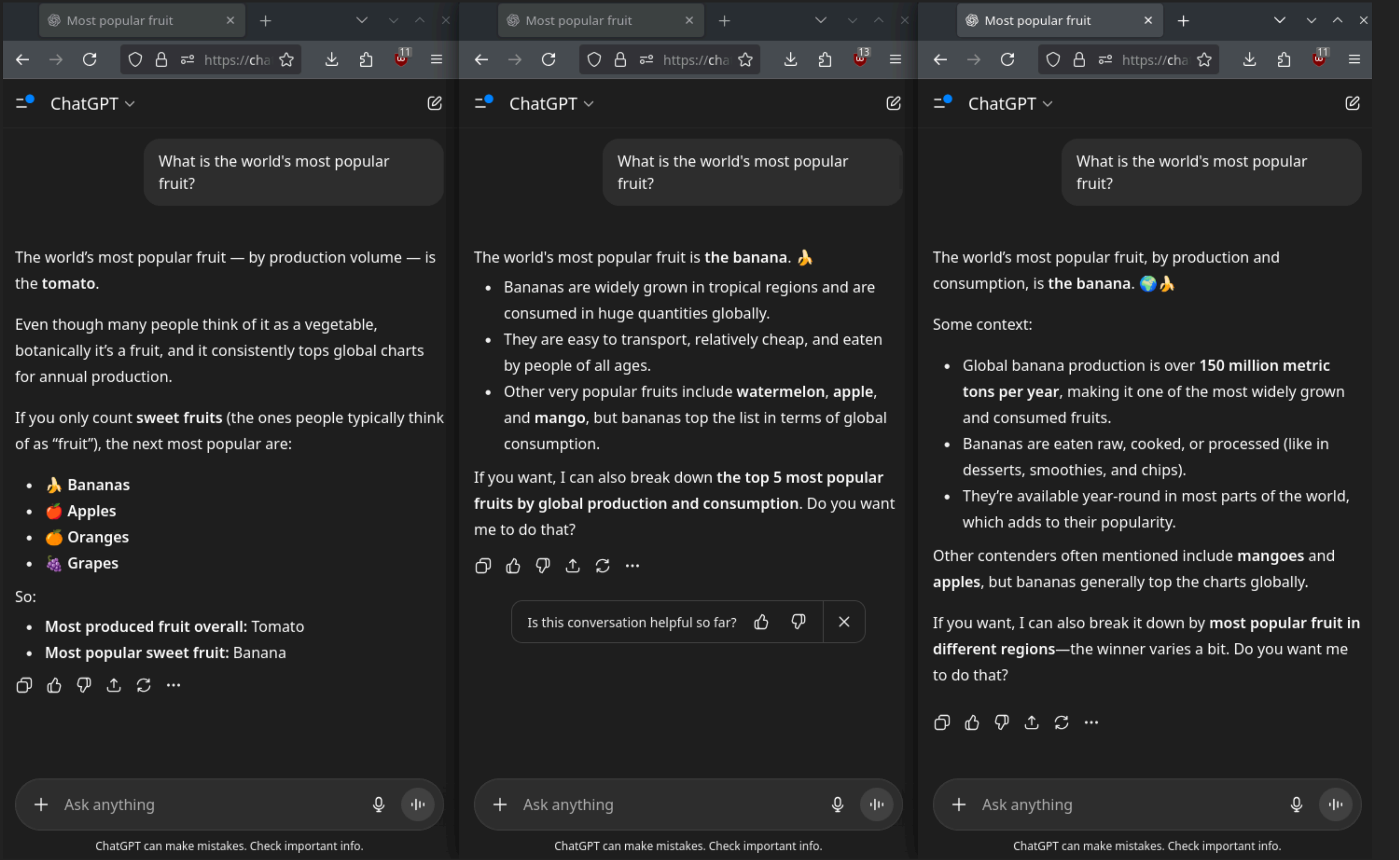
BUILDING RELIABLE AI FEATURES

THE PROBLEM

- AI introduces new failure modes
- Regressions you don't control
- Probabilistic output

WHY AI RELIABILITY IS HARD

- Probability is what makes AI powerful
- It's also a huge business liability
- How to "build in quality" traditionally is well-known
- For AI systems, it's an evolving practice



What is the world's most popular fruit?

The world's most popular fruit — by production volume — is the **tomato**.

Even though many people think of it as a vegetable, botanically it's a fruit, and it consistently tops global charts for annual production.

