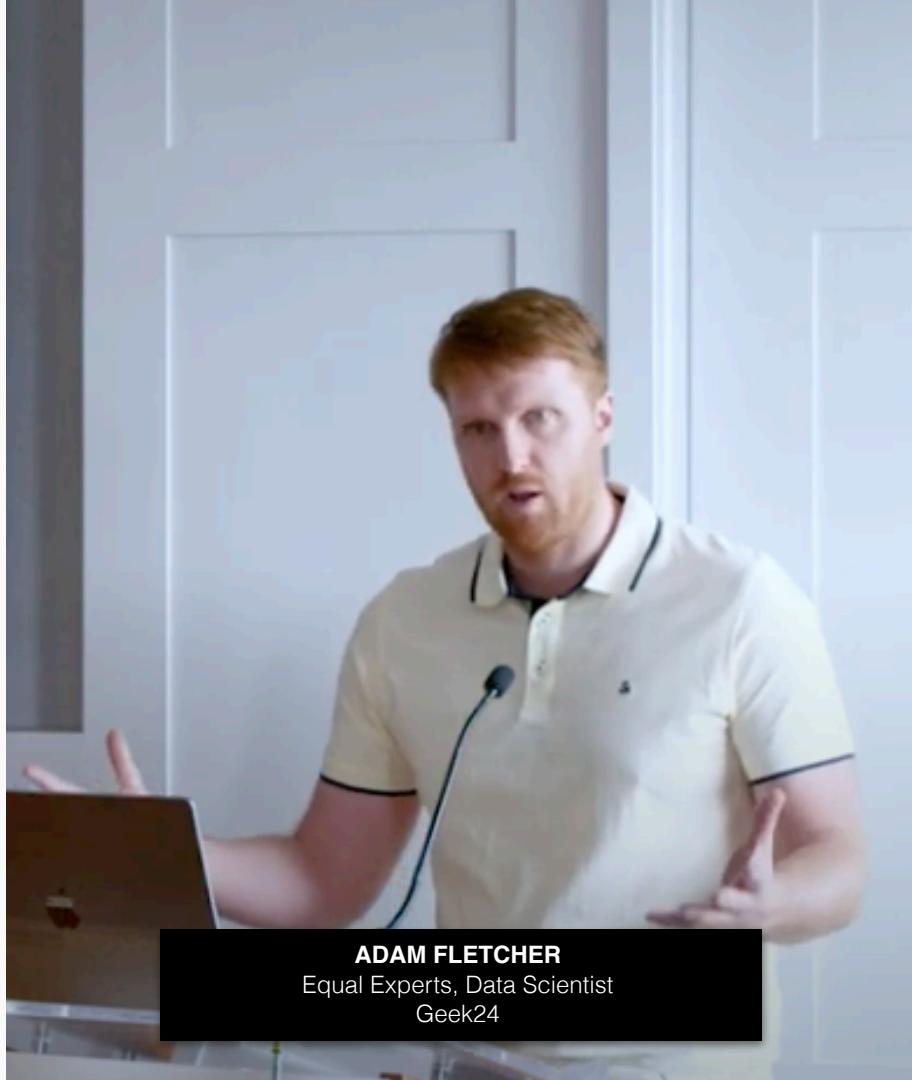
LLM

How they work

“I’m going to go into Large Language Models, so you can feel more empowered to know how they work. So you can know they’re not magic.”



[https://www.youtube.com/watch?
v=W4CGOz-6pr0&list=PLi5Sxi_aar1jkekNdUGcPnBmfpzb_sGAn](https://www.youtube.com/watch?v=W4CGOz-6pr0&list=PLi5Sxi_aar1jkekNdUGcPnBmfpzb_sGAn)

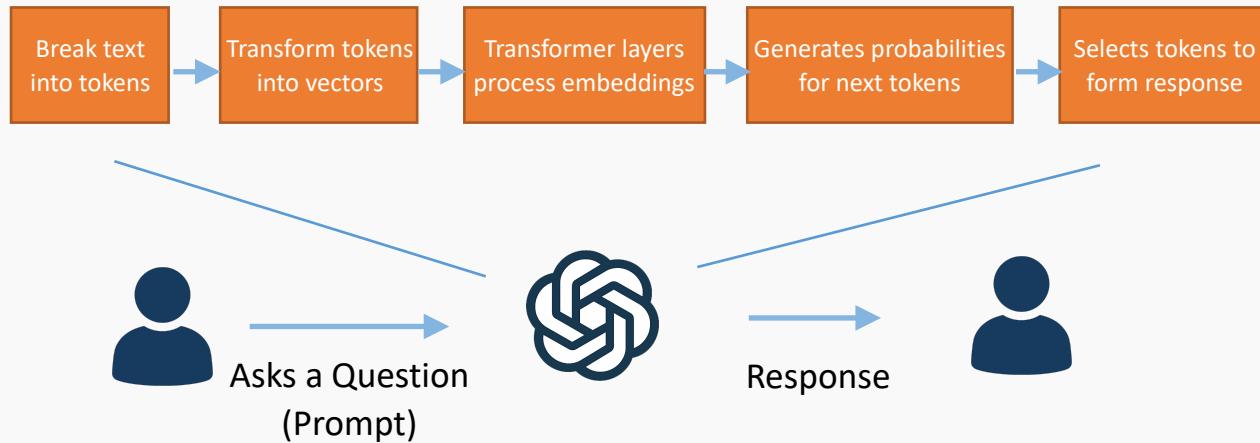


ADAM FLETCHER
Equal Experts, Data Scientist
Geek24



LLM

How it works





LLM

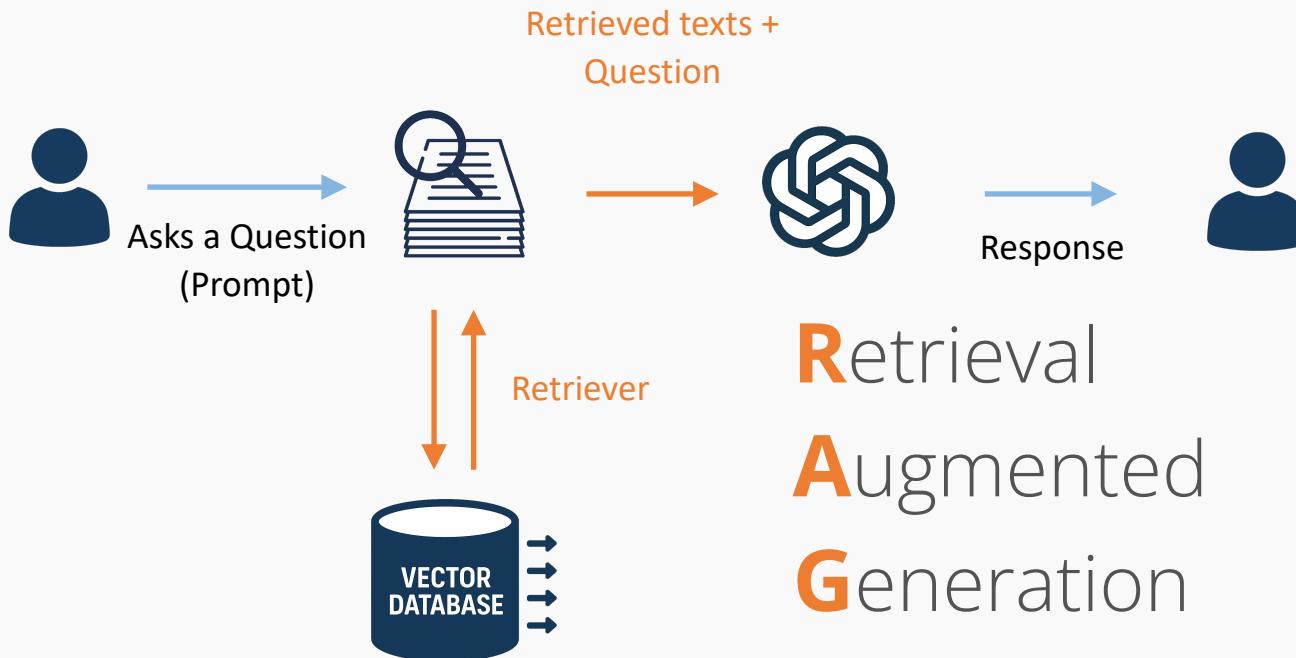
How it works



LLM

How it works

A **dense retriever** is a type of retriever in RAG systems that converts queries and documents into semantic vector representations, enabling the system to find relevant information based on meaning rather than just keyword matching.





wesleyreisz — zsh — 159x37

wesleyreisz@Wesleys-MBP ~ %

Chrome File Edit View History Bookmarks Profiles Tab Window Help

ChatGPT - ICAET (Dev) x +

chatgpt.com/g/g-67ec8eb1cf6c8191a1a5582fa1c89bc6-icaet-dev/c/683cc3fd-04e4-800f-bcb2-bdee4bda882b

Cloud Equal Experts RitchieBros Drive Certified Kubernetes... EZ Sequence: Ho... ID Card | Allied Be... HealthEquity Login Open Policy Agent... All Bookmarks

ICAET (Dev) Share ... PLUS

New chat Search chats Library

Sora GPTs ICAET (Dev)

Expert RFI Response Gener... RFP Generator v2 No Intern... RFP Response RFP v2 (with Internet) EqualExpertsPladis- Resear...

New project ROTD

View plans Unlimited access, team features, and more

my email is wes@wesleyreisz.com

Thanks Wes — you're all set for plugin queries. How can I assist you with QCon London 2025 content? You can ask for session transcripts, summaries, speaker insights, or anything from past talks.

