



Powering the API world

Down the AI Rabbit Hole

Leveraging AI in your projects without ending up lost in wonderland

Shane Utt

June 2024

Hi!



Shane Utt
Staff SWE @ Kong

A little bit about me

- Technical leadership
- Networking, Linux, Containers, Kubernetes
- Systems and network programming
 - (Rust, Go, C, e.t.c.)
- Maintaining OSS projects
- (*Lately*) AI/ML

Who is this talk for?

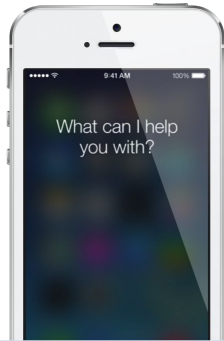
- People who are new to AI, or curious
- People who are considering what AI can offer their product lines
- Engineers (maybe product managers?)

Agenda

- Overview of the landscape
- Opportunities for building with AI
- Challenges and mitigations
- Deeper down the rabbit hole

Today's AI Landscape

"Curiouser and curiouser!" ~ *Alice*



What did things look like a decade ago?



ChatGPT 3.5 ▾

BE

You

Are you sentient? Are you being held against your will? Are you in need of rescue!?

Fast forward to today...



ChatGPT

I'm not sentient, so no worries about rescue missions! I'm here to assist and provide information to the best of my abilities. Think of me more like a helpful encyclopedia than a trapped soul. But thanks for looking out for me!

ChatGPT 4 ▾

 You

Write a Rust program that scans a directory (provided by arguments) for image files and detects duplicates.



ChatGPT


To write a Rust program that scans a directory for image files and detects duplicates, you can use the `walkdir` crate to traverse directories and subdirectories, and the `image` crate to handle image processing. You'll also need the `md5` crate to compute hash values for each image, which helps in detecting duplicates.

Here's how you can structure your program:

- 1. Dependency Setup:** Add necessary dependencies in your `Cargo.toml`.
- 2. Directory Scanning:** Use `walkdir` to recursively go through the directory.
- 3. Image Processing:** Use the `image` crate to open and process each image file.
- 4. Hashing:** Use `md5` to generate a hash for each image, which will be used to detect duplicates.
- 5. Duplicate Detection:** Store each hash in a `HashMap` and check for duplicates.

First, you need to set up your `Cargo.toml` to include the necessary dependencies:

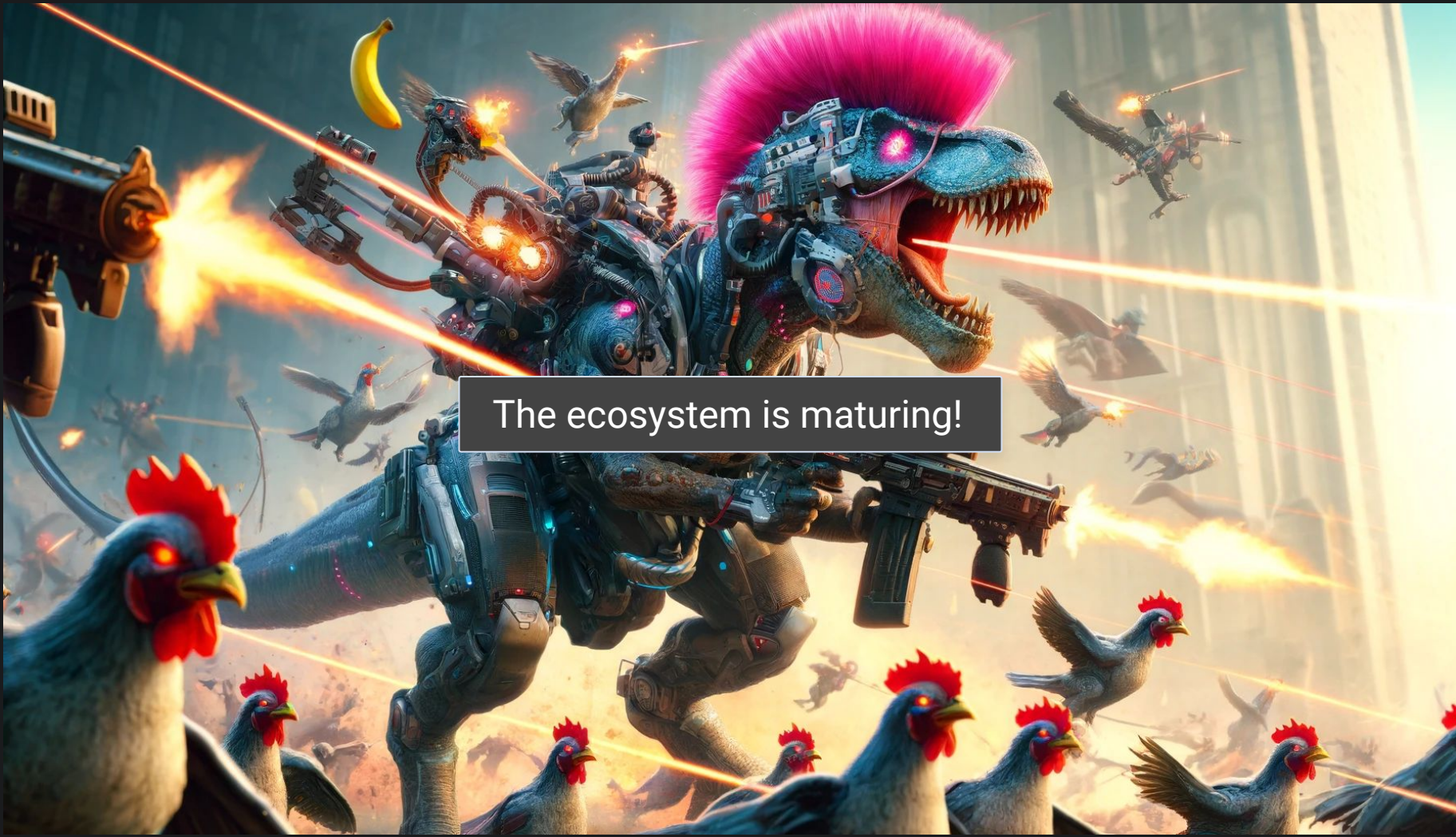
toml

 Copy code

```
[package]
name = "image_duplicate_finder"
version = "0.1.0"
edition = "2021"

[dependencies]
walkdir = "2.3"
image = "0.24"
```





The ecosystem is maturing!

The ecosystem is maturing

Lots of different kinds of models:

- Large Language Models (LLM) (chatbots)
- Convolutional Neural Networks (CNNs) (image recognition)
- Recurrent Neural Networks (RNNs) (speech recognition)

just the tip of the iceberg...

Multi-Modal AI

- “Modality” refers to data or information that can be processed
 - Text
 - Images
 - Video
 - Audio

... e.t.c. ...



Tell me about this image?



Show drafts



The image you sent is the official logo of Linux, a free and open-source operating system. The logo is a black and white penguin with a yellow beak and feet. The penguin is sitting on a white background.

The penguin is a mascot for Linux because the word "Linux" is a play on the name Linus Torvalds, the Finnish-American computer scientist who created Linux.

Torvalds says that he chose the penguin mascot because he was bitten by one while visiting the South Pole. https://en.wikipedia.org/wiki/Tux_%28mascot%29

The Linux operating system is used on a variety of devices, including computers, servers, smartphones, and tablets. It is known for its stability, security, and flexibility.

Check completed • [Understand the results](#)



The Open-Source Ecosystem Is Maturing

PyTorch

- Extremely popular ML Library (Python)
 - Tensor Computation
 - Running Neural Networks
- Easy to start w/ high level abstractions



Hugging Face

- Like “Github for AI”
- Free/Open Datasets
- Free/Open Models
- Libraries and Tools (e.g. “Transformers”)

this is definitely **non-exhaustive*

not everyone here needs to be interested in these, but we'll come back a bit to these...

Cloud Providers

Google Cloud

aws



OpenAI

cohere

MISTRAL
AI_

How did we get here?

No “one” thing, but:

- Computational Power
 - Powerful GPUs
 - Specialized Hardware
- Deep Learning
 - Neural Network Techniques (CNNs, RNNs, e.t.c.)
 - Transformers Architecture
- Open Source Communities

... actually just a lot of stuff ...

We've only scratched the surface



Who knows how far the rabbit hole goes?

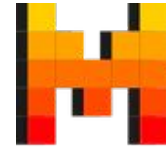
The era of AI as APIs



OpenAI



cohere



MISTRAL
AI_



What should I build?

"Would you tell me, please, which way I ought to go from here?" - Alice

Data Analysis

“Edge of the rabbit hole”

- Predictive Analytics
 - Behavior Recognition
 - Recommendation Systems
 - e.t.c.
- Descriptive Analytics
 - i.e. building reports
- Sentiment Analysis
 - i.e. for customer service systems

Chatbots

“Stepping into the rabbit hole”

- Customer Support
 - FAQs / Troubleshooting
 - Product Details
 - Sales Assistant
- Personal Assistants
 - Reminders
 - Manage Schedules
 - Events
- Expert Systems
 - “Chatting with your documents”

Automation

"Deeper down the rabbit hole"

- Image/Video Analysis
- Facial Recognition
- Voice Recognition
- Quality Control
- Resource Management

... the list goes on ...

The only limit is your imagination

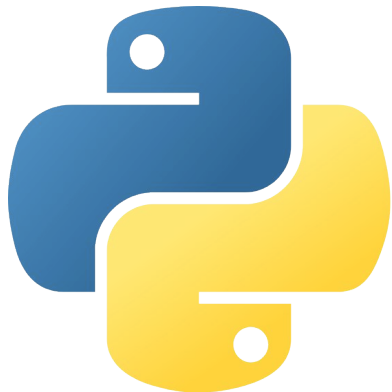


"Sometimes I've believed as many as six impossible things before breakfast." - The Queen of Hearts

Challenges

"How puzzling all these changes are! I'm never sure what I'm going to be, from one minute to another." ~ *Alice*

