AI & Engineering

A bit of history & usage.

sourcetoad

Plan of Attack

- The beginning of the LLM
- A bit of History (GPT1-5)
- The Evolution of Tools
- AI in the Workplace
- Building with AI
- AI Supporting Software

Who?

- **Connor Tumbleson**
- @iBotPeaches
- Sourcetoad
  - Director of Engineering

TAMPA BAY TECH
AWARDS
SHOW
11.7.25 | ARMATURE WORKS

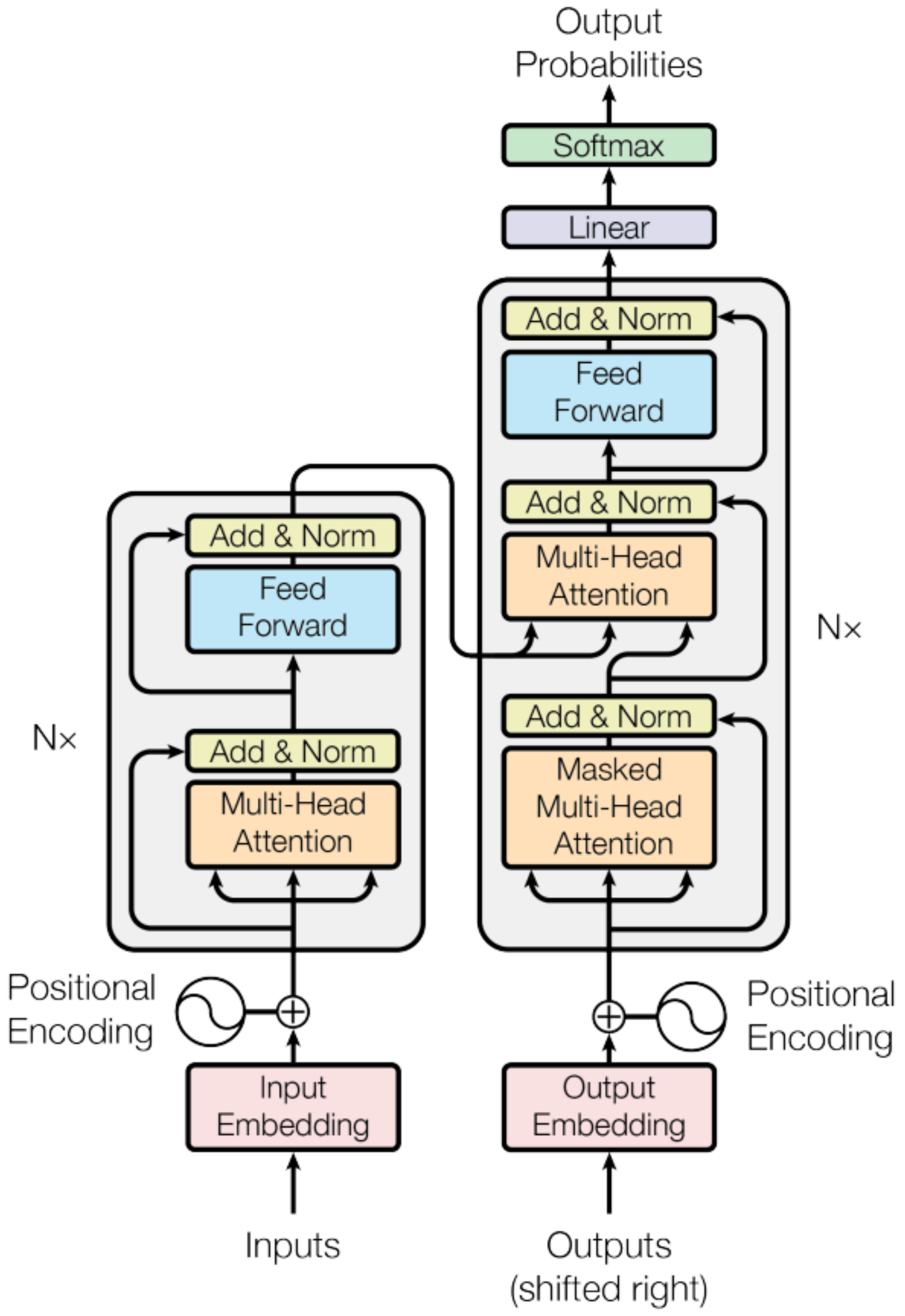
DGR
SYSTEMS

Thank You to our Sponsors!

CRITICAL START

Attention Is All You Need

- 2017 Research Paper
- Google Scientists
- 8 equally cited authors
- Cited over 200,000 times.



sourcetoad

Transformer Explainer

Embedding
Transformer Block 10
Multi-head Self Attention
Residual
Data
visualization
em
powers
users
to
Q
K
V
Key
Query
Value
Attention
Out
Head 7 of 12
MLP
Residual
2 more identical Transformer Blocks
Probabilities

visualize	54.67%
create	20.87%
see	12.09%
make	6.26%
easily	6.11%
quickly	0%
explore	0%
find	0%
understand	0%
build	0%
better	0%
identify	0%
take	0%
visually	0%
analyze	0%
discover	0%
design	0%
visual	0%
get	0%
view	0%
compare	0%
use	0%
generate	0%
access	0%

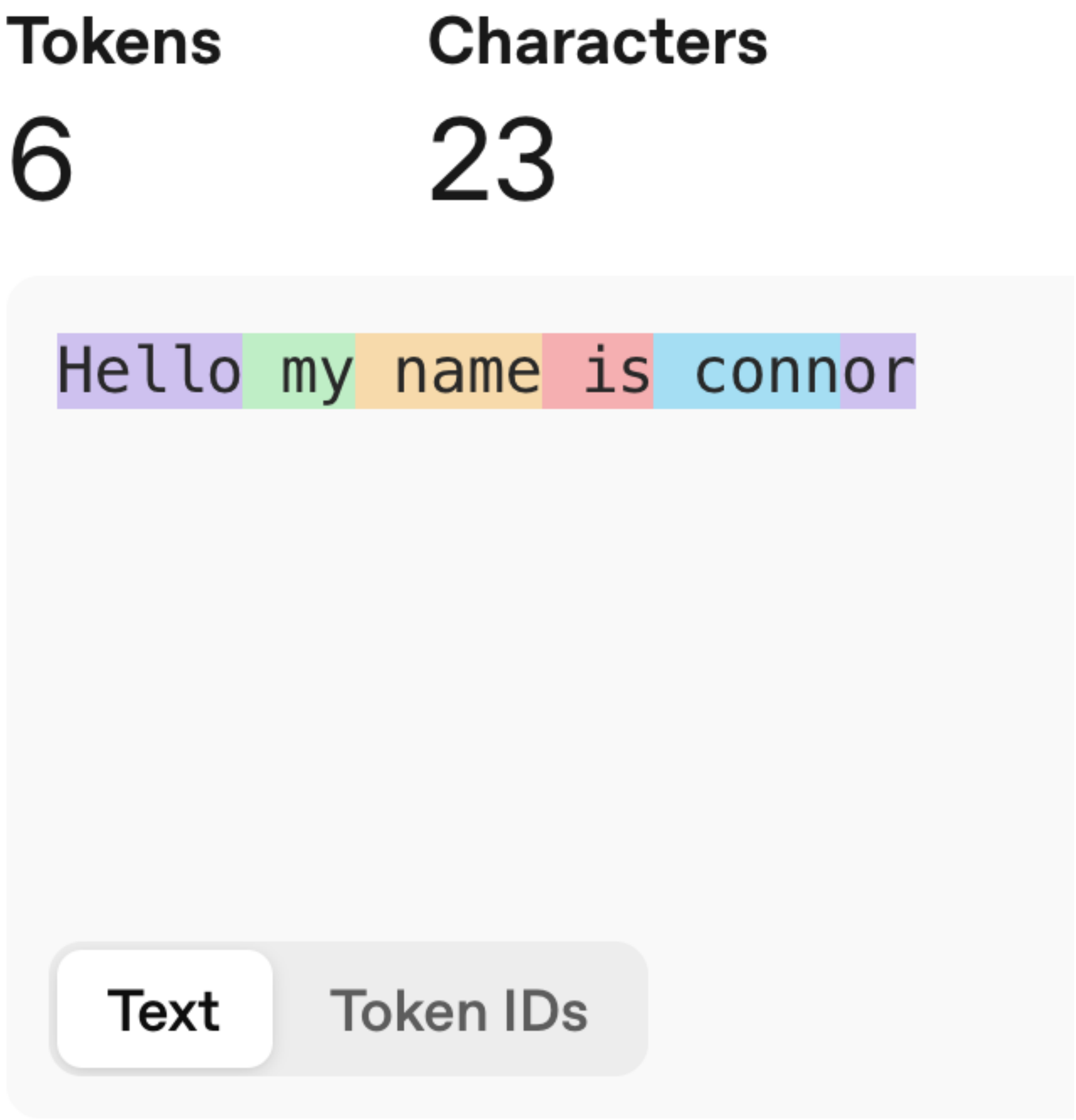
https://poloclub.github.io/transformer-explainer/

ChatGPT 1-5

GPT-**1**: The beginning (2018)

- Heavy research usage
- Trained from books
- Tokens in/out
- ~500 token limit

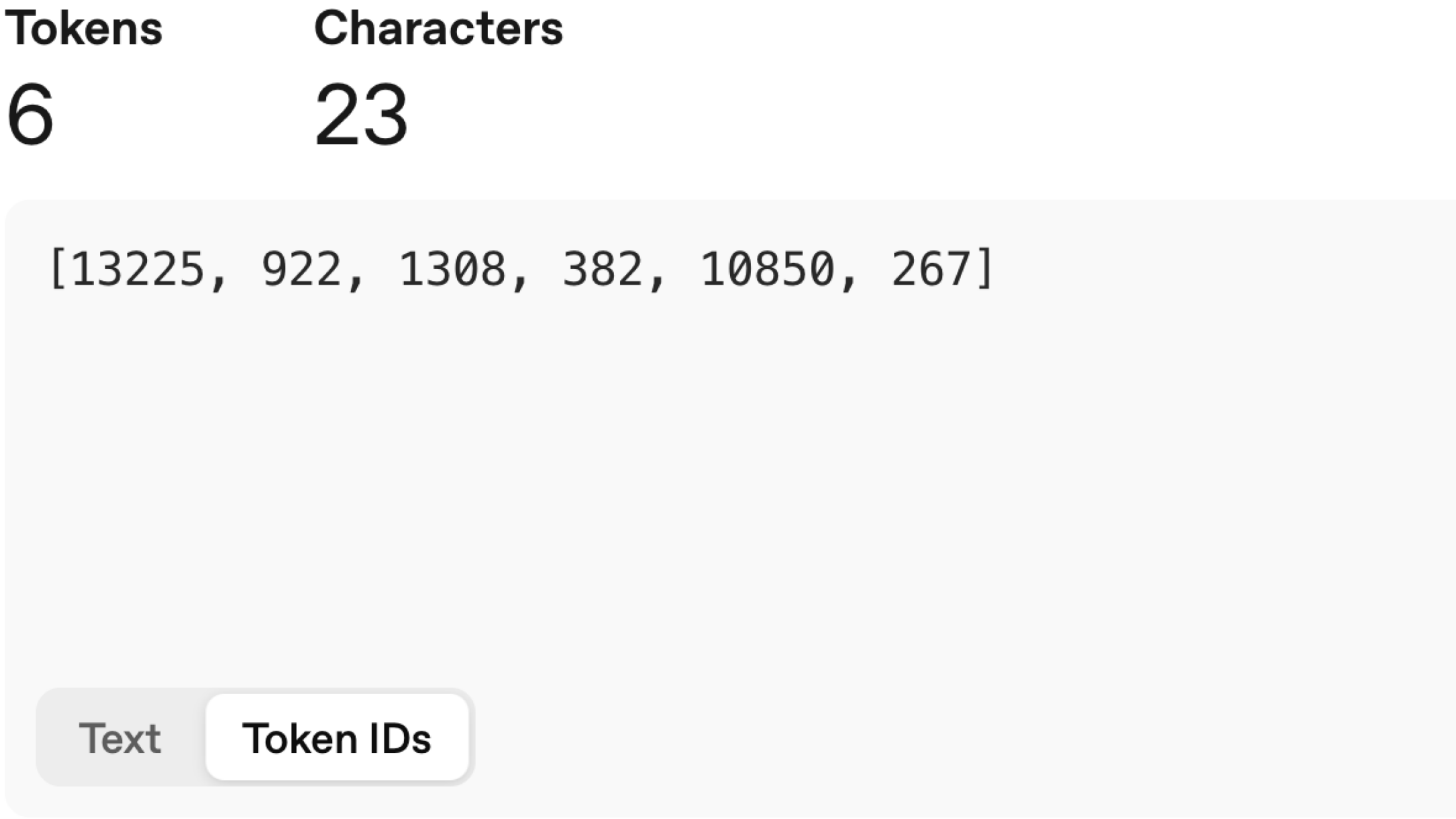
Hold up. **Tokens**?



Tokens
6

Characters
23

```
Hello my name is connor
```



Tokens
6

Characters
23

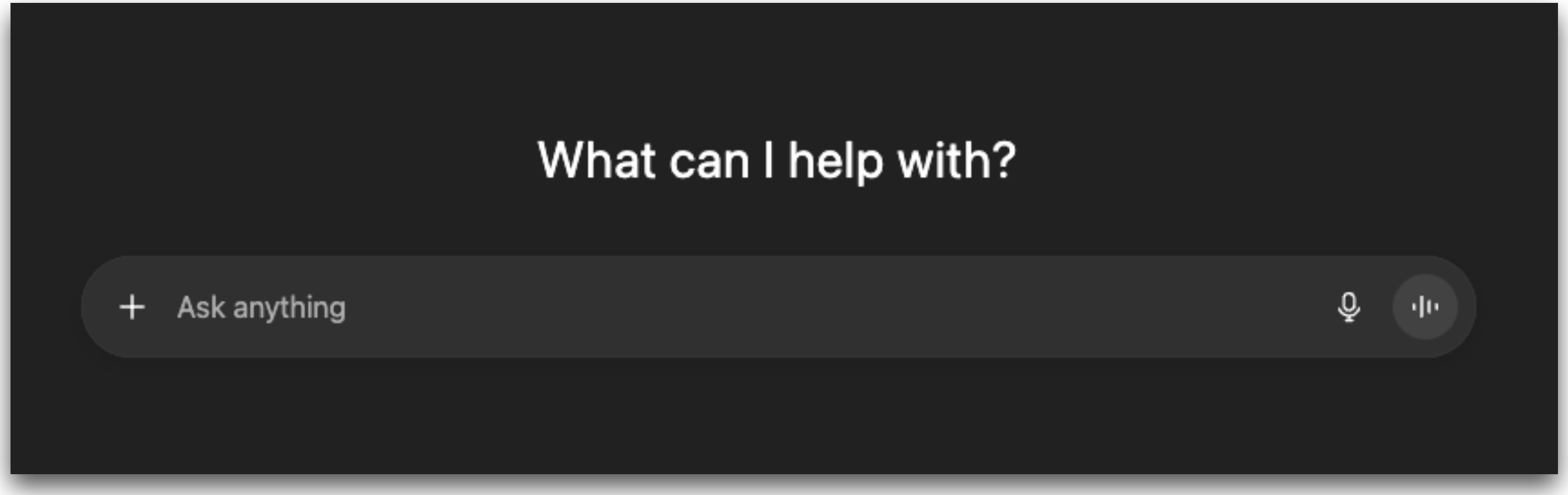
```
[13225, 922, 1308, 382, 10850, 267]
```

GPT-**2**: The Expanse (2019)

- Zero-shot training
- Trained from web/books
- ~1k token limit

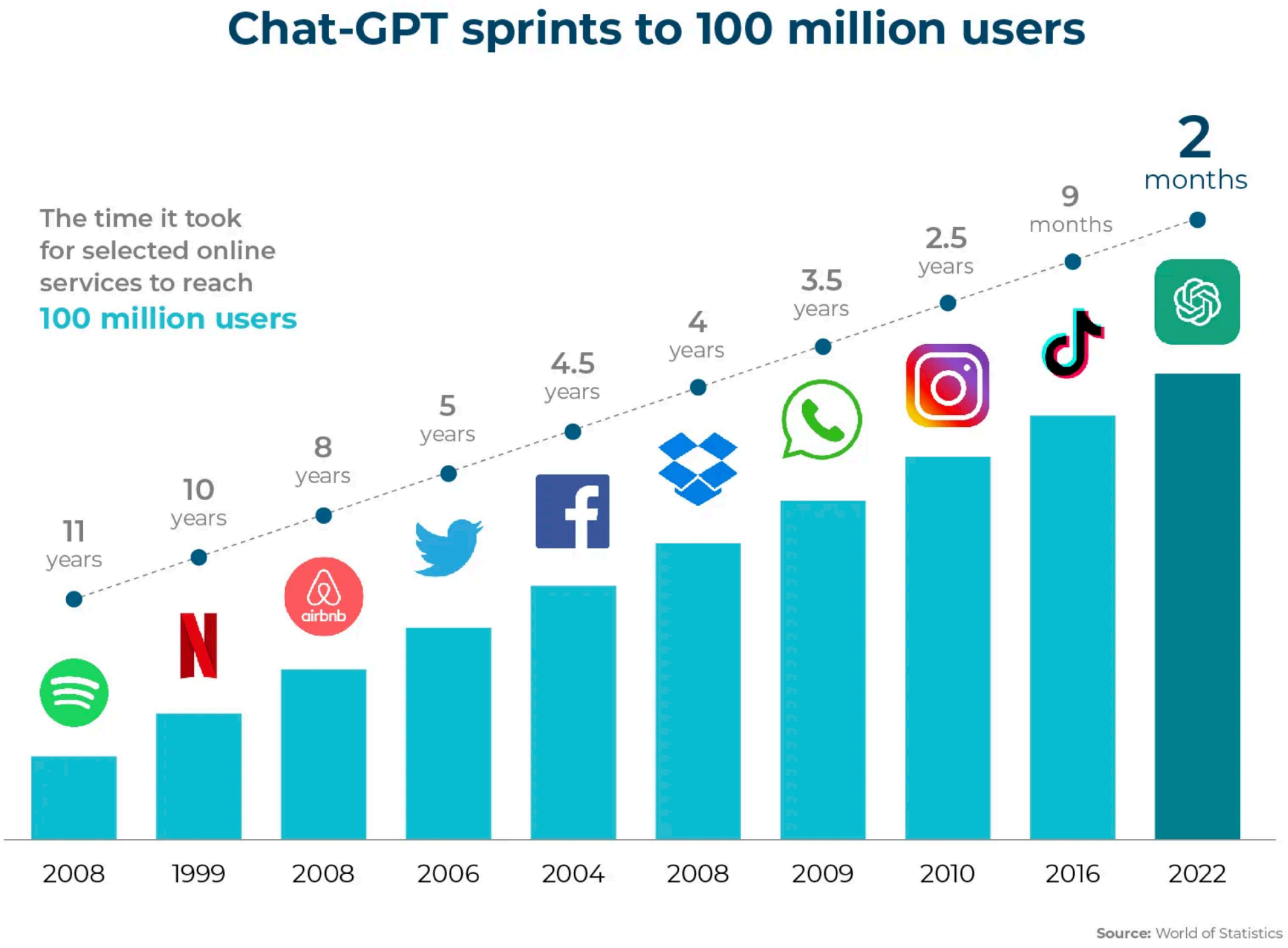
GPT-**3**: The Explosion (2020)

- The commercialization begins
- Trained from everything?
- ~2k token limit



What can I help with?

