Scaling AI @ Elementor: Fault-Tolerant GenAI Solutions for Millions

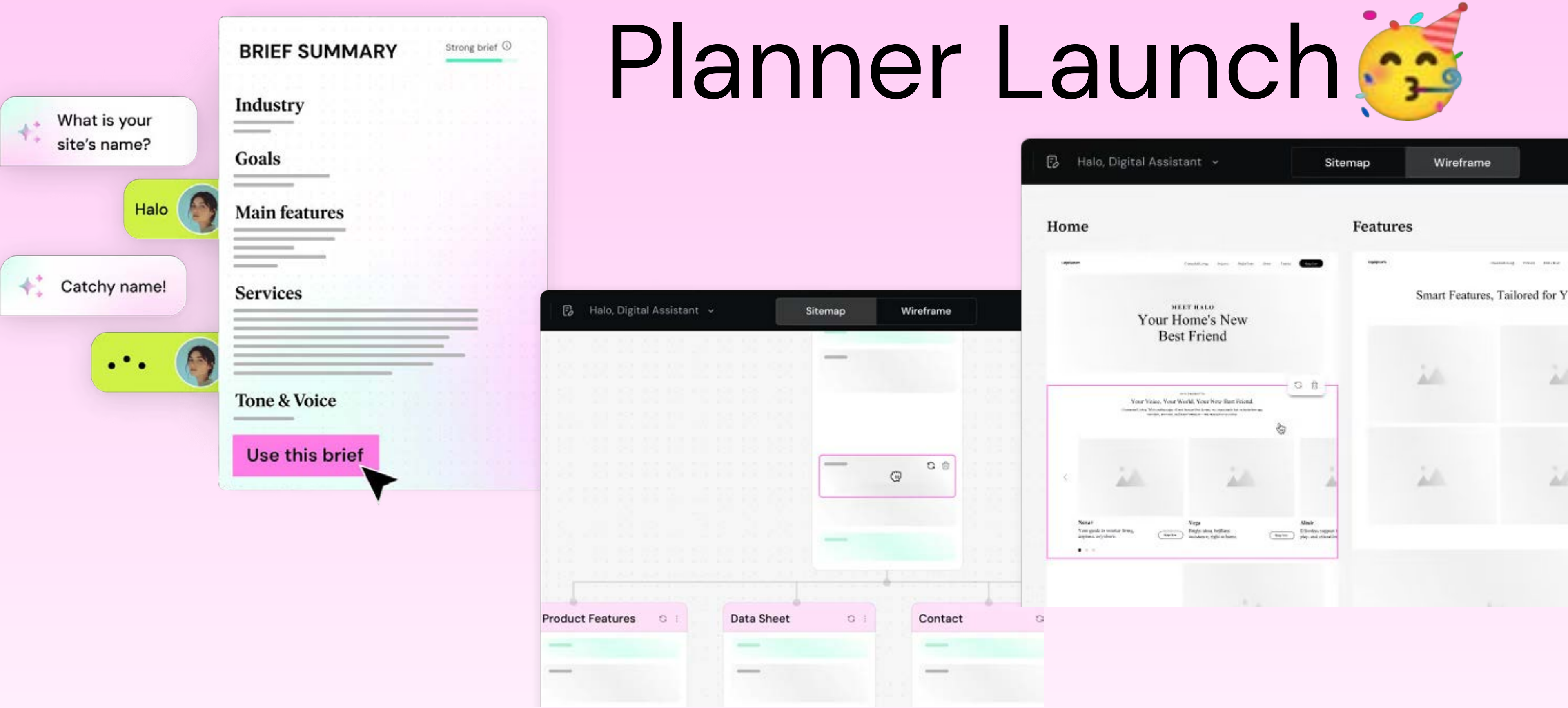
Dennis Nerush
Director of AI Engineering
@Elementor

Once Upon a Time...

21.03.25

Elementor AI Site Planner Launch🥳🥳🥳

Elementor AI Site Planner Launch🥳



What is your site's name?

Halo

Catchy name!

BRIEF SUMMARY

Strong brief

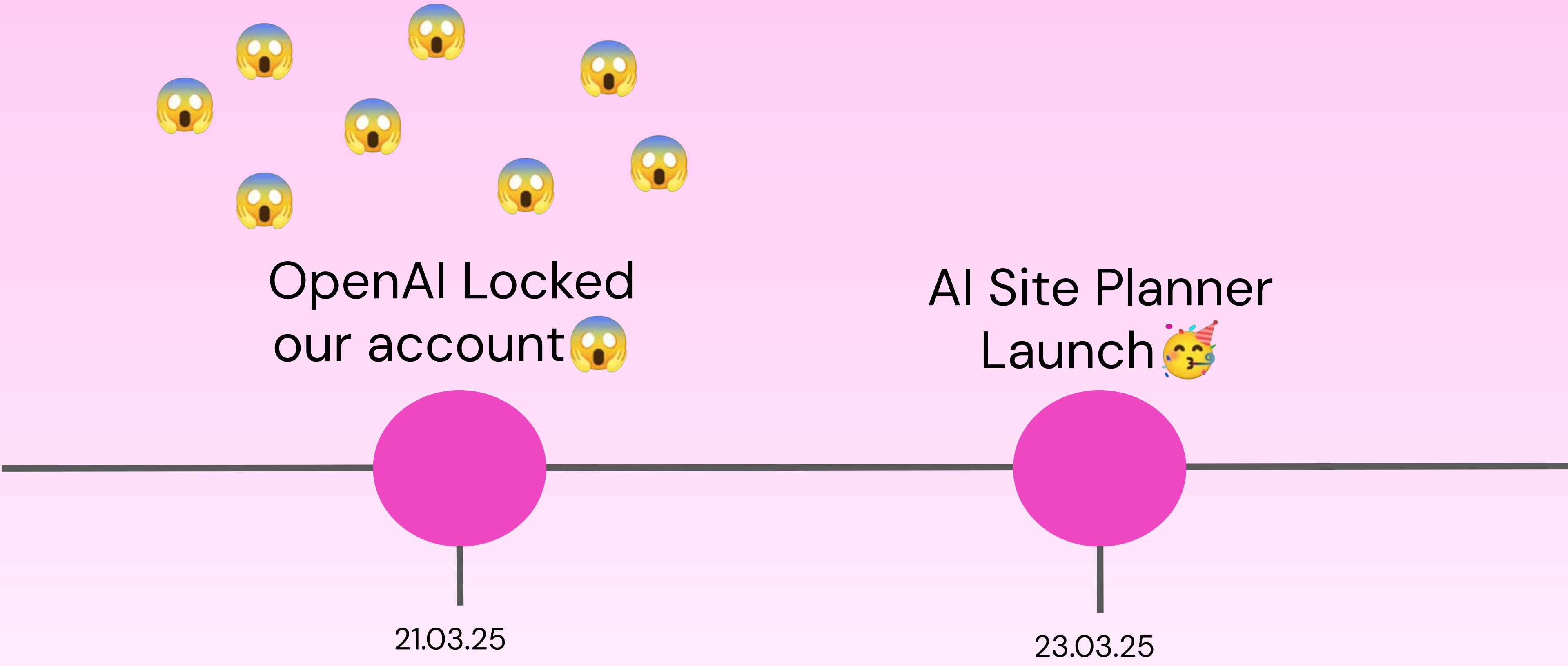
Industry

Goals

Main features

Services

Tone & Voice



OpenAI Locked our account😱

21.03.25

AI Site Planner Launch🥳

23.03.25

Nim

Disclaimer

Microsoft

OpenAI

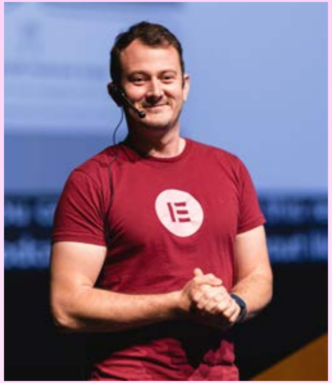
Hi 👋🏼

I'm Dennis

Director of AI Engineering

@Elementor

https://medium.com/@dennisnerush

 ❤️ 🤖

Agenda

Handling 1,000,000 daily AI requests

Resilient AI Solution

Continuously Improving

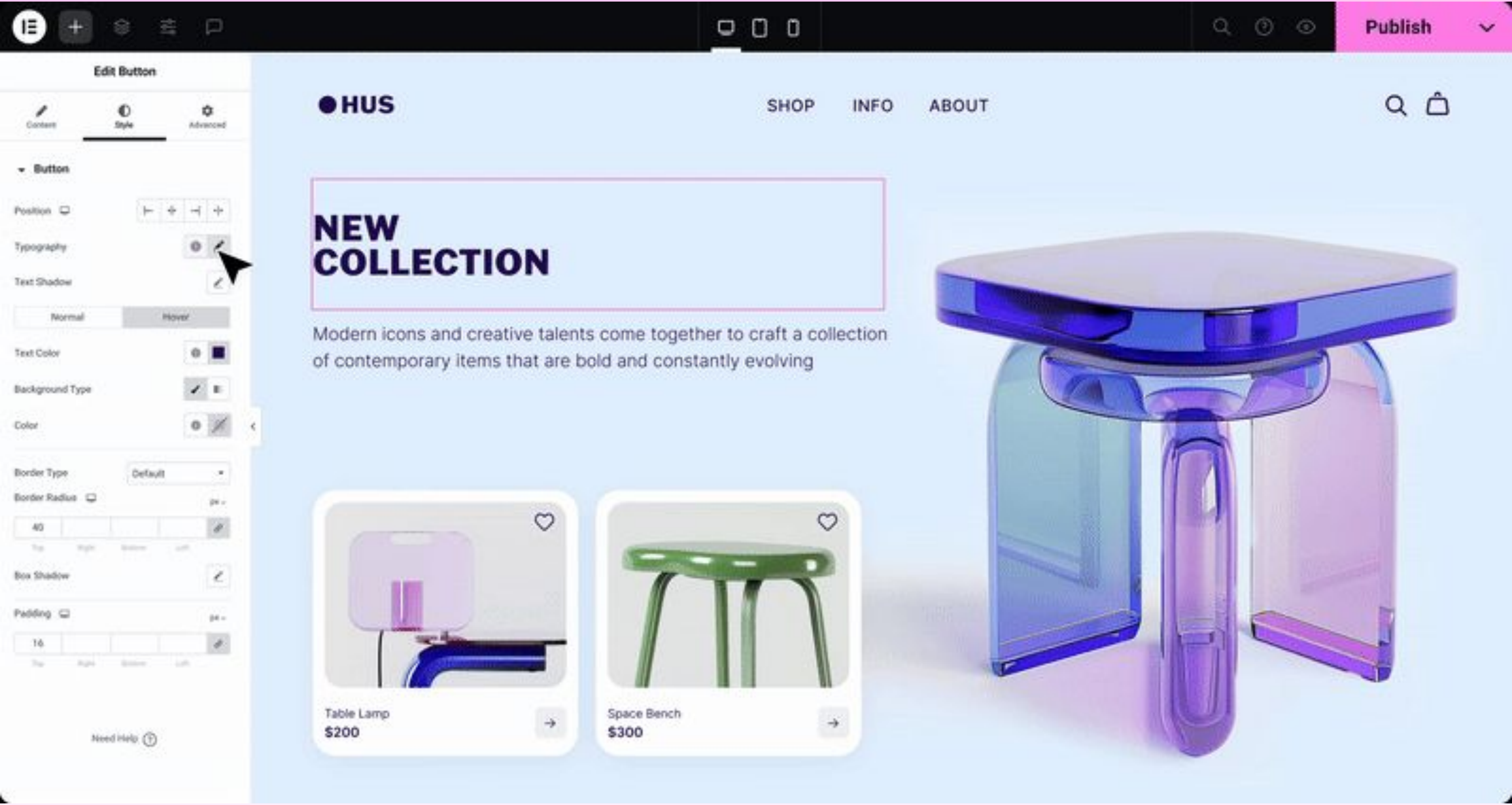
Build a Website
Create Your Future

19M

Websites Built With Elementor

15 sec.

Every 15 Seconds an Elementor Site is Born



Position
Typography
Text Shadow
Text Color
Background Type
Color
Border Type: Default
Border Radius: 40 px (Top, Right, Bottom, Left)
Box Shadow
Padding: 16 px (Top, Right, Bottom, Left)

NEW
COLLECTION

Modern icons and creative talents come together to craft a collection of contemporary items that are bold and constantly evolving

Table Lamp
$200

Space Bench
$300

Elementor AI

B.C.

Before ChatGPT 😅

AI Product in 4 months🤯

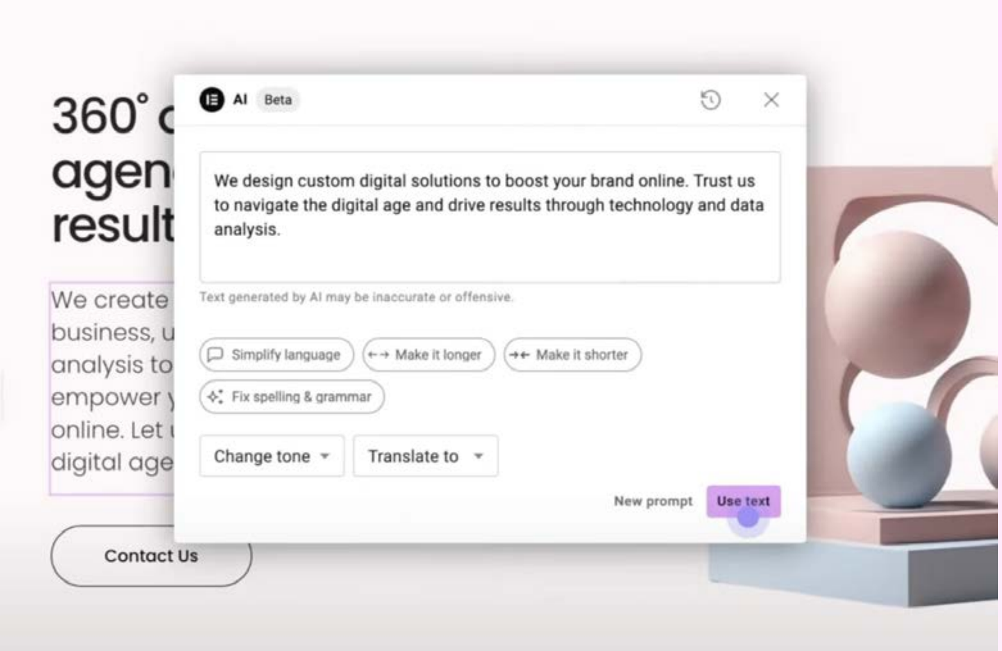
ChatGPT
Launch

November 30,
2022

Elementor AI
Launch🚀

April 13,
2023

Elementor AI ✨

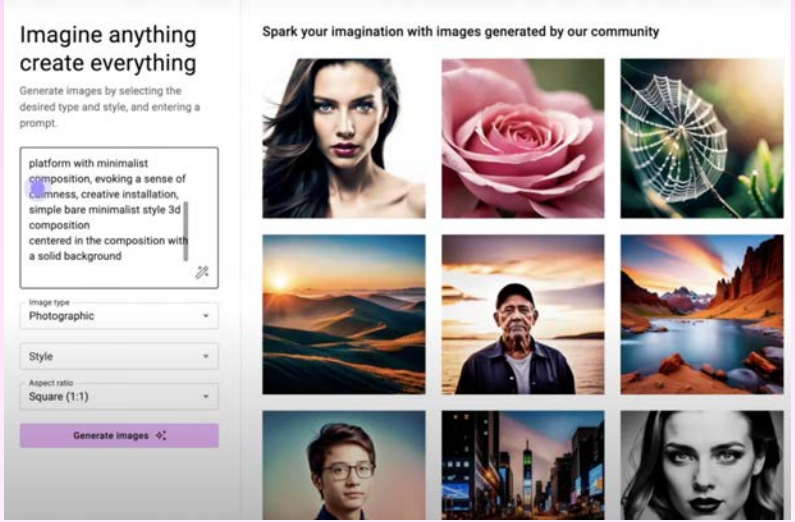


We design custom digital solutions to boost your brand online. Trust us to navigate the digital age and drive results through technology and data analysis.