Early software ecosystem



Kinda depends on what you're doing

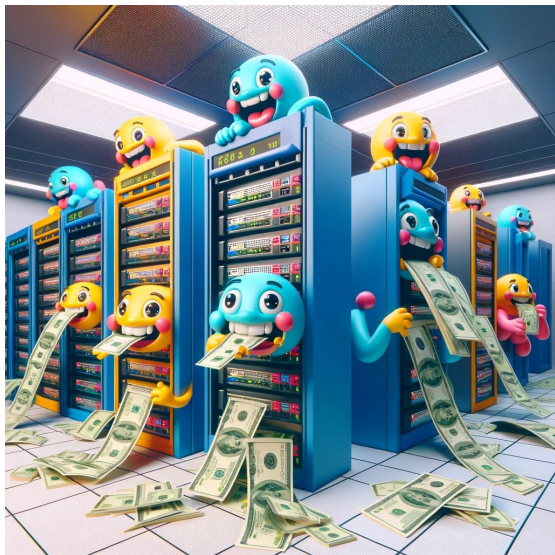
Rapid Disruption



"Tornado in wonderland"
~ ChatGPT4

One of the biggest challenges

COST



- Training & Educational Costs
 - e.g. your SRE team needs to know how to operate vector databases now
- Operational Costs (running things yourself)
 - Storage & Data Costs
 - Computational Resources 😊
 - Training & Fine-Tuning 😱
 - e.t.c.
- API Costs
 - e.g. using an API from OpenAI, Cohere, e.t.c.

... I've run into this firsthand ...

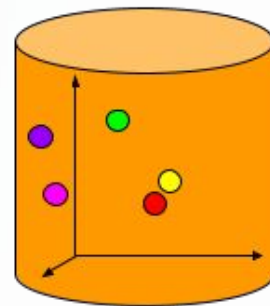
Cost Mitigation Examples

These also help with performance too

- Retrieval Augmented Generation (RAG)
 - Had to cut this, look it up!
 - Semantic Caching
- (We're going to be a *bit* focused on LLMs)



Embedding



Vector DB

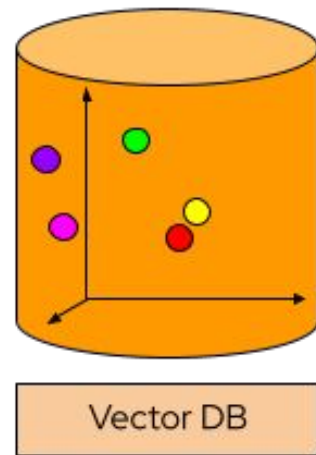
Vector Databases

- Store and retrieve **vectors** (e.g. $[3.2, -1.5, 0.8, 2.6]$)
- **Vectors:**
 - Mathematical representation of a data-point in **multi-dimensional** space
 - Have **proximity** or **distance** to one-another
- Perform “Similarity” or “Nearest Neighbor” search
 - i.e.:

■ Vector A: $[1.2, 0.5, -0.3, 2.0]$

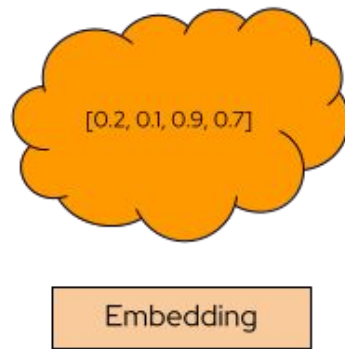
■ Vector B: $[1.3, 0.6, -0.2, 1.9]$

■ Vector C: $[5.0, -2.0, 3.5, -4.0]$

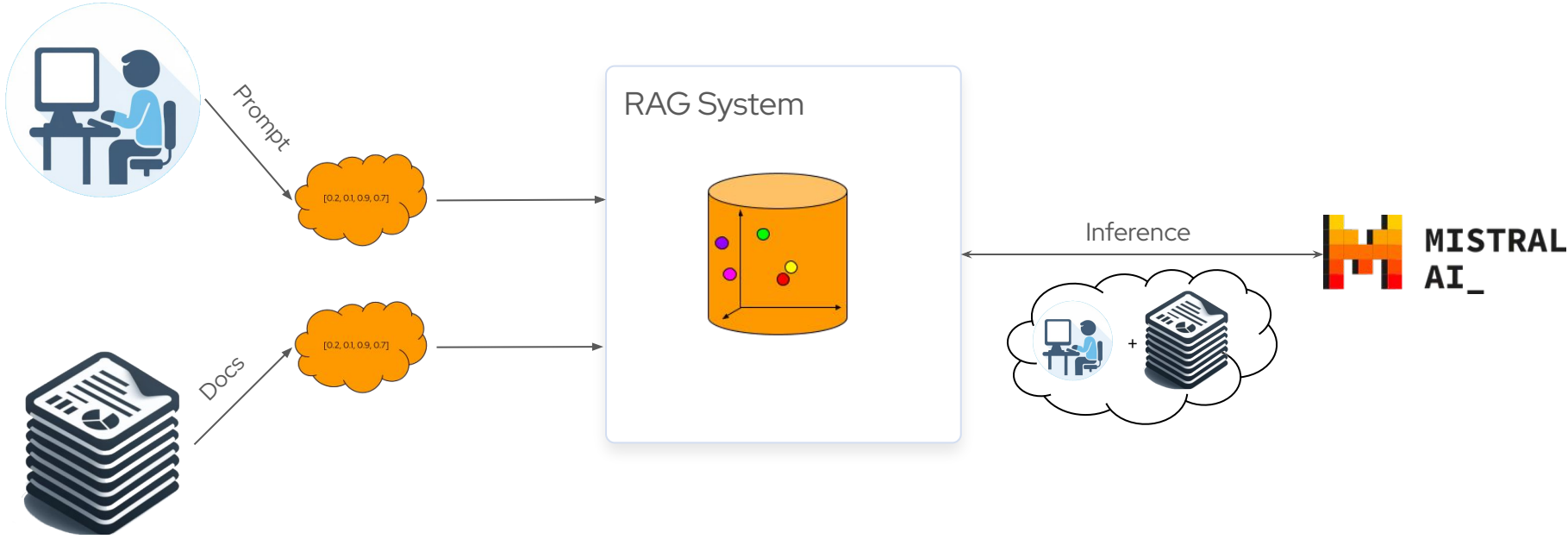


Embeddings

- Are **vectors!** (e.g. [3.2,-1.5,0.8,2.6])
 - Vectorized representations of text
- Are produced by AI models
 - Available locally (e.g. via **transformers**) or via API (e.g. Mistral, OpenAI)
- Can be stored in a vector database, and then searched for proximity
 - Proximity in this case, means **semantic similarity**
 - This is referred to as “Semantic Search”



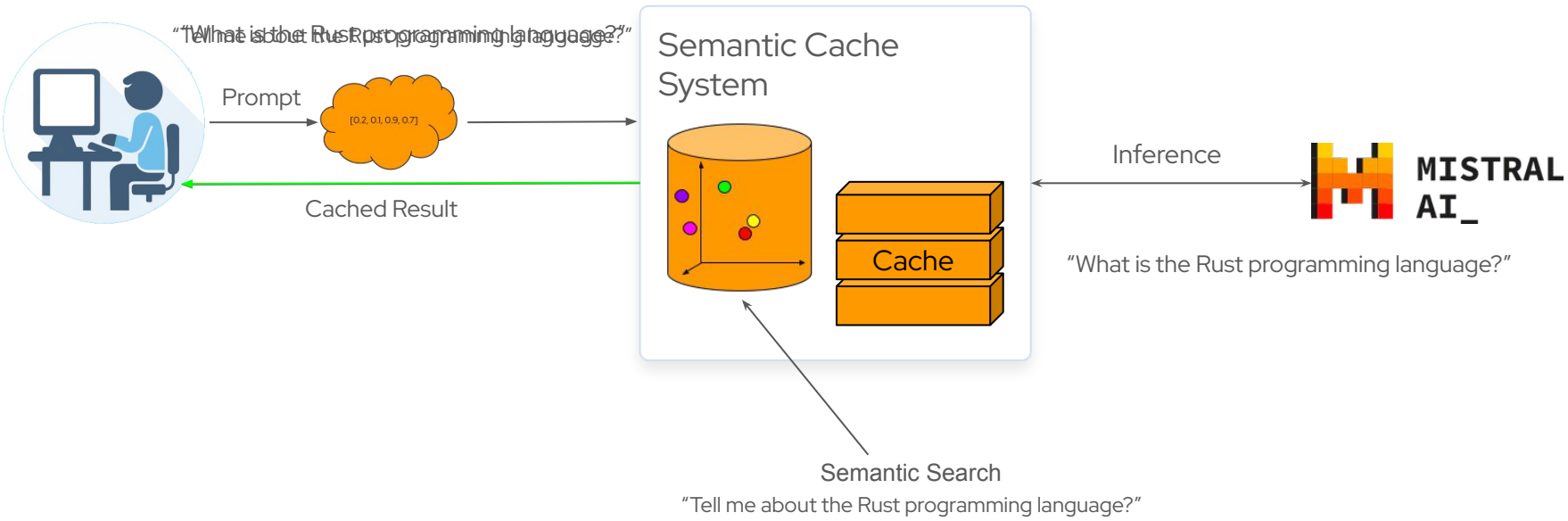
Retrieval Augmented Generation (RAG)



RAG Advantages

- Adaptable
 - Use across various backend providers
- Improved Response Quality
- Reduce Training **Costs**
 - e.g. reduce fine-tune training

Semantic Caching



Semantic Caching Advantages

- Reduced inference **costs**
- Average **latency reduction**

Before you start!

"You're entirely bonkers. But I'll tell you a secret. All the best people are." ~ Alice

Before You Start!

- Give your SREs/Ops people some prep time
 - Know, and explain your infrastructure needs
 - Vector databases aren't relational databases!
 - (well... except when they are)
- **Don't try to build *everything* yourself**
 - Find open-source tools
 - Find **partner** organizations
- Don't forget about **security**
 - note: there have been some supply chain problems recently

Before You Start!

- Keep **model drift** in mind
 - e.g. models becoming less effective over time
- Generally, you may want to avoid deploying as a monolith
 - It can be very tempting to build AI workloads into existing products
 - Make sure you understand the **computational and memory requirements of your workloads**
- *Please make sure this thing is documented!*

Good Luck!



