

Building Massive-Scale Generative AI Services with Kubernetes and Open Source

John McBride - Sr. Software Engineer, OpenSauced



bsky.app/profile/johncodes.com



Copilot, but for git history

StarSearch

Copilot, but for git history

Get information on contributor activity

What type of pull requests has @brandonroberts worked on?

Identify key contributors

Who are the most prevalent contributors to the TypeScript ecosystem?

Find contributors based on their work

Show me the lottery factor for contributors in the remix-run/react-router project?

Find experts

Who are the best developers that know Tailwind and are interested in Rust?

Ask a question



**You**

Who are the best developers that know Tailwind and are interested in Rust?

**StarSearch**

Here are some notable developers who have shown expertise in Tailwind CSS and Rust, based on their contributions to related projects:

Tailwind CSS Contributors:**1. Yashhh999**

- Repository: [Syknapse/Contribute-To-This-Project](#)
- Contribution: Added Tailwind CSS to the project, including library configuration and custom styles.
- [Pull Request](#)

2. OFD16

- Repository: [tailwindtoolbox/Landing-Page](#)
- Contribution: Added custom components and updated configurations using Tailwind CSS.
- [Pull Request](#)

3. ErwinAI

- Repository: [aniftyco/awesome-tailwindcss](#)
- Contribution: Added resources and tools related to Tailwind CSS.
- [Pull Request](#)

4. paulpopus

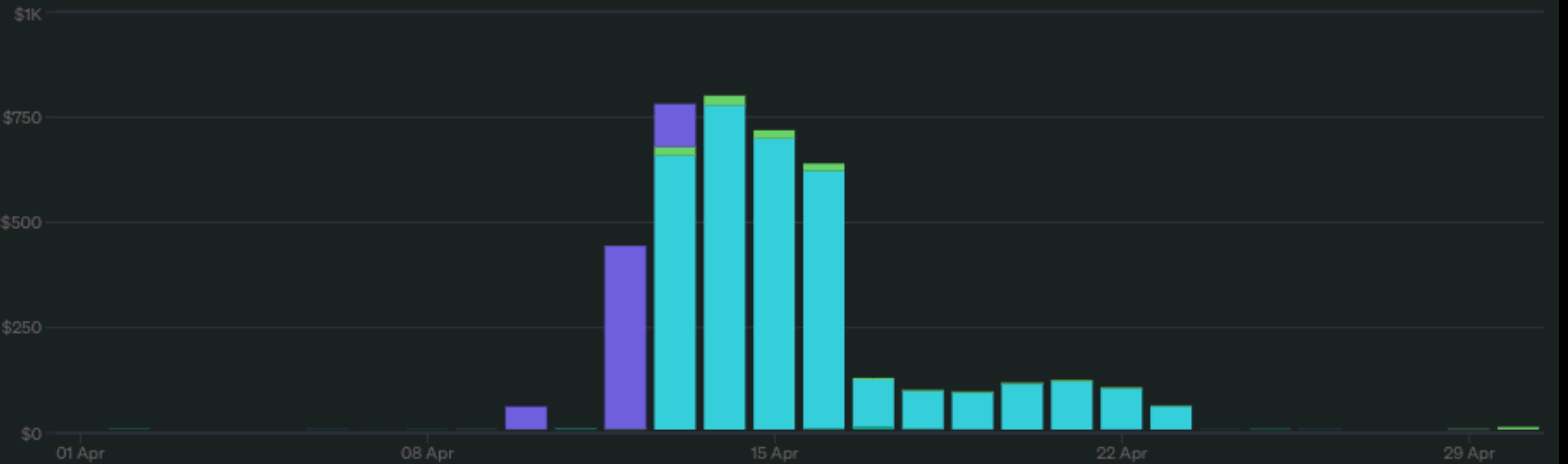
- Repository: [payloadcms/payload](#)
- Contribution: Added Tailwind and shadcn/ui example.

[Clear chat history](#)

Was this response useful?

[Need inspiration? ↑](#)[Ask a question](#)

Monthly Spend \$4,107.12





Back of the napkin math ...

Using gpt4 - 30\$ / 1m input tokens and 60\$ / 1m output tokens

Targeting 40,000+ repos

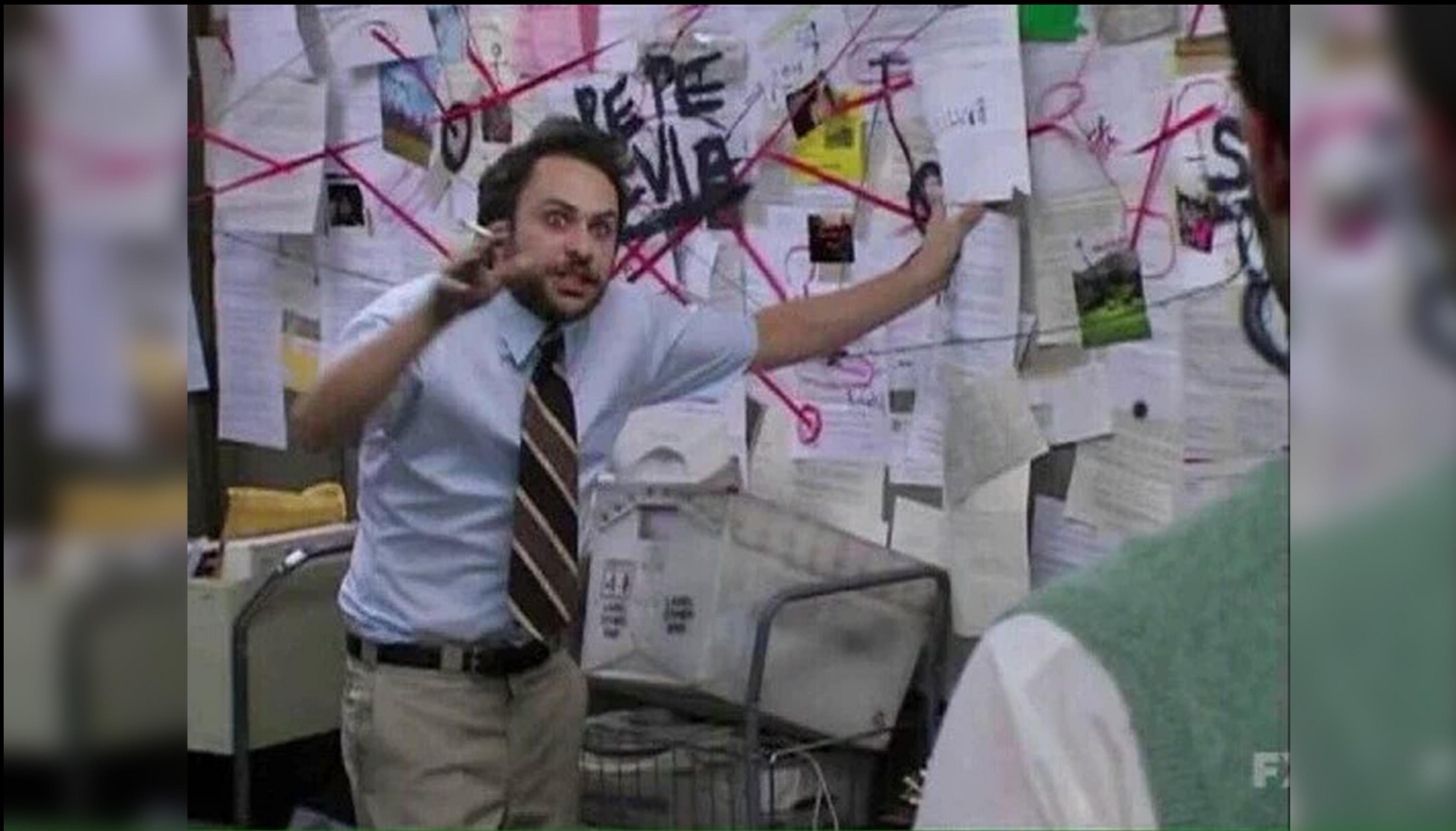
Assuming each repo generates about 190 words / day in issues and pull requests