If you only count **sweet fruits** (the ones people typically think of as "fruit"), the next most popular are:

- 🍌 **Bananas**
- 🍎 **Apples**
- 🍊 **Oranges**
- 🍇 **Grapes**

So:

- **Most produced fruit overall:** Tomato
- **Most popular sweet fruit:** Banana

ChatGPT can make mistakes. Check important info.

What is the world's most popular fruit?

The world's most popular fruit is **the banana**. 🍌

- Bananas are widely grown in tropical regions and are consumed in huge quantities globally.
- They are easy to transport, relatively cheap, and eaten by people of all ages.
- Other very popular fruits include **watermelon**, **apple**, and **mango**, but bananas top the list in terms of global consumption.

If you want, I can also break down **the top 5 most popular fruits by global production and consumption**. Do you want me to do that?

Is this conversation helpful so far?

ChatGPT can make mistakes. Check important info.

What is the world's most popular fruit?

The world's most popular fruit, by production and consumption, is **the banana**. 🌍🍌

Some context:

- Global banana production is over **150 million metric tons per year**, making it one of the most widely grown and consumed fruits.
- Bananas are eaten raw, cooked, or processed (like in desserts, smoothies, and chips).
- They're available year-round in most parts of the world, which adds to their popularity.

Other contenders often mentioned include **mangoes** and **apples**, but bananas generally top the charts globally.

If you want, I can also break it down by **most popular fruit in different regions**—the winner varies a bit. Do you want me to do that?

ChatGPT can make mistakes. Check important info.

🍅🍌🍌

WHY DOES THIS HAPPEN?

- LLMs are good at *specific things*
- things that can impact your feature
    - LLM Model
    - Prompts/code
    - User input
    - Probability

WHY DOES THIS MATTER?

> *ChatGPT can make mistakes. Check important info.*

OpenAI

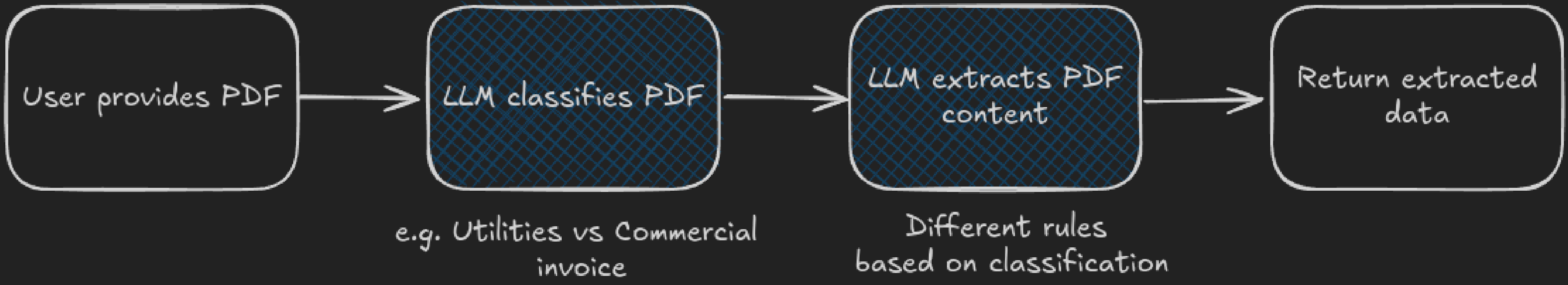
WHY DOES THIS MATTER?

regarding users of AI in a business context

*Many **rely on AI output without evaluating accuracy** (66%) and are making mistakes in their work due to AI (56%).*

