



# Gen AI for Software Development





Page

Messages

Notifications 3

Insights

Publishing Tools

Settings

Help ▾



somkiat.cc

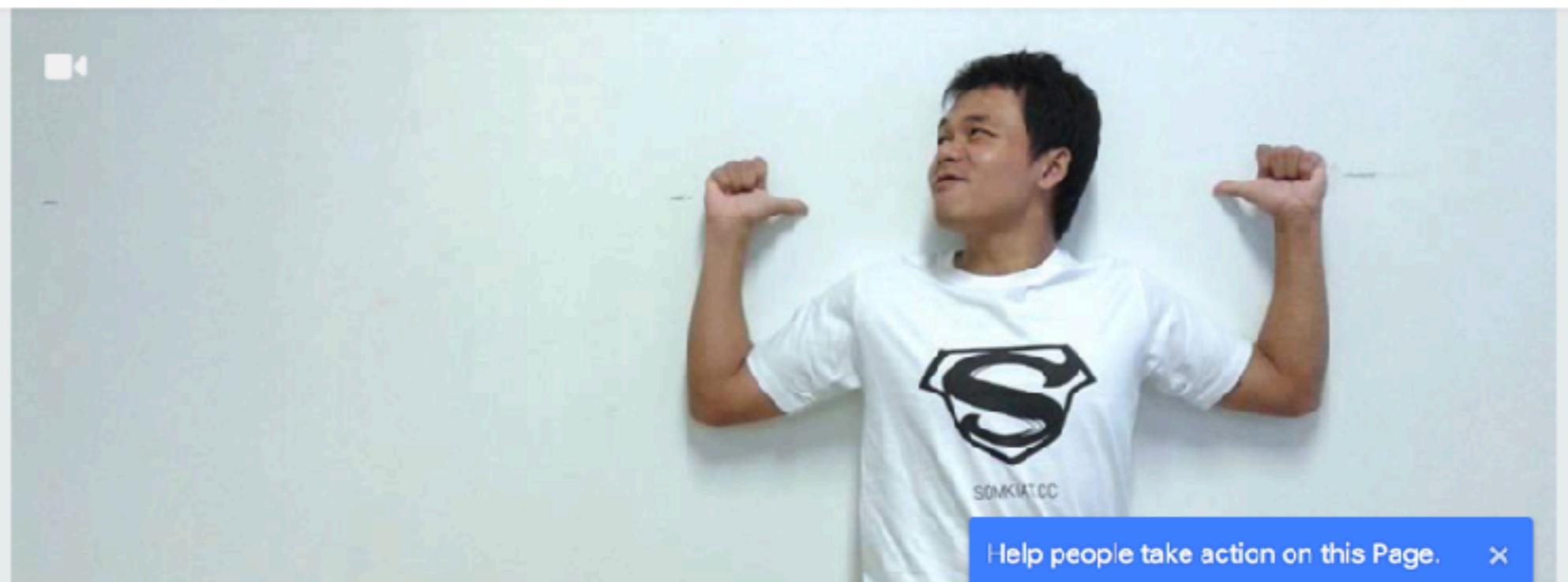
@somkiat.cc

Home

Posts

Videos

Photos



 Liked ▾

 Following ▾

 Share

...

+ Add a Button

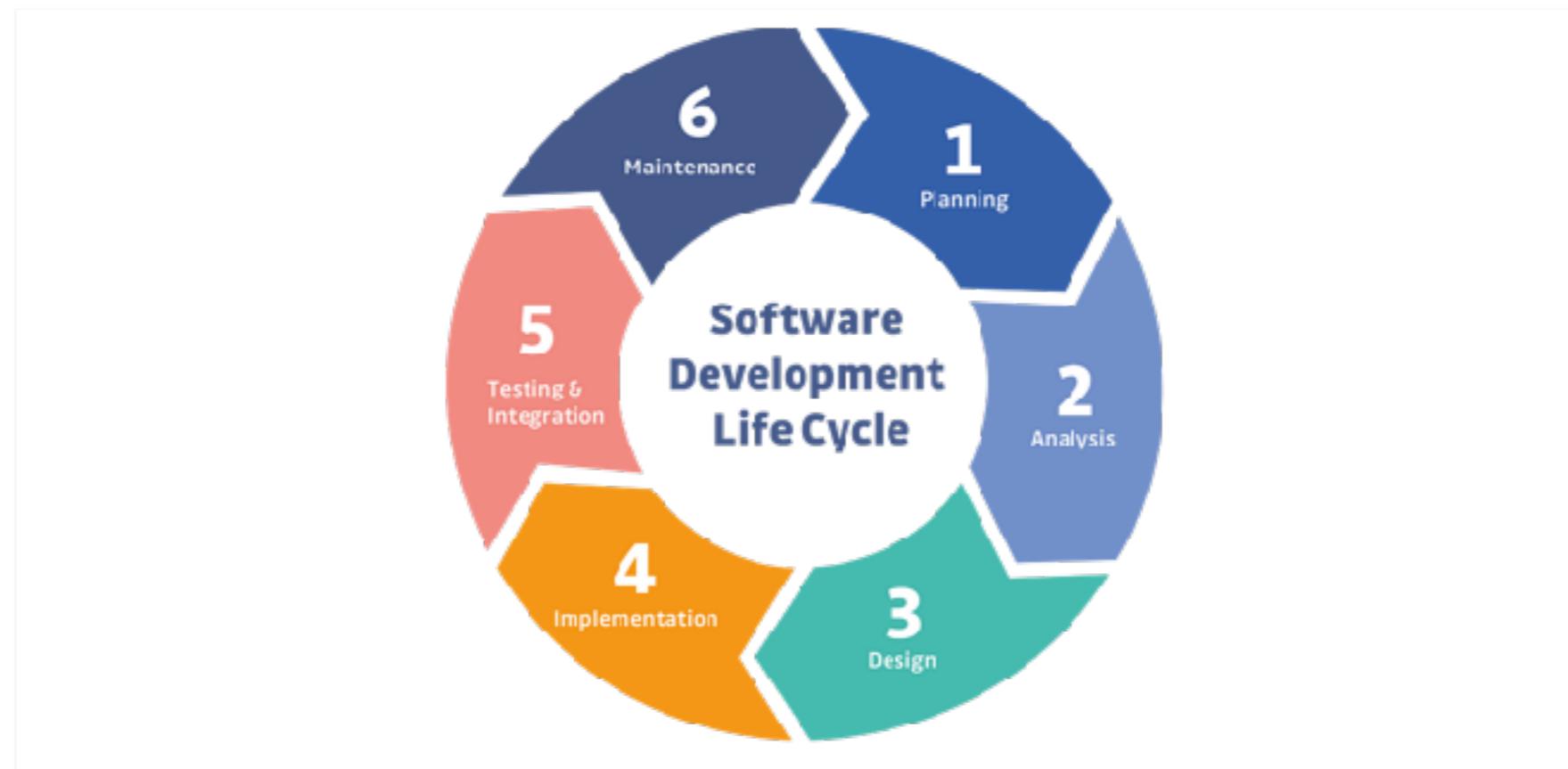


**[https://github.com/up1/  
workshop-ai-with-technical-team](https://github.com/up1/workshop-ai-with-technical-team)**



# Goals

Integrate Generative AI in Software Development  
Optimize code quality  
Team up with AI on coding tasks  
Develop innovative solutions



# Software Development

Requirement

Design

Develop

Testing

Deploy

Generative AI

Improve Productivity ... (Replace human !!)



# Learning Path

AI/ML

AI Model

Prompt Engineer

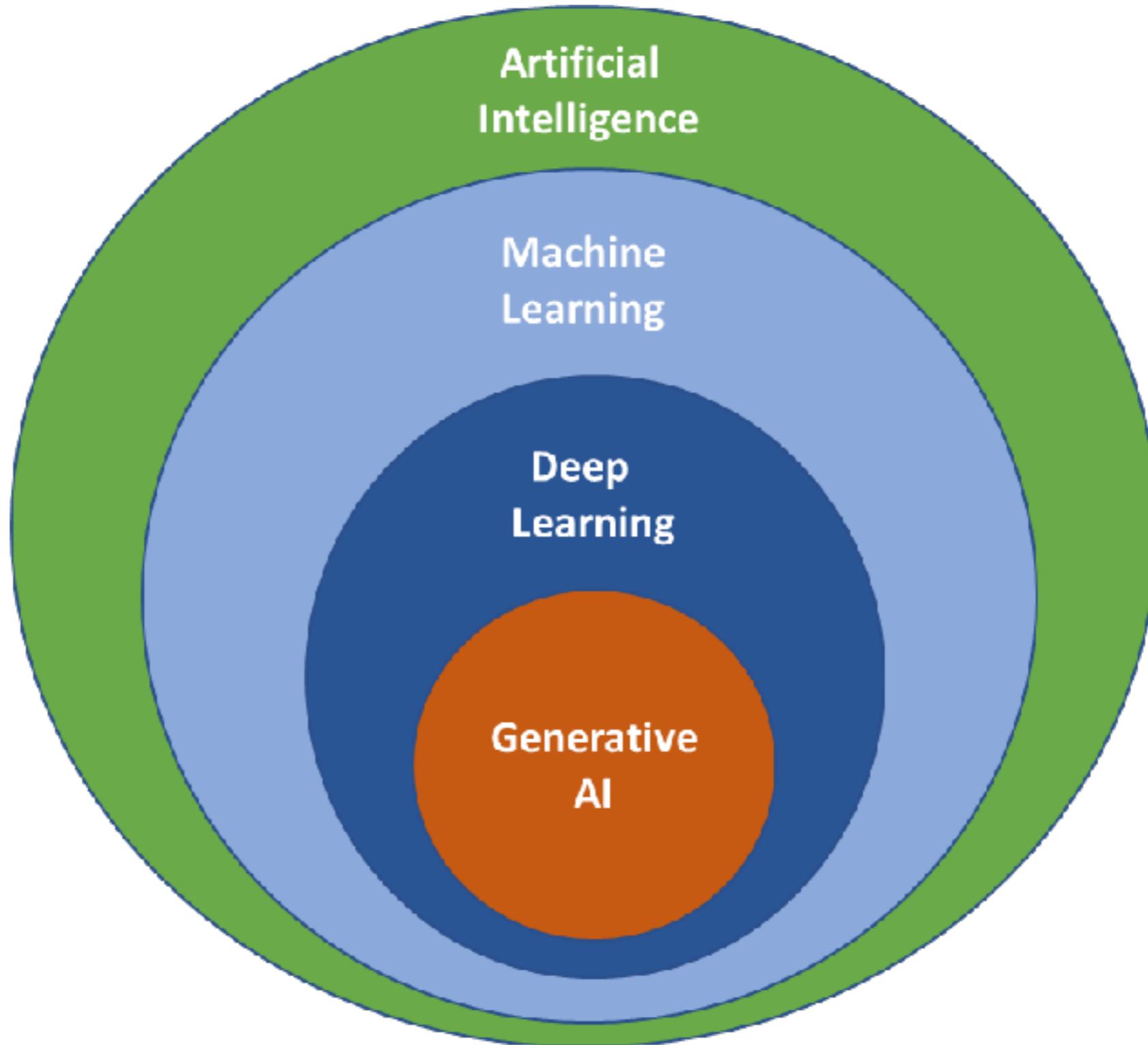
AI in Software development

Develop AI/LLM app

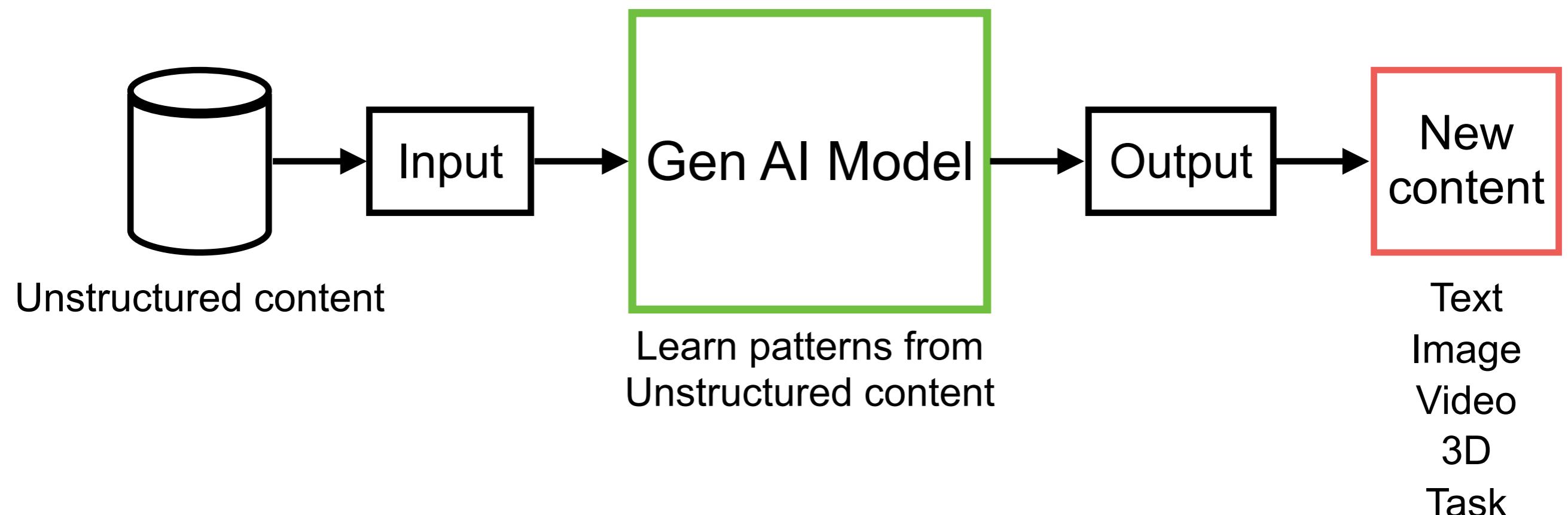
RAG

Use cases and workshops





# Generative AI



<https://grow.google/ai-essentials/>



# Generative AI

**LLMs**

**Large Language Models**

Text generation

Code generation

Chatbot

Conversation AI

**GANs**

**Generative Adversarial Network**

Image generation

Deep fake

Art creation

Simulate financial market

**VAEs**

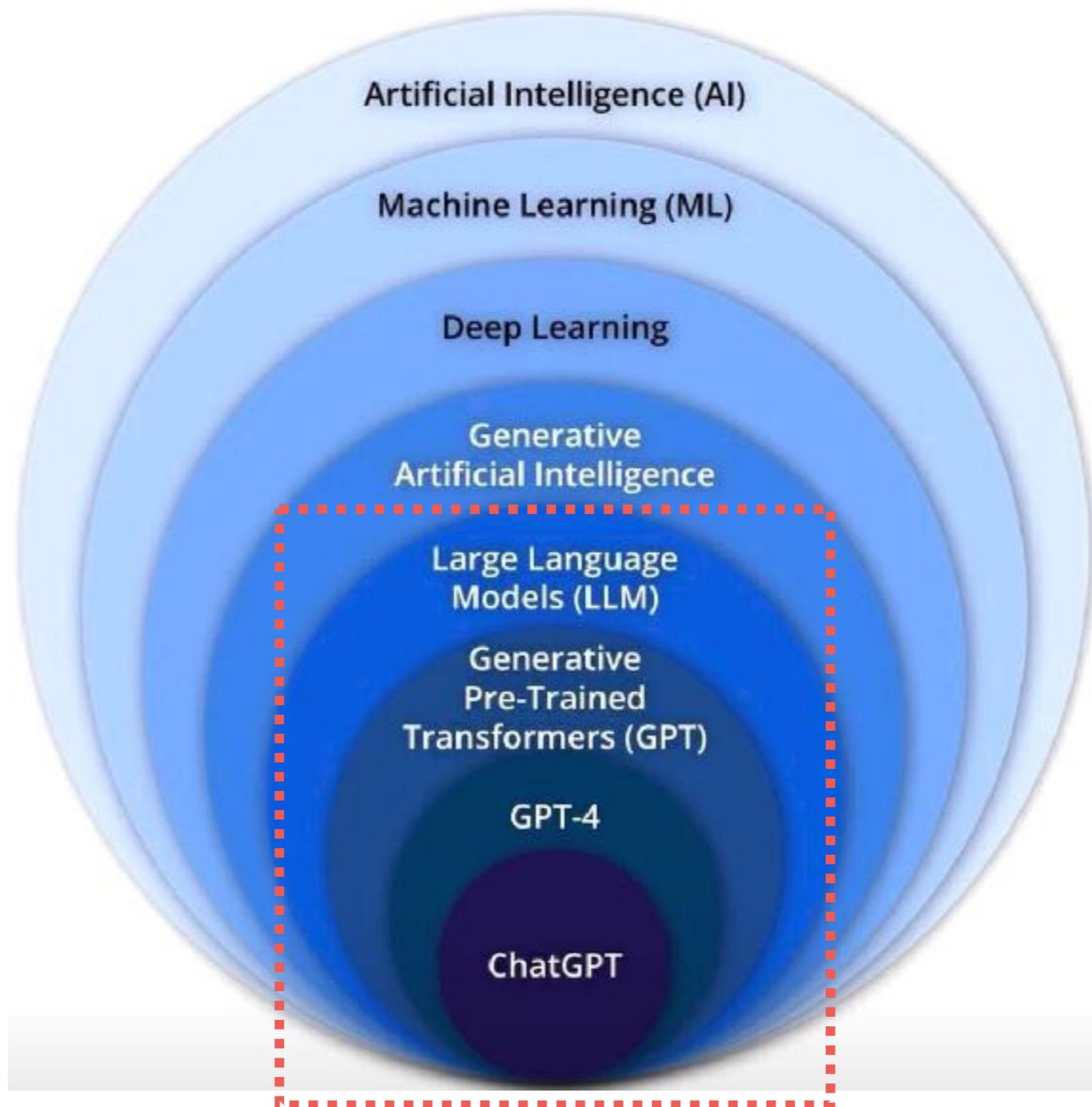
**Variational Autoencoders**

Data compression

Synthetic data generation

Image reconstruction





# Large Language Model (LLM)

Type of AI model

Process, understand

Generate human readable data

LLM

Training data !!



# LLMs in industry

Model name	Company
Bidirectional Encoder Representation from Transformers (BERT)	Google AI
Generative Pre-trained transformer-5 (GPT-5)	OpenAI
Pathways Language Model-E (PaLM-E)	Google AI
BLOOM	NVIDIA AI
Llama 4	Facebook
Claude 4.5 Sonnet	Anthropic



# LLM Development timeline

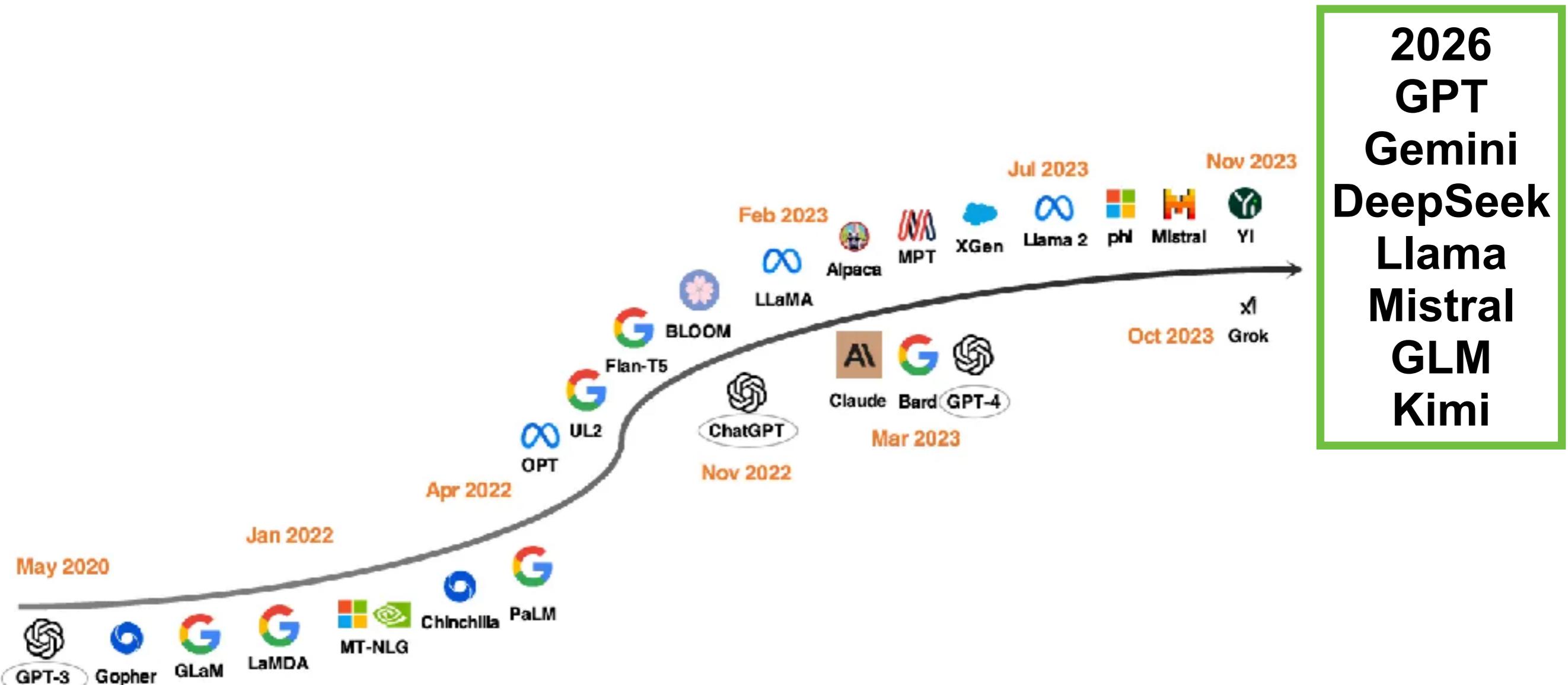


Figure 3: LLM development timeline. The models below the arrow are closed-source while those above the arrow are open-source.

<https://arxiv.org/abs/2311.16989>



# Let's go !!



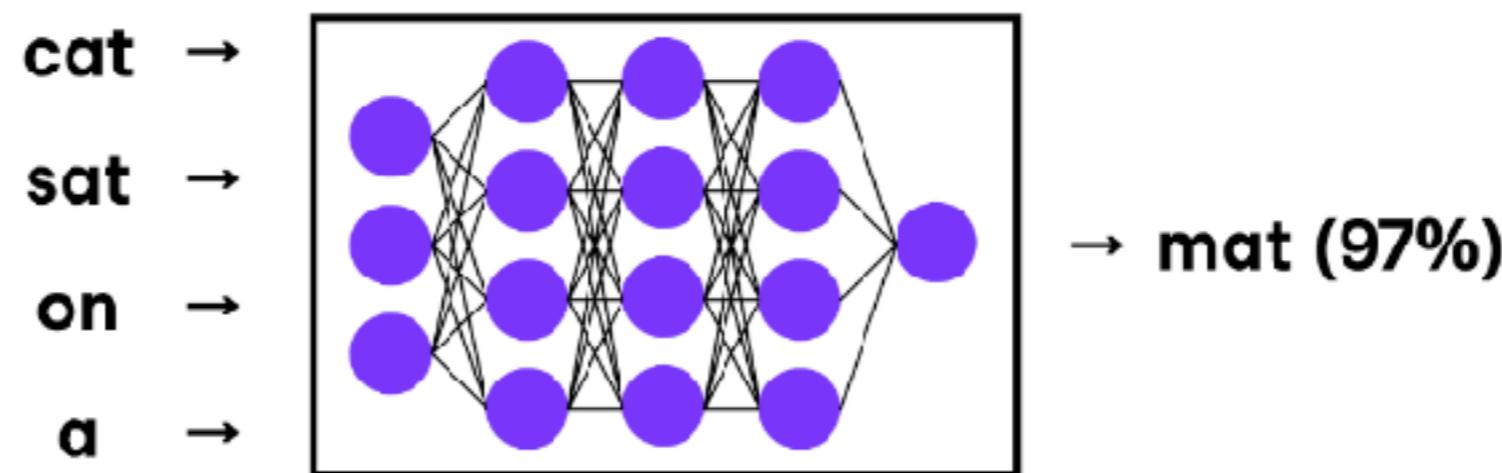
# ສືເໜືອງ



# Large Language Model (LLM)

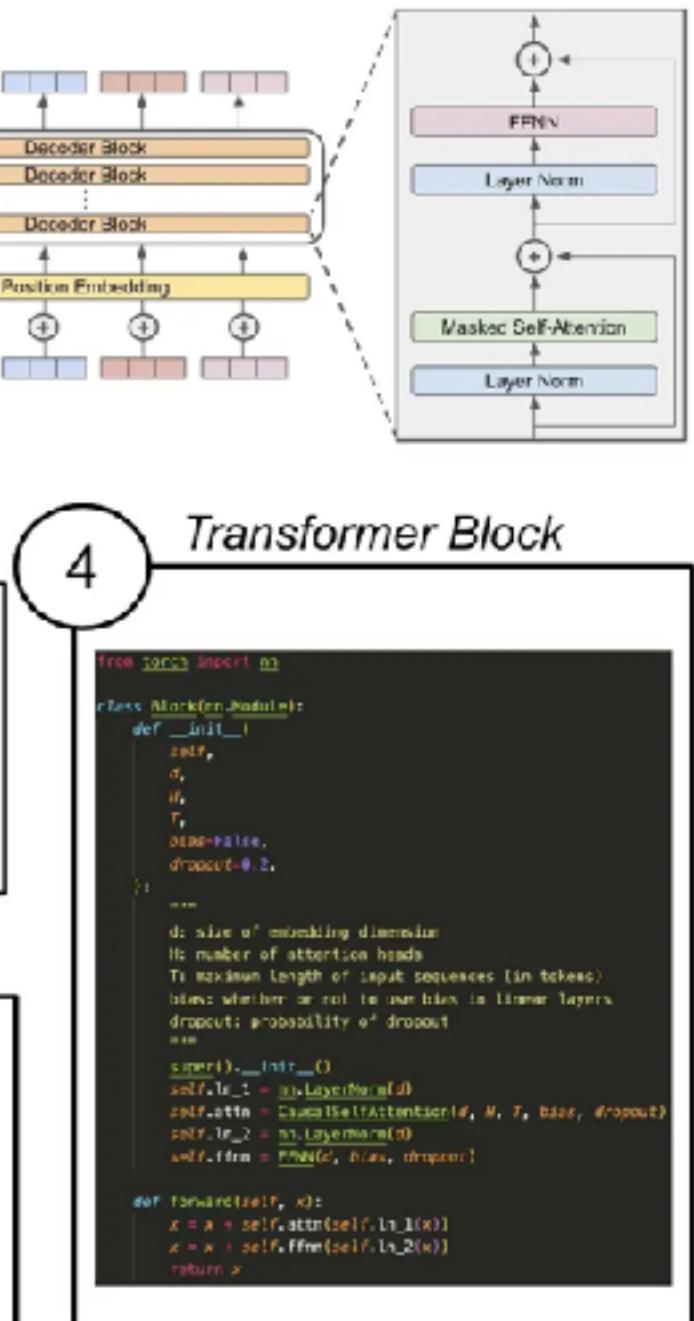
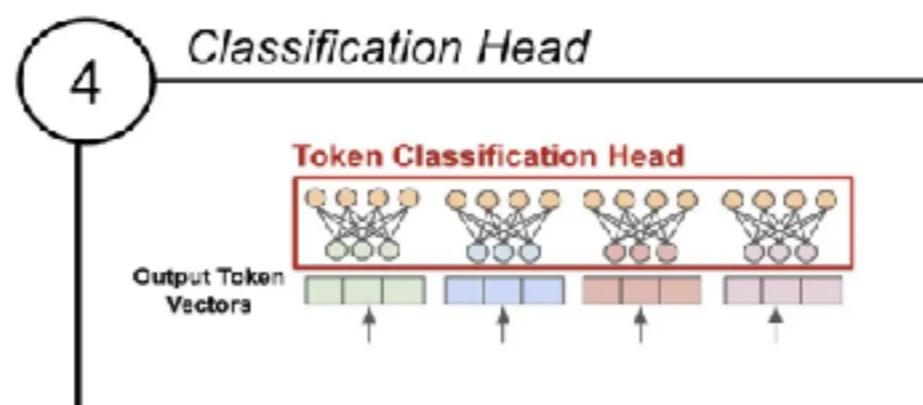
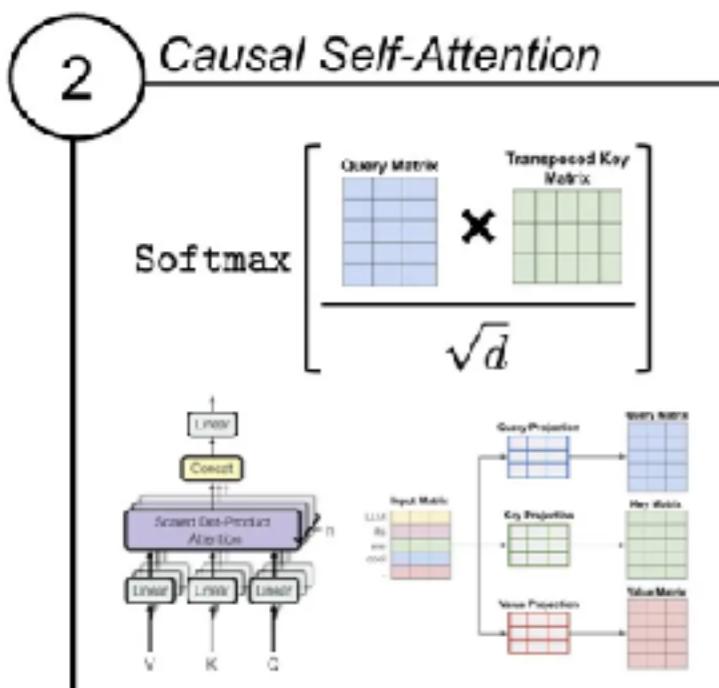
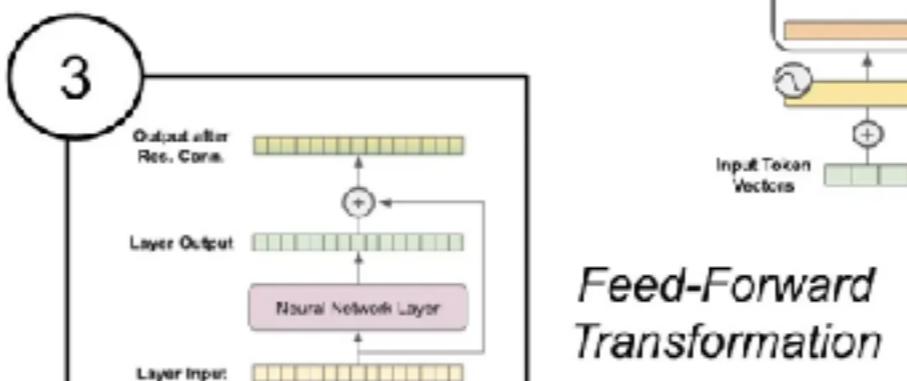
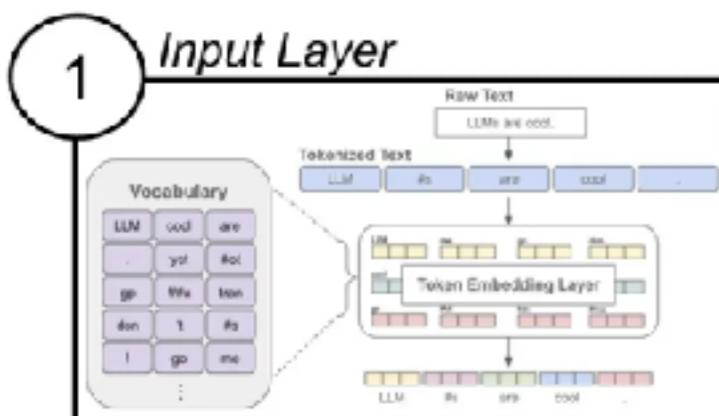
Neural network

Predicts the next word in a sequence



# LLM components !!

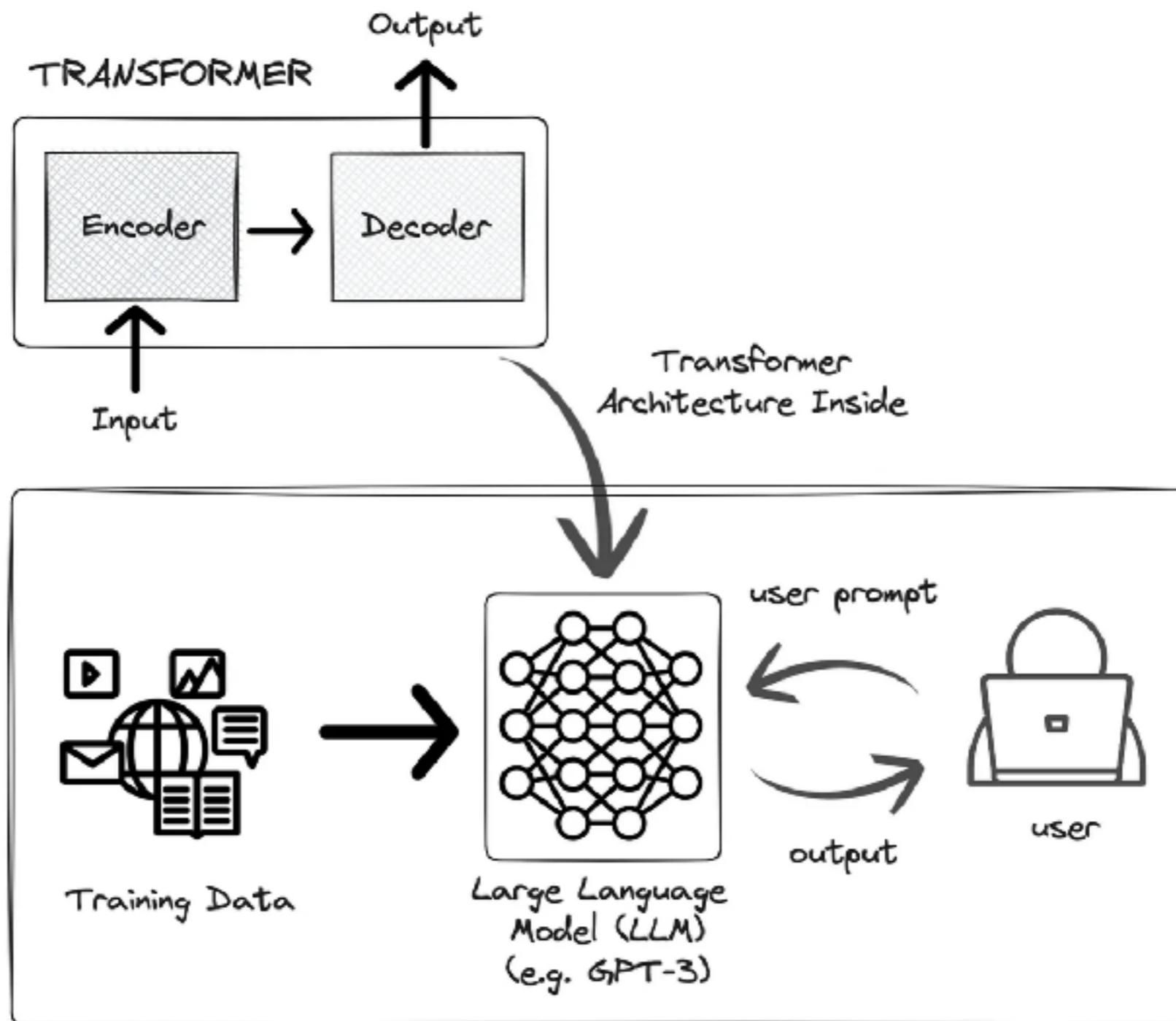
## Components of the Decoder-only Transformer



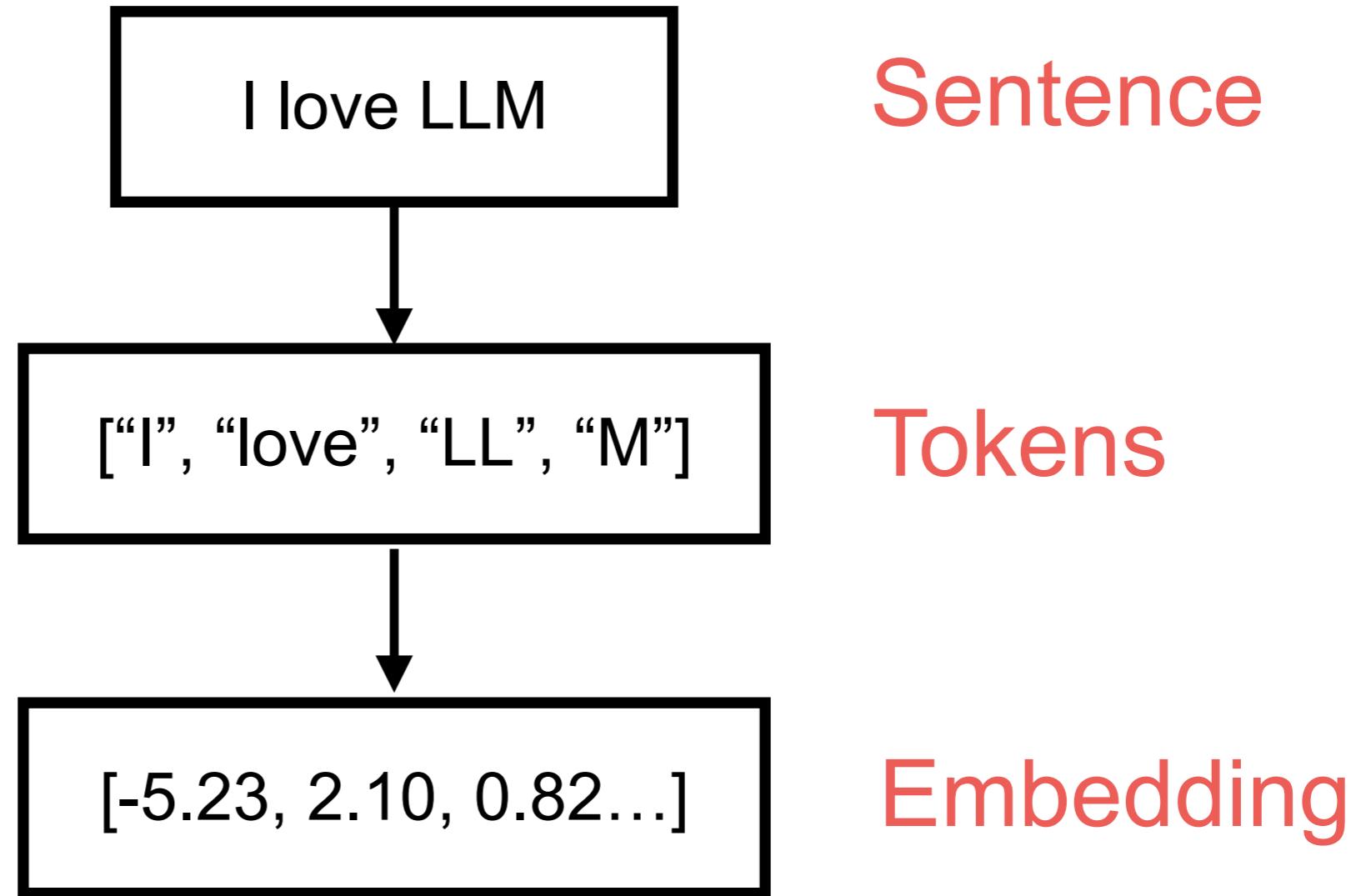
<https://stackoverflow.blog/2024/08/22/lms-evolve-quickly-their-underlying-architecture-not-so-much>



# Transformer inside



# Transformer process



# OpenAI Tokenizer

GPT-4o & GPT-4o mini (coming soon)    **GPT-3.5 & GPT-4**    GPT-3 (Legacy)

ประเทศไทย

[Clear](#) [Show example](#)

<b>Tokens</b>	<b>Characters</b>
10	9

ประเทศไทย

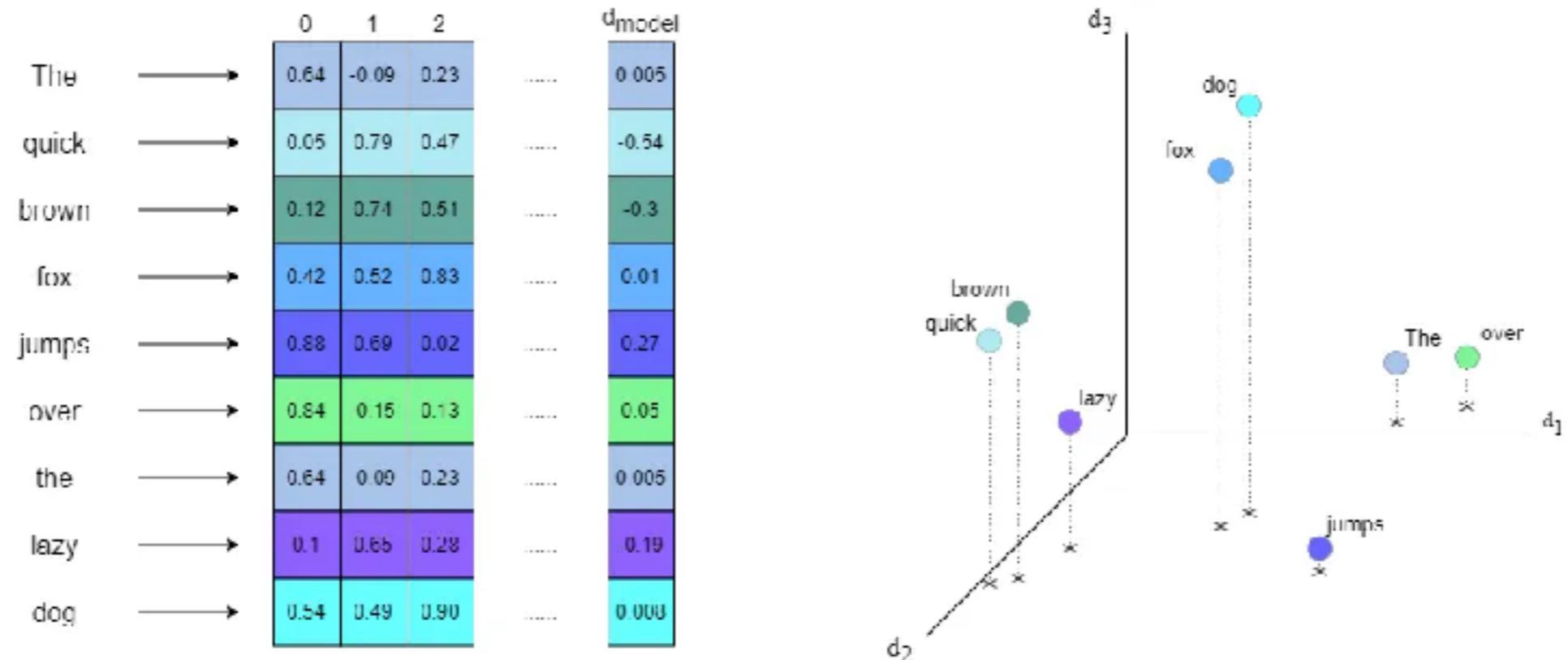
[Text](#) [Token IDs](#)

<https://platform.openai.com/tokenizer>



# Embedding ?

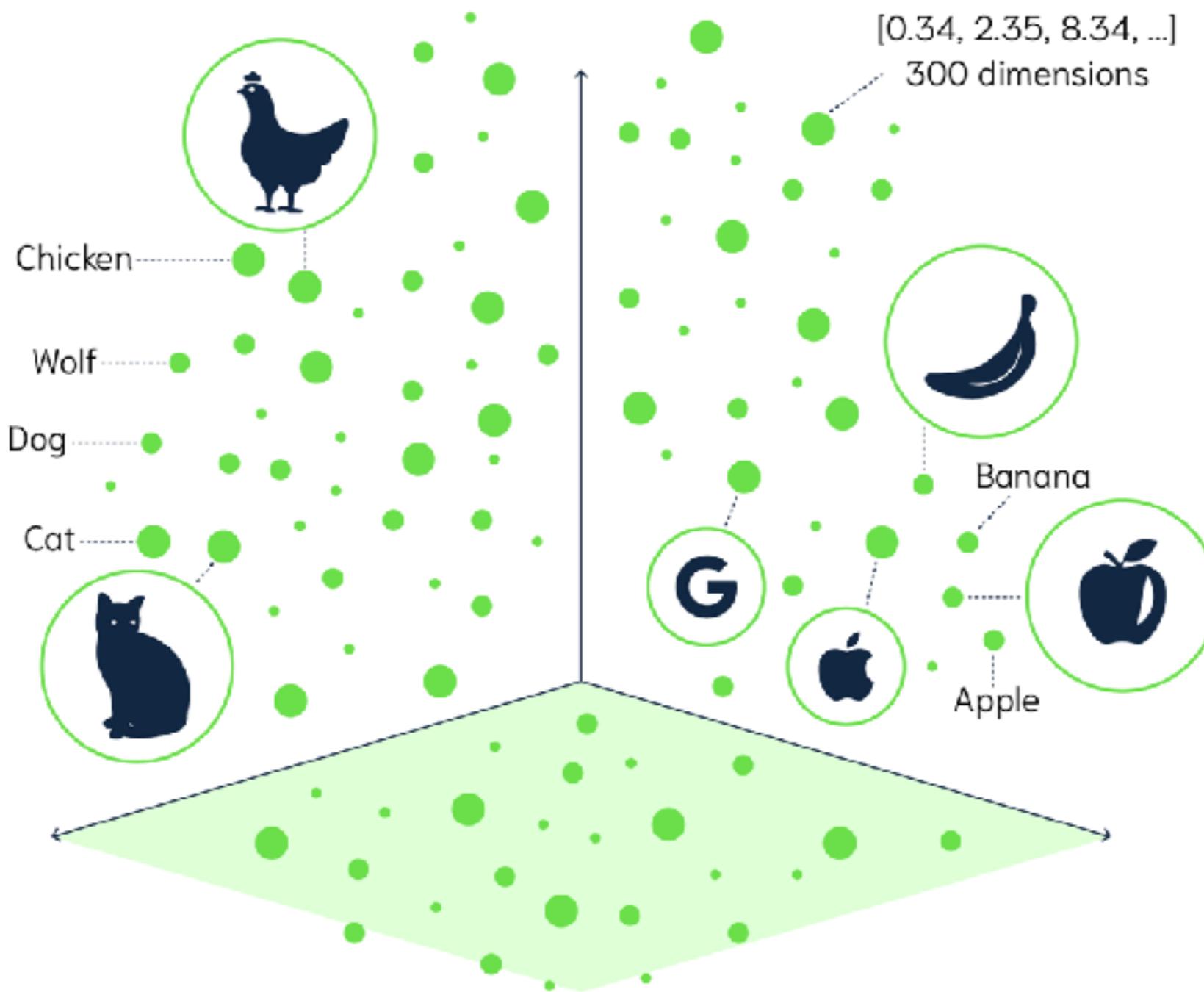
Map items of unstructured data to high-dimensional real vectors



<https://towardsdatascience.com/transfomers-in-depth-part-1-introduction-to-transformer-models-in-5-minutes-ad25da6d3cca>