~= 600\$ / day

~= 216,000\$ / year

**Depending on scale or user generated volume,
can get very, very expensive, quickly.**

Depending on scale or user generated volume,
can get very, very expensive, quickly.



How we built huge scale, AI RAG data pipelines on k8s



- 💡 **Consume GitHub events via events firehose**
- 💿 **Use timescaledb vector storage of embeddings to retrieve relevant information**
- 📝 **Use OpenAI function calling and services with multi-agent strategy**

GitHub events

“Firehose”



GitHub events
“Firehose”



It's pizza time



GitHub events
“Firehose”



It's pizza time



It's pizza time





It's pizza time



event_time, event_type ... metadata

It's pizza time



event_time, event_type ... metadata

timescaledb



It's pizza time



timescaledb



It's pizza time



timescaledb



OpenSauced

OpenSauced Workspace Explore Highlights StarSearch

Users, Repositories... 8K

analogjs/analog

The fullstack meta-framework for Angular. Powered by Vite and Nitro

Add to Workspace Share Range: 30d

Overview Contributors

Contributors 30 days

Total 30 days: **31** Average per day: **1**

Date	Commits
4/25	2
4/26	4
4/27	1
4/28	3
4/29	4
4/30	1
5/1	1
5/2	1
5/3	4
5/4	1
5/5	1
5/6	1
5/7	1
5/8	1
5/9	1
5/10	1
5/11	1
5/12	1
5/13	3
5/14	1
5/15	1
5/16	1
5/17	1
5/18	1
5/19	2

Lottery Factor High

The top contributor of this repository has made **50%** of all commits in the past **30** days.

Contributor	Commits	% of Total
brandonroberts	14	50%
eduardoRoth	2	7%
sonukapoор	2	7%
BoogieMonsta	1	3%
Other contributors		33%





Star-search-embedder





Star-search-embedder





Star-search-embedder





Dear Katie,

Here's to the crazy ones. The ones who are different... The ones who see things differently. The troublemakers. The round pegs who aren't fond of rules. And they have no respect for the status quo. You can quote them, disagree with them, glorify or vilify them. About the only thing you can't do is ignore them. Because they change things.

They push the human race forward. And while some may see them as the crazy ones, we see genius. Because the people who are crazy enough to think they can change the world, are the ones who do.

Take care,
John Appleseed



Small open source LLM on a few GPUs on Kubernetes



Small open source LLM on a few GPUs on Kubernetes



Generate a summary for the content

**Small open source
LLM on a few GPUs
on Kubernetes**



**Small open source
LLM on a few GPUs
on Kubernetes**







Node pool with
T4 GPUs



Node pool with T4 GPUs

Installed Nvidia
drivers



vLLM



vLLM

Daemonset
with vLLM



vLLM

VLLM



VLLM



VLLM

Open
large language model



VLLM

Concurrent
OpenAI compatible API

Open
large language model



VLLM

Open
large language model





Open large language model



Easily and freely available



**Often a more permissible license and able to modify
for your needs (like re-quantize or fine tune)**



Small and very efficient (for most use cases)



vLLM



vLLM

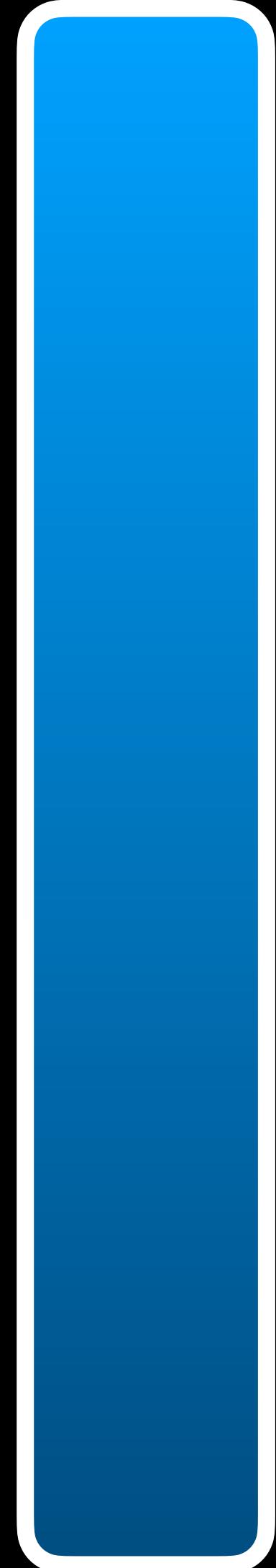


vLLM

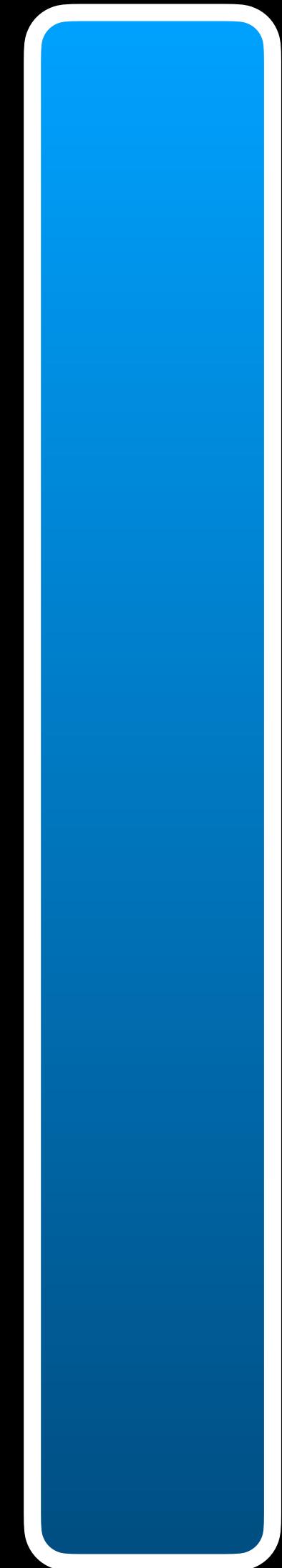


Kubernetes
Service





OpenAI
compatible
API



vllm-service.vllm-namespace.svc.cluster.local



vllm-service.vllm-namespace.svc.cluster.local

v1/chat/completions

How do we manage all this?