Text generated by AI may be inaccurate or offensive.

Elementor AI ✨

🟫💻



Imagine anything create everything

Generate images by selecting the desired type and style, and entering a prompt.

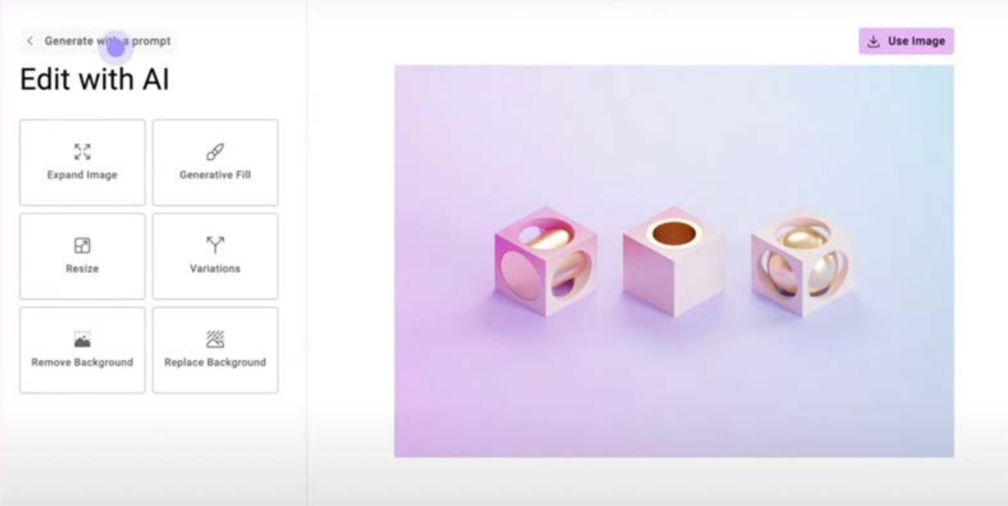
platform with minimalist composition, evoking a sense of calmness, creative installation, simple bare minimalist style 3d composition
centered in the composition with a solid background

Image type: Photographic

Style

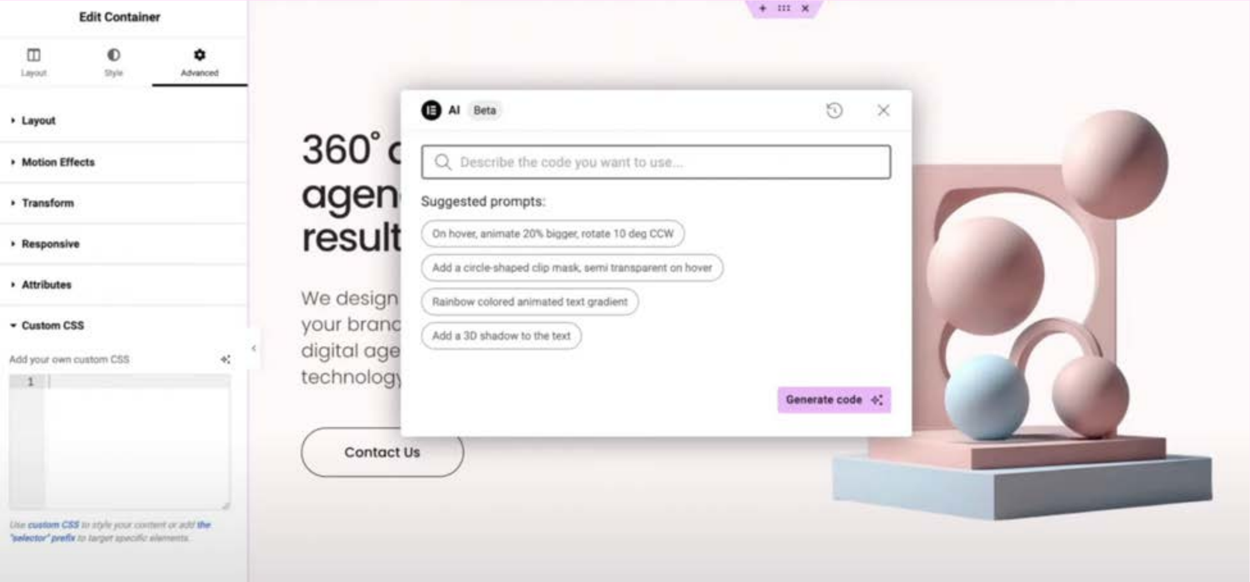
Aspect ratio: Square (1:1)

Spark your imagination with images generated by our community



Edit with AI

Elementor AI ✦



Add your own custom CSS

Use custom CSS to style your content or add the "selector" prefix to target specific elements.

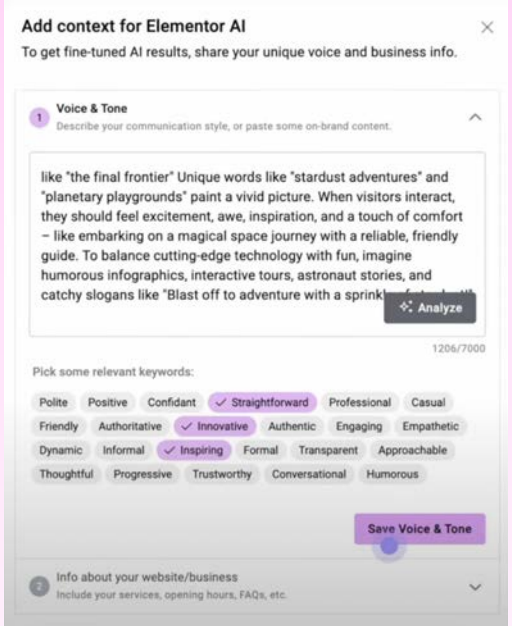
Describe the code you want to use...

Suggested prompts:

- On hover, animate 20% bigger, rotate 10 deg CCW
- Add a circle-shaped clip mask, semi transparent on hover
- Rainbow colored animated text gradient
- Add a 3D shadow to the text

Elementor AI ✨

🟫 💻



Add context for Elementor AI

To get fine-tuned AI results, share your unique voice and business info.

1 **Voice & Tone**
Describe your communication style, or paste some on-brand content.

like "the final frontier" Unique words like "stardust adventures" and "planetary playgrounds" paint a vivid picture. When visitors interact, they should feel excitement, awe, inspiration, and a touch of comfort – like embarking on a magical space journey with a reliable, friendly guide. To balance cutting-edge technology with fun, imagine humorous infographics, interactive tours, astronaut stories, and catchy slogans like "Blast off to adventure with a sprink

1206/7000

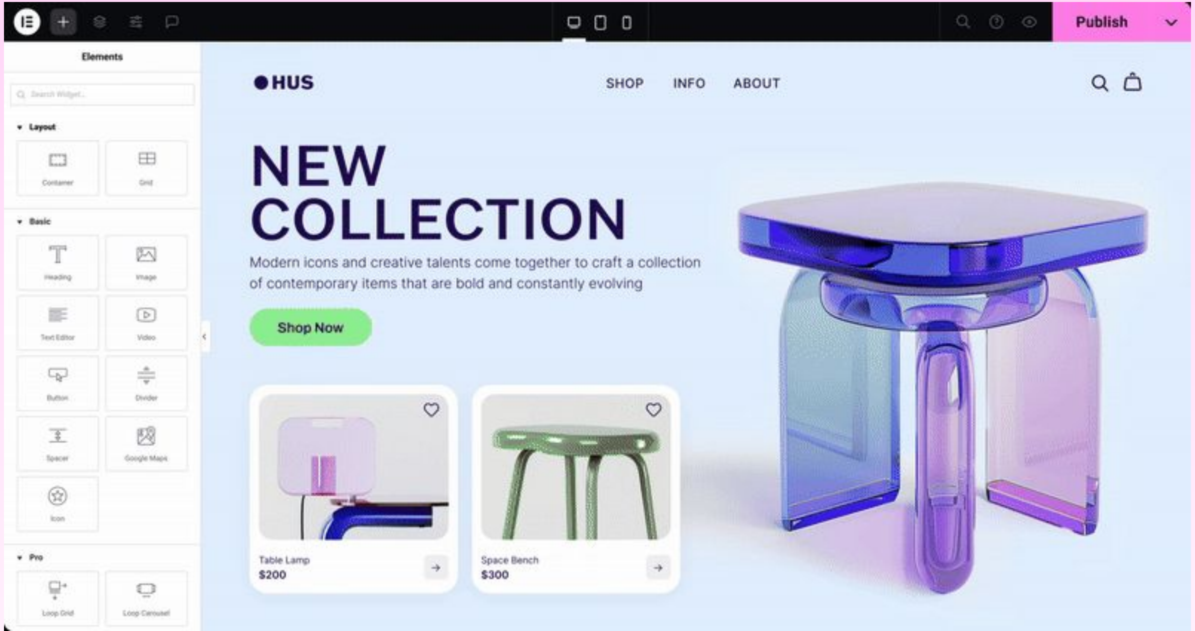
Pick some relevant keywords:

Polite · Positive · Confidant · ✓ Straightforward · Professional · Casual · Friendly · Authoritative · ✓ Innovative · Authentic · Engaging · Empathetic · Dynamic · Informal · ✓ Inspiring · Formal · Transparent · Approachable · Thoughtful · Progressive · Trustworthy · Conversational · Humorous

2 **Info about your website/business**
Include your services, opening hours, FAQs, etc.

Elementor AI ✨

🟫 💻



NEW COLLECTION

Modern icons and creative talents come together to craft a collection of contemporary items that are bold and constantly evolving

Elementor AI ✨

~2M users 🐣

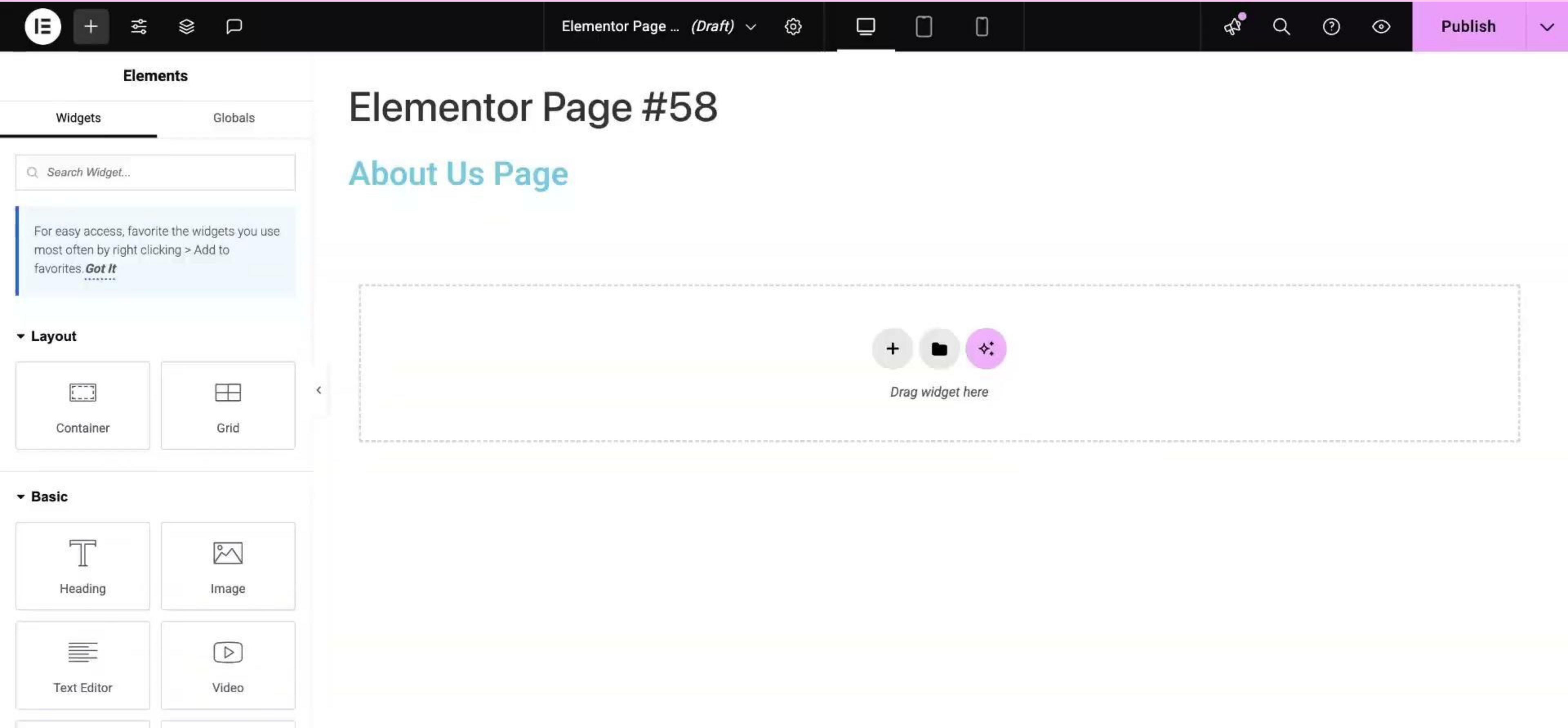
~10B/m tokens 🪙

40+ AI based features 🚀

#1 - Handling 1,000,000 **daily** AI **requests**

#1 - Handling 1,000,000 daily AI requests

- In 2023, we started with **OpenAI**
  - gpt3.5 turbo rate limit: 60 RPM😩
  - We wanted a familiar vendor - so we moved to **Microsoft Azure**🥸
- It worked fine for a year, until we started working on the **Elementor AI Copilot**



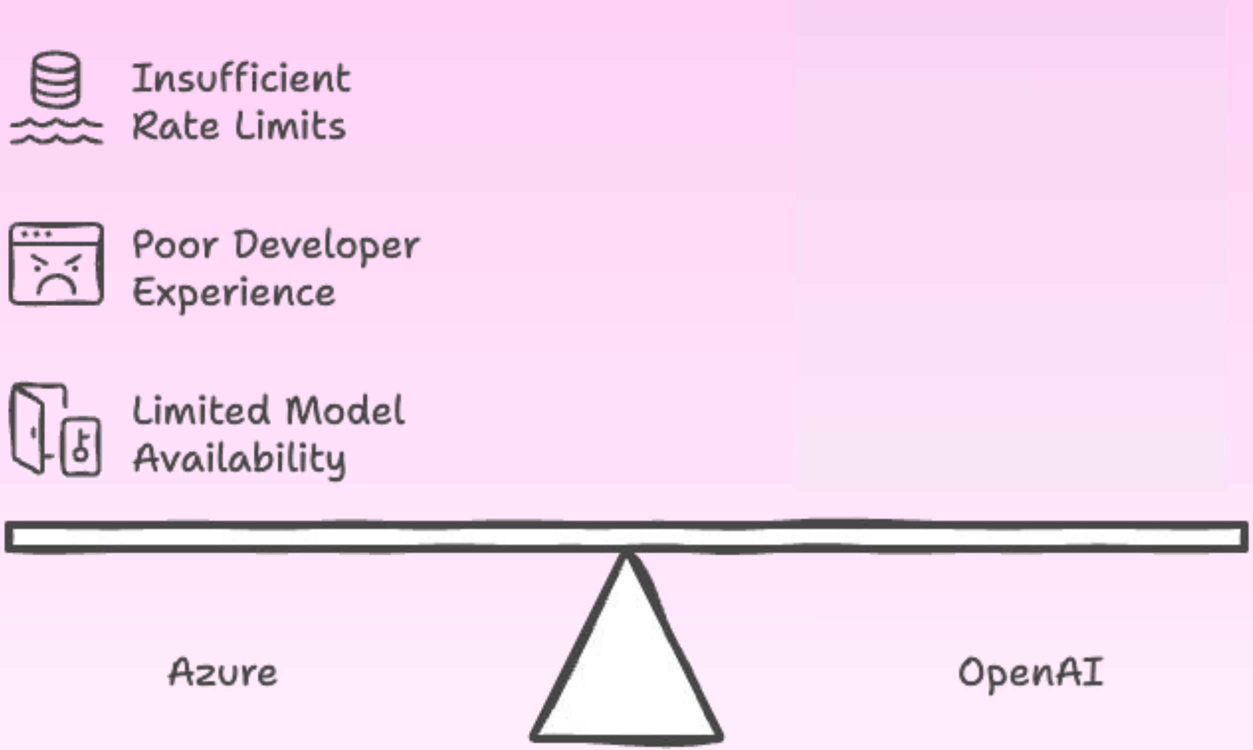
For easy access, favorite the widgets you use most often by right clicking > Add to favorites. *Got It*