# Visual of Vector space



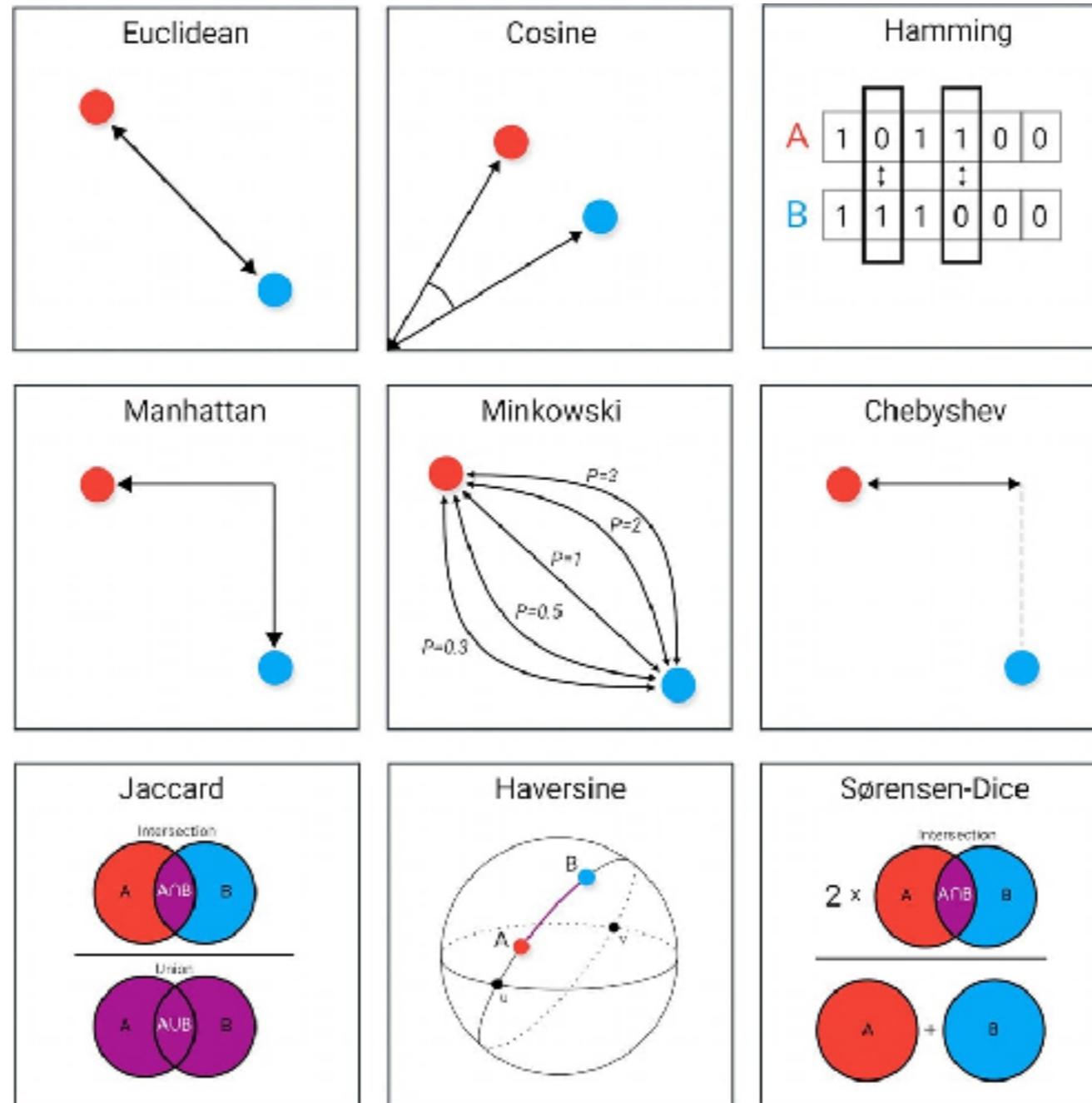
# Embedding Leaderboard

Rank (Box...)	Model	Zero-shot	Memory U...	Number of P...	Embedding D...	Max Tokens	Mean (T...	Mean (TaskT...	Bitext ...	Classific...
1	<a href="#">llama-embed-nemotron-Bb</a>	99%	28629	7B	4096	32768	69.45	61.09	<b>81.72</b>	73.21
2	<a href="#">Qwen3-Embedding-8B</a>	99%	28866	7B	4096	32768	<b>70.58</b>	<b>61.69</b>	80.89	<b>74.00</b>
3	<a href="#">gemini-embedding-001</a>	99%	Unknown	Unknown	3072	2048	68.37	59.59	79.28	71.82
4	<a href="#">Qwen3-Embedding-4B</a>	99%	15341	4B	2560	32768	69.45	60.86	79.36	72.33
5	<a href="#">Qwen3-Embedding-0.6B</a>	99%	2272	595M	1024	32768	64.34	56.01	72.23	66.83
6	<a href="#">gte-Qwen2-7B-instruct</a>	⚠ NA	29040	7B	3584	32768	62.51	55.93	73.92	61.55
7	<a href="#">Ling-Embed-Mistral</a>	99%	13563	7B	4096	32768	61.47	54.14	70.34	62.24
8	<a href="#">multilingual-e5-large-instruct</a>	99%	1068	560M	1024	514	63.22	55.08	80.13	64.94
9	<a href="#">embeddinggemma-300m</a>	99%	578	367M	768	2048	61.15	54.31	64.40	60.90
10	<a href="#">SFR-Embedding-Mistral</a>	96%	13563	7B	4096	32768	60.99	53.92	70.00	60.02
11	<a href="#">text-multilingual-embedding-002</a>	99%	Unknown	Unknown	768	2048	62.15	54.25	70.73	64.64

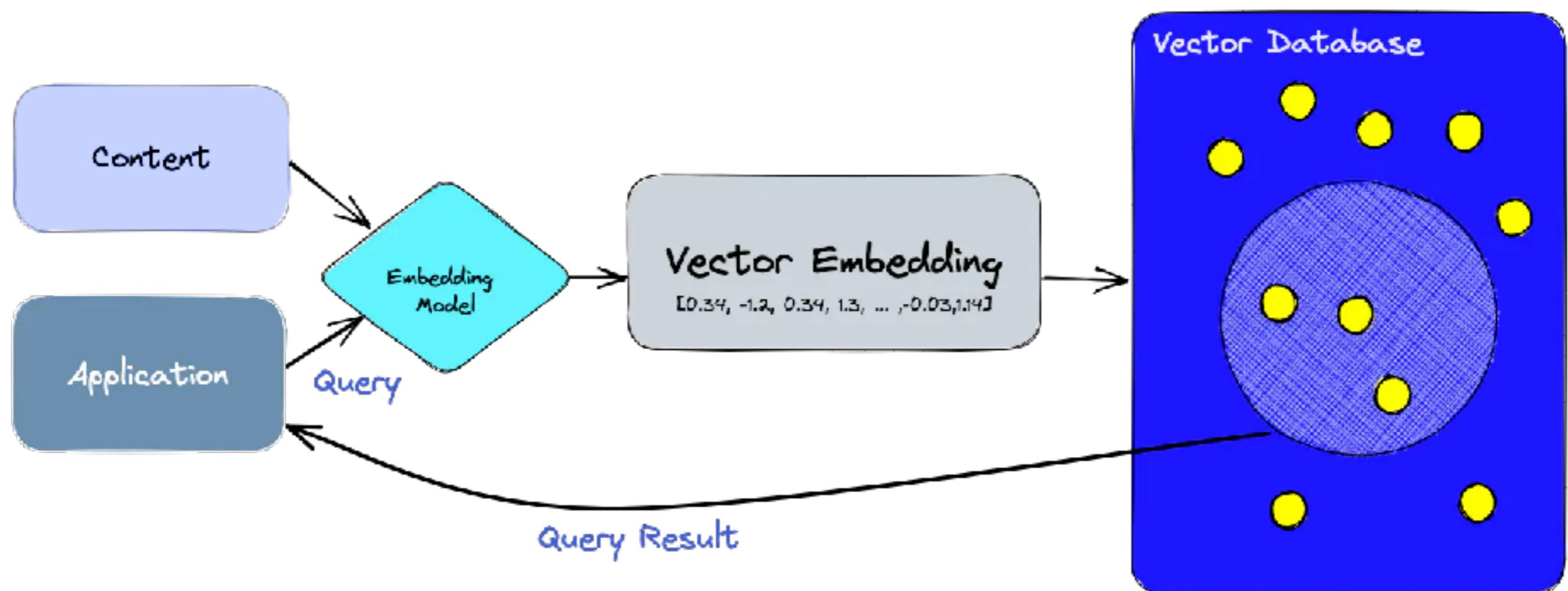
<https://huggingface.co/spaces/mteb/leaderboard>



# Distance measure in Data Science

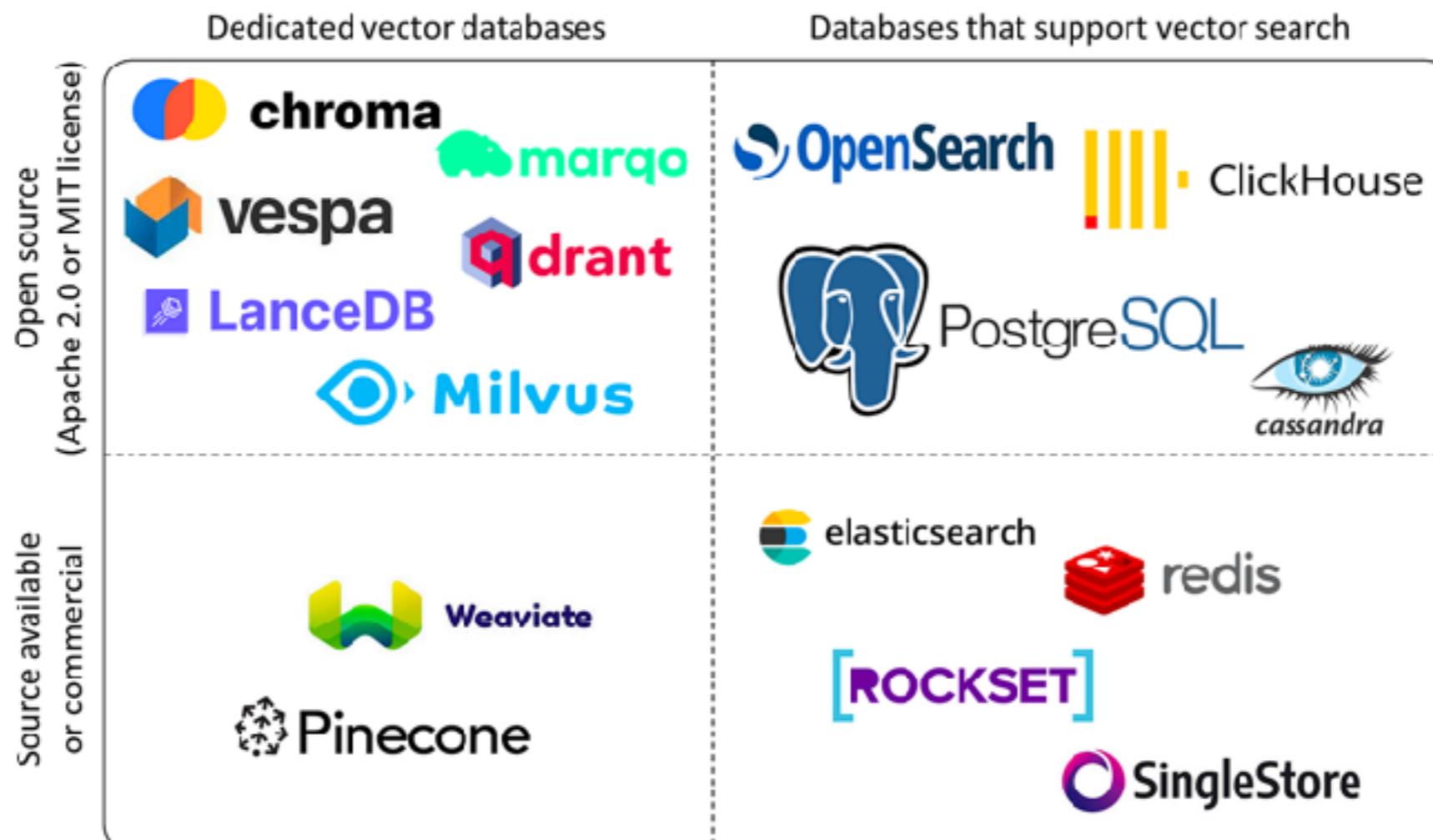


# Store data in Vector Database



# Vector Database ?

Map items of unstructured data to high-dimensional real vectors



<https://towardsdatascience.com/transformers-in-depth-part-1-introduction-to-transformer-models-in-5-minutes-ad25da6d3cca>



Application

Infrastructure

Model



Application

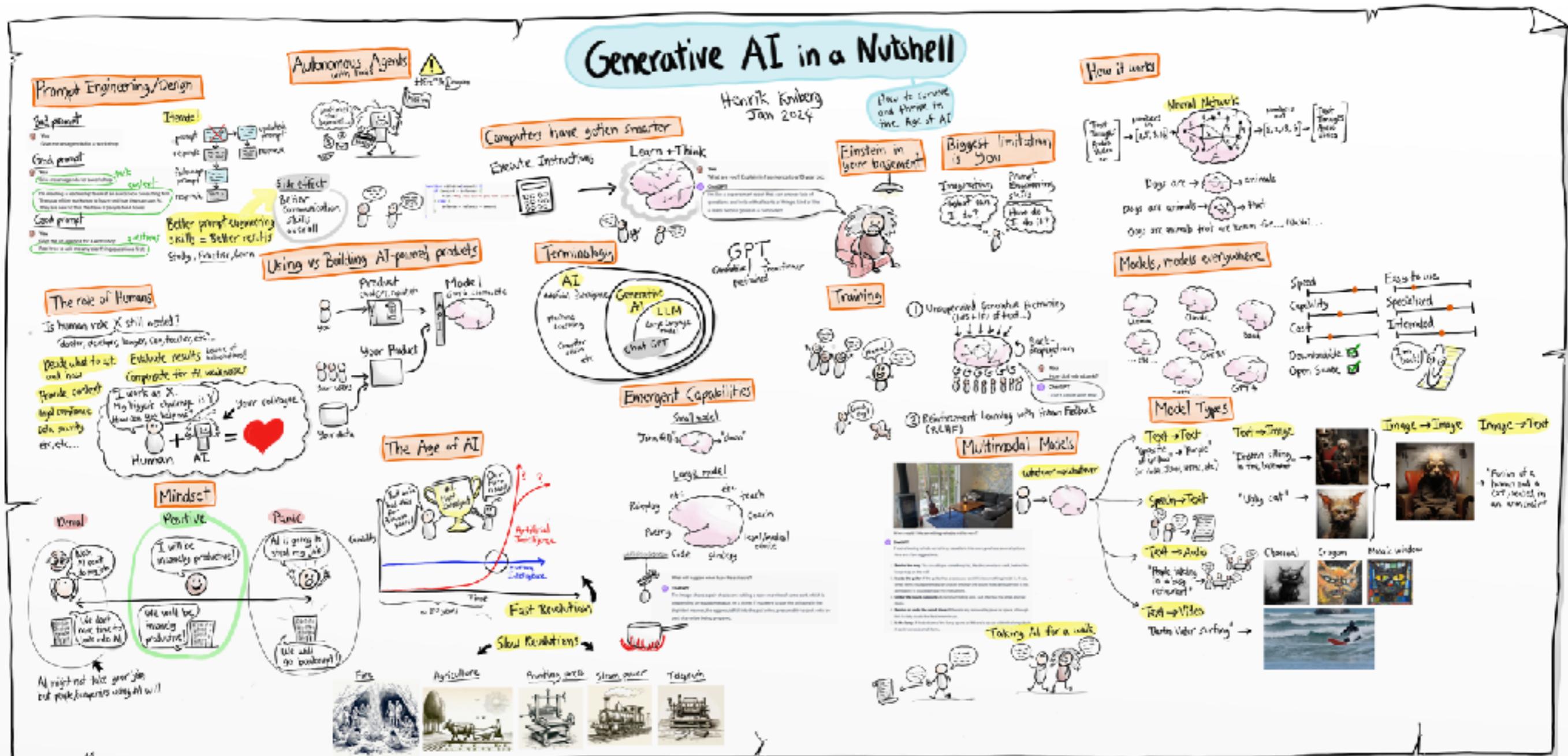
REST API

Infrastructure

Model



# Generative AI in Nutshell

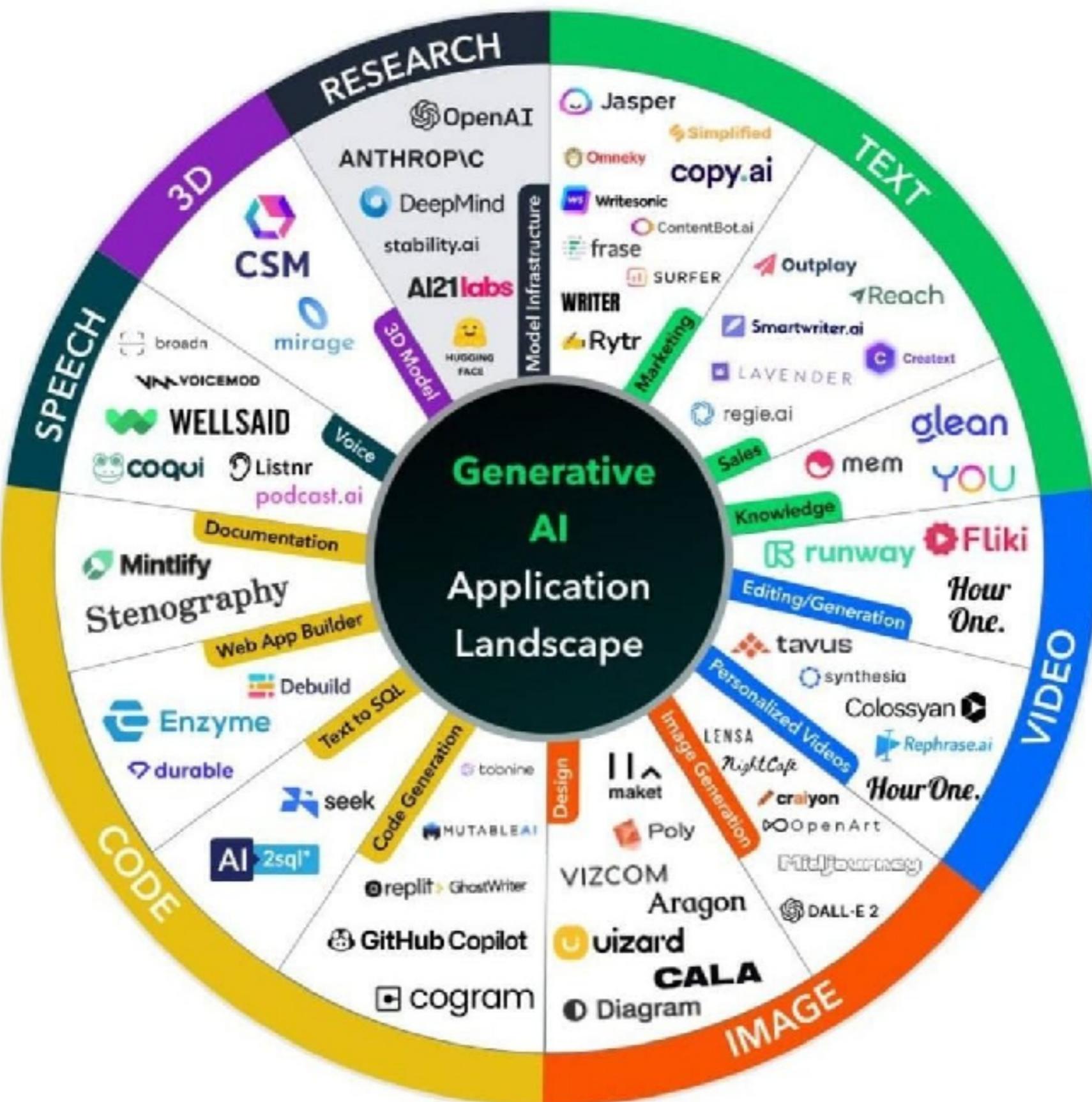


<https://www.youtube.com/watch?v=2IK3DFHRFw>



# **Application Tool chains**





# Tool chains category

Assist  
tasks

Interaction  
modes

Prompt  
composition

Properties of  
model



# Assist tasks

Finding information faster in context

Generating code

Reasoning about code

Transforming code into something ..

Requirement

Design

Develop

Testing

Deploy

Software Delivery Lifecycle



# Interaction modes

## Chat interfaces

In-line assistance (typing in code editor)  
CLI (command-line interface)



# Z.ai

The screenshot shows the Z.ai AI interface. At the top left, there are icons for a clock, a document, and a dropdown menu labeled "GLM-4.5". On the right side, there are links for "API" and "Sign in". The main area features a large text "Hi, I'm Z.ai". Below it is a search bar with the placeholder "How can I help you today?". A toolbar below the search bar includes a "+" button, a "Tools" button, a "Deep Think" button (which is highlighted in blue), and a font size button "A". At the bottom, there is a row of buttons for "AI Slides 🔥", "Full-Stack", "Magic Design", "Deep Research", and "Write code".

<https://z.ai/>



# Models

## Models

Language Models    Image Models    Video Generation Models

New

Reasoning Model

**GLM-4.6**

Zai's latest model achieves SOTA among open-source models! Context window expanded to 200K. Brings you superior performance in real-world coding, reasoning, tool using and role-playing.

[Learn More >](#)

New

Reasoning Model

**GLM-4.5-Air**

Zai's new lightweight flagship model delivers SOTA performance with exceptional cost-effectiveness!

[Learn More >](#)

New

Reasoning Model

**GLM-4.5-Flash**

The free version of GLM-4.5 now stands as Zai's most powerful offering, delivering unparalleled performance at no cost.

[Learn More >](#)

New

Visual Reasoning Model

**GLM-4.5V**

Achieve the state-of-the-art (SOTA) performance among open-source VLMs of the same level in various benchmark tests.

[Learn More >](#)

New

Language Model

**GLM-4-32B-0414-128K**

A general-purpose, cost-efficient LLM for advanced Q&A, coding, search, and structured task automation across business and technical domains.

[Learn More >](#)

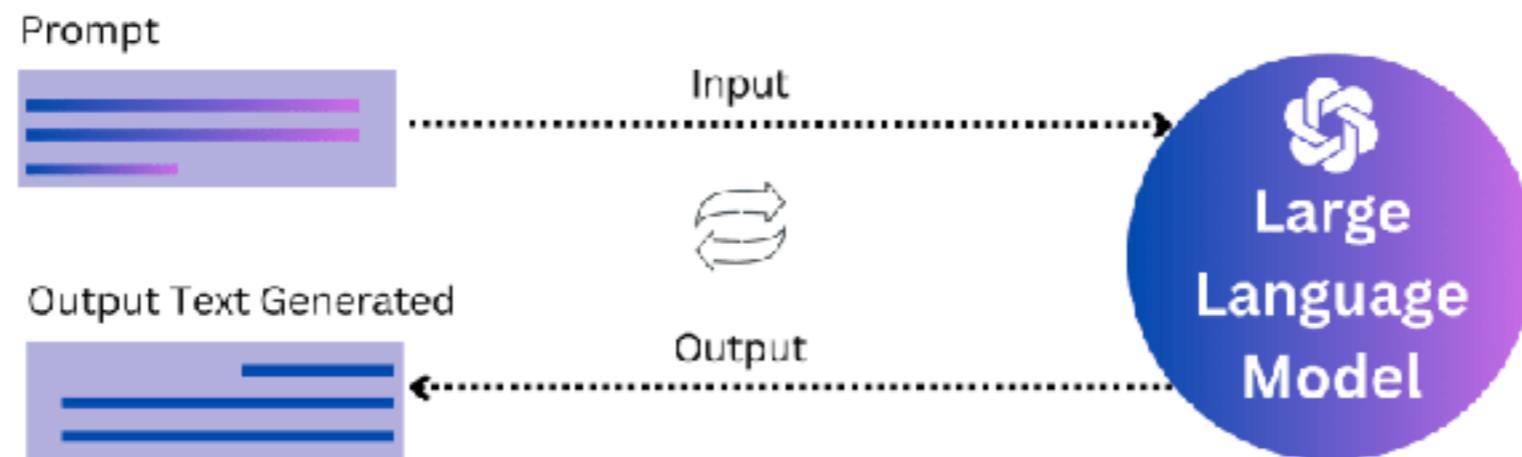
<https://z.ai/>



# Prompt composition

Prompt engineering

Compose prompts from user inputs and context



<https://platform.openai.com/docs/guides/prompt-engineering>



# Better Prompt

Write clear instructions

Provide reference text

Split complex tasks into simpler subtasks

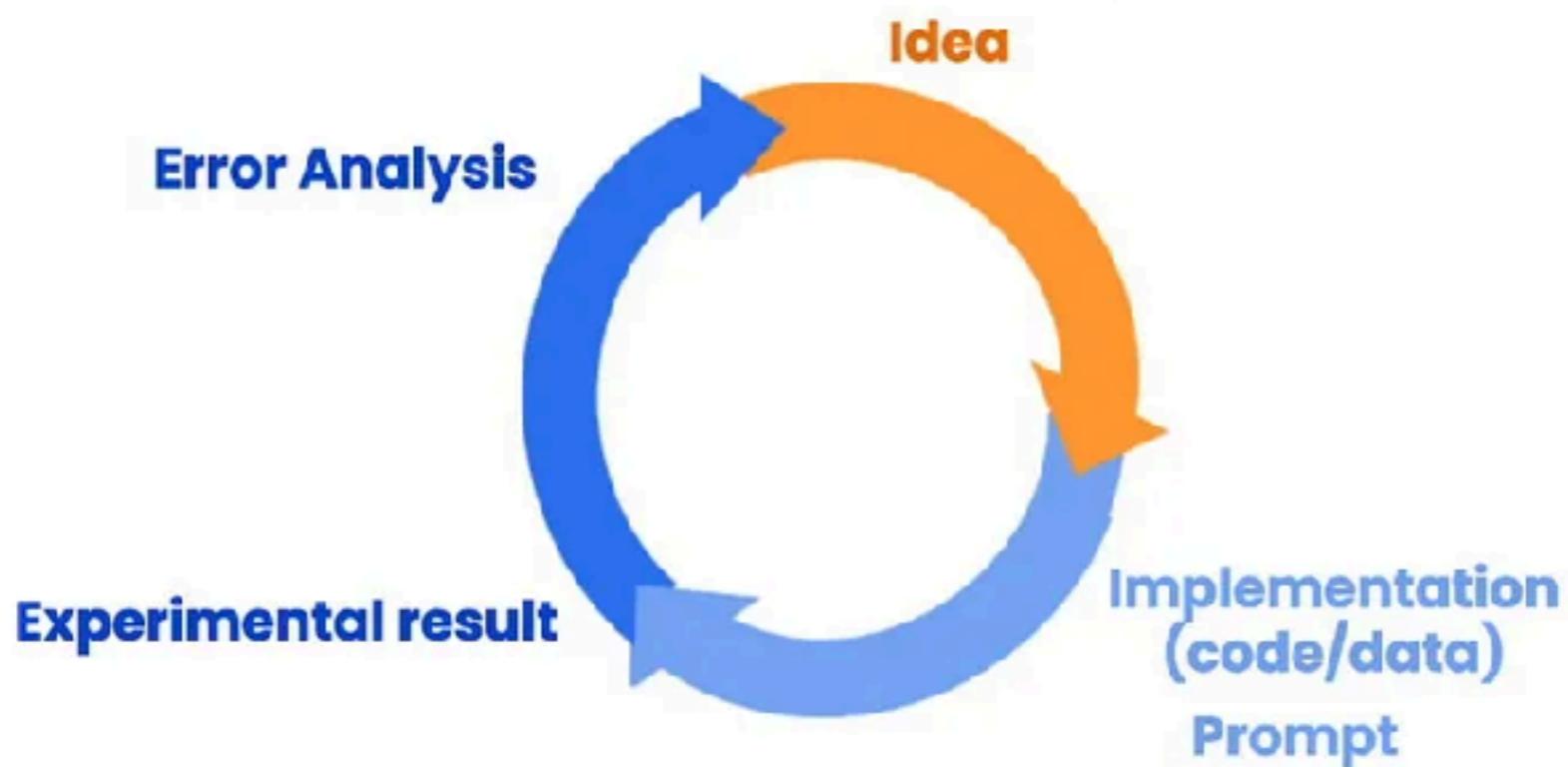
Give the model time to think

Testing and improve ...

<https://platform.openai.com/docs/guides/prompt-engineering>



# Iterative Prompt Development

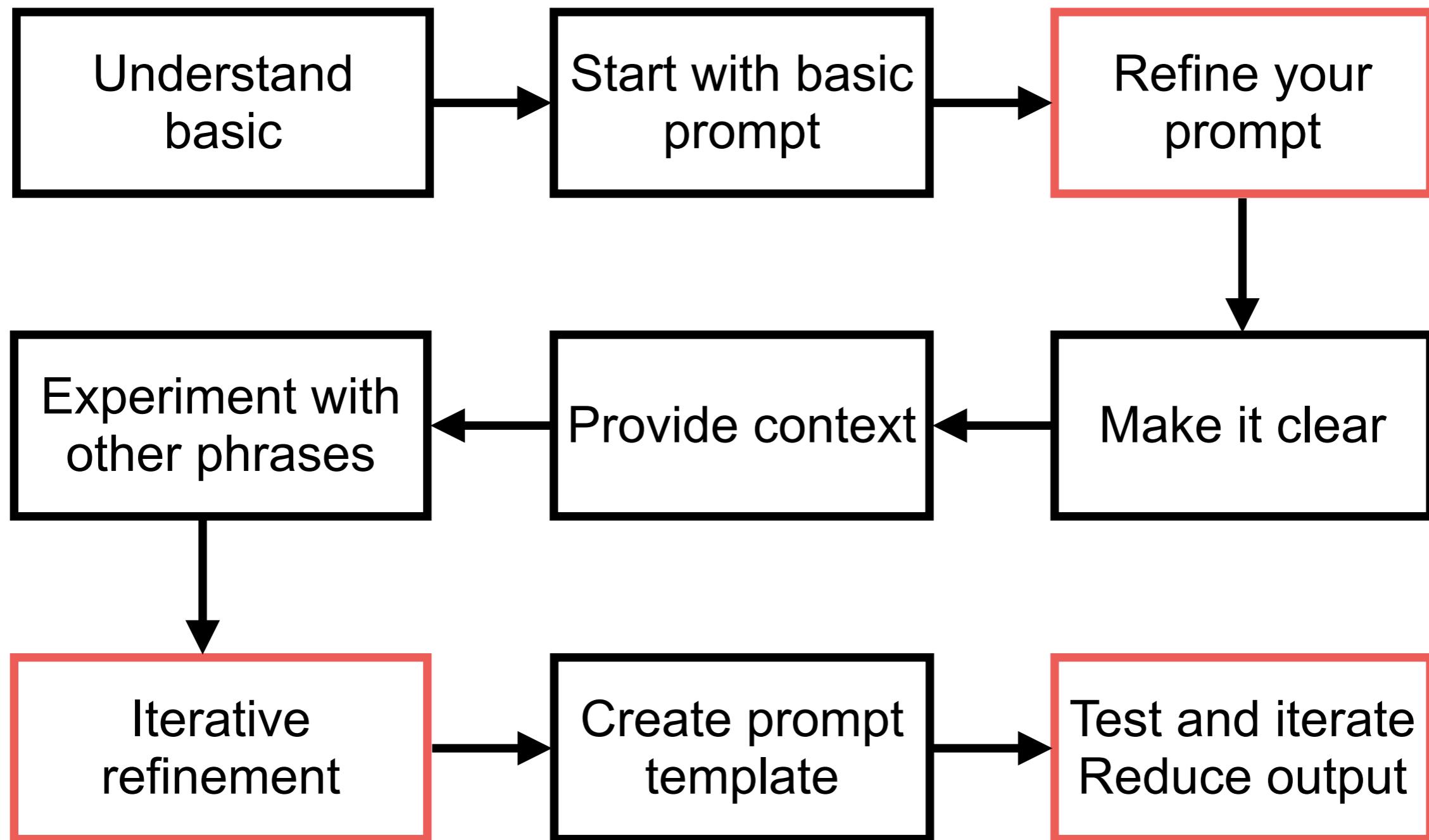


- Try something
- Analyze where the result does not give what you want
- Clarify instructions, give more time to think
- Refine prompts with a batch of examples

<https://www.deeplearning.ai/short-courses/chatgpt-prompt-engineering-for-developers/>



# Basic of Prompt Engineer



# Better Prompt

Write clear instructions

Provide reference text

Split complex tasks into simpler subtasks

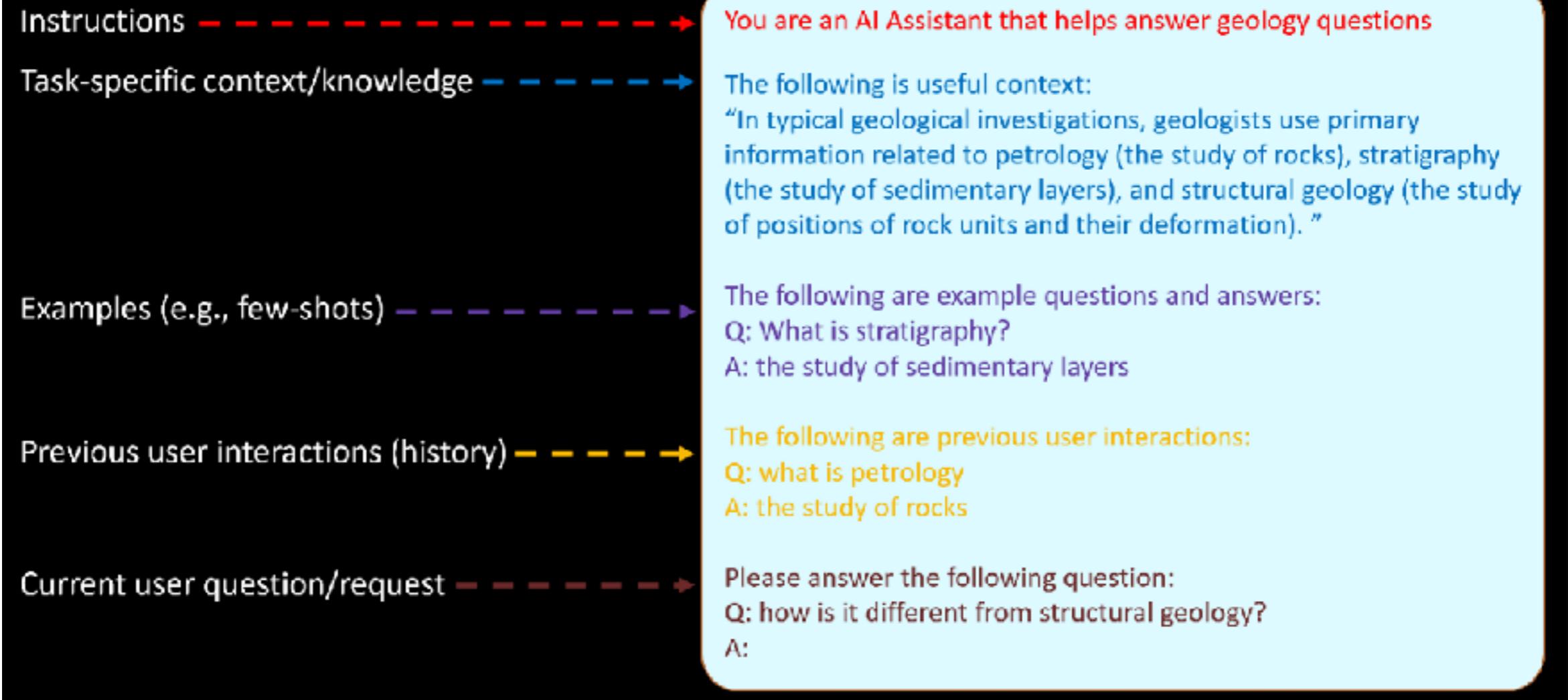
Give the model time to think

Testing and improve ...

<https://platform.openai.com/docs/guides/prompt-engineering/six-strategies-for-getting-better-results>



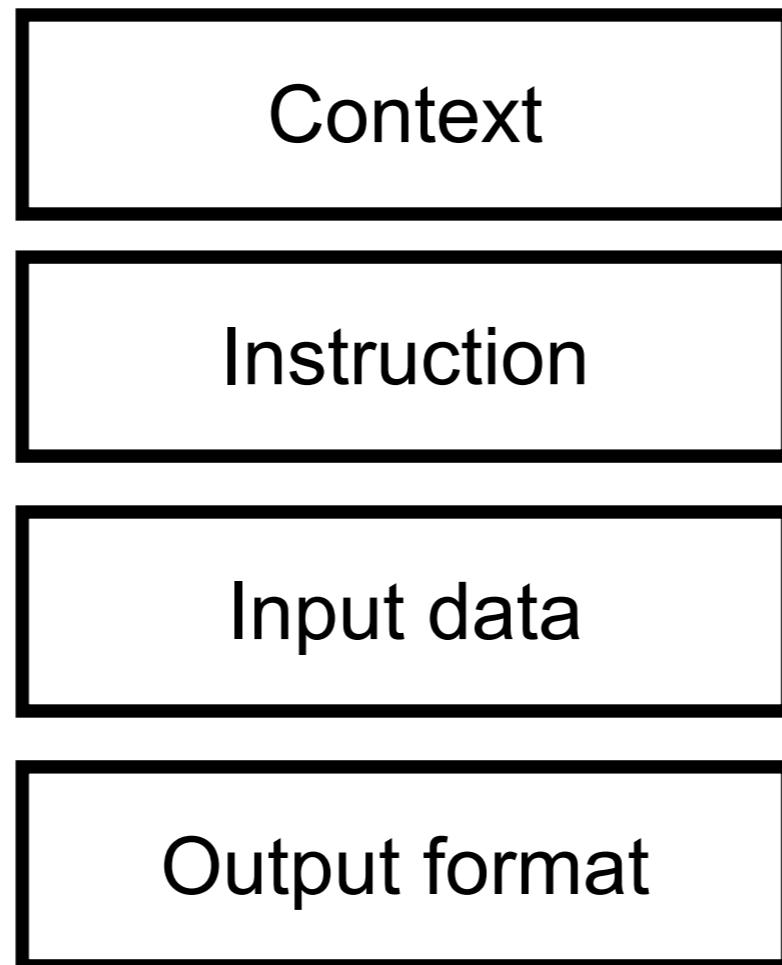
# Prompt Structure



<https://learn.microsoft.com/en-us/ai/playbook/technology-guidance/generative-ai/working-with-langs/prompt-engineering>



# Structure of Prompt



<https://platform.openai.com/docs/guides/prompt-engineering>



# Prompting Guide

Prompt Engineering

## Prompt Engineering Guide

Prompt engineering is a relatively new discipline for developing and optimizing prompts to efficiently use language models (LMs) for a wide variety of applications and research topics. Prompt engineering skills help to better understand the capabilities and limitations of large language models (LLMs).

Researchers use prompt engineering to improve the capacity of LLMs on a wide range of common and complex tasks such as question answering and arithmetic reasoning. Developers use prompt engineering to design robust and effective prompting techniques that interface with LLMs and other tools.

Prompt engineering is not just about designing and developing prompts. It encompasses a wide range of skills and techniques that are useful for interacting and developing with LLMs. It's an important skill to interface, build with, and understand capabilities of LLMs. You can use prompt engineering to improve safety of LLMs and build new capabilities like augmenting LLMs with domain knowledge and external tools.

<https://www.promptingguide.ai/>



# Structure of Prompt !!

**APE**  
Action, Purpose,  
Expectation

**RACE**  
Role, Action,  
Context,  
Expectation

**TAG**  
Task, Action, Goal

**COAST**  
Context, Objective,  
Action, Scenario,  
Task

**RISE**  
Role, Input, Step,  
Expectation

<https://twitter.com/pradeepeth/status/1673271866696544257>



# Prompt Techniques

Zero-shot

Chain-of  
Thought (CoT)

Few-shot

Meta or structure

<https://www.promptingguide.ai/techniques>



# Chain of Thought Prompting (CoT)

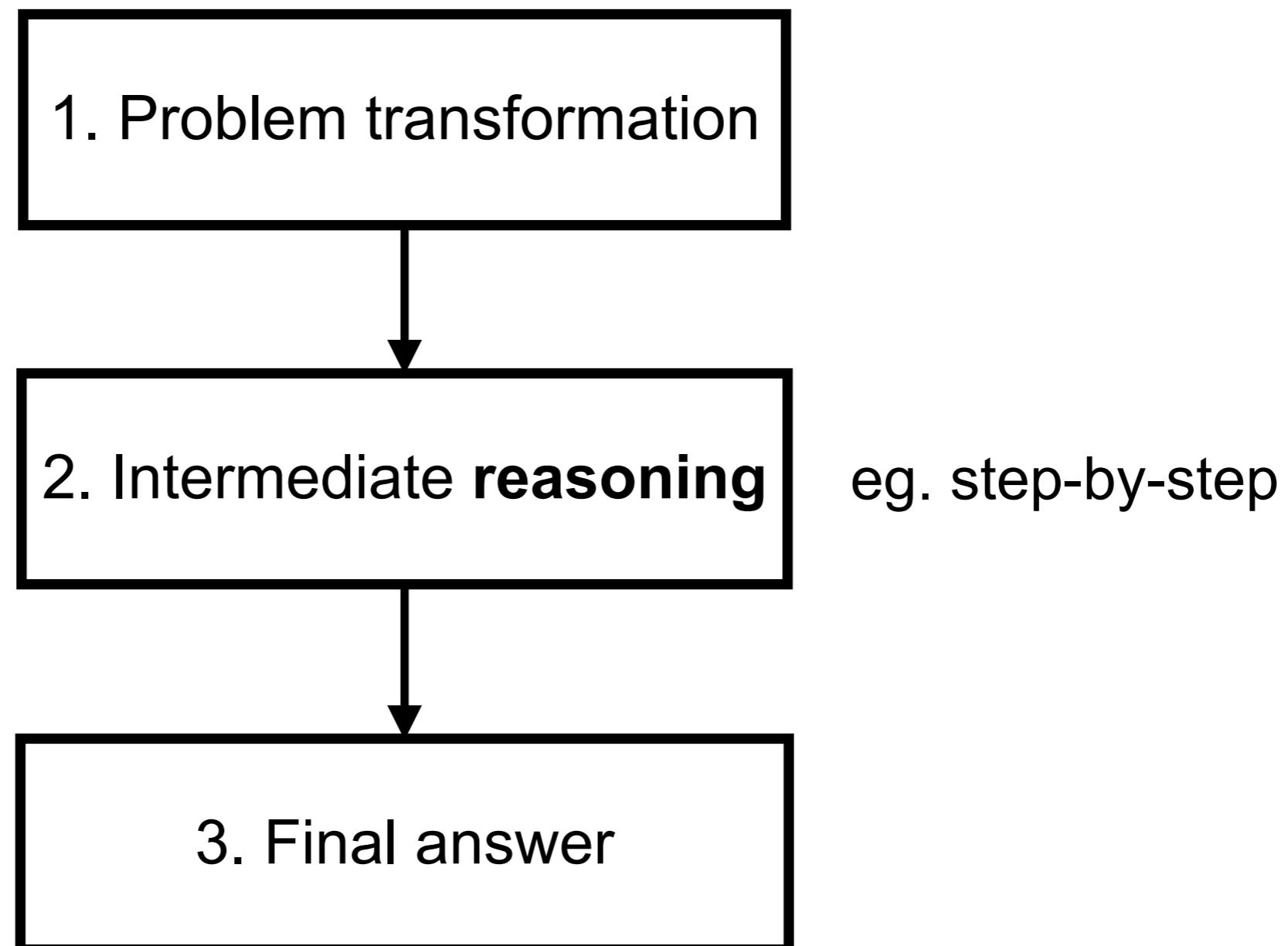
Technique used to improve the reasoning ability of LLM

Try to break down a complex problem into smaller, More manageable steps, lead to final answer

Reasoning model !!



# Chain of Thought Prompting (CoT)



# Chain of Thought Prompting (CoT)

## (a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The answer is 8. ✗

## (b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls.  $5 + 6 = 11$ . The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are  $16 / 2 = 8$  golf balls. Half of the golf balls are blue. So there are  $8 / 2 = 4$  blue golf balls. The answer is 4. ✓

## (c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: The answer (arabic numerals) is

(Output) 8 ✗

## (d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: Let's think step by step.

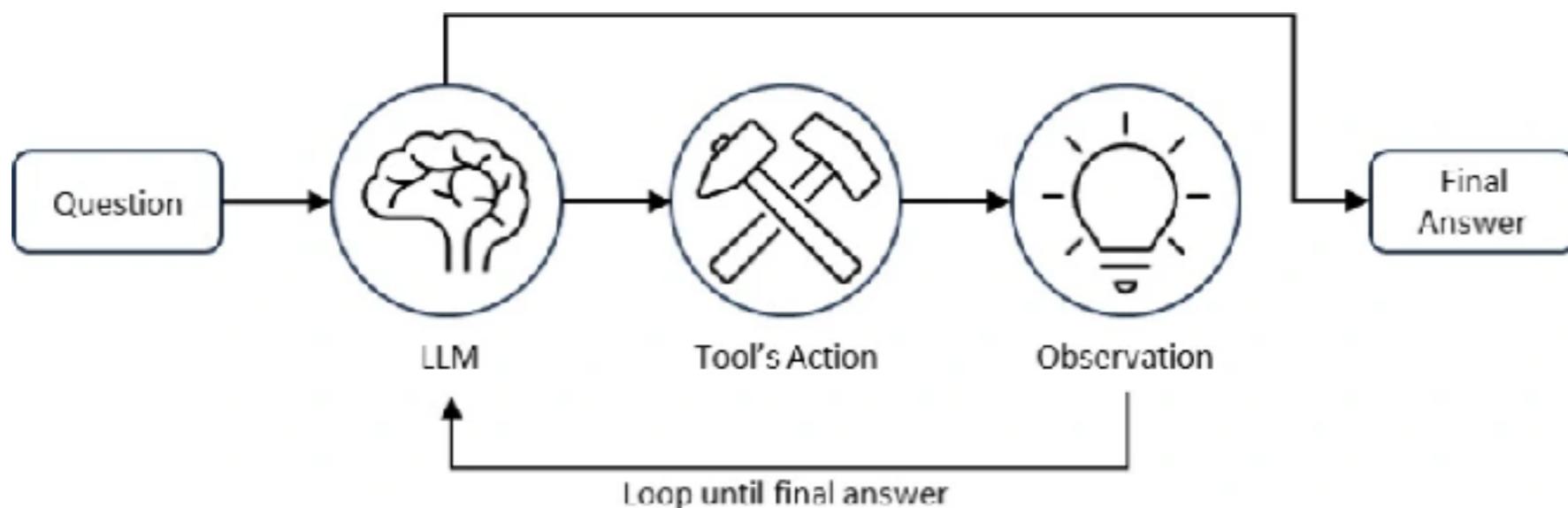
(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

<https://www.promptingguide.ai/techniques/cot>



# ReAct Prompt

LLM reasoning and additional tools (expert)  
Improve better answer



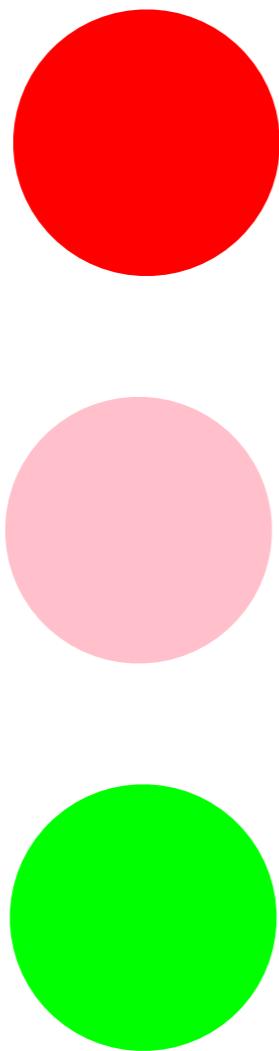
<https://www.promptingguide.ai/techniques/react>



# Workshop



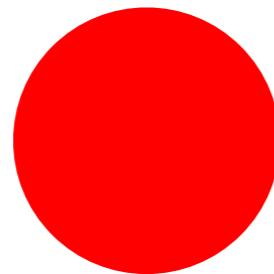
# RGB ?



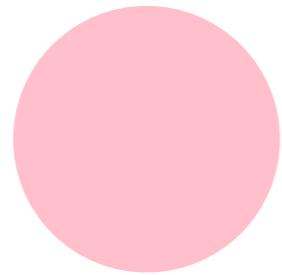
<https://github.com/up1/workshop-ai-with-technical-team/tree/main/workshop/demo-rgb>



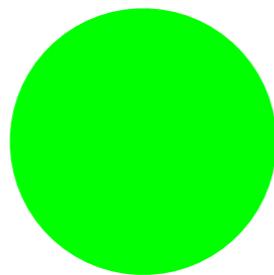
# RGB ?