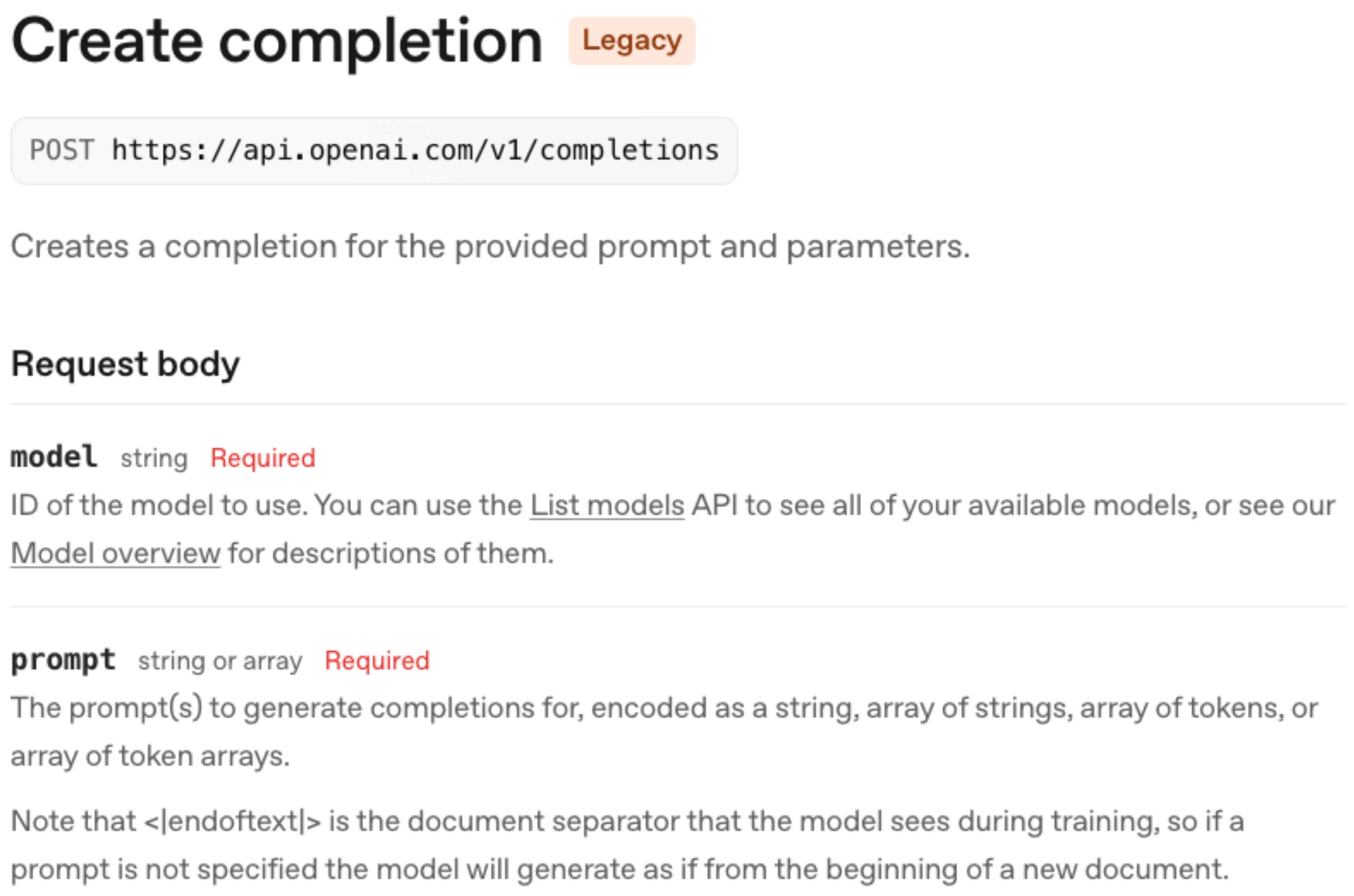
GPT-**3**: The Explosion (2020)



Source: World of Statistics

GPT-**3**: The Explosion (2020)

- /completions
  - `logit_bias`
  - `presence_penalty`
  - `seed`
  - `temperature`
  - `top_p`



Create completion Legacy

`POST https://api.openai.com/v1/completions`

Creates a completion for the provided prompt and parameters.

Request body

`model` string Required

ID of the model to use. You can use the List models API to see all of your available models, or see our Model overview for descriptions of them.

`prompt` string or array Required

The prompt(s) to generate completions for, encoded as a string, array of strings, array of tokens, or array of token arrays.

Note that <|endoftext|> is the document separator that the model sees during training, so if a prompt is not specified the model will generate as if from the beginning of a new document.

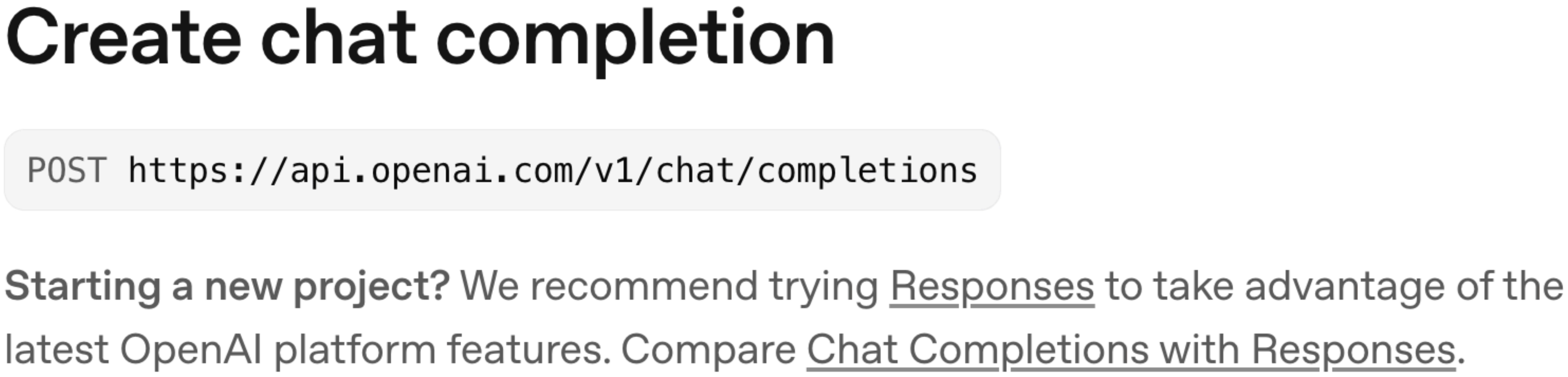
GPT-**4**: The Refinement (2023)

- Multimodal Input
- Trained from more and more.
- ~32k token limit

GPT-4 can accept a prompt of text and images, which—parallel to the text-only setting—lets the user specify any vision or language task.

GPT-4: The Refinement (2023)

- New API `/chat/completions`



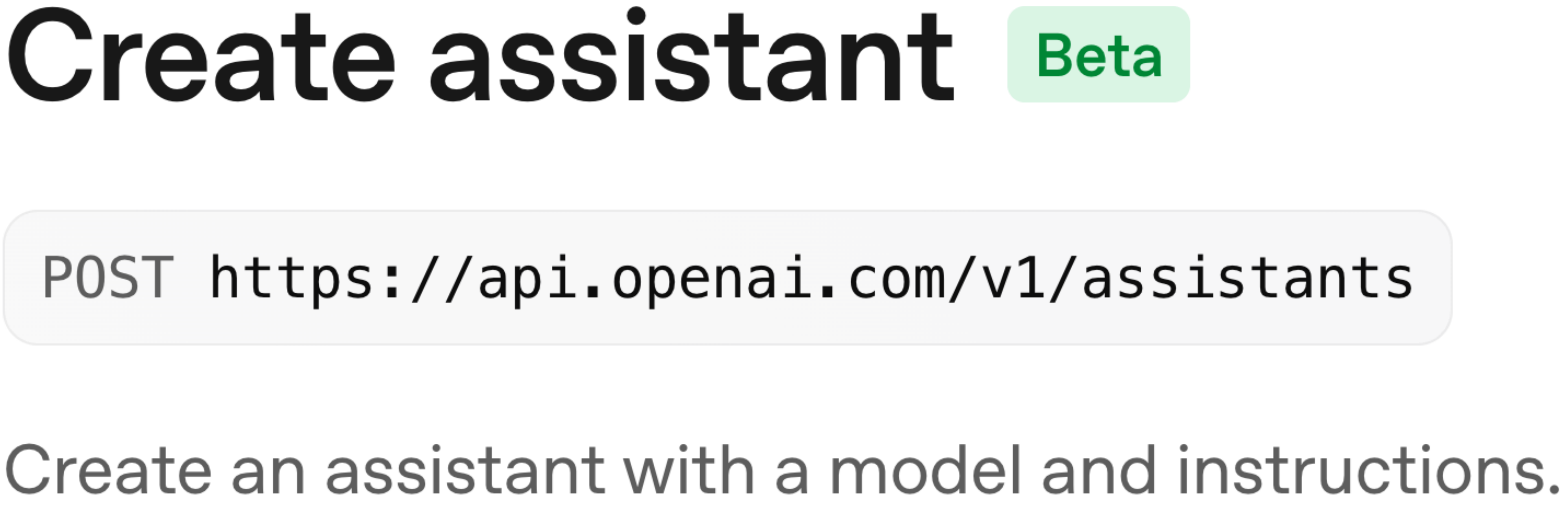
Create chat completion

```
POST https://api.openai.com/v1/chat/completions
```

Starting a new project? We recommend trying Responses to take advantage of the latest OpenAI platform features. Compare Chat Completions with Responses.

GPT-**4**: The Refinement (2023)

- New API `/assistants`



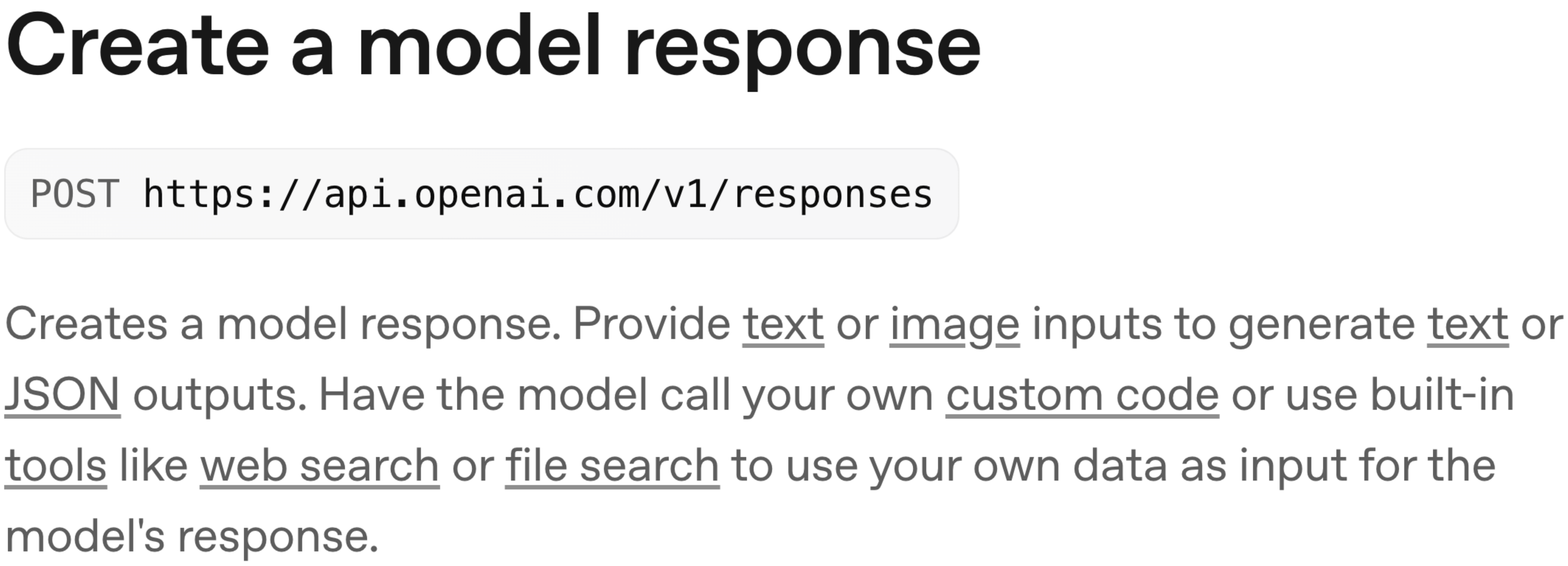
Create assistant Beta

```
POST https://api.openai.com/v1/assistants
```

Create an assistant with a model and instructions.

GPT-4: The Refinement (2024)

- New API `/responses`



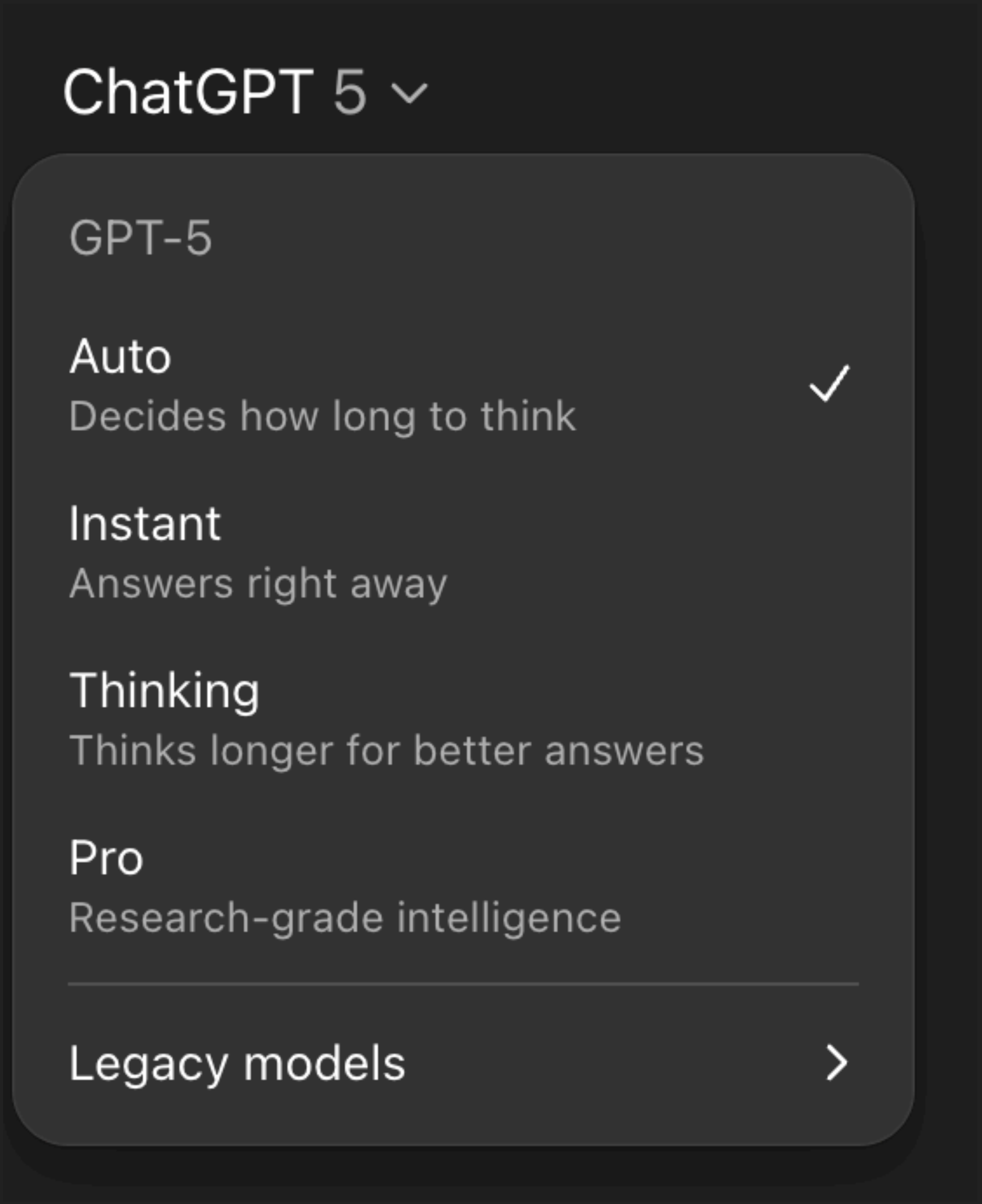
Create a model response

`POST https://api.openai.com/v1/responses`

Creates a model response. Provide text or image inputs to generate text or JSON outputs. Have the model call your own custom code or use built-in tools like web search or file search to use your own data as input for the model's response.