How do we manage all this?

- 🦆 Pulumi infra as code for managing and bootstrapping multiple clusters / ends
- 🎷 GitHub actions for automatic container build and deployment
- 🤖 Grafana alerting, On-call, and basic validation



DEMO



Generate a summary for the content

**Small open source
LLM on a few GPUs
on Kubernetes**



Generate a summary for the content



Dear Kate,

Here's to the crazy ones. The misfits. The rebels. The troublemakers. The round pegs in the square holes. The ones who see things differently. They're not fond of rules. And they have no respect for the status quo. You can quote them, disagree with them, glorify or vilify them. About the only thing you can't do is ignore them. Because they change things.

They push the human race forward. And while some may see them as the crazy ones, we see genius. Because the people who are crazy enough to think they can change the world, are the ones who do.

Take care,
John Appleseed

Text Embedding Model



Dear Kate,

Here's to the crazy ones. The ones who are different. The ones who are膽
holes. The ones who see things differently. They're not fond of rules. And they have no respect for the status quo. You can quote them, disagree with them, glorify or vilify them. About the only thing you can't do is ignore them. Because they change things.

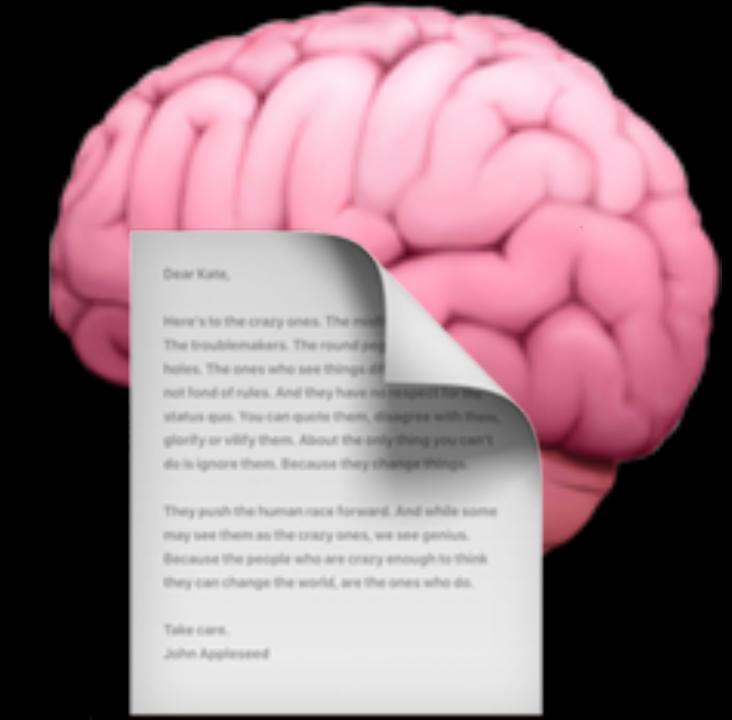
They push the human race forward. And while some may see them as the crazy ones, we see genius. Because the people who are crazy enough to think they can change the world, are the ones who do.

Take care,
John Appleseed

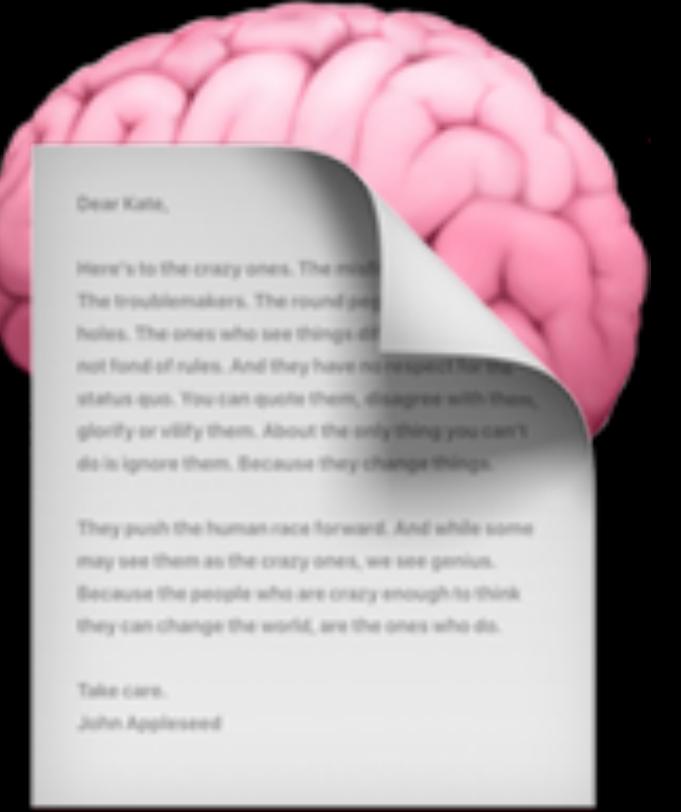
Text Embedding Model



Text Embedding Model



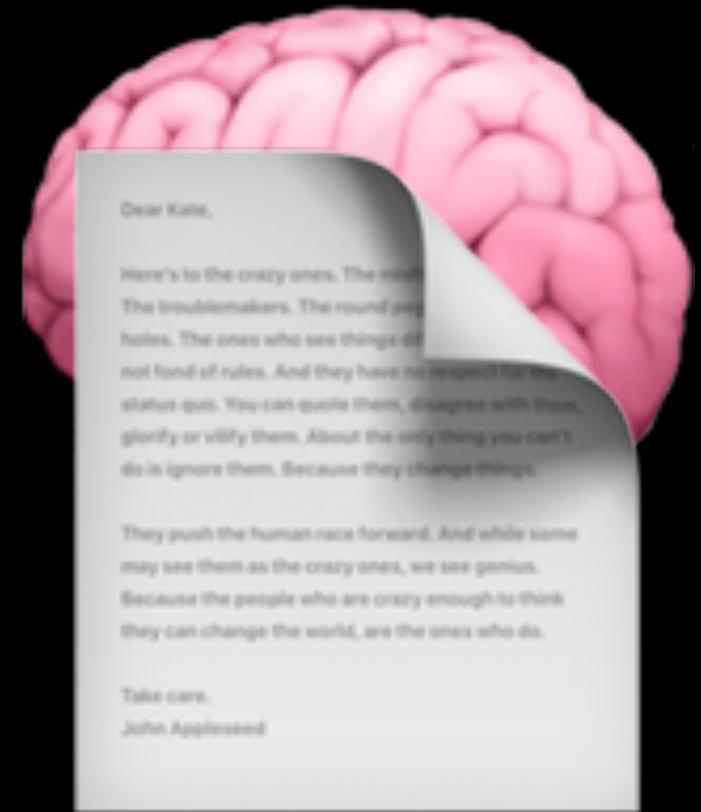
Text Embedding Model



[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]