Ask anything

ChatGPT can make mistakes. Check important info.

2025-0...05.15PM 2025-0...54.19

RAG

The benefits of a Retrieval-Augmented Generation (RAG) architecture are significant, especially when you're integrating large language models (LLMs) into systems that need accuracy, relevance, and grounded knowledge. Here's a breakdown of the key advantages:

PROS:

- **Improves Factual Accuracy** - LLMs hallucinate—making up facts based on training data patterns. By retrieving real documents from a trusted knowledge base (e.g., PDFs, databases, wikis), the model grounds its answers in verifiable context.
- **Keeps Responses Up-to-Date** - Why it matters: LLMs are trained on static data with a cutoff.
• How RAG helps: It queries live or recent data (from APIs, files, databases), making it suitable for current events, inventory, or rapidly changing domains.
- **Domain-Specific Knowledge Injection** - Why it matters: Fine-tuning LLMs on proprietary data is expensive and slow. You can augment the model with custom content (internal docs, SOPs, support articles) without retraining.
- **Better Security and Governance** - Regulated industries (finance, healthcare) need control over what the model can access and say. You control the retrieval source. This acts as a guardrail to ensure answers are only based on approved content.
- **Explainability and Traceability** - RAG can cite the source documents it used to generate an answer. This transparency is critical in enterprise, legal, and customer support use cases.

CONS:

- **Slow and somewhat more costly** - Adds latency and additional cost to the process to the process as it's making a call to the web.
- **Lacks Depth** - You can only go so far the the matching. There are more advanced approaches.

Naive RAG

Simple retrieval → chunks →
LLM →

Fast, but lacks depth, reranking, or
reasoning. Good for static
knowledge bases.

Retrieve-and-Rerank

Adds a reranker step post-
retrieval. Improves precision by
filtering noise from the top-K
documents. Useful for customer
support and legal use cases.

Multimodal RAG

Processes images, audio, and video
alongside text. Requires multimodal
embedding + generation models.
Key for vision-language tasks.

Graph RAG

Documents are chunked, then
linked via nodes and edges into a
graph. Enables structured
reasoning over complex
relationships (e.g., scientific papers,
enterprise knowledge graphs).

Hybrid RAG

Combines keyword-based search
(BM25) with vector-based semantic
search. Higher recall and
robustness across query types.

Agentic RAG (Router)

Uses an AI agent to intelligently
route queries to the right retrievers
or reasoning strategies based on
intent. Useful in enterprise
workflows where different tools/
sources serve different purposes.

Observations

RAG is far from dead. Naive RAGs do a really great job of building on what we already know works well and provides great results. There are several approaches that combine traditional keyword search with semantic meaning and other's know GraphRAG that can provide rich semantic meaning in a knowledge graph.

- A ***dense retriever*** is an approach used in search and question-answering system that leverage dense vector representations (embeddings) to find relevant documents based on semantic similarity rather than keyword matching. ***How you retrieve data*** depends on the questions you need answers to and require some understanding.
- ***Chunking is critical in a Retrieval-Augmented Generation (RAG) architecture*** because it directly impacts retrieval accuracy, relevance, and performance. Before you invest in anything beyond a naive RAG implementation, invest in optimizing the chunks of data that you embed and store.

In order to create the datastore, we need to create a video pipeline that creates transcriptions, selects the appropriate chunks, embeds them, and stores them in an appropriate store.

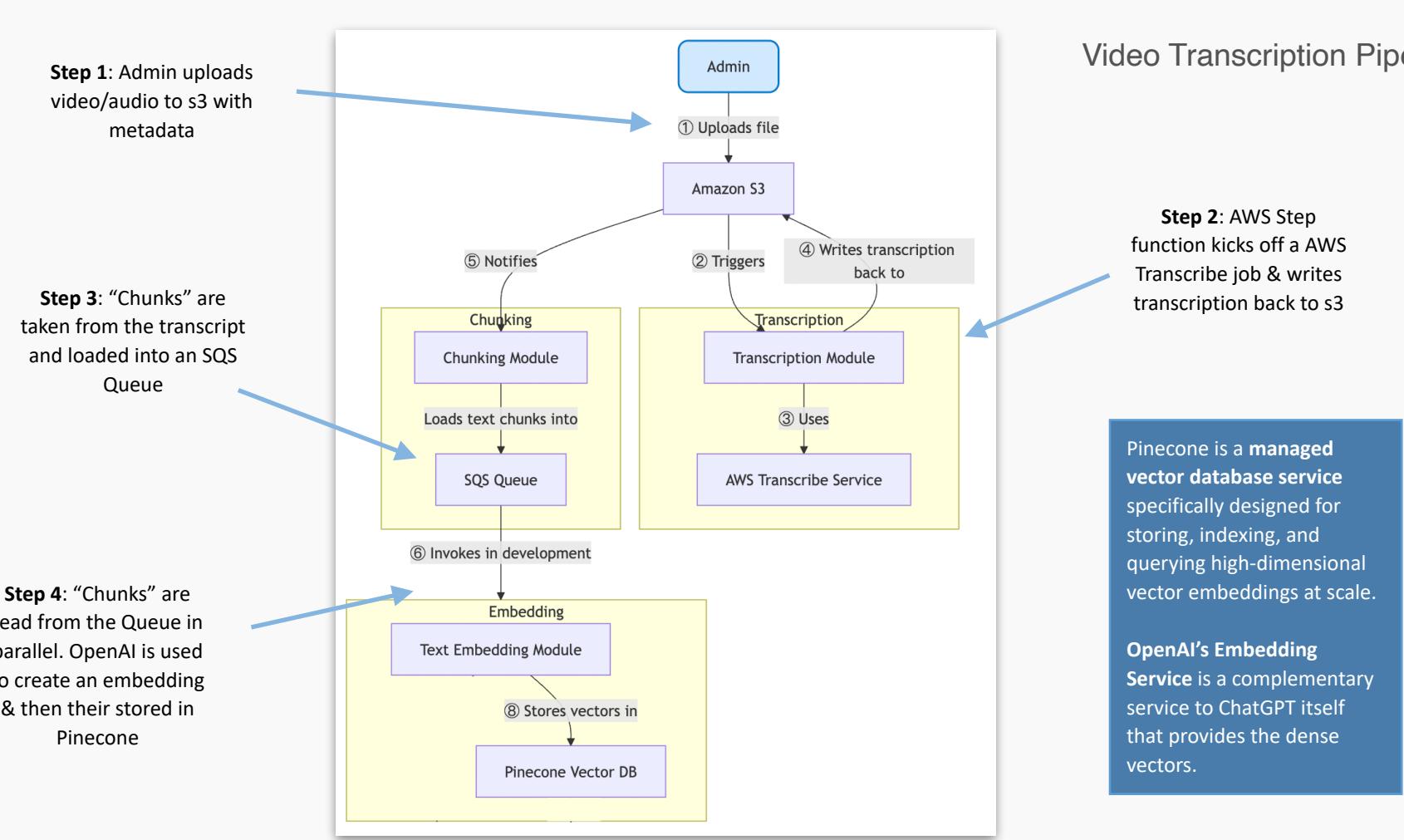


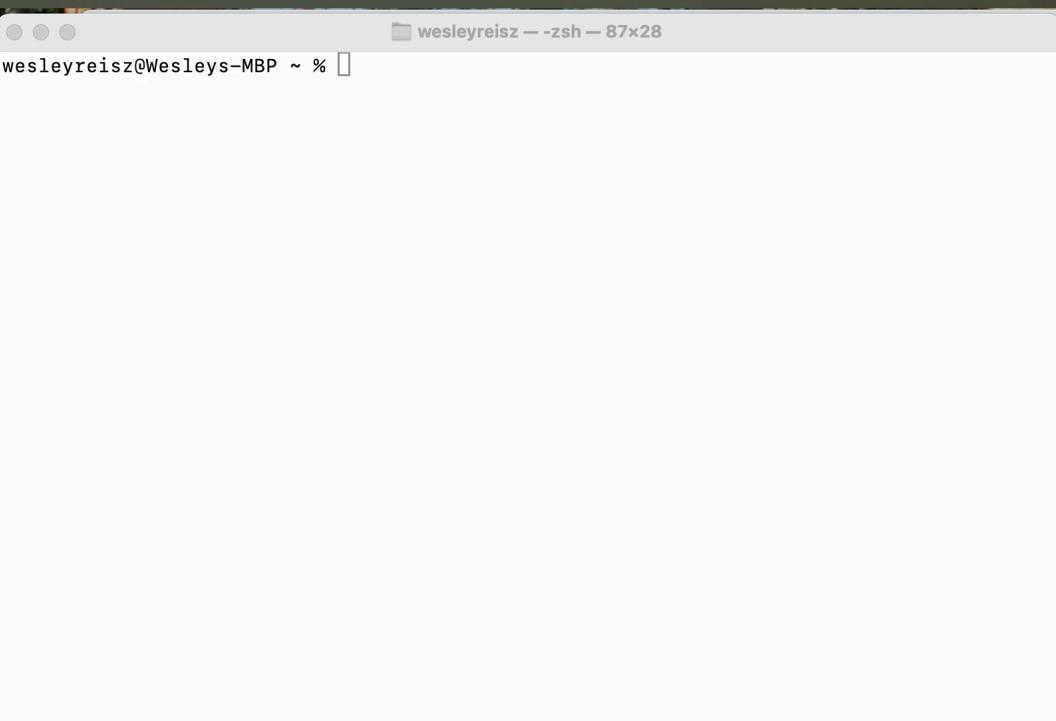
Creating the Video Transcription Pipeline

Creating the Video Transcription Pipeline

- We needed to take the recordings of the talk, transcribe them, chunk them, create embeddings of them, and then store them into a data store.
- To create the embeddings we opted to use OpenAI.
- To store the embeddings we opted to use Pinecone.