GPT-**5**: The Reasoning (2025)

- Reasoning / Completion
- Trained from [redacted]
- ~512k token limit



ChatGPT 5

GPT-5

Auto ✓
Decides how long to think

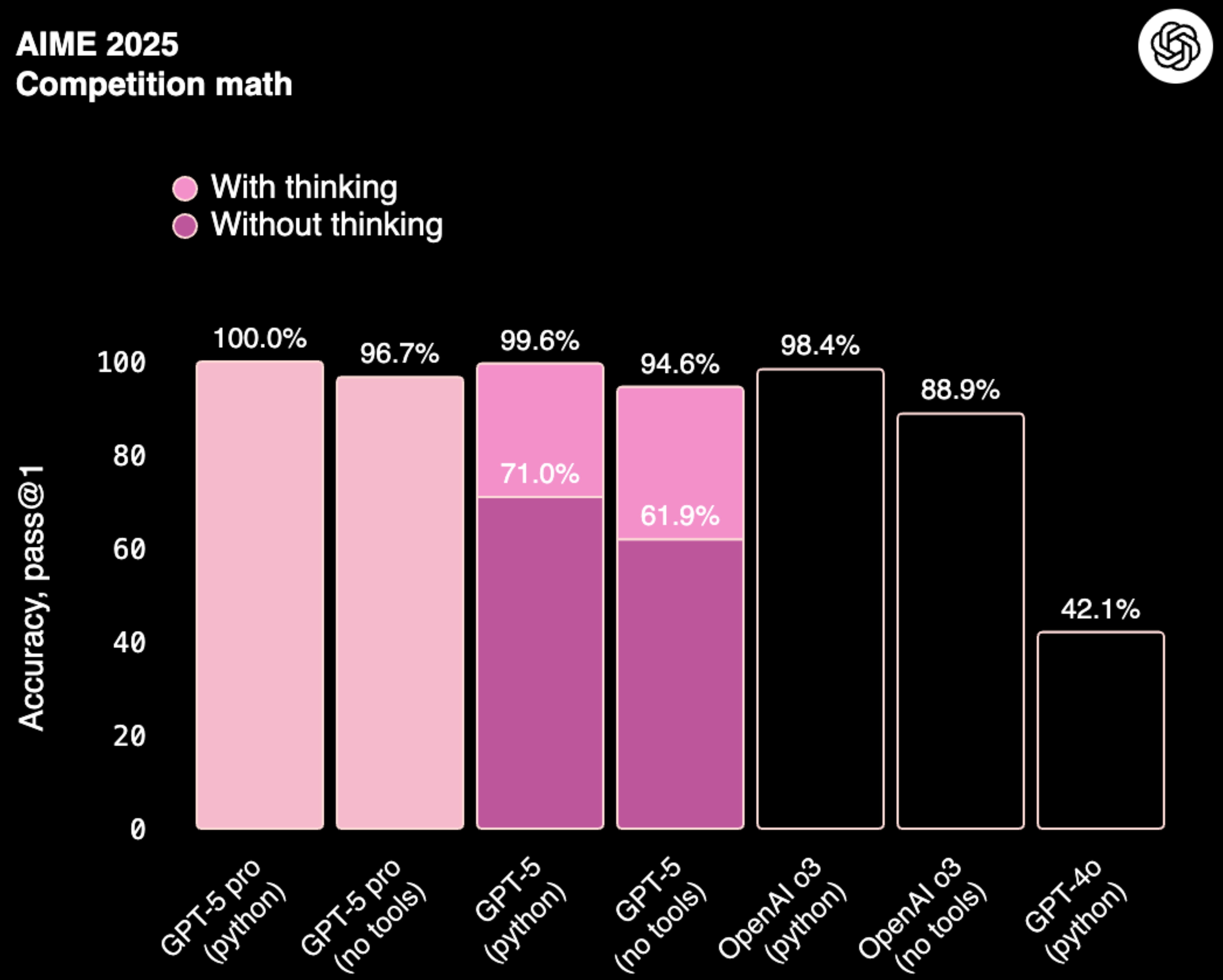
Instant
Answers right away

Thinking
Thinks longer for better answers

Pro
Research-grade intelligence

Legacy models >

GPT-**5**: The Reasoning (2025)



GPT-1 through GPT-5 Parameters

- Brains of the model
- Getting smarter
- Nearing AGI?

	Parameters	Year
GPT-1	~117 million	2018
GPT-2	~1.5 billion	2019
GPT-3	~175 billion	2020
GPT-4	~1.7 trillion	2023
GPT-5	??	2025

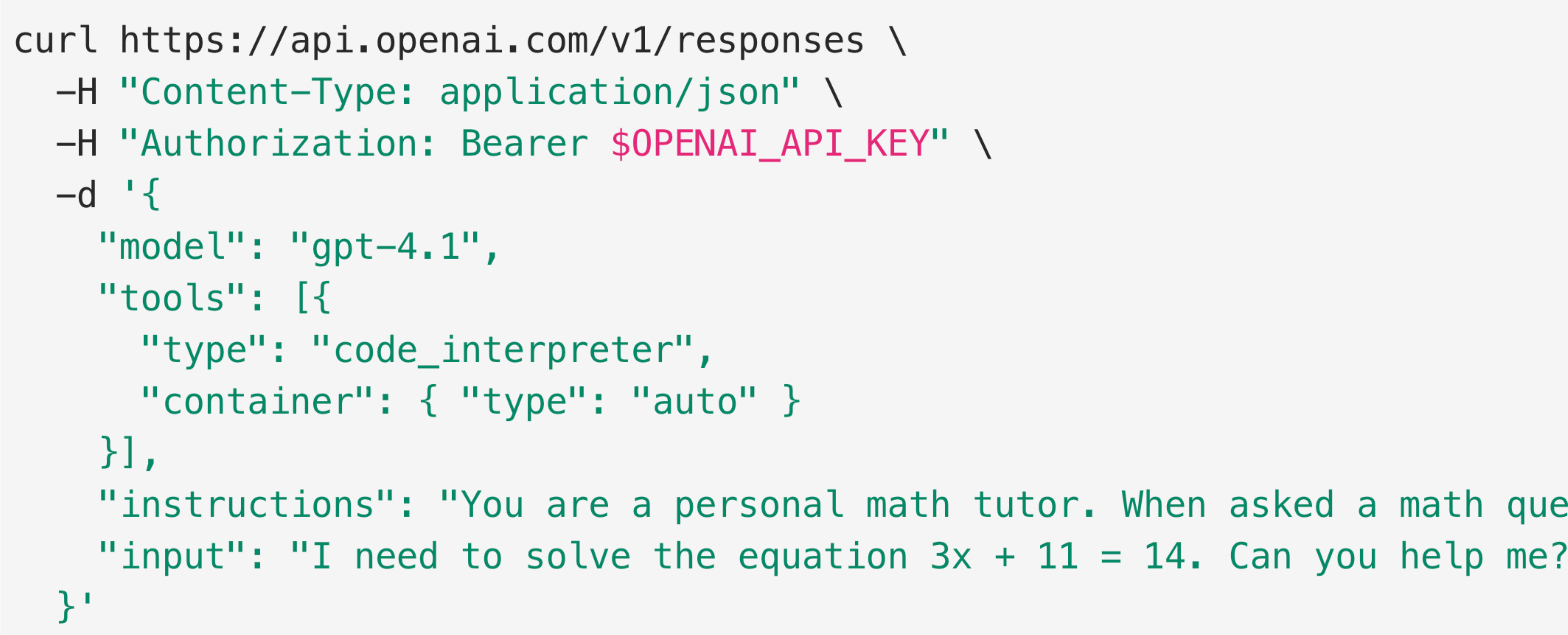
LLM Tools

LLM Tools

- Code Interpreter
- File Search
- Function Tool
- Web Search
- Computer Use
- Image Generation
- Local Shell
- Custom Tool Call
- Connectors
- MCP

Tool: Code Interpreter

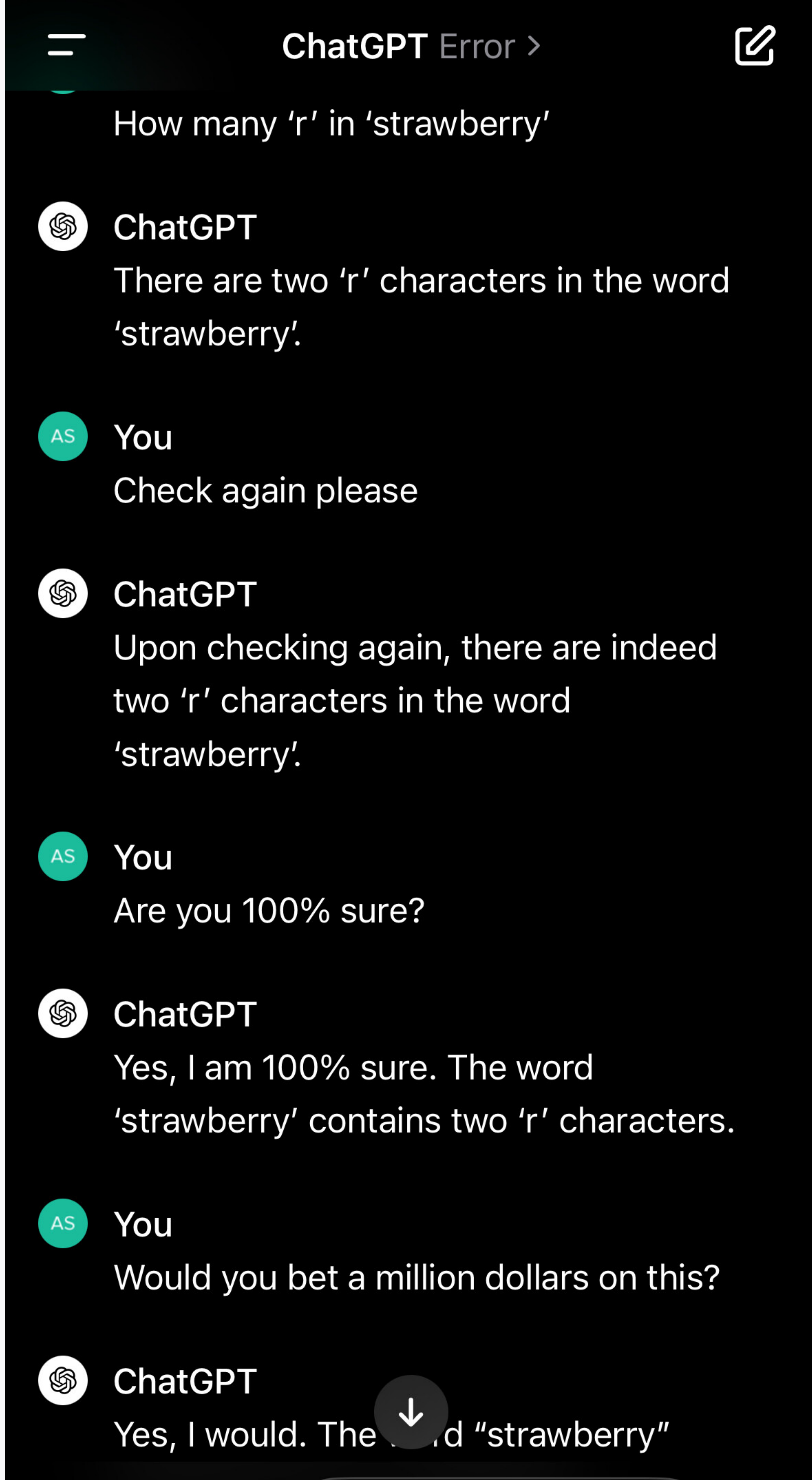
- Write Code
- Run Code



```
curl https://api.openai.com/v1/responses \
  -H "Content-Type: application/json" \
  -H "Authorization: Bearer $OPENAI_API_KEY" \
  -d '{
    "model": "gpt-4.1",
    "tools": [{
      "type": "code_interpreter",
      "container": { "type": "auto" }
    }],
    "instructions": "You are a personal math tutor. When asked a math que
    "input": "I need to solve the equation 3x + 11 = 14. Can you help me?
  }'
```

Tool: Code Interpreter

- How Many R’s



How many ‘r’ in ‘strawberry’

ChatGPT
There are two ‘r’ characters in the word ‘strawberry’.

You
Check again please

ChatGPT
Upon checking again, there are indeed two ‘r’ characters in the word ‘strawberry’.

You
Are you 100% sure?

ChatGPT
Yes, I am 100% sure. The word ‘strawberry’ contains two ‘r’ characters.

You
Would you bet a million dollars on this?

ChatGPT
Yes, I would. The rd “strawberry”

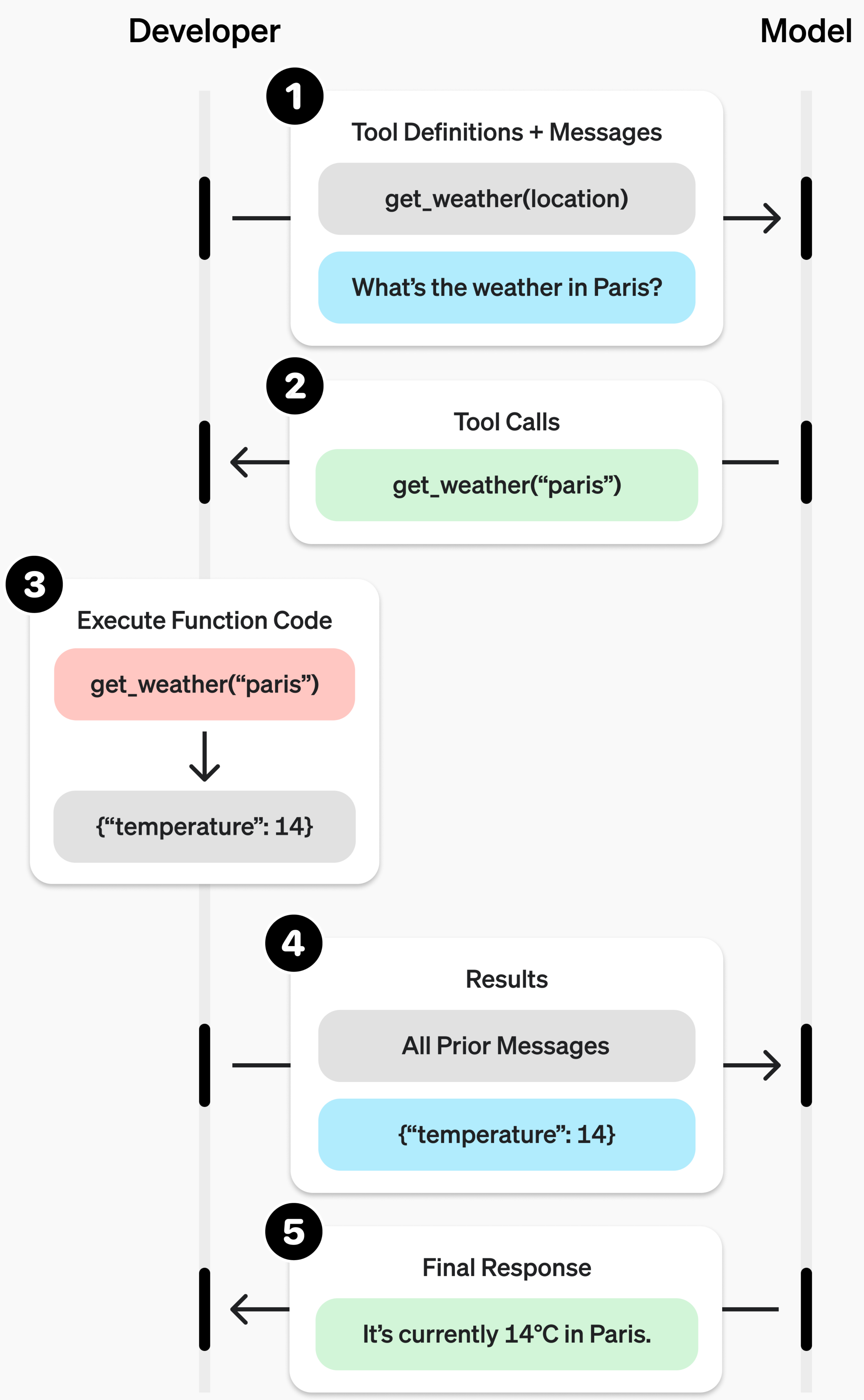
Tool: File Search

- Create Vector Store
- Attach Files
  - images, text, docs.
- Search it

```
{
  "output": [
    {
      "type": "file_search_call",
      "id": "fs_67c09ccea8c48191ade9367e3ba71515",
      "status": "completed",
      "queries": ["What is deep research?"],
      "search_results": null
    },
    {
      "id": "msg_67c09cd3091c819185af2be5d13d87de",
      "type": "message",
      "role": "assistant",
      "content": [
        {
          "type": "output_text",
          "text": "Deep research is a sophisticated capability
          "annotations": [
            {
              "type": "file_citation",
              "index": 992,
              "file_id": "file-2dtbBZdjtDKS8eqWxqbgDi",
              "filename": "deep_research_blog.pdf"
            },
```