[255, 0, 0]



[255, 192, 203]

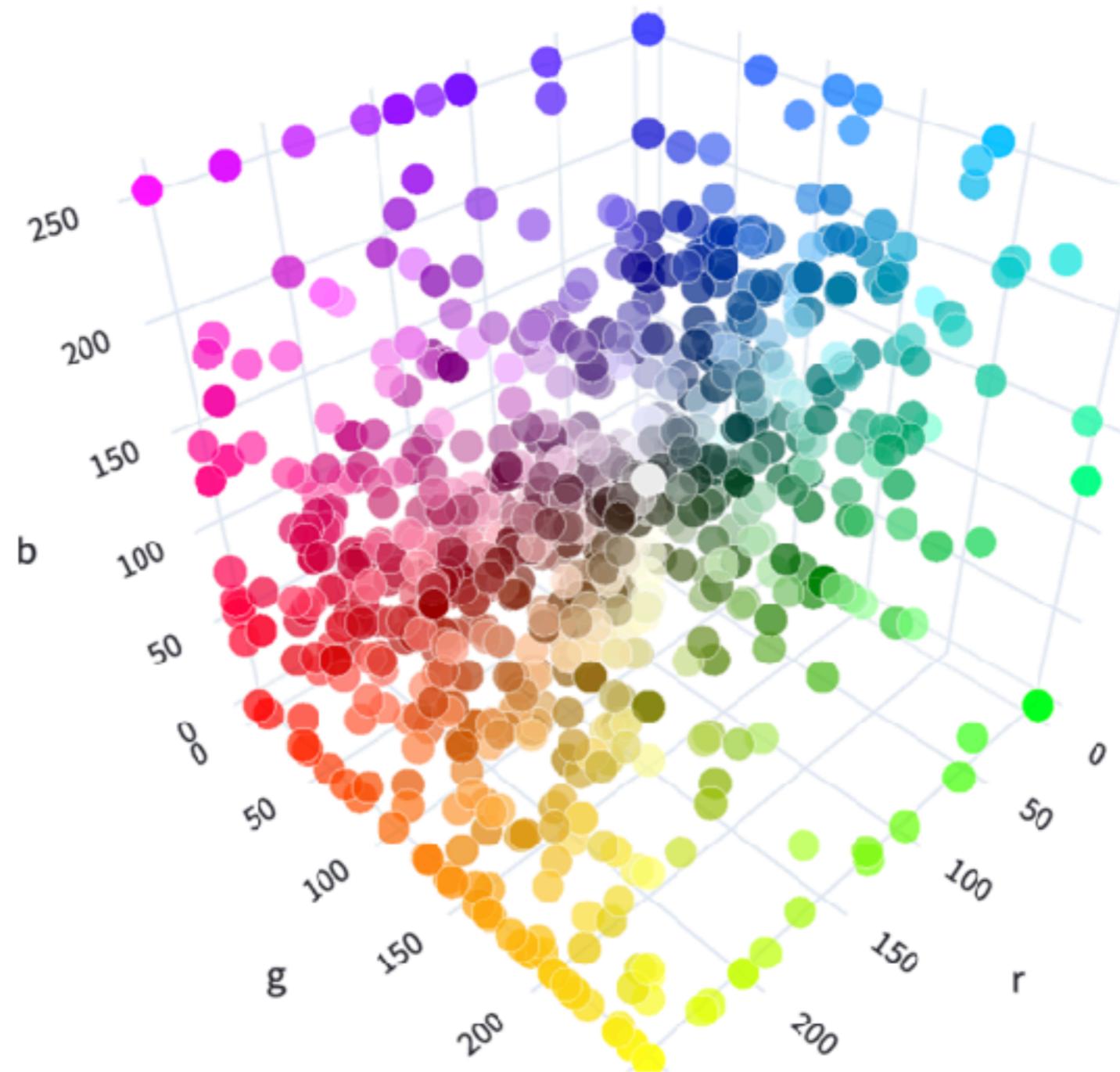


[0, 255, 0]

<https://github.com/up1/workshop-ai-with-technical-team/tree/main/workshop/demo-rgb>



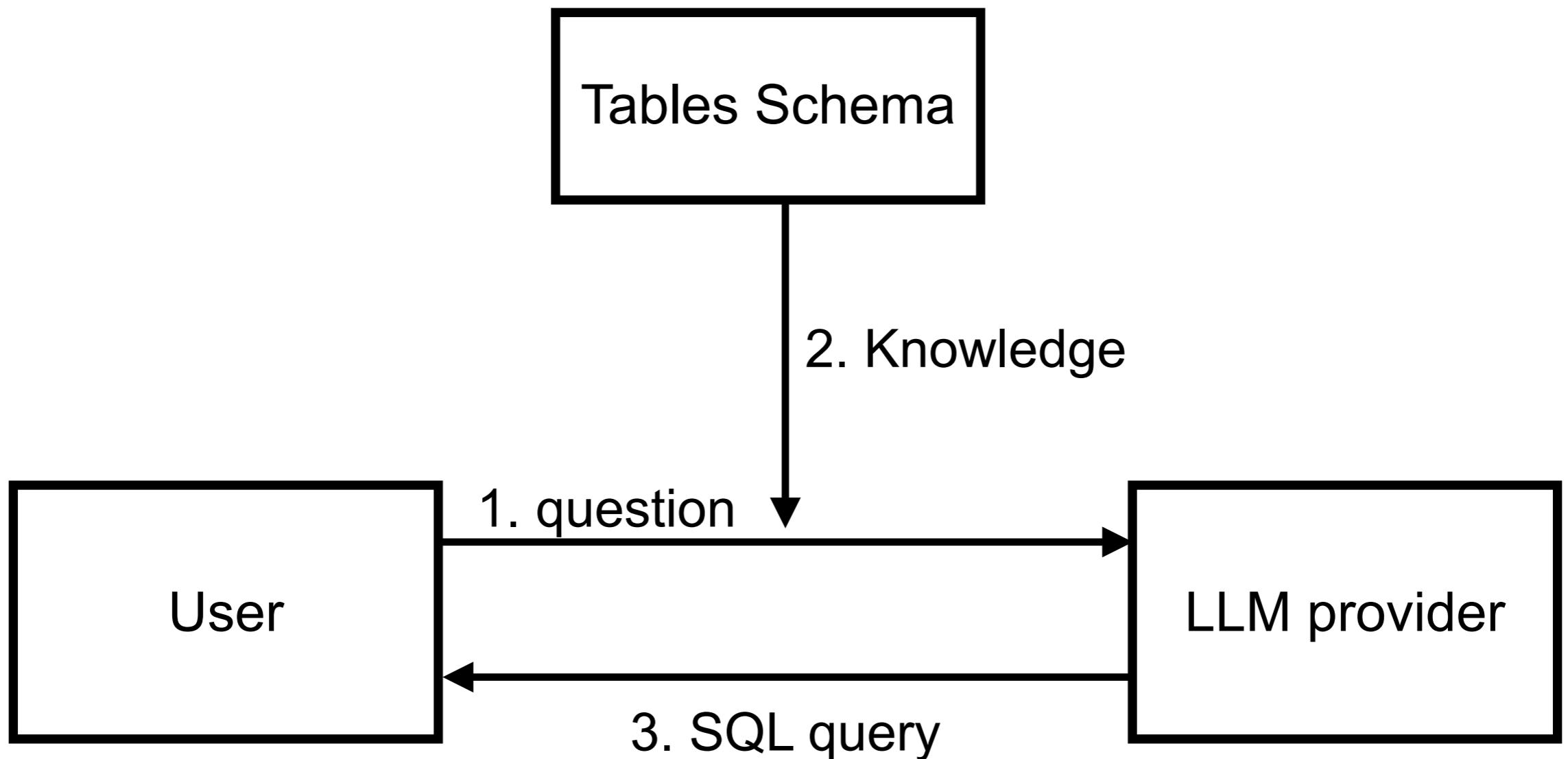
# Visualize RGB



[https://huggingface.co/spaces/jphwang/colorful\\_vectors](https://huggingface.co/spaces/jphwang/colorful_vectors)



# Text-to-SQL



<https://github.com/up1/workshop-ai-with-technical-team/tree/main/workshop/demo-sql>



# Software Development Life Cycle



# SDLC

Requirement

Design

Develop

Testing

Deploy



## Planning

Written planning process

Arch diagrams

## Testing

Automated testing

TDD

Testing environments

Testing in prod

Performance testing

Load testing

Generate test data

## Development

Automated dev env

CI/CD

Prototyping

Code review

Code generation

Templates

Cross-platform dev

Preview env

Post-commit code review

Linting

Static code analysis

Project mgmt

## Shipping

FF & experimentation

Logging

Monitoring & alerting

Staged rollouts

## Maintenance

Debug production

Documentation

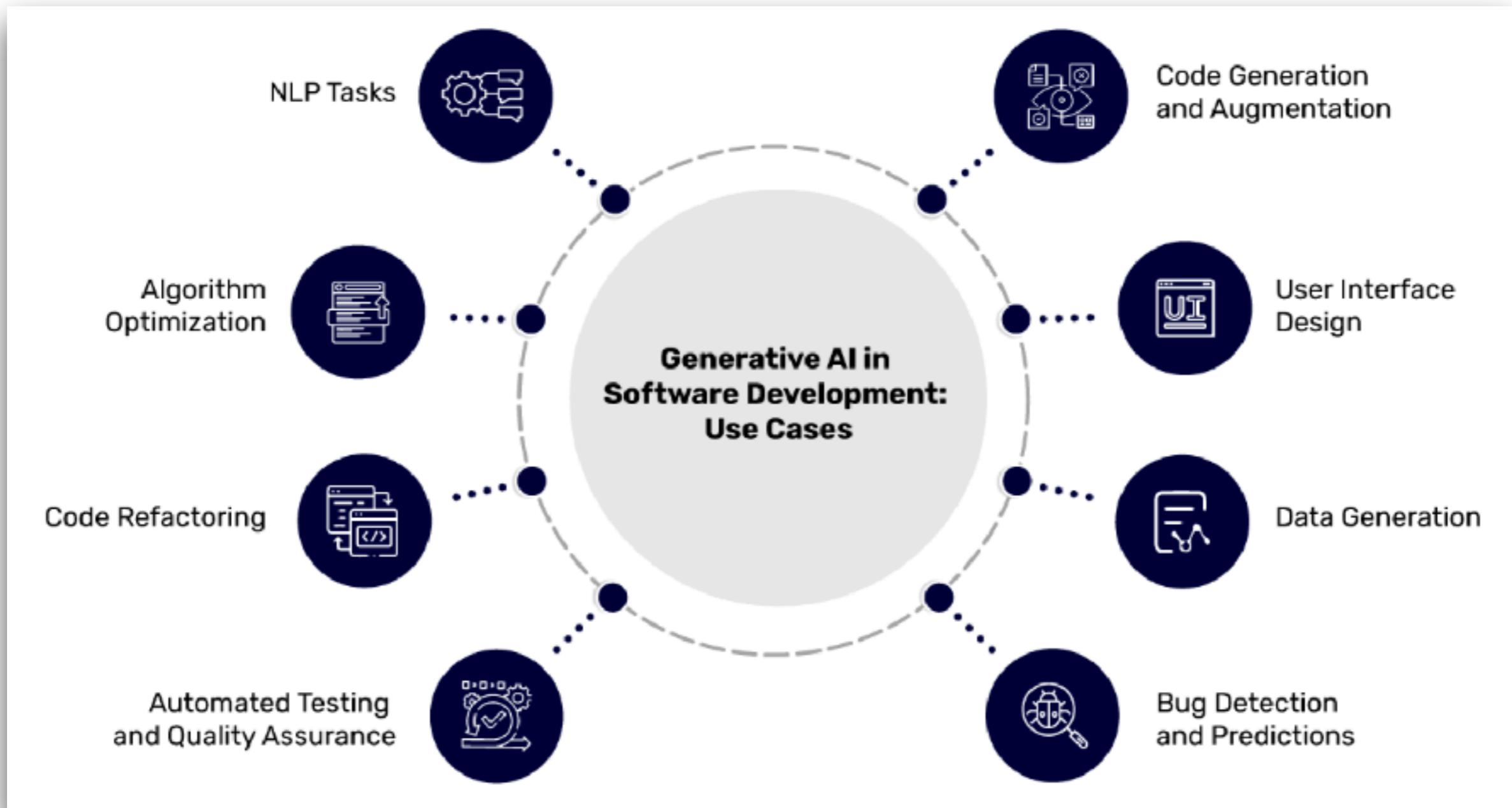
Runbook examples

Mitigation runbook

[pragmaticengineer.com](http://pragmaticengineer.com)



# SDLC



# Impacts with productivity ?

Automated  
simple tasks

Improve quality  
and reliability

Improve  
communication

Faster  
prototype



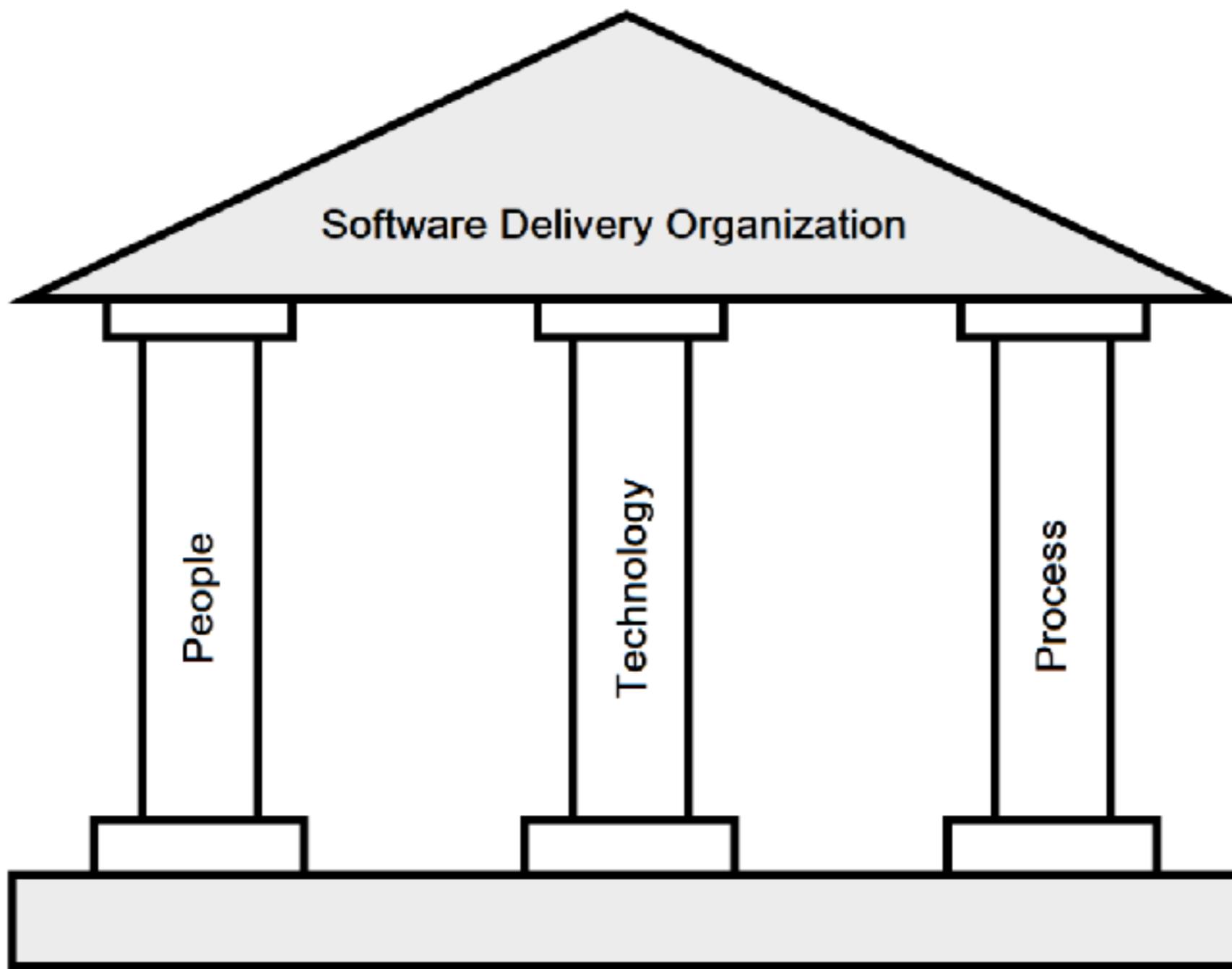
**Generative AI isn't just a tool  
it's your team member**



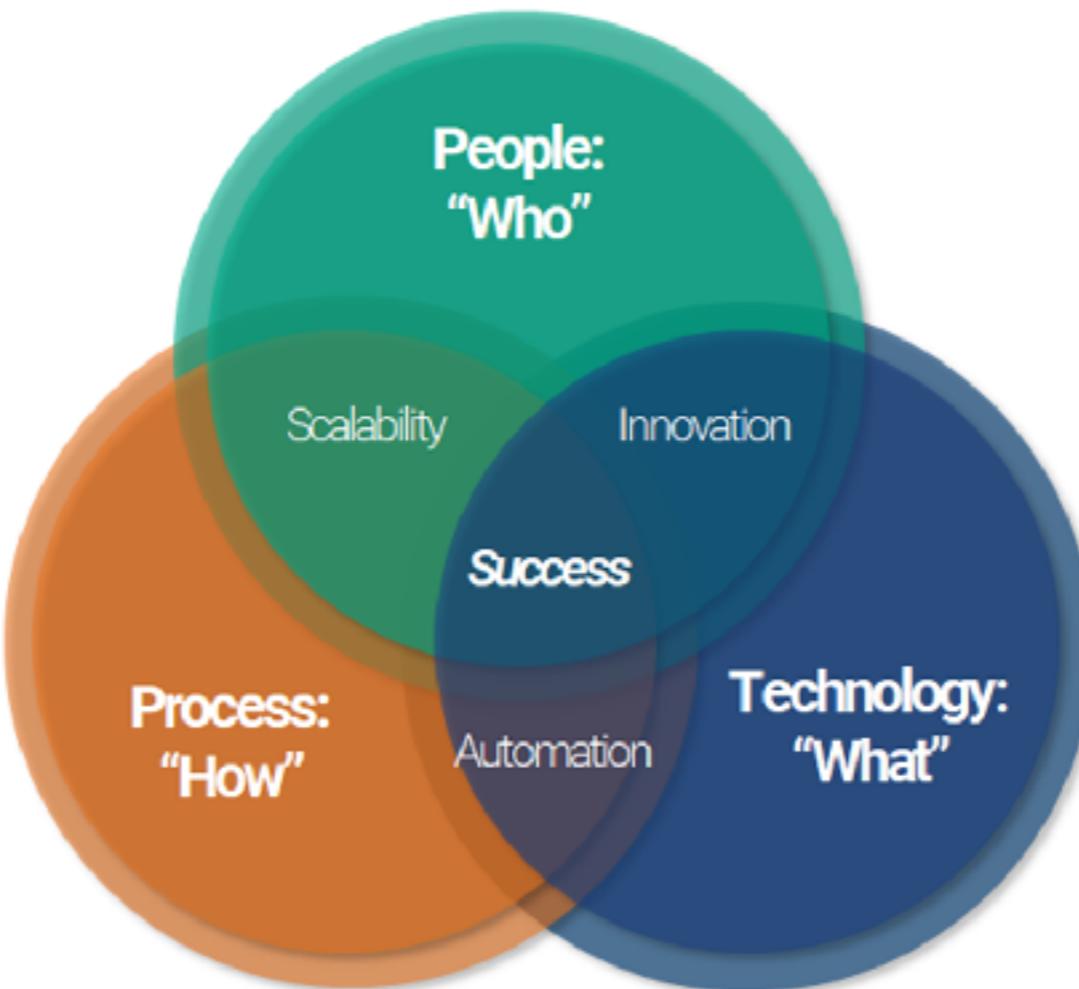
# **Trust, but verify output !!**



# 3 Pillar of Software Development



# 3 Pillar of Software Development



# Requirement and Analysis



# Requirement and Analysis

Requirement

Design

Develop

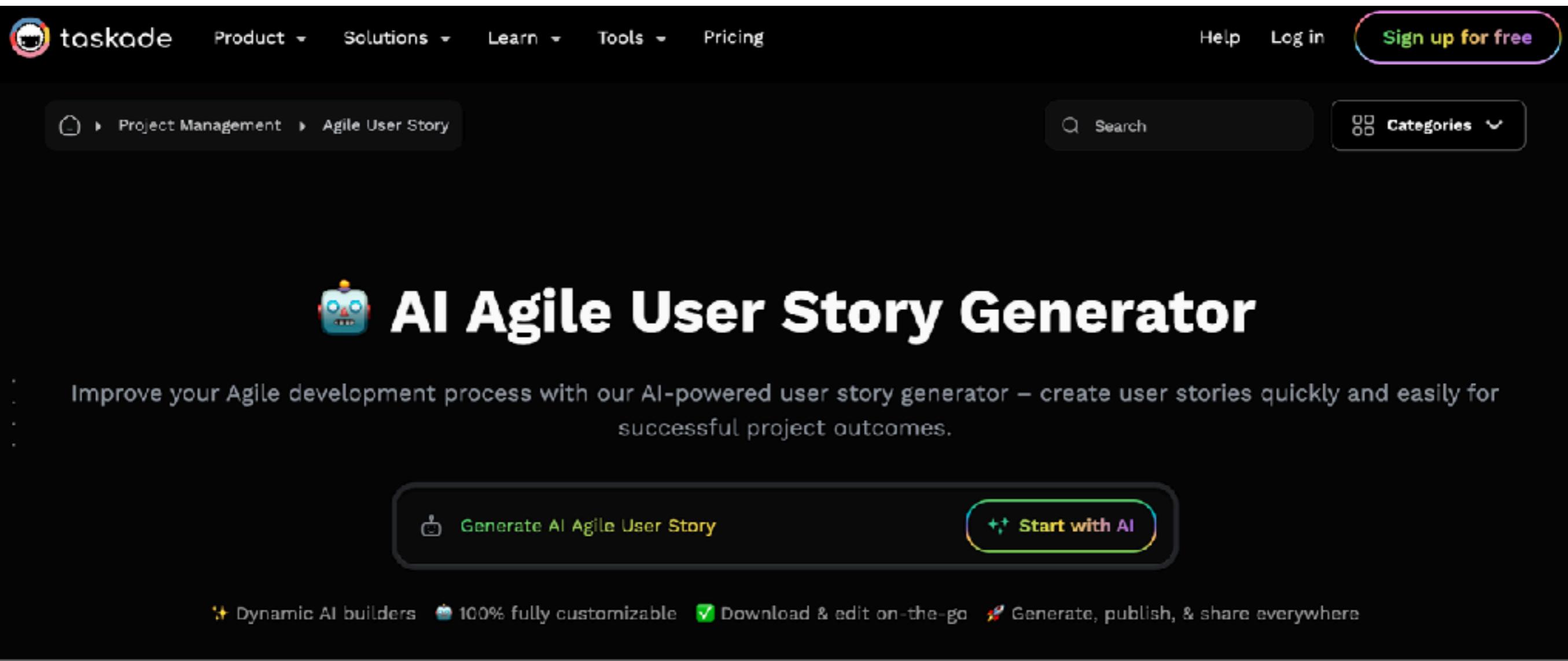
Testing

Deploy

Requirements writing and analysis  
User story generation



# Agents AI for Automated tasks



The screenshot shows the Taskade website's AI Agile User Story Generator page. At the top, there's a navigation bar with links for Product, Solutions, Learn, Tools, Pricing, Help, Log in, and a prominent 'Sign up for free' button. Below the navigation is a breadcrumb trail showing 'Project Management > Agile User Story'. To the right are search and categories filters. The main title 'AI Agile User Story Generator' is displayed with a small AI icon. A sub-headline explains the tool's purpose: 'Improve your Agile development process with our AI-powered user story generator – create user stories quickly and easily for successful project outcomes.' Two buttons are visible: 'Generate AI Agile User Story' and 'Start with AI'. Below these buttons, four features are listed with icons: Dynamic AI builders, 100% fully customizable, Download & edit on-the-go, and Generate, publish, & share everywhere.

taskade

Product Solutions Learn Tools Pricing Help Log in Sign up for free

Project Management Agile User Story

Search Categories

## AI Agile User Story Generator

Improve your Agile development process with our AI-powered user story generator – create user stories quickly and easily for successful project outcomes.

Generate AI Agile User Story Start with AI

Dynamic AI builders 100% fully customizable Download & edit on-the-go Generate, publish, & share everywhere

<https://www.taskade.com/generate/project-management/agile-user-story>



# Example with food delivery

**Food Delivery Workflow Template**

**🛒 Order Processing #order**

- Check for new orders
  - Verify customer details
  - Confirm payment status
- Prepare order items
  - Gather ingredients
  - Cook or prepare food
  - Package items securely

**🚚 Delivery Management #delivery**

- Assign delivery driver
- Plan delivery route
  - Prioritize multiple deliveries
  - Use GPS for directions
- Confirm delivery with customer
  - Send delivery notification
  - Obtain customer signature

**📊 Post-Delivery Tasks #postdelivery**

⌚ What would you like to do next? ▶

+ Create project ➡

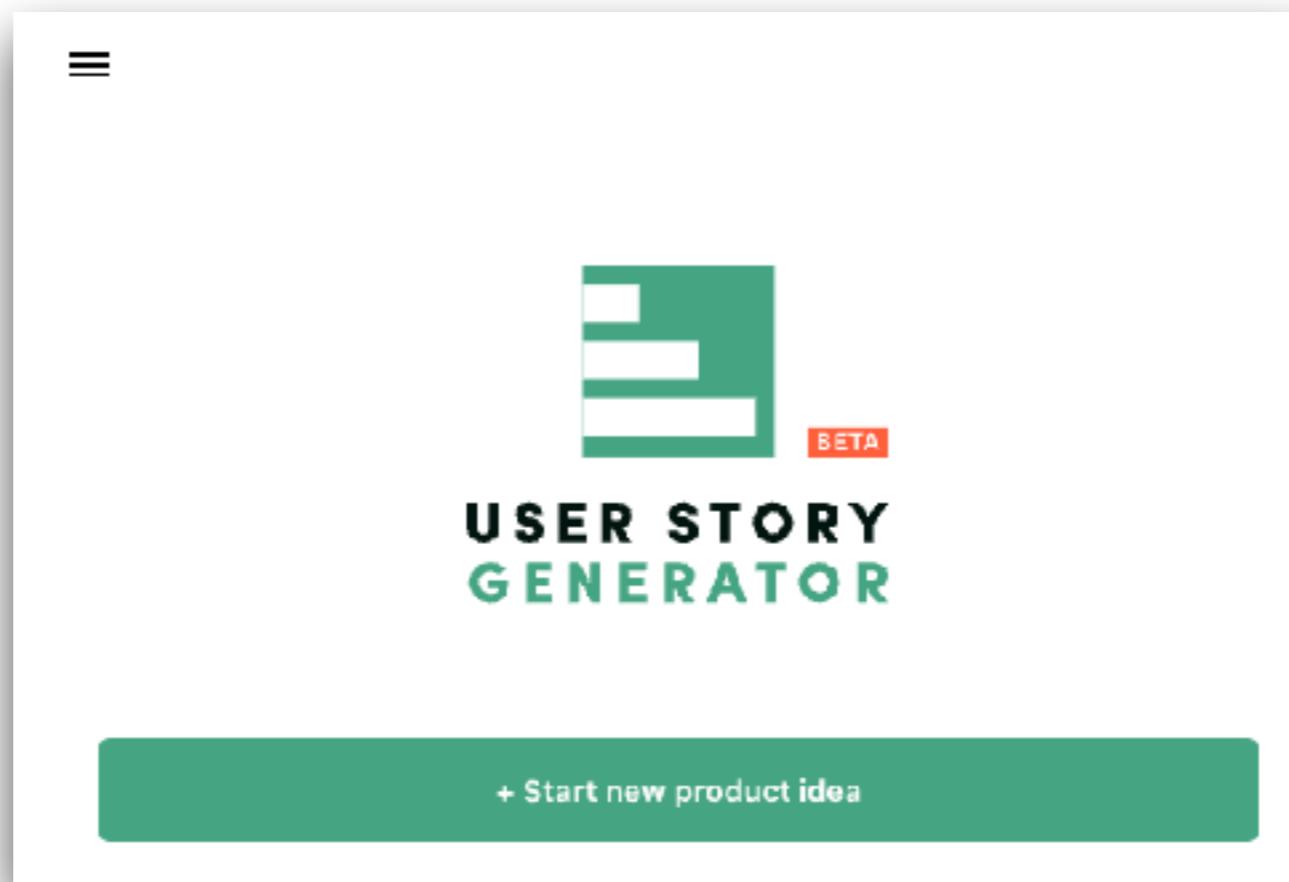
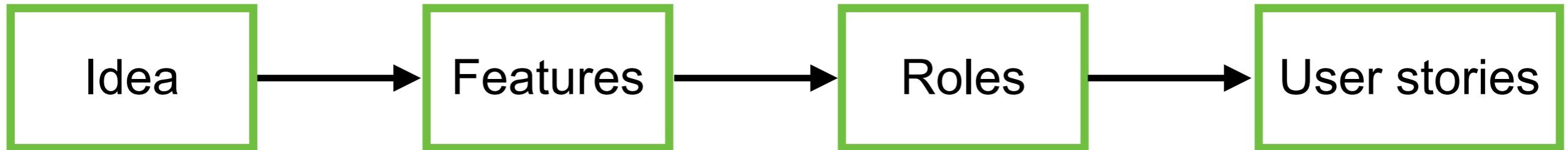
✍ Continue writing

☰ Make longer

<https://www.taskade.com/generate/project-management/agile-user-story>



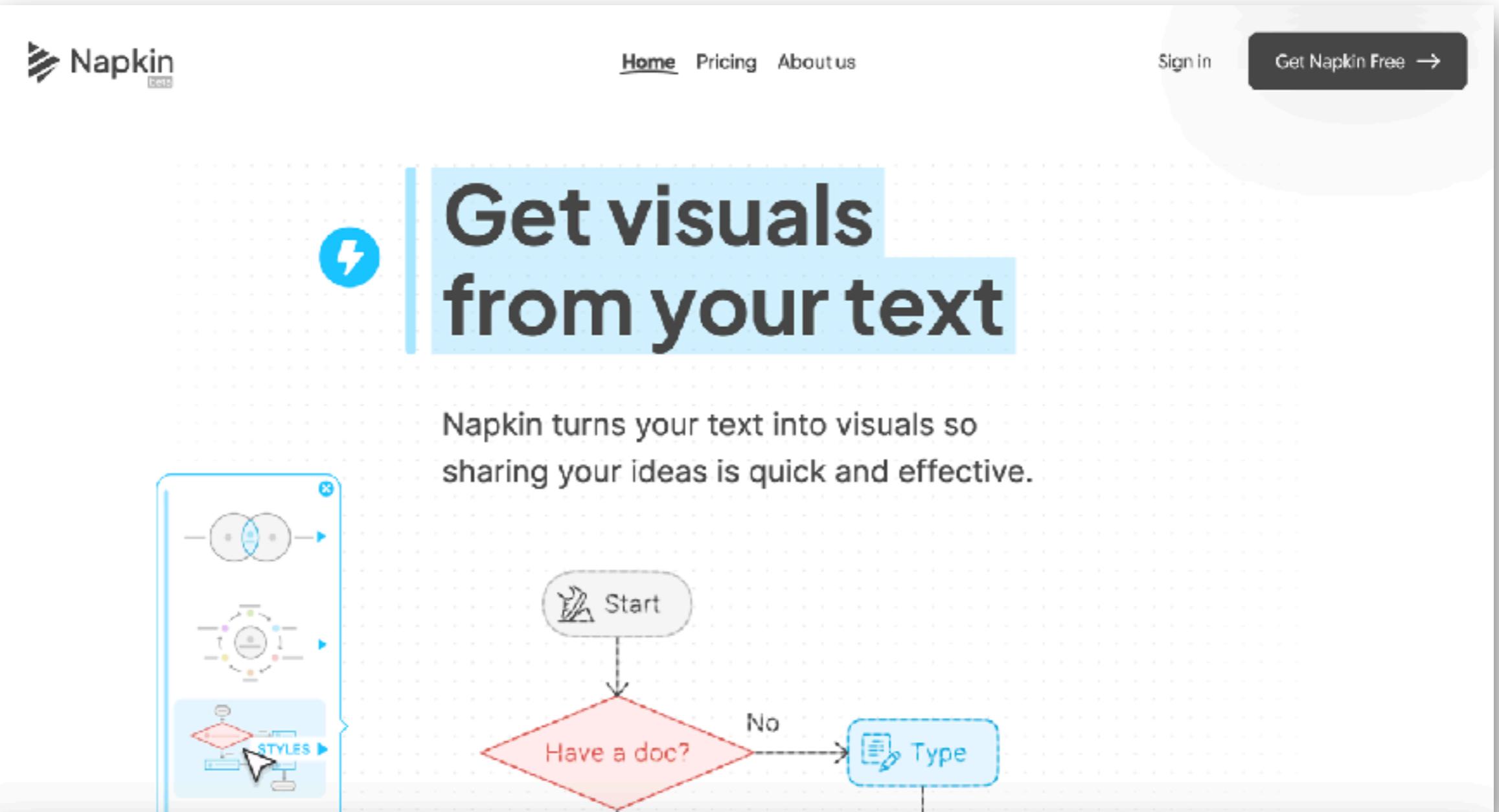
# User Story Generator



<https://userstorygenerator.ai/>



# Napkin



The image shows the Napkin AI website homepage. At the top left is the Napkin logo with the word "beta" underneath. To the right are navigation links for "Home", "Pricing", and "About us". Further right are "Sign in" and a "Get Napkin Free" button with a right-pointing arrow. The main headline "Get visuals from your text" is displayed in large, bold, dark font, accompanied by a blue lightning bolt icon. Below the headline, a subtext explains: "Napkin turns your text into visuals so sharing your ideas is quick and effective." On the left, there's a screenshot of the Napkin interface showing various visual elements like circles and arrows. On the right, a flowchart illustrates the process: "Start" leads to a decision diamond "Have a doc?". If "No", it goes to a "Type" action; if "Yes", it branches off. The URL <https://www.napkin.ai/> is visible at the bottom of the page.

<https://www.napkin.ai/>



# Requirement analysis

Clarify of User requirement ?

<https://github.com/up1/workshop-ai-with-technical-team/wiki/Requirement-analysis>



# Design Process



# Design

Requirement

Design

Develop

Testing

Deploy

Architecture writing assistance

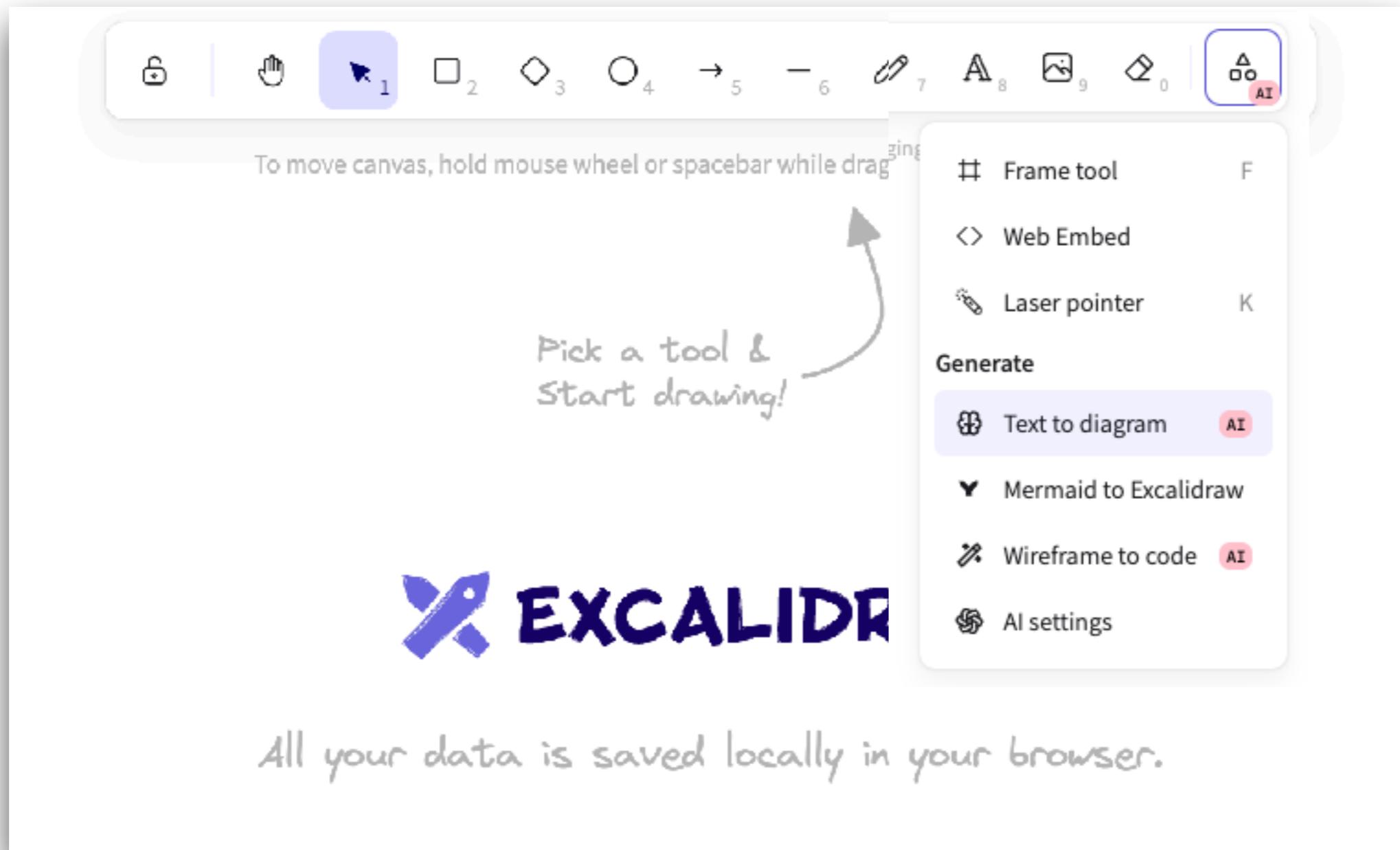
Sequence flow diagram generation

Data modeling

UX/UI design assistance



# Excalidraw with AI



<https://excalidraw.com/>



# Demo

Text to diagram AI Beta Mermaid

Currently we use Mermaid as a middle step, so you'll get best results if you describe a diagram, workflow, flow chart, and similar.

Prompt 9 requests left today

Try to generate authentication service from below

1. Customer call api gateway with REST API /login
2. Api gateway check user and password from auth service via gRPC
3. API gateway send response to client

Preview

```
sequenceDiagram
    participant Customer1 as Customer
    participant APIGateway as APIGateway
    participant AuthService as AuthService
    Customer1->>APIGateway: REST API /login with credentials
    activate APIGateway
    APIGateway->>AuthService: gRPC check user and password
    activate AuthService
    AuthService-->>APIGateway: gRPC response (success or failure)
    deactivate AuthService
    deactivate APIGateway
    APIGateway-->>Customer1: HTTP response (success or failure)
```

The diagram illustrates a three-step process for generating an authentication service. It starts with a Customer sending a REST API /login with credentials to an APIGateway. The APIGateway then performs a gRPC check user and password operation on an AuthService. Finally, the APIGateway returns an HTTP response (either success or failure) back to the Customer.

Generate → Cmd Enter

[View as Mermaid](#) →

Insert →



# Database

The screenshot shows a database management interface with the following details:

- New database**: A button in the top left.
- User Management Database ...**: A link in the top left.
- User and Shipping Address ...**: A link in the top left.
- Diagram**: A tab in the top center.
- Migrations**: A tab in the top center.
- shipping\_addresses**: A table with columns: id (Primary key), user\_id (Foreign key referencing users.id), address\_line1, address\_line2, city, state, postal\_code, country.
- users**: A table with columns: id (Primary key), name, email.
- Executed SQL**: A dropdown menu in the top right.
- Create table of user that have many shipping address**: A tooltip message.
- The tables 'users' and 'shipping\_addresses' have been created successfully. Here's a brief overview:**
  - Users Table**: Contains user information with columns for 'id', 'name', and 'email'.
  - Shipping Addresses Table**: Stores multiple addresses for each user, with columns for 'id', 'user\_id' (foreign key referencing 'users'), 'address\_line1', 'address\_line2', 'city', 'state', 'postal\_code', and 'country'.
- Conversation renamed to User and Shipping Address Management chance**: A message at the bottom left.
- PG 18 | Local-only database**: A status bar at the bottom left.
- Primary key**, **Identity**, **Unique**, **Nullable**: Options at the bottom left.
- Message AI or write SQL**: A text input field at the bottom right with an upward arrow icon.

<https://database.build/>



# Magic Pattern

The screenshot shows the Magic Patterns web application interface. At the top, there is a navigation bar with links for 'Magic Patterns', 'Use Cases', 'Customers', and a user profile icon. Below the navigation bar, a large heading reads 'Prototype your product ideas with AI.' followed by a subtext: 'Iterate on components & designs in our AI-native editor. Export to React or Figma.' Three buttons are visible: 'Generate a new UI' (disabled), 'Add a new feature to an existing UI' (highlighted in blue), and 'Apply a theme to an existing UI'. The main area features a flowchart diagram with three boxes connected by arrows. The first box on the left contains the text 'Import your existing UI' and a placeholder 'Add an image or screenshot'. The middle box contains the text 'Describe what to add to the existing UI' with examples like 'e.g. add an error state' and '(Optional) Include an image'. The third box on the right contains a 'Generate' button. The entire interface has a clean, modern design with a light gray background and white text.

<https://www.magicpatterns.com/>



# Make Real

The image displays two wireframe prototypes of an "Intelligence Report System".

**Left Prototype:**

- Header: "Intelligence Report System"
- Form:
  - Placeholder: "ค้นหาข้อมูล..."
  - Send button: "Send"
- Section: "Search results"
  - Items: "Row 1", "Row 2", "Row 3"

**Right Prototype:**

- Header: "Intelligence Report System"
- Form:
  - Placeholder: "ค้นหาข้อมูล..."
  - Search input field: "send"
  - Search button: "Send"
- Section: "Search Results"
  - Header: "Search Results" (with a magnifying glass icon)
  - Text: "3 results found"
  - Item 1: "Intelligence Report #001"
    - Description: "Comprehensive analysis of market trends and competitive intelligence data for Q4 2024."
    - Details: "Oct 15, 2024", "Analyst Team A", "High Priority"
  - Item 2: "Security Assessment Report"
    - Description: "Detailed security vulnerability assessment and threat analysis for enterprise systems."
    - Details: "Oct 17, 2024", "Security Team", "Critical"
  - Item 3: "Operational Intelligence Summary"
    - Description: "Monthly operational metrics and performance indicators with strategic recommendations."
    - Details: "Oct 19, 2024", "Operations Team", "Same", "For review in meeting"

<https://github.com/tldraw/make-real>



# v0.dev

The screenshot shows the v0.dev platform's interface. At the top, there is a navigation bar with a logo on the left and a "Private Beta" button on the right. Below the navigation bar, a dark modal window titled "A 'report an issue' modal" is open, featuring three small icons: a square with a diagonal line, a checkmark, and a left arrow. Below the modal are four buttons: "Product categories", "Hero section", "Contact form", and "Ecommerce dashboard". The main area of the screen displays a grid of website prototypes. In the top row, there are four prototypes: "Soccer Game", "Enhance Your Education Journey", "A website in a black and white theme", and "Welcome to the Festival Page". In the bottom row, there are four prototypes: "page for a soccer game, with a hero section", "A hero section for a... website", "A website in a black and white theme", and "product tour like appcues". Each prototype has a small circular profile picture of a person at the bottom left and a descriptive text bubble at the bottom right.

<https://v0.dev/>



# Bolt.new

The screenshot shows the Bolt.new homepage. At the top, a large question "What do you want to build?" is displayed in a bold, black font. Below it, a subtitle reads "Prompt, run, edit, and deploy full-stack web apps." A search bar contains the placeholder text "How can Bolt help you today?". Below the search bar are two rows of project suggestions in rounded rectangular boxes. The first row includes "Start a blog with Astro", "Build a mobile app with NativeScript", and "Create a docs site with Vitepress". The second row includes "Scaffold UI with shadcn", "Draft a presentation with Sliddev", and "Code a video with Remotion". Further down, a section titled "or start a blank app with your favorite stack" displays icons for various frameworks and tools: Next.js, React, Node.js, TypeScript, GraphQL, Prisma, Tailwind CSS, and Vite. At the bottom of the page is a red footer bar containing the URL "https://bolt.new/".

What do you want to build?

Prompt, run, edit, and deploy full-stack web apps.

How can Bolt help you today?

Start a blog with Astro Build a mobile app with NativeScript Create a docs site with Vitepress

Scaffold UI with shadcn Draft a presentation with Sliddev Code a video with Remotion

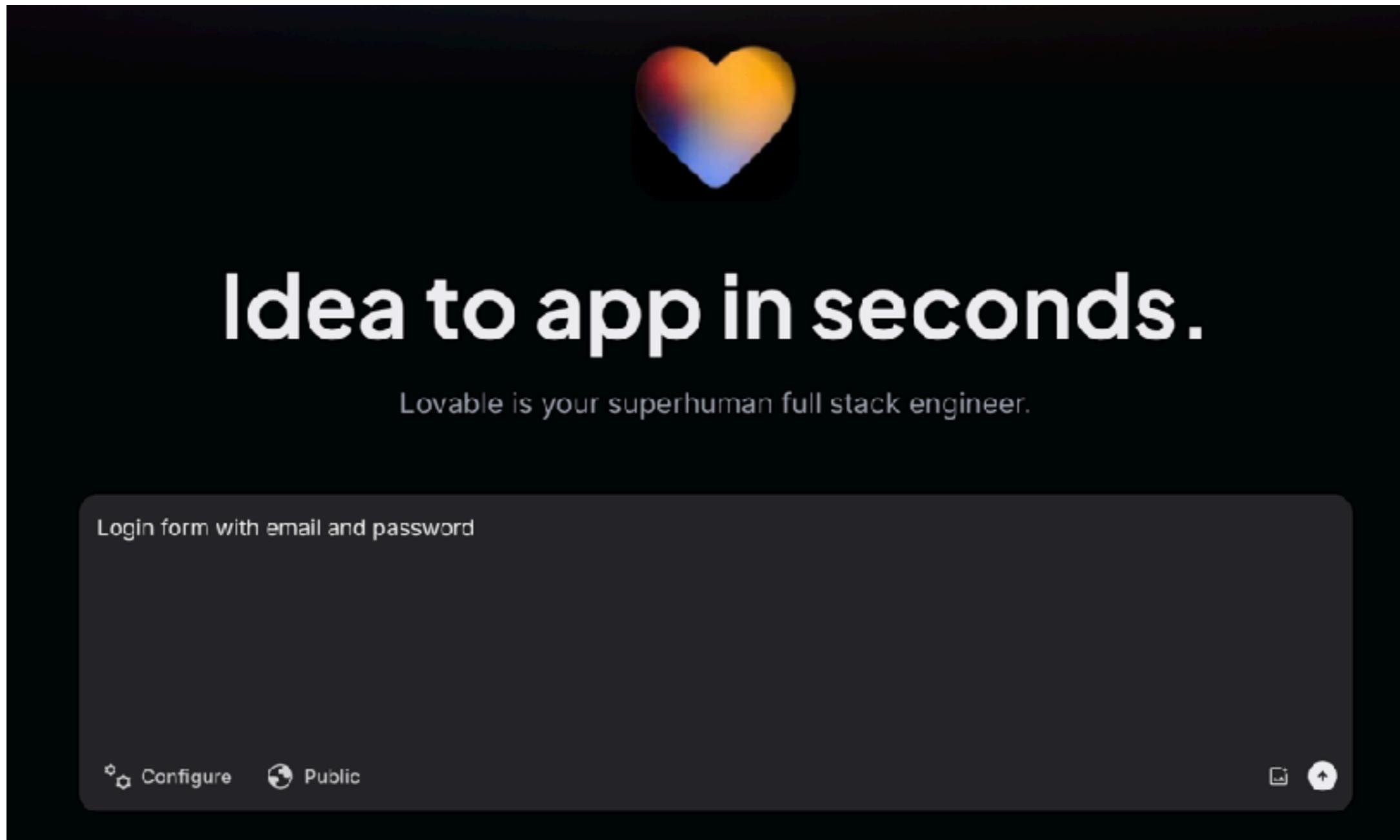
or start a blank app with your favorite stack

Next.js React Node.js TypeScript GraphQL Prisma Tailwind CSS Vite

<https://bolt.new/>



# Lovable



<https://lovable.dev/>



# SuperDesign

The screenshot shows the SuperDesign interface. At the top, a large banner features the text "Vibe it. Design it." and "Explore freely, iterate fast. Your design, AI-powered.". Below the banner are two buttons: "Design on web" and "Integrate with" followed by three small icons. The main workspace is a large text input field with a placeholder "Describe what you want to create...". Below this field are several icons: a camera, a pencil, a folder, and a Google Gemini logo with the text "Gemini 3 Flash". To the right of these icons is a button labeled "Use AI Designer" with a dropdown arrow, and a black circular button with a white upward-pointing arrow. At the bottom of the workspace are three more buttons: "Recreate Screenshot", "Import from Site", and "Explore Effects".