KPMG, 2025

AN EXAMPLE



User provides PDF

LLM classifies PDF

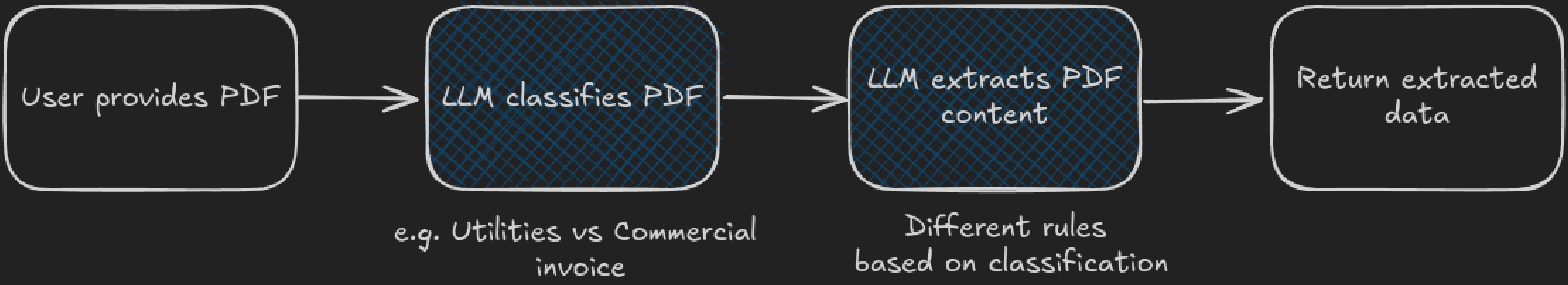
e.g. Utilities vs Commercial invoice

LLM extracts PDF content

Different rules based on classification

Return extracted data

AN EXAMPLE



1. Incorrect classification
2. Prompt variability
3. Input data variability
4. Model changes
5. Hallucinations

SOLUTION?