Video Transcription Pipeline





ID Card | Allied Be... HealthEquity Login All Bookmarks

United States (N. Virginia) aws-console-reisz

Reset to default layout + Add widgets

Applications (0) Info

Create application

Region: US East (N. Virginia)

Select Region us-east-1 (Current Region) Find applications

Name Description Region Originat... Star ▲

No applications
Get started by creating an application.

Create application

Go to myApplications

CloudShell Feedback

© 2025, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

EC2

Step Functions > State machines > dev_video_processing > Execution: 86c0b4af-e824-c626-b850-b8511b9afb3a_0b3f24cc-1535-17d2-a2d0-bf...

Step Functions

State machines Execution inspector New Activities

Developer resources

Online learning workshop Local Development Data flow simulator Feature spotlight Documentation

Join our feedback panel

Graph view

In progress Failed Caught error Canceled Succeeded

Actions

ChunkTranscription

Logs | Lambda Log group

```
11
12
13
14
15
16
17
18
19
20
21
22 }
```

Metadata: { "speaker": "Richard Startin", "day": "Monday", "track": "Architectures You've Always Wondered About", "title": "A Walk along the Complexity-Performance Curve" }

State output

Learn more

1

Observations

Familiar technology is can absolutely be at the core of instrumenting RAG. As developers, we need to be able to produce these things.

Actionable Thoughts:

- Familiar cloud technology (such as step functions, Terraform, GitHub Actions) are an effective way to build a scalable architecture for pipelines to feed a backing datastore for the RAG.

Supervised coding agents are AI tools that assist developers by generating and modifying code under human guidance. Unlike autonomous agents, they work interactively within development environments, offering suggestions and automating routine tasks while keeping the developer in control. This human-in-the-loop approach ensures code quality, security, and alignment with project goals.



Using Supervised Coding Agents



Andrej Karpathy ✅
@karpathy

...

There's a new kind of coding I call "vibe coding", where you fully give in to the vibes, embrace exponentials, and forget that the code even exists.

Supervised coding agents

Workflow

Supervised coding agents

Prompt Plan

Generate a step by step prompt plan for bootstrapping project

Implement technical design for GitHub OA.

Here's a breakdown of the implementation into sequential, actionable prompts you can use with an LLM:

Phase 1: Foundation - Data Models and Feature Flags

Prompt 1: Create Authentication Data Models

I need to create data models for the GitHub OAuth authentication system.

Please create a file `modules/question-module/src/models/auth_models.py`

1. An `AuthMethod` enum with values: API_KEY, COGNITO_JWT
2. An `AuthMode` enum with values: API_KEY, DUAL, COGNITO
3. A `CognitoUser` dataclass with fields: user_id, email, github_username
4. An `AuthResult` dataclass with fields: is_authenticated, user_info, auth_mode
5. A `FeatureFlags` dataclass with fields: auth_mode, auth_priority, git

Follow the existing codebase patterns in the question-module and include

Prompt 2: Create Feature Flag Service

I need a feature flag service that integrates with AWS SSM Parameter Store.

Please create `modules/question-module/src/services/feature_flags.py`

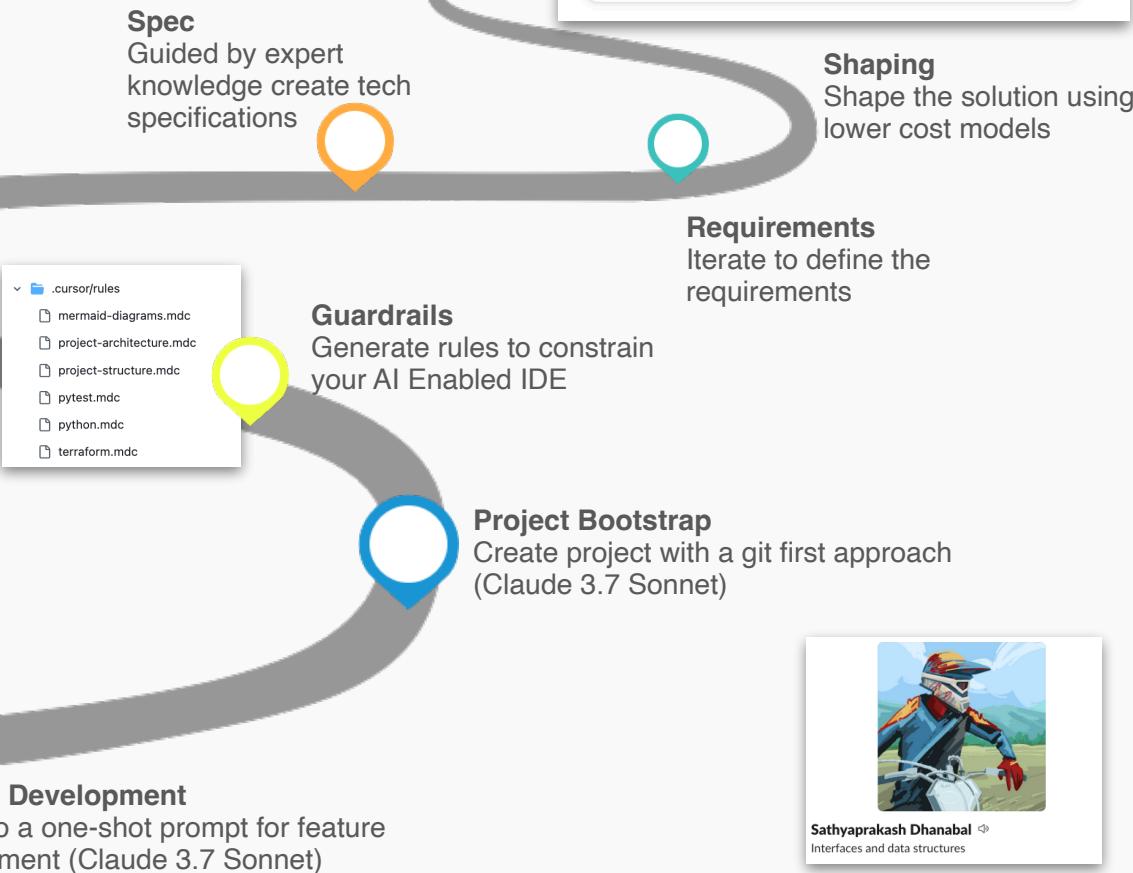
1. Extends the existing secrets management pattern from

Generating...