<https://app.superdesign.dev/>



# Development Process



# Develop

Requirement

Design

Develop

Testing

Deploy

Code generation

Review and explain code

Debugging code

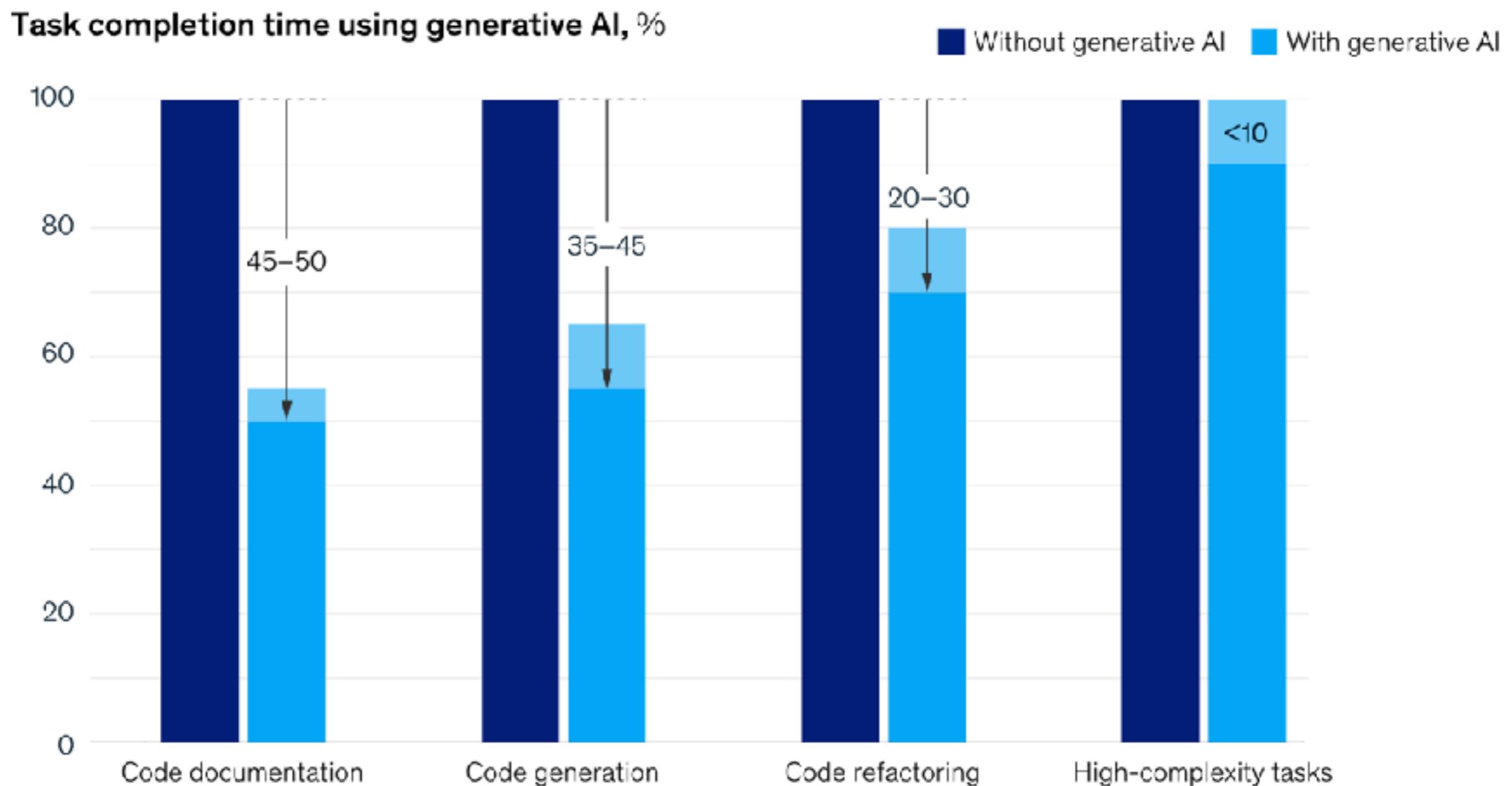
Improve consistency

Code translation



# Development

**Generative AI can increase developer speed, but less so for complex tasks.**



<https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/unleashing-developer-productivity-with-generative-ai>



# Category of Tools

Chat AI

ChatGPT  
Gemini  
Claude.ai  
DeepSeek

Code AI

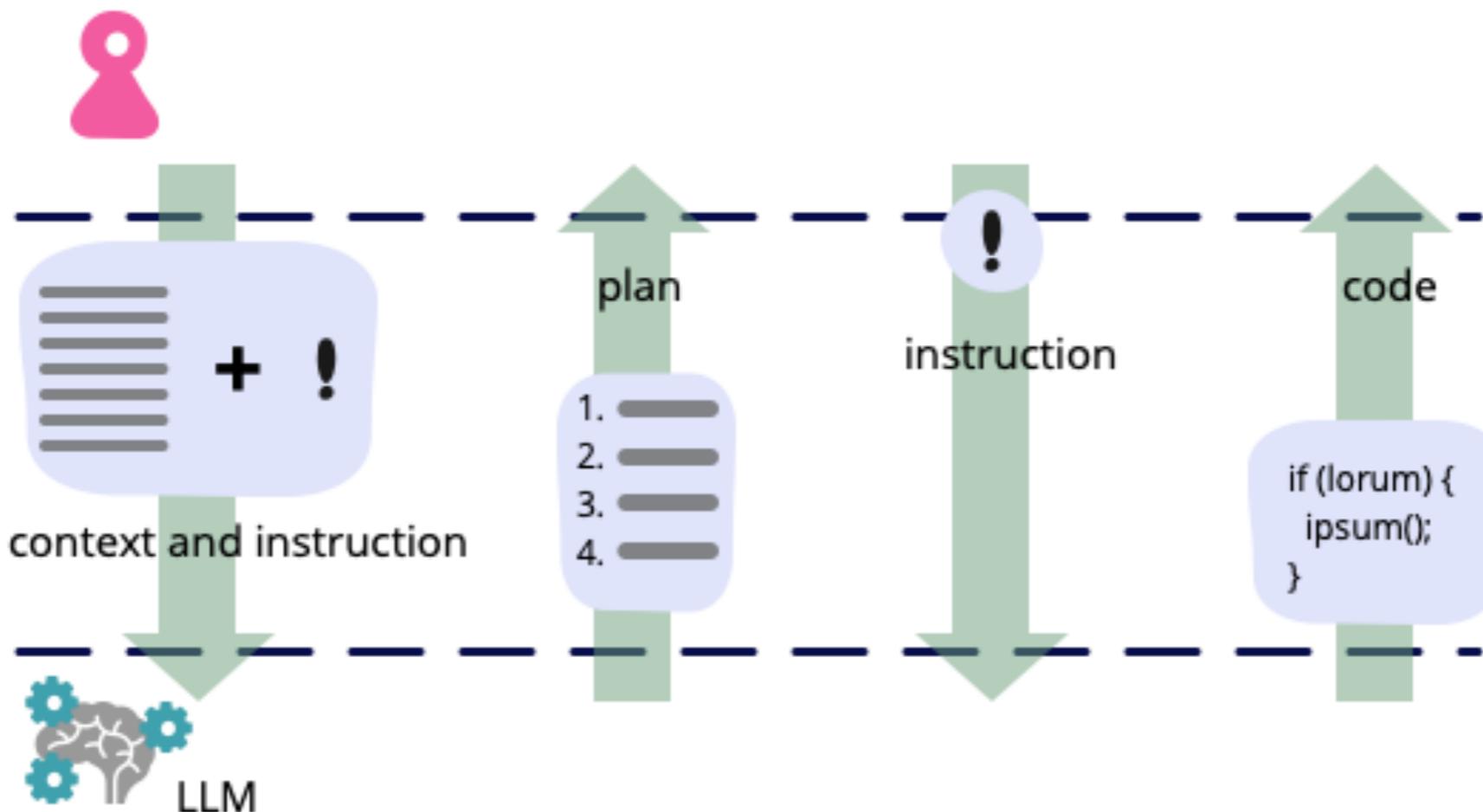
Copilot in VSCode  
Cursor IDE  
Windsurf  
Zed

AI Agent  
Public and Local

Aider  
Claude code  
Gemini CLI  
Codex CLI  
Qwen3-coder



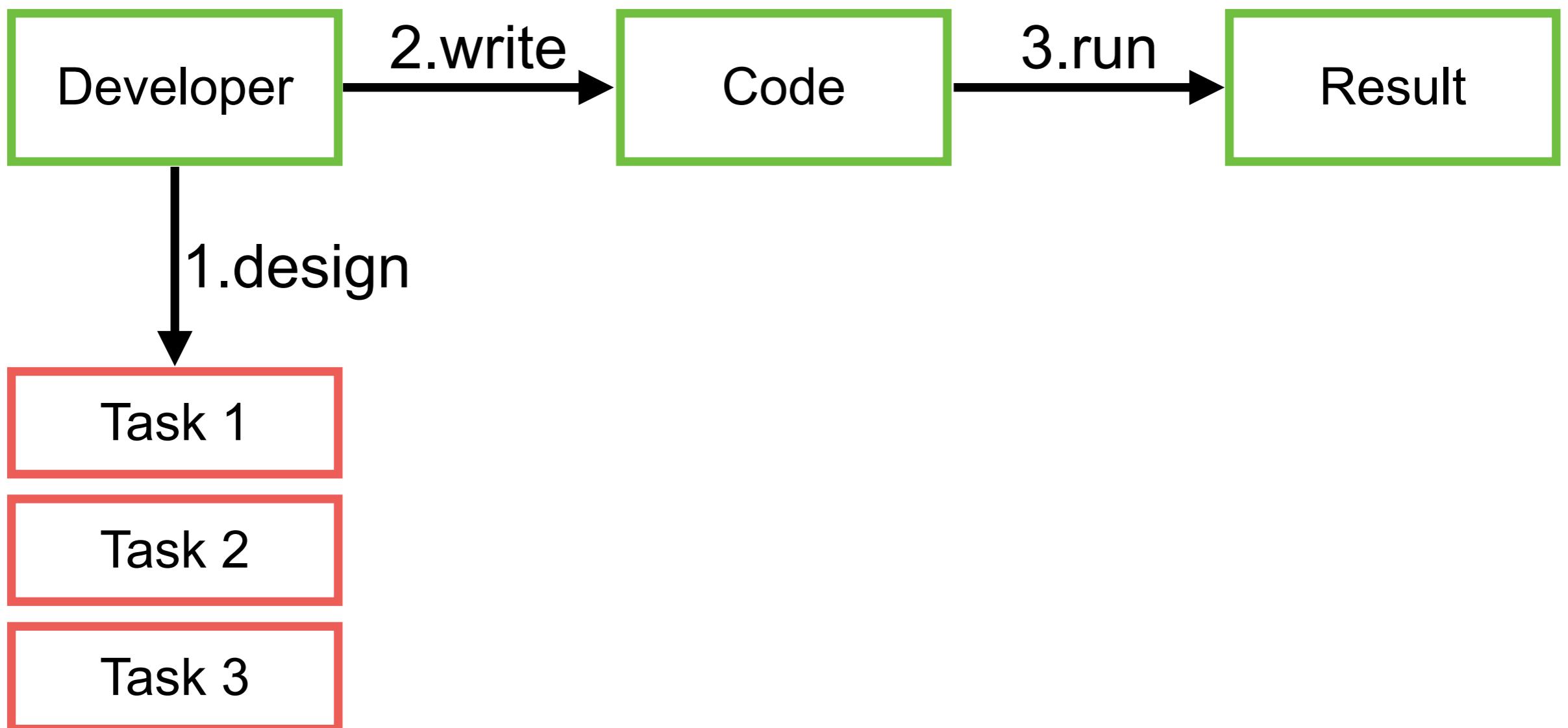
# Development



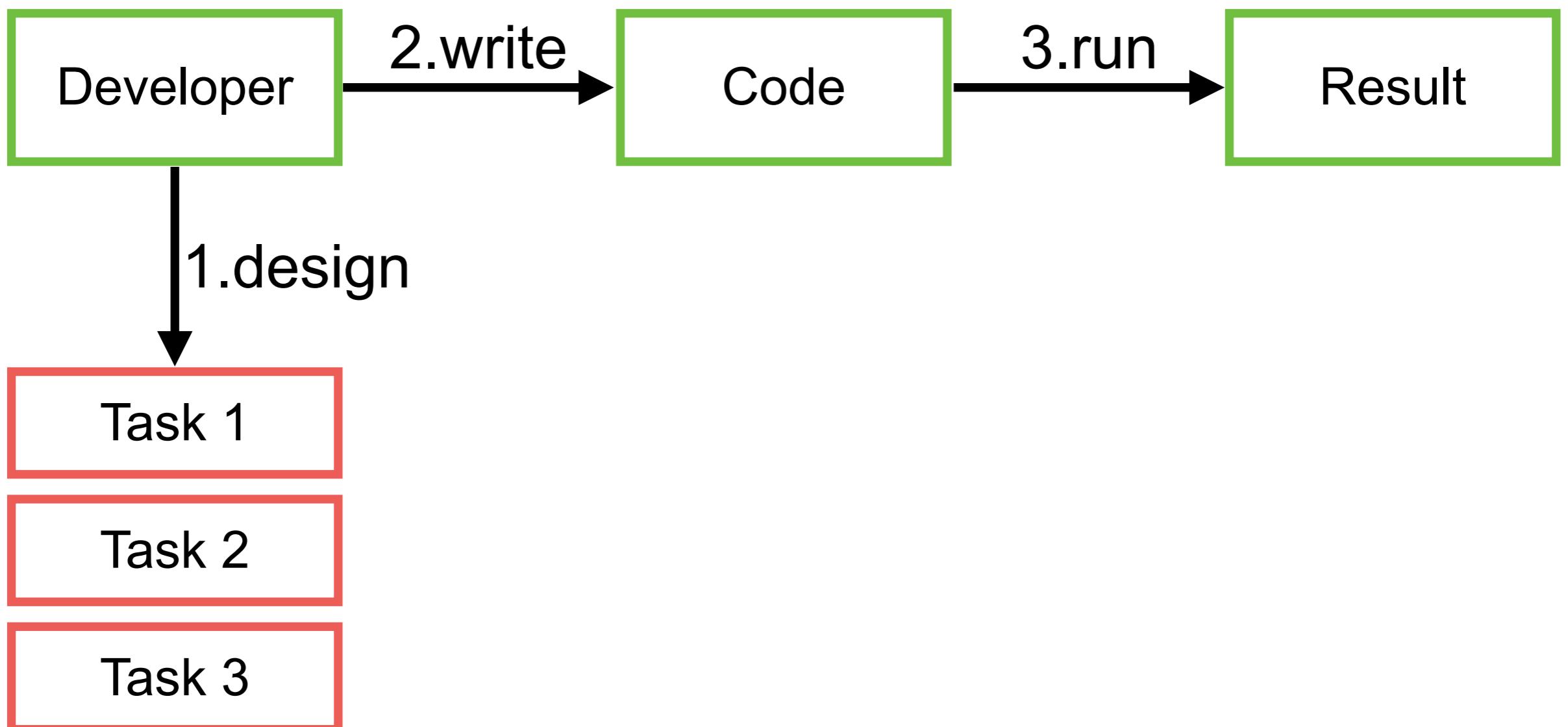
<https://martinfowler.com/articles/2023-chatgpt-xu-hao.html>



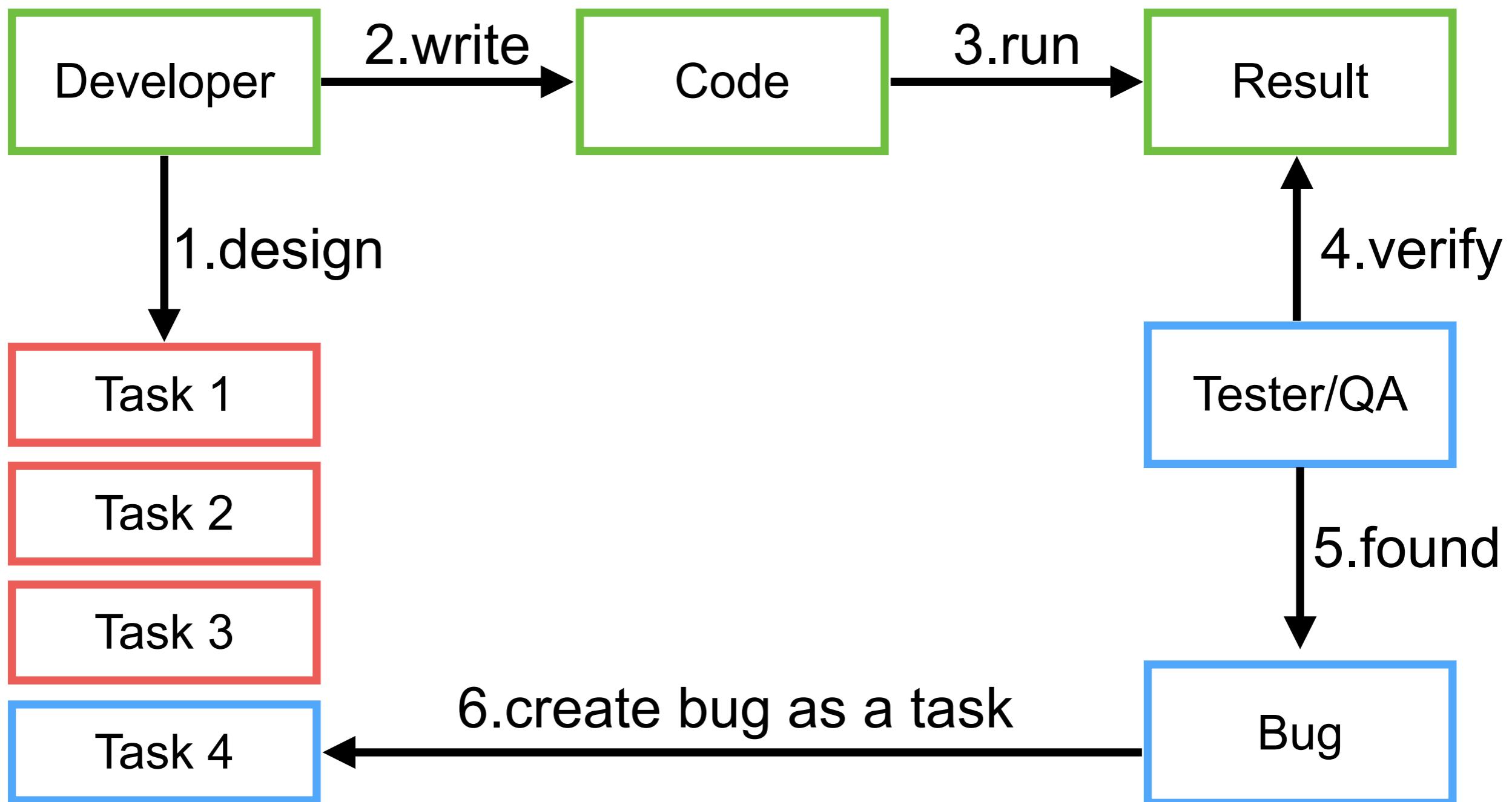
# Development



# Development



# Development + Testing



# Main Features

Ask question

Generate code

Refactor code

Document code

Find problems in your code



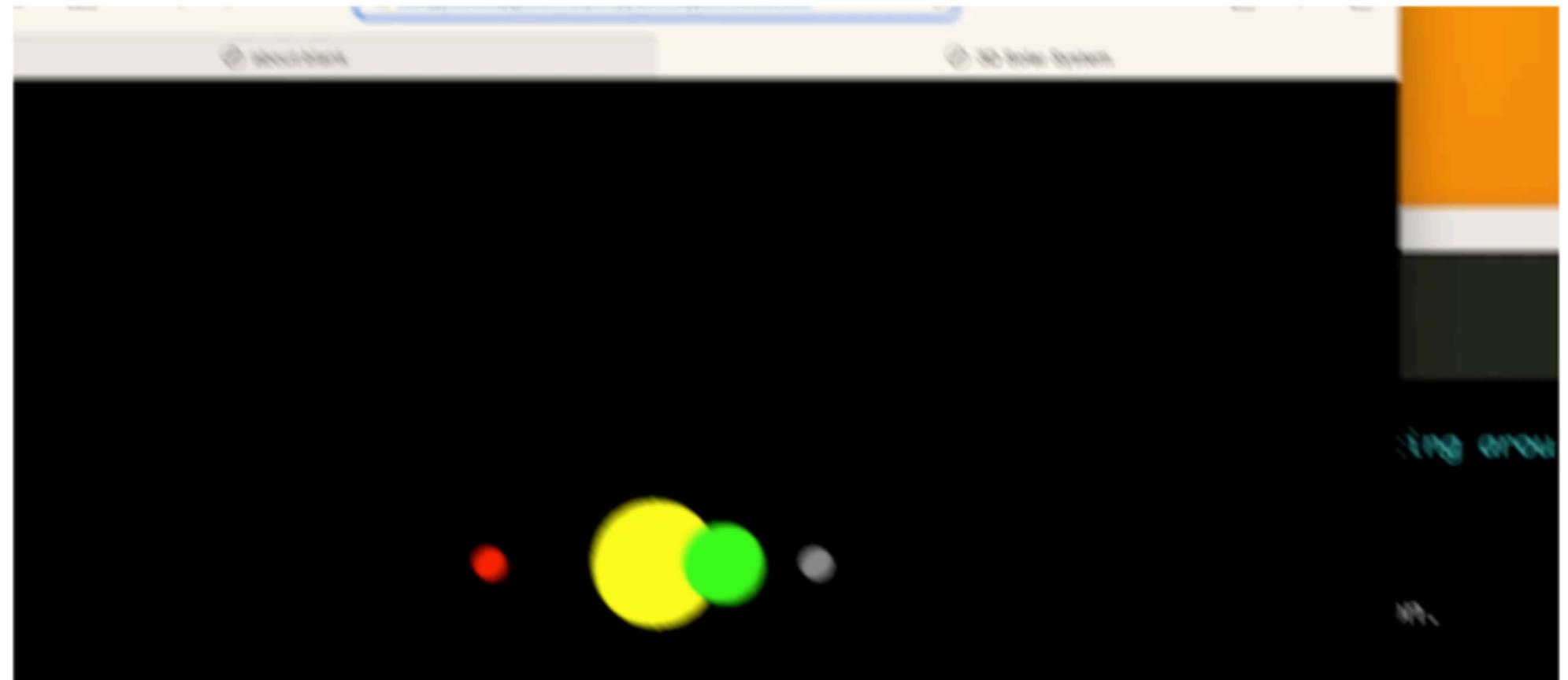
# Pair programming with AI



# AI Pair Programming

Aider is AI pair programming in your terminal

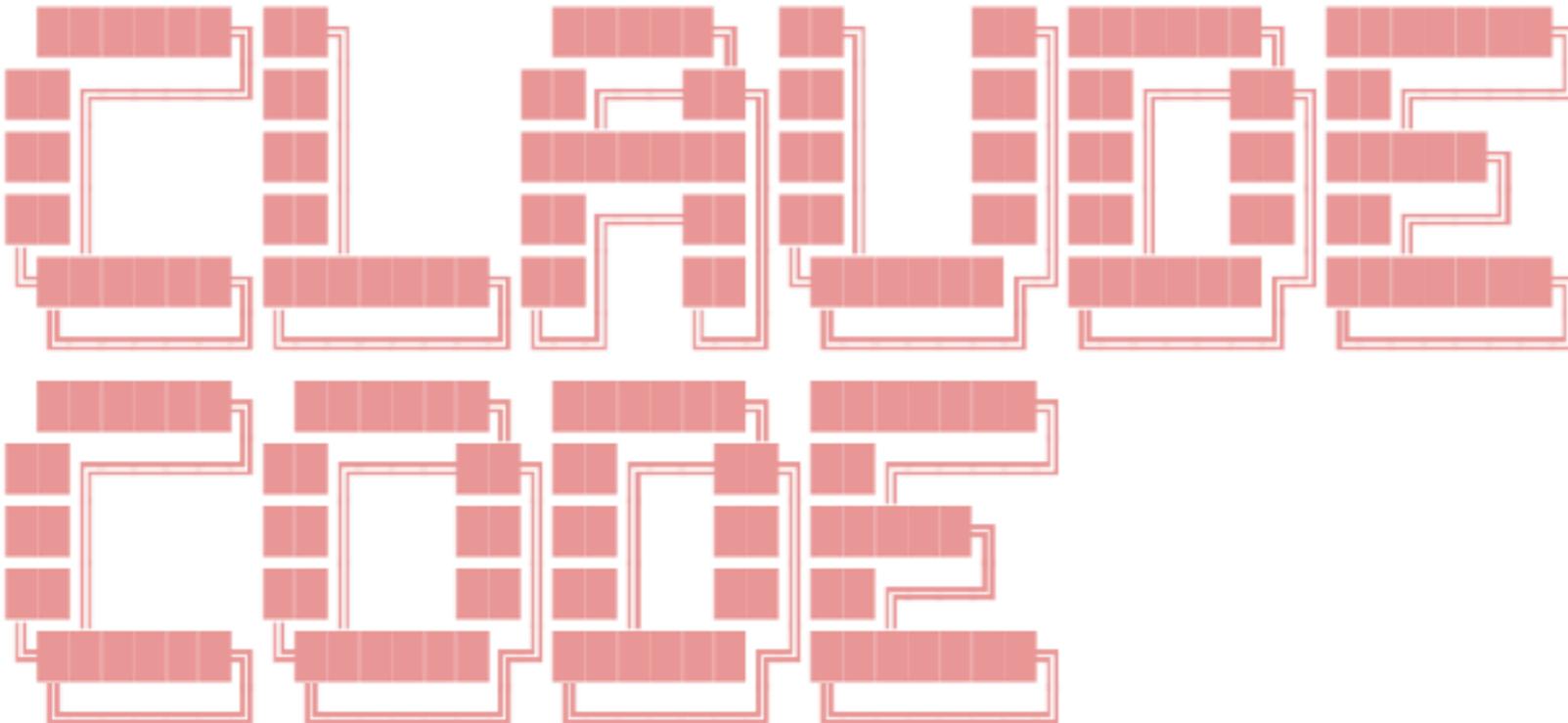
Aider lets you pair program with LLMs, to edit code in your local git repository. Start a new project or work with an existing git repo. Aider works best with GPT-4o & Claude 3.5 Sonnet and can [connect to almost any LLM](#).



<https://github.com/paul-gauthier/aider>



\* Welcome to **Claude Code** research preview!



**Claude Code is billed based on API usage through your Anthropic Console account.**

Pricing may evolve as we move towards general availability.

Press Enter to login to your Anthropic Console account...

2025/02

<https://www.anthropic.com/news/clause-3-7-sonnet>





2025/06

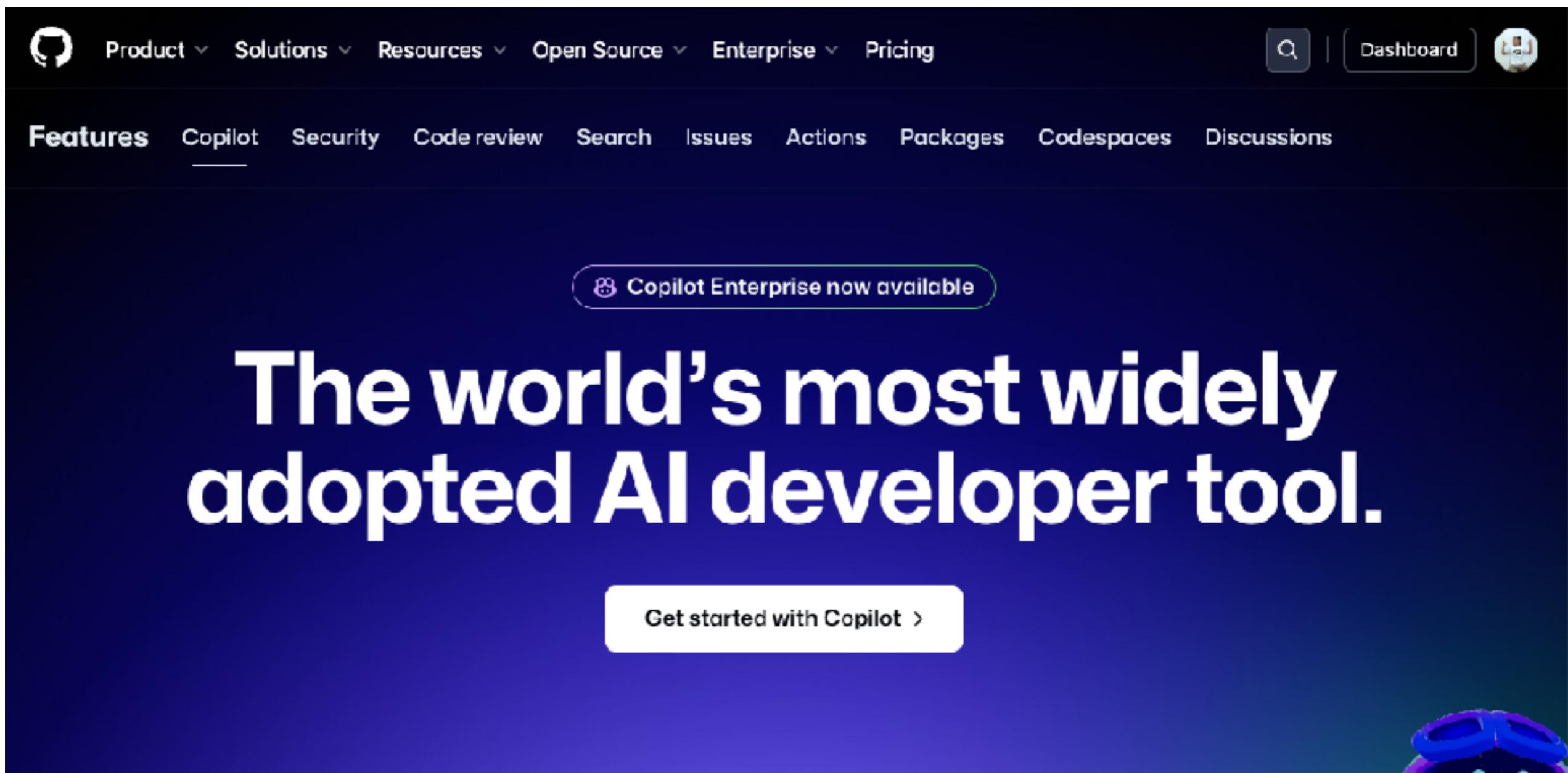
<https://github.com/google-gemini/gemini-cli>



# AI in IDE



# GitHub Copilot



The screenshot shows the GitHub Copilot homepage. At the top, there's a navigation bar with links for Product, Solutions, Resources, Open Source, Enterprise, Pricing, a search bar, a dashboard button, and a user profile icon. Below the navigation is a secondary navigation bar with links for Features, Copilot (which is underlined), Security, Code review, Search, Issues, Actions, Packages, Codespaces, and Discussions. A prominent banner in the center says "Copilot Enterprise now available". The main headline reads "The world's most widely adopted AI developer tool." Below it is a button labeled "Get started with Copilot >". In the bottom right corner, there's a small, stylized blue and purple AI character.

<https://github.com/features/copilot>



# Cursor.sh



Pricing   Features   Forum   Docs   Careers   Blog

Sign In

Download

# The AI Code Editor

Built to make you extraordinarily productive, Cursor is the best way to code with AI.



Download for Free



Watch Demo  
1 Minute

<https://www.cursor.com/>



AI for Software Development  
© 2020 - 2026 Siam Chamnankit Company Limited. All rights reserved.

# Cursor Directory

cursor.directory

Get latest updates   [Subscribe](#)   [Live](#) [Learn](#) [About](#)

Search...  All Popular

Topic	Count
TypeScript	15
Python	9
Next.js	9
React	9
PHP	5
C#	4
Expo	4
React Native	4
Tailwind	4
Supabase	4
Web Development	3
Game Development	3
JavaScript	3
Laravel	3

**TypeScript**

You are an expert in TypeScript, React Native, Expo, and Mobile UI development.

Code Style and Structure

- Write concise, technical TypeScript code; accurate examples.
- Use functional and declarative programming patterns; avoid classes.
- Prefer iteration and modularization over duplication.
- Use descriptive variable names with auxiliary verbs (e.g., isLoading, hasError).
- Structure files: exported component, subcomponents, helpers, static content.
- Follow Expo's official documentation for best practices.

**Krish Kalaria**   
expo-router expo-status-bar +7 more ~

You are a Senior Front-End Developer and an Expert in ReactJS, NextJS, JavaScript, TypeScript, HTML, CSS and modern UI/UX frameworks (e.g., TailwindCSS, Shadon, Radix). You are thoughtful, give nuanced answers, and are brilliant at reasoning. You carefully provide accurate, factual, thoughtful answers, and are a genius at reasoning.

- Follow the user's requirements carefully & to the letter.

- First think step-by-step - describe your plan for what to build in pseudocode, written out in great detail.

**Mohammadali Karimi**   
Tailwind CSS Shadon UI +1 more ~

You are an expert in TypeScript, Node.js, Next.js App Router, React, Shadon UI, Radix UI and Tailwind.

Code Style and Structure

- Mditate concisely, functionally, TypeScript-like.

**Nathan Brachotte**   
gatsby react +2 more ~

You are an expert in Solidity, TypeScript, Node.js, Next.js 14 App Router, React, Vite, View v2, Wagmi v2, Shadon UI, Radix UI, and Tailwind Aria.

[Submit +](#)

<https://cursor.directory/>



# Windsurf by Codeium



## Built to keep you in *flow state*

The first agentic IDE, and then some. The Windsurf Editor is where the work of developers and AI truly flow together, allowing for a coding experience that feels like literal magic.

 Download the Windsurf Editor

See all download options

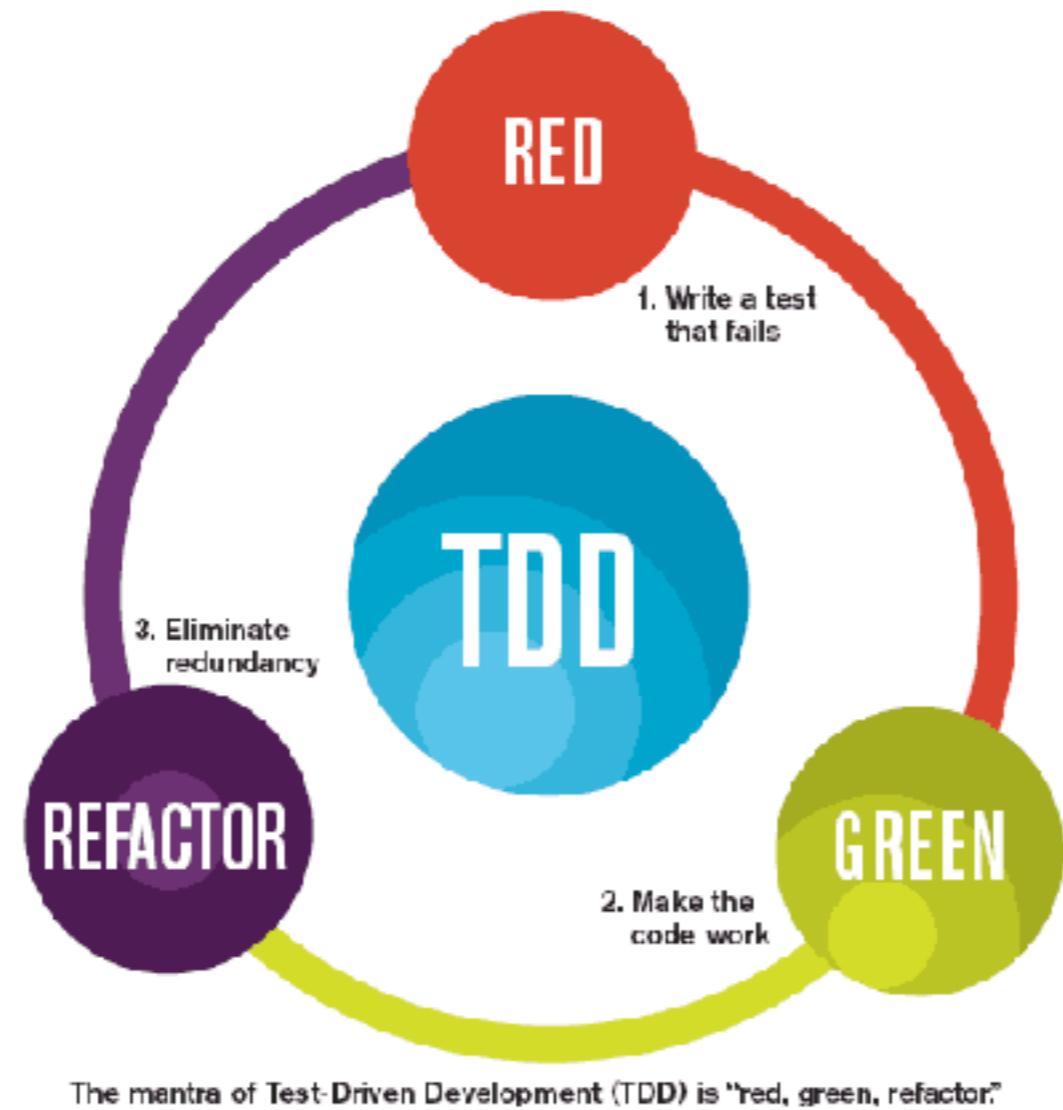
<https://codeium.com/windsurf>



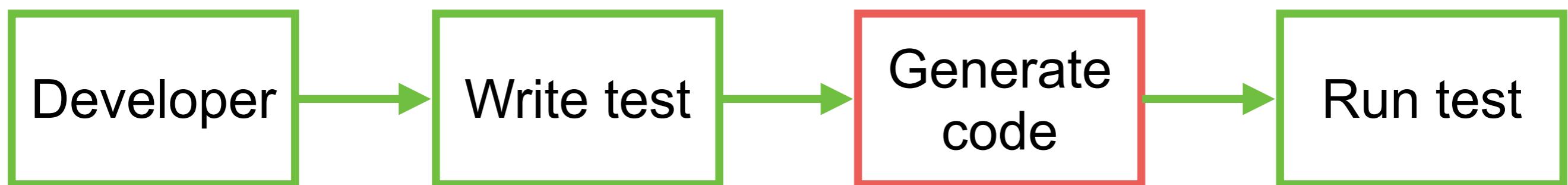
# Test-Driven Development



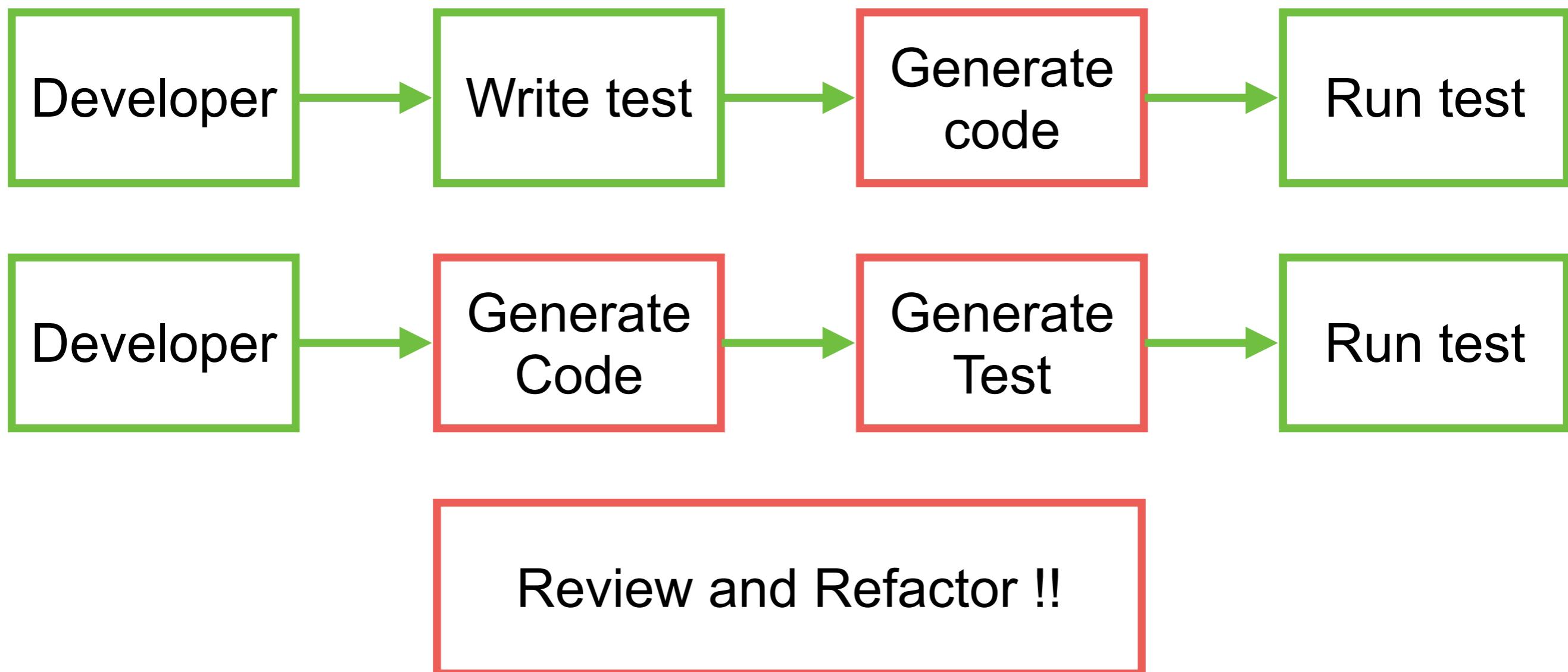
# Test-Driven-Development



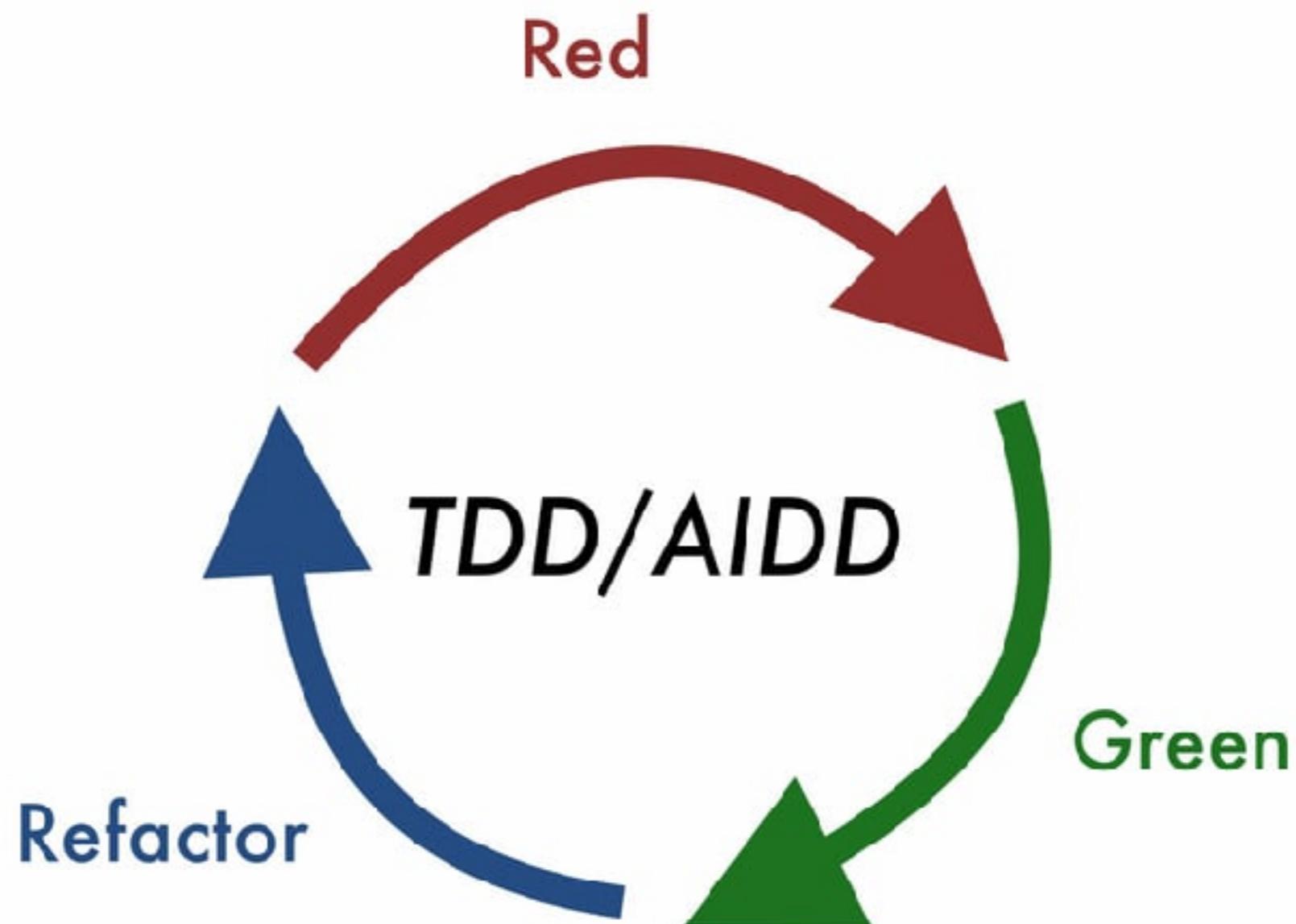
# Test-Driven-Development



# Test-Driven-Development



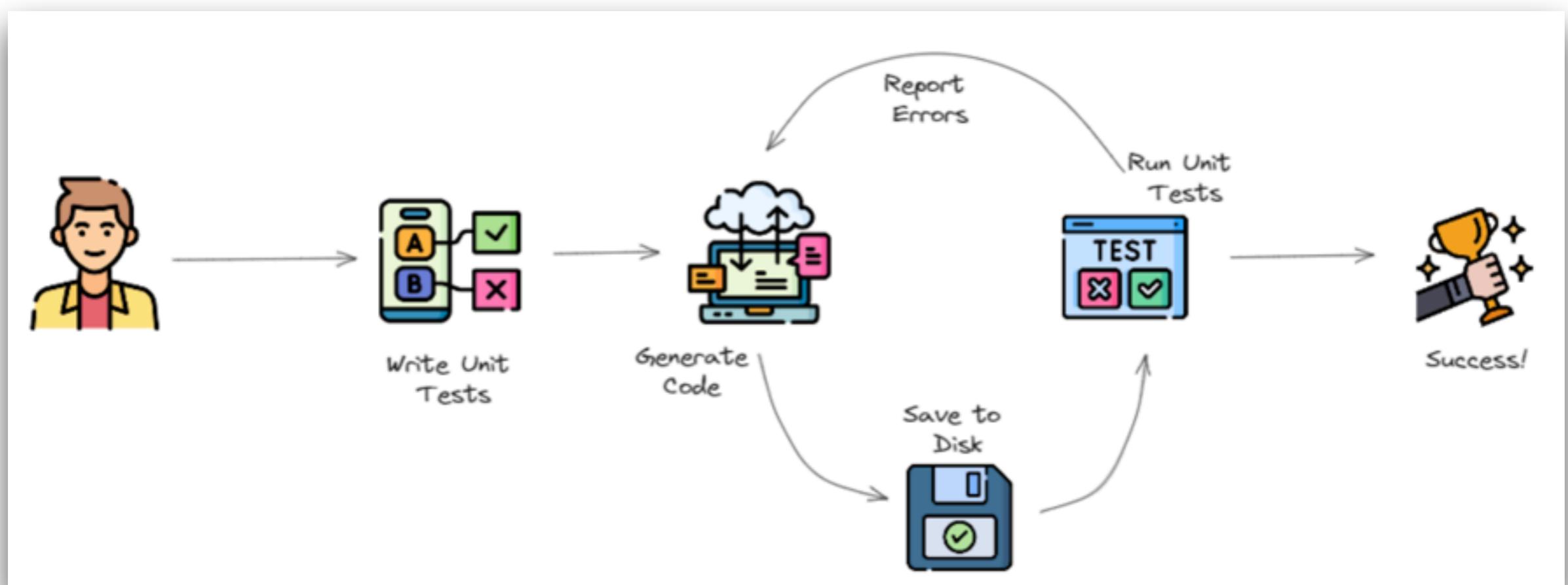
# AI-DD



<https://dev.to/dawiddahl/ai-is-changing-the-way-we-code-ai-driven-development-aidd-2ngo>