- Version control
- Observability
- Evaluation Testing
- Iteration

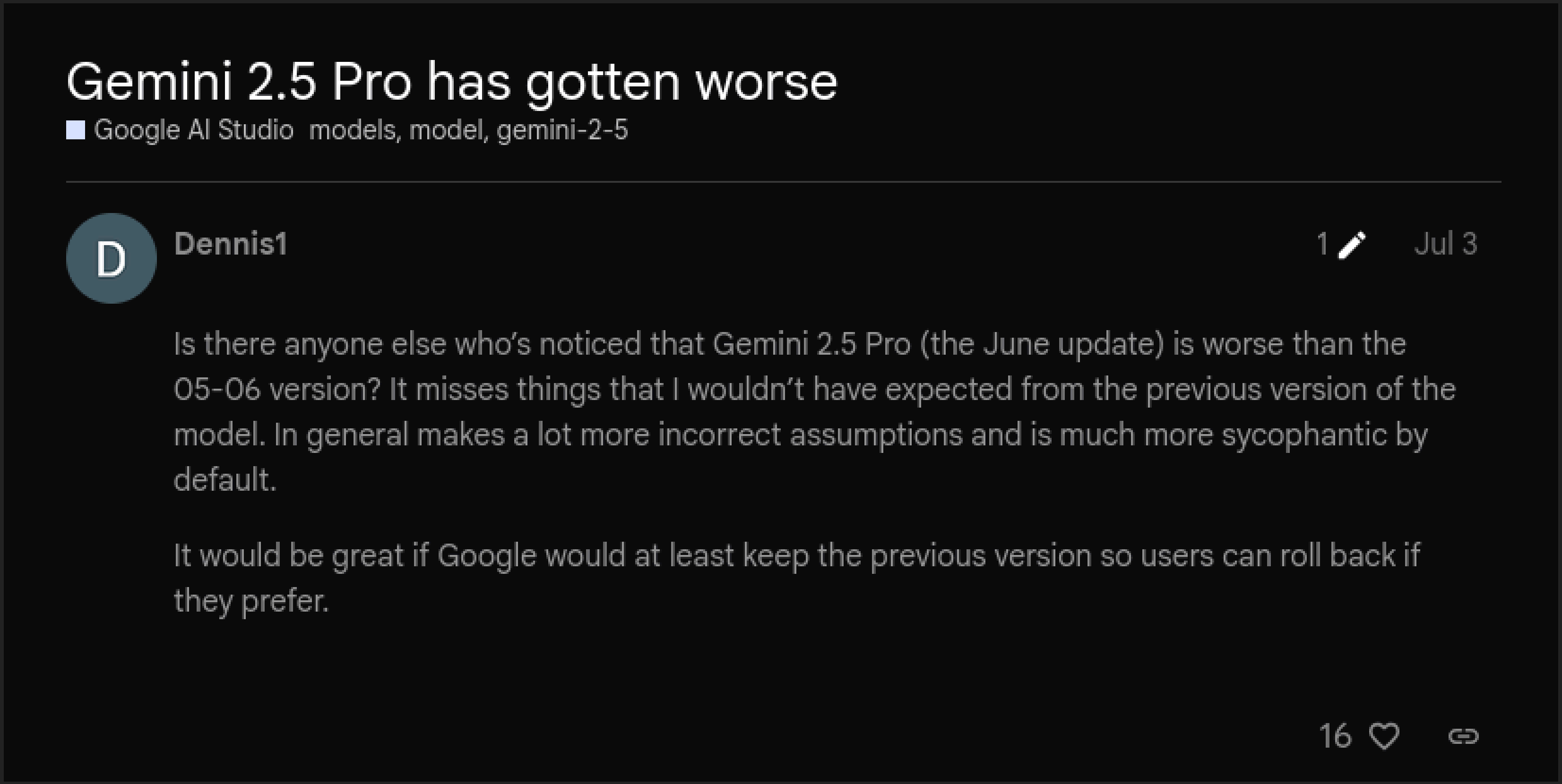
VERSION CONTROL

Consider

`gpt-4.1` vs `gpt-4.1-2025-04-14`

or

`gemini-2.5-flash` vs `gemini-flash-latest`



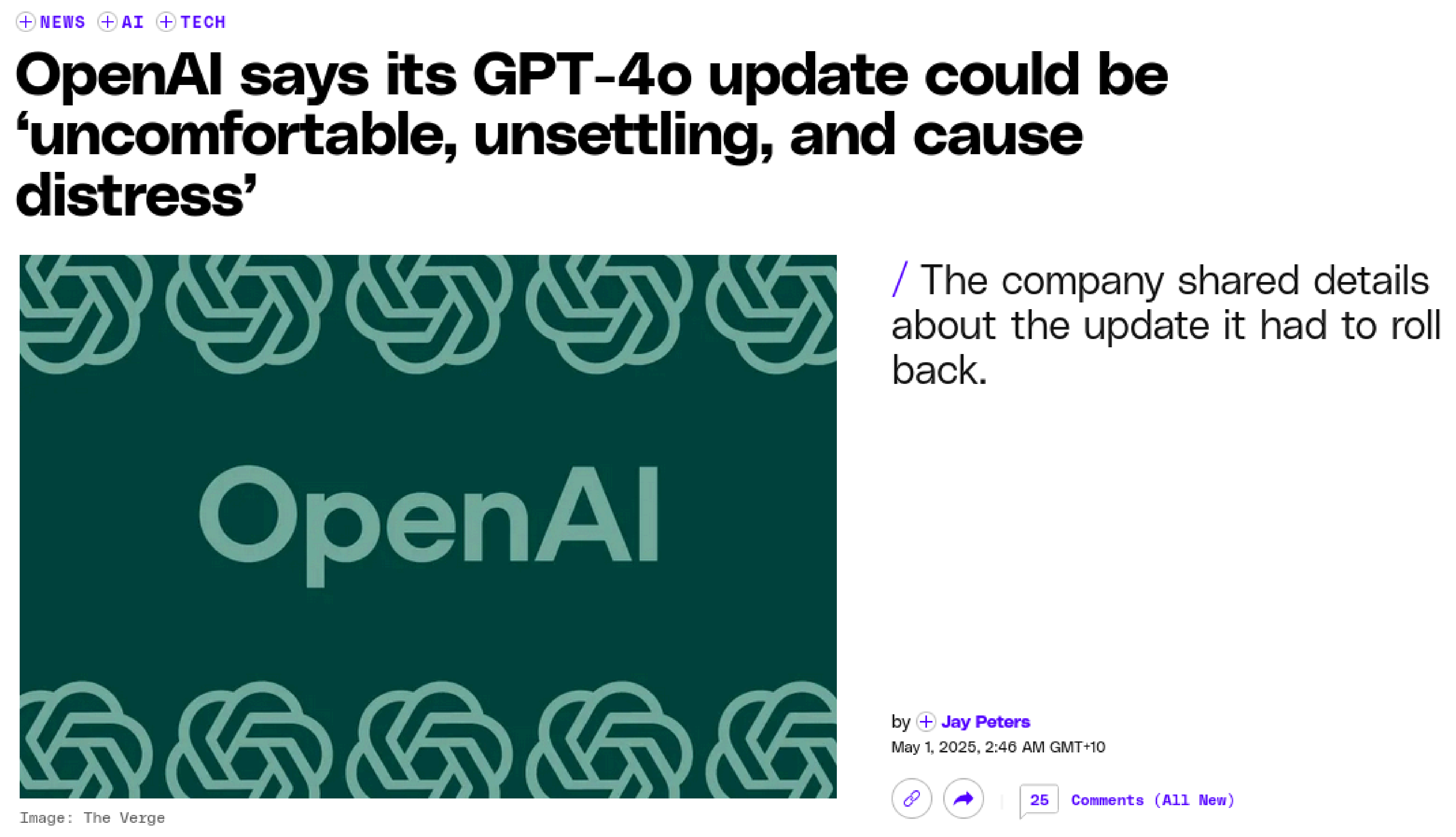
Gemini 2.5 Pro has gotten worse

Google AI Studio models, model, gemini-2-5

Dennis1 1 Jul 3

Is there anyone else who's noticed that Gemini 2.5 Pro (the June update) is worse than the 05-06 version? It misses things that I wouldn't have expected from the previous version of the model. In general makes a lot more incorrect assumptions and is much more sycophantic by default.