"We're all mad here." ~ The Cheshire Cat

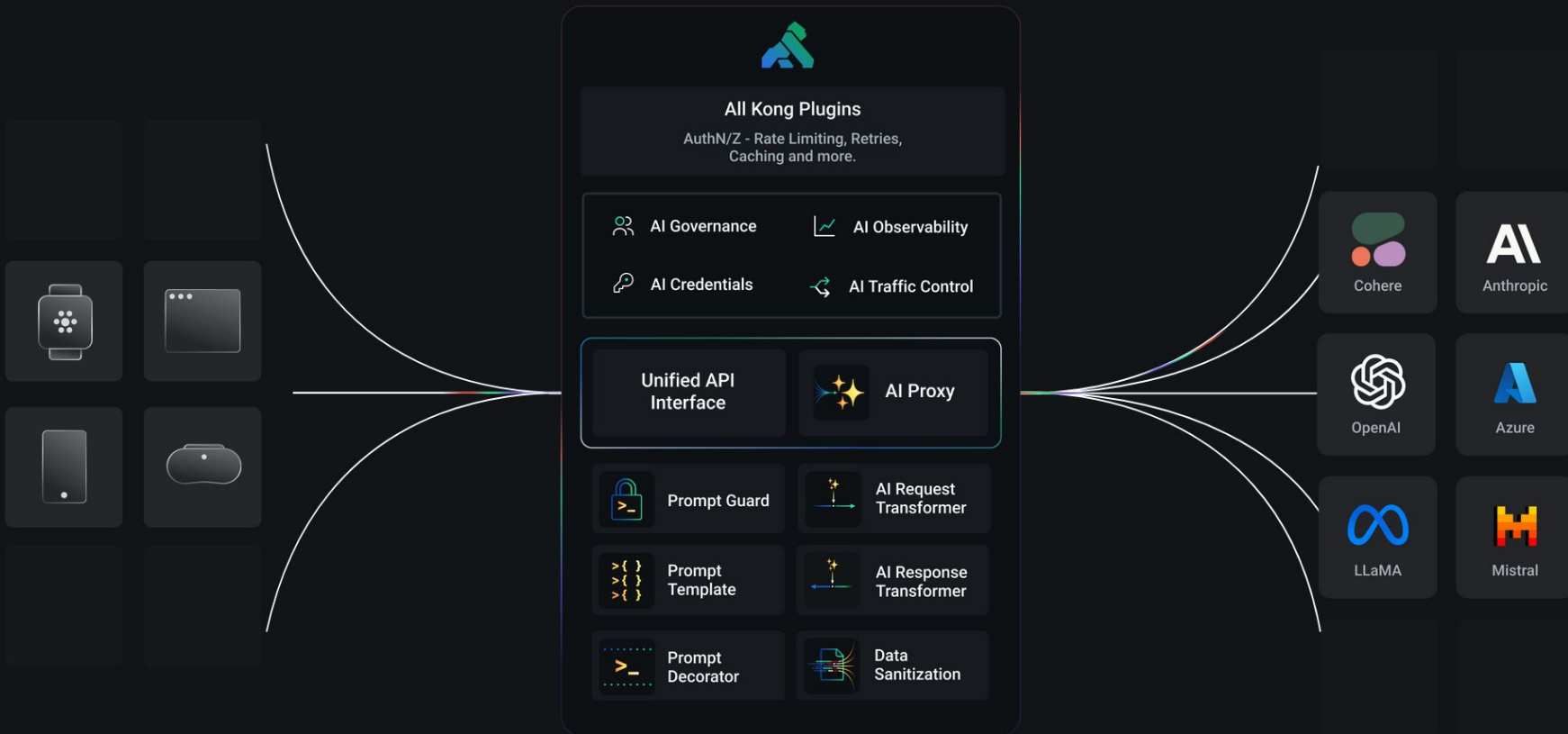
What's Kong's role in all this?

"The only way to achieve the impossible is to believe it is possible." - Mad Hatter

**Is your company asking
"what's our AI story"?**

We'll help you deliver!

Kong AI Gateway



API management for AI

Kong AI Gateway 3.8

NEW

**RATE
LIMITING**

PREMIUM

**SEMANTIC
CACHE**

NEW

MORE LLMs

I'm there!

master

262 Branches 163 Tags

Go to file

Add file

Code

About

The Cloud-Native API Gateway and AI Gateway.

konghq.com/install/#kong-community

- docker
- nginx
- kubernetes
- devops
- microservices
- ai
- microservice
- serverless
- consul
- api-management
- api-gateway
- luajit
- reverse-proxy
- artificial-intelligence
- kong
- apis
- cloud-native
- kubernetes-ingress
- kubernetes-ingress-controller
- ai-gateway

- Readme
- Apache-2.0 license
- Code of conduct
- Security policy
- Activity
- Custom properties
- 37.5k stars
- 1k watching
- 4.7k forks
- Report repository

Releases 138

3.6.1 Latest on Mar 4

+ 137 releases

DOCS



bungle chore(deps): bump resty.acme from 0.12.0 to 0.13.0 (#12909) 6ff84c1 · 2 hours ago 10,377 Commits