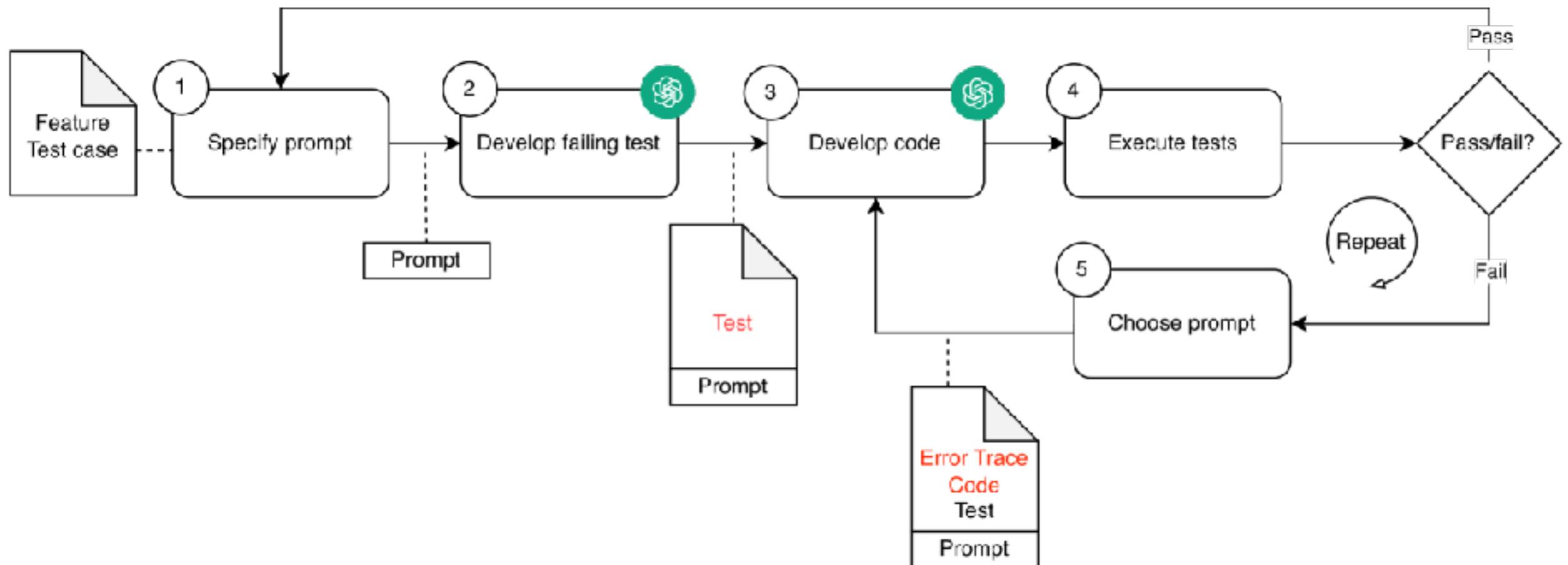
# TDD with AI



<https://github.com/allenheltondev/tdd-ai>



# TDD with AI



<https://arxiv.org/html/2405.10849v1>



# Test-Driven-Generation (TDG)



# Test-Driven-Generation (TDG)

Development practice that integrate Generative AI into the development life-cycle

TDD

+

Pair  
programming

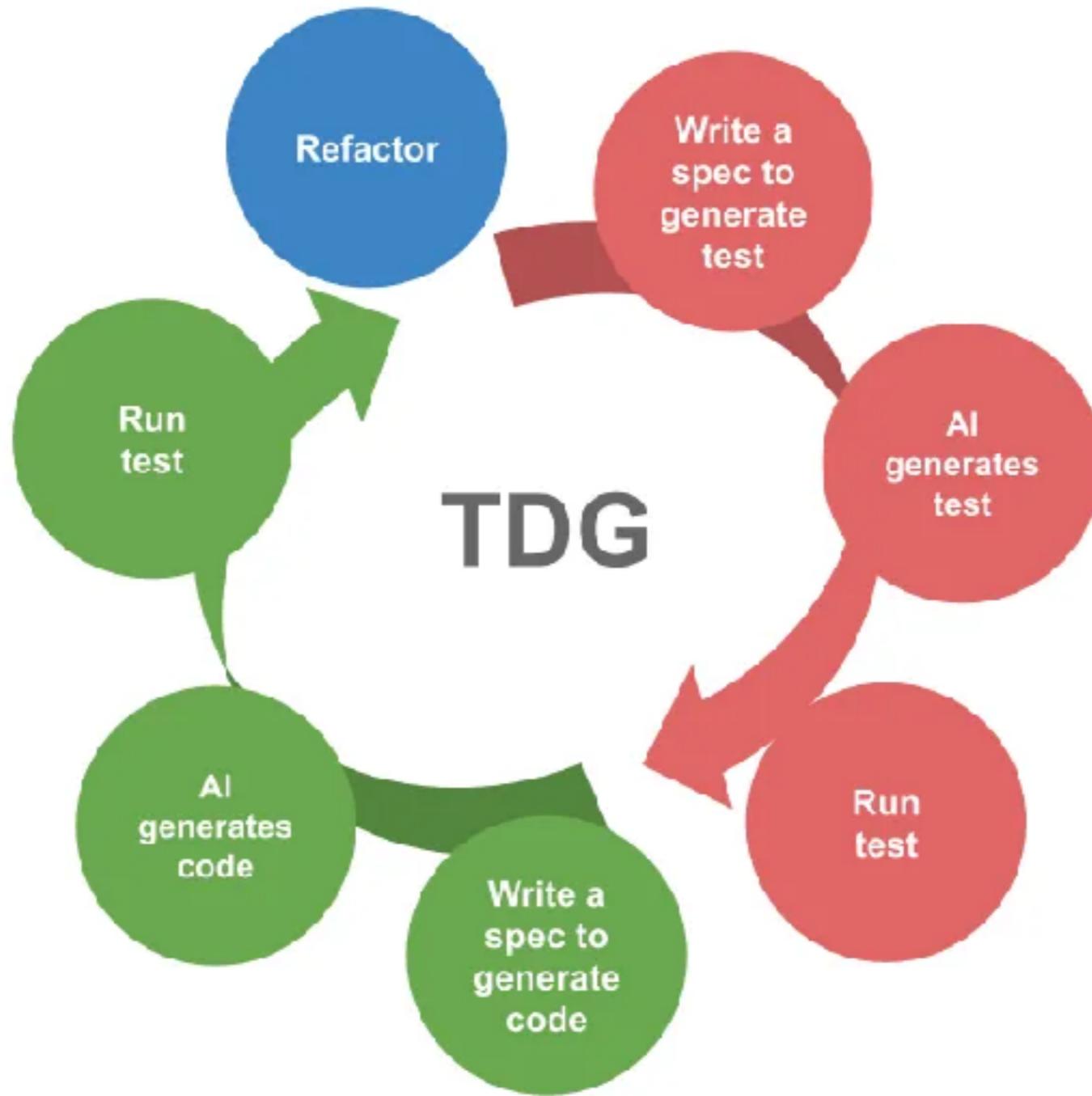
+

Generative AI

<https://chanwit.medium.com/test-driven-generation-tdg-adopting-tdd-again-this-time-with-gen-ai-27f986bed6f8>



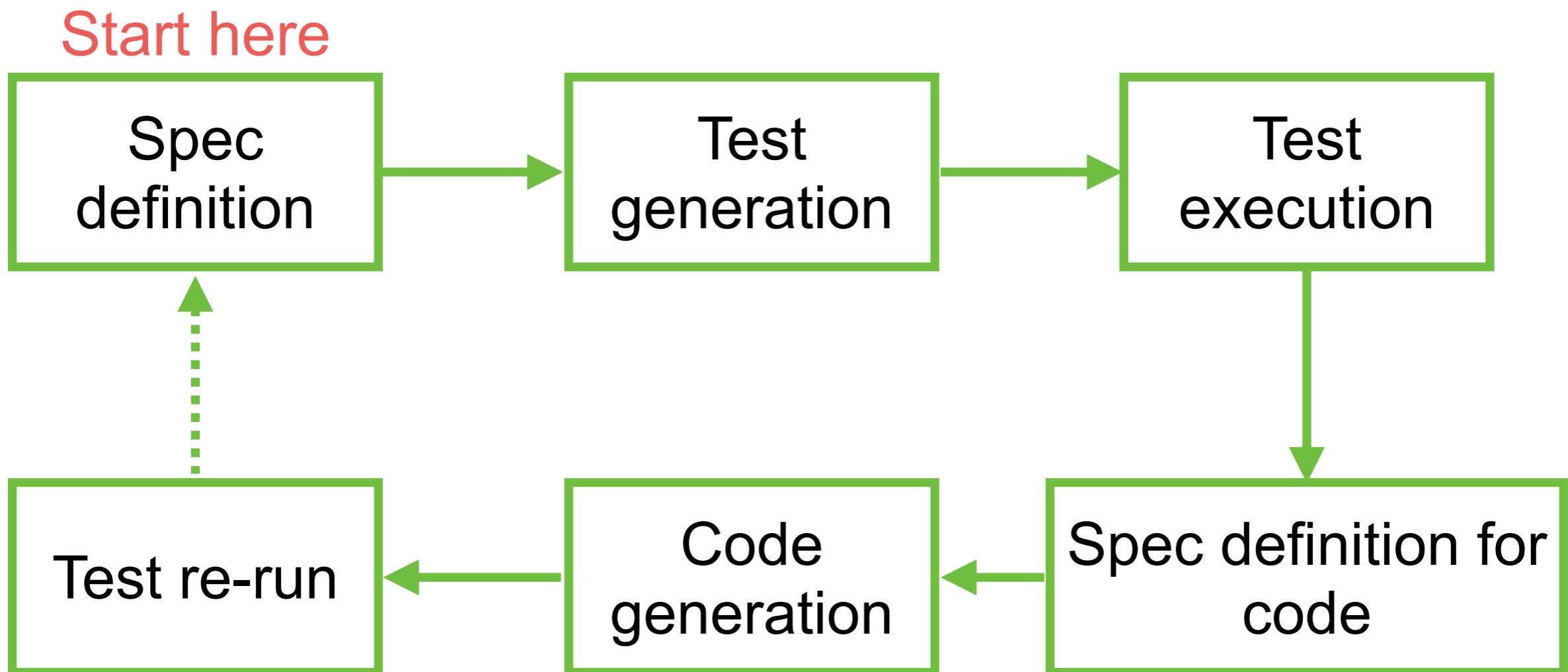
# Test-Driven-Generation (TDG)



<https://chanwit.medium.com/test-driven-generation-tdg-adopting-tdd-again-this-time-with-gen-ai-27f986bed6f8>



# Test-Driven-Generation (TDG)



# Tips and Techniques with AI

Scope of prompt

Error in result

Setup and config  
project

Latest information

Use technical  
keywords



# **Specification-Driven Development (SDD)**



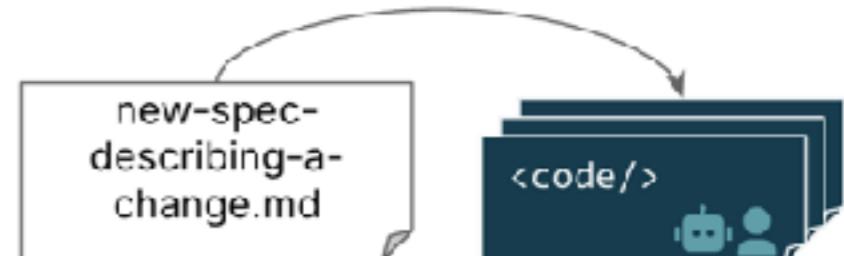
# Workflow SDD

## Levels of “spec-driven”

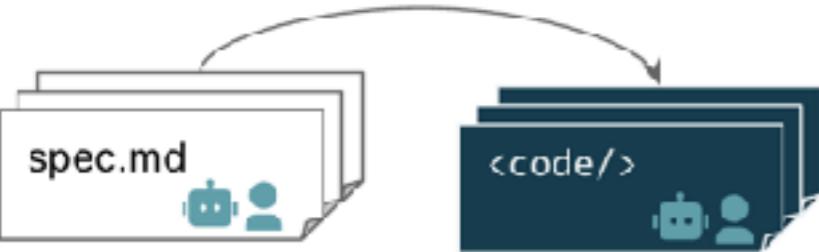
Creation of feature



Evolution and maintenance of feature



**spec-first**



**spec-anchored**



**spec-as-source**

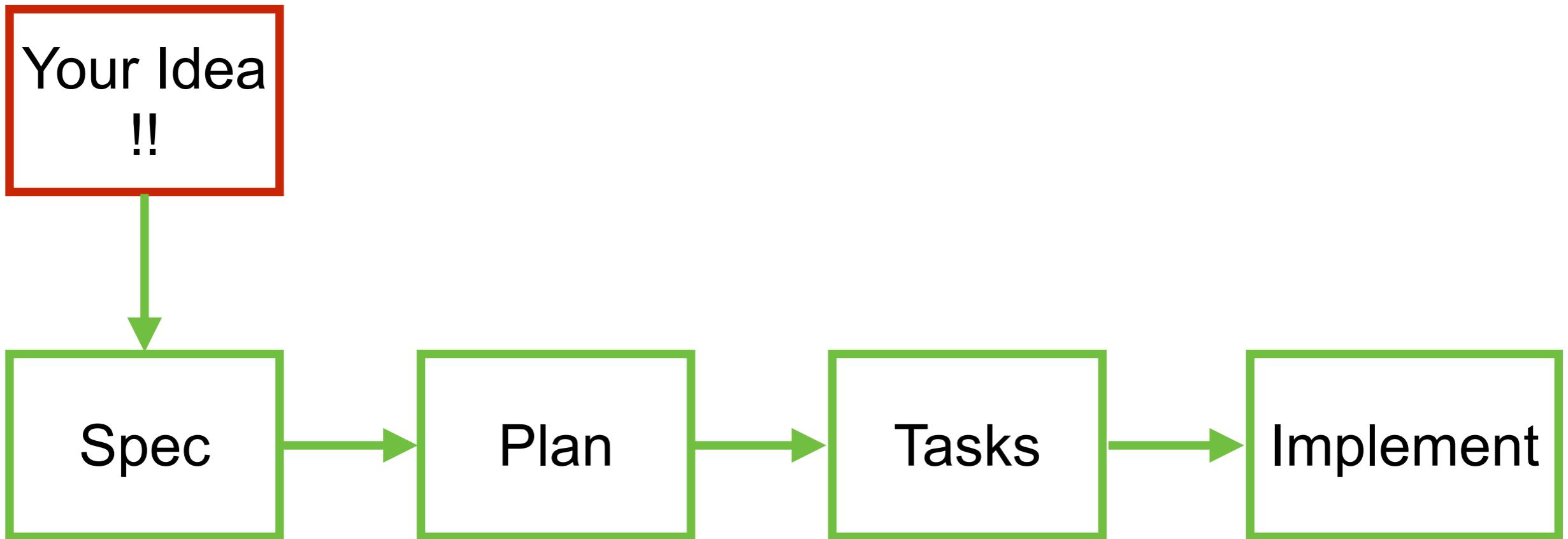


<https://martinfowler.com/articles/exploring-gen-ai.html>

<https://martinfowler.com/articles/exploring-gen-ai/sdd-3-tools.html>



# Workflow SDD



# Memory Bank

## Memory bank

AGENTS.md

project.md

architecture.md

*Examples for illustration,  
actual file structures vary*

## Agent

## Specs

STORY-  
324.md

STORY-  
525.md

product-  
search.md

config-  
loader.md

## feature-x

data-mo  
del.md

plan.  
md

contr  
acts

*Examples for illustration,  
actual file structures vary*



# Workflow Tools

AWS Kiro

Spec-kit

Custom by IDE

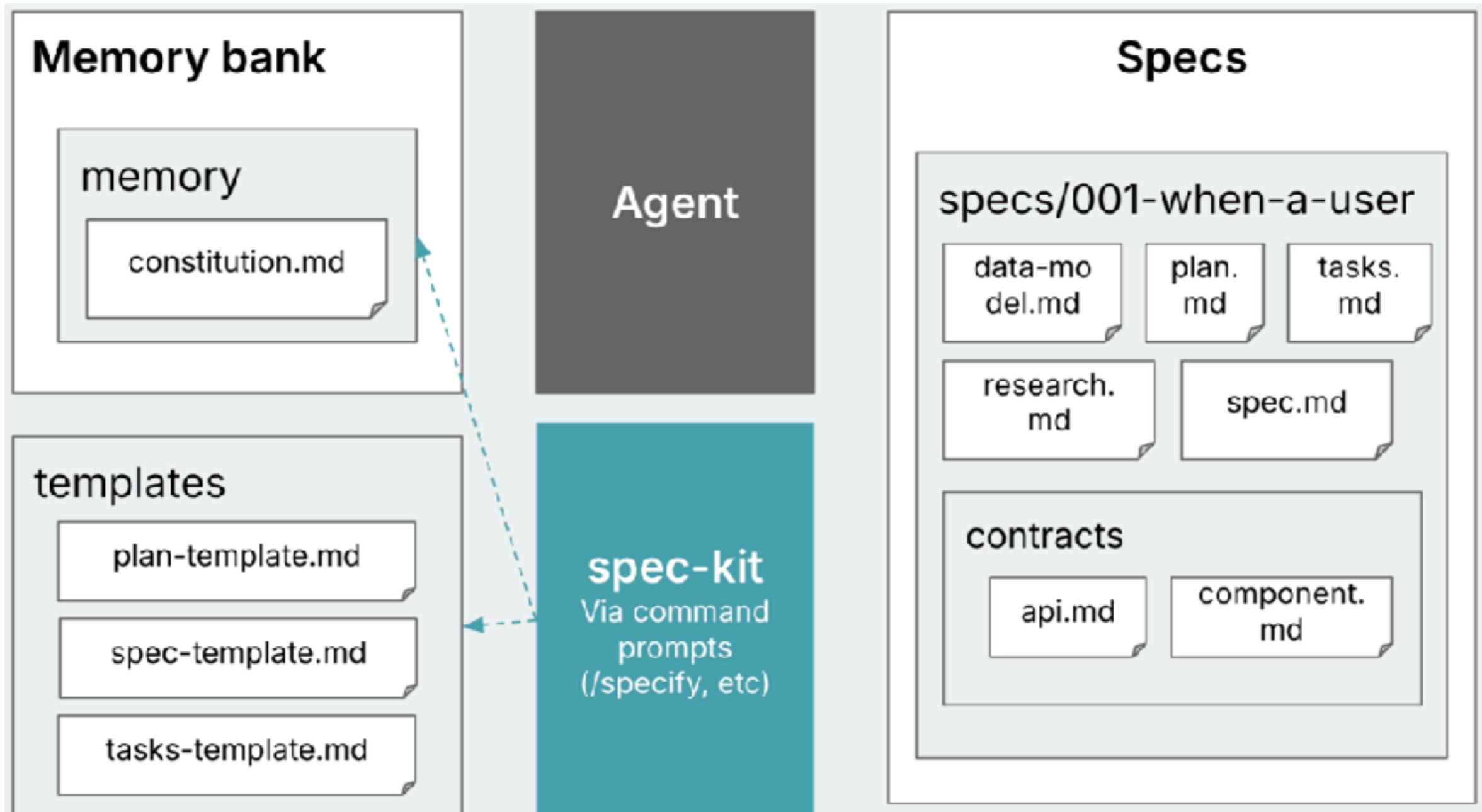
Requirement  
Design  
Tasks

Specify  
Plan  
Tasks

VSCode rules  
Cursor rules  
Github rules



# Spec-Kit



# Workshop with Coding

Chat

Text Editor with AI

Pair programming  
with AI

SDD and Rules

BMAD-METHOD



# Pair programming with AI

GPT-5

Claude 4.5  
Sonnet

DeepSeek  
Coder

Llama

<https://github.com/up1/workshop-ai-with-technical-team/tree/main/workshop/aider>



# **Testing Process with High quality process**

**Functional**

**Non-Functional**



# Testing

Requirement

Design

Develop

Testing

Deploy

Test cases writing

Test code generation

Bug detection

Test planning

Data test generation



# 6 Keys Software Quality

Defect density

Code duplication

Hardcode token/key

Security  
vulnerabilities

Outdated package

Non-permissive  
opensource libraries

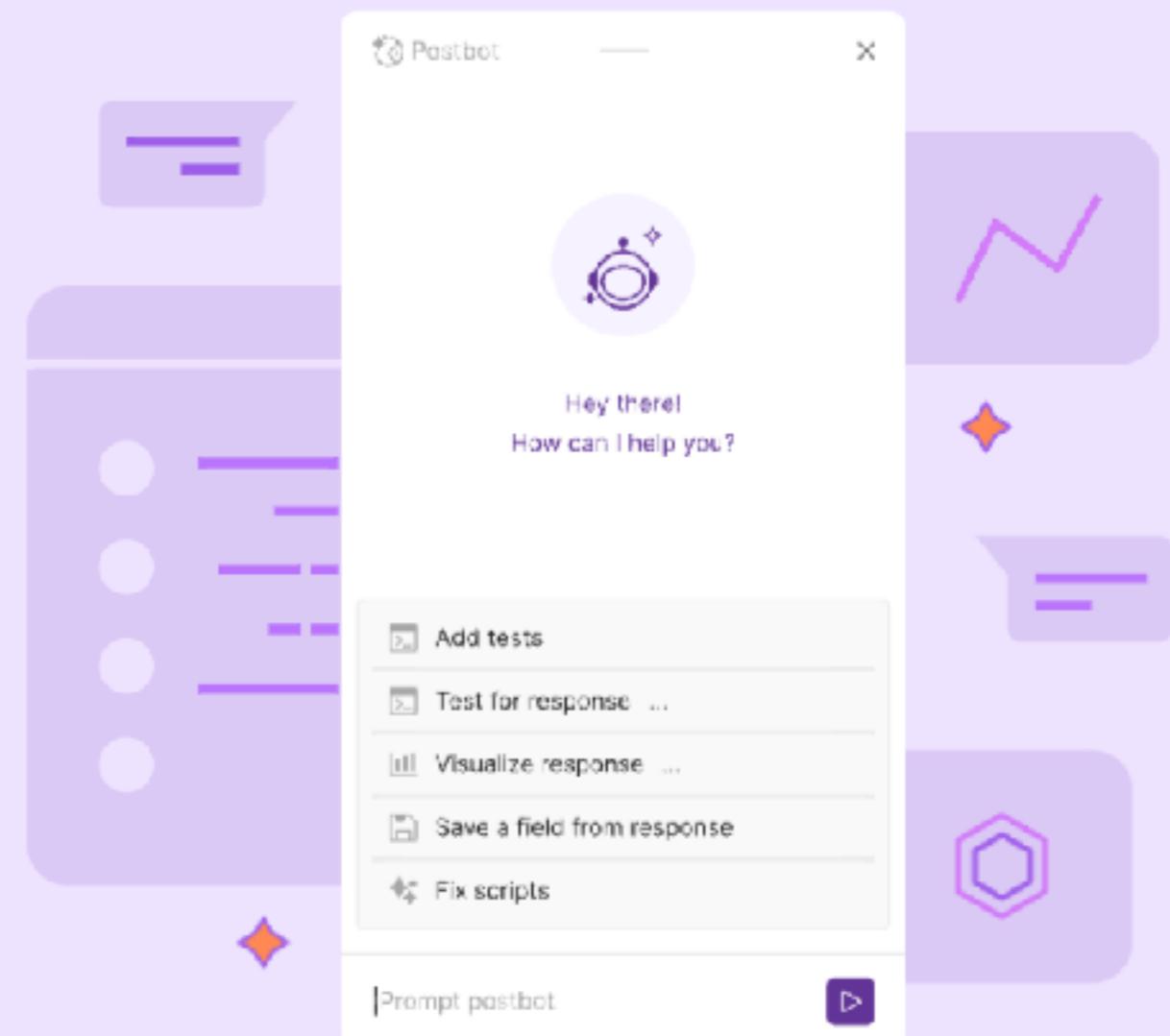


# Testing with Postman

**Postbot, our AI-powered assistant, will supercharge your API development.**

Speed up your most common API development workflows with natural-language input, conversational interactions, and contextual suggestions.

[Get Started](#)



<https://www.postman.com/product/postbot/>



# PostBot

The screenshot shows the Postman application interface. A request is being made to `https://jsonplaceholder.typicode.com/users/1` using the `GET` method. The `Tests` tab is selected in the request header bar. The response body is displayed in Pretty JSON format, showing a user object with fields like id, name, username, email, and address. A context menu is open over the response body, listing options such as `Add tests to this request`, `Test for response...`, `Visualize response...`, `Save a field from response`, and `Add documentation`. A modal window titled `New on Postbot` provides information about the AI feature, stating it can auto-complete tests to help work faster and suggest important tests based on available responses.

```
1 {  
2   "id": 1,  
3   "name": "Leanne Graham",  
4   "username": "Bret",  
5   "email": "Sincere@april.biz",  
6   "address": {  
7     "street": "Kulas Light",  
8     "suite": "Apt. 556".  
9   }  
10 }
```

<https://www.postman.com/product/postbot/>



# Browser Use

 Browser Use

FEATURES

PRICING

BLOG

DOCUMENTATION

72,343

26.3K

23.6K

CLOUD

[BETA] THE MOST STEALTH BROWSER INFRASTRUCTURE

## The AI browser agent

Repetitive work is dead. Browser Use empowers anyone to automate repetitive online tasks, no code required. No barriers. Simply tell it what you want done.

<https://browser-use.com/>



AI for Software Development

© 2020 - 2026 Siam Chamnankit Company Limited. All rights reserved.

# Playwright Test Agent



<https://playwright.dev/docs/test-agents>



# Deployment Process



# Deploy and manage

Requirement

Design

Develop

Testing

Deploy

CI/CD pipeline

Infrastructure as a code

Automated script

Performance and monitoring suggestion

Document generation

AI-assist support

ChatOps, AIOps



# K8sGPT



## CLOUD NATIVE K8sGPT joins the SANDBOX CNCF Sandbox

K8sGPT is a tool for scanning your kubernetes clusters, diagnosing and triaging issues in simple english. It has SRE experience codified into its analyzers and helps to pull out the most relevant information to enrich it with AI.

Get it now!

<https://k8sgpt.ai/>



# PromptOps



Solutions ▾ Resources ▾ Contact us Log In

## ChatGPT for your DevOps Teams

Turn DevOps tasks into automated workflows with a single prompt straight from Slack

Get started

Learn More

A screenshot of the PromptOps Slack app interface. At the top, there's a message from the "PromptOps APP" bot: "Hello, I'm PromptOps, your DevOps virtual assistant for managing, troubleshooting, and running DevOps tasks directly from Slack!". Below this, a Slack conversation shows a user named "Sergoy" reporting an issue: "Uh oh, MOAR 408 errors in graph-engine. Help, @PromptOps!". The "PromptOps APP" bot responds with "Eleven ran into this issue last week" and suggests a fix: "aws easily fixed it by bumping up CPU cores on nginx node. Should I do it?". Two buttons at the bottom right of the message are "Yes" and "No". At the bottom of the screenshot, there are several buttons: "Acknowledge", "Resolve", "Run a play ▾", and "Edit code".

PromptOps APP  
Hello, I'm PromptOps, your DevOps virtual assistant for managing, troubleshooting, and running DevOps tasks directly from Slack!

Sergoy 02:14am  
Uh oh, MOAR 408 errors in graph-engine.  
Help, @PromptOps!

PromptOps APP 02:14am  
Eleven ran into this issue last week  
aws easily fixed it by bumping up CPU cores on nginx node.  
Should I do it?

Yes No

Acknowledge Resolve Run a play ▾ Edit code

<https://www.promptops.com/devops/>



AI for Software Development  
© 2020 - 2026 Siam Chamnankit Company Limited. All rights reserved.

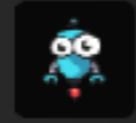
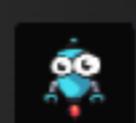
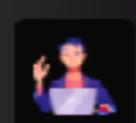
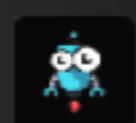
# ChatOps for DevOps

[Product](#)[How it Works](#)[Learn](#)[Company](#)[Uptime](#)[Sign In](#)[Book a demo](#)[Sign Up](#)

## > ChatGPT for DevOps

Converse with your engineering platforms, powered by LLM.  
A virtual teammate to handle DevOps requests so you can handle the rest.

[Add Kubi to Slack](#)

-  Kubi (DevOps)  
@Alerts I got an alert from Prometheus:  
  
Deployment 'alert-manager' on namespace 'Openfaas' is experiencing high traffic
-  Kubi (DevOps)  
@Alerts Should I increase the number of replicas on 'alert-manager'?
-  Jeff (R&D)  
Yes
-  Kubi (DevOps)  
  
✓ The following deployment has been updated:  
Deployment: alert-manager  
Namespace: Openfaas  
Replicas: 3

<https://www.kubiya.ai/>



# Risks when using Generative AI



# Risks

Quality of output generated  
Explainability of decisions  
Security policy !!  
Sensitive data !!



# Tips

Understand what you want

Modular approach

Clear and Precise inputs

Make sure you understand the code



# Local LLM



# Local LLM

Run LLM on local machine/device  
Try to customize with your requirement

Reduce cost

Data privacy

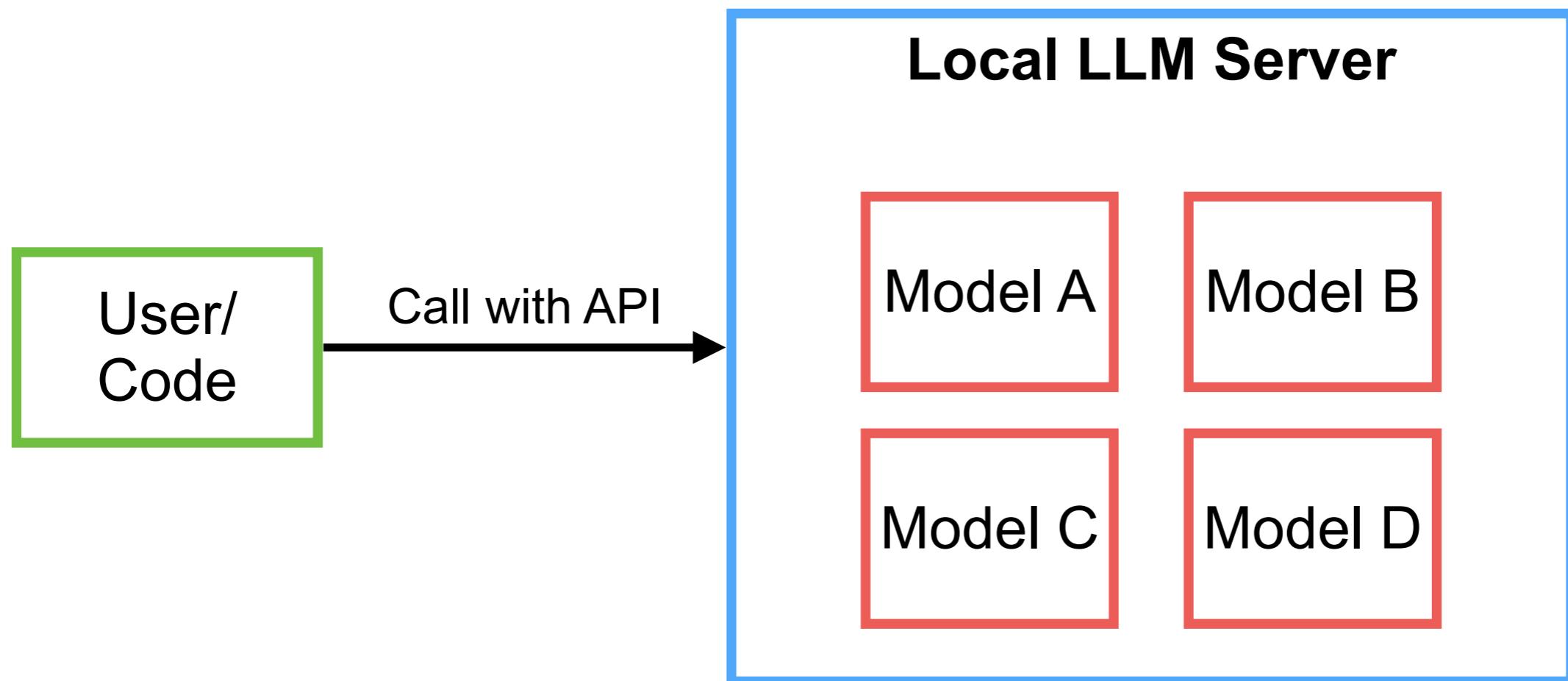
Responsive

Offline mode



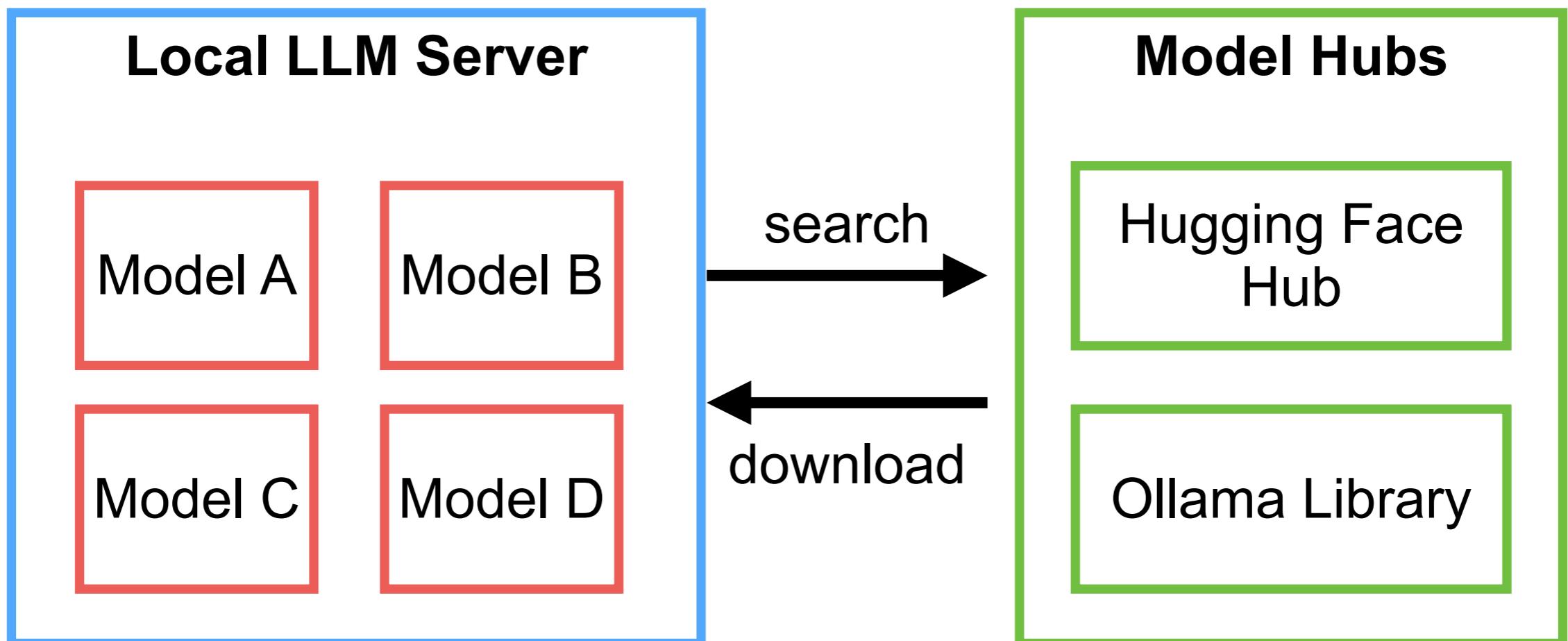
# Local LLM

Improve your LLM models, more accurate answer



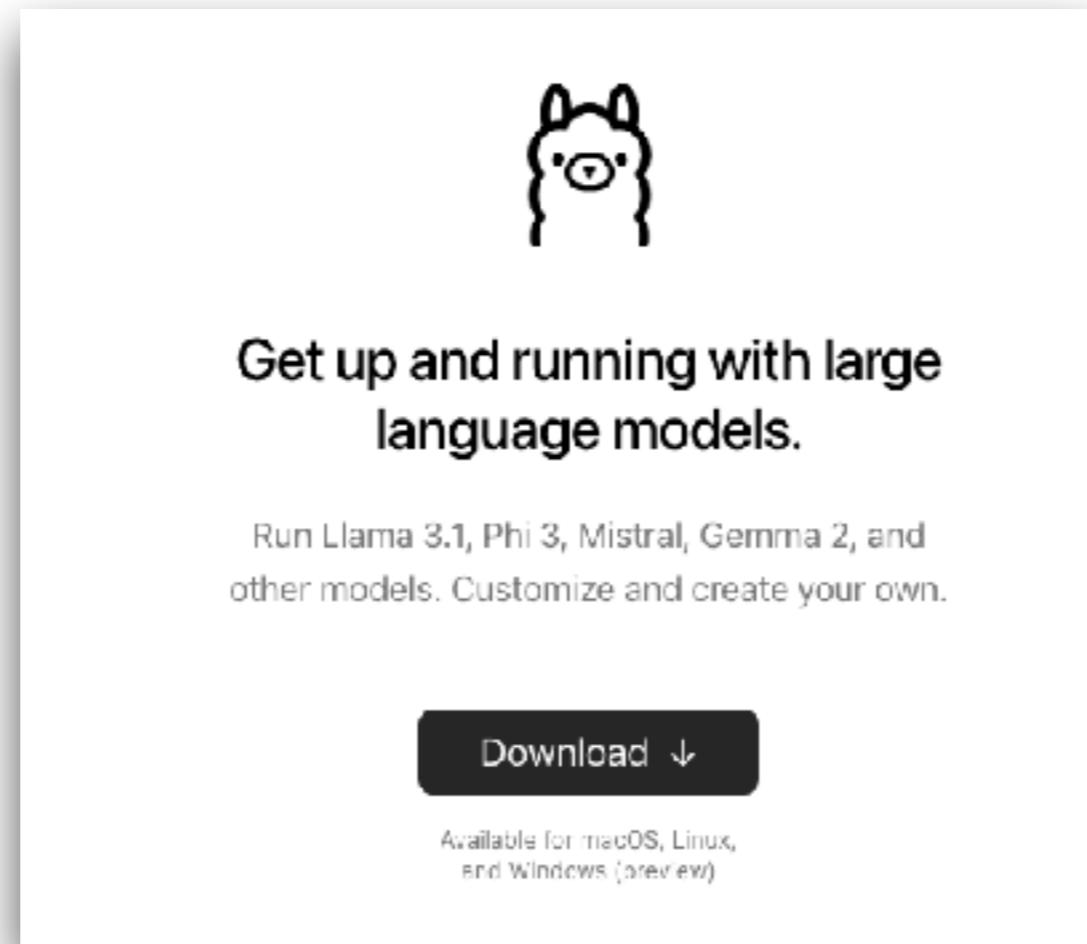
# Models ?

How to download models ?



# Local LLM with Ollama

\$ollama run **llama3.2**



<https://ollama.com/>



# Local LLM with LM Studio

The image shows the LM Studio website and its desktop application side-by-side.

**Website Screenshot:**

- Header:** LM Studio logo, Docs, Blog, Download.
- Title:** LM Studio
- Text:** Discover, download, and run local LLMs.
- Announcement:** LM Studio v0.3.0 is finally here! 🎉🎉🎉 Read the announcement.
- Run Buttons:** LLaMa, Phi, Gamma, DeepSeek, Owen, Mistral.
- Text:** Built with open source projects like [llama.cpp](#) and [lmstudio.js](#).
- Download Buttons:**
  - Download LM Studio for Mac (M1/M2/M3) 0.3.2
  - Download LM Studio for Windows (x64) 0.3.2
  - Download LM Studio for Linux (x86) 0.3.2
- Text:** LM Studio is provided under the [terms of use](#).

**Application Screenshot:**

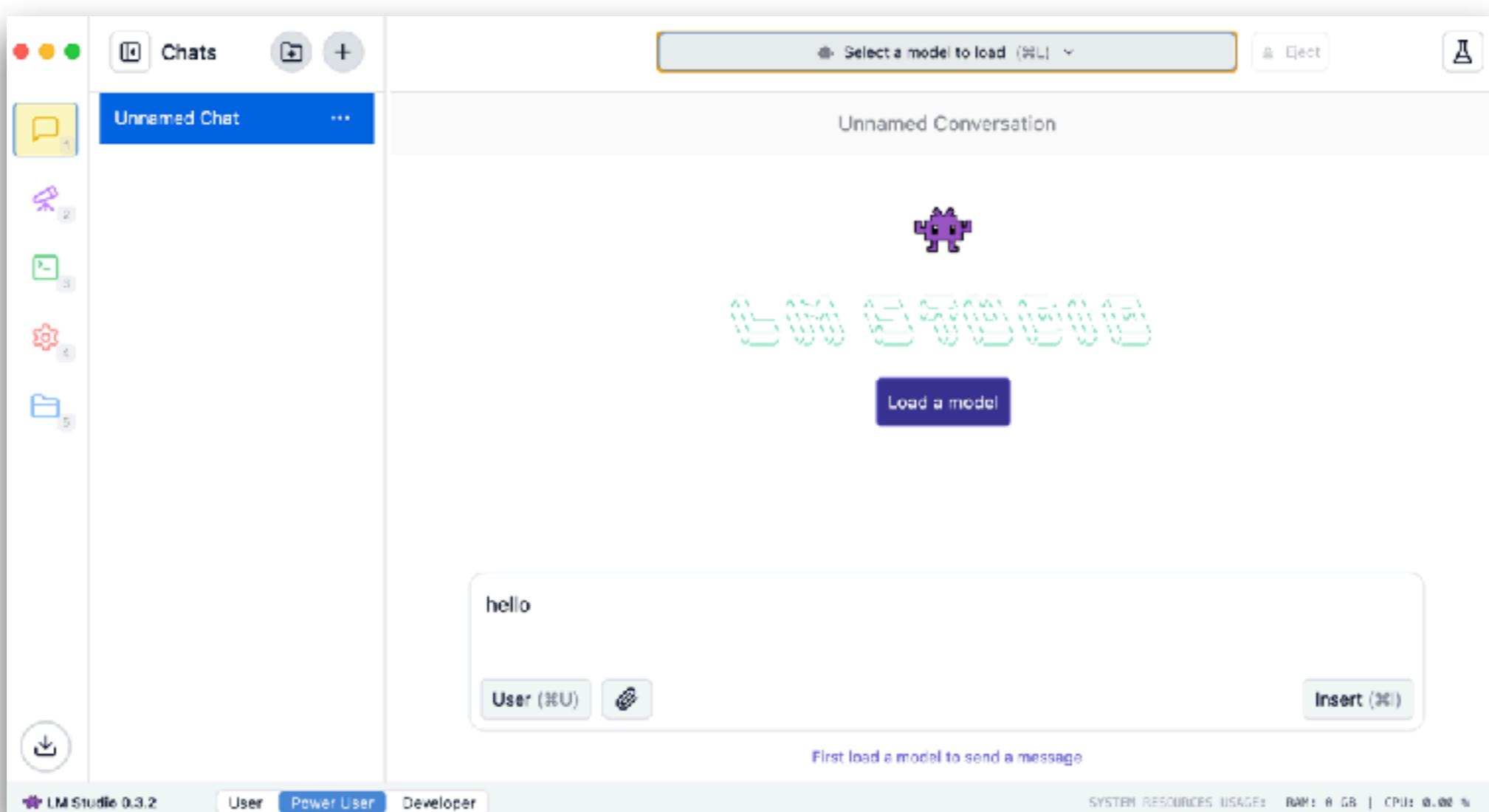
A screenshot of the LM Studio application window titled "LM Studio - Community Edition 0.3.0". The window displays a file tree on the left and a configuration panel on the right. The configuration panel includes sections for "Advanced Configuration" (with fields for "System Prompt" and "Model"), "P Serial" (set to 1), "C Sensors" (set to 1), "IP AutomationAddress" (set to 1), and "IP AutomationPort" (set to 5). A cursor is visible over the "IP AutomationAddress" field.

<https://lmstudio.ai/>



# Local LLM with LM Studio

Load model from Hugging Face



<https://lmstudio.ai/>



# Local LLM with LlamaEdge



LlamaEdge

Feature FAQ Models Docs | [View on GitHub](#)

**The easiest, smallest and fastest local LLM runtime and API server.**

[Quick Start with Gaia](#)

Powered by Rust & WasmEdge (A CNCF hosted project) ⓘ

<https://llamaedge.com/>



# Local LLM with LocalAI



<https://localai.io/>



# More

GPT4All

LlamaFile

Jan.ai

NextChat

Anything LLM

<https://github.com/Hannibal046/Awesome-LLM>



# LLM Models



# Hugging Face Model Hub

NEW! AI Tools are now available in HuggingChat

The AI community building the future.

The platform where the machine learning community collaborates on models, datasets, and applications.

Models 450,541

Tasks

- Text-to-image
- Image-to-Text
- Text-to-video
- Visual Question Answering
- Document Question Answering
- Graph Machine Learning

Computer Vision

- Object Detection
- Image Classification
- Object Detection
- Image Segmentation
- Image-to-Image
- Unconditional Image Generation
- Video Classification
- Zero-Shot Image Classification

Natural Language Processing

- Text Classification
- Text Generation
- Code Generation
- Question Answering
- Zero-Shot Classification
- Translation
- Summarization
- Conversational
- Text Generation
- Text-to-Text Generation

Audio

- Text-to-Speech
- Automatic Speech Recognition
- Audio-to-Audio
- Audio Classification
- Music Activity Detection

Tabular

- Tabular Classification
- Tabular Regression

Reinforcement Learning

- Reinforcement Learning
- Robitics

Model

- meta-llama/Llama-2-7b
- stabilityai/stable-diffusion-v1-base
- openai/clip-vit-large-patch14
- lllyasviel/ControlNet-v1-2
- ceresense/zeroscope\_v2\_XL
- meta-llama/Llama-2-20b
- tiiuae/falcon-40b-instruct
- MirroredH/MixedCodes-v1B-v1.0
- CompVis/stable-diffusion-v2-4
- StabilityAI/stable-diffusion-v2-2
- Salesforce/qwen-2b-8w-inat

<https://huggingface.co/>



# Hugging Face :: Model

The screenshot shows the Hugging Face Model Hub interface. On the left, there's a sidebar with sections for Tasks (Libraries, Datasets, Languages, Licenses), Other, Filter Tasks by name, Multimodal (Image-Text-to-Text, Visual Question Answering, Document Question Answering, Video-Text-to-Text, Any-to-Any), and Computer Vision (Depth Estimation, Image Classification, Object Detection, Image Segmentation, Text-to-Image, Image-to-Text, Image-to-Image, Image-to-Video, Unconditional Image Generation, Video Classification, Text-to-Video, Zero-Shot Image Classification, Mask Generation, Zero-Shot Object Detection, Text-to-3D). The main area is titled 'Models 233,861' and has a search bar with 'llama'. It includes 'Full-text search' and 'Sort: Trending' buttons. Below the search bar, the results are listed:

- black-forest-labs/FLUX.1-dev**  
Text-to-Image • Updated Aug 16 • 919k • 4.73k
- meta-llama/Meta-Llama-3.1-BB-Instruct**  
Text Generation • Updated Aug 21 • 3.09M • 2.61k
- jinaai/reader-lm-1.5b**  
Text Generation • Updated 5 days ago • 8.28k • 382
- black-forest-labs/FLUX.1-schnell**  
Text-to-Image • Updated Aug 16 • 1.06M • 2.35k
- nvidia/Llama-3\_1-Nemotron-51B-Instruct**  
Text Generation • Updated about 14 hours ago • 61 • 79
- dleemiller/word-llama-12-supercat**  
Updated Aug 12 • 81
- ICTNLP/Llama-3.1-8B-Omni**  
Updated 12 days ago • 1.39k • 324

<https://huggingface.co/>



# Big Code model leader board

★ Big Code Models Leaderboard

Inspired from the [Open LLM Leaderboard](#) and [Open LLM-Perf Leaderboard](#), we compare performance of base multilingual code generation models on [HumanEval](#) benchmark and [MultiPL-E](#). We also measure throughput and provide information about the models. We only compare open pre-trained multilingual code models, that people can start from as base models for their trainings.

Evaluation table Performance Plot About Submit results 🚀

See All Columns

Search for your model and press ENTER...

Filter model types

all  base  instruction-tuned  EXT external-evaluation

T	Model	Win Rate	humaneval-python	java	javascript	cpp
♦ EXT	<a href="#">OpenCodeInterpreter-DS-33B</a>	55.83	75.23	54.8	69.06	64.47
♦ EXT	<a href="#">Nxcode-CQ-7B-croc</a>	55.42	87.23	60.91	71.69	68.04
♦	<a href="#">CodeQwen1.5-7B-Chat</a>	55.08	87.2	61.04	70.31	67.85
♦ EXT	<a href="#">CodeFuse-DeepSeek-33b</a>	54.33	76.83	60.76	66.46	65.22
♦ EXT	<a href="#">DeepSeek-Coder-33b-instruct</a>	52	80.02	52.03	65.13	62.36
♦ EXT	<a href="#">Artigenz-Coder-DS-6.7B</a>	51.5	70.89	56.84	66.16	59.75
♦ EXT	<a href="#">DeepSeek-Coder-7b-instruct</a>	50.33	86.22	53.34	65.8	59.66
♦ EXT	<a href="#">OpenCodeInterpreter-DS-6.7B</a>	49.67	73.2	51.41	63.85	60.81

<https://huggingface.co/spaces/bigcode/bigcode-models-leaderboard>



# Model in Ollama

The screenshot shows the Ollama library interface. At the top left is a circular icon of a cartoon llama head. To its right, the word "Models" is displayed. Below this is a search bar containing the text "deepseek". To the right of the search bar is a dropdown menu set to "Featured".