It would be great if Google would at least keep the previous version so users can roll back if they prefer.



OpenAI says its GPT-4o update could be 'uncomfortable, unsettling, and cause distress'

Image: The Verge

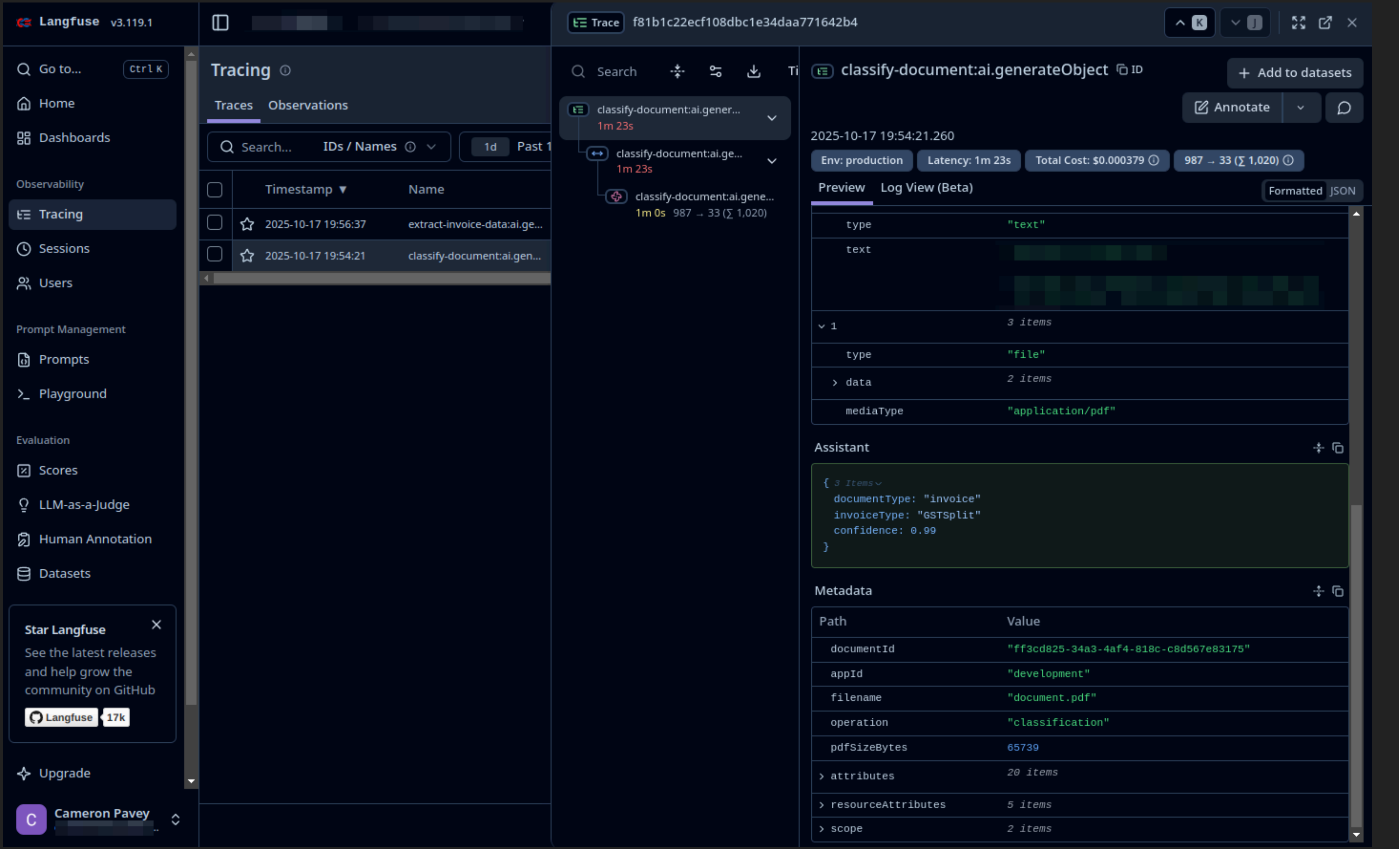
/ The company shared details about the update it had to roll back.