Text Embedding Model

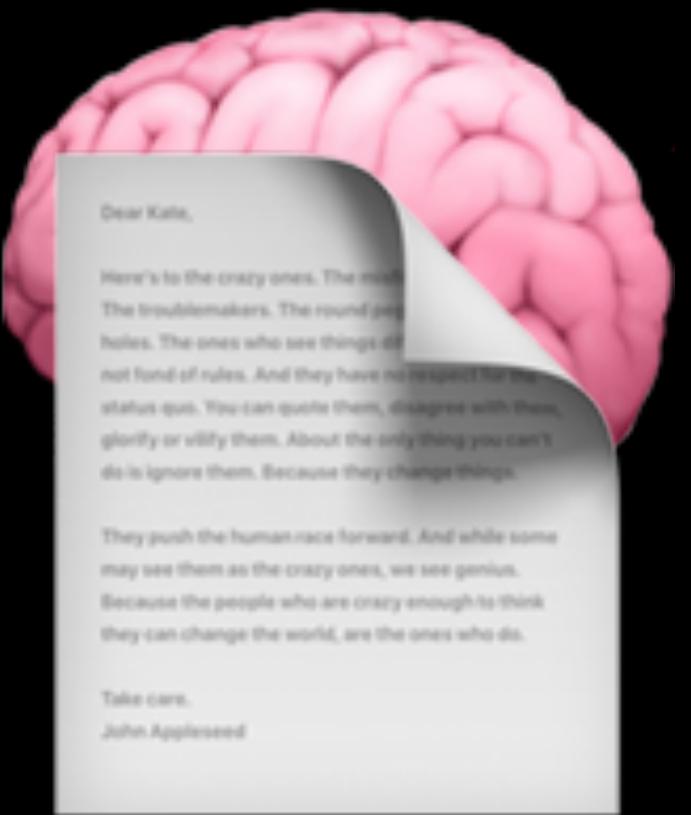
pgvector



[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]



Text Embedding Model



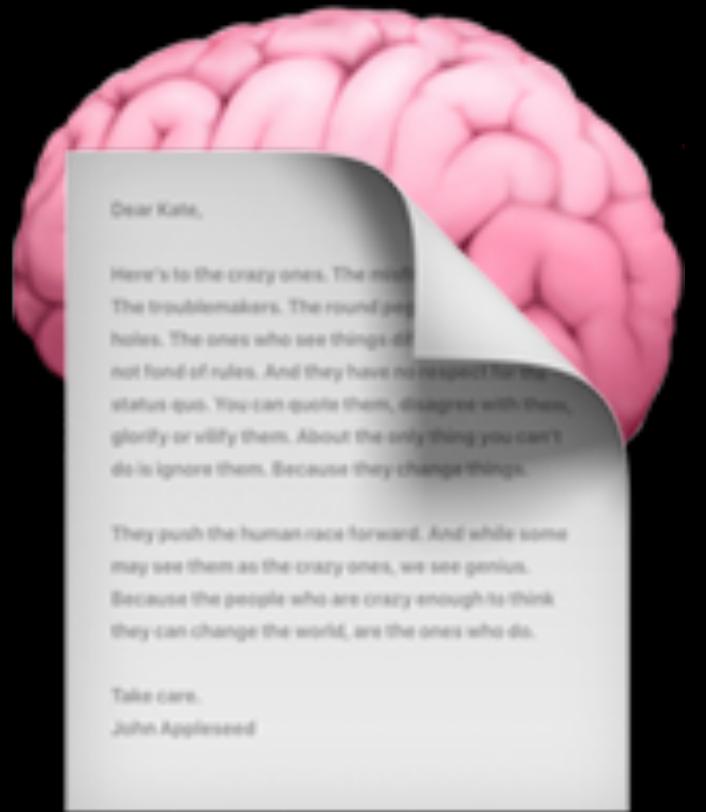
pgvector



[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]



Text Embedding Model



pgvector



[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]

pgvector



```
[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]  
[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]  
[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]  
...  
[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]  
[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]  
[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]
```

pgvector



[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]

[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]

[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]

...

[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]

[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]

[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]

event_time, pr_number ... metadata

event_time, pr_number ... metadata

event_time, pr_number ... metadata

...

event_time, pr_number ... metadata

event_time, pr_number ... metadata

event_time, pr_number ... metadata

pgvector



[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]

[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]

[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]

...

[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]

[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]

[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]

event_time, pr_number ... metadata

event_time, pr_number ... metadata

event_time, pr_number ... metadata

...

event_time, pr_number ... metadata

event_time, pr_number ... metadata

event_time, pr_number ... metadata

Cosine similarity
Hierarchical Navigable Small Worlds (HNSW) index

pgvector



[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]

[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]

[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]

...

[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]

[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]

[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]

event_time, pr_number ... metadata

event_time, pr_number ... metadata

event_time, pr_number ... metadata

...

event_time, pr_number ... metadata

event_time, pr_number ... metadata

event_time, pr_number ... metadata

Cosine similarity
Hierarchical Navigable Small Worlds (HNSW) index

pgvector



- Separate tables
- Separate embeddings

[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]

[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]

[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]

...

[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]

[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]

[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]

event_time, pr_number ... metadata

event_time, pr_number ... metadata

event_time, pr_number ... metadata

...

event_time, pr_number ... metadata

event_time, pr_number ... metadata

event_time, pr_number ... metadata

Cosine similarity
Hierarchical Navigable Small Worlds (HNSW) index

StarSearch



StarSearch



StarSearch



A multi-agent approach using
OpenAI function calling
and our existing services architecture