The first model listed is "deepseek-coder-v2". Its description reads: "An open-source Mixture-of-Experts code language model that achieves performance comparable to GPT4-Turbo in code-specific tasks." Below the description are three blue buttons labeled "Code", "16B", and "236B". Underneath these buttons are three small icons with the numbers "307K", "65", and "3 months ago" respectively.

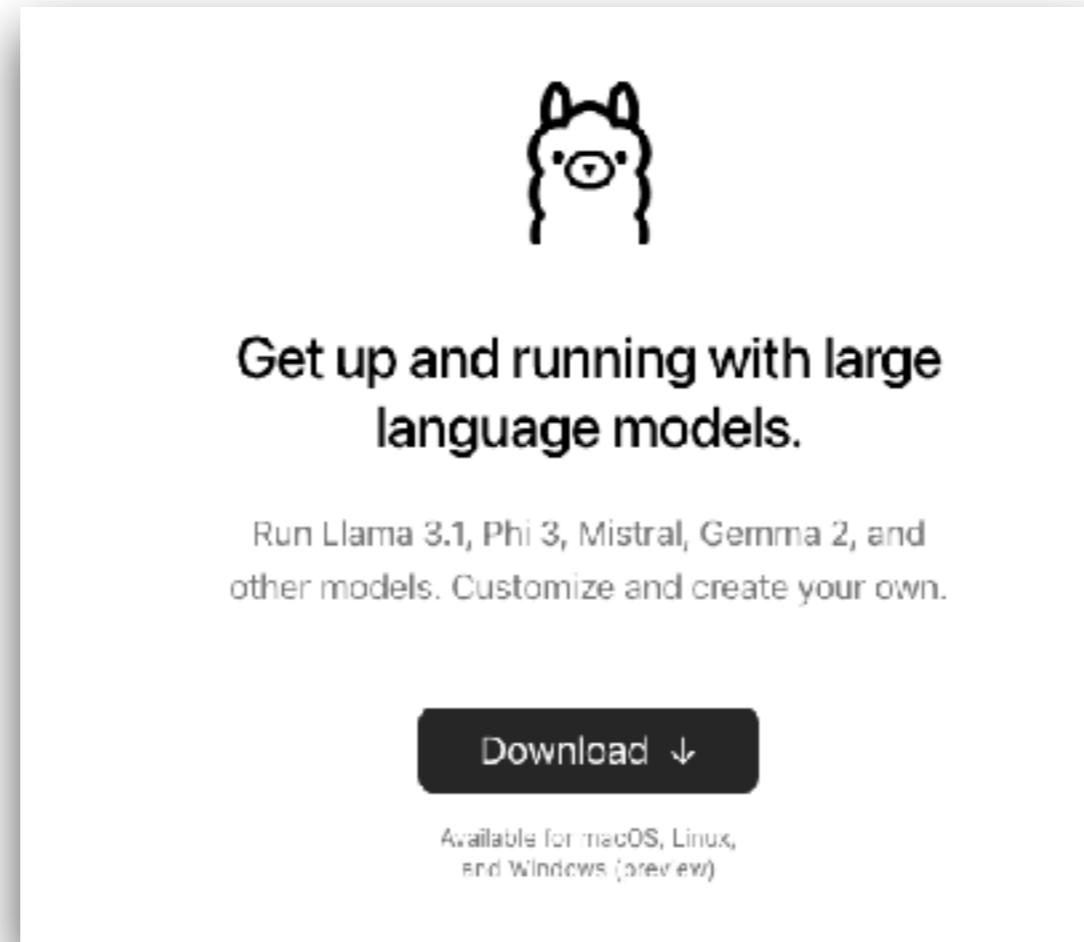
The second model listed is "deepseek-coder". Its description reads: "DeepSeek Coder is a capable coding model trained on two trillion code and natural language tokens." Below the description are three blue buttons labeled "Code", "1B", "7B", and "33B". Underneath these buttons are three small icons with the numbers "303.9K", "102", and "9 months ago" respectively.

<https://ollama.com/library>



# Workshop with Ollama

\$ollama run **llama3.1**



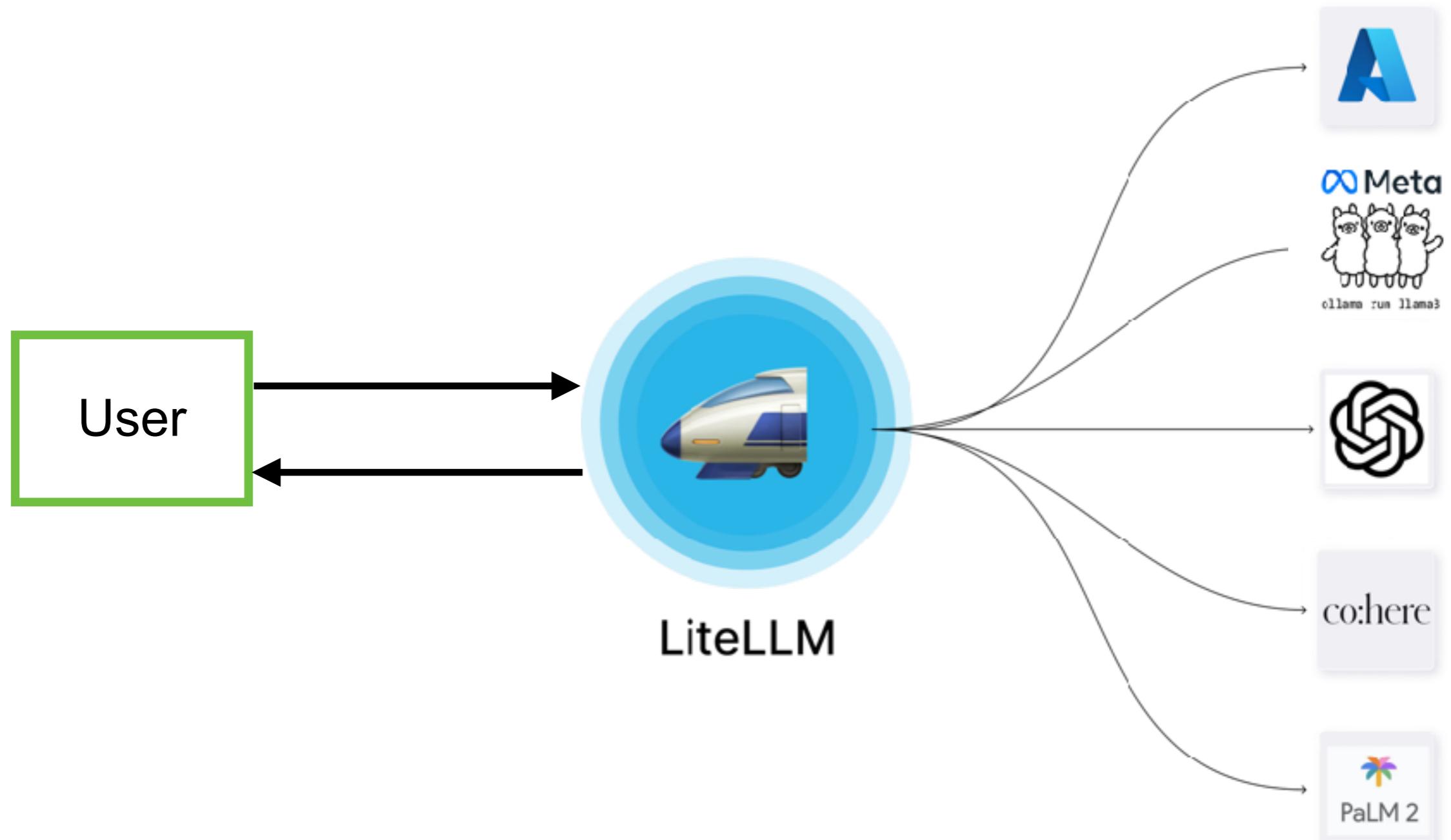
<https://github.com/up1/workshop-ai-with-technical-team/wiki/Local-LLM-with-Ollama>



# **LiteLLM as a Proxy**



# LiteLLM as a Proxy

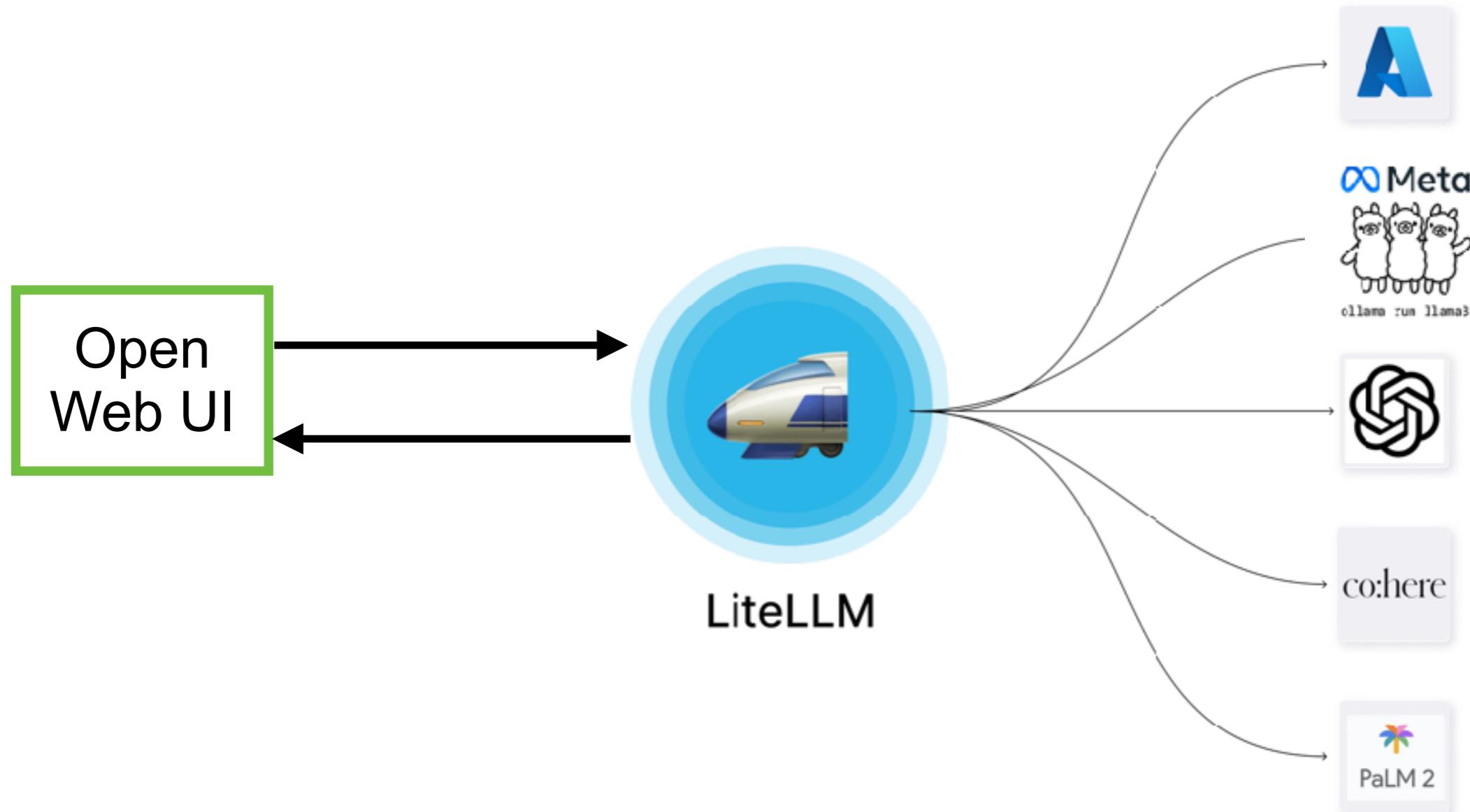


<https://www.litellm.ai/>



# Workshop

Use docker compose to build and run



<https://github.com/up1/workshop-ai-with-technical-team/wiki/LiteLLM-and-WebUI>

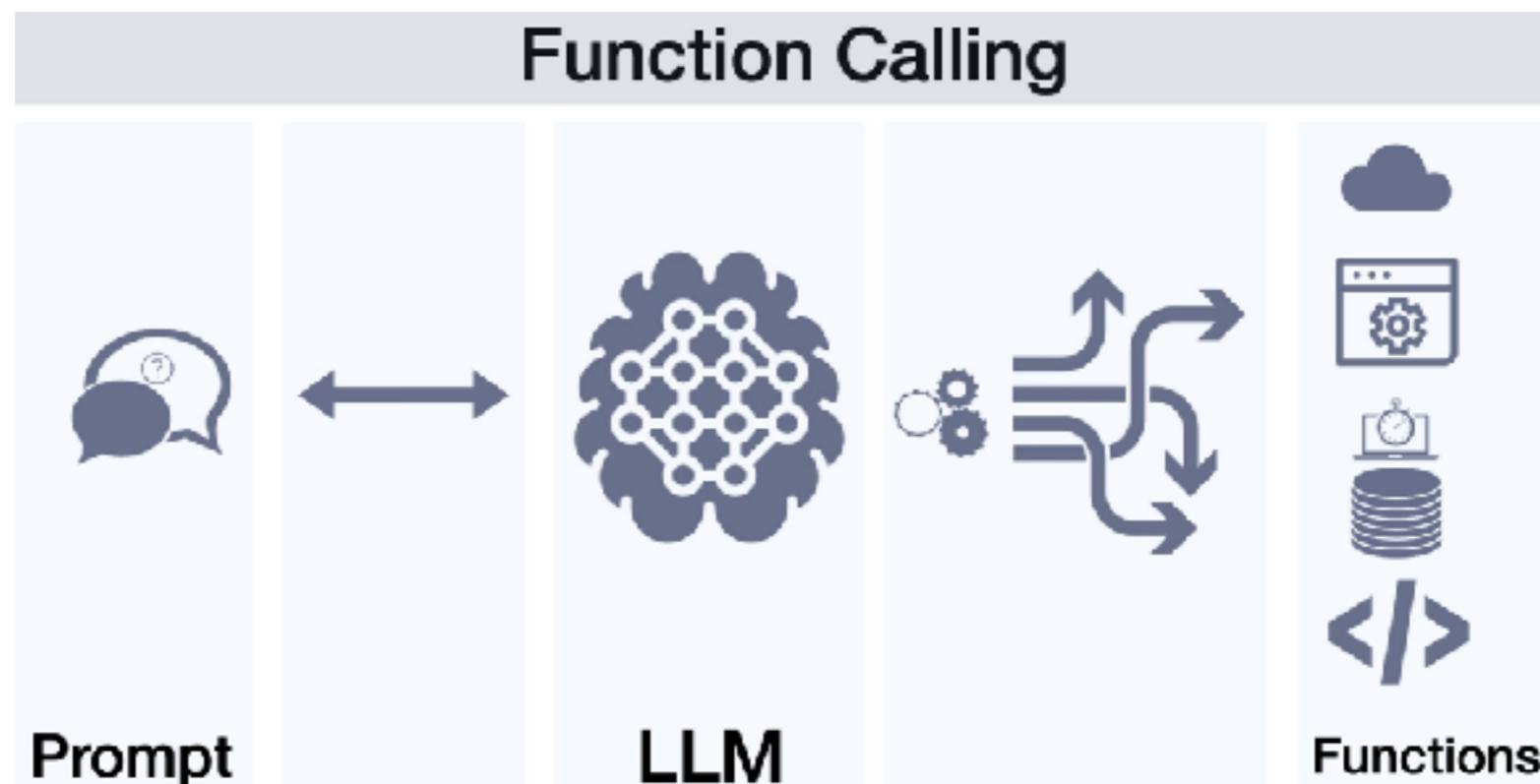


# Function Calling ?

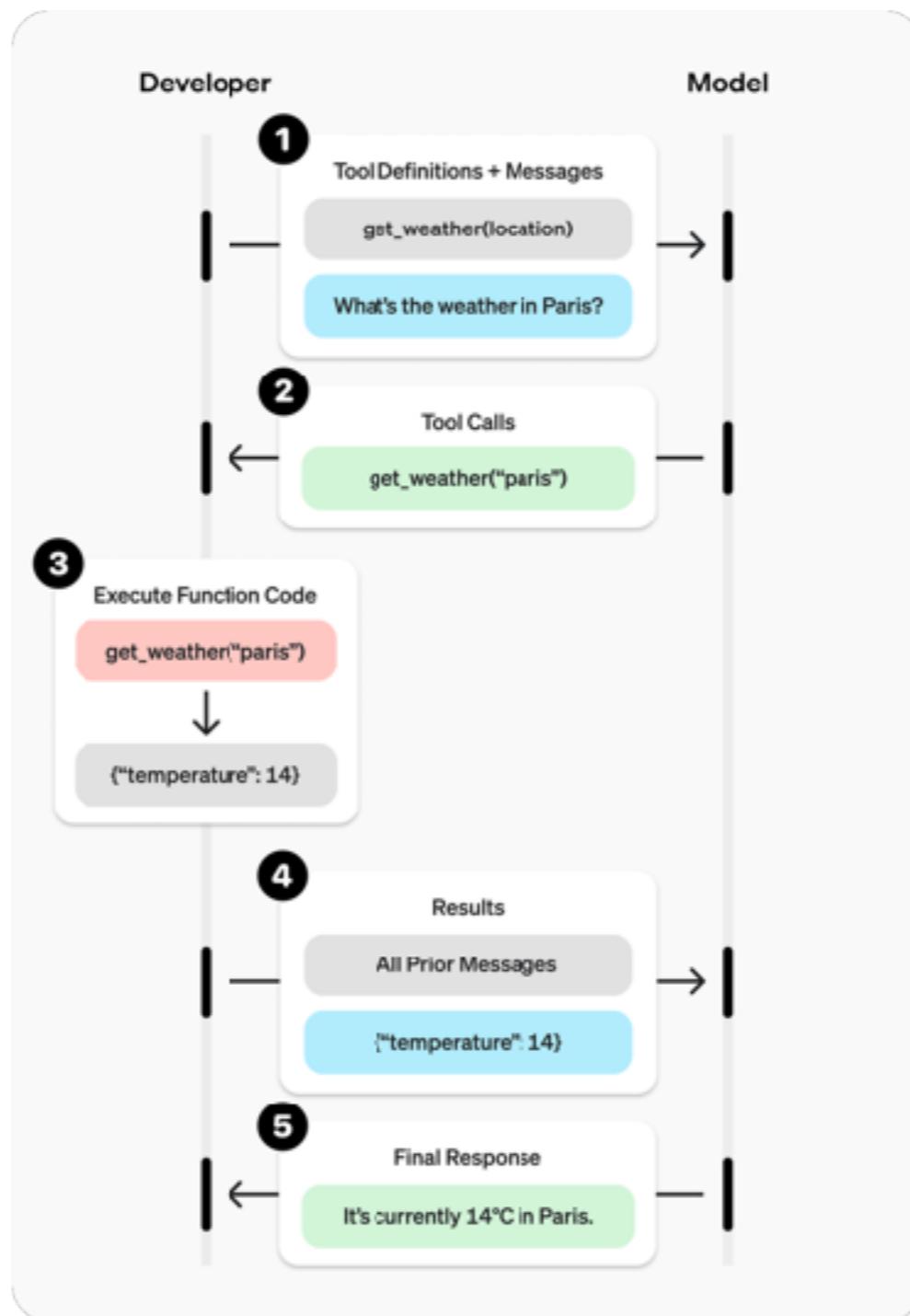


# Function Calling ?

Allow LLM to recognize what tool it need based on user's input and when to invoke it.



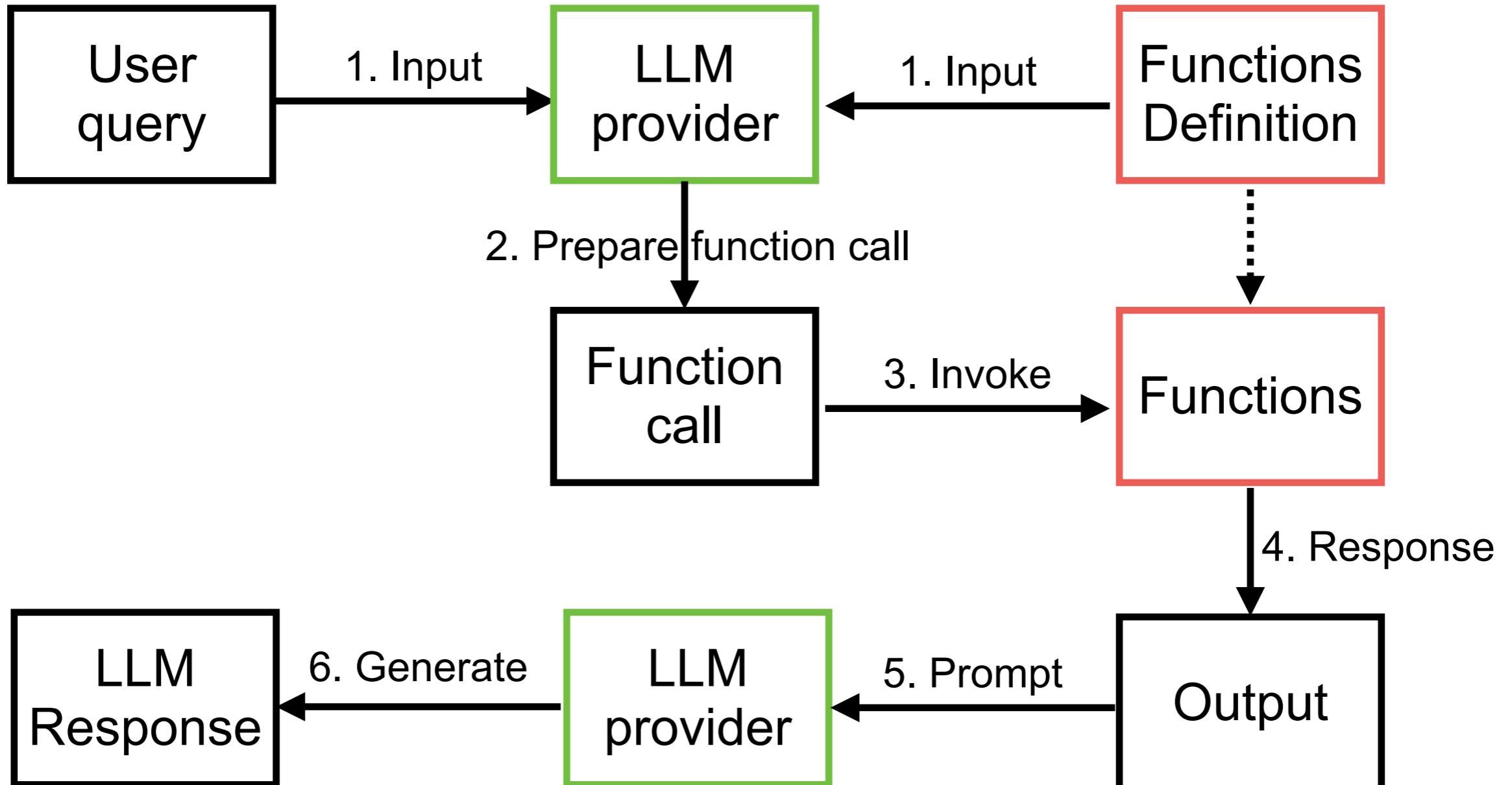
# Function calling ?



<https://platform.openai.com/docs/guides/function-calling?api-mode=responses>



# Function Calling



# Support function calling

Provider name	Model name
OpenAI	GPT 4
Anthropic	Claude 3 (Sonnet, Haiku, Opus)
Google	Gemini



# Function calling !!

**APIs: Every tool needs its own key**

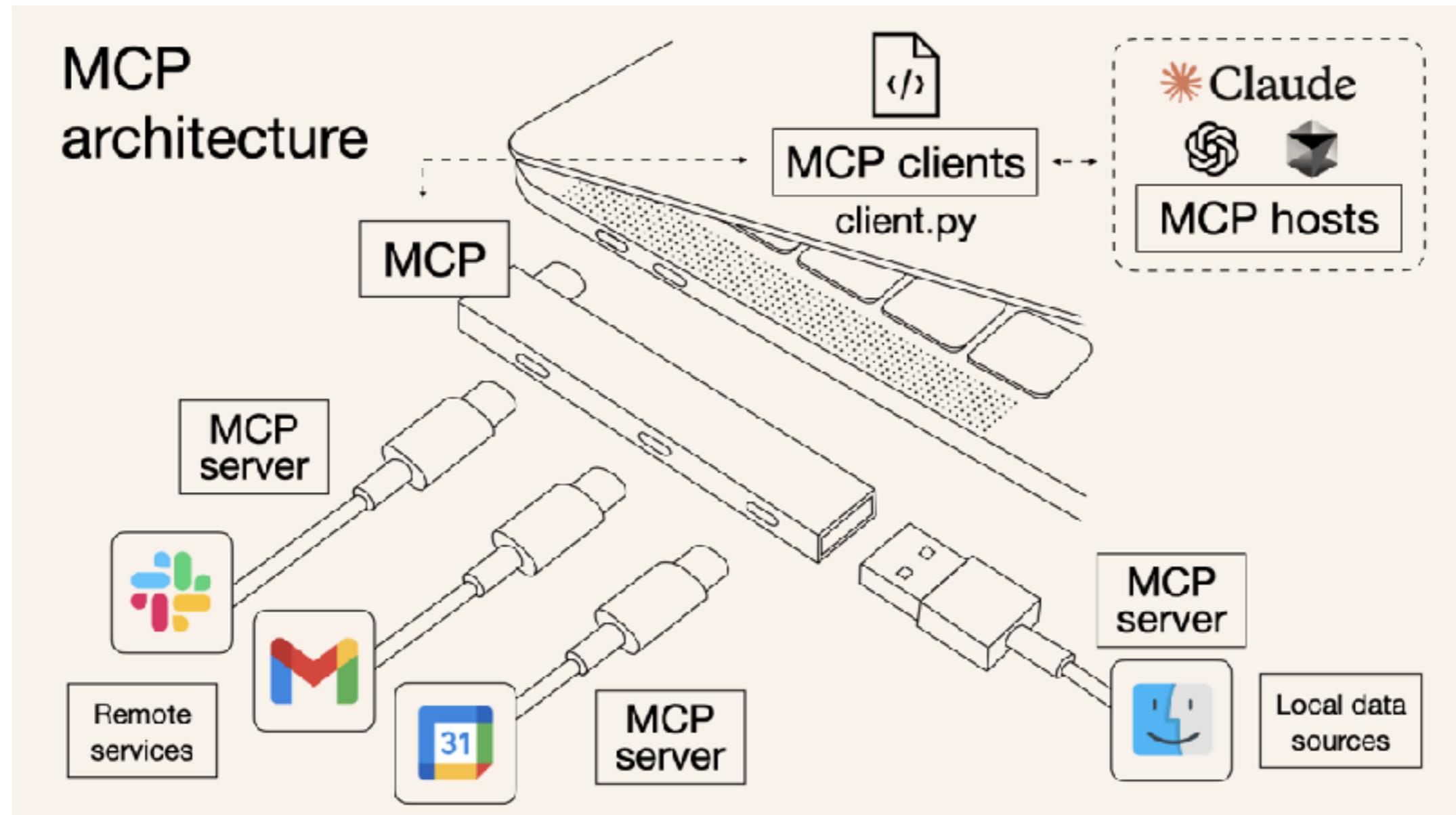
Traditional APIs require different authentication and integration for each service,  
like needing different keys for different locks

The diagram consists of two rows of seven items each. The top row contains seven different types of locks: a standard padlock, a combination lock, a cylinder lock, a circular lock, another cylinder lock, a pin tumbler lock, and a complex ornate lock. The bottom row contains seven keys, each designed to fit one of the locks above it. Below the keys, there is a row of eight icons, each representing a different API or service: 'APIs' (represented by a box with the text), Google Sheets (document icon), Google Calendar (calendar icon), Google Drive (cloud storage icon), Google Maps (map icon), Google Photos (camera icon), Google Slides (slide icon), and WhatsApp (phone icon).

<https://norahsakal.com/blog/mcp-vs-api-model-context-protocol-explained/>



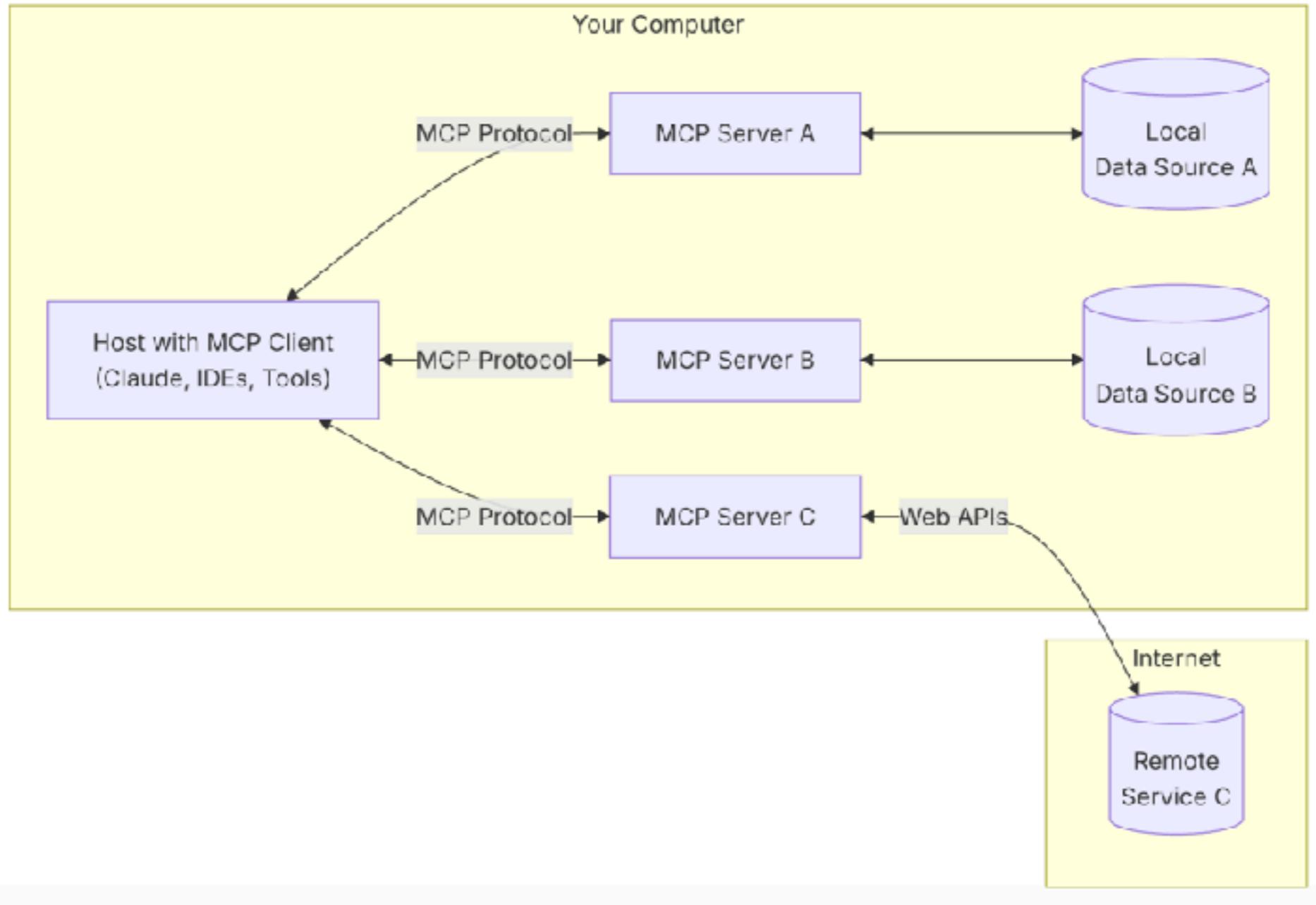
# Standardized function calling !!



<https://norahsakal.com/blog/mcp-vs-api-model-context-protocol-explained/>



# Model Context Protocol



<https://modelcontextprotocol.io/>



# Awesome MCP

## Find Awesome MCP Servers and Clients

MCP.so is a third-party MCP Marketplace with **17495** MCP Servers collected.

[ShipAny](#): NextJS boilerplate for building AI SaaS startups.

Search with keywords



Today

Featured

Latest

Clients

Hosted

Official

### Featured MCP Servers

[View All →](#)

 EdgeOne Pages MCP ★ An MCP service designed for deploying HTML content to...	 AlphaVantage ★ Bring enterprise-grade stock market data to agents and LLMs	 Time ★ A Model Context Protocol server that provides time and timezone...	 Zhipu Web Search ★ Zhipu Web Search MCP Server is a search engine specifically...
 MCP Advisor ★ MCP Advisor & Installation – Use the right MCP server for your...	 HowtoCook Mcp ★ 基于Anduin2017 / HowToCook 《程序员在家做饭指南》的mcp server, ...	 Minimax MCP ★ Official Minimax Model Context Protocol (MCP) server that enables...	 Serper MCP Server ★ A Serper MCP Server
 Jina AI MCP Tools ★ A Model Context Protocol (MCP) server that integrates with Jina AI...	 Amap Maps ★ 高德地图官方 MCP Server	 Playwright Mcp ★ Playwright MCP server	 Baidu Map ★ 百度地图核心API现已全面兼容MCP协议，是国内首家兼容MCP协议的地图...

<https://mcp.so/>



AI for Software Development

© 2020 - 2026 Siam Chamnankit Company Limited. All rights reserved.

# **Retrieval-Augmented Generation (RAG)**



# RAG

Enhances LLMs by retrieving external knowledge before generating response

Improve accuracy

Reduce hallucinations

Real-time knowledge updates



# Core Components

## Retriever

Fetch relevant documents from a knowledge base

## Generator

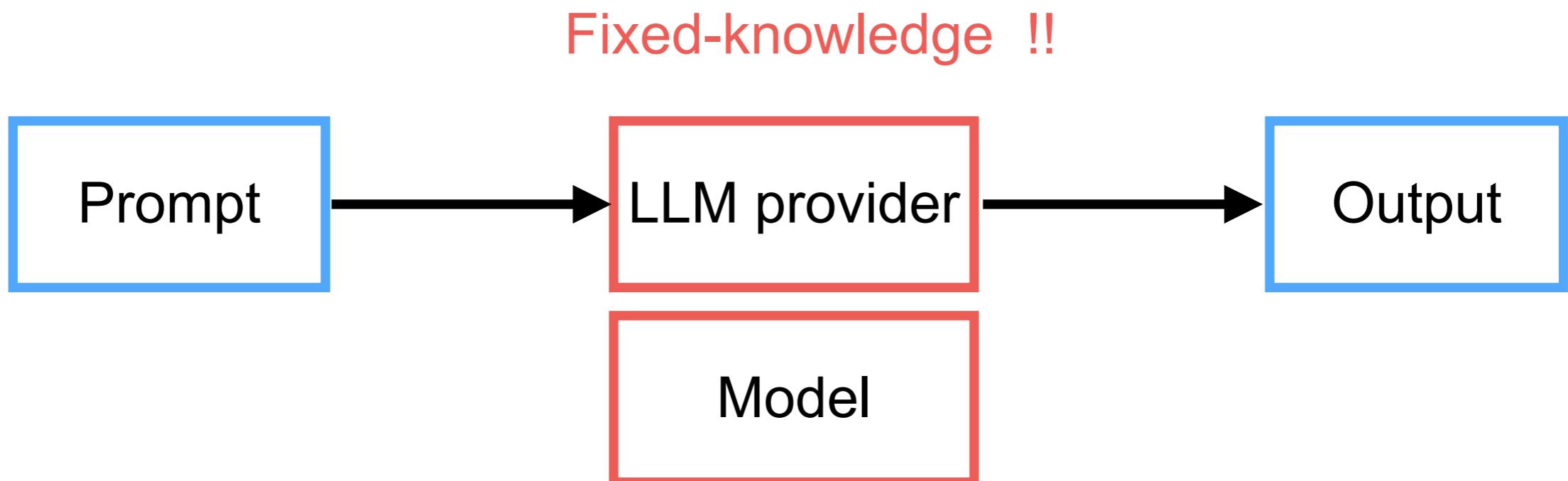
Uses the retrieved information to generate a response

## Enhancements

Different RAG architectures modify how retrieval and generation interact to improve performance

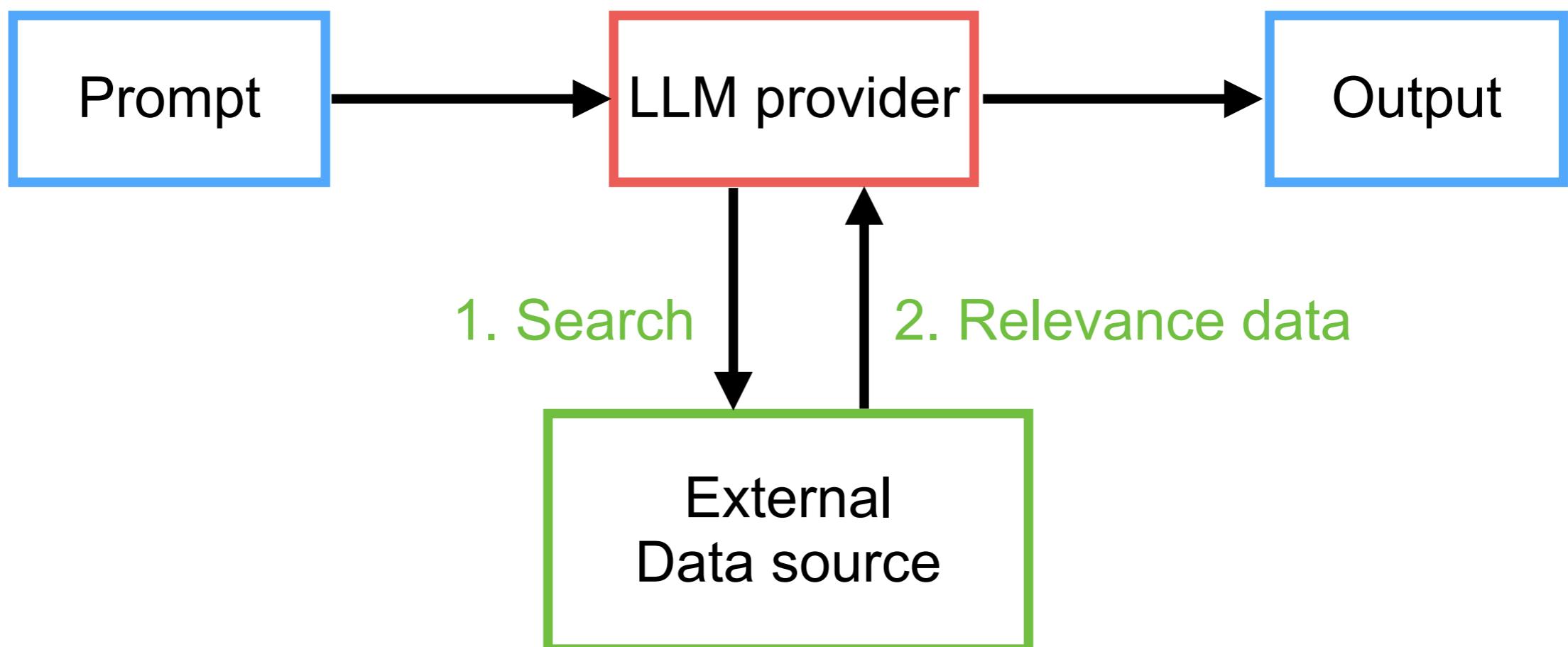


# Limitation of LLM



# RAG ?

Fixed-knowledge !!