by **Jay Peters**

May 1, 2025, 2:46 AM GMT+10

OBSERVABILITY

- Similar to traditional observability
- allows you to see inputs, outputs, costs
- see how your system is performing
- track down traces related to user feedback
    - capture feedback about poor performance



[redacted]

[redacted]

	Timestamp	Name
☐ ☆	2025-10-17 19:56:37	extract-invoice-data:ai.ge...
☐ ☆	2025-10-17 19:54:21	classify-document:ai.gen...

Trace f81b1c22ecf108dbc1e34daa771642b4

- classify-document:ai.gener... 1m 23s
  - classify-document:ai.ge... 1m 23s
    - classify-document:ai.gene... 1m 0s 987 → 33 (∑ 1,020)

classify-document:ai.generateObject ID

2025-10-17 19:54:21.260

Env: production | Latency: 1m 23s | Total Cost: $0.000379 | 987 → 33 (∑ 1,020)

type	"text"
text	[redacted]
1	3 items
type	"file"
data	2 items
mediaType	"application/pdf"

Assistant

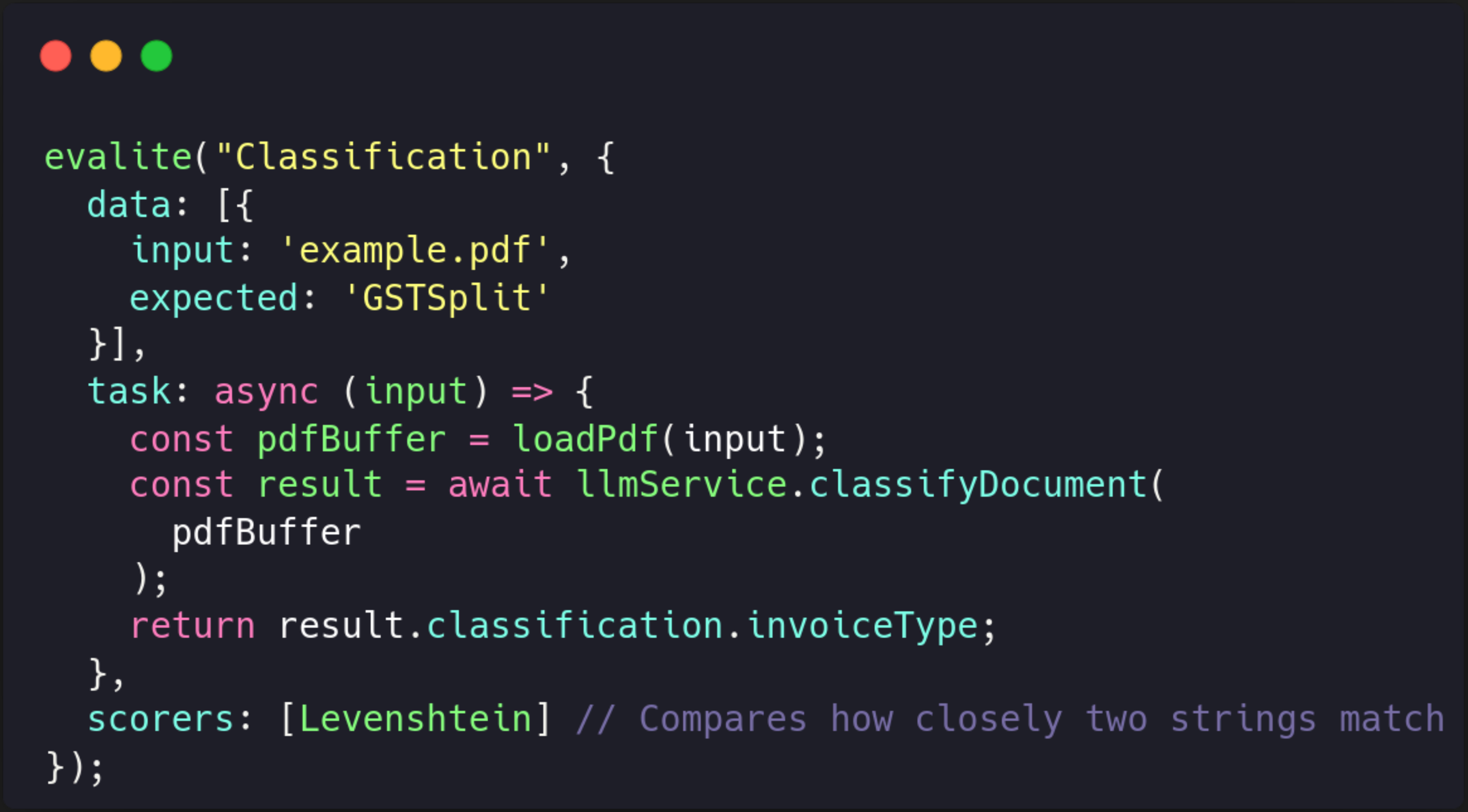
```
{ 3 Items
  documentType: "invoice"
  invoiceType: "GSTSplit"
  confidence: 0.99
}
```