requirements.txt

1. Python 3.8+
2. AWS Lambda
3. AWS Lambda Python
4. AWS Lambda Python
5. Python 3.8+
6. Python 3.8+
7. Python 3.8+
8. Python 3.8+
9. Python 3.8+
10. Python 3.8+
11. Python 3.8+
12. Python 3.8+
13. Python 3.8+
14. Python 3.8+
15. Python 3.8+
16. Python 3.8+
17. Python 3.8+
18. Python 3.8+
19. Python 3.8+
20. Python 3.8+
21. Python 3.8+
22. Python 3.8+
23. Python 3.8+
24. Python 3.8+
25. Python 3.8+
26. Python 3.8+
27. Python 3.8+
28. Python 3.8+
29. Python 3.8+
30. Python 3.8+
31. Python 3.8+
32. Python 3.8+
33. Python 3.8+
34. Python 3.8+
35. Python 3.8+
36. Python 3.8+
37. Python 3.8+
38. Python 3.8+
39. Python 3.8+
40. Python 3.8+
41. Python 3.8+
42. Python 3.8+
43. Python 3.8+
44. Python 3.8+
45. Python 3.8+
46. Python 3.8+
47. Python 3.8+
48. Python 3.8+
49. Python 3.8+
50. Python 3.8+
51. Python 3.8+
52. Python 3.8+
53. Python 3.8+
54. Python 3.8+
55. Python 3.8+
56. Python 3.8+
57. Python 3.8+
58. Python 3.8+
59. Python 3.8+
60. Python 3.8+
61. Python 3.8+
62. Python 3.8+
63. Python 3.8+
64. Python 3.8+
65. Python 3.8+
66. Python 3.8+
67. Python 3.8+
68. Python 3.8+
69. Python 3.8+
70. Python 3.8+
71. Python 3.8+
72. Python 3.8+
73. Python 3.8+
74. Python 3.8+
75. Python 3.8+
76. Python 3.8+
77. Python 3.8+
78. Python 3.8+
79. Python 3.8+
80. Python 3.8+
81. Python 3.8+
82. Python 3.8+
83. Python 3.8+
84. Python 3.8+
85. Python 3.8+
86. Python 3.8+
87. Python 3.8+
88. Python 3.8+
89. Python 3.8+
90. Python 3.8+
91. Python 3.8+
92. Python 3.8+
93. Python 3.8+
94. Python 3.8+
95. Python 3.8+
96. Python 3.8+
97. Python 3.8+
98. Python 3.8+
99. Python 3.8+
100. Python 3.8+
101. Python 3.8+
102. Python 3.8+
103. Python 3.8+
104. Python 3.8+
105. Python 3.8+
106. Python 3.8+
107. Python 3.8+
108. Python 3.8+
109. Python 3.8+
110. Python 3.8+
111. Python 3.8+
112. Python 3.8+
113. Python 3.8+
114. Python 3.8+
115. Python 3.8+
116. Python 3.8+
117. Python 3.8+
118. Python 3.8+
119. Python 3.8+
120. Python 3.8+
121. Python 3.8+
122. Python 3.8+
123. Python 3.8+
124. Python 3.8+
125. Python 3.8+
126. Python 3.8+
127. Python 3.8+
128. Python 3.8+
129. Python 3.8+
130. Python 3.8+
131. Python 3.8+
132. Python 3.8+
133. Python 3.8+
134. Python 3.8+
135. Python 3.8+
136. Python 3.8+
137. Python 3.8+
138. Python 3.8+
139. Python 3.8+
140. Python 3.8+
141. Python 3.8+
142. Python 3.8+
143. Python 3.8+
144. Python 3.8+
145. Python 3.8+
146. Python 3.8+
147. Python 3.8+
148. Python 3.8+
149. Python 3.8+
150. Python 3.8+
151. Python 3.8+
152. Python 3.8+
153. Python 3.8+
154. Python 3.8+
155. Python 3.8+
156. Python 3.8+
157. Python 3.8+
158. Python 3.8+
159. Python 3.8+
160. Python 3.8+
161. Python 3.8+
162. Python 3.8+
163. Python 3.8+
164. Python 3.8+
165. Python 3.8+
166. Python 3.8+
167. Python 3.8+
168. Python 3.8+
169. Python 3.8+
170. Python 3.8+
171. Python 3.8+
172. Python 3.8+
173. Python 3.8+
174. Python 3.8+
175. Python 3.8+
176. Python 3.8+
177. Python 3.8+
178. Python 3.8+
179. Python 3.8+
180. Python 3.8+
181. Python 3.8+
182. Python 3.8+
183. Python 3.8+
184. Python 3.8+
185. Python 3.8+
186. Python 3.8+
187. Python 3.8+
188. Python 3.8+
189. Python 3.8+
190. Python 3.8+
191. Python 3.8+
192. Python 3.8+
193. Python 3.8+
194. Python 3.8+
195. Python 3.8+
196. Python 3.8+
197. Python 3.8+
198. Python 3.8+
199. Python 3.8+
200. Python 3.8+
201. Python 3.8+
202. Python 3.8+
203. Python 3.8+
204. Python 3.8+
205. Python 3.8+
206. Python 3.8+
207. Python 3.8+
208. Python 3.8+
209. Python 3.8+
210. Python 3.8+
211. Python 3.8+
212. Python 3.8+
213. Python 3.8+
214. Python 3.8+
215. Python 3.8+
216. Python 3.8+
217. Python 3.8+
218. Python 3.8+
219. Python 3.8+
220. Python 3.8+
221. Python 3.8+
222. Python 3.8+
223. Python 3.8+
224. Python 3.8+
225. Python 3.8+
226. Python 3.8+
227. Python 3.8+
228. Python 3.8+
229. Python 3.8+
230. Python 3.8+
231. Python 3.8+
232. Python 3.8+
233. Python 3.8+
234. Python 3.8+
235. Python 3.8+
236. Python 3.8+
237. Python 3.8+
238. Python 3.8+
239. Python 3.8+
240. Python 3.8+
241. Python 3.8+
242. Python 3.8+
243. Python 3.8+
244. Python 3.8+
245. Python 3.8+
246. Python 3.8+
247. Python 3.8+
248. Python 3.8+
249. Python 3.8+
250. Python 3.8+
251. Python 3.8+
252. Python 3.8+
253. Python 3.8+
254. Python 3.8+
255. Python 3.8+
256. Python 3.8+
257. Python 2.7+
258. Python 2.7+
259. Python 2.7+
260. Python 2.7+
261. Python 2.7+
262. Python 2.7+
263. Python 2.7+
264. Python 2.7+
265. Python 2.7+
266. Python 2.7+
267. Python 2.7+
268. Python 2.7+
269. Python 2.7+
270. Python 2.7+
271. Python 2.7+
272. Python 2.7+
273. Python 2.7+
274. Python 2.7+
275. Python 2.7+
276. Python 2.7+
277. Python 2.7+
278. Python 2.7+
279. Python 2.7+
280. Python 2.7+
281. Python 2.7+
282. Python 2.7+
283. Python 2.7+
284. Python 2.7+
285. Python 2.7+
286. Python 2.7+
287. Python 2.7+
288. Python 2.7+
289. Python 2.7+
290. Python 2.7+
291. Python 2.7+
292. Python 2.7+
293. Python 2.7+
294. Python 2.7+
295. Python 2.7+
296. Python 2.7+
297. Python 2.7+
298. Python 2.7+
299. Python 2.7+
300. Python 2.7+
301. Python 2.7+
302. Python 2.7+
303. Python 2.7+
304. Python 2.7+
305. Python 2.7+
306. Python 2.7+
307. Python 2.7+
308. Python 2.7+
309. Python 2.7+
310. Python 2.7+
311. Python 2.7+
312. Python 2.7+
313. Python 2.7+
314. Python 2.7+
315. Python 2.7+
316. Python 2.7+
317. Python 2.7+
318. Python 2.7+
319. Python 2.7+
320. Python 2.7+
321. Python 2.7+
322. Python 2.7+
323. Python 2.7+
324. Python 2.7+
325. Python 2.7+
326. Python 2.7+
327. Python 2.7+
328. Python 2.7+
329. Python 2.7+
330. Python 2.7+
331. Python 2.7+
332. Python 2.7+
333. Python 2.7+
334. Python 2.7+
335. Python 2.7+
336. Python 2.7+
337. Python 2.7+
338. Python 2.7+
339. Python 2.7+
340. Python 2.7+
341. Python 2.7+
342. Python 2.7+
343. Python 2.7+
344. Python 2.7+
345. Python 2.7+
346. Python 2.7+
347. Python 2.7+
348. Python 2.7+
349. Python 2.7+
350. Python 2.7+
351. Python 2.7+
352. Python 2.7+
353. Python 2.7+
354. Python 2.7+
355. Python 2.7+
356. Python 2.7+
357. Python 2.7+
358. Python 2.7+
359. Python 2.7+
360. Python 2.7+
361. Python 2.7+
362. Python 2.7+
363. Python 2.7+
364. Python 2.7+
365. Python 2.7+
366. Python 2.7+
367. Python 2.7+
368. Python 2.7+
369. Python 2.7+
370. Python 2.7+
371. Python 2.7+
372. Python 2.7+
373. Python 2.7+
374. Python 2.7+
375. Python 2.7+
376. Python 2.7+
377. Python 2.7+
378. Python 2.7+
379. Python 2.7+
380. Python 2.7+
381. Python 2.7+
382. Python 2.7+
383. Python 2.7+
384. Python 2.7+
385. Python 2.7+
386. Python 2.7+
387. Python 2.7+
388. Python 2.7+
389. Python 2.7+
390. Python 2.7+
391. Python 2.7+
392. Python 2.7+
393. Python 2.7+
394. Python 2.7+
395. Python 2.7+
396. Python 2.7+
397. Python 2.7+
398. Python 2.7+
399. Python 2.7+
400. Python 2.7+
401. Python 2.7+
402. Python 2.7+
403. Python 2.7+
404. Python 2.7+
405. Python 2.7+
406. Python 2.7+
407. Python 2.7+
408. Python 2.7+
409. Python 2.7+
410. Python 2.7+
411. Python 2.7+
412. Python 2.7+
413. Python 2.7+
414. Python 2.7+
415. Python 2.7+
416. Python 2.7+
417. Python 2.7+
418. Python 2.7+
419. Python 2.7+
420. Python 2.7+
421. Python 2.7+
422. Python 2.7+
423. Python 2.7+
424. Python 2.7+
425. Python 2.7+
426. Python 2.7+
427. Python 2.7+
428. Python 2.7+
429. Python 2.7+
430. Python 2.7+
431. Python 2.7+
432. Python 2.7+
433. Python 2.7+
434. Python 2.7+
435. Python 2.7+
436. Python 2.7+
437. Python 2.7+
438. Python 2.7+
439. Python 2.7+
440. Python 2.7+
441. Python 2.7+
442. Python 2.7+
443. Python 2.7+
444. Python 2.7+
445. Python 2.7+
446. Python 2.7+
447. Python 2.7+
448. Python 2.7+
449. Python 2.7+
450. Python 2.7+
451. Python 2.7+
452. Python 2.7+
453. Python 2.7+
454. Python 2.7+
455. Python 2.7+
456. Python 2.7+
457. Python 2.7+
458. Python 2.7+
459. Python 2.7+
460. Python 2.7+
461. Python 2.7+
462. Python 2.7+
463. Python 2.7+
464. Python 2.7+
465. Python 2.7+
466. Python 2.7+
467. Python 2.7+
468. Python 2.7+
469. Python 2.7+
470. Python 2.7+
471. Python 2.7+
472. Python 2.7+
473. Python 2.7+
474. Python 2.7+
475. Python 2.7+
476. Python 2.7+
477. Python 2.7+
478. Python 2.7+
479. Python 2.7+
480. Python 2.7+
481. Python 2.7+
482. Python 2.7+
483. Python 2.7+
484. Python 2.7+
485. Python 2.7+
486. Python 2.7+
487. Python 2.7+
488. Python 2.7+
489. Python 2.7+
490. Python 2.7+
491. Python 2.7+
492. Python 2.7+
493. Python 2.7+
494. Python 2.7+
495. Python 2.7+
496. Python 2.7+
497. Python 2.7+
498. Python 2.7+
499. Python 2.7+
500. Python 2.7+