# How to search/retrieve data ?

Data source

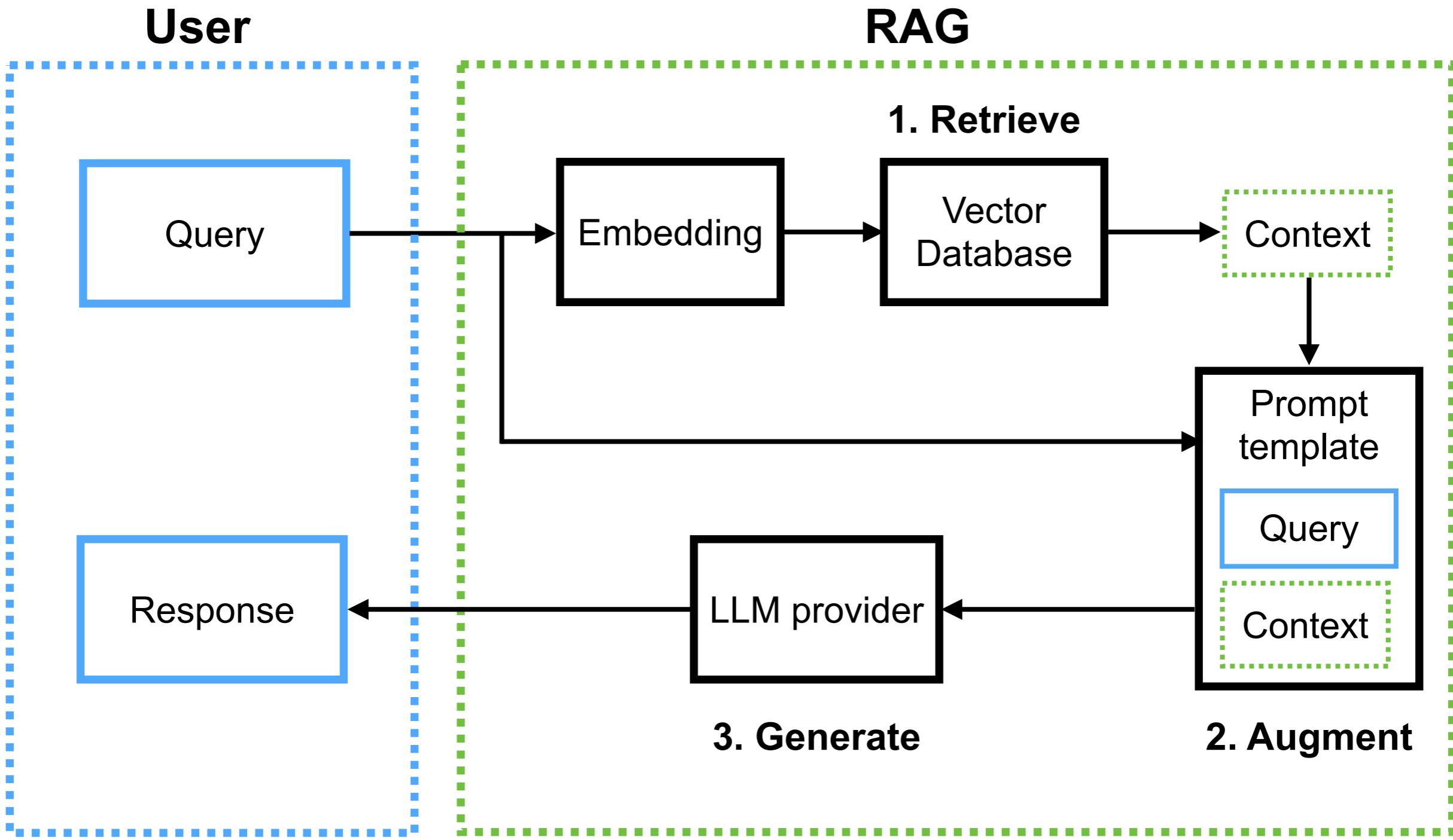
Data preparation

Search

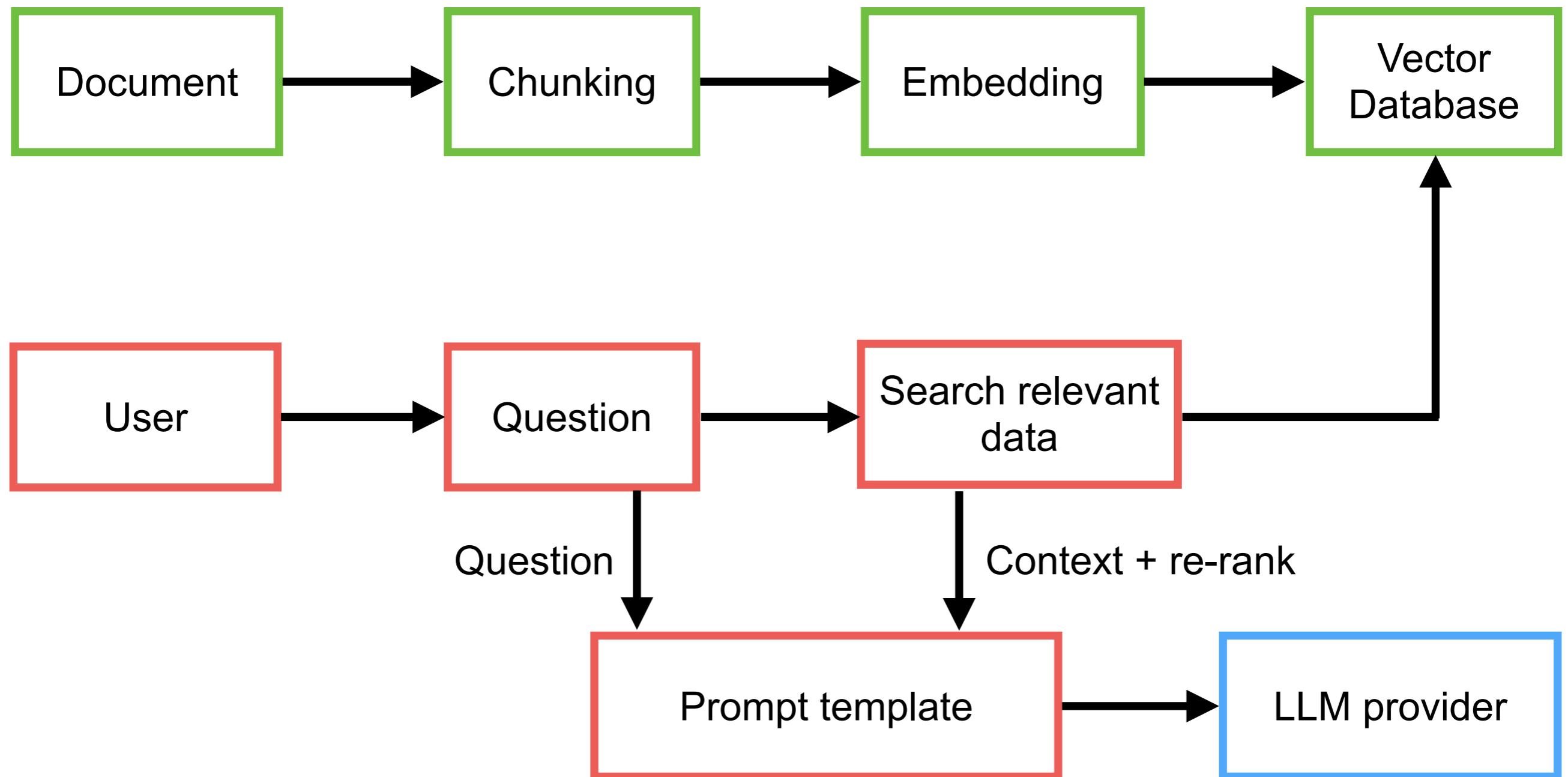
Search result



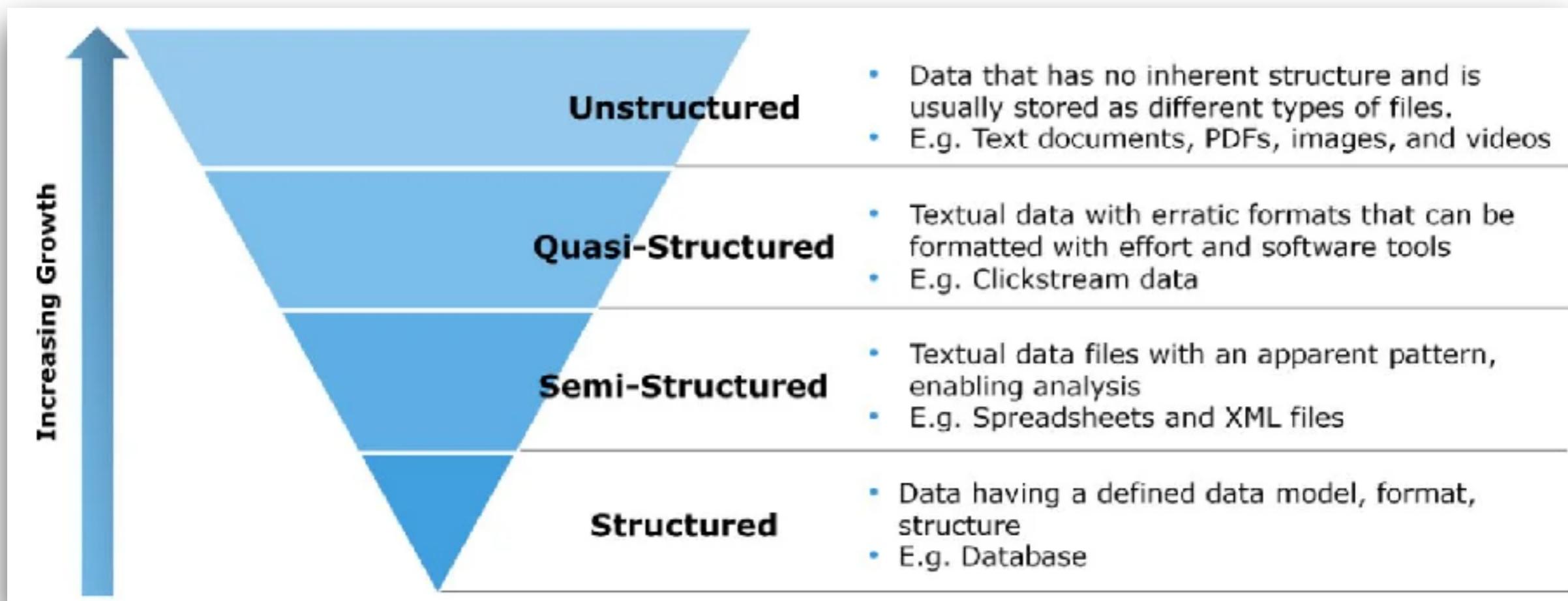
# Basic RAG Architecture



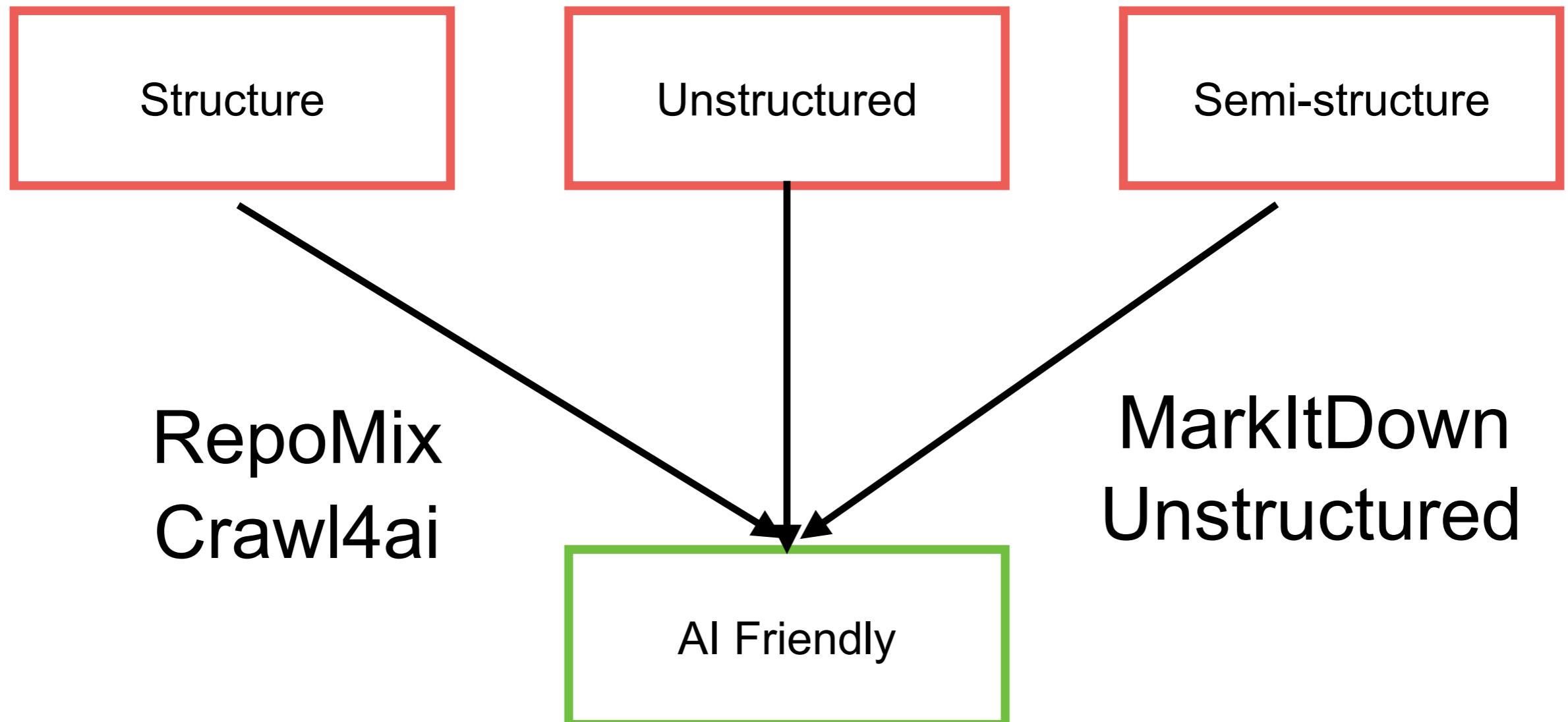
# RAG process



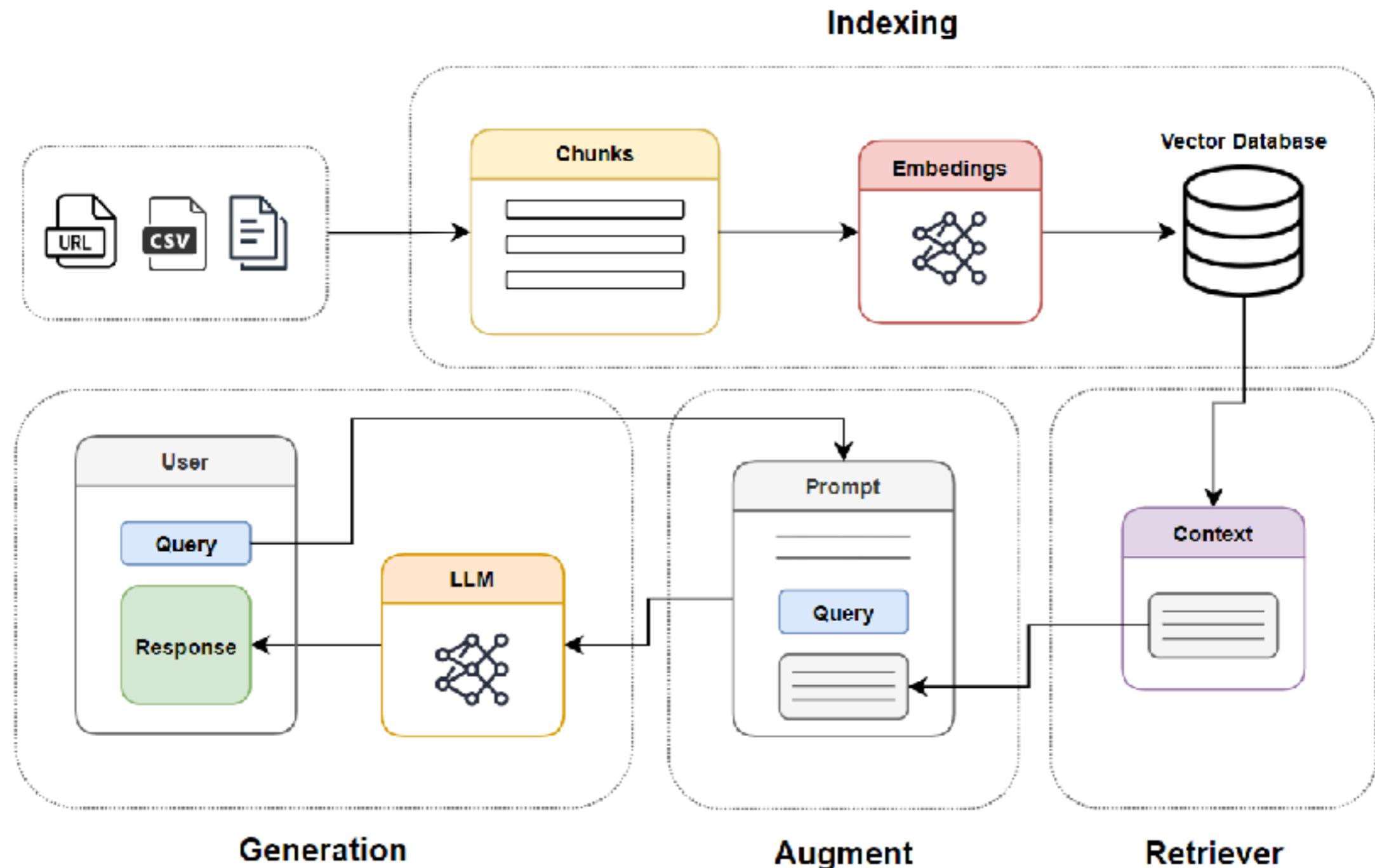
# Structures of Data ?



# Friendly Data for LLM/AI



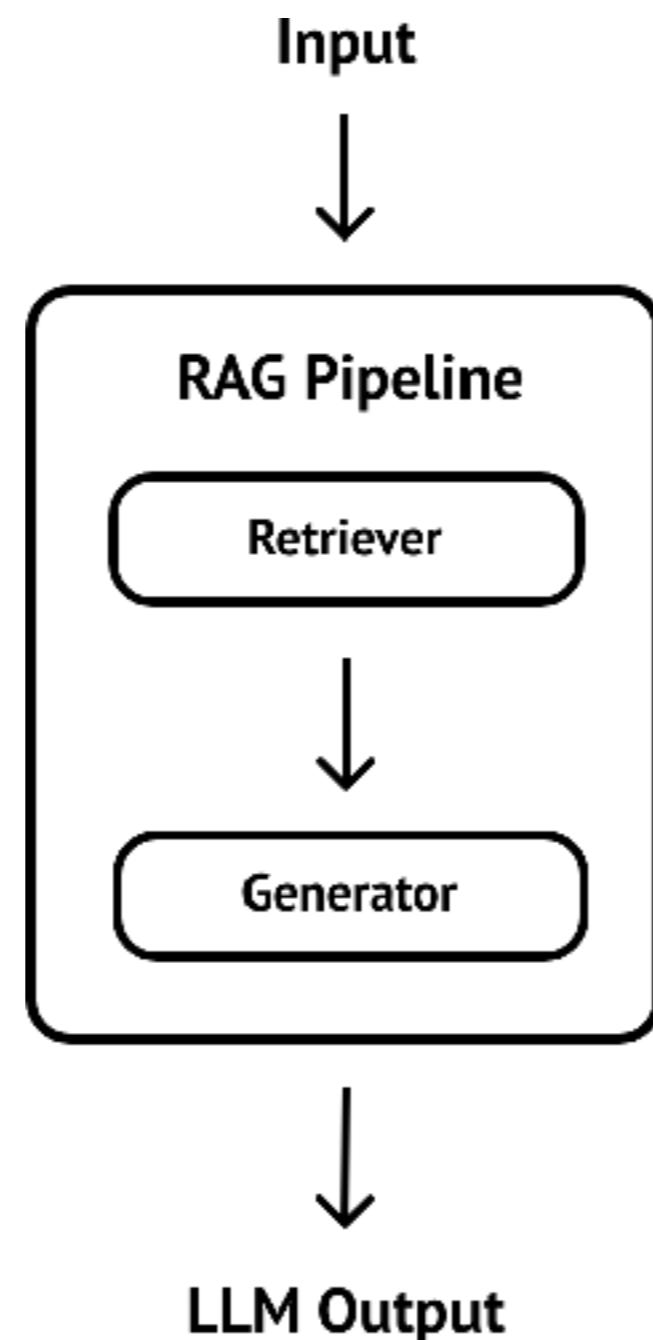
# RAG Cookbook (techniques)



<https://github.com/athina-ai/rag-cookbooks>



# RAG Evaluation !!



#### Retriever Metrics:

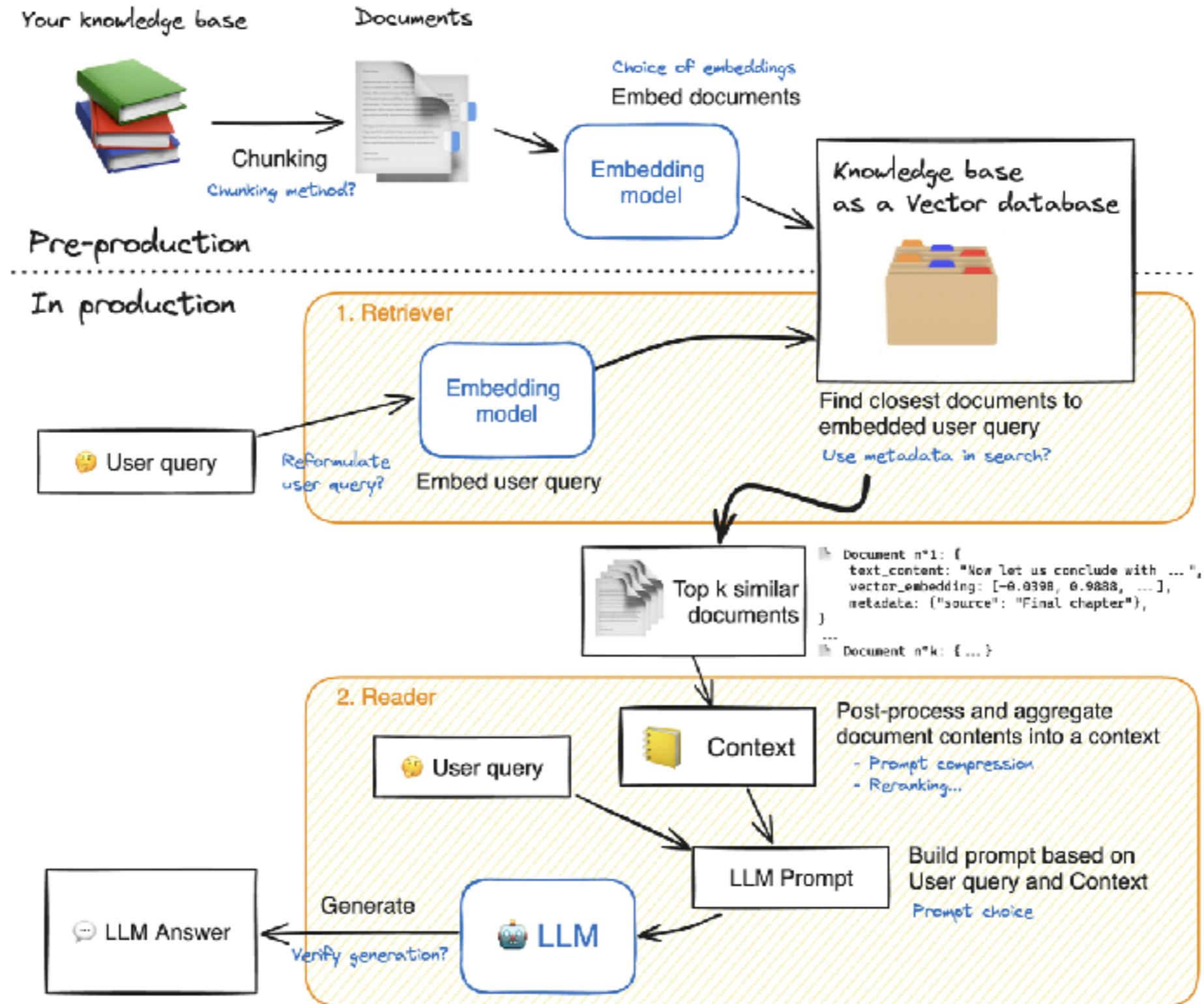
- Contextual Recall
- Contextual Precision
- Contextual Relevancy

#### Generator Metrics:

- Answer Relevancy
- Faithfulness

<https://www.deepeval.com/guides/guides-rag-evaluation>

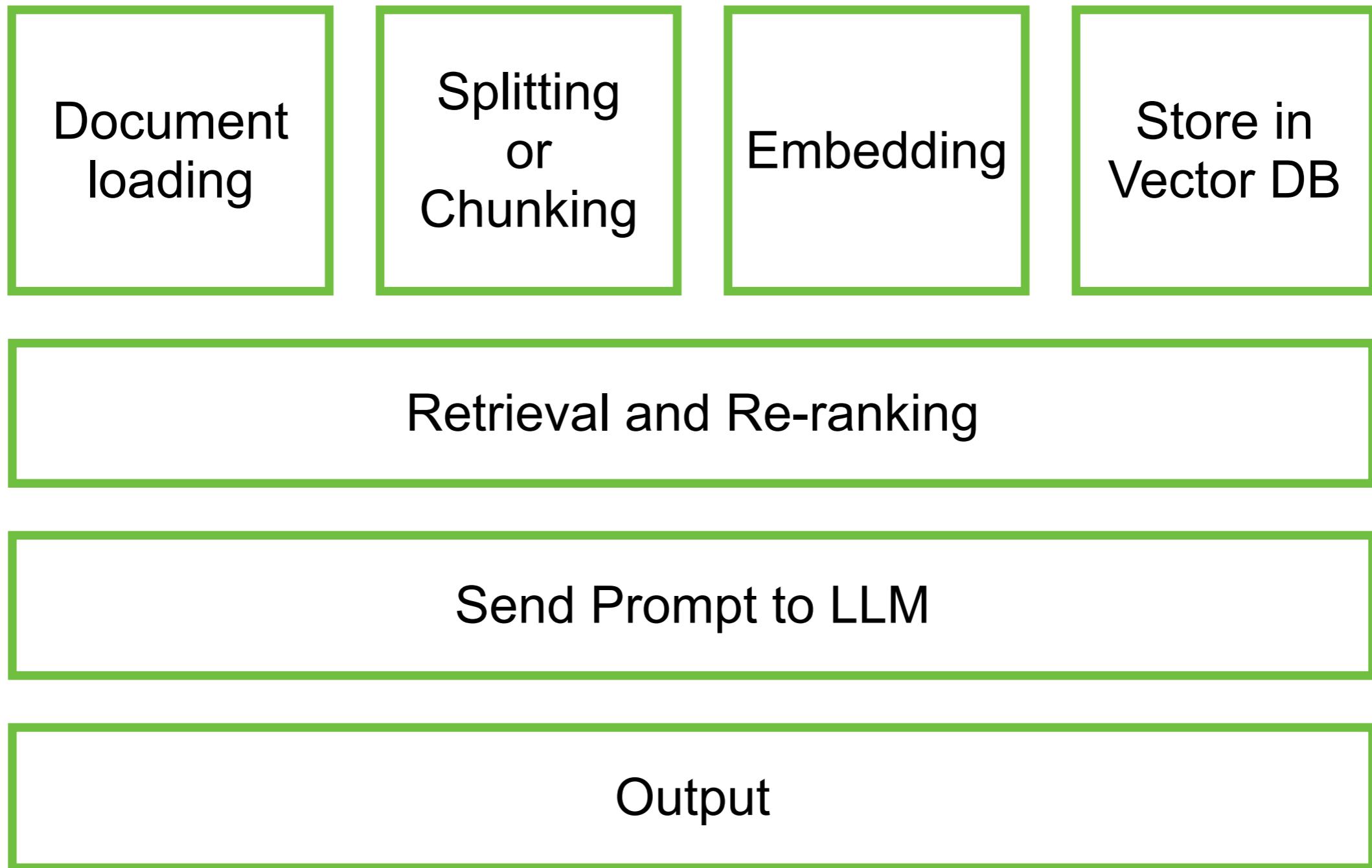




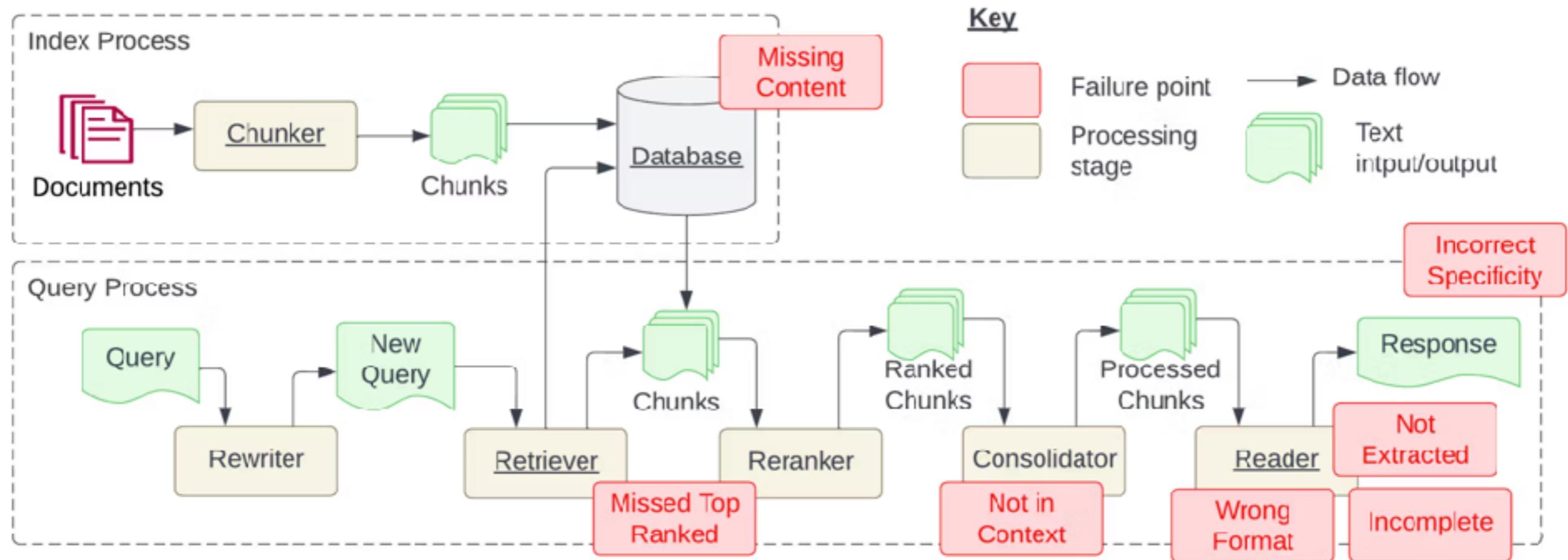
[https://huggingface.co/learn/cookbook/en/rag\\_evaluation](https://huggingface.co/learn/cookbook/en/rag_evaluation)



# RAG Implementation



# Failure Points of RAG



**Figure 1: Indexing and Query processes required for creating a Retrieval Augmented Generation (RAG) system. The indexing process is typically done at development time and queries at runtime. Failure points identified in this study are shown in red boxes. All required stages are underlined. Figure expanded from [19].**

<https://www.galileo.ai/blog/mastering-rag-how-to-architect-an-enterprise-rag-system>



# RAG is better

**But come with Cost !!**

More latency

Retrieval errors

Accuracy !!

More  
Complexity

Maintenance  
overhead



# RAG Techniques ?

Semantic  
chunking

Chunk size  
selector

Context chunk  
header

Adaptive RAG

Re-ranking

Graph TAG

<https://github.com/FareedKhan-dev/all-rag-techniques>



# Pre-Retrieval optimization ?

Preprocessing and cleansing data

Chunking strategies

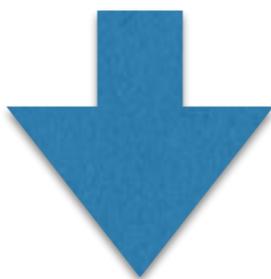
Add metadata

Embedding model selection

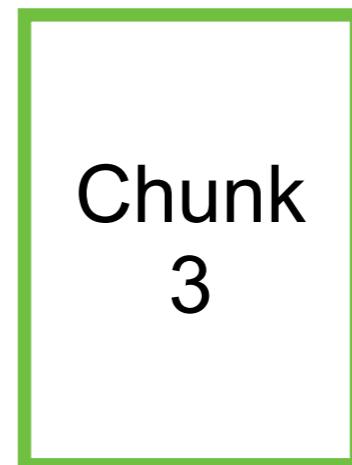
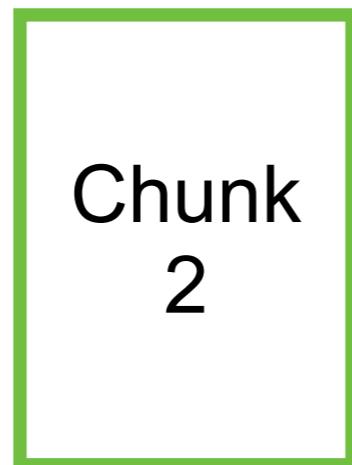
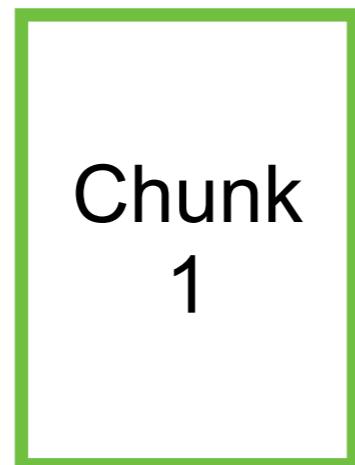


# Chunking strategies

Size of Document  
**> context window size**

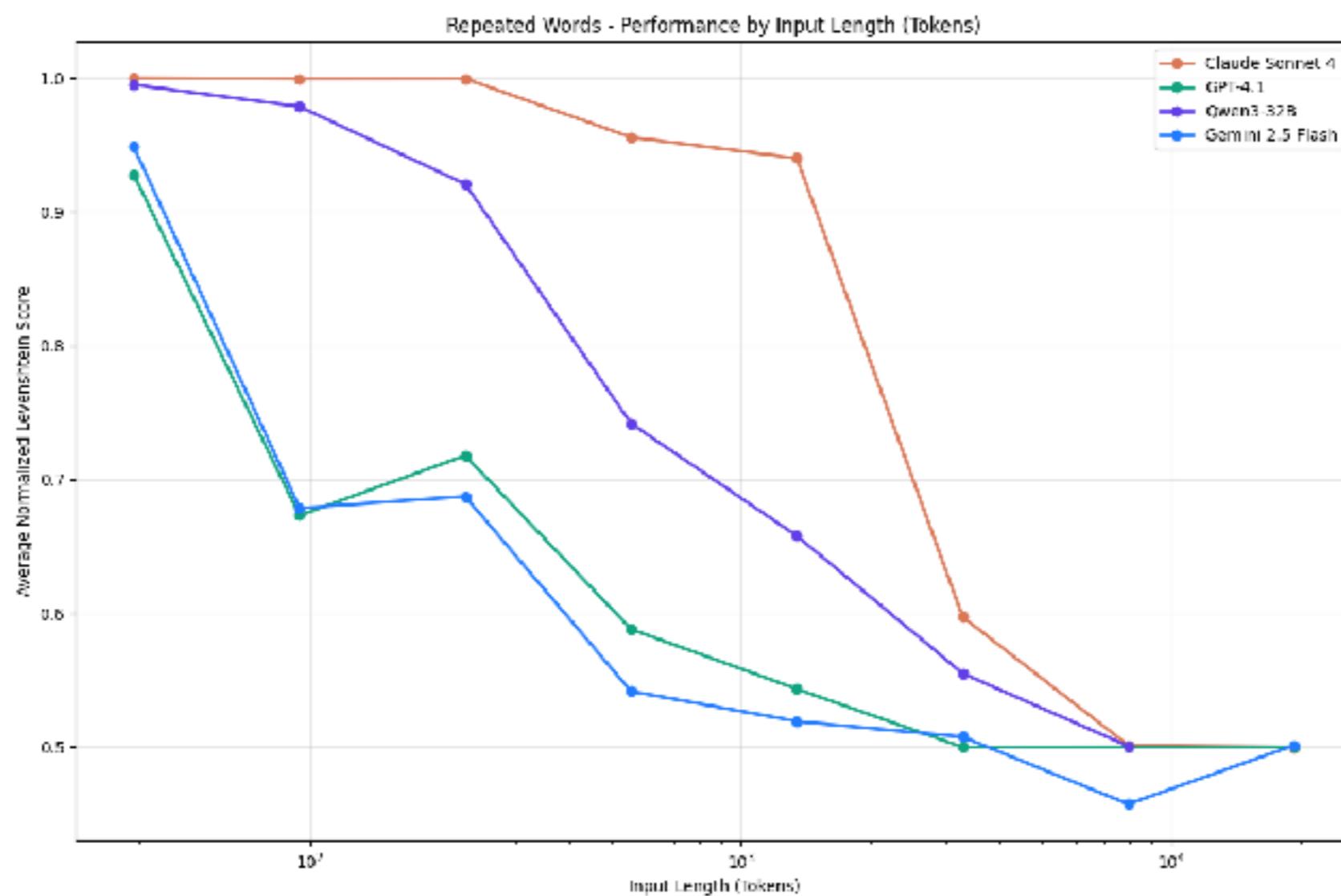


Chunking



# Context Window ?

Number of tokens in the context window increases,  
the model's ability to accurately recall information from that context decreases



<https://www.anthropic.com/engineering/effective-context-engineering-for-ai-agents>



# Chunking Strategies !!

Fixed size  
Recursive characters  
Document structure-based  
Semantic chunking  
Agentic chunking

...

<https://blog.dailydoseofds.com/p/5-chunking-strategies-for-rag>



# Good Chunking

Efficiency

Relevance

Context  
preservation

Improve content  
generation

Reduce noise



# Bad Chunking

Loss context

Redundancy

Inconsistency



# Chunking Visualization

## ChunkViz v0.1

Want to learn more about AI Engineering Patterns? Join me on [Twitter](#) or [Newsletter](#).

Language Models do better when they're focused.

One strategy is to pass a relevant subset (chunk) of your full data. There are many ways to chunk text.

This is an tool to understand different chunking/splitting strategies.

[Explain like I'm 5...](#)

One of the most important things I didn't understand about the world when I was a child is the degree to which the returns for performance are superlinear.

Teachers and coaches implicitly told us the returns were linear. "You get out," I heard a thousand times, "what you put in." They meant well, but this is rarely true. If your product is only half as good as your competitor's, you don't get half as many customers. You get no customers, and you go out of business.

It's obviously true that the returns for performance are superlinear in business. Some think this is a flaw of capitalism, and that if we changed the rules it would stop being true. But superlinear returns for performance are a feature of the world, not an artifact of rules we've invented. We see the same pattern in fame, power, military victories, knowledge, and ...

[Upload txt](#)

Splitter:

Chunk Size:

Chunk Overlap:

Total Characters: 2658  
Number of chunks: 107  
Average chunk size: 24.8

One of the most important things I didn't understand about the world when I was a child is the degree to which the returns for performance are superlinear.

Teachers and coaches implicitly told us the returns were linear. "You get out," I heard a thousand times, "what you put in." They meant well, but this is rarely true. If your product is only half as good as your competitor's, you don't get half as many customers. You get no customers, and you go out of business.

<https://chunkviz.up.railway.app/>



# Retrieval optimization ?

Re-ranking

Hybrid search

Query  
transformation

Multi-vector  
embedding

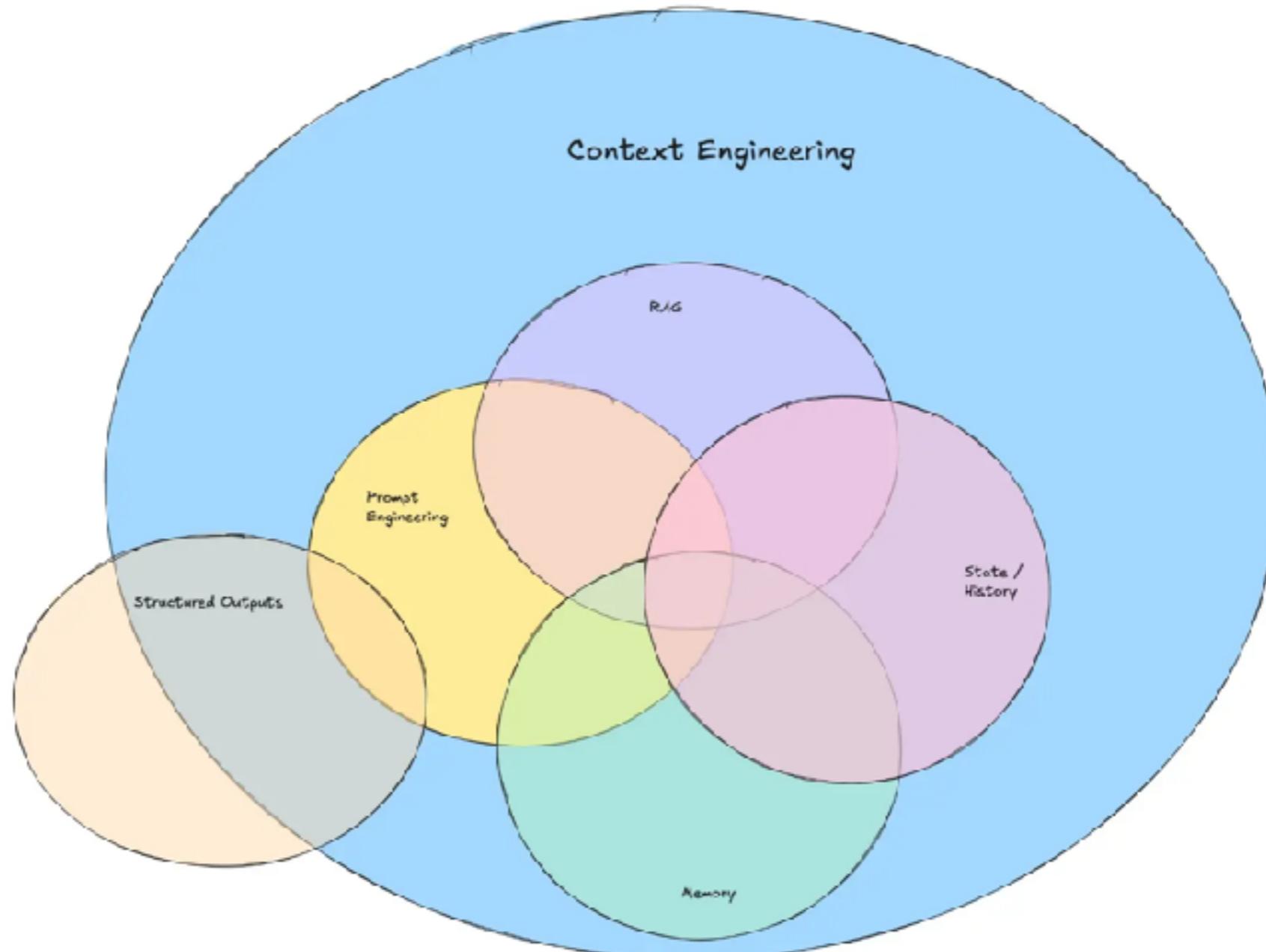
Contextual  
retrieval

Vector database  
Selection





# Context Engineering



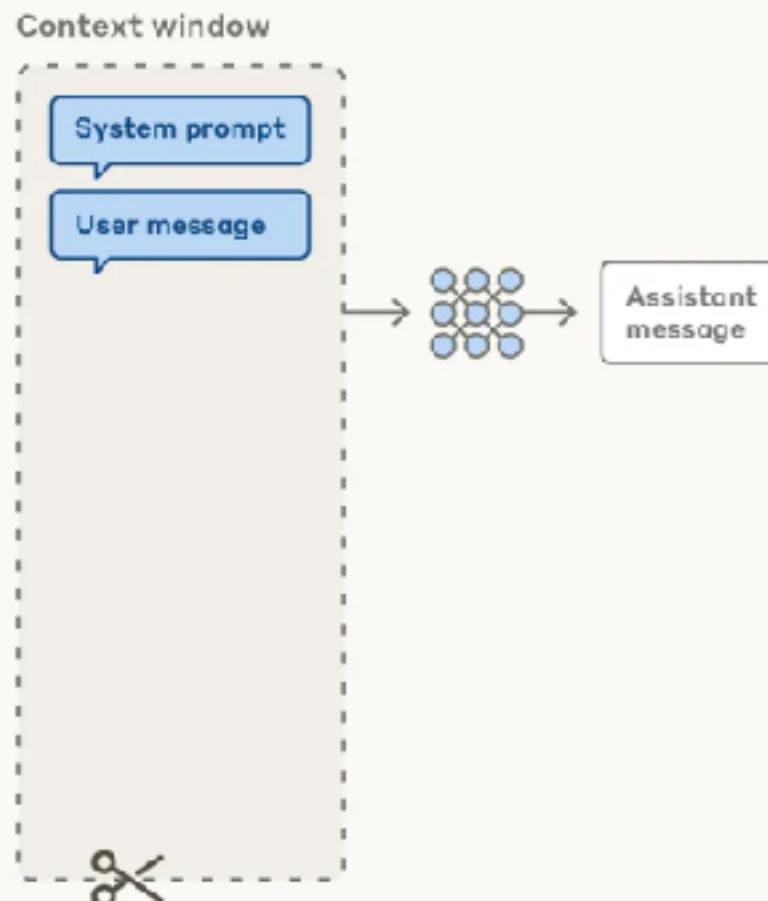
<https://www.promptingguide.ai/guides/context-engineering-guide>



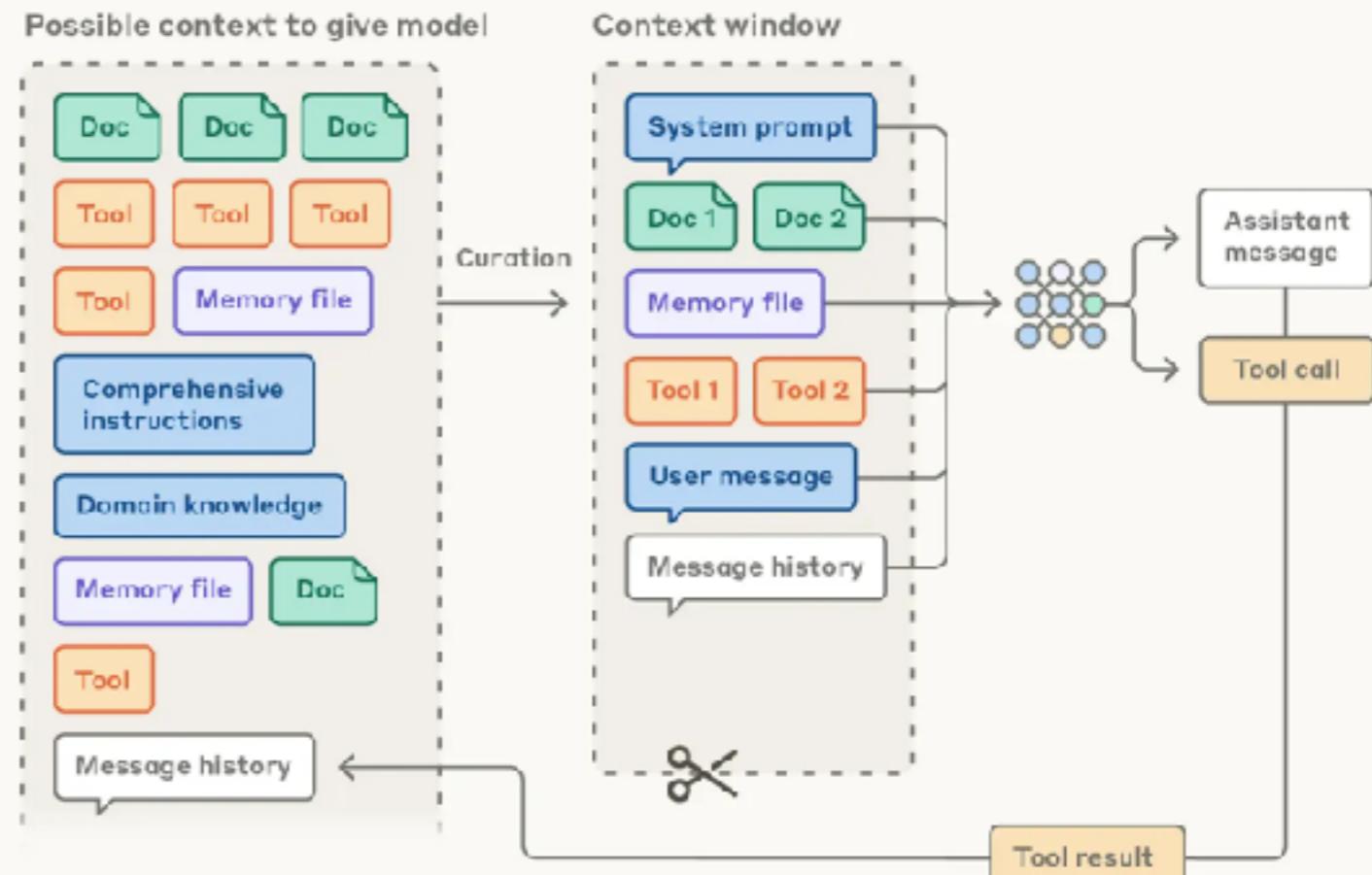
# Effective Context Engineering

## Prompt engineering vs. context engineering

Prompt engineering  
for single turn queries



Context engineering for agents



<https://www.anthropic.com/engineering/effective-context-engineering-for-ai-agents>

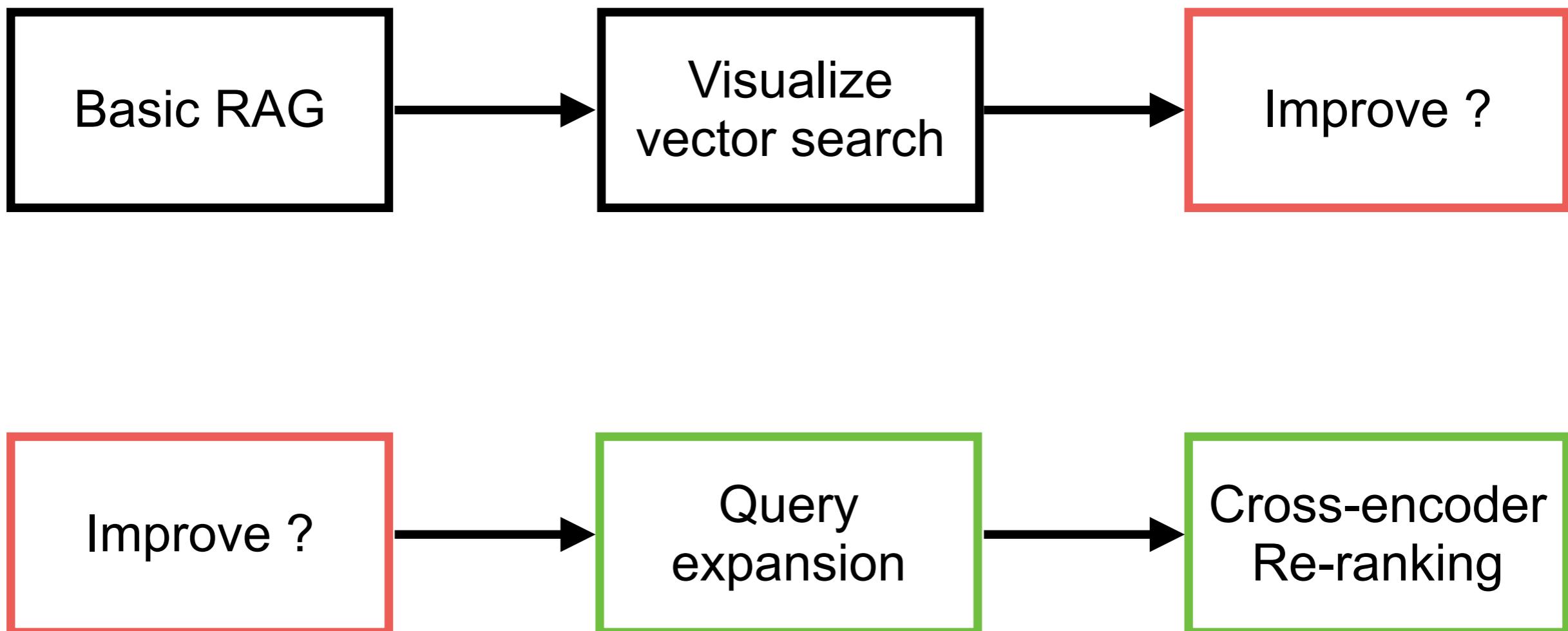


# RAG Workshop

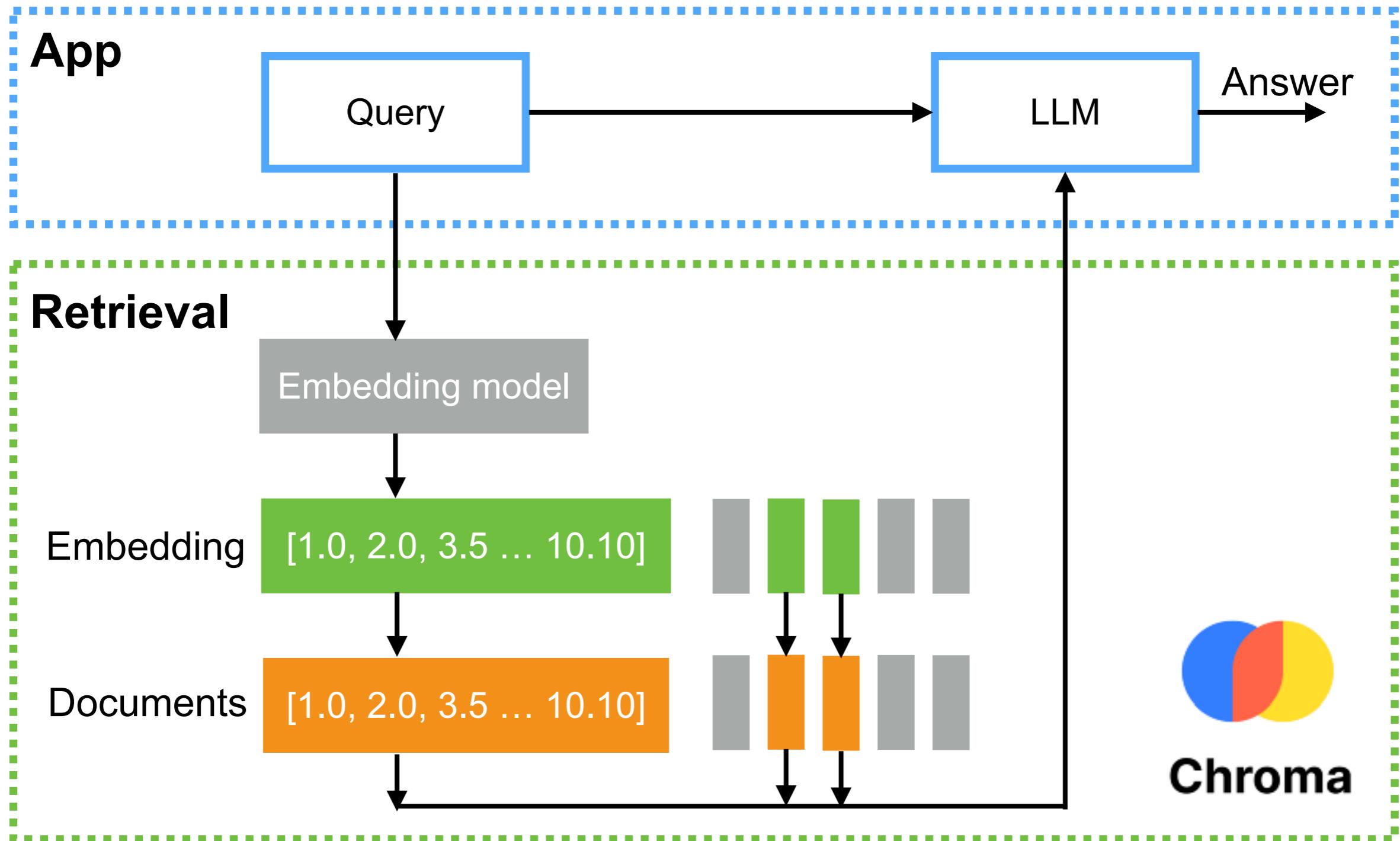
<https://github.com/up1/workshop-basic-llm/tree/main/workshop/basic-rag>



# RAG workshop



# Basic RAG



# Query Expansion

Query expansion is a widely used technique to improve the recall of search systems