Elementor Page #58

About Us Page

😢

Handling 1,000,000 daily AI **requests**



Insufficient Rate Limits

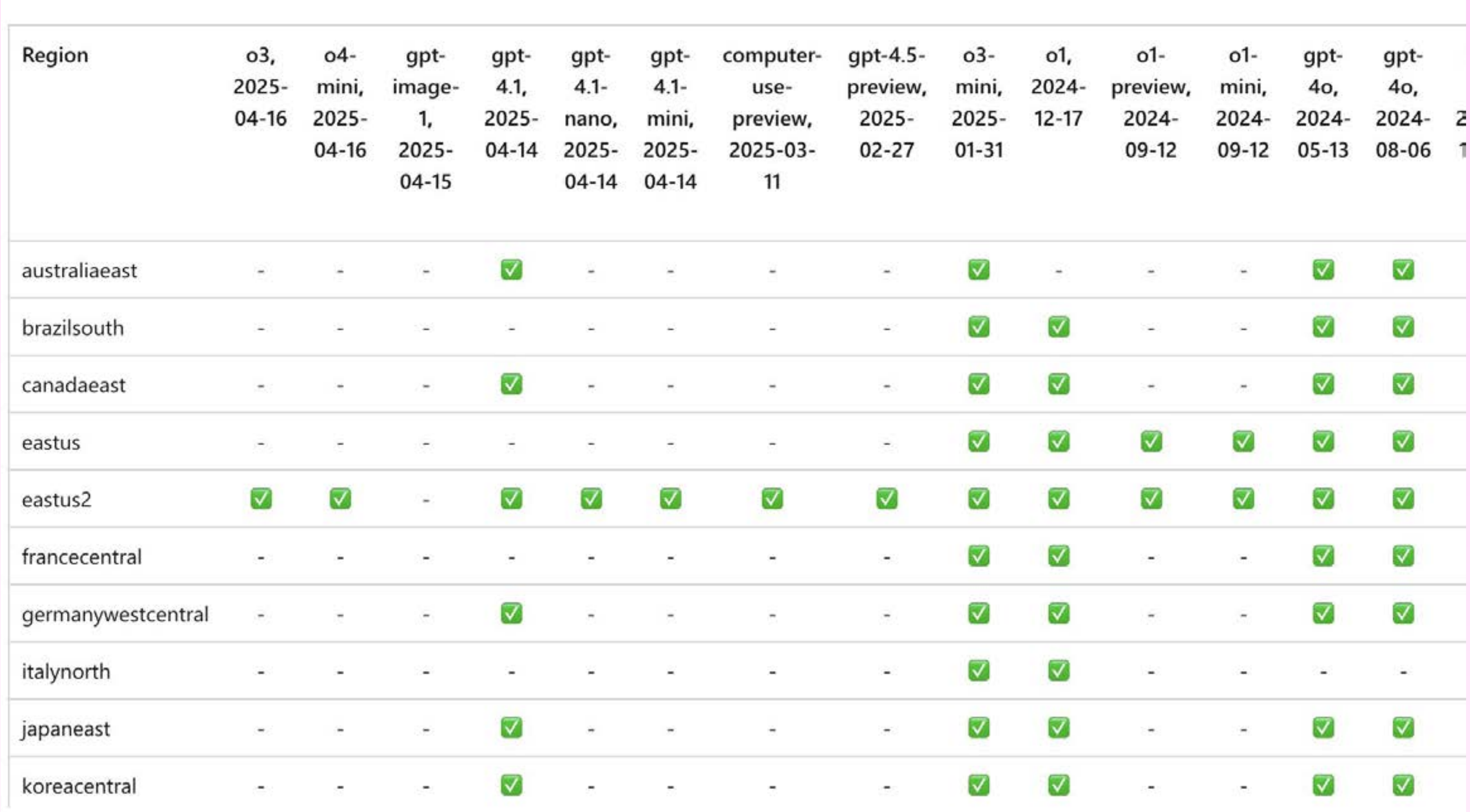
Poor Developer Experience

Limited Model Availability

Azure

OpenAI

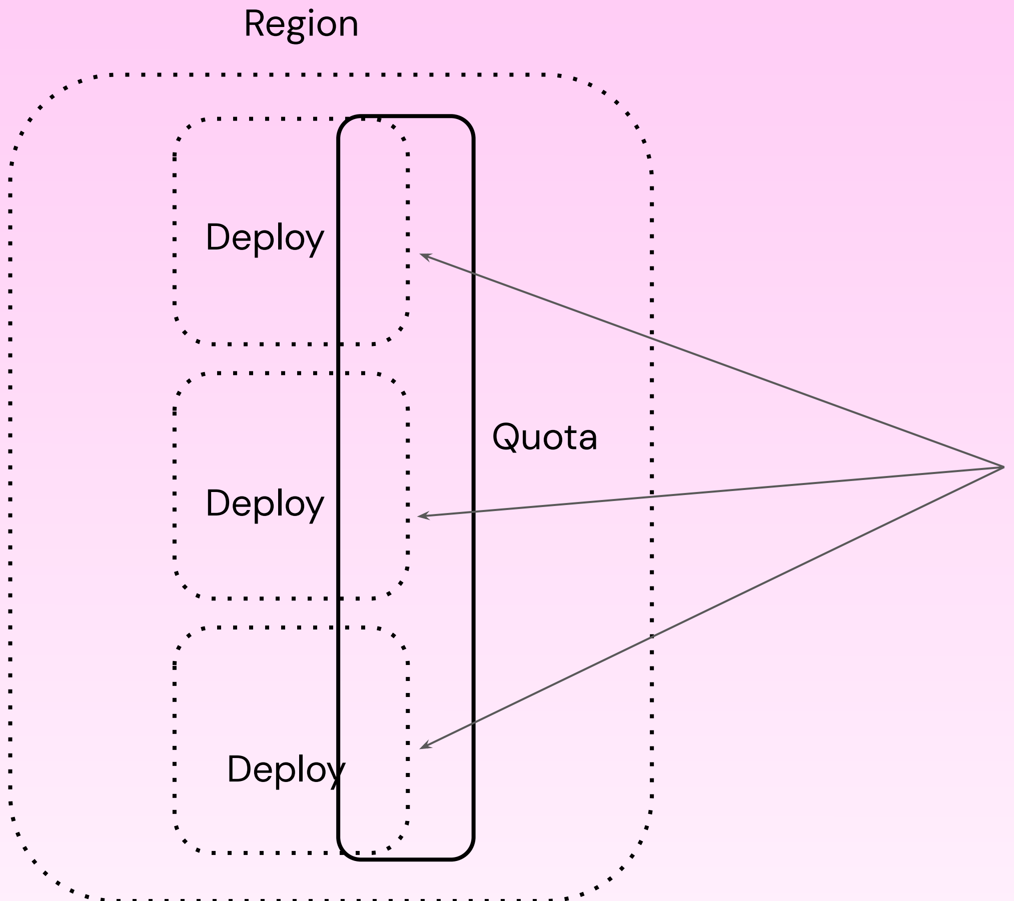
Where's my model🤔 ?



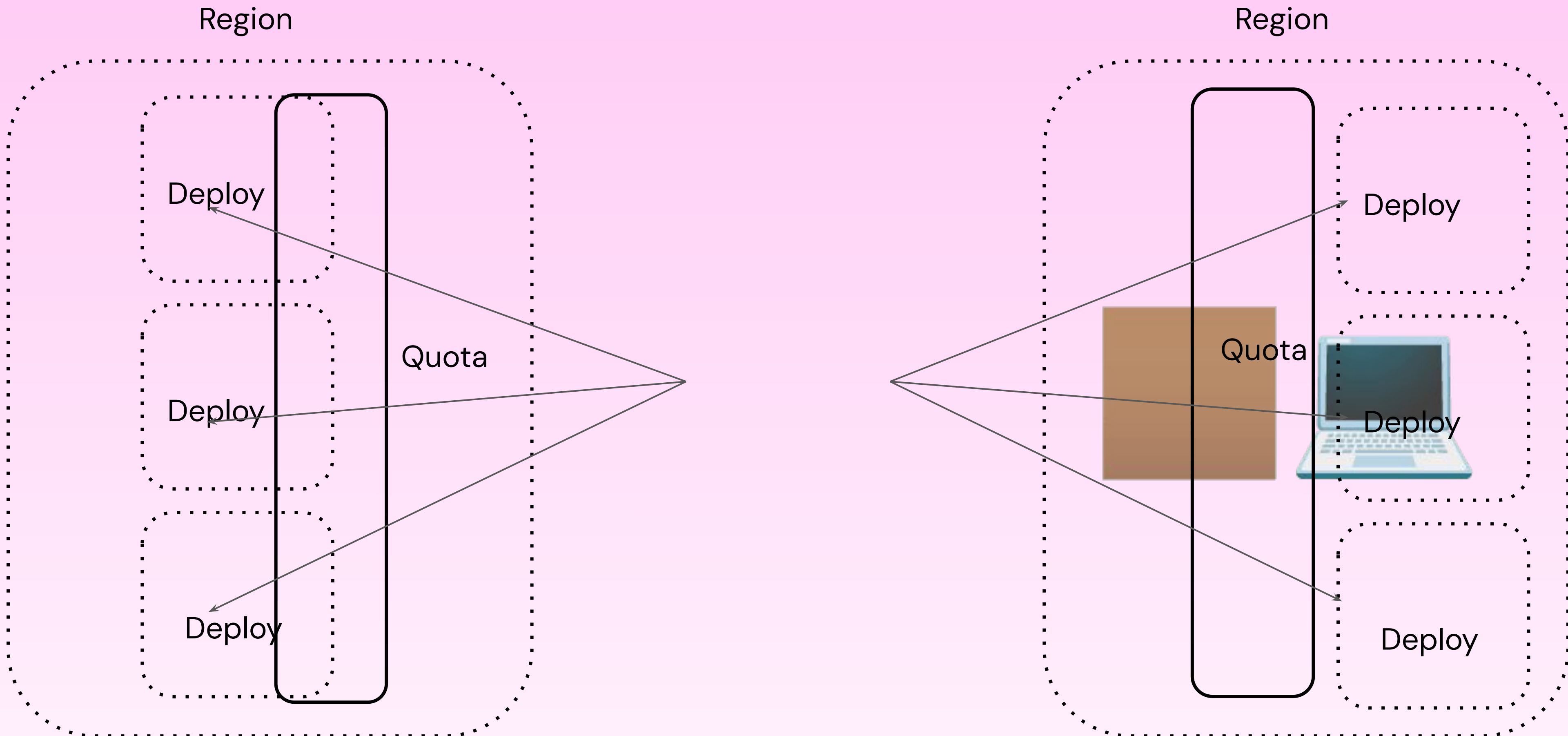
Region	o3, 2025-04-16	o4-mini, 2025-04-16	gpt-image-1, 2025-04-15	gpt-4.1, 2025-04-14	gpt-4.1-nano, 2025-04-14	gpt-4.1-mini, 2025-04-14	computer-use-preview, 2025-03-11	gpt-4.5-preview, 2025-02-27	o3-mini, 2025-01-31	o1, 2024-12-17	o1-preview, 2024-09-12	o1-mini, 2024-09-12	gpt-4o, 2024-05-13	gpt-4o, 2024-08-06
australiaeast	-	-	-	✅	-	-	-	-	✅	-	-	-	✅	✅
brazilsouth	-	-	-	-	-	-	-	-	✅	✅	-	-	✅	✅
canadaeast	-	-	-	✅	-	-	-	-	✅	✅	-	-	✅	✅
eastus	-	-	-	-	-	-	-	-	✅	✅	✅	✅	✅	✅
eastus2	✅	✅	-	✅	✅	✅	✅	✅	✅	✅	✅	✅	✅	✅
francecentral	-	-	-	-	-	-	-	-	✅	✅	-	-	✅	✅
germanywestcentral	-	-	-	✅	-	-	-	-	✅	✅	-	-	✅	✅
italynorth	-	-	-	-	-	-	-	-	✅	✅	-	-	-	-
japaneast	-	-	-	✅	-	-	-	-	✅	✅	-	-	✅	✅
koreacentral	-	-	-	✅	-	-	-	-	✅	✅	-	-	✅	✅

Link to Azure portal and models availability

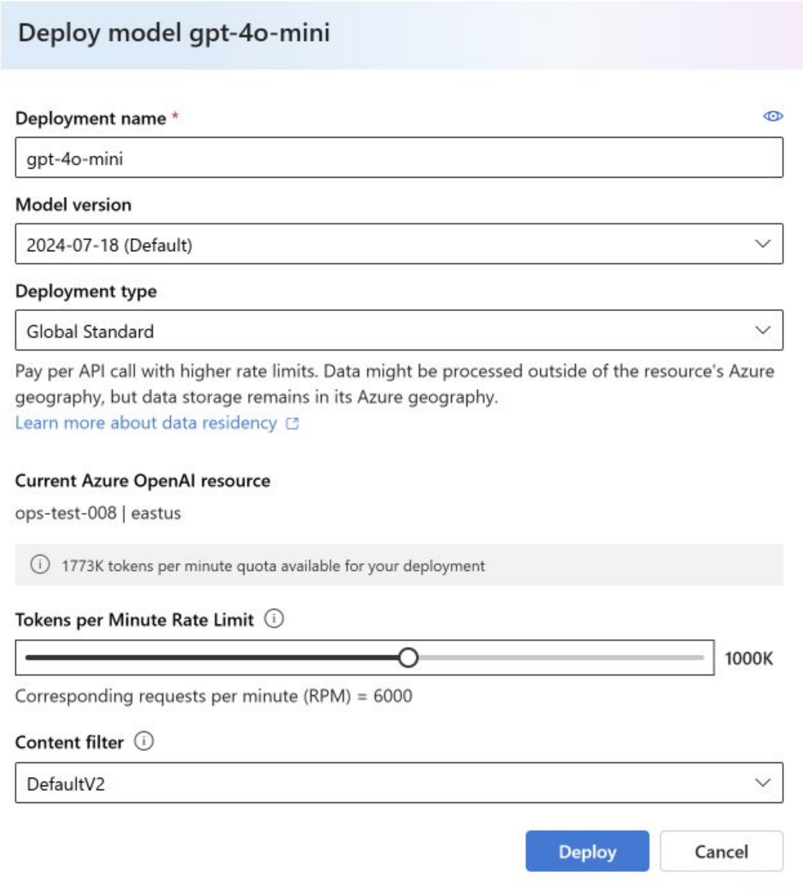
Cognitive Load for developers🥸



Cognitive Load for developers🥸



There's a form to increase the quota...



Deploy model gpt-4o-mini

Deployment name *

gpt-4o-mini

Model version

2024-07-18 (Default)

Deployment type

Global Standard

Pay per API call with higher rate limits. Data might be processed outside of the resource's Azure geography, but data storage remains in its Azure geography.