Pre-processor agent







Detect prompt injection attacks

Cleans up user's prompt

Attempt to correct spelling mistakes



Detect prompt injection attacks

Cleans up user's prompt

Attempt to correct spelling mistakes



✓ Detect prompt injection attacks

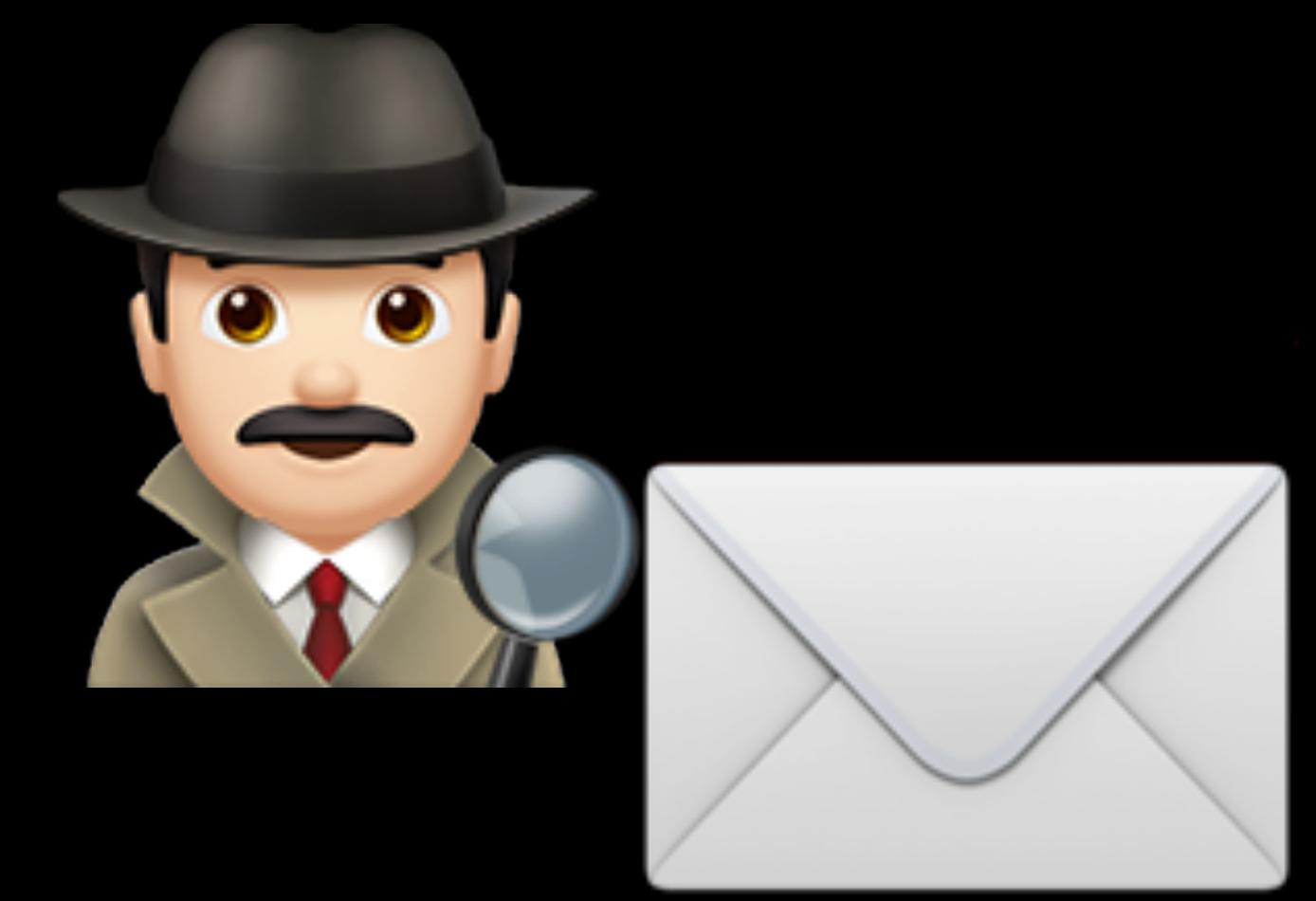
✓ Cleans up user's prompt

Attempt to correct spelling mistakes



- ✓ Detect prompt injection attacks
- ✓ Cleans up user's prompt
- ✓ Attempt to correct spelling mistakes