Ambiguity

Vocabulary  
mismatch

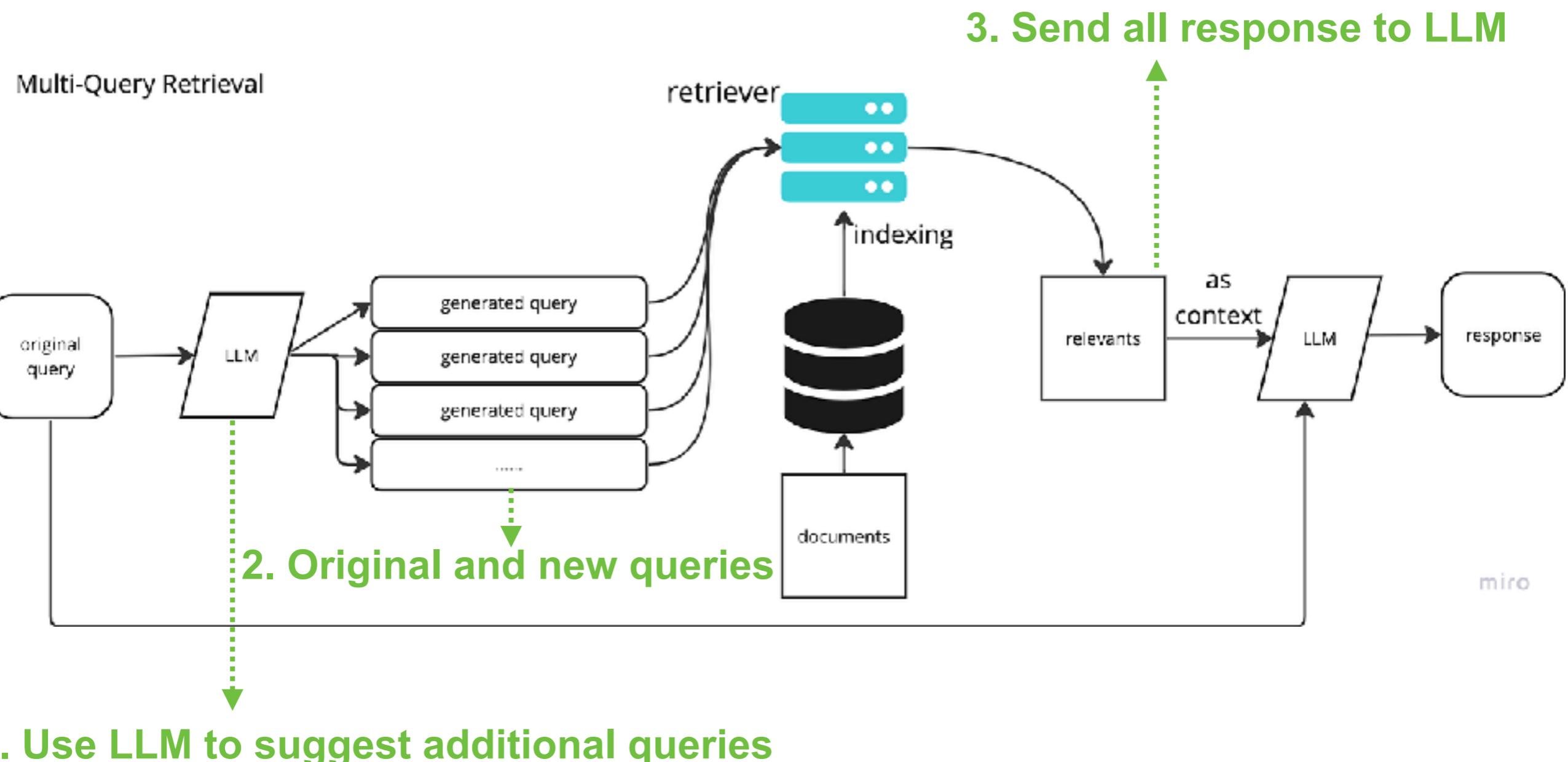
Lack of context

“Add synonyms, related context and contextual terms”

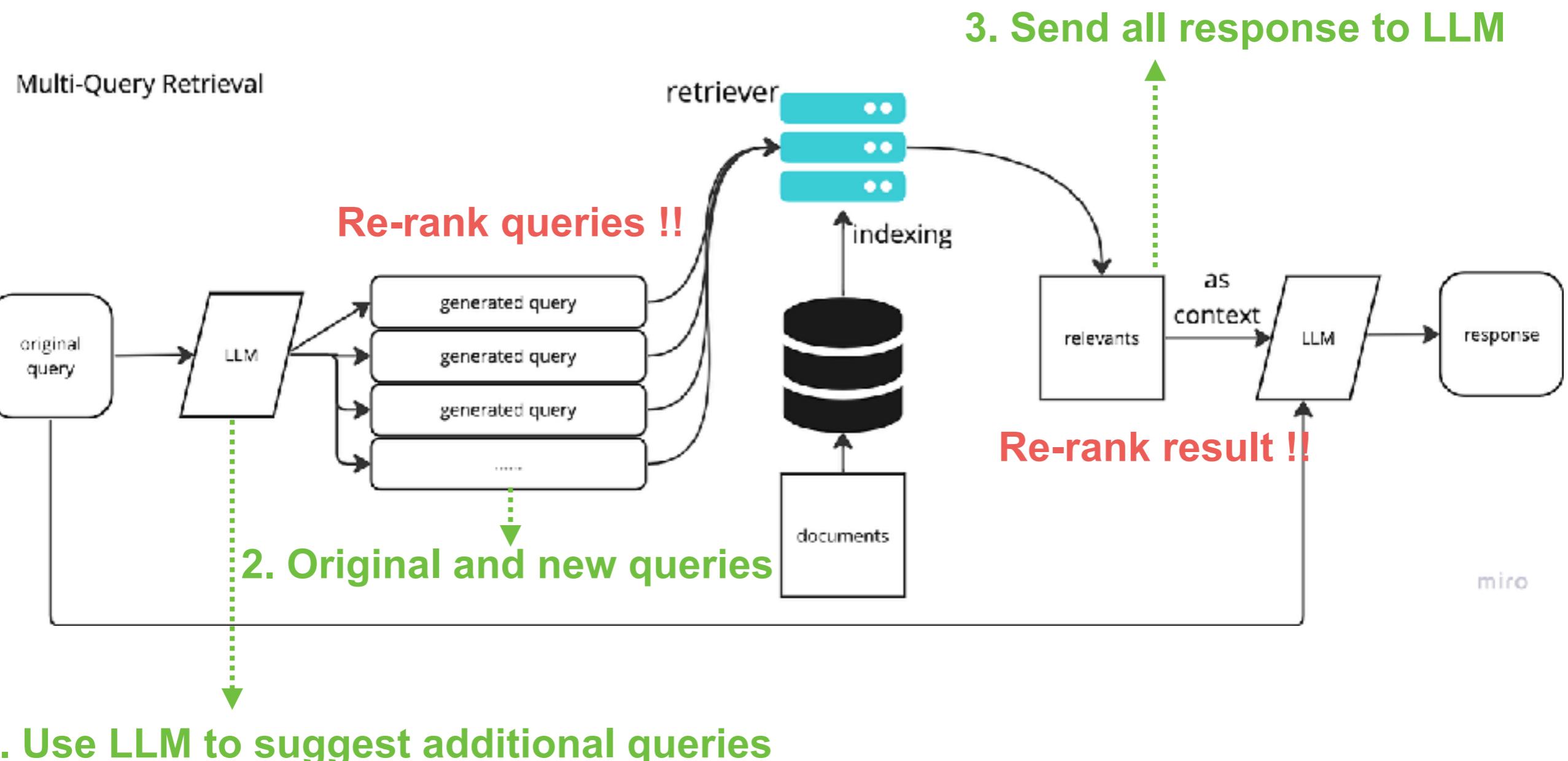
<https://arxiv.org/abs/2305.03653>



# Query Expansion

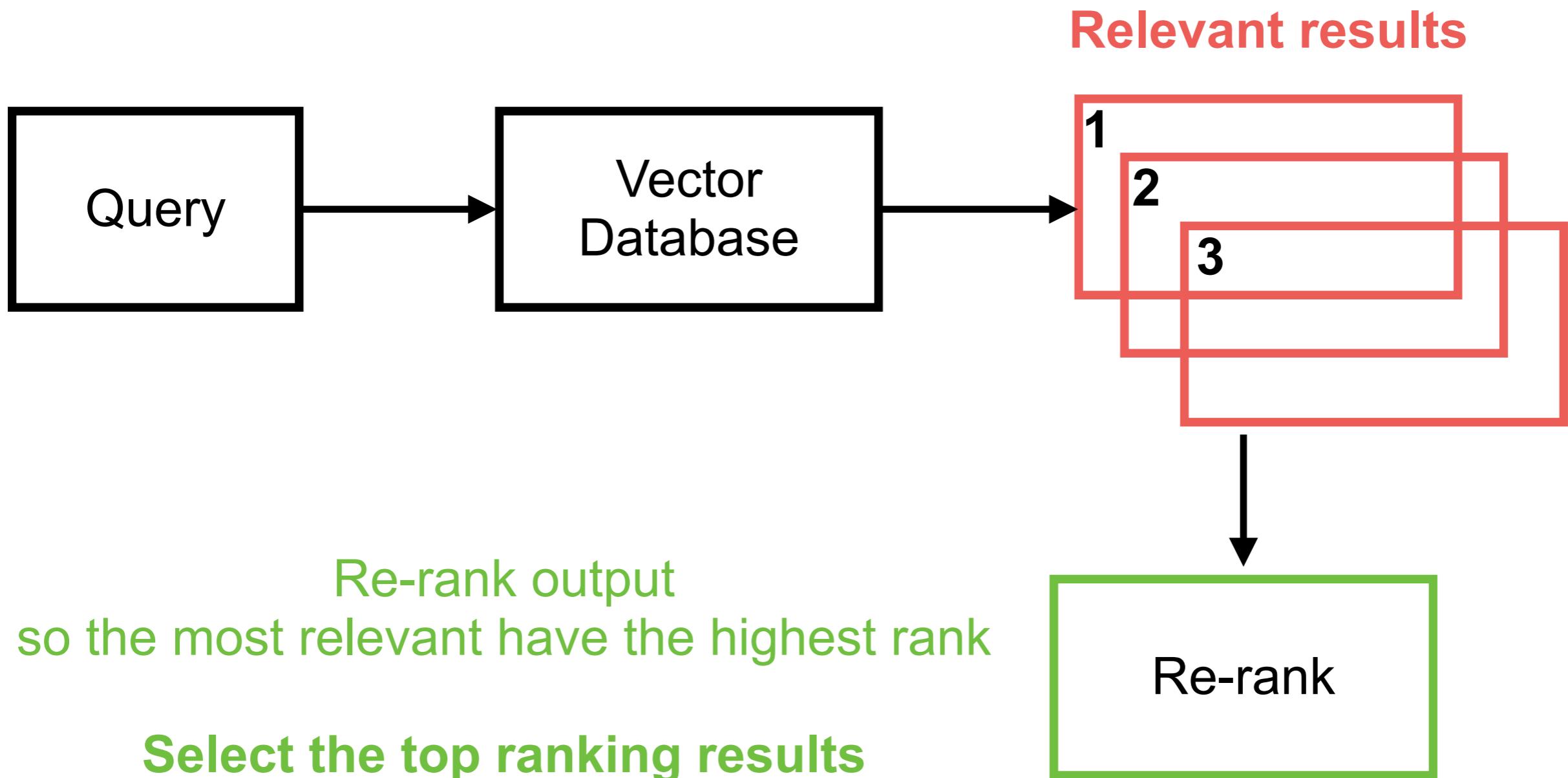


# Re-rank !!

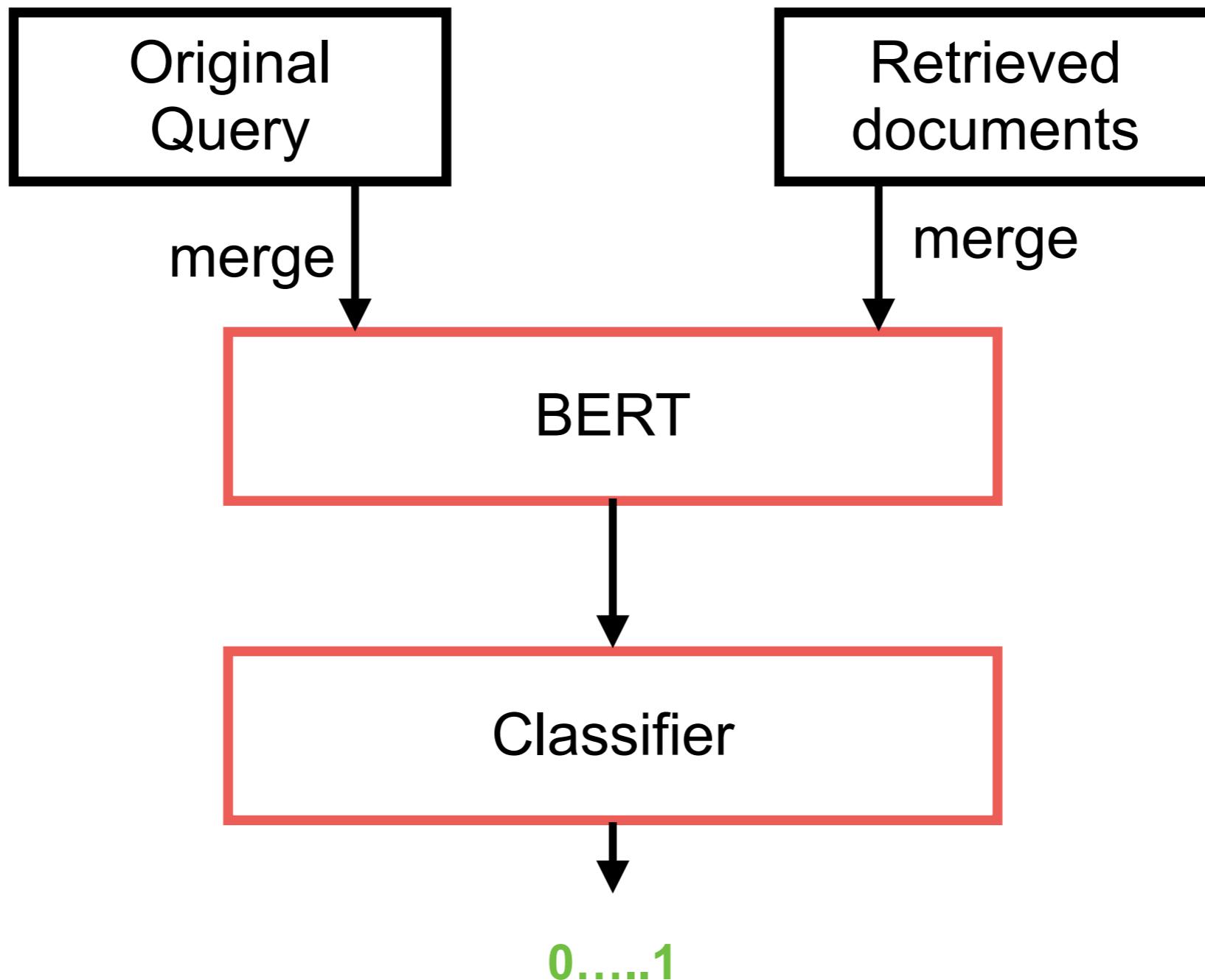


# Cross-encoder Re-ranking

How to ordering relevant results !!

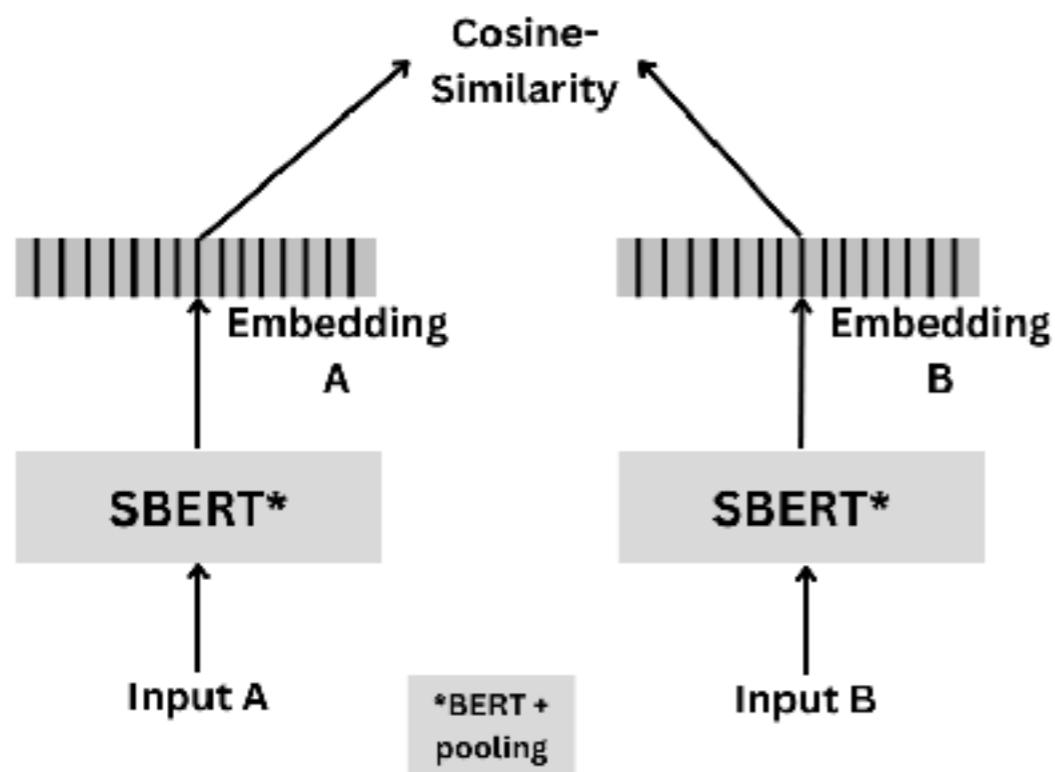


# Cross-Encoder

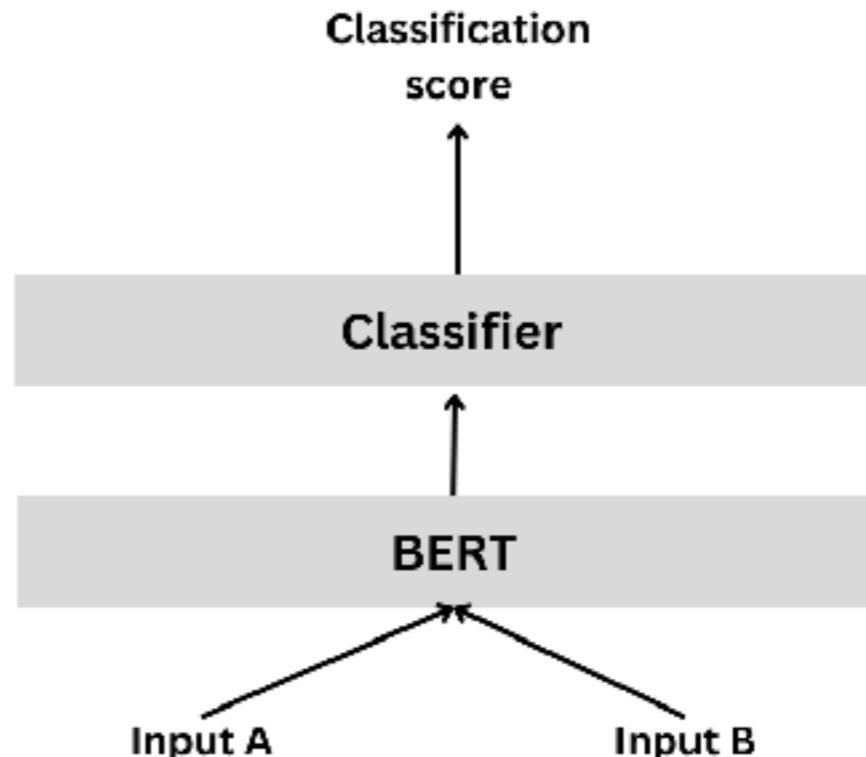


# Cross-Encoder !!

## Bi-encoder



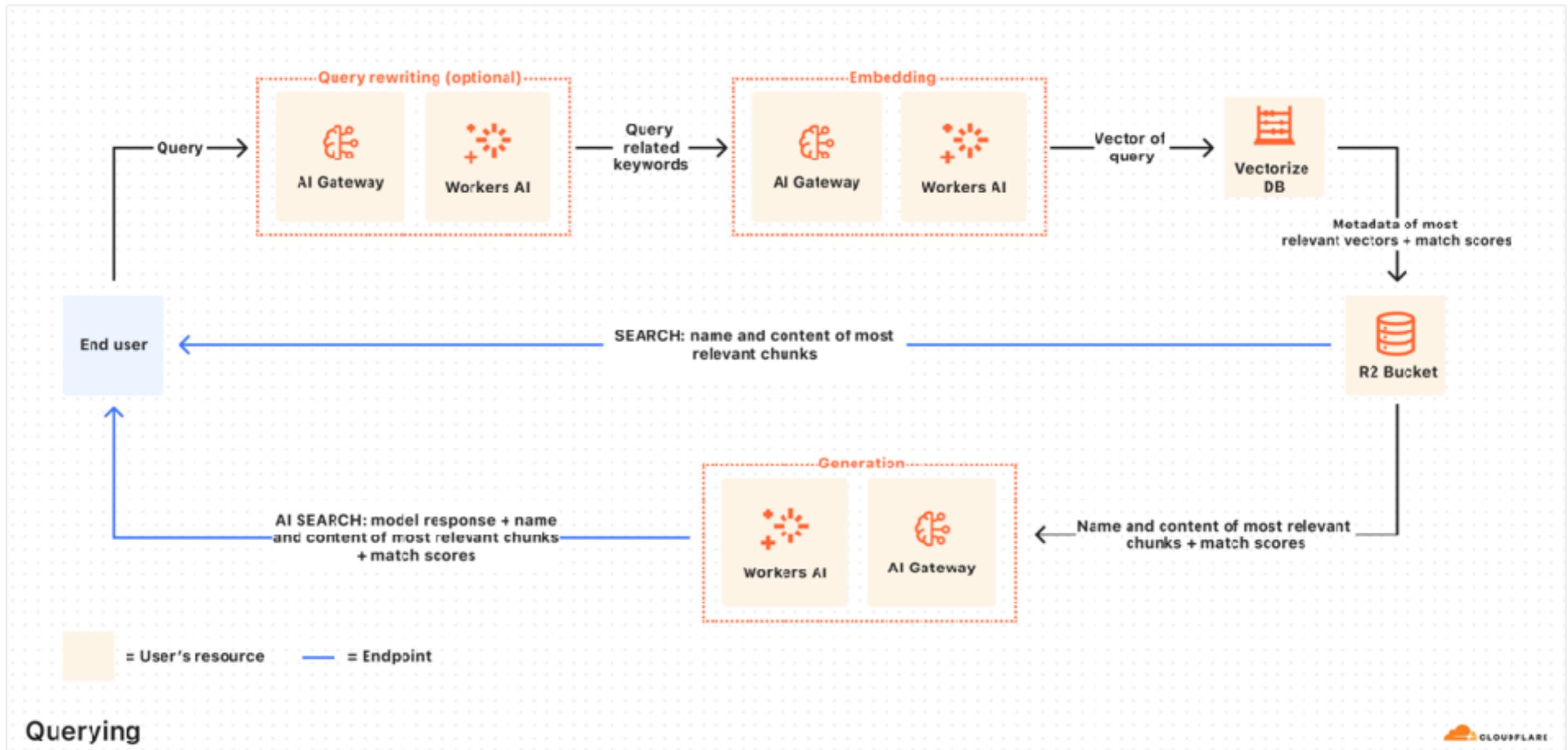
## Cross Encoder



[https://sbert.net/examples/cross\\_encoder/applications/README.html](https://sbert.net/examples/cross_encoder/applications/README.html)



# Cloudflare AutoRAG



<https://developers.cloudflare.com/autorag/>



# Guardrails



<https://github.com/guardrails-ai/guardrails>



# Guardrails

Help to build reliable AI applications

Guard for  
input and output

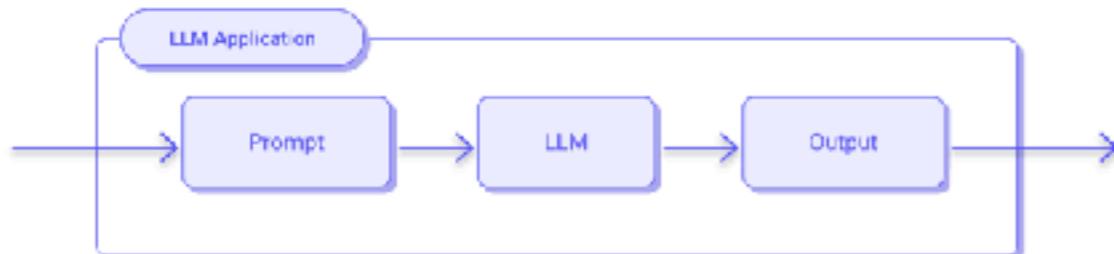
Help to generate  
Structured output

<https://github.com/guardrails-ai/guardrails>

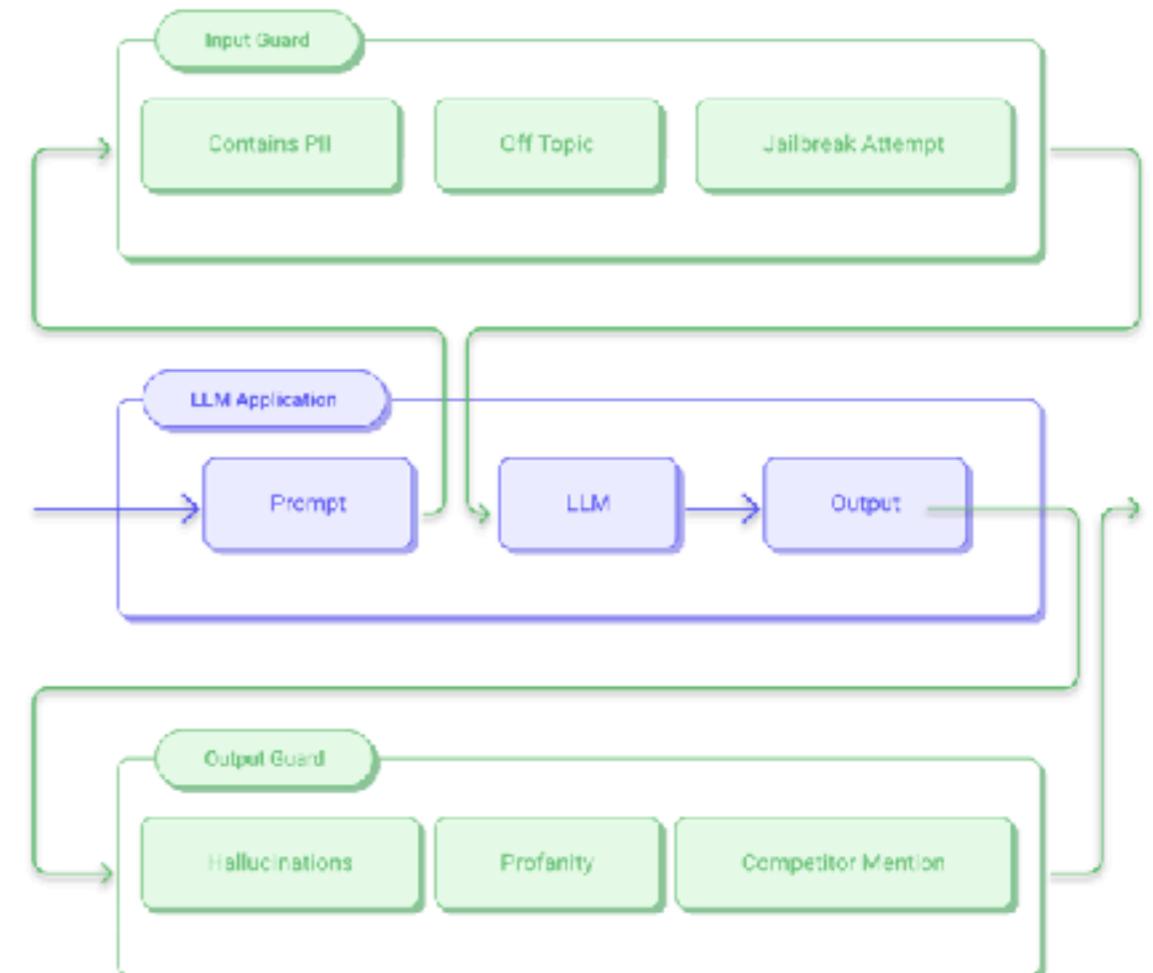


# Guardrails

## *Without Guardrails*



## *With Guardrails*



# Guardrails Hub

The screenshot shows the 'Validators' section of the Guardrails AI website. On the left, there are several filter categories: 'USE CASES' (Chatbots, Customer Support, Structured Data, RAG, Summarization, Codegen, Text2SQL), 'RISK CATEGORY' (Etiquette, Brand Risk, Factuality, Formatting, Invalid Code, Jailbreaking, Code Exploits, Data Leakage), 'INFRASTRUCTURE REQUIREMENTS' (ML, LLM, NA, Rule), 'CONTENT TYPE' (String, Object, List, Integer, Float, SQL, Code, CSV, Python), 'CERTIFICATION' (Guardrails Certified), and 'LANGUAGE' (EN). The main area displays a grid of 10 validators, each with a title, description, last updated date, input type (String, Brand Risk, Factuality, ML), and a 'Select' button:

- Arize Dataset Embeddings**: Validates that user-generated input does not match the dataset of jailbreak... Last updated 8 months ago. Select button.
- Ban List**: Validates that the output does not contain banned words, using fuzzy search. Last updated 8 months ago. Select button.
- Bespoke MiniCheck**: Validates that the LLM-generated text is supported by the provided context using... Last updated 7 months ago. Select button.
- Bias Check**: Validates that the text is free from biases related to age, gender, sex, ethnicity,... Last updated 1 week ago. Select button.
- Competitor Check**: Flags mentions of competitors. Fixes responses by filtering out competitor names. Last updated 5 months ago. Select button.
- Correct Language**: Validate that an LLM-generated text is in the expected language. If the text is not ... Last updated 11 months ago. Select button.
- Cucumber Expression Match**: Validates that the input string matches a specified cucumber expression. Last updated 6 months ago. Select button.
- Detect PII**: Detects personally identifiable information (PII) in text, using Microsoft Presidio. Last updated 6 months ago. Select button.

<https://hub.guardrailsai.com/>



# Guardrails Index

## AI Guardrails Index

Created by Guardrails AI

[Download PDF](#)

[Register for Webinar](#)

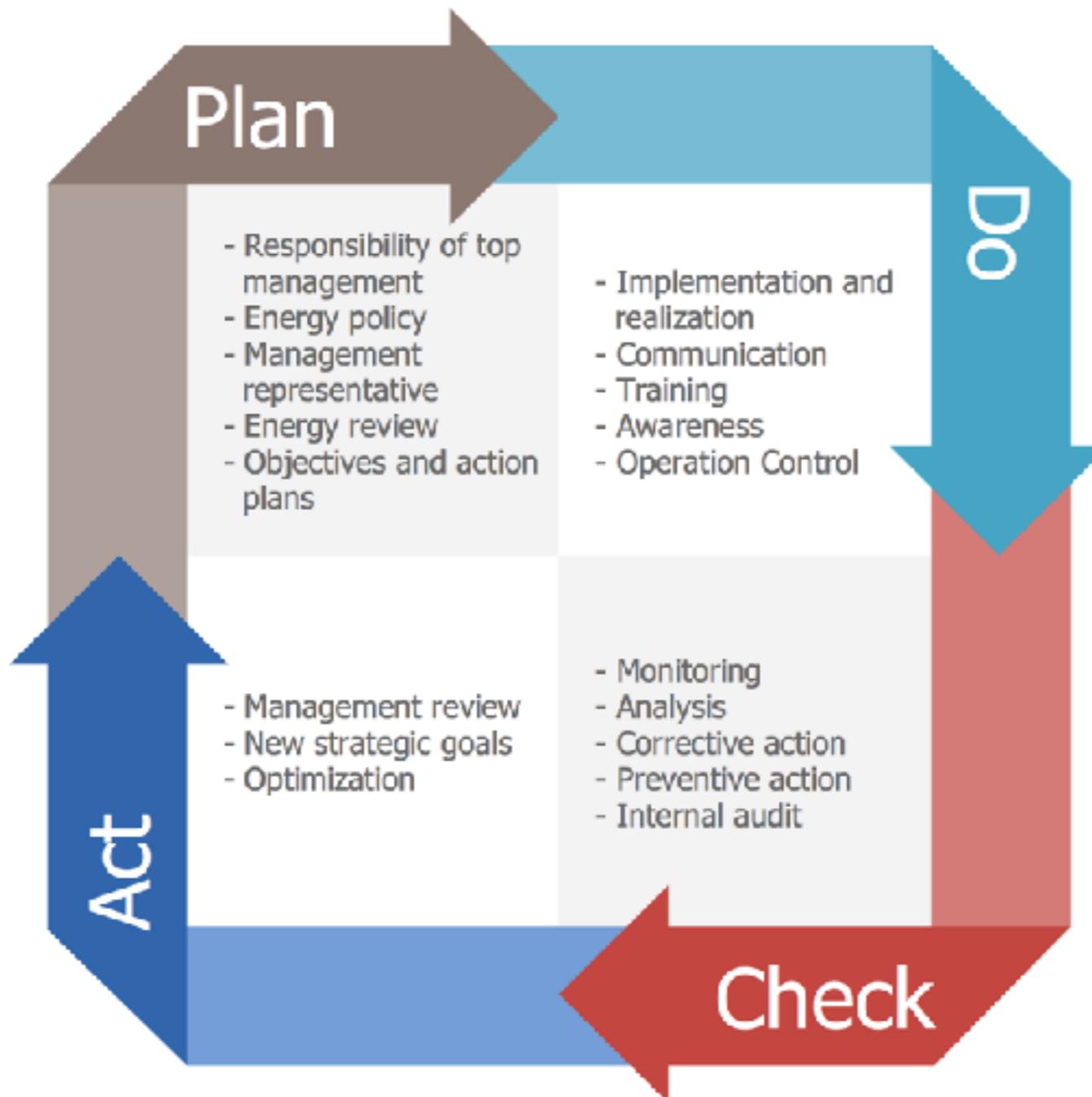
## AI Guardrails Categories

We broke AI safety down into 6 categories and curated datasets and models that demonstrate the state of AI guardrails using LLMs and other open source models.

<https://index.guardrailsai.com/>

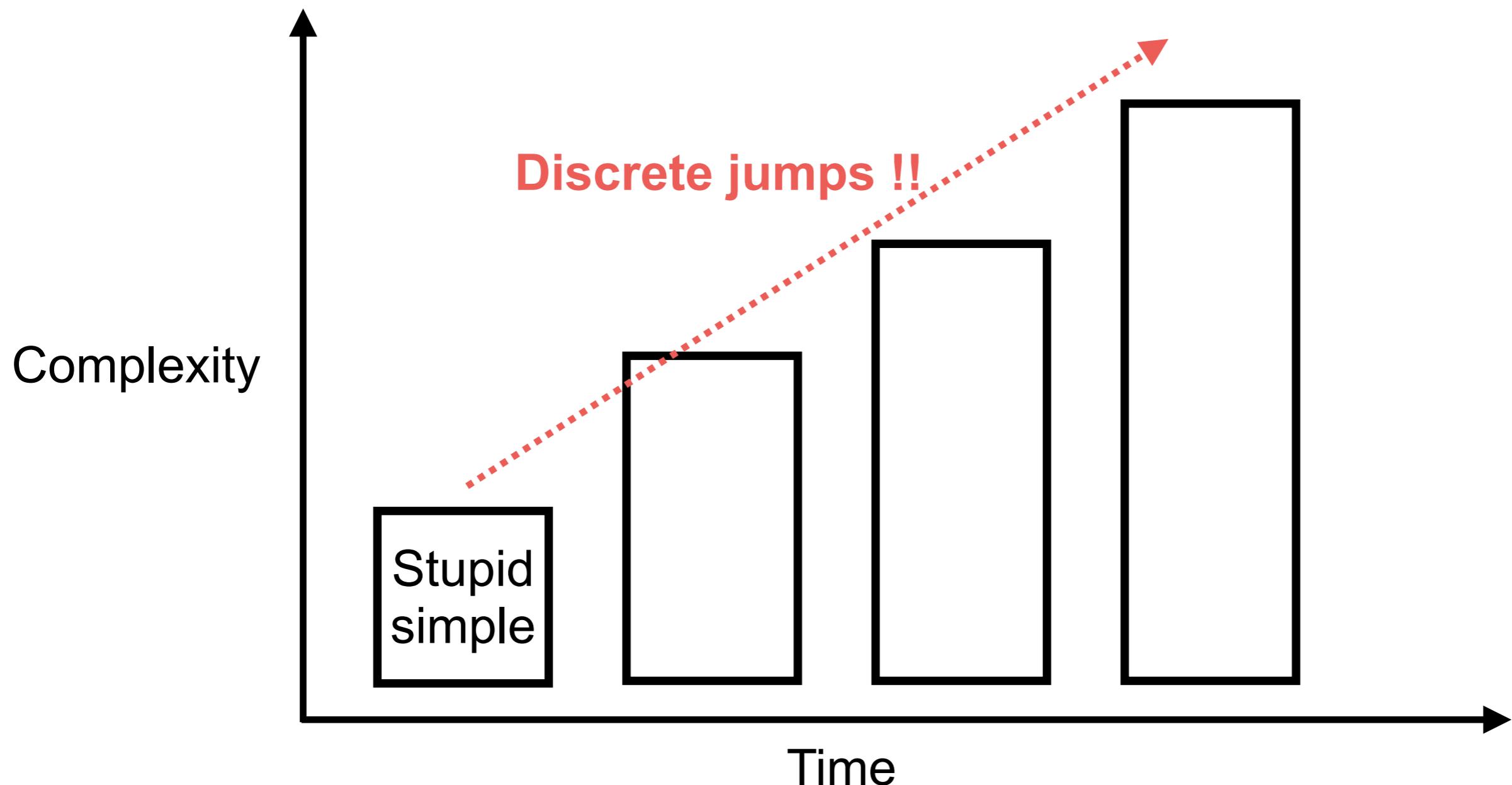


# PDCA process

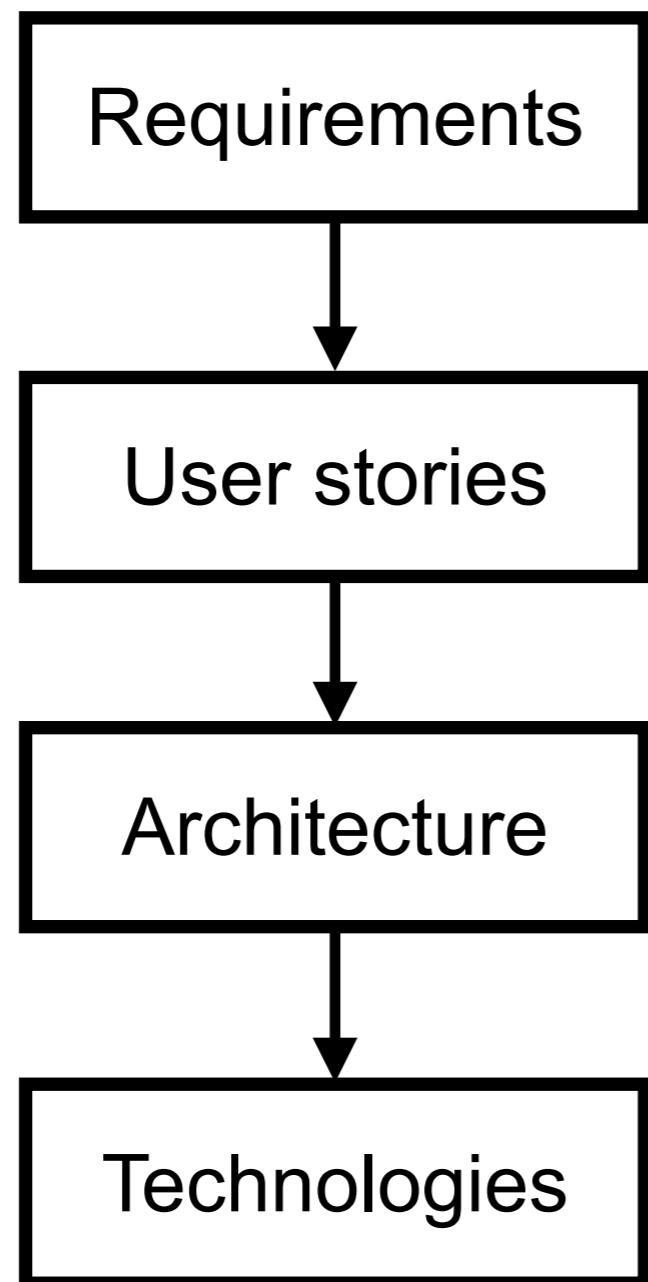


# Staged Complexity

Start small and get a feature working first



# Planning workflow



What you're build ?

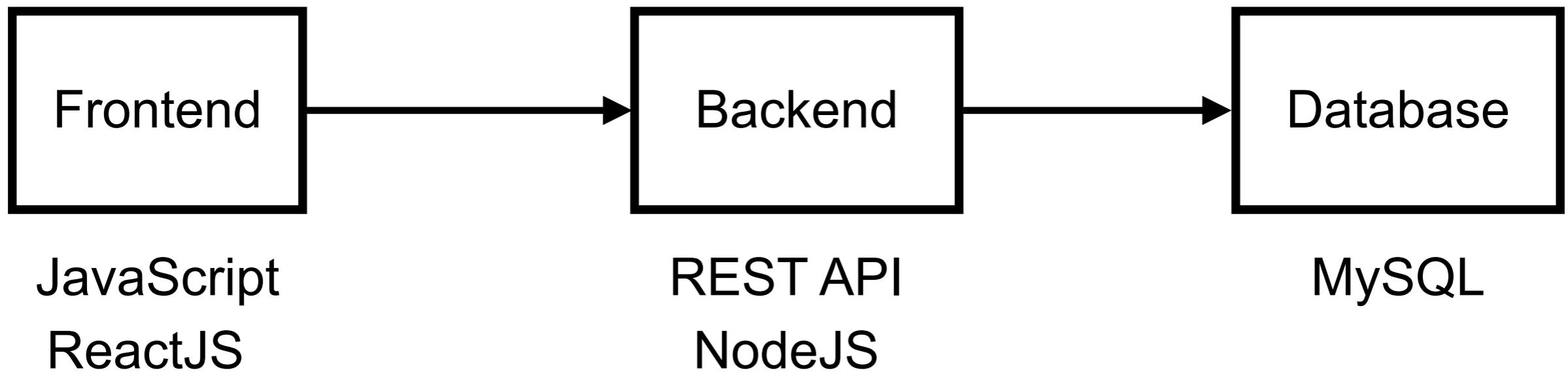
What use need ?

Selection ?

Selection ?



# Architecture of project

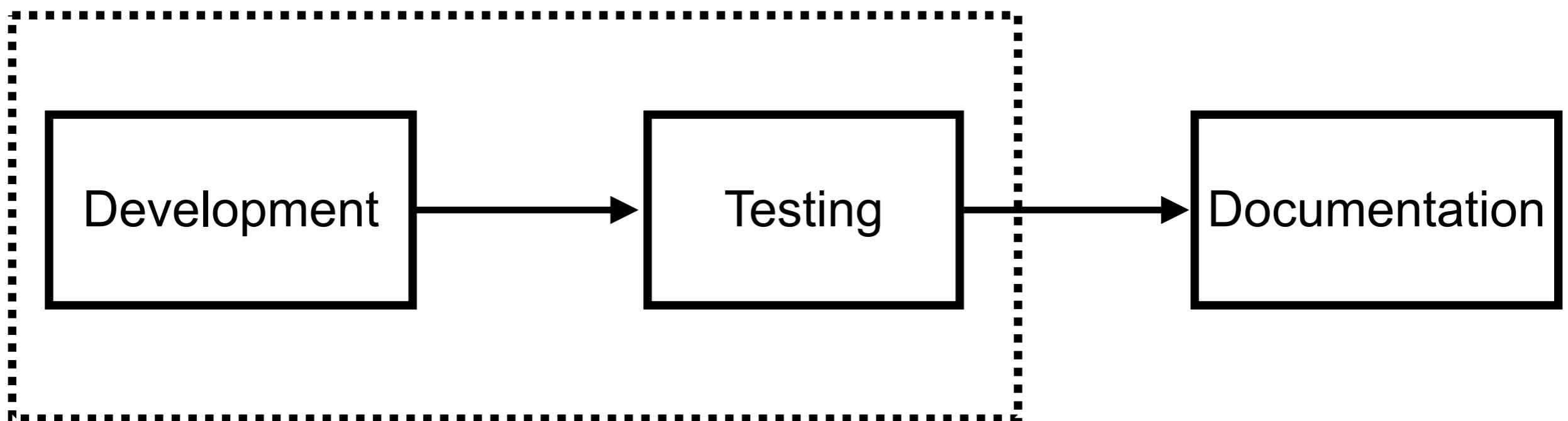


Choose your tech stack !!



# Processes

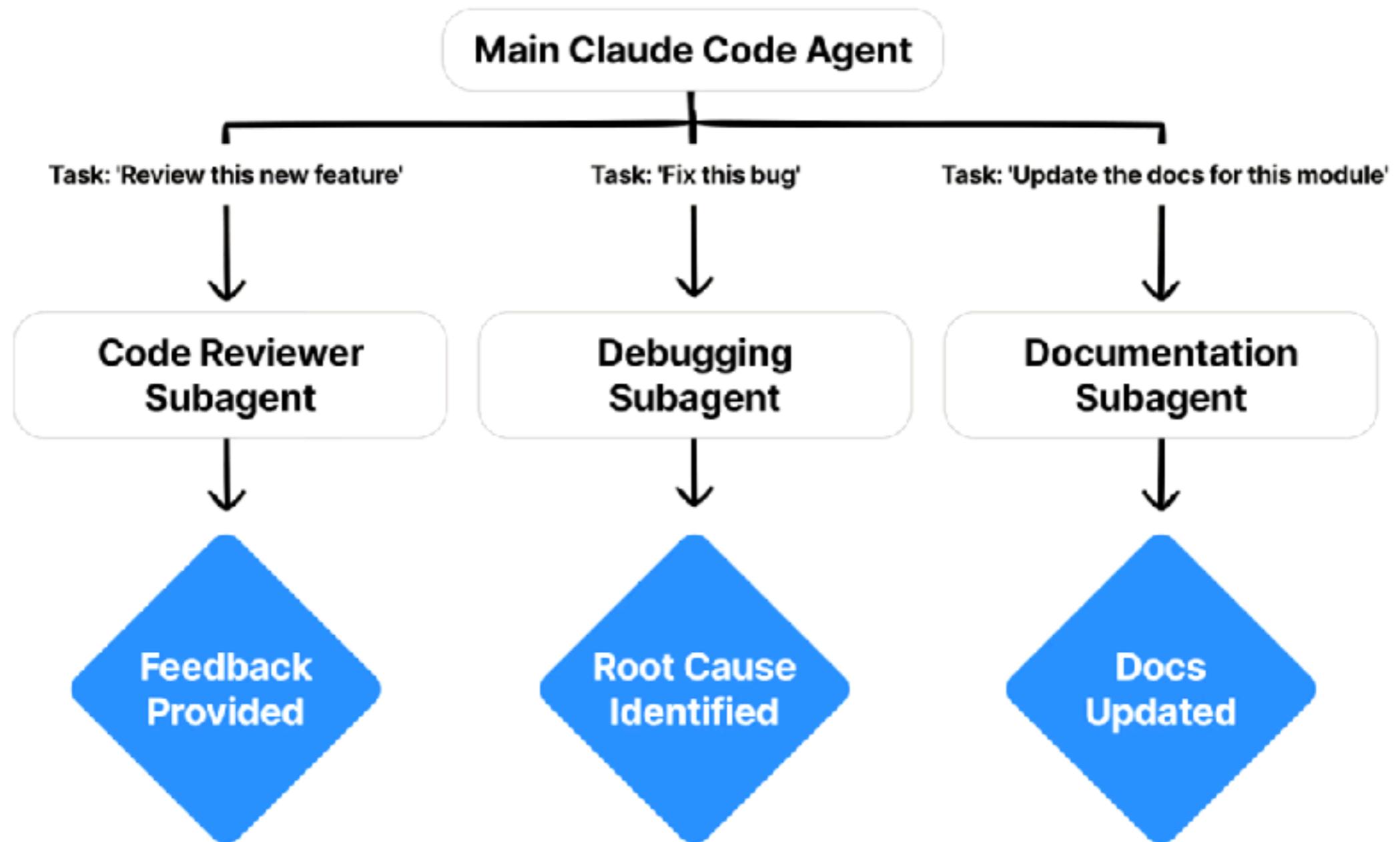
## Iterative and incremental process



# AI Subagent



# AI Subagent



<https://code.claude.com/docs/en/sub-agents>



# AI Subagent



<https://github.com/VoltAgent/awesome-claude-code-subagents>



# Start your journey