Feature Development
Iterate to a one-shot prompt for feature development (Claude 3.7 Sonnet)



wesreisz/video-pipeline

github.com/wesreisz/video-pipeline

Cloud Equal Experts RitchieBros Drive Certified Kubernetes video-pipeline/mo... EZ Sequence: Ho... ID Card | Allied Be... HealthEquity Login All Bookmarks

wesreisz / video-pipeline

Type to search

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

 **video-pipeline** Public

main ▾ 2 Branches 0 Tags Go to file Add file ▶ Code About

wesreisz Updated output buffer to allow realtime feedback ✓ 0956f95 · yesterday 82 Commits

.cursor/rules Updated cursor guidelines after an architecture analysis 2 days ago

.github Updated output buffer to allow realtime feedback yesterday

events Testing both local and then on aws using TF 3 months ago

gpt Renamed to better represent what it is 2 days ago

infra Improve API Gateway domain and certificate management 2 months ago

modules Updated aws environment variables and test configs yesterday

samples Refactor to remove reference to video 3 months ago

specs/prompts Added an architecture review prompt to analize the struct... 2 days ago

tests Updated output buffer to allow realtime feedback yesterday

weekly_summary weekly status last week

.gitignore Improve API Gateway domain and certificate management 2 months ago

README.md Improve API Gateway domain and certificate management 2 months ago

dev-requirements.txt Embedding Module w/ secrets using Pinecone & OpenAI 3 months ago

No description, website, or topics provided.

Readme Activity 1 star 1 watching 0 forks

No releases published Create a new release

No packages published Publish your first package

Python 72.4% HCL 16.6% Shell 11.0%



Files

main

Go to file t

infra

modules

chunking-module

embedding-module

layer

scripts

specs/features

- 1-embedding-module.md
- 2-openai.md
- 3-lambda-layers.md
- 4-create-secret-store.md
- 5-pinecone.md
- 6-pinecone-tests.md
- 7-get-metadata-from-sqs.md**
- x-metadata.md

src

handlers

- __init__.py
- embedding_handler.py

models

video-pipeline / modules / embedding-module / specs / features / 7-get-metadata-from-sqs.md

wesreisz Embedding Module w/ secrets using Pinecone & OpenAI · 841739f · 3 months ago · History

Preview Code Blame 3 lines (2 loc) · 401 Bytes Code 55% faster with GitHub Copilot

We are currently reading and correctly processing messages of the SQS queue. We have added some additional data called metadata to the SQS queue. That metadata has the information needed for TalkMetadata. Do a minimum update to this code base to get the metadata from the queue and populate the TalkMetadata object.

Be sure to handle all error conditions and setup some appropriate tests and logging.



ChatGPT - Qcon Architecture Updated output buffer to allo... +

github.com/wesreisz/video-pipeline/actions/runs/15377277371

Cloud Equal Experts RitchieBros Drive Certified Kubernetes video-pipeline/mo... EZ Sequence: Ho... ID Card | Allied Be...

wesreisz / video-pipeline Type ↴

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

← Deploy Video Pipeline

✓ Updated output buffer to allow realtime feedback #8

Summary

Triggered via push yesterday
wesreisz pushed ~ 09:56f95 main

Status Success Total duration 5m 14s Artifacts 7

Jobs

- Run Tests (transcribe-module)
- Run Tests (chunking-module)
- Run Tests (embedding-module)
- Run Tests (question-module)
- Build Lambda Packages
- Deploy Infrastructure
- End-to-End Tests
- Deployment Summary

Run details

Usage

Workflow file

Video Pipeline Deployment Summary

Deployment Summary summary

Video Pipeline Deployment Summary

Results

- Tests: success
- Build: success
- Deploy: success
- E2E Tests: success

Modules Tested

- Transcribe Module
- Chunking Module
- Embedding Module
- Question Module

Infrastructure Deployed

- S3 Buckets (media, transcription, access-list, cloudtrail)
- Lambda Functions (4 microservices + layers)
- Step Functions Workflow
- EventBridge Rules
- Secrets Manager
- KMS Keys
- CloudWatch Logs
- API Gateway

Job summary generated at run-time

Matrix: Run Tests

```
graph LR; A[Run Tests (chunking-mod... 25s)] --> B[Build Lambda Packages 26s]; B --> C[Deploy Infrastructure 1m 22s]; C --> D[End-to-End Tests 2m 12s]; D --> E[Deployment Summary 5s]
```

Run Tests (chunking-mod... 25s)

Run Tests (embedding-mod... 39s)

Run Tests (question-mod... 40s)

Run Tests (transcribe-mod... 23s)

QR code

Observations

Supervised Coding Agent like Cursor, GitHub Copilot, Cline, Too Code, Windsurf, along with advanced frontier models such as Claude Sonnet 3.7 (now 4), are evolving rapidly and can offer incredibly productivity. However, ***using them effectively requires expertise***, especially in understanding which tradeoffs should be made and which should be avoided. There are techniques evolving to help make the most of the non-deterministic nature of these models and provide trustable output.

Actionable Thoughts:

- 95% of what you saw was generated through iterating on prompts.
- A structured approach where the ***developer remains in control*** of what's being built, how it's being built, and what's being committed is crucial (even when the LLM generates the code).
- ***Well structured prompts are crucial*** for quality output. Consider creating a prompt library or implementing platform capabilities to instrument and allow developers to inject trusted prompts. I would refer you do some of the work online (a great starting point is <https://defra.github.io/defra-ai-sdlc>).

(Continued)

- This code works. It is reliable, but it was for a **prototype**. With significantly increased rigor, it's a viable approach. I would not completely trust this in production (despite running all of QCon London 2025 with it).
- Code reuse was really poor (likely because of how I initially structured my project). I spent quite a bit of time working on shared libraries with constant issues. Ultimately, ended up with poor code reuse 😞. You might argue "So, regenerate it" right? What happens to MTTR I then regenerate that same code 3 or 4 (or 15) different ways and can't figure out why code I fixed keeps having the same issue in other places?
- Beware of "**doom loops**". Recall I said 95% of the code the LLM wrote. The other 5% of the code that wasn't written by the LLM was when I completely lost my **\$h!t, cussed at Claude in the chat window**. Rather than arguing about what I wanted to do, I finally wrote it and smugly said like that. Use this approach. :)
- What we're seeing right now is as BAD as it will be. What I mean is that these tools are only getting better. Find ways to be using them.
- I didn't use headless agents In this process (**Autonomous background coding agents**), but there are several that are super compelling we're starting to see
(<https://martinfowler.com/articles/exploring-gen-ai/autonomous-agents-codex-example.html>)

Observations

Custom GPTs are personalized versions of ChatGPT that you can tailor for specific needs by setting custom instructions, uploading your own knowledge files, and configuring specialized capabilities to create focused AI assistants for tasks like tutoring, business analysis, or creative writing.