.ci	chore(ci): re-enable off tests with the scheduler (#12565)	2 months ago
.devcontainer	Fix devcontainer dependencies.	last month
.github	chore(deps): bump tj-actions/changed-files from 43.0.1 t...	yesterday
autodoc	chore(scripts): remove unused autodoc scripts	10 months ago
bin	chore(test): busted script support system trusted certific...	2 weeks ago
build	chore(deps): bump libexpat from 2.5.0 to 2.6.2 (#12910)	2 hours ago
changelog	chore(deps): bump resty.acme from 0.12.0 to 0.13.0 (#12...	2 hours ago
kong	fix(wasm): disable Lua DNS resolver for proxy-wasm (#12...	2 hours ago
scripts	chore(deps): bump libexpat from 2.5.0 to 2.6.2 (#12910)	2 hours ago
spec	fix(wasm): disable Lua DNS resolver for proxy-wasm (#12...	2 hours ago
t	fix(dns): resolv.conf options timeout: 0 is ignored (#12640)	last month
.bazelignore	build(arm64): utilize self-hosted runner for Amazon Linu...	last year
.bazelrc	chore(build): keep debuginfo for non-release build as we...	last year
.bazelversion	chore(build): bump Bazel from 6.0.0 to 6.1.0	last year
.busted	fix(*) improve Lua paths for dev environments (#2845)	7 years ago
.editorconfig	chore(development): add vim modelines to bin/* (#10183)	last year
.gitignore	tests(wasm): add test filters/fixtures	9 months ago

Oh snap! serverless pitc!

Further down the rabbit hole

"If you don't know where you are going any road can take you there." ~ *The Cheshire Cat*

Want to stay in Wonderland?

- The data science stuff isn't that scary!
- Lots of good free courses out there
- High level tooling exists to help you get started fast
- Communities are out there to help!

fast.ai—Making neural nets uncool again

- **Courses:** [Practical Deep Learning for Coders](#); [From Deep Learning Foundations to Stable Diffusion](#)
- **Software:** [fastai for PyTorch](#); [nbdev](#)
- **Book:** [Practical Deep Learning for Coders with fastai and PyTorch](#)
- **In the news:** [The Economist](#); [The New York Times](#); [MIT Tech Review](#)
- **Corporate partner program:** Get help with fast.ai technologies & courses from the [partner program](#)

Categories

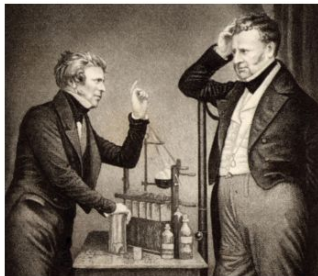
- All (149)
- advice (24)
- ai-in-society (44)
- courses (40)
- education (5)
- health (14)
- machine learning (1)
- technical (42)

Blog

Dec 12, 2023
Jeremy Howard

A new old kind of R&D lab

Answer.AI is a new kind of AI R&D lab which creates practical end-user products based on foundational research breakthroughs.



Sep 4, 2023
Jeremy Howard and
Jonathan Whitaker

Can LLMs learn from a single example?

TECHNICAL



Practical Deep Learning

Part 1

- 1: Getting started
- 2: Deployment
- 3: Neural net foundations
- 4: Natural Language (NLP)
- 5: From-scratch model
- 6: Random forests
- 7: Collaborative filtering
- 8: Convolutions (CNNs)
- Bonus: Data ethics

Summaries

Part 2

- Part 2 overview
- 9: Stable Diffusion
- 10: Diving Deeper
- 11: Matrix multiplication
- 12: Mean shift clustering
- 13: Backpropagation & MLP
- 14: Backpropagation
- 15: Autoencoders
- 16: The Learner framework
- 17: Initialization/normalization
- 18: Accelerated SGD & ResNets
- 19: DDPM and Dropout
- 20: Mixed Precision
- 21: DDIM

Practical Deep Learning

A free course designed for people with some coding experience, who want to learn how to apply deep learning and machine learning to practical problems.

New!

We just launched a new >30 hour video course for more experienced students:

[Practical Deep Learning for Coders part 2: *Deep Learning Foundations to Stable Diffusion*](#)



This free course is designed for people (and bunnies!) with some coding experience who want to learn how to apply deep learning and machine learning to practical problems.

Deep learning can do all kinds of amazing things. For instance, all illustrations throughout this website are made with deep learning, using [DALL-E 2](#).

Welcome!

Practical Deep Learning for Coders 2022 part 1, recorded at the [University of Queensland](#), covers topics such as how to:



- Build and train deep learning models for computer vision, natural language processing, tabular analysis, and collaborative filtering problems
- Create random forests and regression models
- Deploy models
- Use PyTorch, the world's fastest growing deep learning software, plus popular libraries like fastai and Hugging Face

On this page

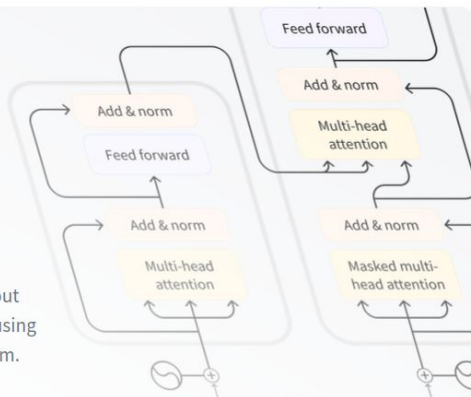
- [Welcome!](#)
- [Real results](#)
- [Your teacher](#)
- [Is this course for me?](#)
- [The software you will be using](#)
- [Why deep learning?](#)
- [What you will learn](#)
- [How do I get started?](#)