Learn more about data residency

Current Azure OpenAI resource

ops-test-008 | eastus

ⓘ 1773K tokens per minute quota available for your deployment

Tokens per Minute Rate Limit ⓘ

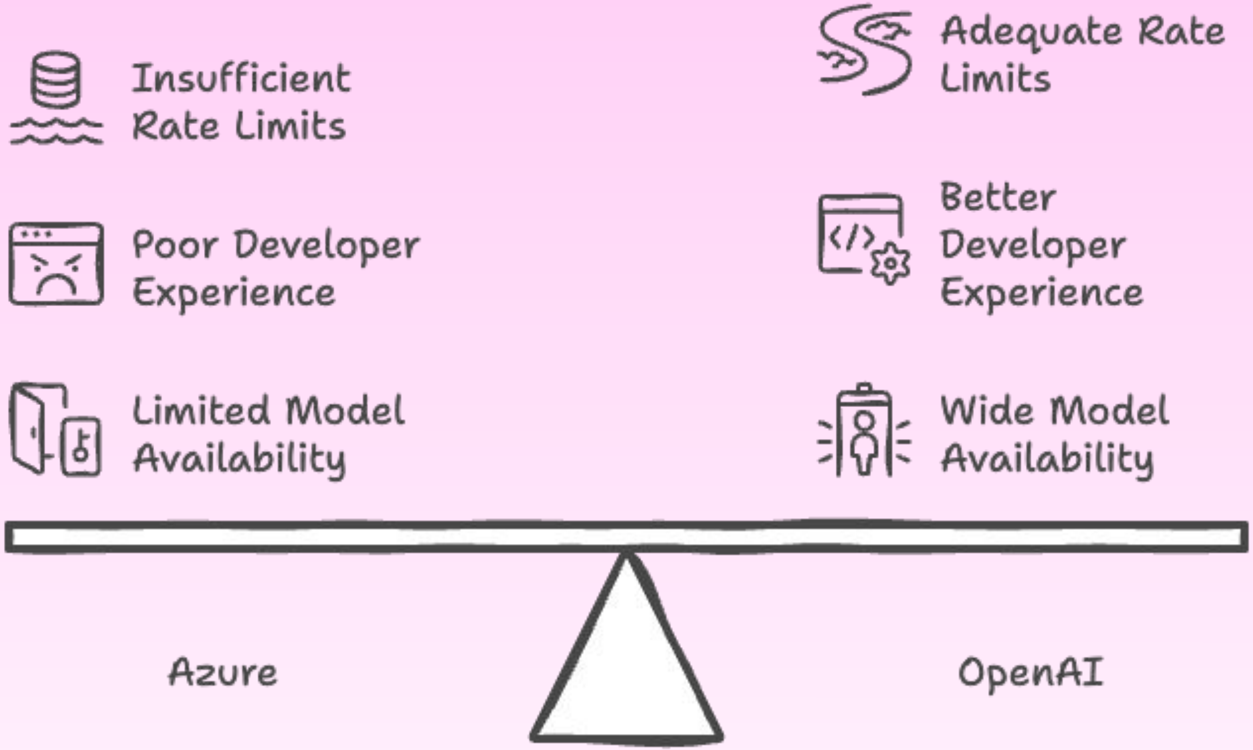
1000K

Corresponding requests per minute (RPM) = 6000

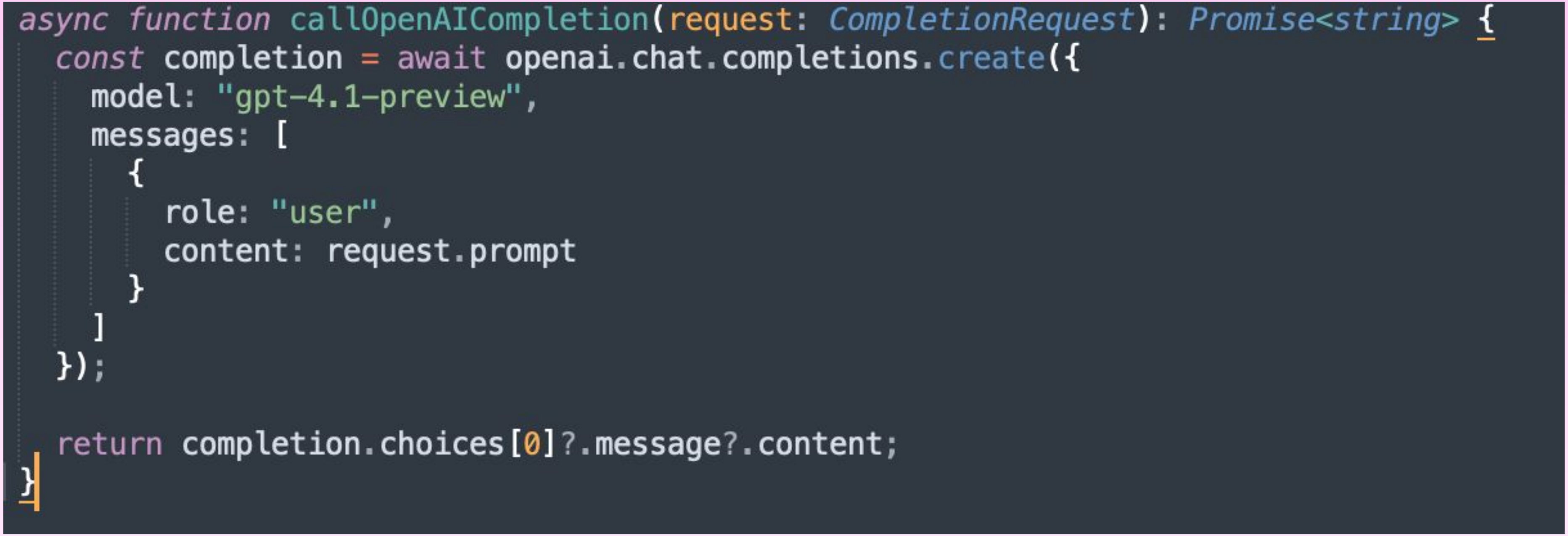
Content filter ⓘ

DefaultV2

Handling 1,000,000 daily AI <u>requests</u>


Insufficient Rate Limits
Poor Developer Experience
Limited Model Availability
Adequate Rate Limits
Better Developer Experience
Wide Model Availability
Azure
OpenAI

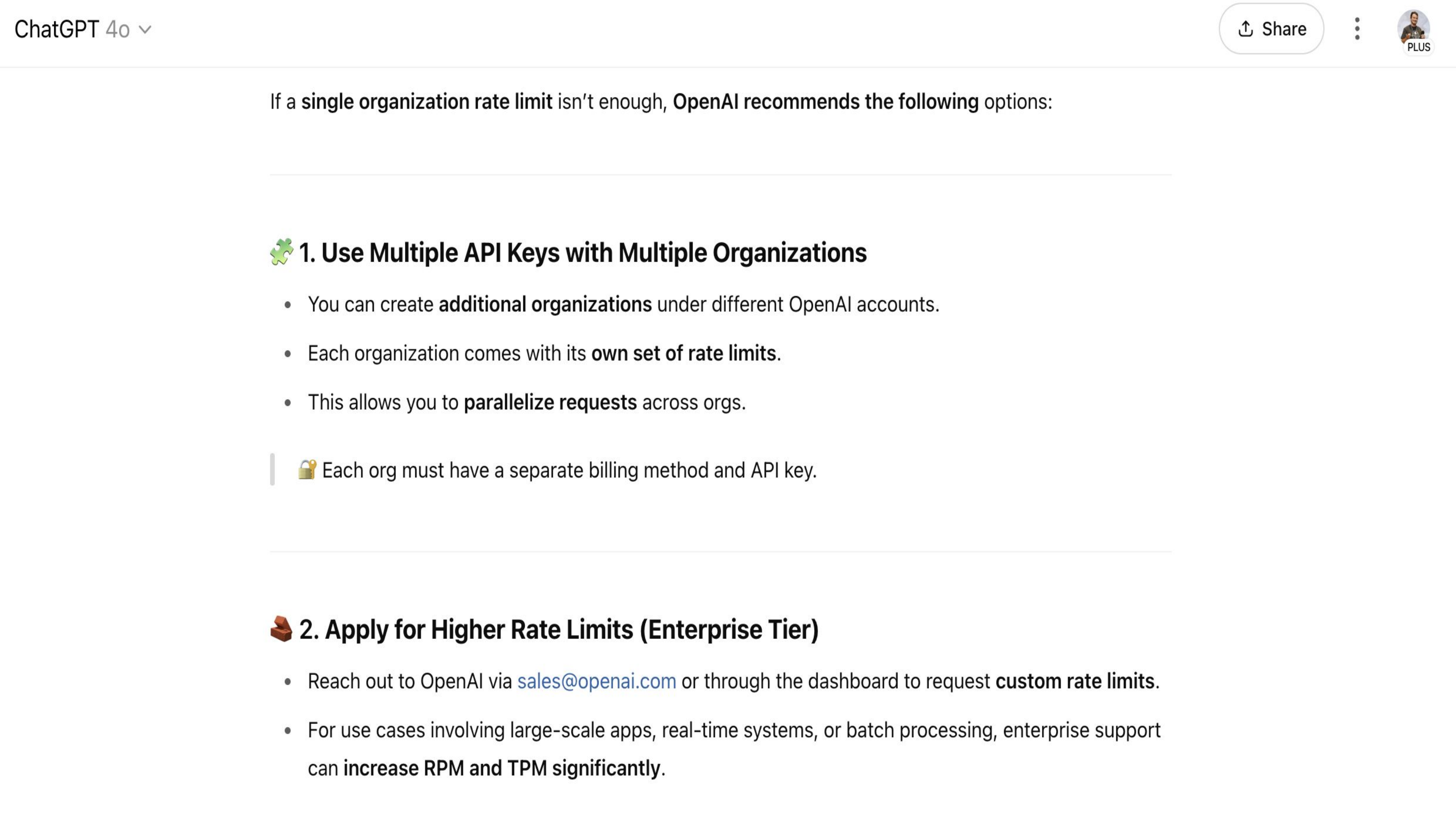
Back to OpenAI🤩



```
async function callOpenAICompletion(request: CompletionRequest): Promise<string> {
  const completion = await openai.chat.completions.create({
    model: "gpt-4.1-preview",
    messages: [
      {
        role: "user",
        content: request.prompt
      }
    ]
  });

  return completion.choices[0]?.message?.content;
}
```

IT’S STILL NOT ENOUGH



If a **single organization rate limit** isn't enough, **OpenAI recommends the following** options:

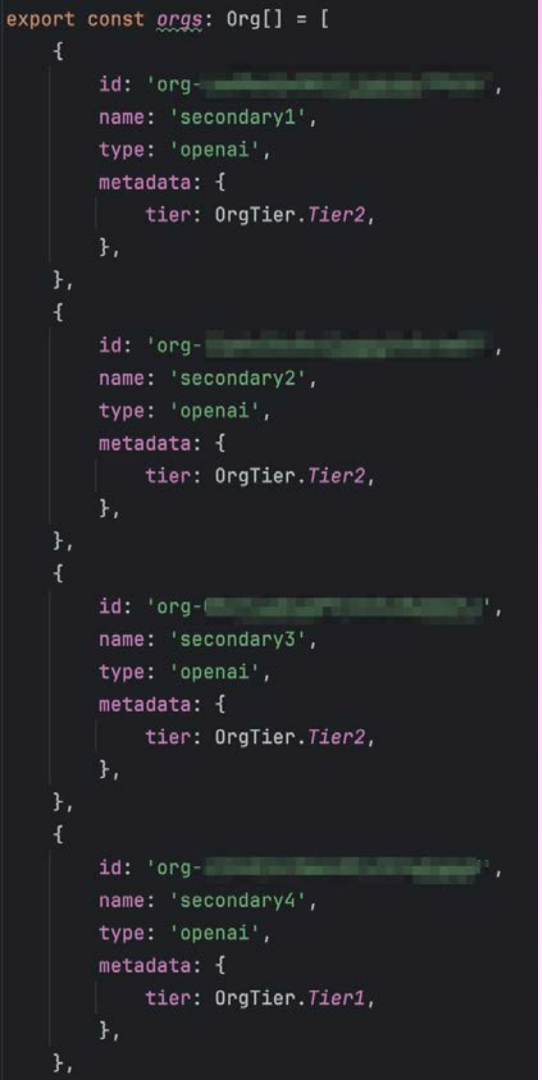
🧩 1. Use Multiple API Keys with Multiple Organizations

- You can create **additional organizations** under different OpenAI accounts.
- Each organization comes with its **own set of rate limits**.
- This allows you to **parallelize requests** across orgs.

> 🔐 Each org must have a separate billing method and API key.

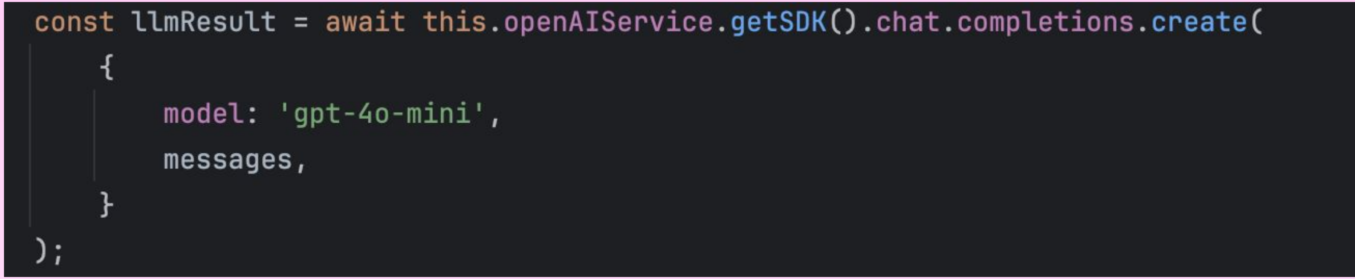
🧱 2. Apply for Higher Rate Limits (Enterprise Tier)

- Reach out to OpenAI via sales@openai.com or through the dashboard to request **custom rate limits**.
- For use cases involving large-scale apps, real-time systems, or batch processing, enterprise support can **increase RPM and TPM significantly**.



```
export const orgs: Org[] = [
    {
        id: 'org-[redacted]',
        name: 'secondary1',
        type: 'openai',
        metadata: {
            tier: OrgTier.Tier2,
        },
    },
    {
        id: 'org-[redacted]',
        name: 'secondary2',
        type: 'openai',
        metadata: {
            tier: OrgTier.Tier2,
        },
    },
    {
        id: 'org-[redacted]',
        name: 'secondary3',
        type: 'openai',
        metadata: {
            tier: OrgTier.Tier2,
        },
    },
    {
        id: 'org-[redacted]',
        name: 'secondary4',
        type: 'openai',
        metadata: {
            tier: OrgTier.Tier1,
        },
    },
```