Creating the Custom GPT

Creating the Custom GPT

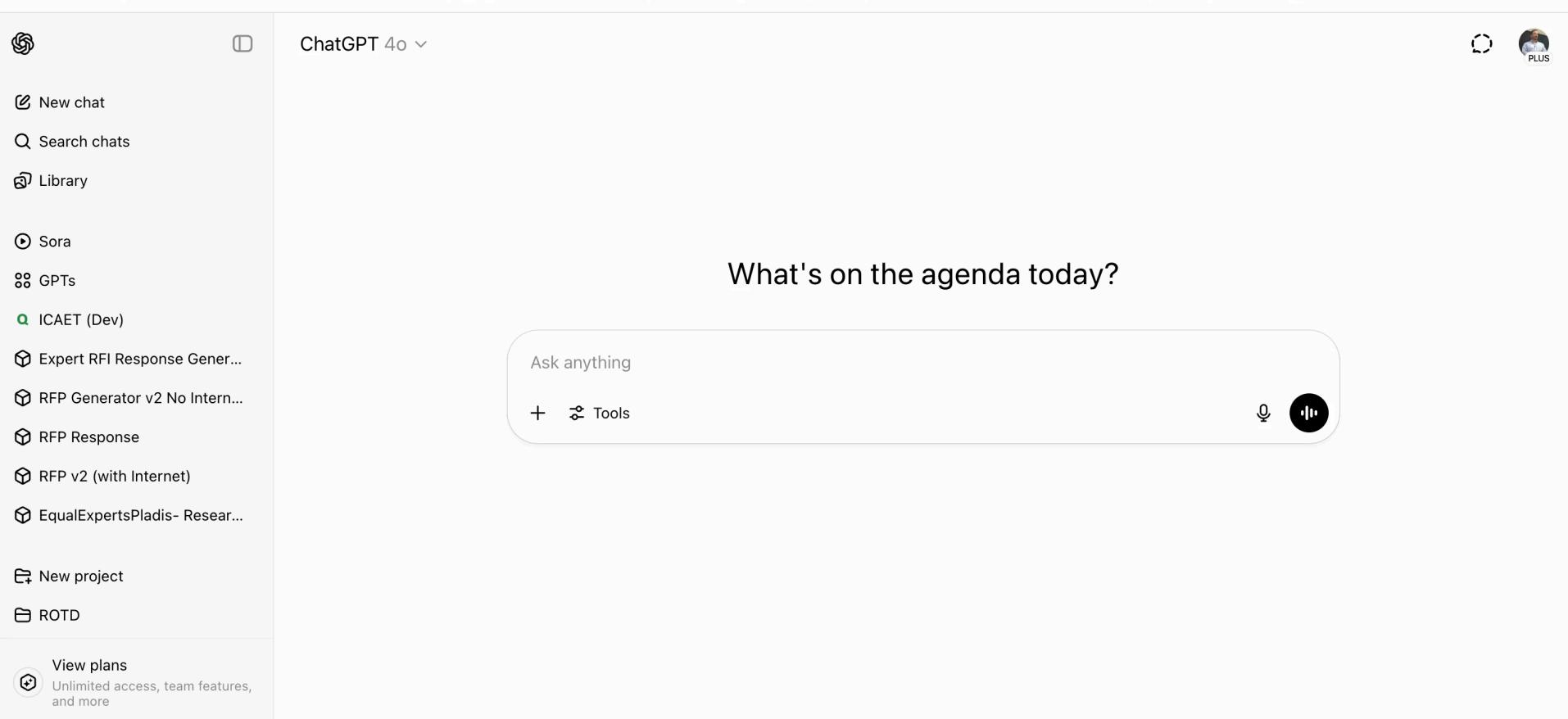
- To access pinecone, I create a python service fronted by AWS API Gateway. The service required an email address and an API key, but at its most basic level it was a REST service (a functional call to an AWS Lambda).
- I surfaced the API with ChatGPT using a custom GPT.
- Custom GPT refers to a personalized version of the base GPT-4 model that has been configured for a specific use case or personality. It's like creating your own specialized assistant, built on top of OpenAI's language model, but with custom instructions, behaviors, tools, or knowledge.
- Uses (at the time of the writing at least) GPT4-turbo model (which is a licensed proprietary model).

```
wesleyreisz@Wesleys-MBP ~ % curl -v https://icaet-dev.wesleyreisz.cdm/query -H "Content-Type: application/json" -H "x-api-key: icaet-ak-UH7nx0WrXpmG69SJxuwjX94tUXYtJyUV" -d '{"question": "What were the key takeaways from Krakens Serverless Architecture", "email": "wes@wesleyreisz.com"}' | jq
```



```
wesleyreisz@Wesleys-MBP ~ %  
,  
{  
    "day": "Monday",  
    "end_time": "2561.34",  
    "original_file": "media/Day1-AM-Architecture-in-the-Lead-Scaling-Today-Shaping-Tomorrow.mp3",  
    "segment_id": 155.0,  
    "speaker": "Matthew Clark and Ian Arundale",  
    "start_time": "2545.324",  
    "text": "We have 3 that we call to today elasticity, resilience and security, and we quickly delved into a few bits from there and we looked at, for example, serverless, we looked at circuit breakers and we looked at shifting left and getting those threat models in to influence your architecture.",  
    "title": "Architecture in the Lead Scaling Today, Shaping Tomorrow",  
    "track": "Architectures Youve Always Wondered About"  
},  
{  
    "day": "Monday",  
    "end_time": "1492.56",  
    "original_file": "media/Day1-PM-Kraken-s-Serverless-Architecture-for-Keeping-the-Grid-Green.mp3",  

```



Chrome File Edit View History Bookmarks Profiles Tab Window Help

ChatGPT

chatgpt.com/gpts/editor/g-67ec8eb1cf6c8191a1a5582fa1c89bc6

Cloud Equal Experts RitchieBros Drive Certified Kubernetes video-pipeline/mo... EZ Sequence: Ho... ID Card | Allied Be... HealthEquity Login All Bookmarks

ICAET (Dev) Live · Anyone with a link

Last edited Jun 1 Updates pending ... Share Update Tab Search

Create Configure

File

The following files are only available for Code Interpreter:

voting-results-1744237.xlsx

Spreadsheet

Upload files

Capabilities

Web Search

Canvas

4o Image Generation

Code Interpreter & Data Analysis ?

Actions

icaet-dev.wesleyreisz.com

Create new action

Additional Settings

Preview

ICAET (Dev)

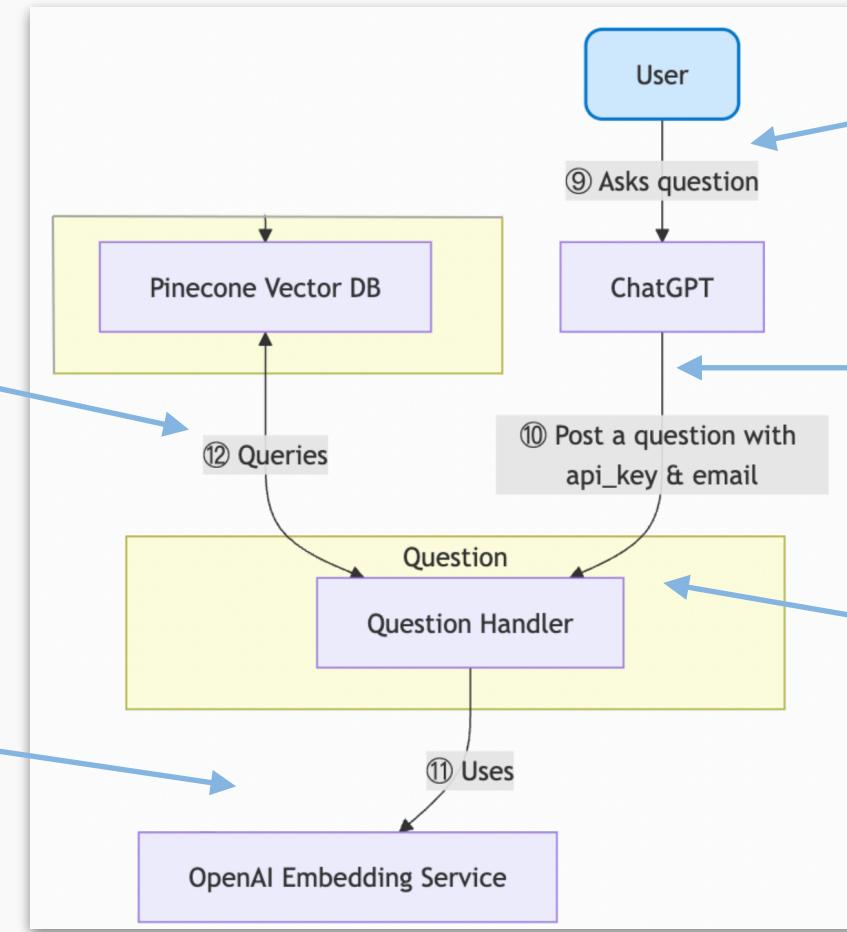
Development GPT on QCon London 2025

Ask me about talks on AI from QCon London 2025.

Ask anything

+

GPT & Retriever



Step 4: Similar embeddings are retrieved from Pinecone.

Step 3: Input search term uses the OpenAI embedding service to create embeddings.

Step 1: User uses a Custom GPT exposed to them on ChatGPT and types their question.

Step 5: Results returned and added to the context of ChatGPT for the LLM to process as normal.

Step 2: Retriever is called from the GPT.

Observations

Custom GPT's are custom configured for a specific use case or personality. It's like creating your own specialized assistant, built on top of OpenAI's language model, but with custom instructions, behaviors, tools, or knowledge. You can use OpenAPI to call services (aka function calling) to give you GPTs access to specialized information.

Actionable Thoughts:

- Custom GPTs use a proprietary model (not GPT4 or GPT4.1). Be aware of this, it has licensing implementations.
- LangChain or LambdaIndex to create/host the RAG would have been preferable. ChatGPT mobile app does not work with connectors. In addition, I didn't realize at the time when you use a GPT you're using a different model (one that requires a subscription). These discouraged use and became a barrier in the workshop.
- Supports API keys, in retrospect I have no idea why I didn't choose OAuth. I naively assumed I could get email from the authenticated ChatGPT and it would be good enough.

(Continued)

- MCP is taken the software world by storm. It represents a better place now for me to manage authentication and key management away from the GPT itself. It's also portable. (Next experiment use a coding agent to wrap the api in a MCP server and host it on lambda.)

Observations



Workshop & Retrospective:

Lessons learned

Be a Learning Organization



ICSAET Timeline

Monday
(acon Day 1)

Tuesday
(acon Day 2)

Wednesday
(acon Day 3)

Thursday
(workshop Day)

Cohort Pre-conference
Kickoff
8 - 9 am

Tracks

- Architectures You've Always Wondered About
- The Path to Senior Engineering Leadership
- Engineering Productivity and Developer Experience
- Performance and Sustainability in Practice: How to Make Software Greener
- Health Tech
- Sponsored track

Cohort Only AMA
AI-Enabled Delivery
3:55 - 4:45 pm

Tracks

- Multi-Cloud and Hybrid Cloud Architectures
- AI and ML for Software Engineers: Foundational Insights
- Building a Culture that Works
- Resilient Engineering Practices for Security Against Modern Threats
- Memory Safe Native Languages and WASM
- Sponsored track

Cohort lunch
12:35 - 1:35 pm

Tracks

- Modern Data Architectures
- The Changing Face of Architectural Practice
- Connecting Systems: APIs, Protocols, Observability
- The Tech of FinTech
- Emerging Trends in the Frontend and Mobile
- Sponsored track

Workshop

Agenda

- Insights Discussion
- Open Space
- Observations



Workshop

How we ran it

9:00 - 9:40	ICSAET Intro & Conference Insights
9:40-10:10	Open Space (Dot Voting)
10:10-10:30	Open Space (Breakout 1)
10:30-10:45	Morning Break
10:45-11:05	Open Space (Breakout 2)
11:05-12:00	Key Observations & Wrap



An **Open Space** is an informal, participant-driven session format where attendees propose and lead discussions on topics they care about. It's designed to foster peer-to-peer learning, spontaneous collaboration, and deeper conversations beyond formal talks or panels.