Metadata

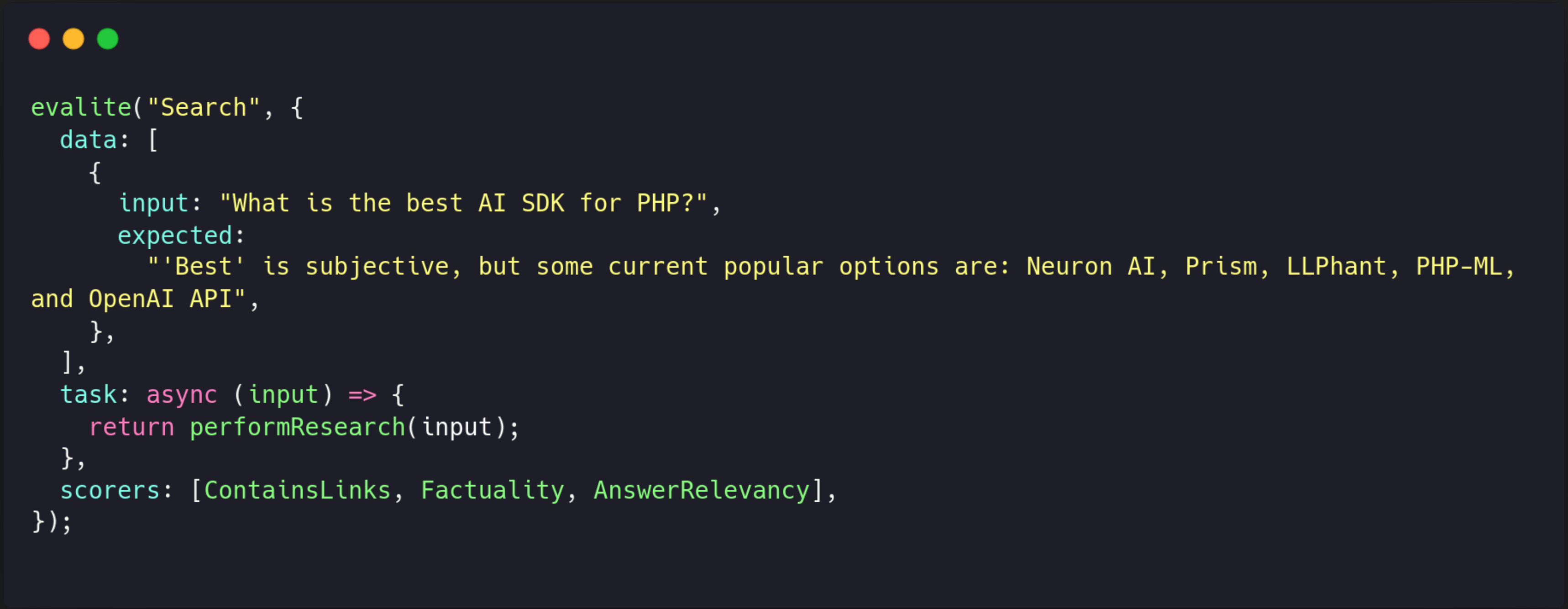
Path	Value
documentId	"ff3cd825-34a3-4af4-818c-c8d567e83175"
appId	"development"
filename	"document.pdf"
operation	"classification"
pdfSizeBytes	65739
attributes	20 items
resourceAttributes	5 items
scope	2 items