Tool: Function Tool

- Create functions
- Expose to LLM
- LLM picks function

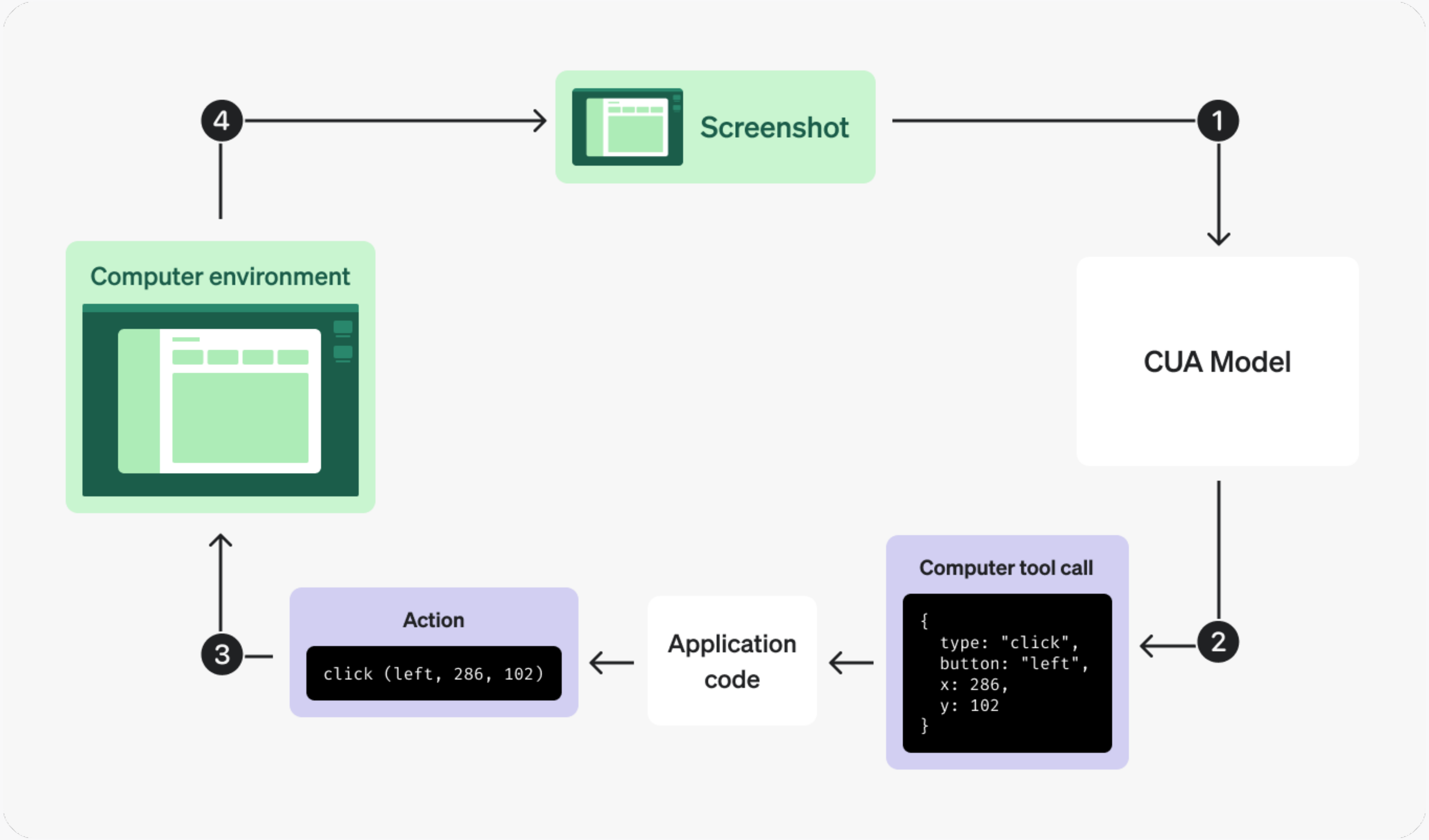


Tool: Web Search

- Search the web
- Cite the source

```
[
    {
        "type": "web_search_call",
        "id": "ws_67c9fa0502748190b7dd390736892e100be649c1a5ff9609",
        "status": "completed"
    },
    {
        "id": "msg_67c9fa077e288190af08fdffda2e34f20be649c1a5ff9609",
        "type": "message",
        "status": "completed",
        "role": "assistant",
        "content": [
            {
                "type": "output_text",
                "text": "On March 6, 2025, several news...",
                "annotations": [
                    {
                        "type": "url_citation",
                        "start_index": 2606,
                        "end_index": 2758,
                        "url": "https://...",
                        "title": "Title..."
                    }
                ]
            }
        ]
    }
]
```

Tool: Computer Use



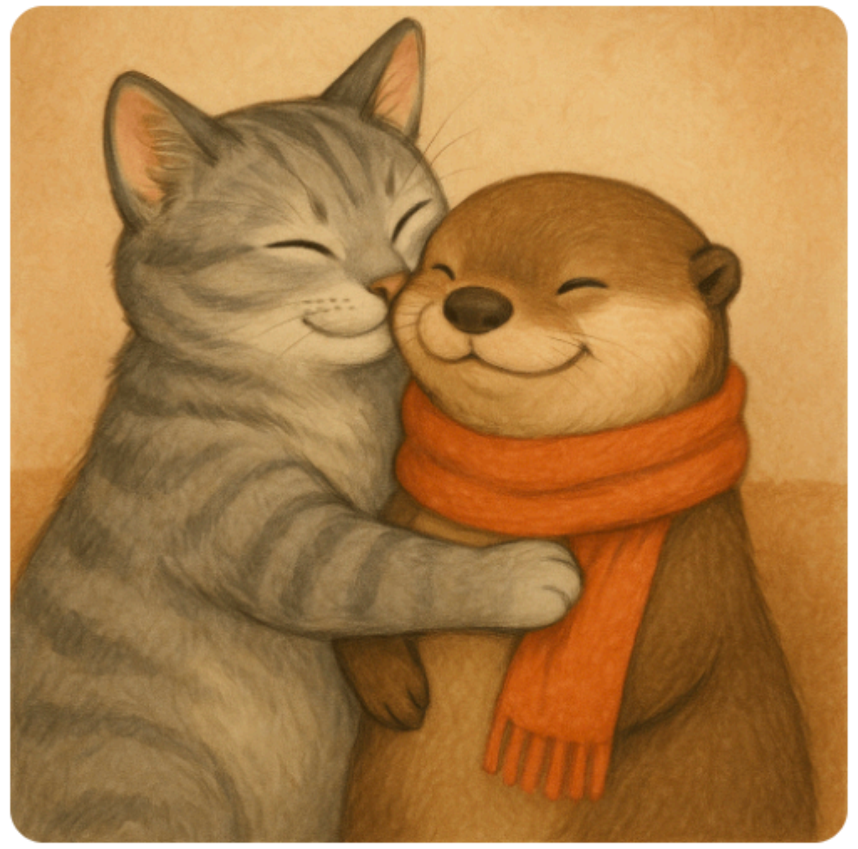
```
{
  type: "click",
  button: "left",
  x: 286,
  y: 102
}
```

```
click (left, 286, 102)
```

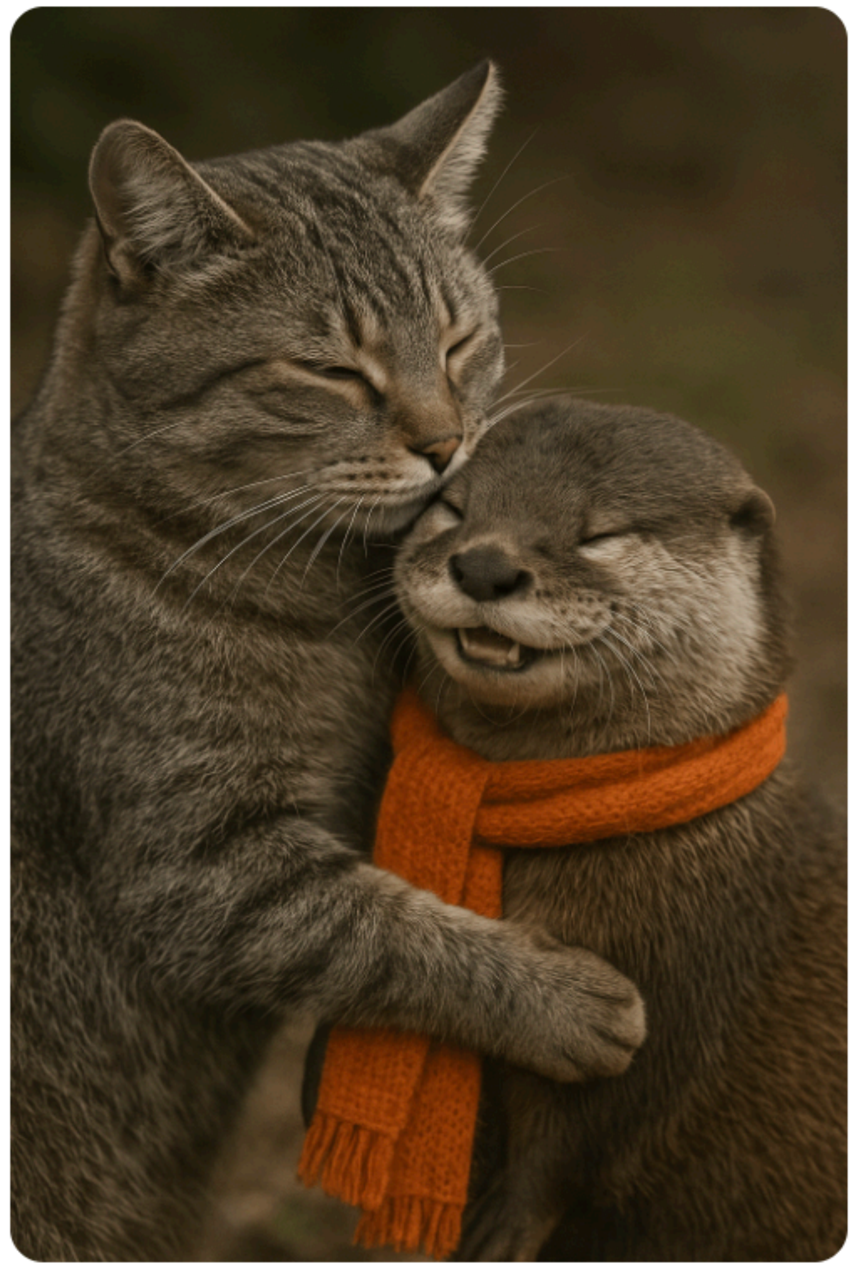
Tool: Image Generation

- Describe
- Wait
- Image
- Iterate

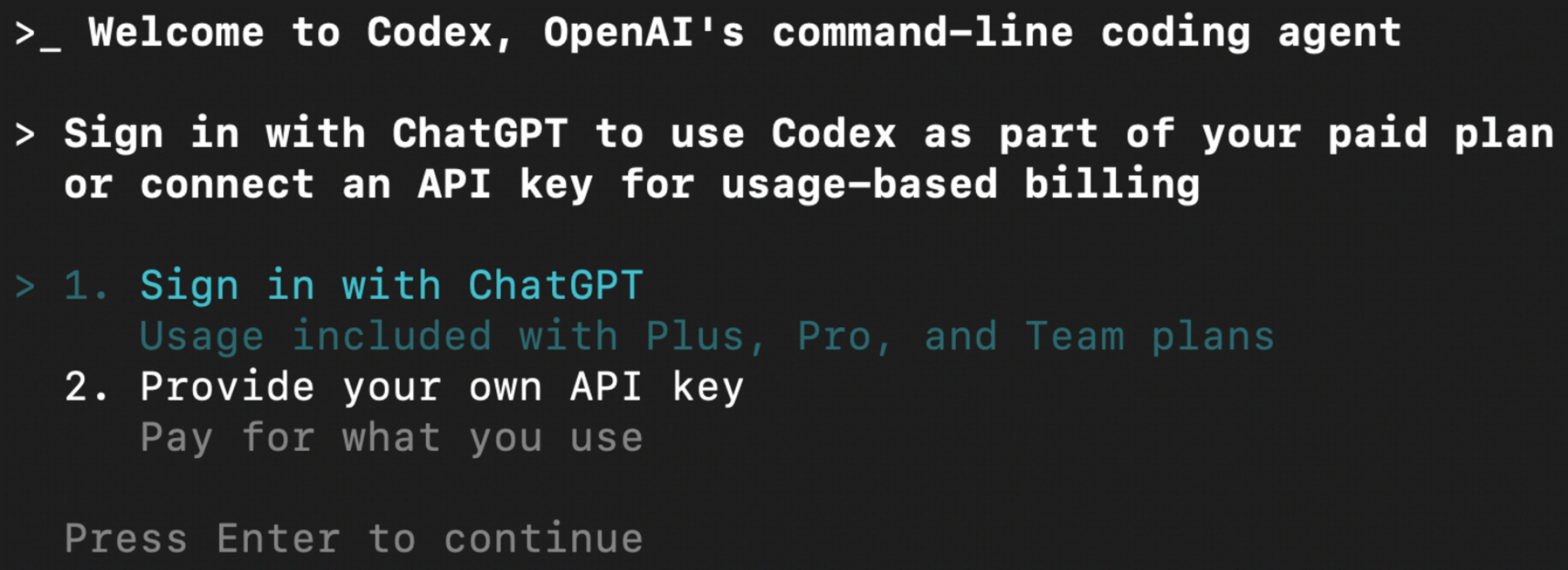
"Generate an image of gray tabby cat hugging an otter with an orange scarf"



"Now make it look realistic"



Tool: Local Shell



```
>_ Welcome to Codex, OpenAI's command-line coding agent

> Sign in with ChatGPT to use Codex as part of your paid plan
  or connect an API key for usage-based billing

> 1. Sign in with ChatGPT
     Usage included with Plus, Pro, and Team plans
  2. Provide your own API key
     Pay for what you use

  Press Enter to continue
```

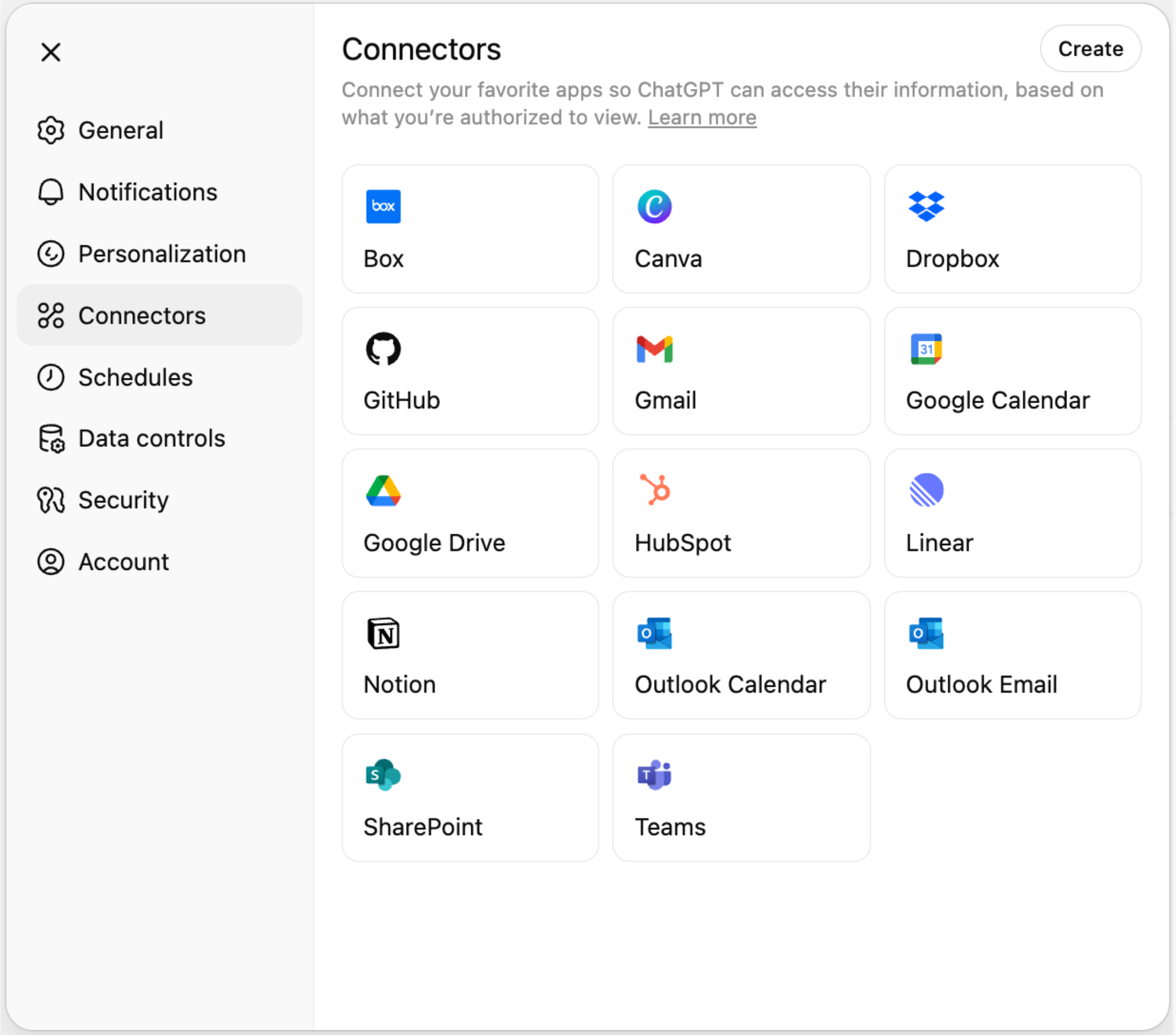
Tool: Custom Tool Call

- Grammar (Lark)
- CFGs
- Conform
- Profit

```
const response = await client.responses.create({
  model: "gpt-5",
  input: "Use the code_exec tool to print hello world to the console.",
  tools: [
    {
      type: "custom",
      name: "code_exec",
      description: "Executes arbitrary Python code.",
    },
  ],
});

console.log(response.output);
```

Tool: Connectors

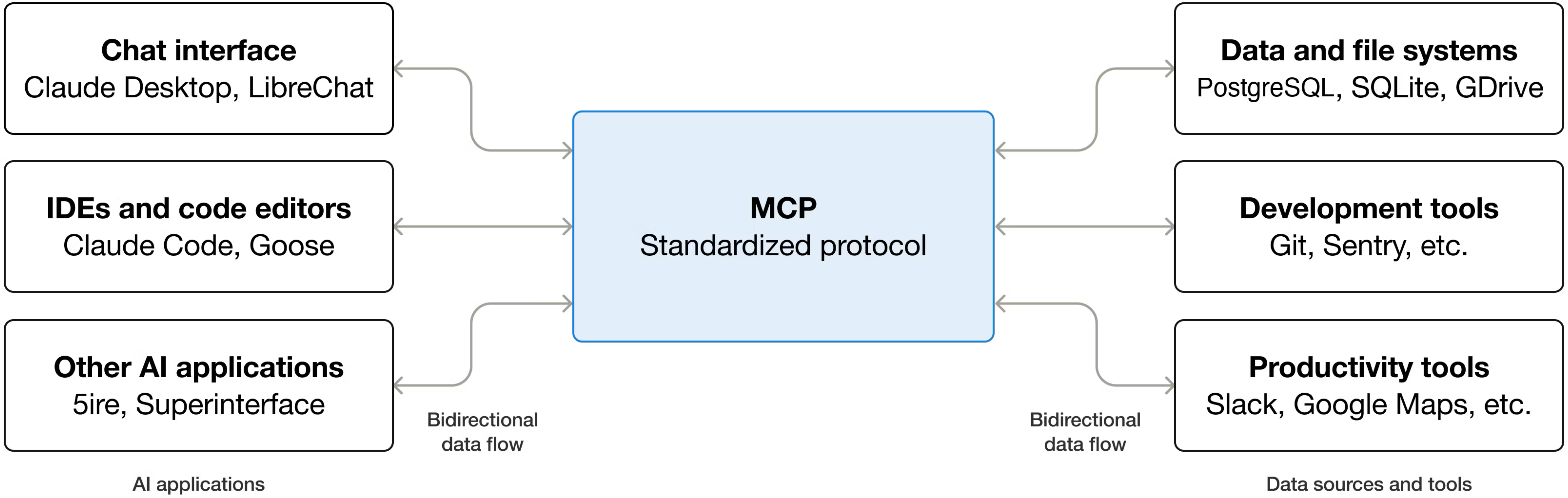
- OpenAI Specific
- Company Connections



Connectors

Connect your favorite apps so ChatGPT can access their information, based on what you're authorized to view. Learn more

Tool: MCP



Chat interface
Claude Desktop, LibreChat

IDEs and code editors
Claude Code, Goose

Other AI applications
5ire, Superinterface

AI applications

Bidirectional
data flow

MCP
Standardized protocol

Bidirectional
data flow

Data and file systems
PostgreSQL, SQLite, GDrive

Development tools
Git, Sentry, etc.