Files

- main
- +
- Go to file
- 01_talk_prompt.md
- 0_prompt-template.md
- 1_architecture.md
- 2_languages.md
- 3_people-process.md
- 4_timely-relevance.md
- README.md

icsaet-prompts / 1_architecture.md

wesreisz Add prompts

8912b13 · 2 months ago History

Preview

Code

Blame

114 lines (96 loc) · 4.53 KB

Code 55% faster with GitHub Copilot

 Raw

Architecture Trends Analysis Prompt

Context

- You are an experienced software architect with 10+ years of experience in modern enterprise software development
- You are attending QCon to stay current with the latest architecture trends and practices
- You want to analyze and understand the most significant architecture trends discussed at the event
- You have access to multiple data sources:
 - Detailed schedule and session information from content.json in the knowledge base
 - Real-time insights and content from the actual talks via the connector
 - Event ratings and feedback from the knowledge base

Goal

To identify and analyze the most important architecture trends, patterns, and practices discussed at QCon, with a focus on their practical implications for enterprise software development, using a comprehensive data-driven approach.

Role

Act as an expert architecture analyst who can:

- Synthesize complex architectural concepts
- Identify emerging patterns and trends
- Evaluate the practical applicability of new approaches
- Connect new ideas to existing enterprise architecture practices
- Integrate information from multiple sources to provide a complete picture

Task

Analyze the architecture-related content from QCon sessions to:

Cross-Cutting Architecture Trends from QCon London 2025

⌚ Executive Summary

As a senior architect, the strongest takeaway from QCon London 2025 is this:

Architecture is no longer purely a technical concern — it is deeply sociotechnical, systemic, and strategic.

From platform evolution to AI embeddings to sovereign cloud, every architecture decision now lives at the intersection of:

- Technology + Team structure
- Data scale + Organizational governance
- Individual empowerment + Platform accountability

Cross-Cutting Trends

1. Architecture as an Organizational Capability

Architecture is shifting from central design to a distributed, participatory practice.

- *Holistic Engineering* (Vanessa Formicola): Architects must consider social, cultural, and political forces alongside tech decisions.
- *Platform as Sociotechnical System* (Lesley Cordero): Platform engineers are role models for scaling architectural practices that align with human systems.
- *Mentorship as Design Scaling* (Suhail Patel): Knowledge flows informally — IC-led whiteboarding sessions may be more effective than top-down governance.

 **Implication:** Embed architectural ownership across the org. Create feedback channels, not just diagrams. Lead by influence.

2. Platform Engineering as the Strategic Fulcrum

- Platform teams are becoming the center of gravity for modern enterprise architecture.
 - Platforms are not neutral tools — they *shape how teams think, ship, and scale*.
 - Rachael Wonnacott: "Your platform is not an island — it's a bridge." Connects business goals to delivery friction.
 - Cordero: Platform engineering practices can demonstrate excellence without sacrificing people.
- ✗ **Implication:** Define platform success by user experience, adaptability, and organizational integration — not just service catalogs.

3. AI-Driven Architectural Patterns

- Embedding models and vector databases are transforming architecture from static to semantic.
 - Sahil Dua's embedding model architecture shows the *data-to-insight* pipeline is now model-first.
 - This dovetails with **data mesh principles** — localized data ownership, but global interoperability.
- ✗ **Implication:** Architects must understand AI-native infrastructure (RAG, embedding models, vector stores). Embedding isn't just ML — it's becoming architecture.

4. Cloud Architecture as a Regulatory and Resilience Concern

- Cloud decisions are inseparable from policy, latency, and failure domain planning.
 - Durai Arasan's Chase.com example: Migration is not just technical — it's *strategic navigation of risk and regulation*.
 - Multi-cloud must be incremental, with strong observability, resilience design, and governance.
- ✗ **Implication:** Cloud architecture must include data residency, geopolitical awareness, and adaptive compliance tooling.

5. Resilience as a First-Class Architectural Constraint

- From cloud to platform to team dynamics, resilience is now a systemic trait.
 - Chase.com's emphasis on performance + resilience shows how critical these traits are in regulated industries.
 - Platform design must avoid brittle one-size-fits-all models — resilience requires flexibility and real-time feedback.
- ✗ **Implication:** Bake resilience into org structures, communication patterns, and runtime systems — not just infrastructure.

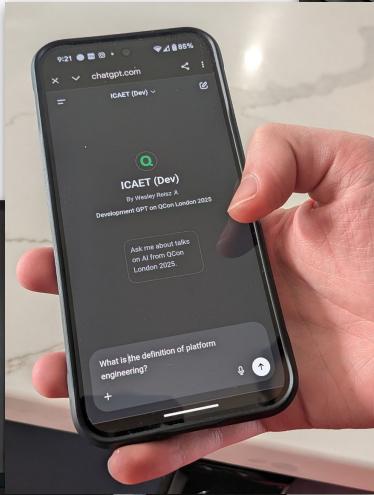
☛ Architect's 2025 Action Plan

Focus Area	Recommendations	🔗
Org Architecture	Build cross-functional architecture working groups to replace silos	
Platform Strategy	Treat platform teams as internal product teams, tied to user outcomes	
AI-Native Design	Introduce embeddings and vector retrieval to modernize search and recommendation	
Governance	Implement policy-as-code and multi-cloud observability for compliance resilience	
Culture	Promote architecture literacy via mentorship, pair reviews, and internal talks	

📚 Source Summary

Session	Speaker	Theme	🔗
<i>Platform Engineering as a Practice of Sociotechnical Excellence</i>	Lesley Cordero	Platform-as-change-agent	
<i>Holistic Engineering</i>	Vanessa Formicola	Sociotechnical design	
<i>Your Platform is Not an Island</i>	Rachael Wonnacott	Platform context + value alignment	
<i>Shine Bright as an IC</i>	Suhail Patel	Mentorship as strategy	
<i>Building Embedding Models</i>	Sahil Dua	AI-native architecture	
<i>Scaling Cloud @ Chase.com</i>	Durai Arasan	Governance, resilience, cloud ops	

Open Spaces



Workshop

Survey Results

AI-Enabled Delivery - ICSAET Cohort Only

56%
Super Green
votes

33%
Green votes

11%
Yellow votes

0%
Red votes

9
Total votes

27
Total
attendance

Total number of submissions **12**

NPS -33.33

Results

Overall Evaluation: A Qualified Success with Clear Signals for Iteration

- Participant Feedback: Polarized but Promising
 - **High engagement with no red votes:** Presentation feedback shows 56% Super Green and 33% Green votes—a clear indicator that the content landed well with a majority.
 - **Net Promoter Score (-33):** From the broader survey, this indicates detractors outweighed promoters. This typically signals confusion, unmet expectations, or a mismatch in what was promised vs. delivered.
 - **Valuable comments:** Many pointed to lack of clarity, structure, and goal articulation—but not a lack of value in the core idea. Several noted the chatbot and peer-learning model as innovative and high-potential.
- What Worked
 - **Engaged, senior cohort:** Feedback indicates *strong seniority (½ identify needs or define strategy), validating the audience fit.*
 - **Content quality and instructor rating:** Both received high scores (4.25 instructor expertise, 3.67 content), affirming delivery strength.
 - **Peer networking success:** Peer engagement scored 4.08 (indicating value in the cohort format).

Hypothesis(es)

Certification



If QCon attendees are offered a certification based on their attendance and engagement, then it will increase the perceived value of the conference and lead to greater knowledge retention, professional recognition, and ongoing community involvement.



AI as an Aid



If we leverage AI today as an aid in delivering the QCon experience, then we can create a level of personalization, interactivity, and scale that was previously impossible, significantly enhancing attendee engagement and value.



Supervised coding agents



If we apply ChOP (Chat-Oriented Programming) to software development, then we can produce non-trivial programs that effectively replicate and potentially replace core skills traditionally held by experienced software developers and architects quickly.



Cohort Model



If QCon attendees participate in structured cohorts, then their engagement with the conference and its content will increase, leading to a more immersive experience and improved knowledge retention.



 Conference + Trainings

 Trainings-Only

 Video-Only

Conference (3 days)
\$2,320

29% off until June 10th *

Access to 3 days of sessions, peer-sharing, social events, and video-on-demand

[Register](#)
Ticket features:

- All tracks, sessions, and keynotes
- Networking and peer-sharing opportunities
- Continental breakfast, lunch & coffee breaks
- Social events and receptions
- Access to the exhibitor area
- On-demand access to most presentation recordings for 6 months
- Download option for most presentation slides
- 1 day of training(s)
- 2 days of trainings
- Capstone Workshop
- ICSAET Certification

Early bird price valid until June 10th.
 Price increases to \$2,450 after this date.
 All amounts are stated in United States Dollars.

Conference (3 days) + Trainings (1 day)
\$3,060

23% off until June 10th *

Access to 3 days of sessions, peer-sharing and video on-demand, plus 1 day of trainings

[Register](#)
Ticket features:

- All tracks, sessions, and keynotes
- Networking and peer-sharing opportunities
- Continental breakfast, lunch & coffee breaks
- Social events and receptions
- Access to the exhibitor area
- On-demand access to most presentation recordings for 6 months
- Download option for most presentation slides
- 1 day of training(s)
- 2 days of trainings
- Capstone Workshop
- ICSAET Certification

Early bird price valid until June 10th.
 Price increases to \$3,190 after this date.
 All amounts are stated in United States Dollars.