Manager agent









Bing Search Agent





Bing Search Agent

Pull Request Agent



Bing Search Agent

Pull Request Agent

Other service agents

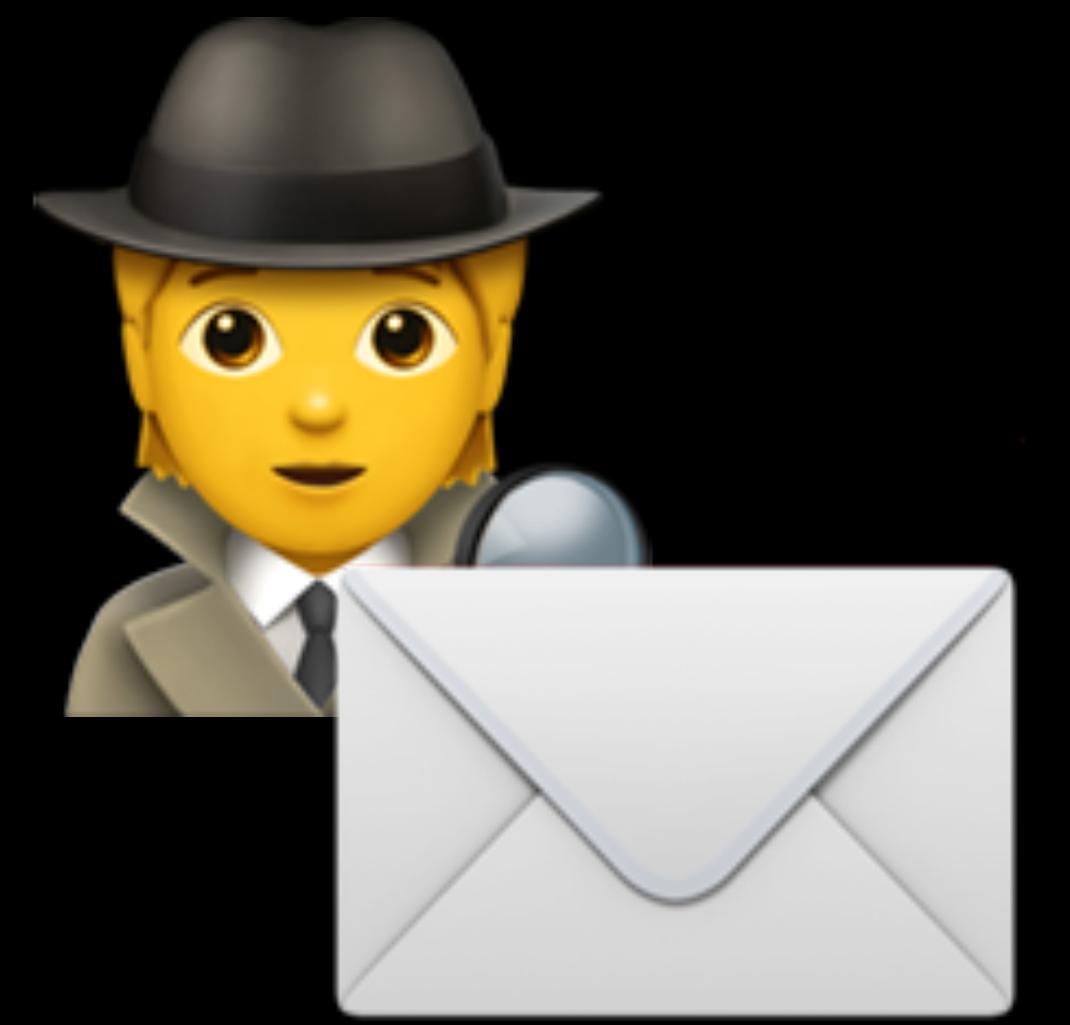


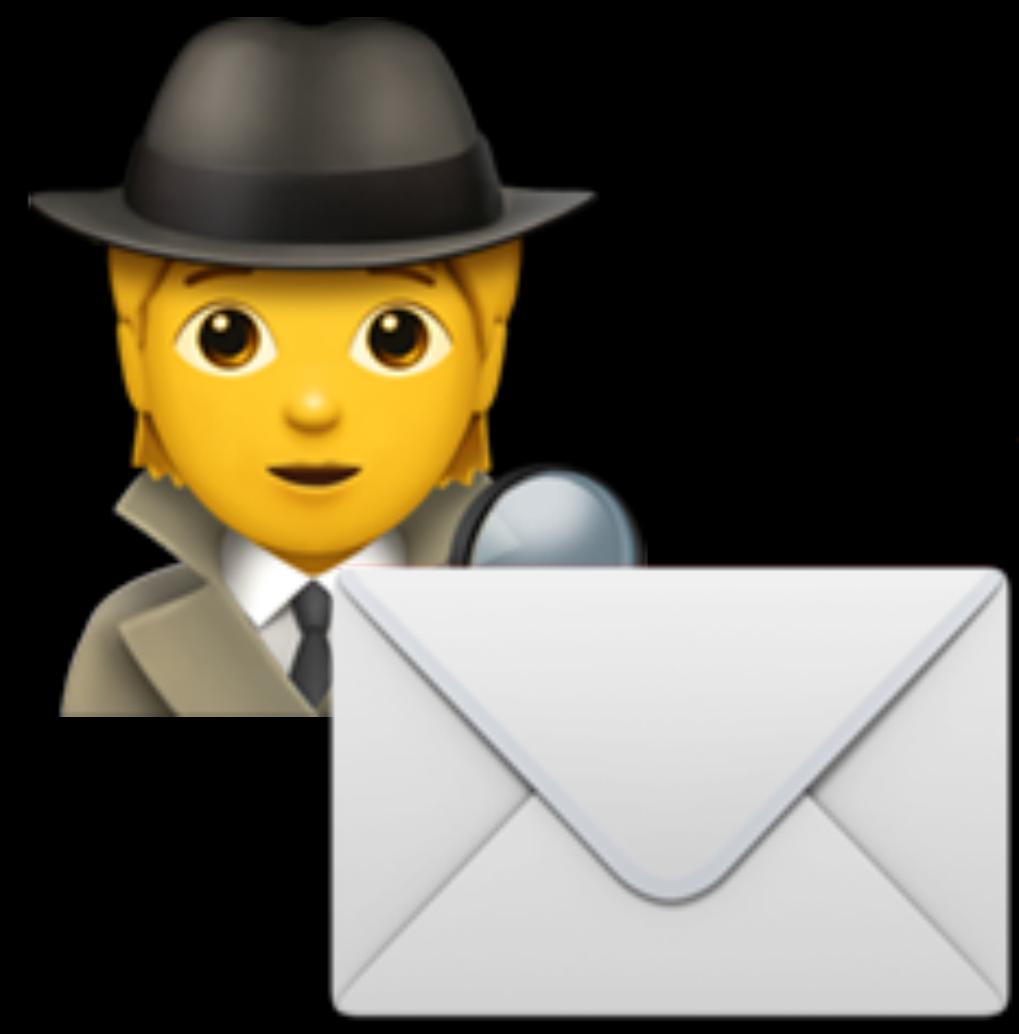
Pull Request Agent

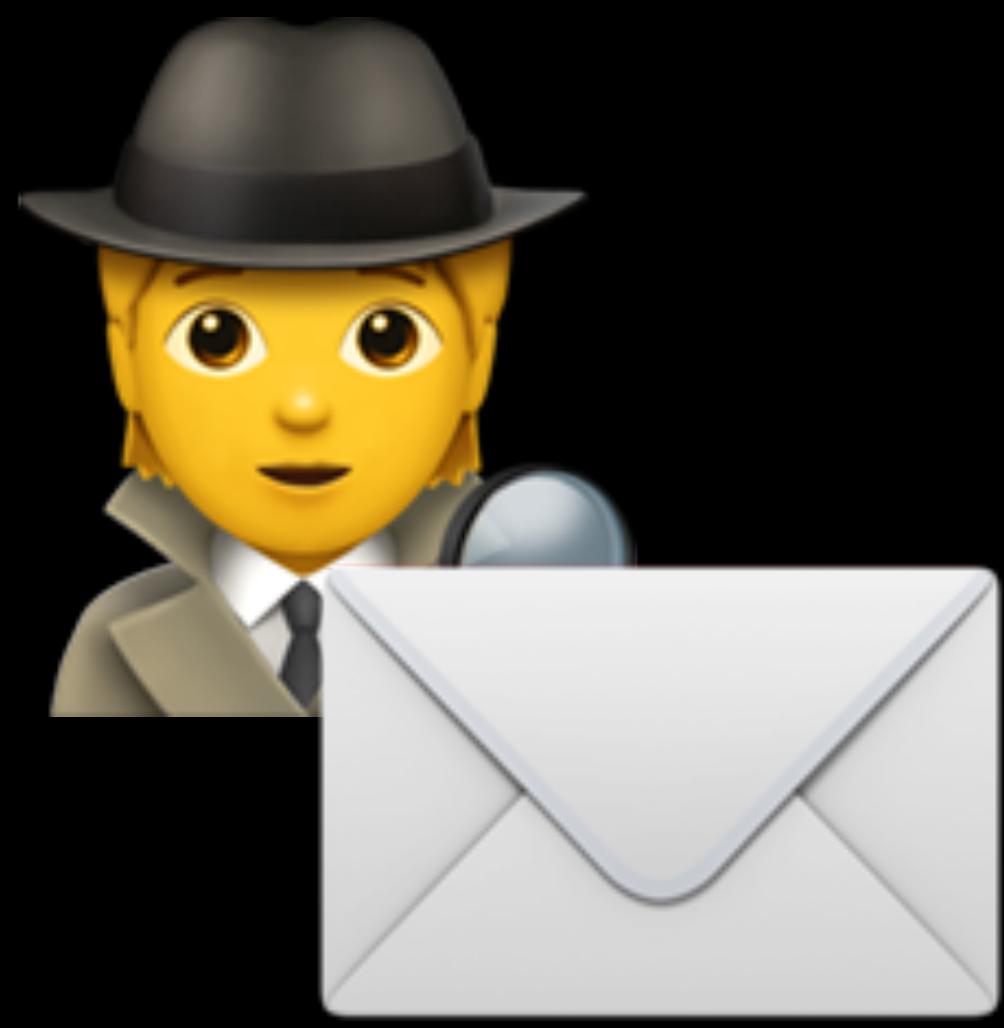




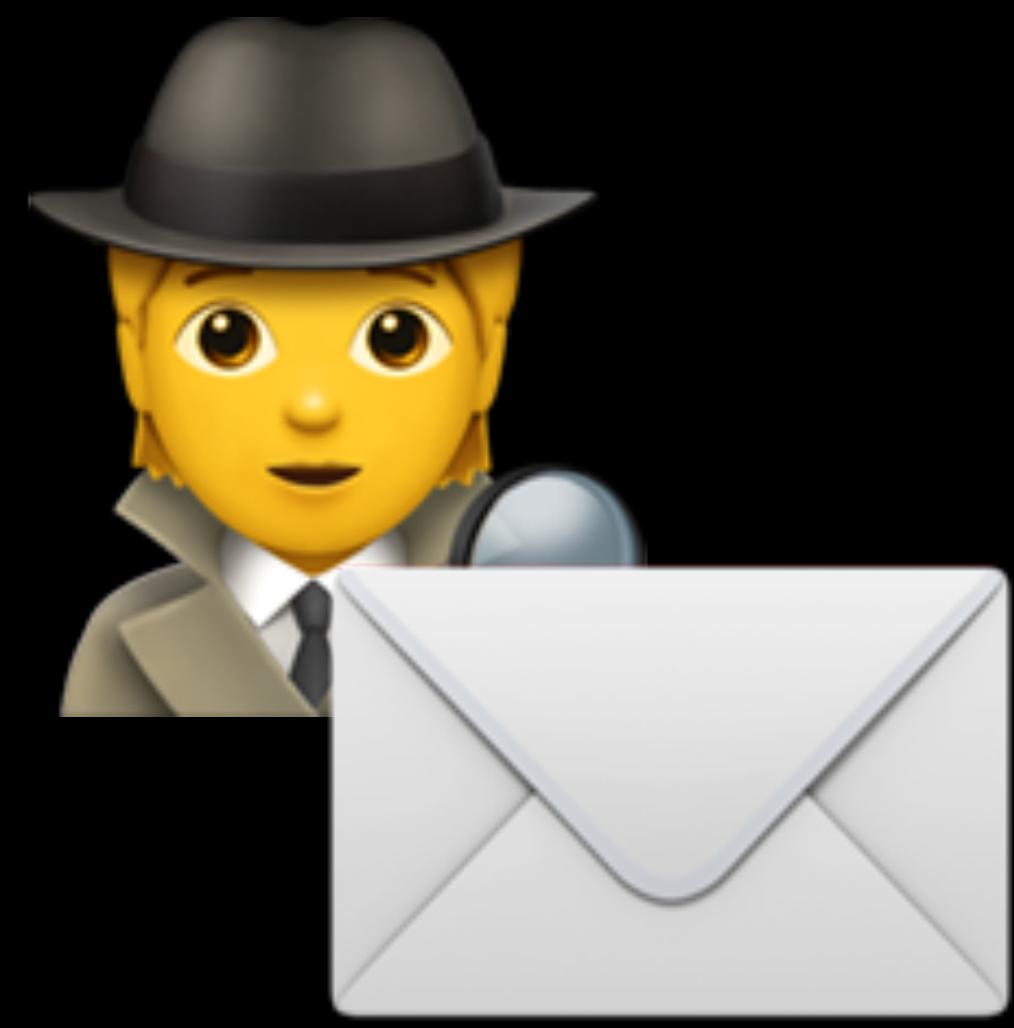