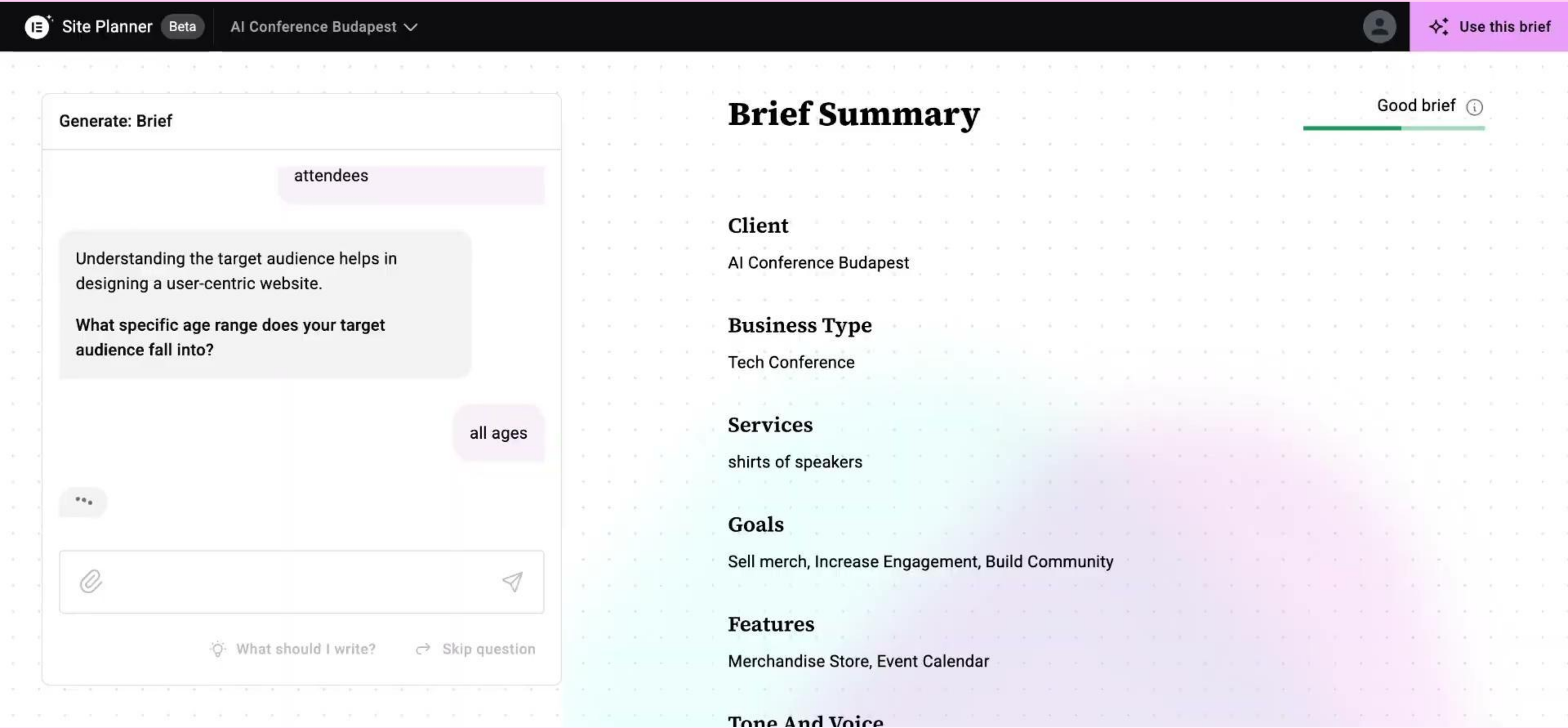
OpenAI MultiOrg Proxy



```
const llmResult = await this.openAIService.getSDK().chat.completions.create(
    {
        model: 'gpt-4o-mini',
        messages,
    }
);
```

```
public getS
    return t                                   ].sdk;
}
```

Handling 1,000,000 daily AI requests



Generate: Brief

attendees

Understanding the target audience helps in designing a user-centric website.

What specific age range does your target audience fall into?

all ages

•••

Brief Summary

Client

AI Conference Budapest

Business Type

Tech Conference

Services

shirts of speakers

Goals

Sell merch, Increase Engagement, Build Community

Features

Merchandise Store, Event Calendar

Tone And Voice

Takeaways - Handling 1,000,000 daily AI requests

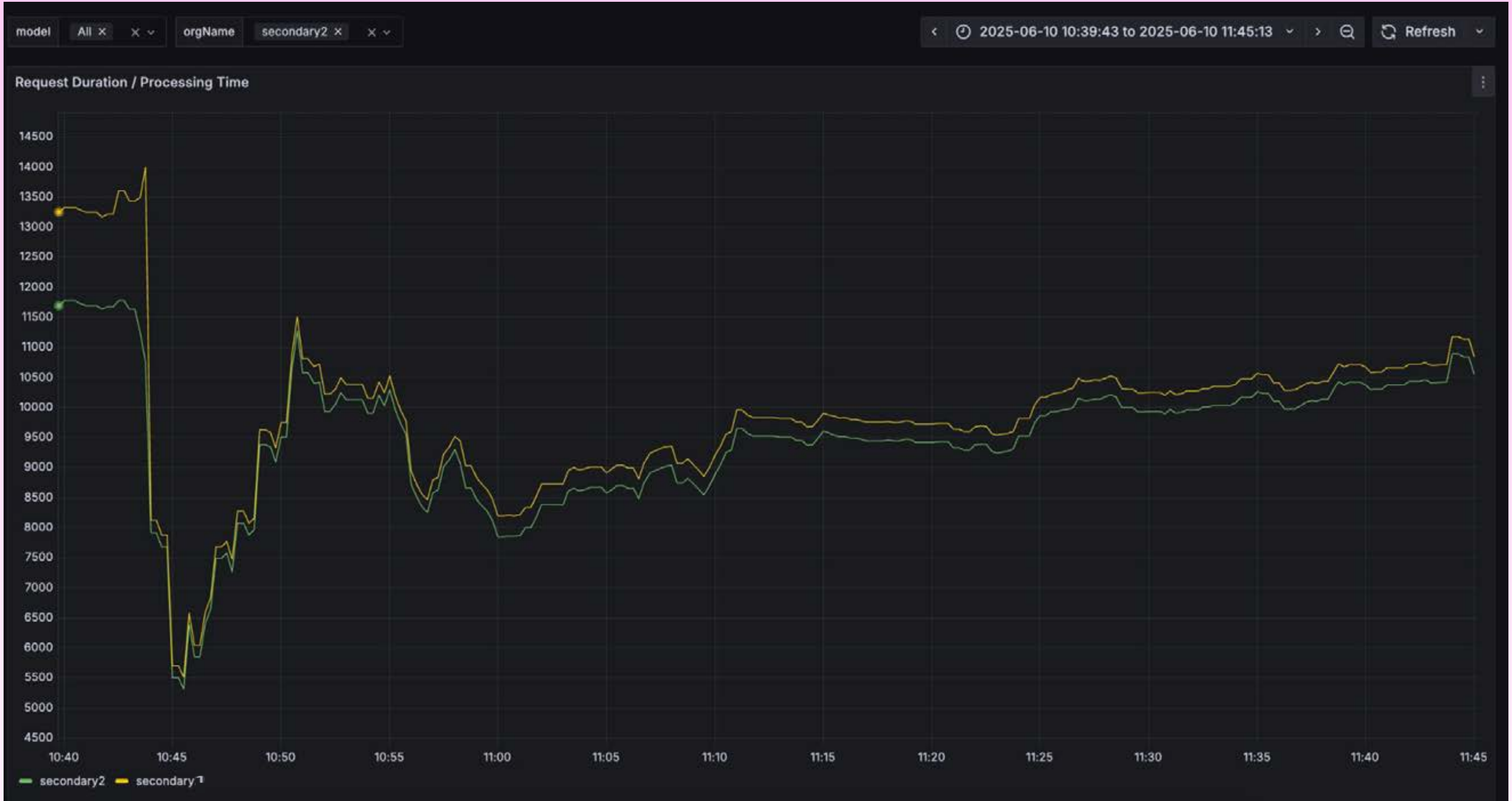
1. Get familiar with the rate limits – monitor your usage
2. Don’t get too close 😅
3. Add additional accounts/organizations as needed

(that’s helpful not only for rate limit 😉)

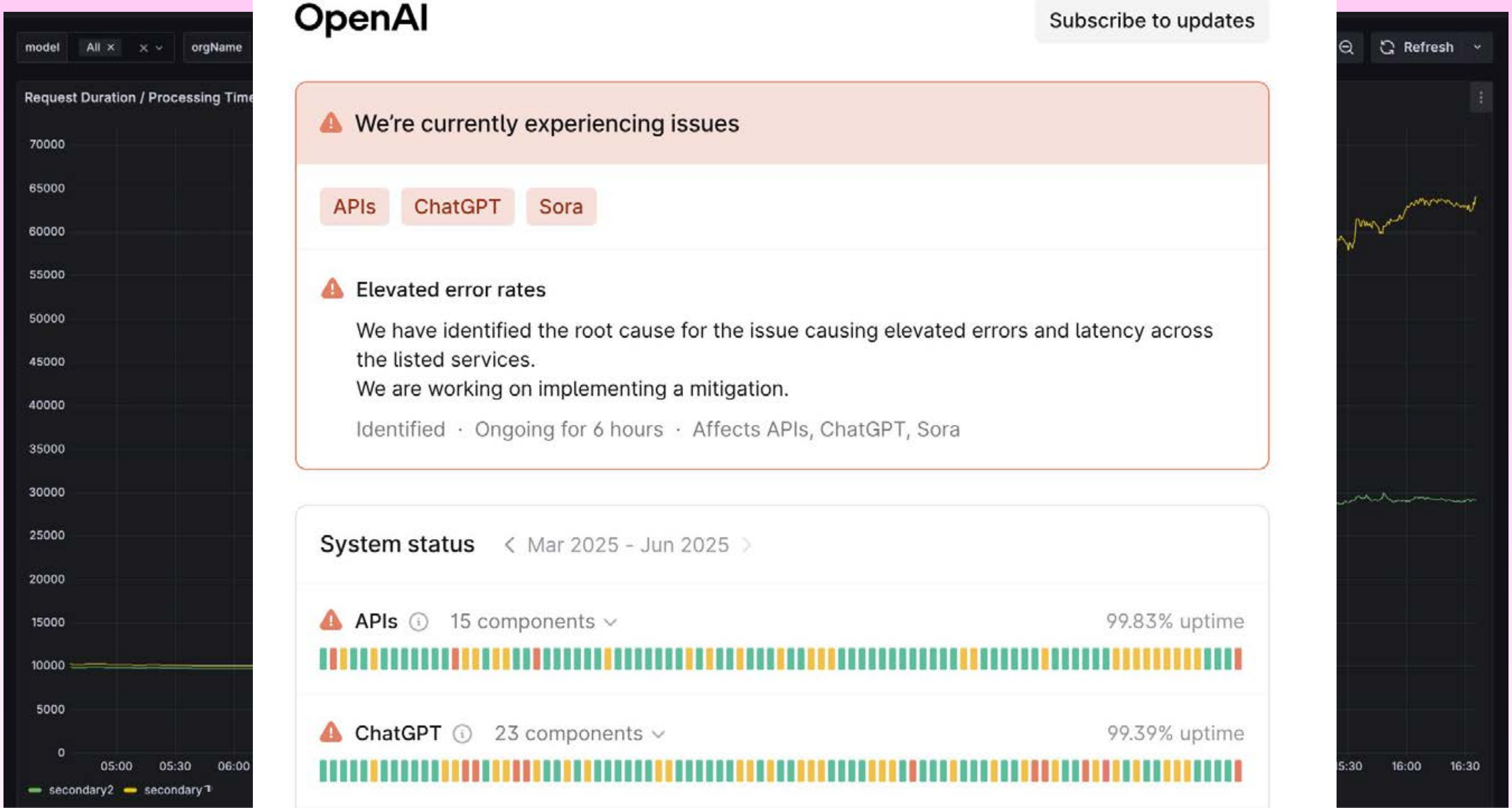
#2 - Being a Resilient AI Solution

“Everyone has a plan 'till they get punched in the mouth.” - Mike Tyson

How we think services work in Production?



How services ACTUALLY work in Production



OpenAI

Subscribe to updates

We're currently experiencing issues

APIs | ChatGPT | Sora

Elevated error rates

We have identified the root cause for the issue causing elevated errors and latency across the listed services.
We are working on implementing a mitigation.

Identified · Ongoing for 6 hours · Affects APIs, ChatGPT, Sora

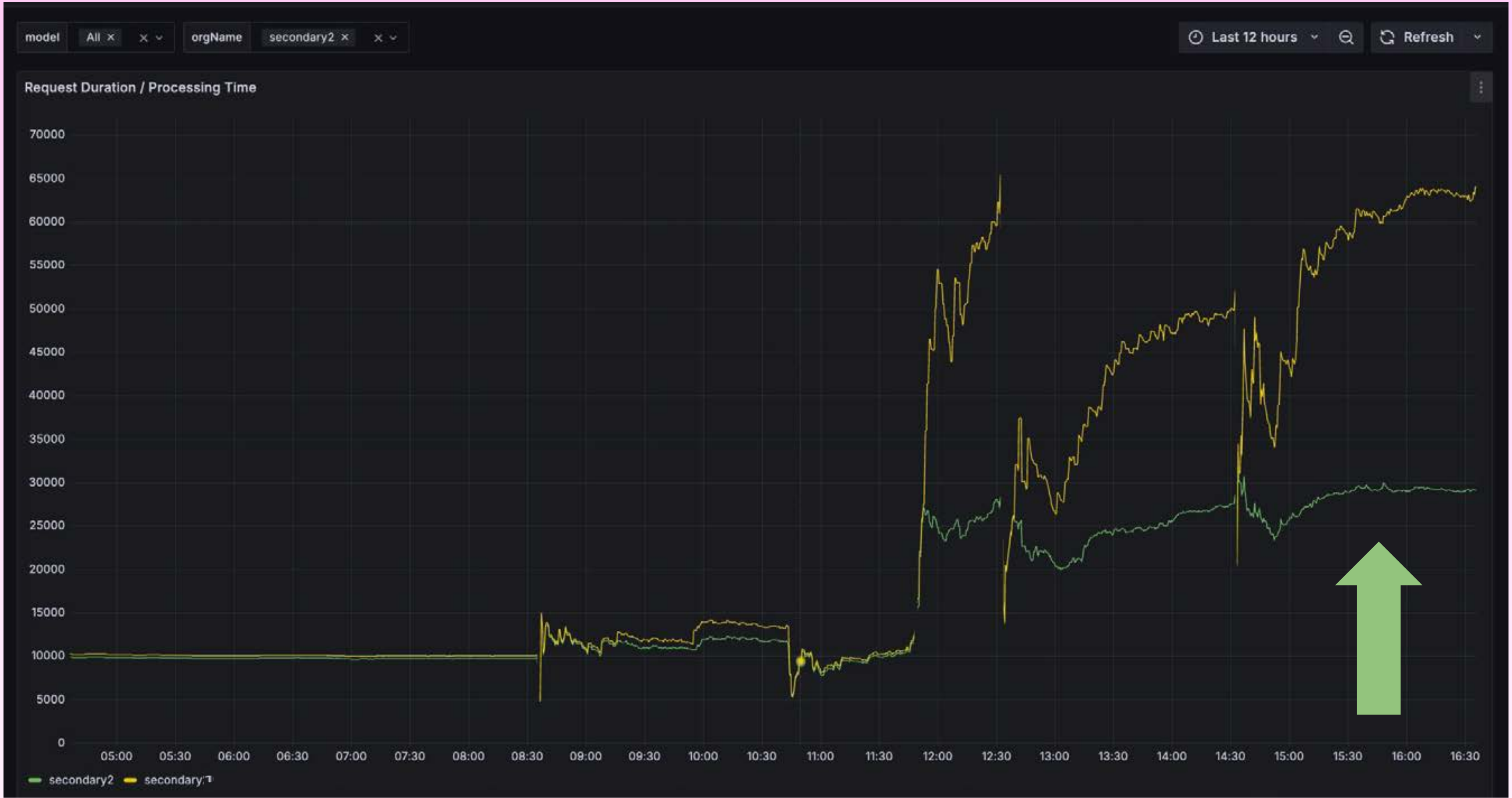
System status ‹ Mar 2025 - Jun 2025 ›

APIs ⓘ 15 components ⌄ 99.83% uptime

ChatGPT ⓘ 23 components ⌄ 99.39% uptime

What our users expect?