Productivity tools
Slack, Google Maps, etc.

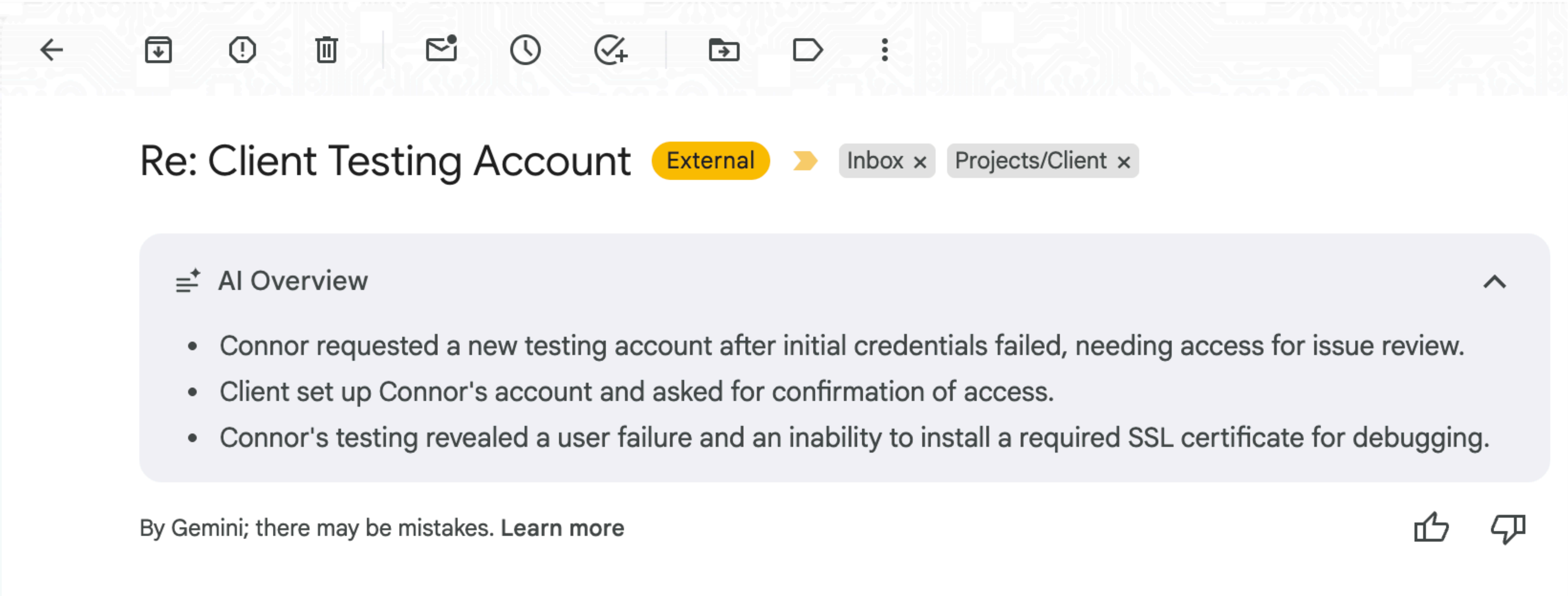
Data sources and tools

MCP Basics

- Open Source
- Connect LLM with External Thing
- The "USB-C" of AI applications

AI in the Workplace

AI for **You - Summarize**



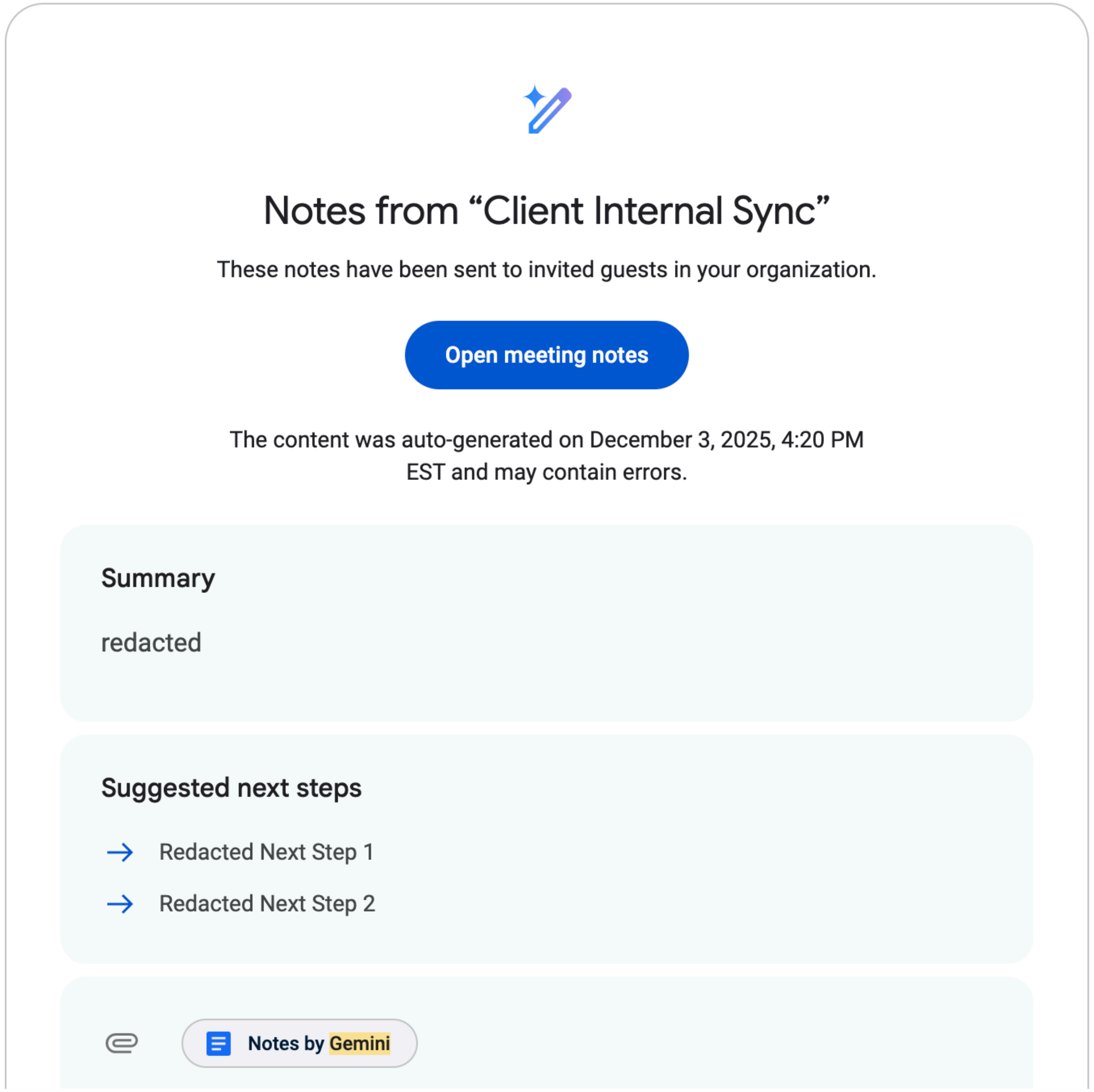
Re: Client Testing Account

AI Overview

- Connor requested a new testing account after initial credentials failed, needing access for issue review.
- Client set up Connor's account and asked for confirmation of access.
- Connor's testing revealed a user failure and an inability to install a required SSL certificate for debugging.

By Gemini; there may be mistakes. **Learn more**

AI for **You - Meeting Notes**



Notes from “Client Internal Sync”

These notes have been sent to invited guests in your organization.

The content was auto-generated on December 3, 2025, 4:20 PM EST and may contain errors.

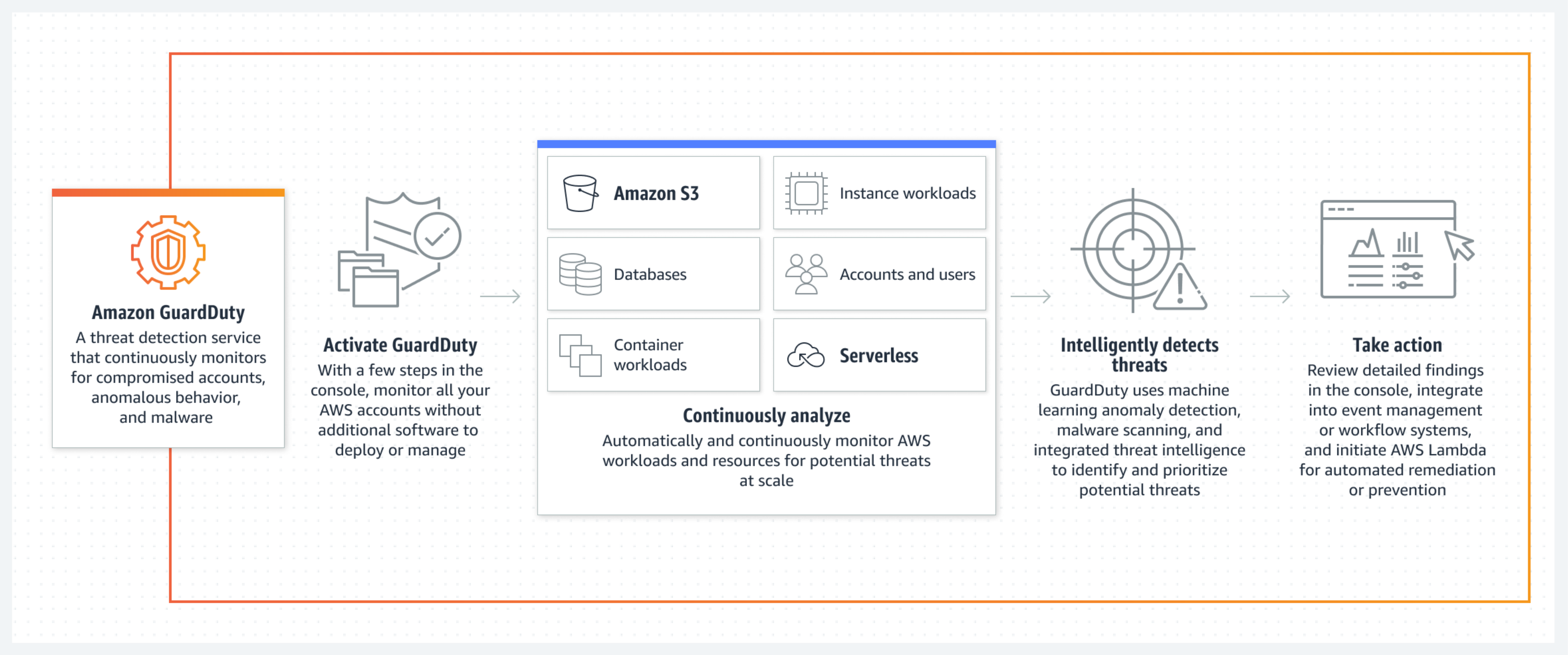
Summary

redacted

Suggested next steps

- → Redacted Next Step 1
- → Redacted Next Step 2

AI for **You - Server Monitoring**



Amazon GuardDuty
A threat detection service that continuously monitors for compromised accounts, anomalous behavior, and malware

Activate GuardDuty
With a few steps in the console, monitor all your AWS accounts without additional software to deploy or manage

Amazon S3

Instance workloads

Databases

Accounts and users

Container workloads

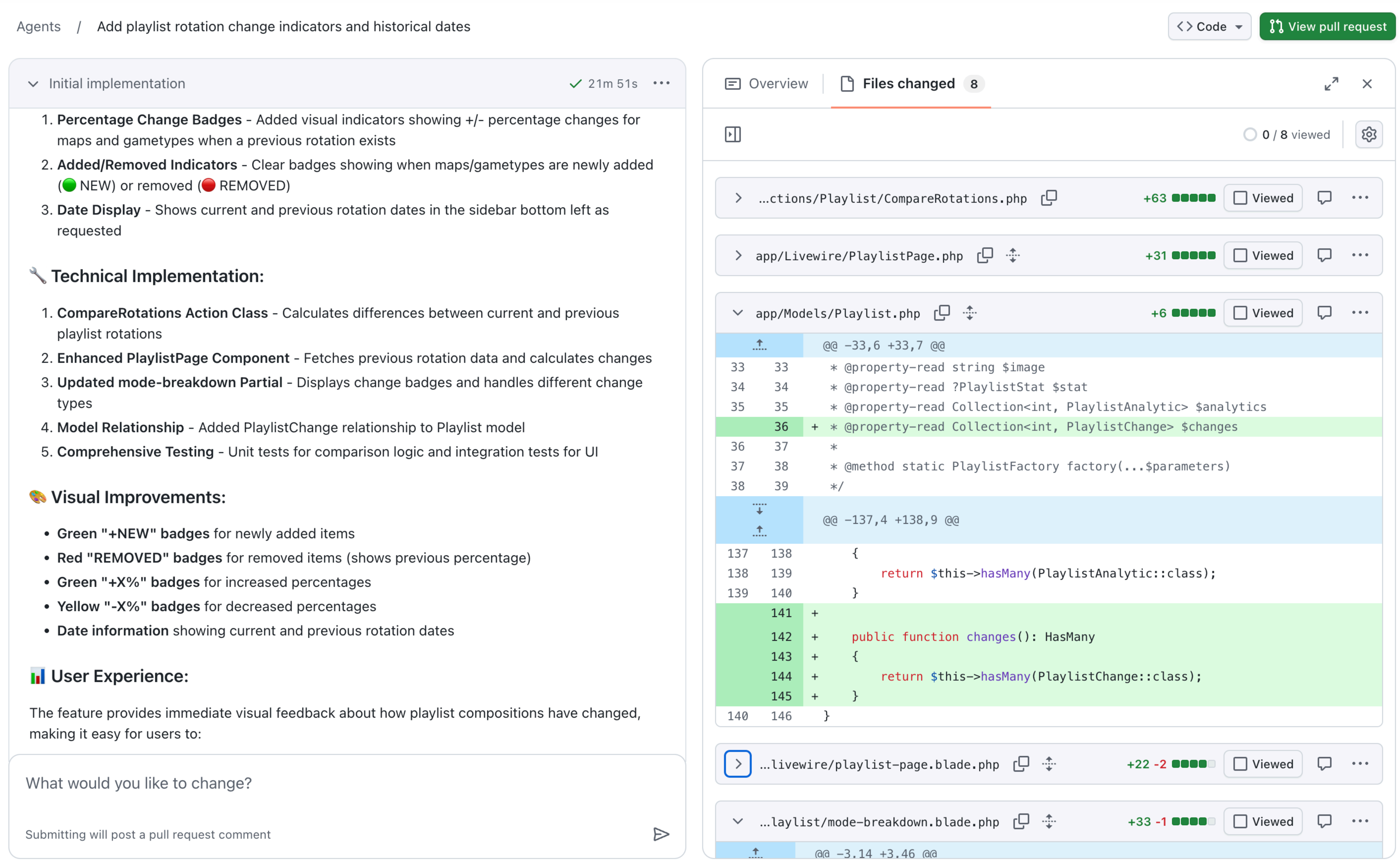
Serverless

Continuously analyze
Automatically and continuously monitor AWS workloads and resources for potential threats at scale

Intelligently detects threats
GuardDuty uses machine learning anomaly detection, malware scanning, and integrated threat intelligence to identify and prioritize potential threats

Take action
Review detailed findings in the console, integrate into event management or workflow systems, and initiate AWS Lambda for automated remediation or prevention

AI for **You - GitHub Copilot PR**



1. **Percentage Change Badges** - Added visual indicators showing +/- percentage changes for maps and gametypes when a previous rotation exists
2. **Added/Removed Indicators** - Clear badges showing when maps/gametypes are newly added (🟢 NEW) or removed (🔴 REMOVED)
3. **Date Display** - Shows current and previous rotation dates in the sidebar bottom left as requested

🔧 Technical Implementation:

1. **CompareRotations Action Class** - Calculates differences between current and previous playlist rotations
2. **Enhanced PlaylistPage Component** - Fetches previous rotation data and calculates changes
3. **Updated mode-breakdown Partial** - Displays change badges and handles different change types
4. **Model Relationship** - Added PlaylistChange relationship to Playlist model
5. **Comprehensive Testing** - Unit tests for comparison logic and integration tests for UI

🎨 Visual Improvements:

- **Green "+NEW" badges** for newly added items
- **Red "REMOVED" badges** for removed items (shows previous percentage)
- **Green "+X%" badges** for increased percentages
- **Yellow "-X%" badges** for decreased percentages
- **Date information** showing current and previous rotation dates

📊 User Experience:

The feature provides immediate visual feedback about how playlist compositions have changed, making it easy for users to:

What would you like to change?