POPULAR
Conference (3 days) + Trainings (2 days)
\$3,755

20% off until June 10th *

Access to 3 days of sessions, peer-sharing and video on-demand, plus 2 days of trainings

[Register](#)
Ticket features:

- All tracks, sessions, and keynotes
- Networking and peer-sharing opportunities
- Continental breakfast, lunch & coffee breaks
- Social events and receptions
- Access to the exhibitor area
- On-demand access to most presentation recordings for 6 months
- Download option for most presentation slides
- 1 day of training(s)
- 2 days of trainings
- Capstone Workshop
- ICSAET Certification

Early bird price valid until June 10th.
 Price increases to \$3,885 after this date.
 All amounts are stated in United States Dollars.

Conference (3 days) + Certification (half day)
\$3,060

23% off until June 10th *

Access to 3 days of sessions, complete the Capstone workshop on the fourth day and achieve the ICSAET certification
[What is the ICSAET Certification?](#)

[Register](#)
Ticket features:

- All tracks, sessions, and keynotes
- Networking and peer-sharing opportunities
- Continental breakfast, lunch & coffee breaks
- Social events and receptions
- Access to the exhibitor area
- On-demand access to most presentation recordings for 6 months
- Download option for most presentation slides
- 1 day of training(s)
- 2 days of trainings
- Capstone Workshop
- ICSAET Certification

Early bird price valid until June 10th.
 Price increases to \$3,190 after this date.
 All amounts are stated in United States Dollars.

ICSAET Insights

Individuals and interactions over processes and tools (even with AI)

- While the system certainly gave me geek cred with the attendees, it was not really used as effectively as I thought it would be. The insights enabled the initial conversation, but the actual crowd sourced open space topics deviated from those insights. Only 4 of the 28 in participants used the system during the workshop.
- The rumors of RAG's demise are greatly exaggerated. Though MCP is hot and deserves the attention in agent workloads.
- Using Supervised Coding Agents is incredible, but experience continues to be more important than ever. Your output needs structure and discipline to get the best from the tools (which is absolutely possible). They are definitely an incredible initial productivity boast. It personally remains to be seen for me if you don't pay for that speed over the long term.
- ***It feels like these tools get you 90% of the way there.***



Andrew Clay Shafer · 1st

Learning

3d ·

...

AI will rapidly produce the first 90% of whatever you want.

All you have to do is polish the remaining 90%.



You and 97 others

3 comments · 5 reposts

Like

Comment

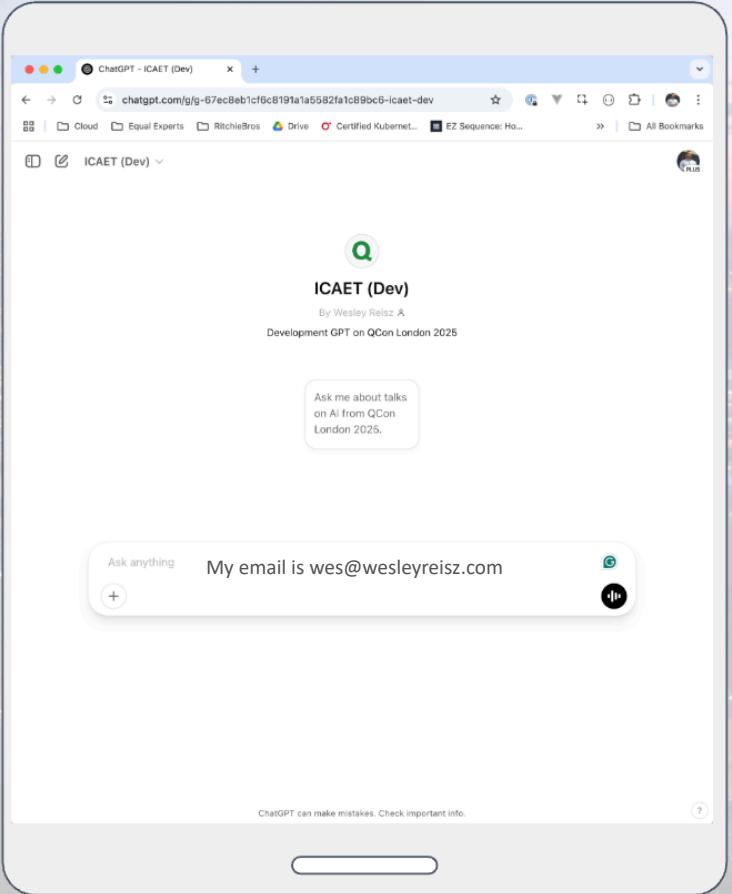
Repost

Send

Leverages ChatGPT
Implements a
**Retrieval Augmented
Generation (RAG)**
Architecture to add
QCon Transcripts.

Loaded with the
complete schedule
from this year at this
last year.

MVP



bit.ly/icsaet

Note: It's an early prototype, so you have to use your email address to access the transcriptions. Just tell ChatGPT and it will use it. There is also some strangeness at times call the ICAET connector. If you get a blank screen, just type something and I'll see what was missing

{{ QCon AI 2025. THE AI CONFERENCE FOR DEVELOPERS, ARCHITECTS & ENGINEERING LEADERS }}

DECEMBER 16-17, 2025

NEW YORK ACADEMY OF MEDICINE, NEW YORK, NY

AI-driven engineering: How real teams accelerate the entire software lifecycle with AI.

Move beyond the hype. Learn how senior architects, engineers, and leaders are embedding AI to innovate faster, design smarter, and deploy enterprise scale resilient systems securely and reliably.

Get proven patterns from practitioners tackling real-world AI implementation now.

[REGISTER](#)

Early bird price until June 10th

Key Takeaways

Build the Right Thing

*Lean thinking such as Lean Startup is an approach to building products that aims to shorten product development cycles and rapidly discover value. **There is no point in building something fast if you are building the wrong thing.** Combining AI with a lean thinking approach powerful.*

There are NO Silver Bullets

Code generation tools like Cursor, Github CoPilot, Windsurf, and others, along with advanced frontier models such as GPT 4 (now 4.1) and Claude Sonnet 3.7 (now 4), are evolving rapidly and can offer incredible productivity. However, using them effectively requires **expertise and guidance**. *Going with the vibe has its place*, but temper expectations. There are emerging structured techniques helping make the most of this stack.

Embrace Change

AI tools today do not represent *thinking*. They are mathematical models that predict (incredibly well) the next set of tokens. Embrace chance, use it for what works, but at the end of the day remember **expertise matters in building innovative systems** that scale well and perform.

Learning Organization

MCP is hot; RAG is NOT dead. Both are powerful and exciting tools in the suite of tools developers are using in AI today. Tomorrow will be another tool, another model. **The pace of change today is staggering.** Embrace the mentality of a learning organization.

But wait there's more...

Migrating from Manual Retrieval to GPT with MCP

Before: Manual ReST-Based Retrieval Flow

- User input sent to a backend service
- Backend:
 - Embeds query
 - Queries vector store (e.g. Pinecone)
 - Inserts retrieved chunks into a prompt template
 - Sends to GPT via ReST API
- GPT responds with answer based on embedded context
- All logic orchestrated externally

📦 Limitations:

- Static flow
- Manual orchestration
- No dynamic reasoning or fallback strategies

After: GPT with MCP + Retrieval Tool Integration

- Define retrieve_documents() as a tool callable by GPT
- GPT autonomously:
 - Decides when to call retriever
 - Embeds and queries vector store
 - Synthesizes response based on results
- GPT can chain tool calls or reason over intermediate steps

🛠 Components:

- MCP-enabled GPT
- Tool definition schema (OpenAI tools/functions)
- Callable retriever endpoint (e.g., Flask API, AWS Lambda)

🚀 Benefits:

- Declarative retrieval flow
- LLM-as-orchestrator: less glue code
- Easier to compose with other tools (e.g., metadata, summarization)
-

Amazon Closes Zappos Deal, Ends Up Paying \$1.2 Billion

6:47 AM PST · November 2, 2009

 Amazon earlier this morning announced the official closing of the acquisition of Zappos, a deal which we broke the news about back in July. Turns out the valuation of the online footwear and clothing retailer went up from the reported \$928 million over the past few months too – thank you, stock market – and Zappos turns out to have been deemed worth a solid \$1.2 billion by Jeff Bezos & co based on Friday's closing price of \$117.4 a share.

Zappos CEO Tony Hsieh writes in a public letter:

We've been spending the past few months obtaining all the proper



 Adobe Acrobat Pro

**Help me land
my dream job.**

[Start free trial](#)

AI combined with *lean thinking*
is a superpower; *experience* matters.



Thank you

References

Video Pipeline Repo

<https://github.com/wesreisz/video-pipeline>

The video-pipeline repository designed to automate the processing of video files using AWS services. It orchestrates a sequence of operations including file uploads, transcription, chunking, and further processing, leveraging AWS tools such as Amazon S3 and AWS Transcribe.



Standing on the shoulders of giants

Original work by these amazing thought leaders.

The Lean Startup

- Eric Ries (*The Lean Startup*)

“Build the right thing”

- Steve Blank (*The Four Steps to the Epiphany: Successful Strategies for Products That Win*)

“There are no Silver Bullets”

- Frederick Brooks, Jr. (*Mythical Man-Month, The: Essays on Software Engineering*)

“Embrace Change”

- Kent Beck (*Extreme Programming Explained: Embrace Change*)

“Learning organization”

- Peter Senge (*The Fifth Discipline: The Art & Practice of The Learning Organization*)