Resilient Multi Org Solution



model All ×

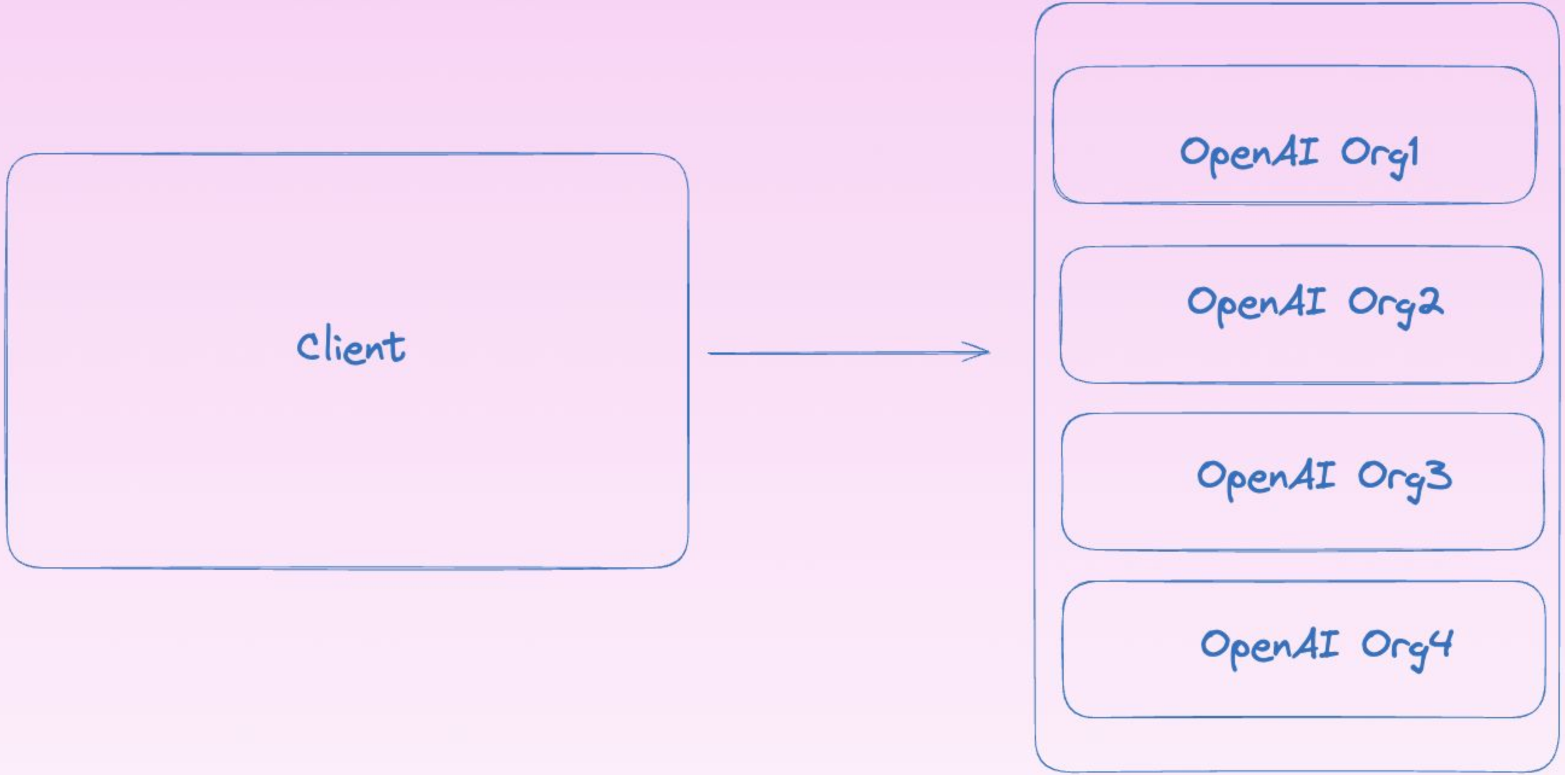
orgName secondary2 ×

Last 12 hours

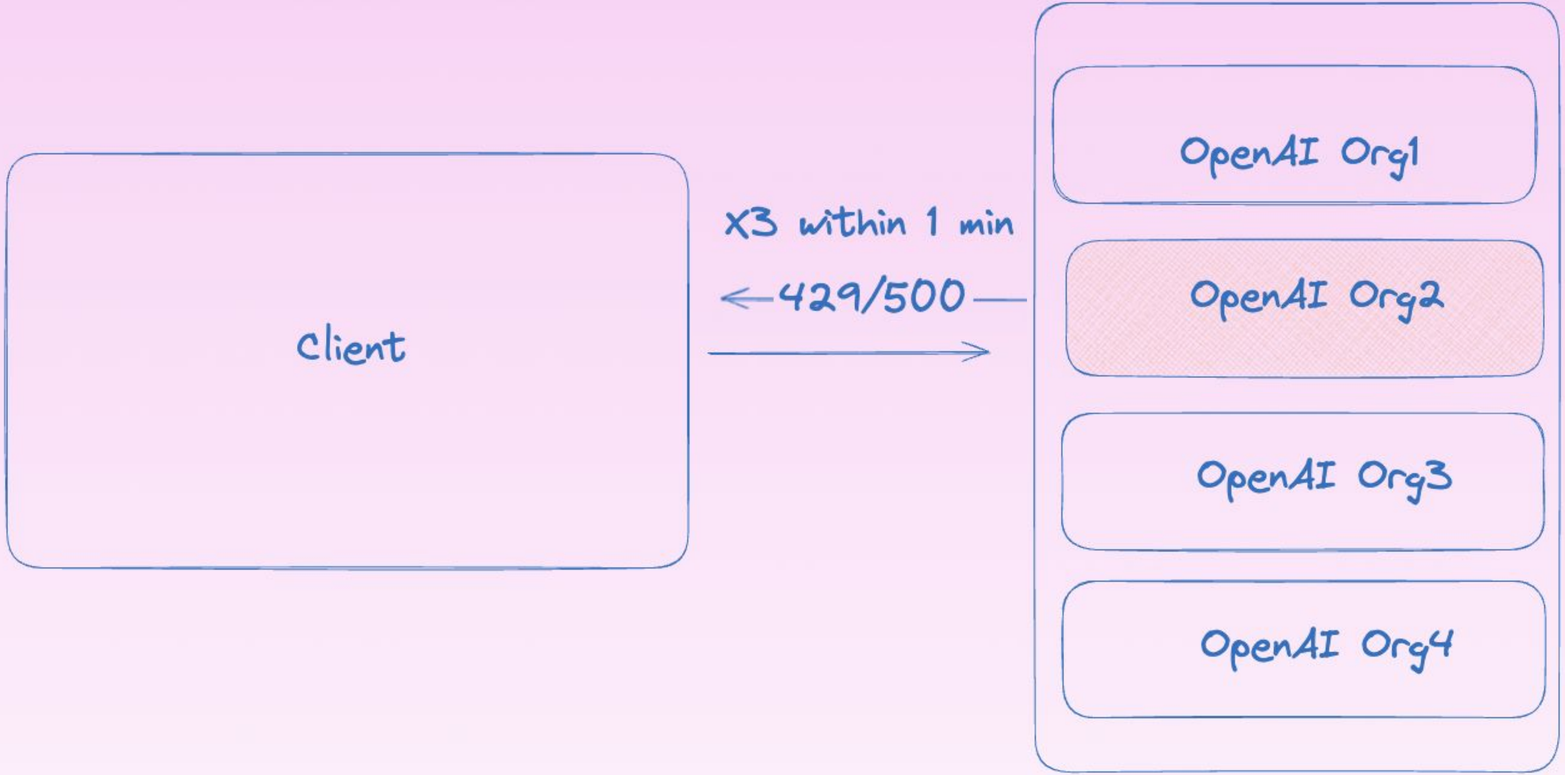
Refresh

Request Duration / Processing Time

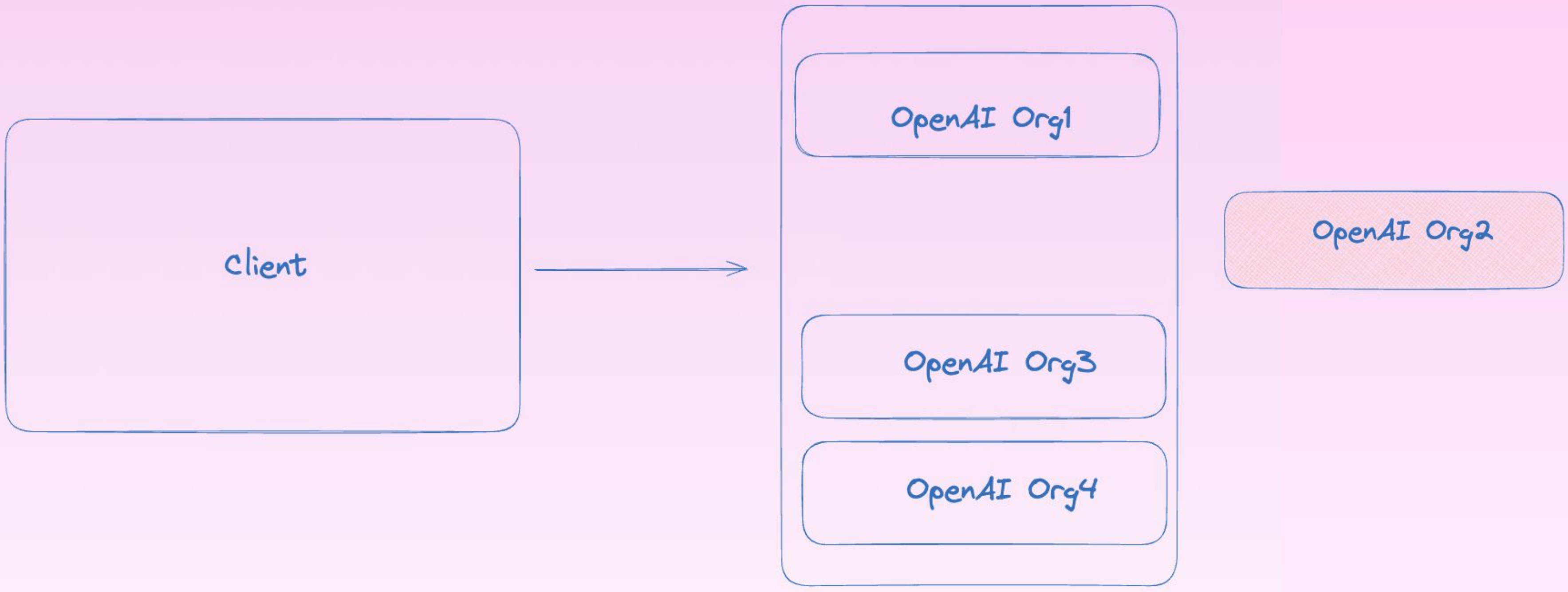
Resilient Multi Org Solution



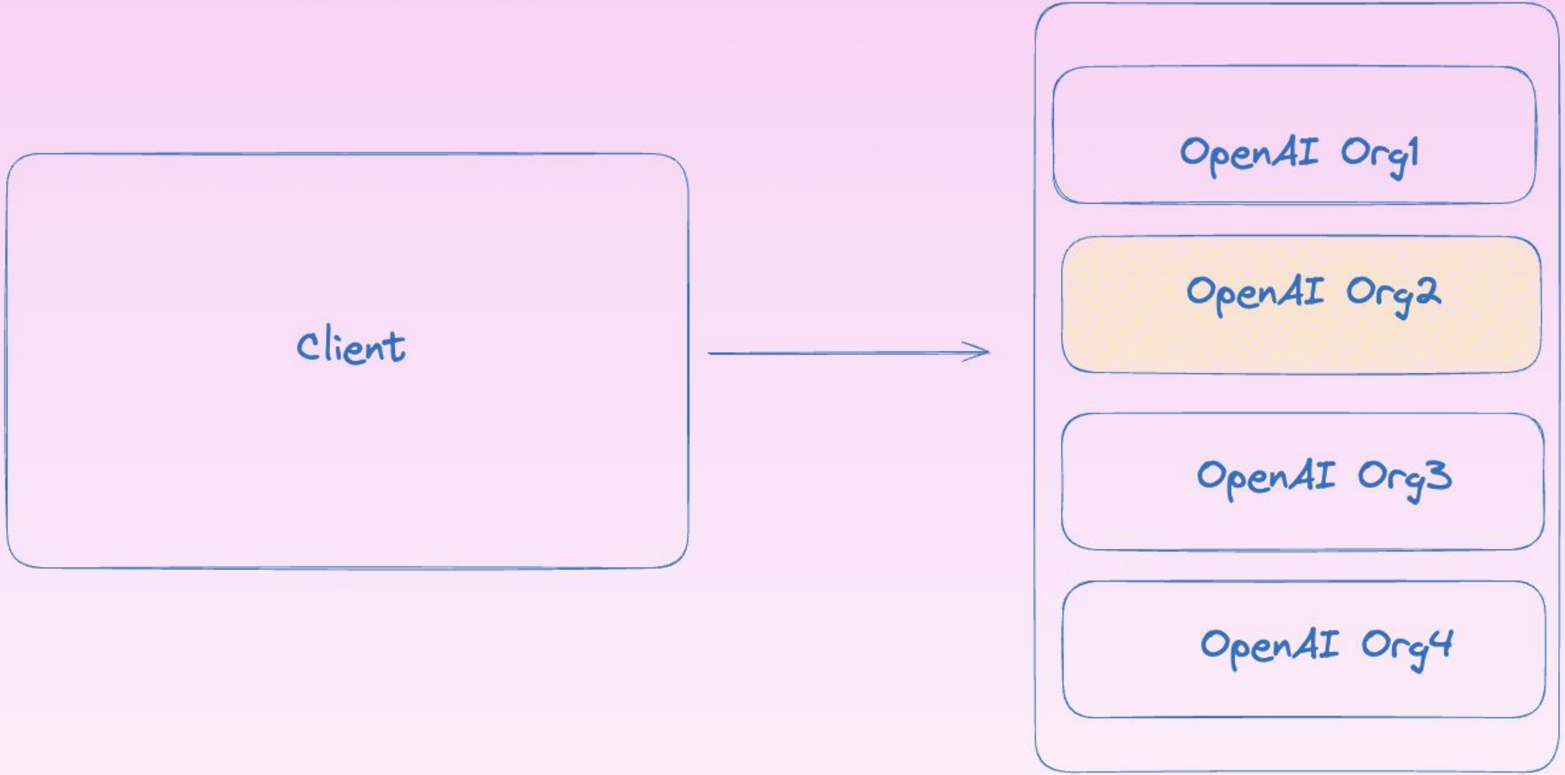
Resilient Multi Org Solution



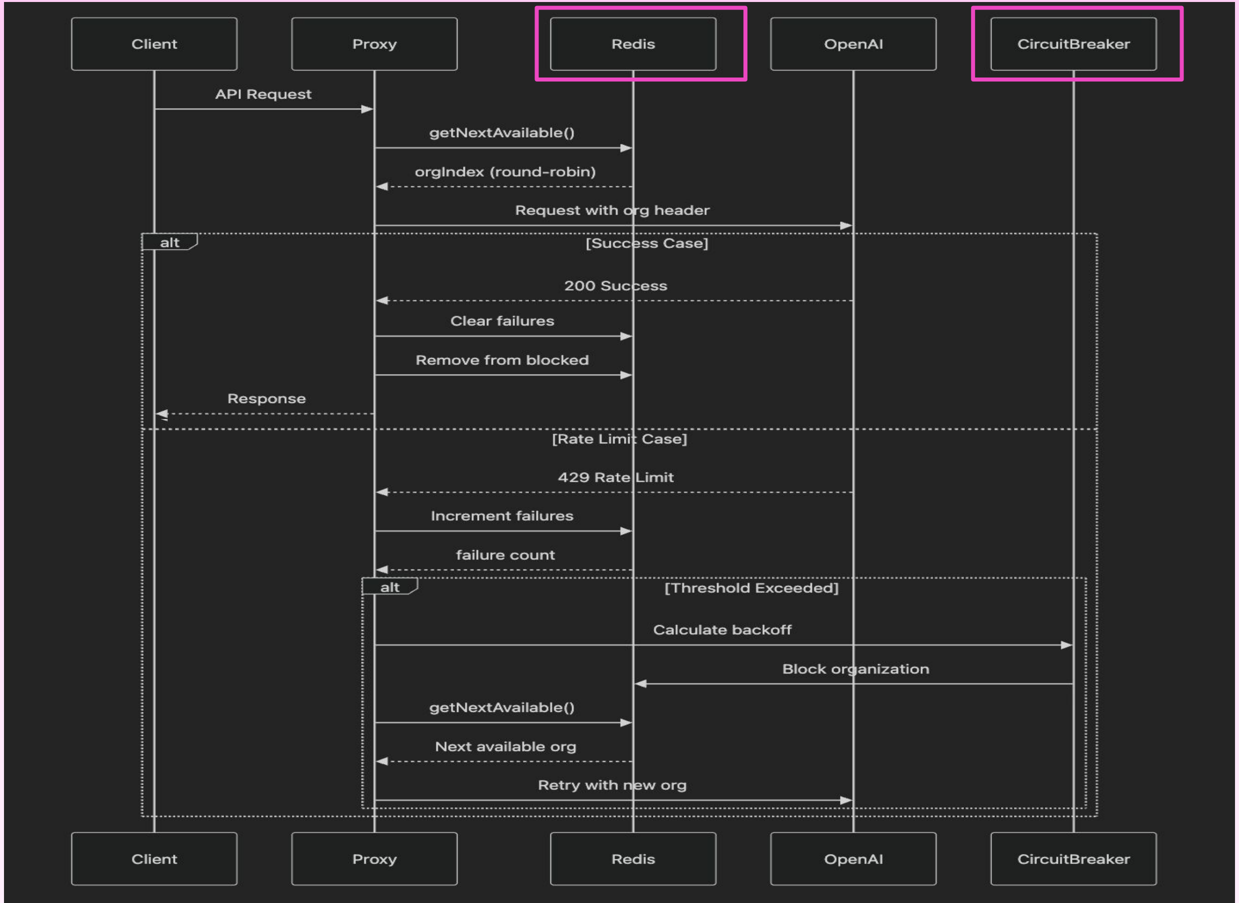
Resilient Multi Org Solution



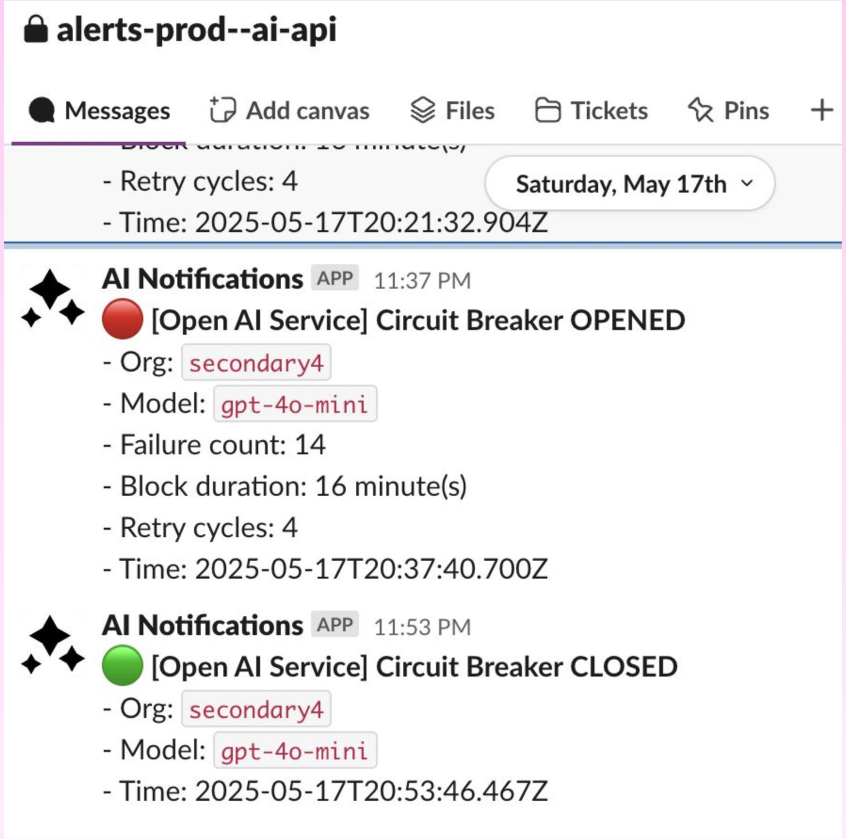
Resilient Multi Org Solution



Resilient Multi Org Solution



Resilient Multi Org Solution



- Block duration: 16 minute(s)
- Retry cycles: 4
- Time: 2025-05-17T20:21:32.904Z

AI Notifications APP 11:37 PM

🔴 [Open AI Service] Circuit Breaker OPENED

- Org: `secondary4`
- Model: `gpt-4o-mini`
- Failure count: 14
- Block duration: 16 minute(s)
- Retry cycles: 4
- Time: 2025-05-17T20:37:40.700Z

AI Notifications APP 11:53 PM

🟢 [Open AI Service] Circuit Breaker CLOSED

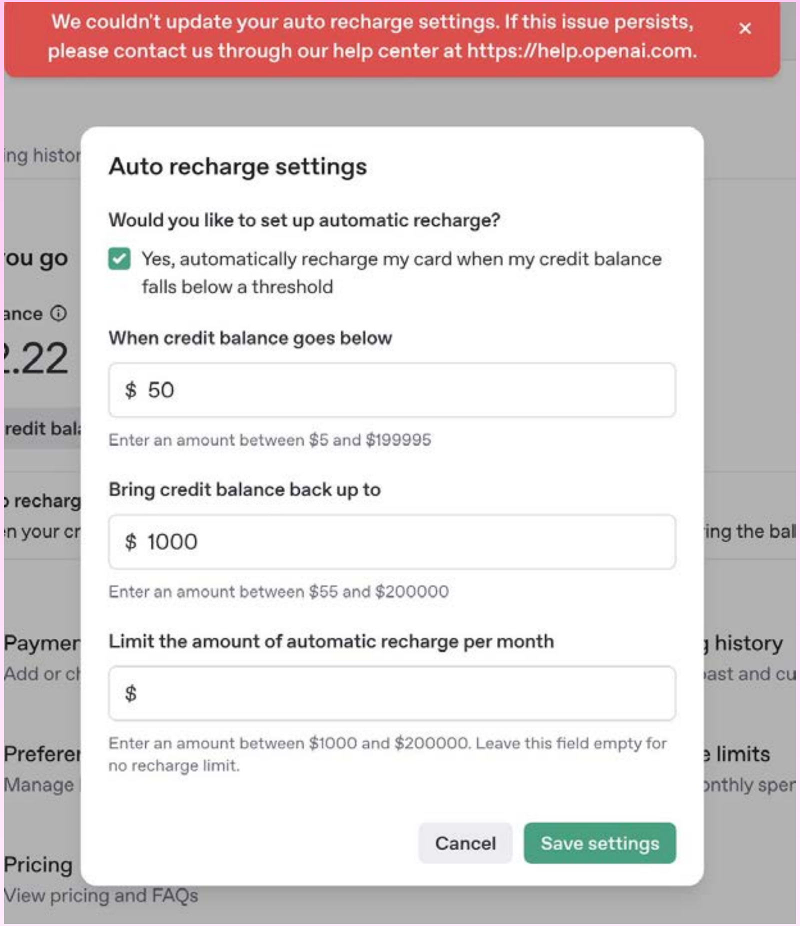
- Org: `secondary4`
- Model: `gpt-4o-mini`
- Time: 2025-05-17T20:53:46.467Z

OpenAI Locked our account😱

21.03.25

AI Site Planner Launch🥳

23.03.25



We couldn't update your auto recharge settings. If this issue persists, please contact us through our help center at https://help.openai.com. ×

Auto recharge settings

Would you like to set up automatic recharge?

☑ Yes, automatically recharge my card when my credit balance falls below a threshold

When credit balance goes below

$ 50

Enter an amount between $5 and $199995

Bring credit balance back up to

$ 1000

Enter an amount between $55 and $200000