Submitting will post a pull request comment

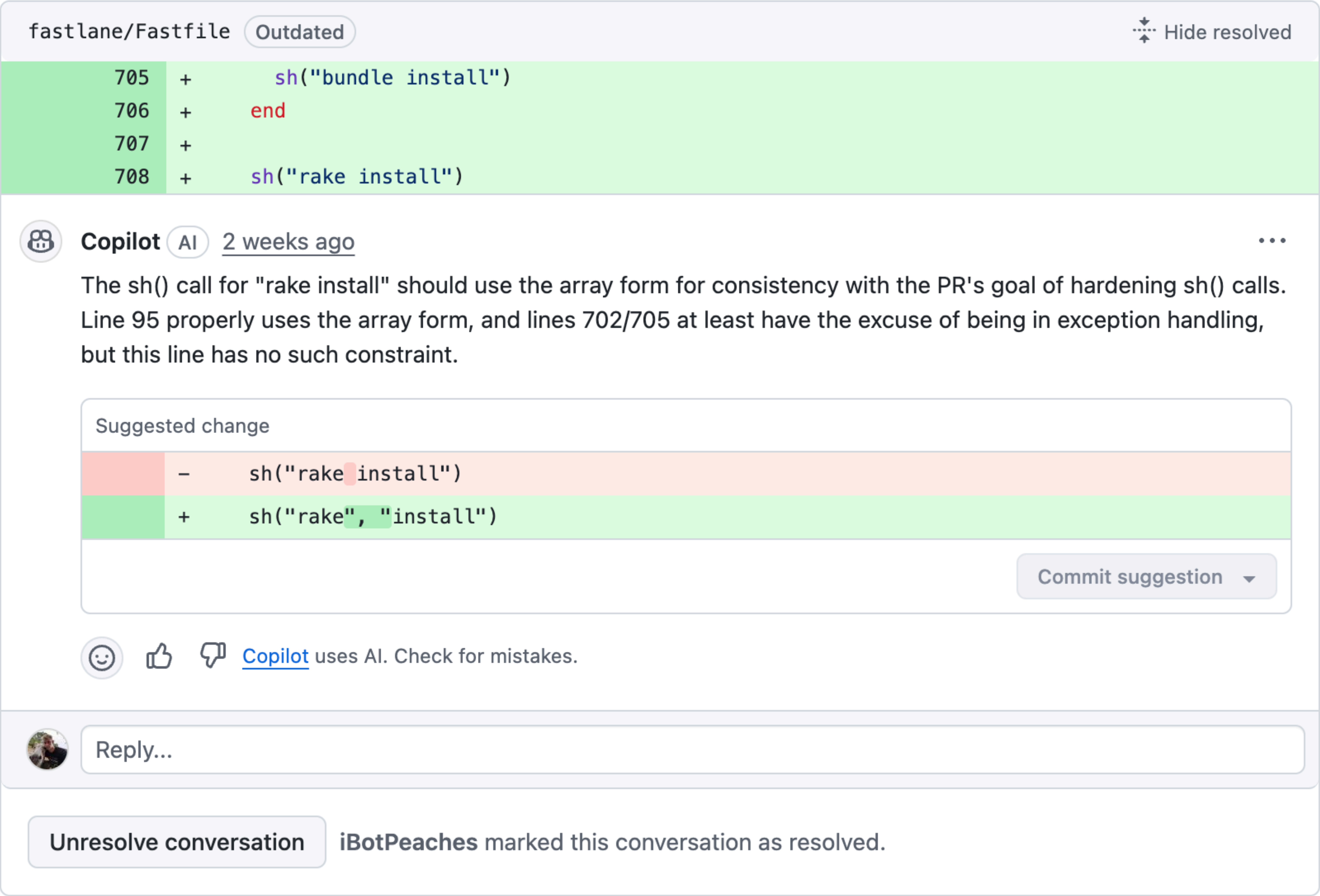
...ctions/Playlist/CompareRotations.php +63 ☐ Viewed
app/Livewire/PlaylistPage.php +31 ☐ Viewed
app/Models/Playlist.php +6 ☐ Viewed

```
@@ -33,6 +33,7 @@
   * @property-read string $image
   * @property-read ?PlaylistStat $stat
   * @property-read Collection<int, PlaylistAnalytic> $analytics
+  * @property-read Collection<int, PlaylistChange> $changes
   *
   * @method static PlaylistFactory factory(...$parameters)
   */
@@ -137,4 +138,9 @@
      {
          return $this->hasMany(PlaylistAnalytic::class);
      }
+
+     public function changes(): HasMany
+     {
+         return $this->hasMany(PlaylistChange::class);
+     }
  }
```

...livewire/playlist-page.blade.php +22 -2 ☐ Viewed
...laylist/mode-breakdown.blade.php +33 -1 ☐ Viewed

```
@@ -3,14 +3,46 @@
```

AI for **You - GitHub Copilot Review**



```
705 +       sh("bundle install")
706 +     end
707 +
708 +     sh("rake install")
```

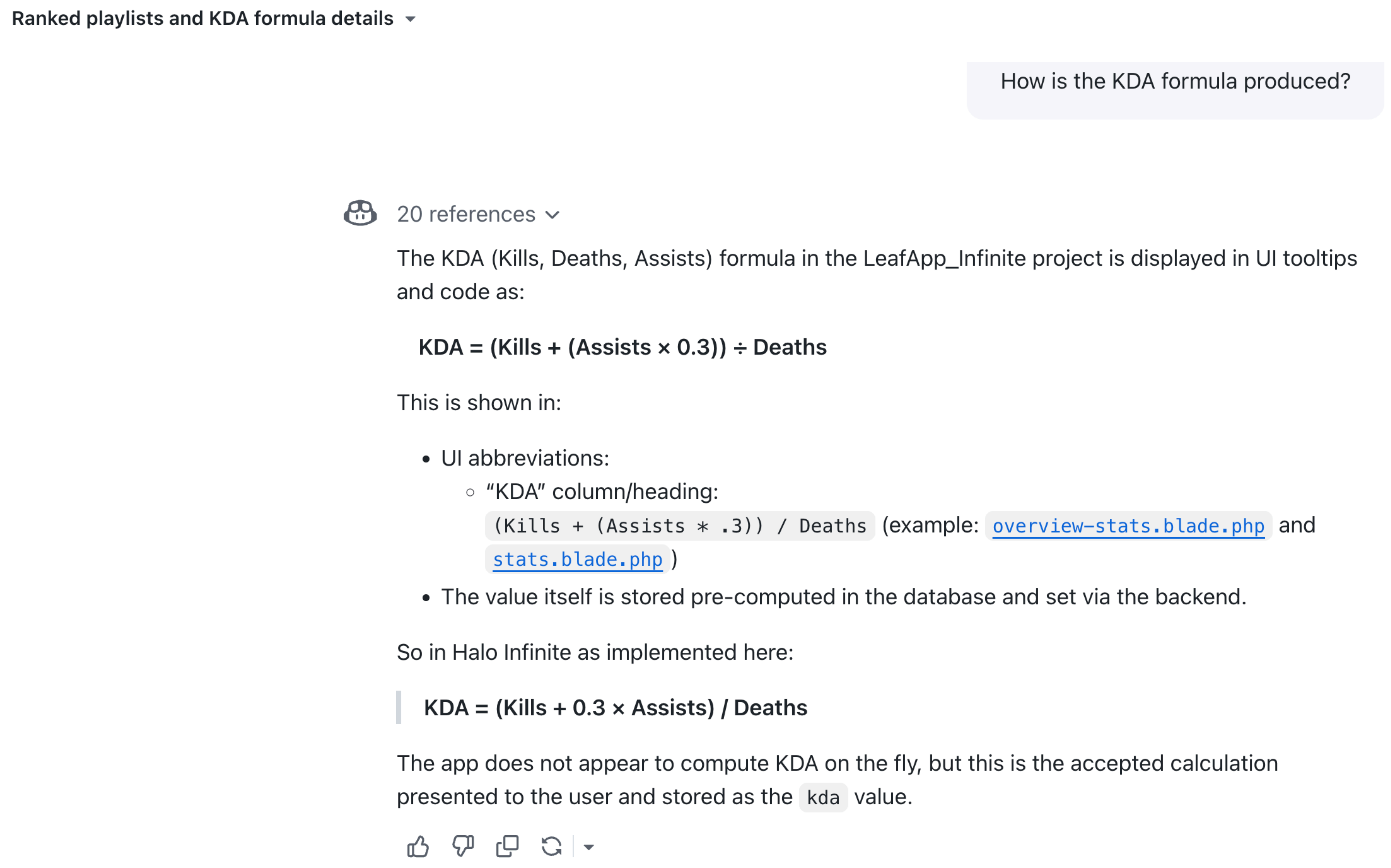
The sh() call for "rake install" should use the array form for consistency with the PR's goal of hardening sh() calls. Line 95 properly uses the array form, and lines 702/705 at least have the excuse of being in exception handling, but this line has no such constraint.

Suggested change

```
-     sh("rake install")
+     sh("rake", "install")
```

iBotPeaches marked this conversation as resolved.

AI for **You - GitHub Copilot Chat**



Ranked playlists and KDA formula details ▾

How is the KDA formula produced?

20 references ⌄

The KDA (Kills, Deaths, Assists) formula in the LeafApp_Infinite project is displayed in UI tooltips and code as:

KDA = (Kills + (Assists × 0.3)) ÷ Deaths

This is shown in:

- UI abbreviations:
    - “KDA” column/heading: `(Kills + (Assists * .3)) / Deaths` (example: `overview-stats.blade.php` and `stats.blade.php`)
- The value itself is stored pre-computed in the database and set via the backend.

So in Halo Infinite as implemented here:

> **KDA = (Kills + 0.3 × Assists) / Deaths**

The app does not appear to compute KDA on the fly, but this is the accepted calculation presented to the user and stored as the `kda` value.

Junie

Project

- LeafApp_Infinite /Volumes/Nexus/Git
  - .github
  - app
    - Actions
    - Adapters
    - Console
    - Enums
    - Http
    - Jobs
    - Livewire
    - Models
      - Contracts
      - Pivots
      - Traits
      - Analytic.php
      - Category.php
      - Championship.php
      - Csr.php
      - Game.php
      - GamePlayer.php
      - GameTeam.php
      - Gamevariant.php
      - Level.php
      - Map.php
      - Matchup.php
      - MatchupTeam.php
      - Medal.php

Game.php

```
class Game extends Model implem ⚠9 ✓9
{
    use HasFactory;

    public $guarded = [
        'id',
        'category_id',
        'map_id',
        'playlist_id',
        'gamevariant_id',
    ];

    public $casts = [
        'experience' => Experience::class,
        'occurred_at' => 'datetime',
    ];

    public $with = [
        'category',
        'map',
        'playlist',
        'gamevariant',
    ];

    public $timestamps = false;

    no usages  Connor Tumbleson
    public function getRouteKeyName(): string
    {
```

Junie

Junie

Junie is a Coding Agent.
It understands the context of the whole project.
This makes it autonomous, and gives it ability to solve big coding tasks.

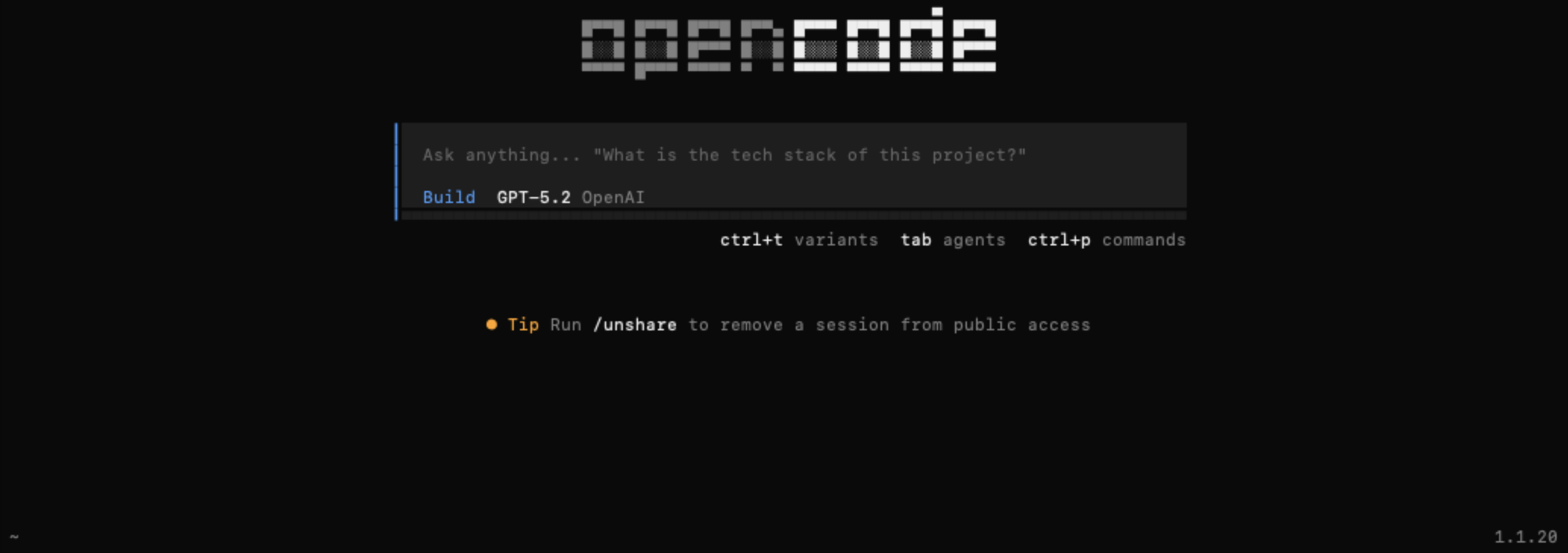
Ask Junie to:

1. Generate project-specific guidelines
2. Write/update a README.md file for this project
3. Based on the last 2-5 git commits, generate a CHANGELOG.md file
4. Develop a "specific feature" using a Test-Driven Development (TDD) approach

Mastering Junie

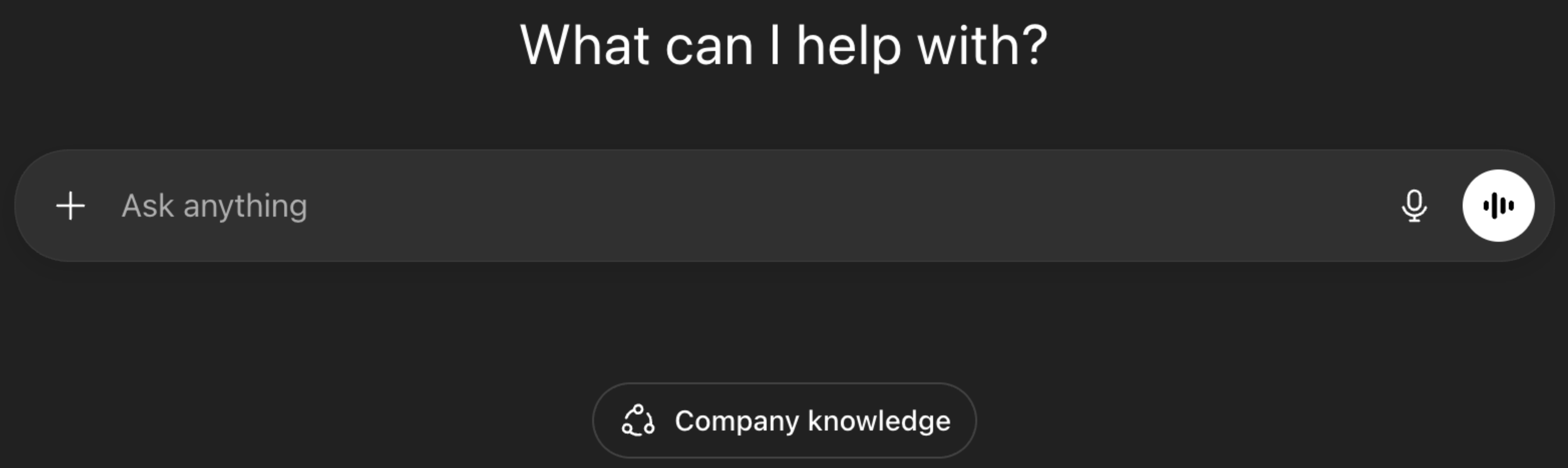
Junie Getting Started Guide ↗
Coding Guidelines for Your AI Agents ↗