Perform vector search and summarize results



**Perform vector search and summarize
results**

**Interface with other services
In parallel in toolkit**



Perform vector search and summarize results

Provided query

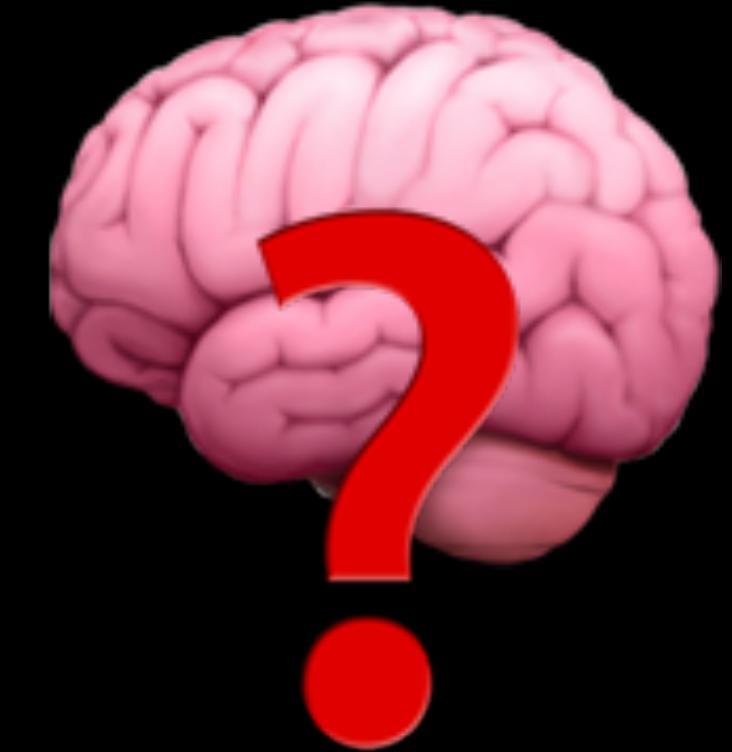


Text Embedding Model

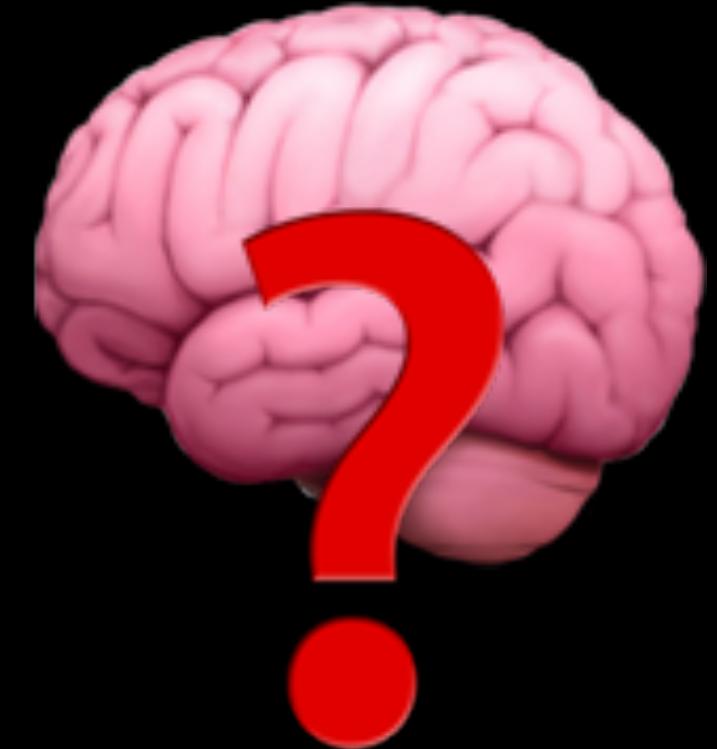


?