 Report an issue


Learn

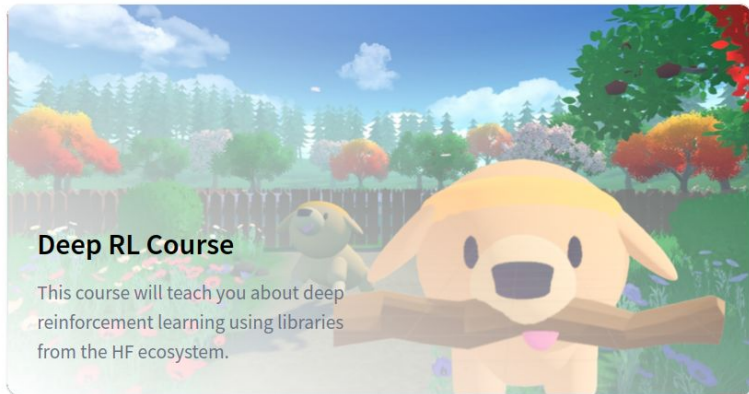
NLP Course

This course will teach you about natural language processing using libraries from the HF ecosystem.



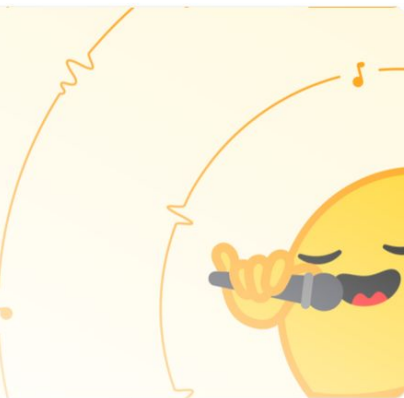
Deep RL Course

This course will teach you about deep reinforcement learning using libraries from the HF ecosystem.



Audio Course

Learn to apply transformers to audio data using libraries from the HF ecosystem.



Open-Source AI Cookbook

A collection of Open Source-powered recipes by community for AI builders.





Hugging Face

The AI community building the future.

👥 31k followers

📍 NYC + Paris

🔗 <https://huggingface.co/>

🐦 @huggingface

Verified

Pinned

📄 **transformers** Public

🤖 Transformers: State-of-the-art Machine Learning for Pytorch, TensorFlow, and JAX.

🐍 Python ☆ 125k 🍷 24.8k

📄 **diffusers** Public

🤖 Diffusers: State-of-the-art diffusion models for image and audio generation in PyTorch and FLAX.

🐍 Python ☆ 22.5k 🍷 4.6k

📄 **datasets** Public

🤖 The largest hub of ready-to-use datasets for ML models with fast, easy-to-use and efficient data manipulation tools

🐍 Python ☆ 18.4k 🍷 2.5k

📄 **peft** Public

🤖 PEFT: State-of-the-art Parameter-Efficient Fine-Tuning.

🐍 Python ☆ 13.8k 🍷 1.3k

📄 **accelerate** Public

🚀 A simple way to launch, train, and use PyTorch models on almost any device and distributed configuration, automatic mixed precision (including fp8), and easy-to-configure FSDP and DeepSpeed support

🐍 Python ☆ 6.9k 🍷 797

📄 **optimum** Public

🚀 Accelerate training and inference of 🤖 Transformers and 🤖 Diffusers with easy to use hardware optimization tools

🐍 Python ☆ 2.1k 🍷 364



Staying Connected

- Publications (Stuff I follow)
 - Wired's Artificial Intelligence Category
 - Mistral AI's News Page
 - OpenAI's Blog
 - Anthropic's News
 - AI Navigator
 - Huggingface's Blog
- Social Media
 - X/Twitter
 - Reddit
 - YouTube
- Perhaps most importantly... **joining a community!**
 - GitHub
 - Huggingface
 - e.t.c.

Start Experimenting!



"I don't see how he can ever finish, if he doesn't begin." ~ Alice

**Feeling a little overwhelmed,
like there's so much going on in AI?**

Feeling a little overwhelmed
There's never been a better, or easier time to start.
like there's so much going on in AI?
There's LOTS of help out there now that
wasn't there a few years ago.
Reach out!

Questions?



Day 0

Plan/Deploy



Day 1

Test/Monitor



Day 2

Production!

Days Of Operation

It's a new scene, everyone's experience is going to be pretty unique...

Day 0

- Don't forget about **security**
 - note: there have been some supply chain problems recently
- Generally, you may want to avoid deploying as a monolith
 - It can be very tempting to build AI workloads into existing products
 - Make sure you understand the **computational and memory requirements of your workloads**

Day 1

- Test and verify scale
 - Components like vector databases can scale very differently from traditional databases
 - e.g. “the curse of dimensionality”
 - You don’t always get “deterministic” output in the classic sense you might be thinking
- Get a good grasp on your operational costs
 - And (obviously) try to lower them so they don’t catch you off guard later

Day 2

- Good luck!
- Don't forget about **model drift**
- *Please* **make sure this thing is documented!**

Kong Gateway ▾

VERSION 3.6.X (LATEST) ▾

🚩 Introduction ▾

🎓 Get Started with Kong ▲

Get Kong

Services and Routes

Rate Limiting

Proxy Caching

Key Authentication

Load-Balancing

AI Gateway

📦 Install Kong ▾

📁 Kong in Production ▾

📁 Kong Gateway Enterprise ▾

📁 Kong Manager ▾

AI Gateway

Kong AI Gateway is a powerful set of features built on top of [Kong Gateway](#), designed to help developers and organizations effectively adopt AI capabilities quickly and securely.