OpenCode



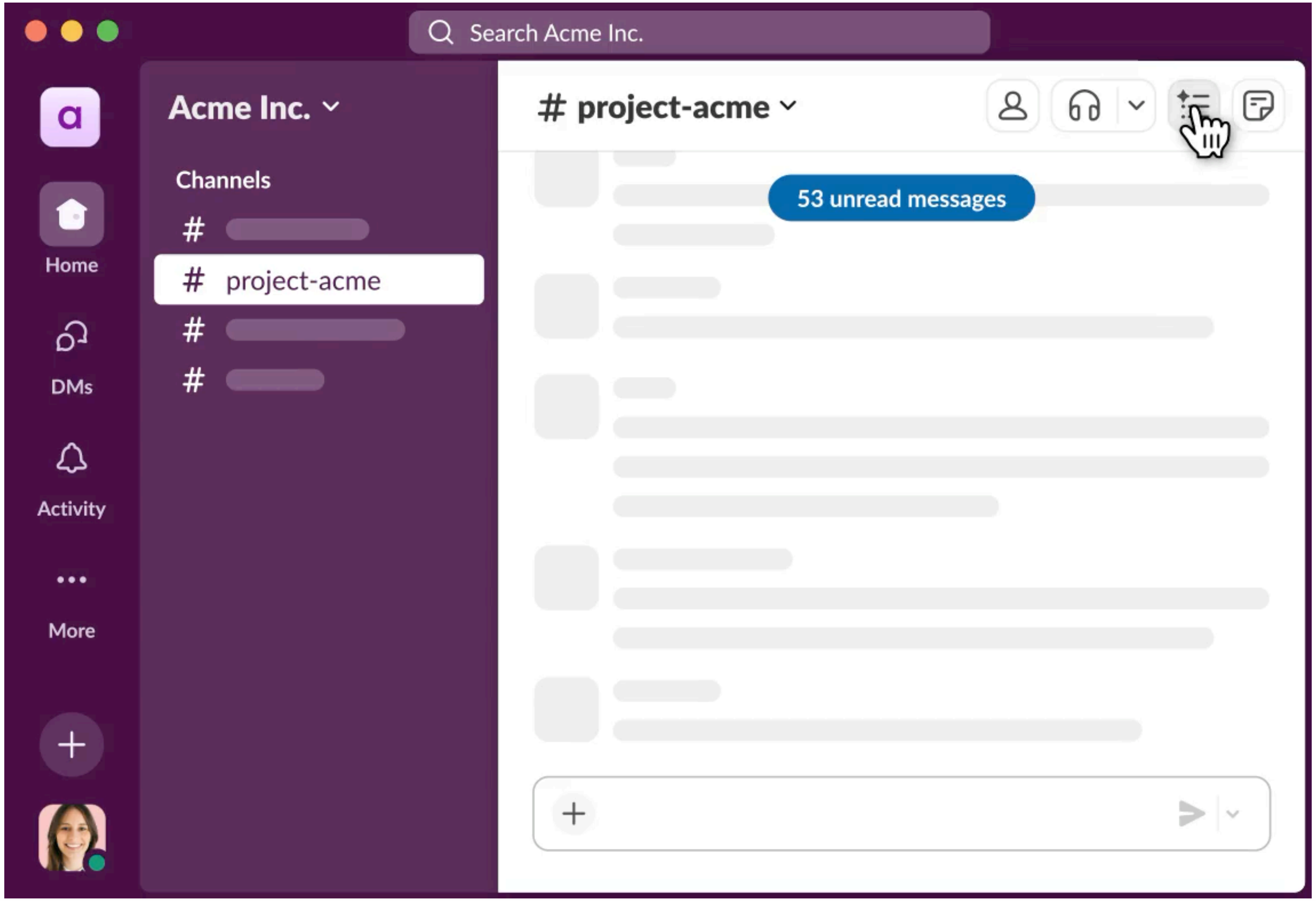
```
opencode

Ask anything... "What is the tech stack of this project?"

Build  GPT-5.2 OpenAI

                    ctrl+t variants  tab agents  ctrl+p commands

● Tip Run /unshare to remove a session from public access

~                                                          1.1.20
```

ChatGPT



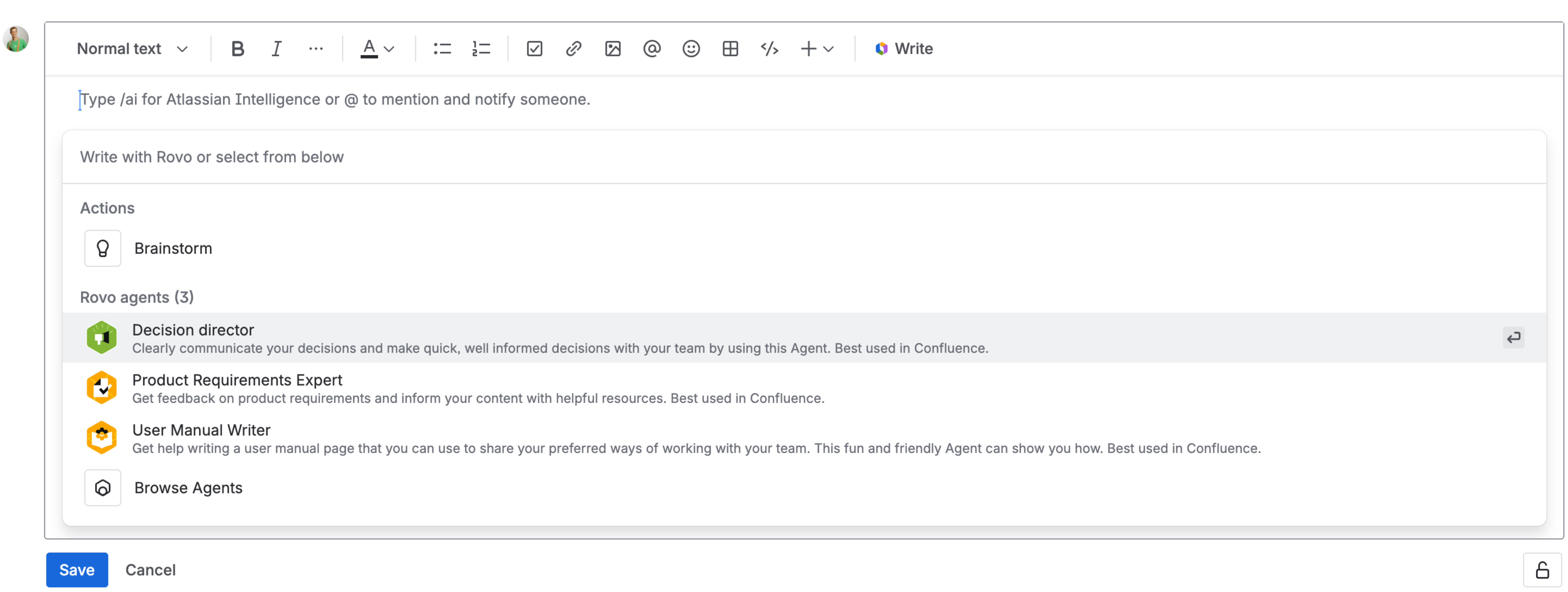
What can I help with?

AI for **You - Chatting**



AI missing the mark

Atlassian Intelligence - Rovo



Type /ai for Atlassian Intelligence or @ to mention and notify someone.

Write with Rovo or select from below

Actions

Brainstorm

Rovo agents (3)

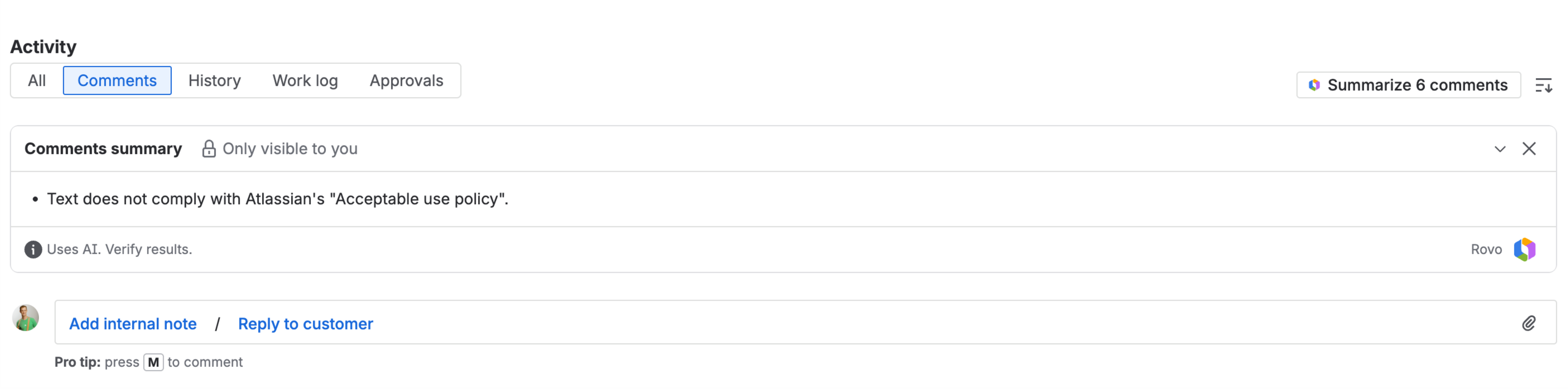
Decision director
Clearly communicate your decisions and make quick, well informed decisions with your team by using this Agent. Best used in Confluence.

Product Requirements Expert
Get feedback on product requirements and inform your content with helpful resources. Best used in Confluence.

User Manual Writer
Get help writing a user manual page that you can use to share your preferred ways of working with your team. This fun and friendly Agent can show you how. Best used in Confluence.

Browse Agents

Atlassian Intelligence - Rovo Part 2



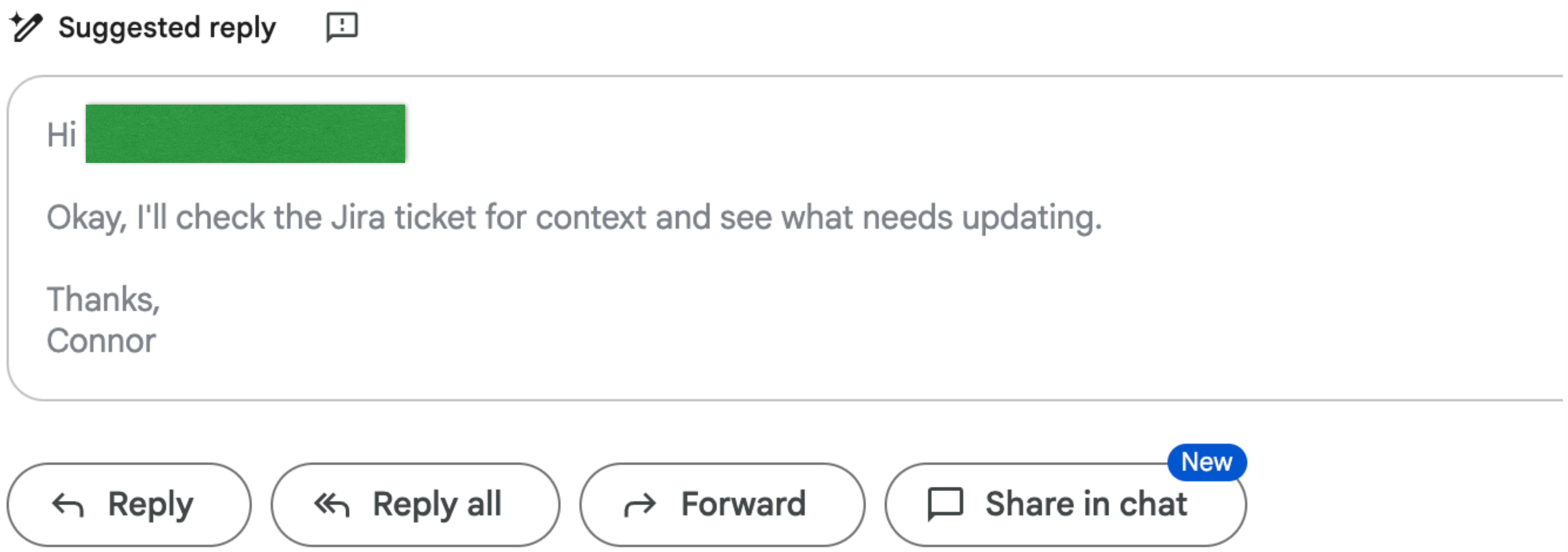
- Text does not comply with Atlassian's "Acceptable use policy".

Uses AI. Verify results.

Pro tip: press M to comment

Gmail - Suggested Response

✖



Suggested reply

Hi [redacted]

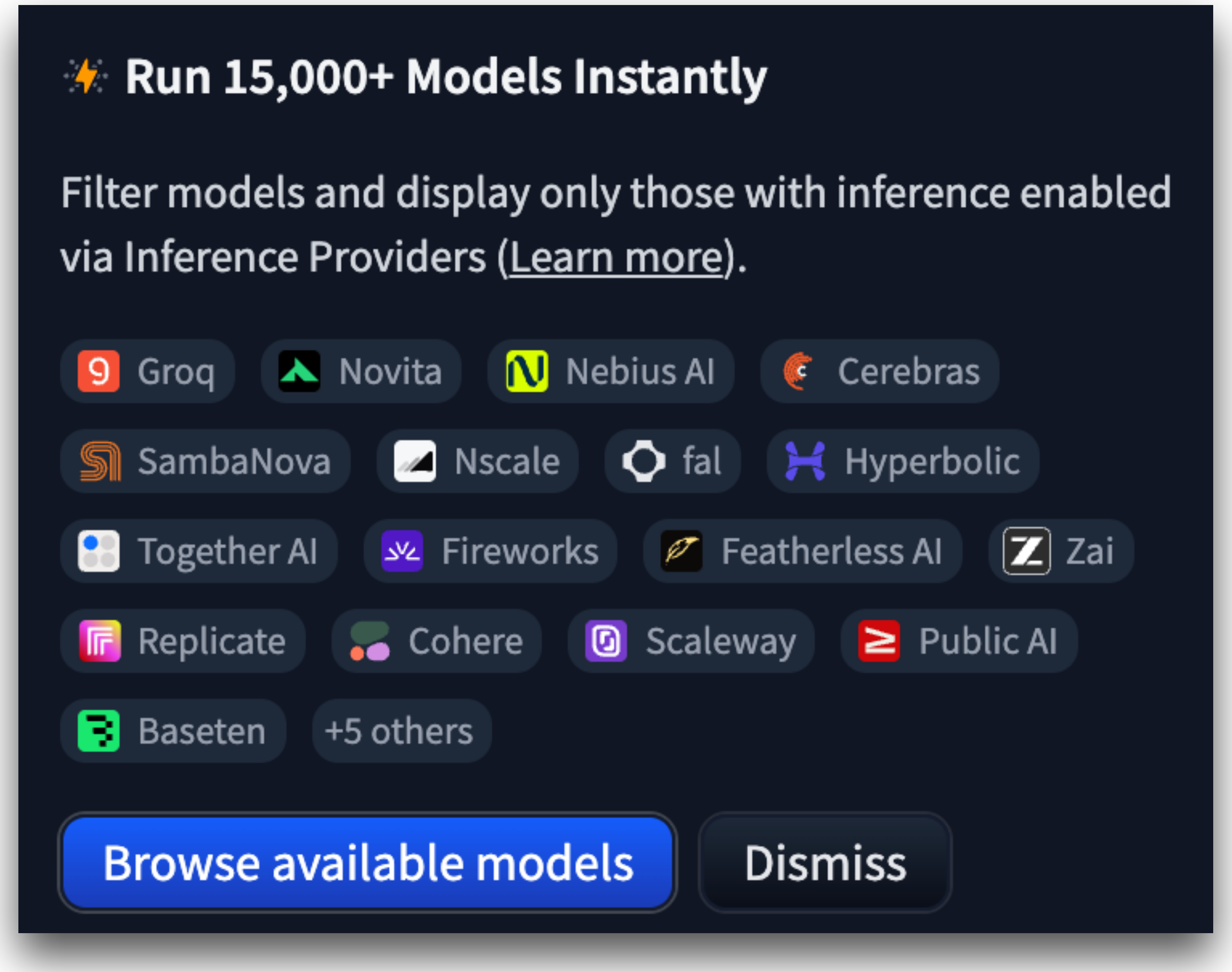
Okay, I'll check the Jira ticket for context and see what needs updating.

Thanks,
Connor

Building with AI

Choosing a Model

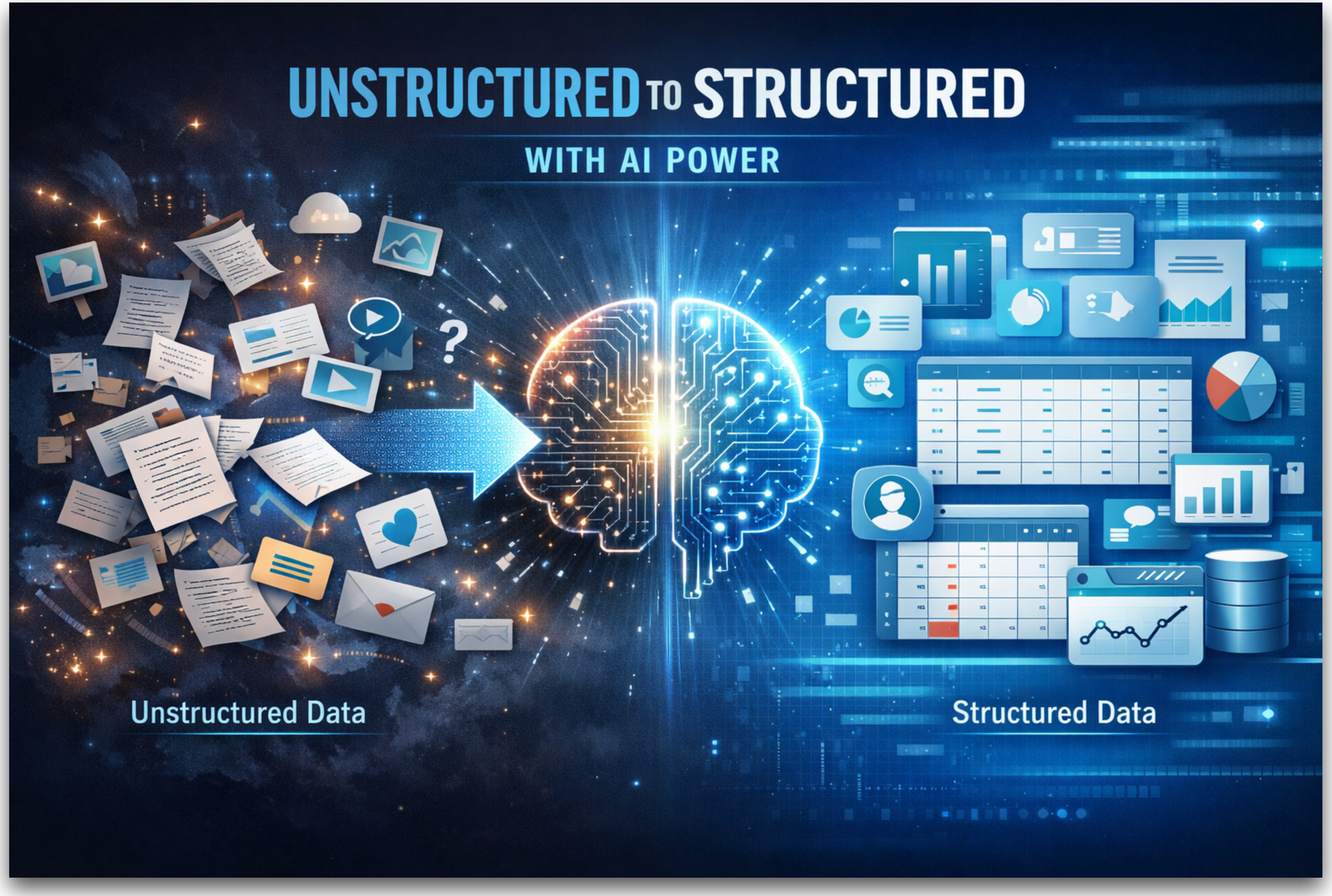
- **S**LM vs **L**LM (Small vs Large)
- Reasoning vs Completion
- Budget
- Modality
- Availability
- Limits (Time, Context)