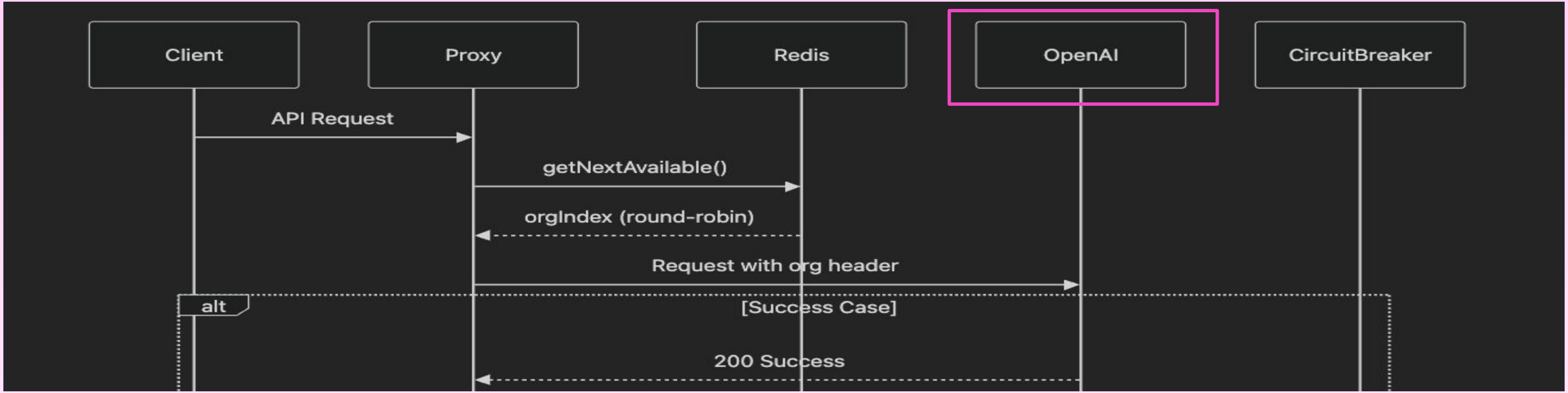
Limit the amount of automatic recharge per month

$

Enter an amount between $1000 and $200000. Leave this field empty for no recharge limit.

Claude

	OpenAI	Claude
Structured output	Yes	No
Schema validation	Native	No
Cost	Less expensive	More expensive
System prompts	Existing	Need adjustments
Token context window	Smaller	Larger



Switch the LLM Provider in 1 click

OPEN
SOURCE

Takeaways - Being a Resilient AI Solution

1. Failure is inevitable
2. Use Circuit Breaker for fault tolerance
3. If needed, have another AI vendor for redundancy

#3 - Continuously Improving AI Solution

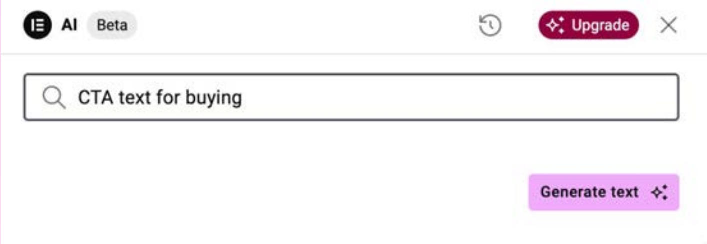
What our users expect?

What our users ACTUALLY expect?

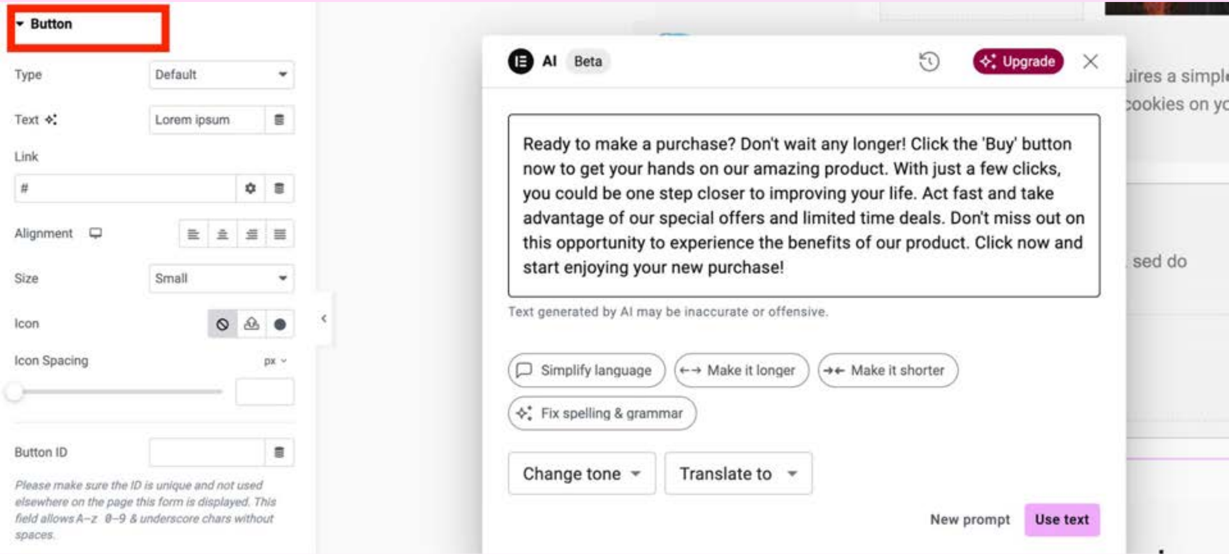
Are we doing a good job? 🤔

How to evaluate AI Performance?

Text



CTA text for buying



Type: Default

Text: Lorem ipsum

Link: #

Size: Small

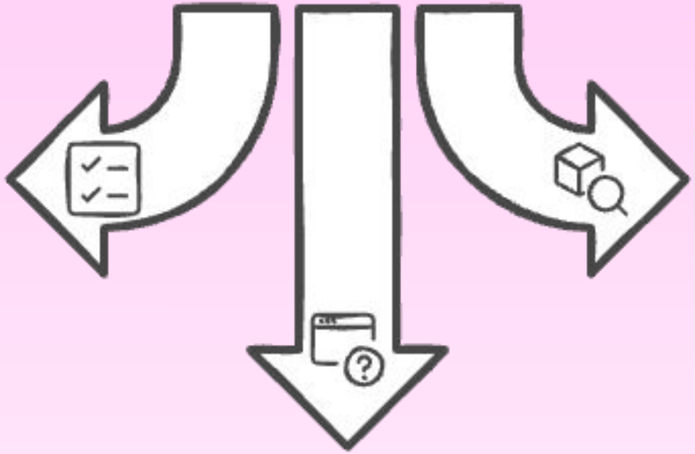
Button ID

Please make sure the ID is unique and not used elsewhere on the page this form is displayed. This field allows A-z 0-9 & underscore chars without spaces.

Ready to make a purchase? Don't wait any longer! Click the 'Buy' button now to get your hands on our amazing product. With just a few clicks, you could be one step closer to improving your life. Act fast and take advantage of our special offers and limited time deals. Don't miss out on this opportunity to experience the benefits of our product. Click now and start enjoying your new purchase!

Text generated by AI may be inaccurate or offensive.

How to evaluate AI Performance?