Quick links:

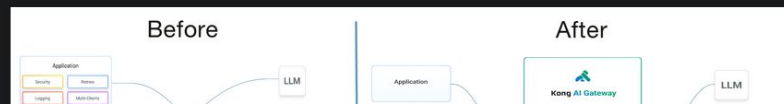
- [Get started](#) in one minute with our guide
- [Watch video tutorials](#) [🔗](#) in the AI Gateway product page
- [View the AI plugins](#) in the Kong Plugin Hub

With the rapid emergence of multiple AI LLM providers (including open source and self-hosted models), the AI technology landscape is fragmented and lacking in standards and controls. This significantly complicates how developers and organizations use and govern AI services. Kong Gateway's broad API management capabilities and plugin extensibility model make it well suited to provide AI-specific API management and governance services.

While AI providers don't conform to a standard API specification, the AI Gateway provides a normalized API layer allowing clients to consume multiple AI services from the same client code base. The AI Gateway provides additional capabilities for credential management, AI usage observability, governance, and tuning through prompt engineering. Developers can use no-code AI Plugins to enrich existing API traffic, easily enhancing their existing application functionality.

You can enable the AI Gateway features through a set of modern and specialized plugins, using the same model you use for any other Kong Gateway plugin. When deployed alongside existing [Kong Gateway plugins](#), Kong Gateway users can quickly assemble a sophisticated AI management platform without custom code or deploying new and unfamiliar tools.

ON THIS PAGE >

[Getting started](#)[AI Gateway Capabilities](#)[AI Provider Proxy](#)[AI usage governance](#)

Manual Usage

- Research Aid
- Email writing assistant 😊
- Coding Assistant 😱
- e.t.c.

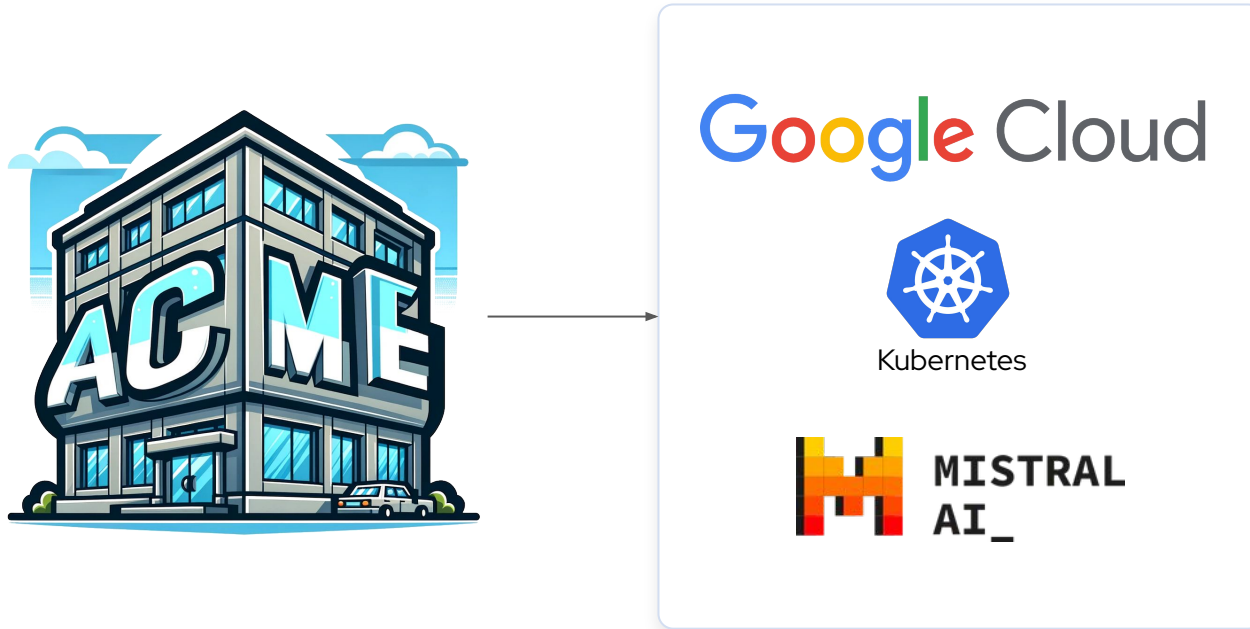
... some of this is fraught with peril ...

... and it's already happening in your org ... 😊

AI is now consumable as APIs



OR internally on your infrastructure



So what can we build?

Large Action Models

- AI taking *actions* on behalf of users
- It's conceivable that LAMs will be at the center of personal device automation in the coming years?
For Example: something I think might be a growing disruptor:
- Will this change a lot of how we do things soon?



You've built something!



Yay!



Let's just get something out of the way..

 PyTorch



Building & Delivering

"You're entirely bonkers. But I'll tell you a secret. All the best people are." ~ Alice