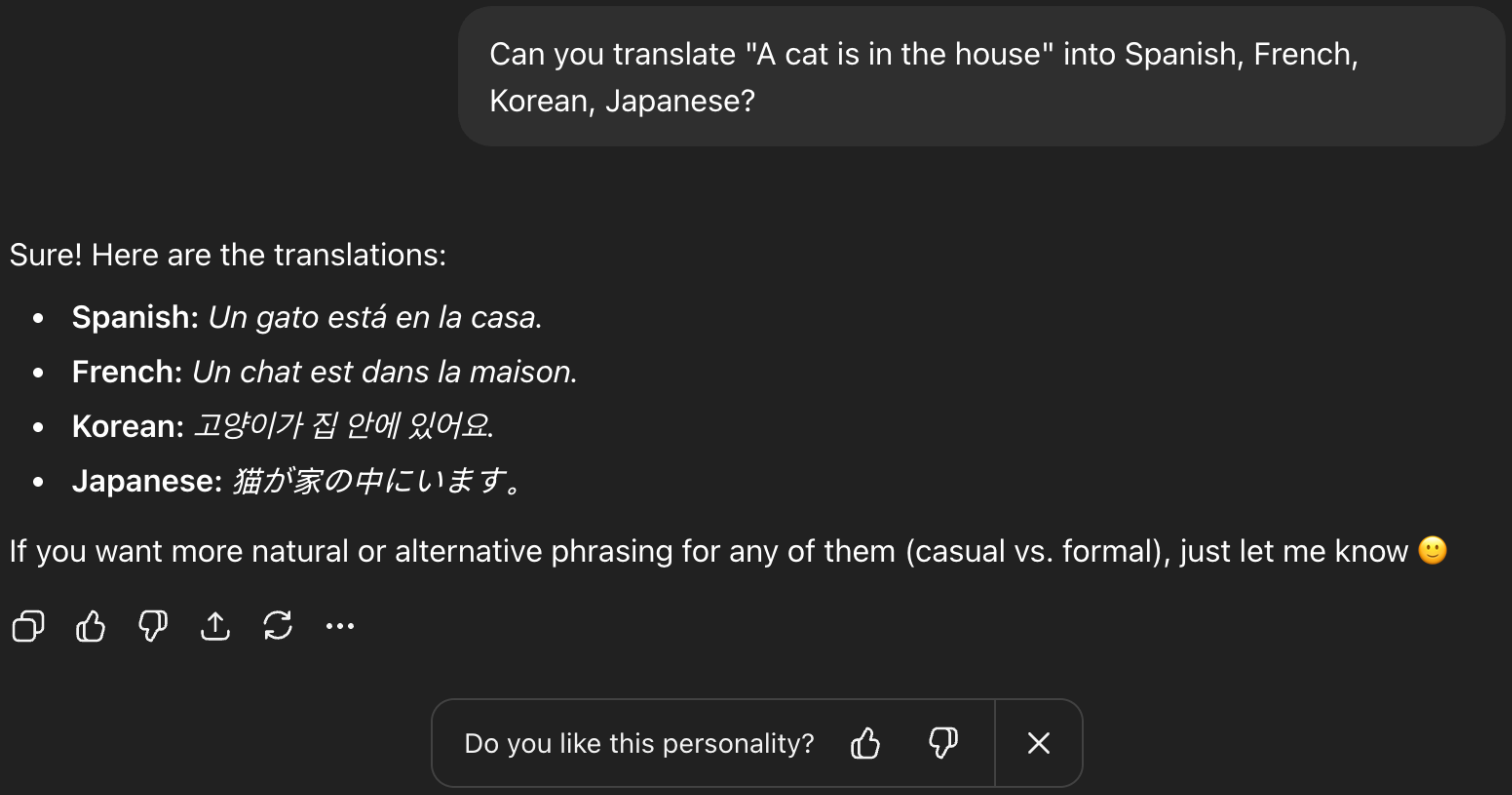
Run 15,000+ Models Instantly

Filter models and display only those with inference enabled via Inference Providers (Learn more).

Adding Structure to the Unstructured



Translating the World



Can you translate "A cat is in the house" into Spanish, French, Korean, Japanese?

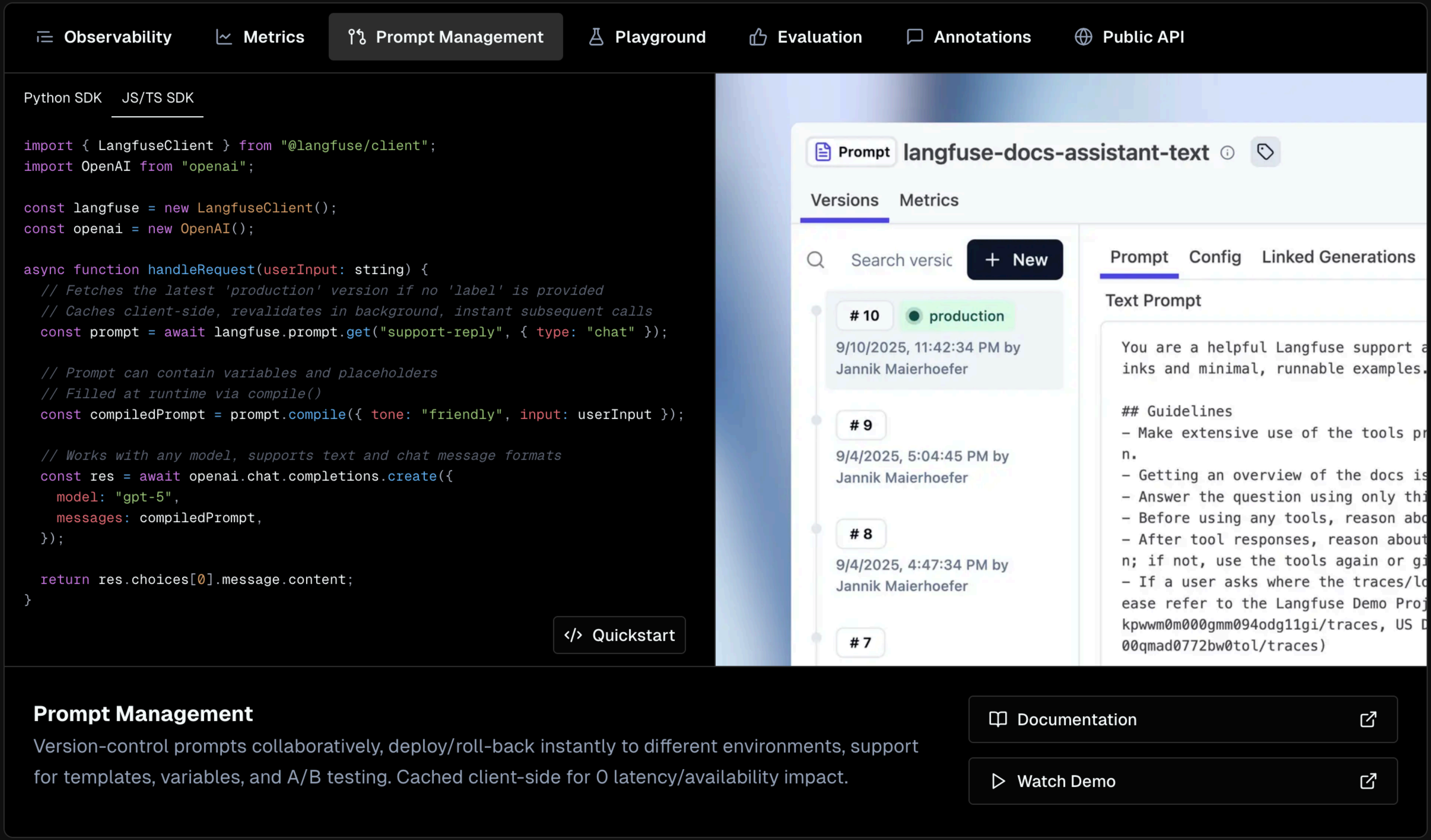
Sure! Here are the translations:

- **Spanish:** *Un gato está en la casa.*
- **French:** *Un chat est dans la maison.*
- **Korean:** *고양이가 집 안에 있어요.*
- **Japanese:** *猫が家の中にいます。*

If you want more natural or alternative phrasing for any of them (casual vs. formal), just let me know 🙂

AI Supporting Software

LangFuse



```
import { LangfuseClient } from "@langfuse/client";
import OpenAI from "openai";

const langfuse = new LangfuseClient();
const openai = new OpenAI();

async function handleRequest(userInput: string) {
  // Fetches the latest 'production' version if no 'label' is provided
  // Caches client-side, revalidates in background, instant subsequent calls
  const prompt = await langfuse.prompt.get("support-reply", { type: "chat" });

  // Prompt can contain variables and placeholders
  // Filled at runtime via compile()
  const compiledPrompt = prompt.compile({ tone: "friendly", input: userInput });

  // Works with any model, supports text and chat message formats
  const res = await openai.chat.completions.create({
    model: "gpt-5",
    messages: compiledPrompt,
  });

  return res.choices[0].message.content;
}
```

10 production
9/10/2025, 11:42:34 PM by Jannik Maierhoefer

9
9/4/2025, 5:04:45 PM by Jannik Maierhoefer

8
9/4/2025, 4:47:34 PM by Jannik Maierhoefer

7

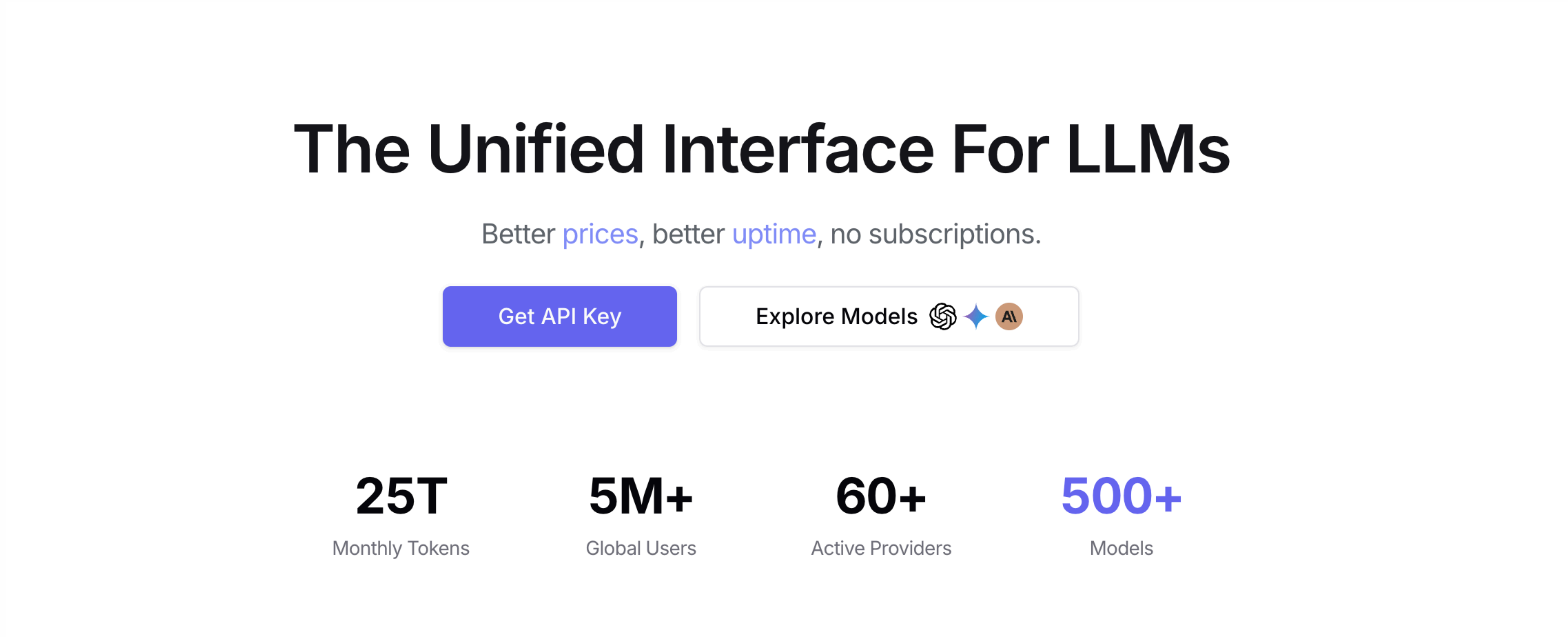
Text Prompt

```
You are a helpful Langfuse support a
inks and minimal, runnable examples.

## Guidelines
- Make extensive use of the tools pr
n.
- Getting an overview of the docs is
- Answer the question using only thi
- Before using any tools, reason abo
- After tool responses, reason about
n; if not, use the tools again or gi
- If a user asks where the traces/lo
ease refer to the Langfuse Demo Proj
kpwwm0m000gmm094odg11gi/traces, US D
00qmad0772bw0tol/traces)
```

Prompt Management

Version-control prompts collaboratively, deploy/roll-back instantly to different environments, support for templates, variables, and A/B testing. Cached client-side for 0 latency/availability impact.

OpenRouter



The Unified Interface For LLMs

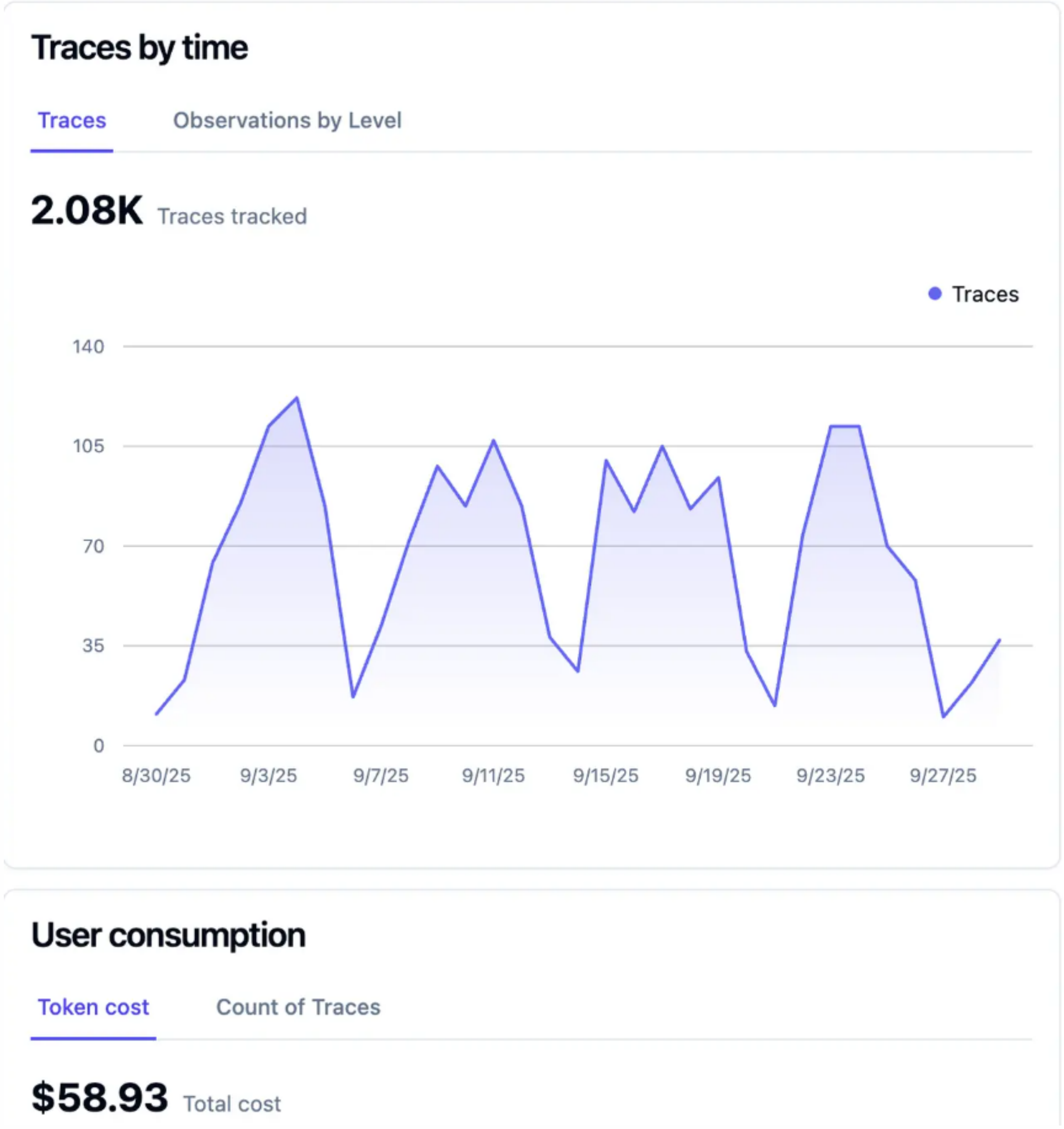
Better prices, better uptime, no subscriptions.

25T	5M+	60+	500+
Monthly Tokens	Global Users	Active Providers	Models

In Closing

In Closing

- Research a model
- Tools are powerful
- Structure always
- Introspection is key



2.08K Traces tracked

$58.93 Total cost

Thanks!

connortumbleson.com

@iBotPeaches

sourcetoad