Text Embedding Model

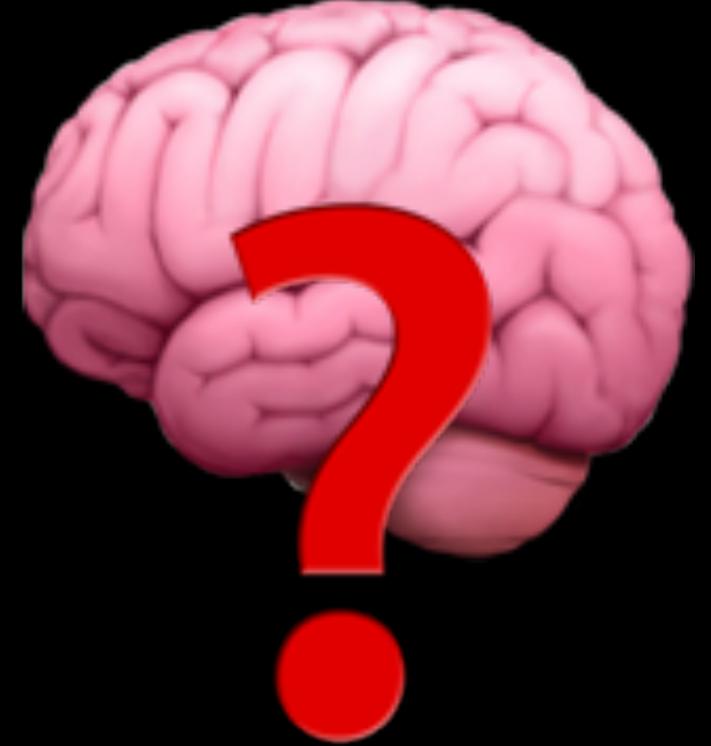


Text Embedding Model



[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]

Text Embedding Model



pgvector



```
[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]    event_time, pr_number ... metadata  
[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]    event_time, pr_number ... metadata
```

[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]

pgvector



```
[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]   event_time, pr_number ... metadata  
[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]   event_time, pr_number ... metadata
```

**Cosine similarly,
nearest neighbor search**

[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]

pgvector

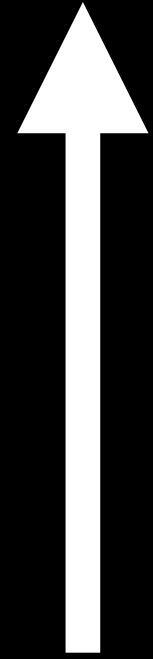


[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]

event_time, pr_number ... metadata

[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]

event_time, pr_number ... metadata



Nearest neighbor

[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]



[1.123, 0.456, 2.789 ... 1.123, 0.456, 2.789]

event_time, pr_number ... metadata

Preamble

Context

Question



Preamble

Context

Question

`event_time, pr_number ... metadata`

OpenAI function calling loop



Preamble

Context

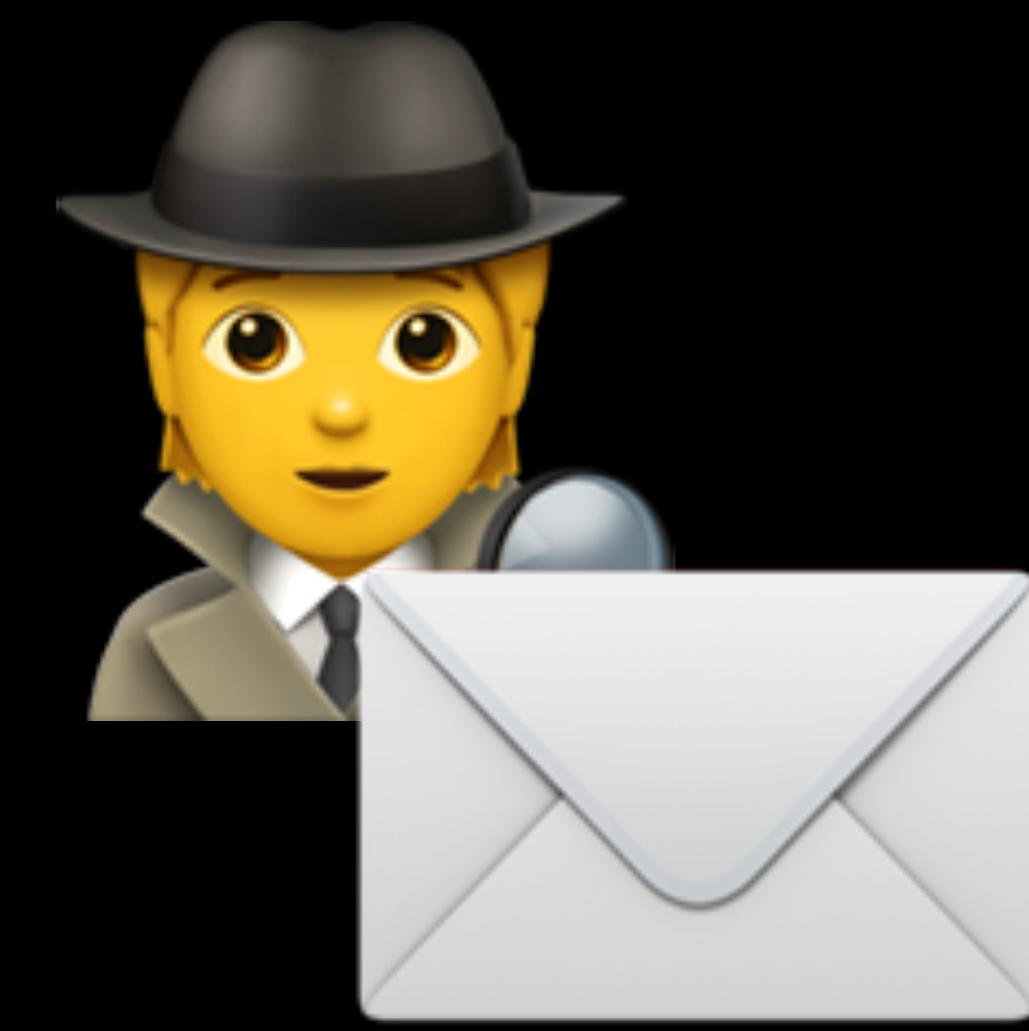
Question

`event_time, pr_number ... metadata`

OpenAI function calling loop















Context aware answer



Context aware answer



Practical lessons and tips



- 🌟 Use LLMs to pre-bake “summaries” for more complex content (like pull requests!) can drastically reduce noise
- 🌴 Use indexes if you can trade off accuracy
- 🏎️ Huge cost saving in using small LLM inference

Challenges



- 🍿 AI / ML / LLM observability and validation seems to still be an open question
- 🚜 Narrower scope and focus



**Please give feedback
on Sched!**



Q&A