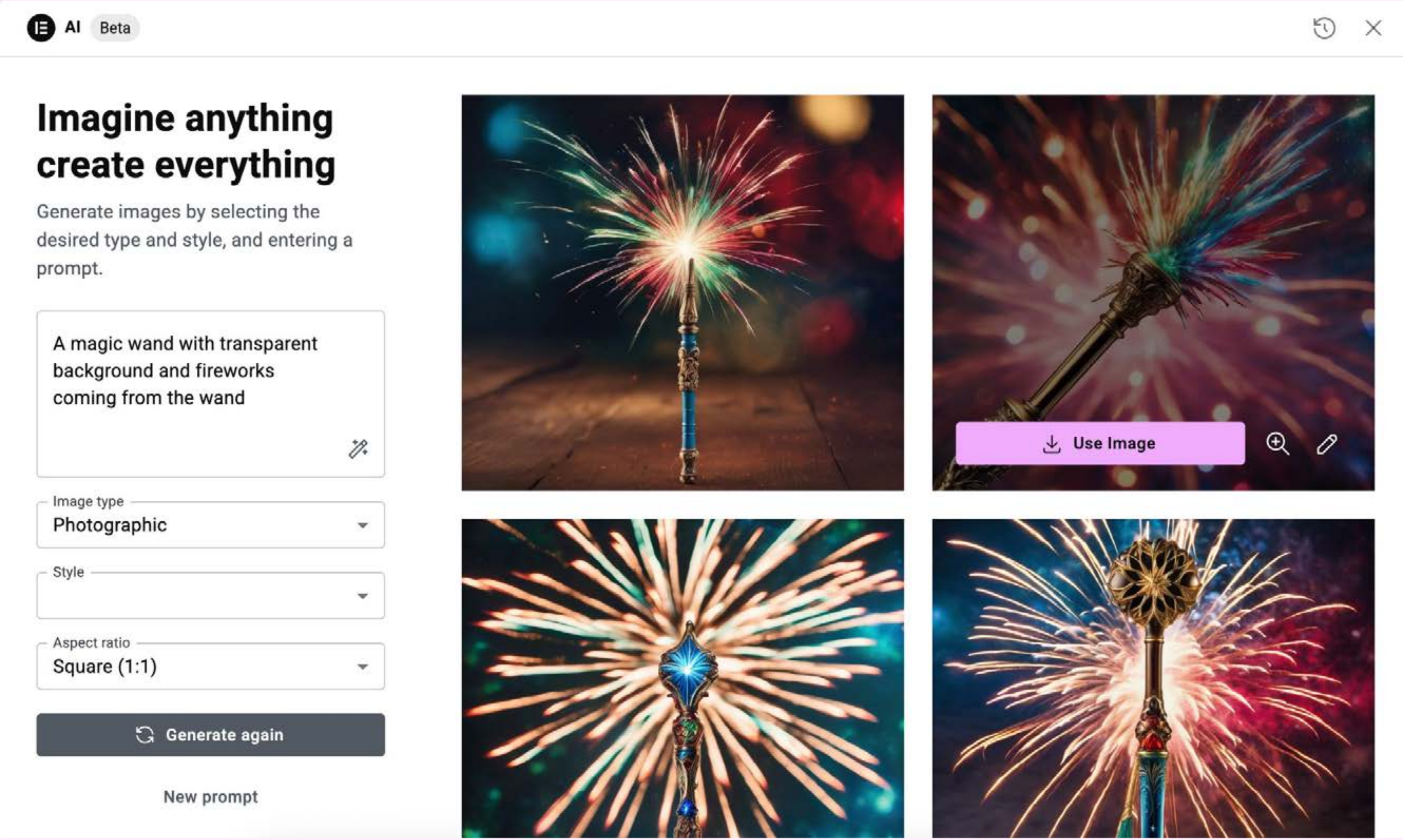


Text Evaluation

Easy to evaluate using test cases

Define Success Metric

“Insert Rate”



Imagine anything create everything

Generate images by selecting the desired type and style, and entering a prompt.

A magic wand with transparent background and fireworks coming from the wand

Image type
Photographic

Style

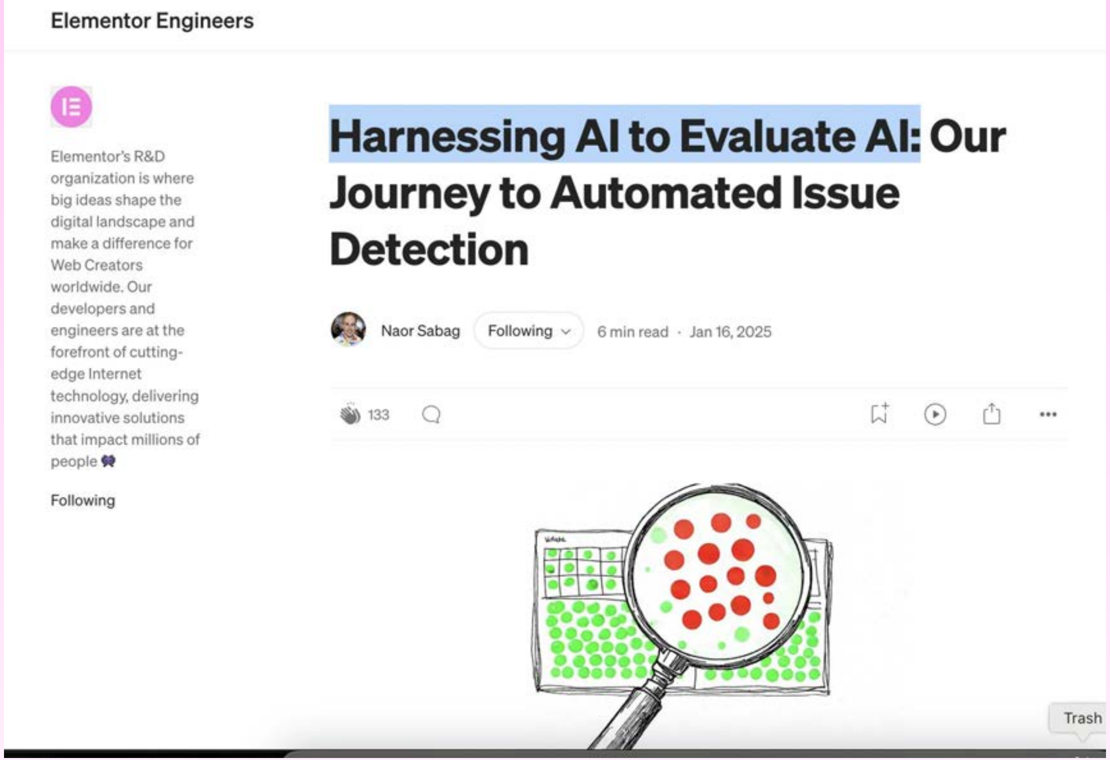
Aspect ratio
Square (1:1)

Insert Rate

User Input	Enhancent Prompt	Full Prompt	Result	Inserted

1. Manually going over the results
2. Putting it in ChatGPT
3. Clustering Job😎

Harnessing AI to Evaluate AI



Elementor Engineers

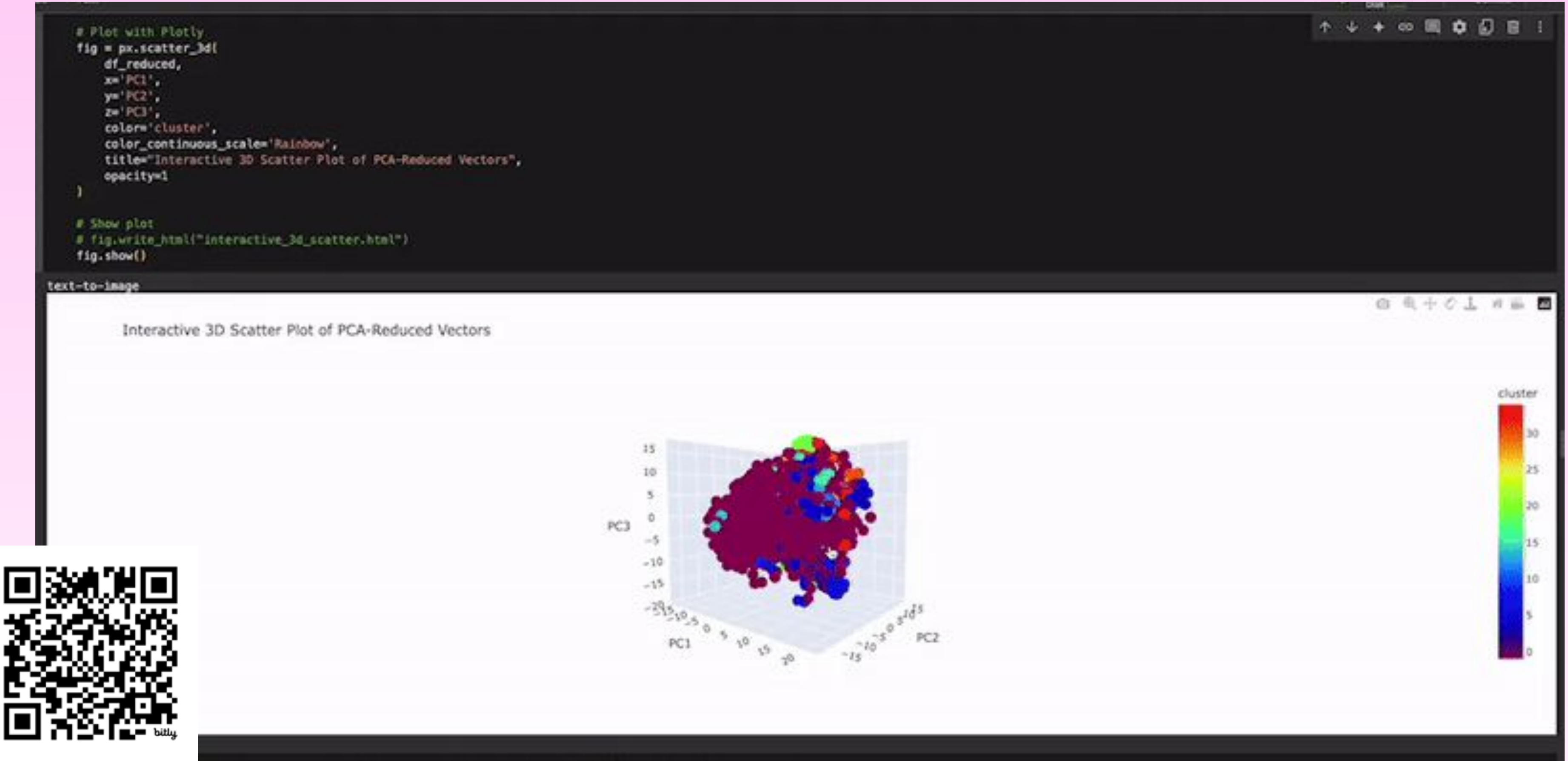
Elementor's R&D organization is where big ideas shape the digital landscape and make a difference for Web Creators worldwide. Our developers and engineers are at the forefront of cutting-edge Internet technology, delivering innovative solutions that impact millions of people 💜

Following

Harnessing AI to Evaluate AI: Our Journey to Automated Issue Detection

Naor Sabag · Following · 6 min read · Jan 16, 2025

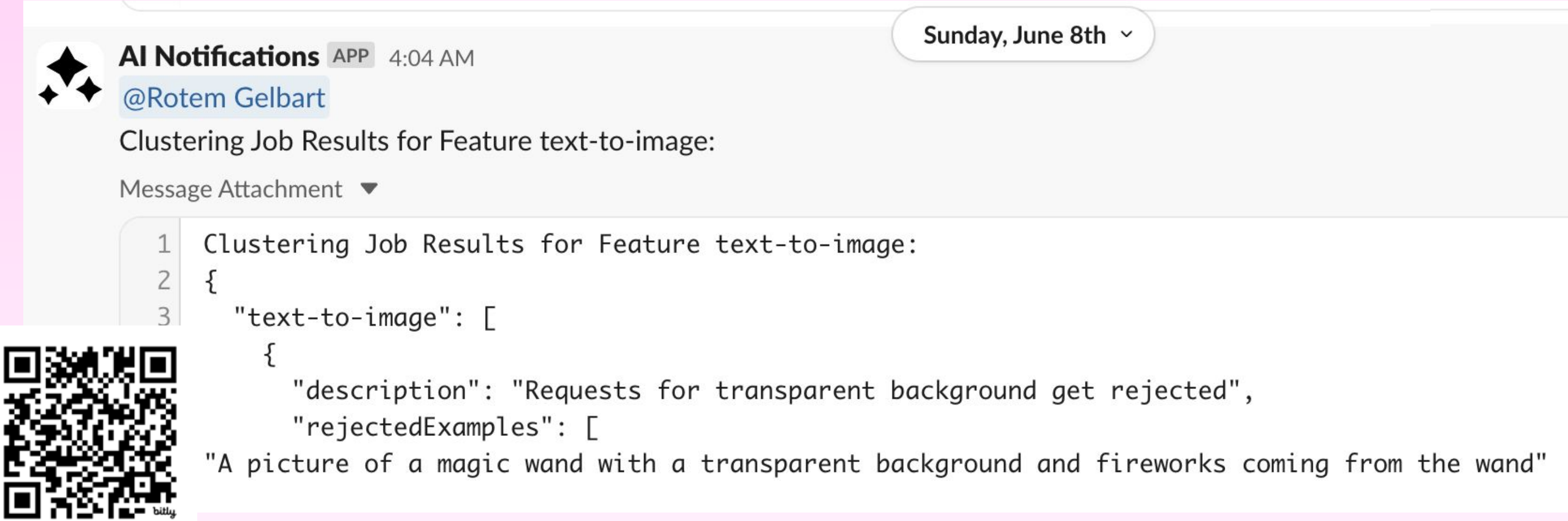
Viewing the Clusters in 3D



```
# Plot with Plotly
fig = px.scatter_3d(
    df_reduced,
    x='PC1',
    y='PC2',
    z='PC3',
    color='cluster',
    color_continuous_scale='Rainbow',
    title="Interactive 3D Scatter Plot of PCA-Reduced Vectors",
    opacity=1
)

# Show plot
# fig.write_html("interactive_3d_scatter.html")
fig.show()
```

Sharing Results in Slack



Sunday, June 8th

AI Notifications APP 4:04 AM

@Rotem Gelbart

Clustering Job Results for Feature text-to-image:

Message Attachment

```
Clustering Job Results for Feature text-to-image:
{
  "text-to-image": [
    {
      "description": "Requests for transparent background get rejected",
      "rejectedExamples": [
"A picture of a magic wand with a transparent background and fireworks coming from the wand"
```

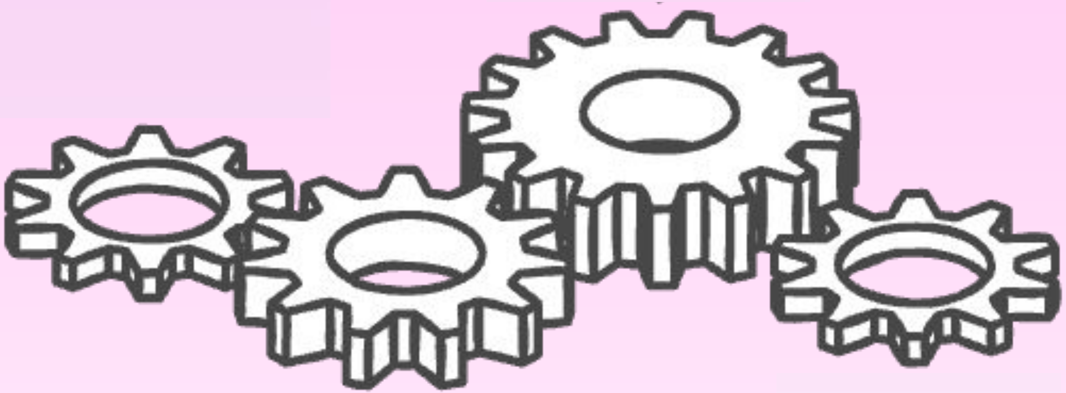
AI Found REAL issues!

- Japanese support
- Transparent background
- Text length
- Layouts for online courses
- Conditional Forms support
- ...

Takeaways - Continuously Improving AI Solution

- Persist all of your data
- Define a success metric
- Measure it
- Pick a way to evaluate success
    - Start manually
    - Gradually automate

To Summarize...



Scan to get the list of tools
and my contact details

May the AI be with you