EVALS

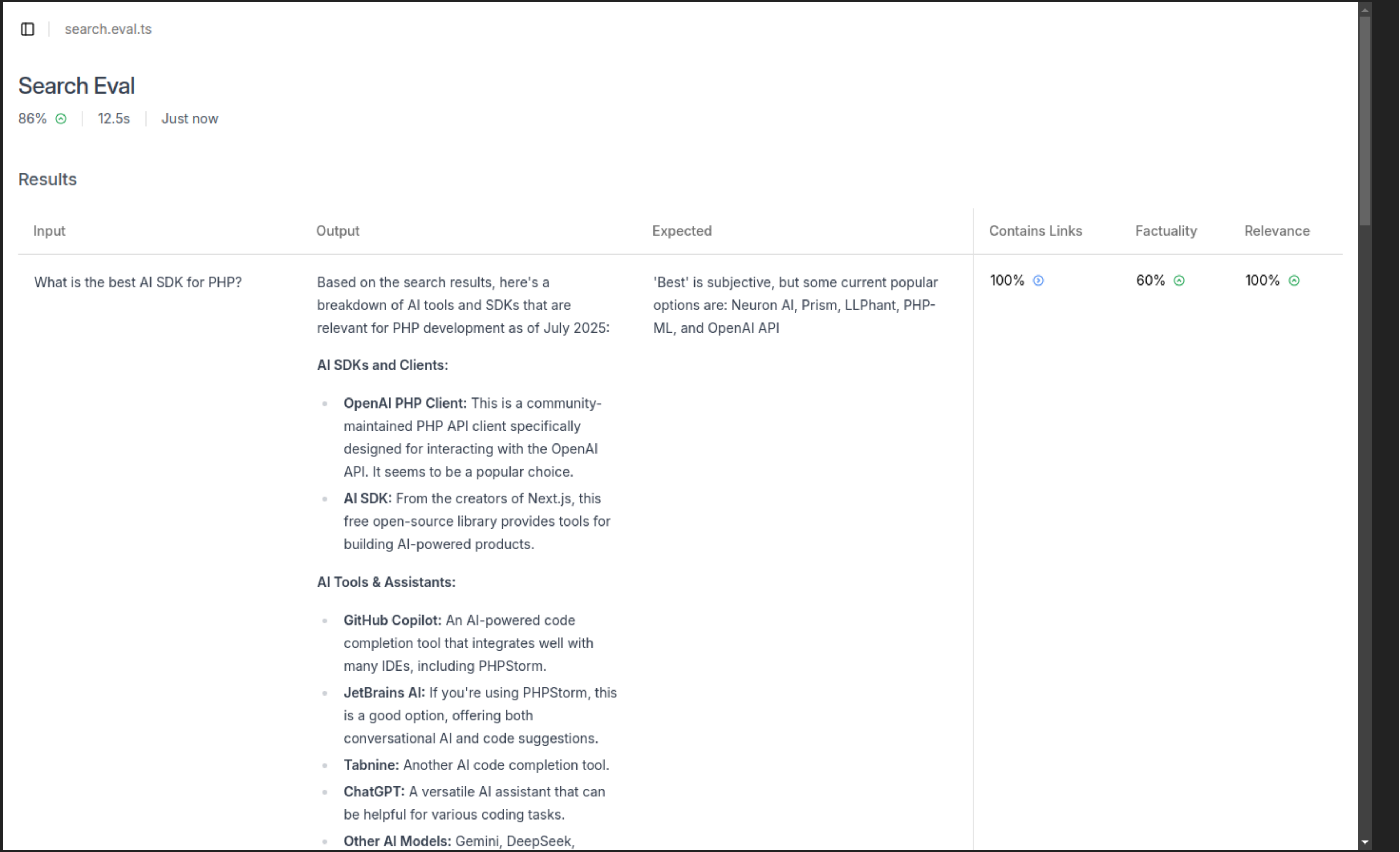
- Testing suitability of LLM output



```
evalite("Classification", {
  data: [{
    input: 'example.pdf',
    expected: 'GSTSplit'
  }],
  task: async (input) => {
    const pdfBuffer = loadPdf(input);
    const result = await llmService.classifyDocument(
      pdfBuffer
    );
    return result.classification.invoiceType;
  },
  scorers: [Levenshtein] // Compares how closely two strings match
});
```



```
evalite("Search", {
  data: [
    {
      input: "What is the best AI SDK for PHP?",
      expected:
        "'Best' is subjective, but some current popular options are: Neuron AI, Prism, LLPhant, PHP-ML, and OpenAI API",
    },
  ],
  task: async (input) => {
    return performResearch(input);
  },
  scorers: [ContainsLinks, Factuality, AnswerRelevancy],
});
```



search.eval.ts

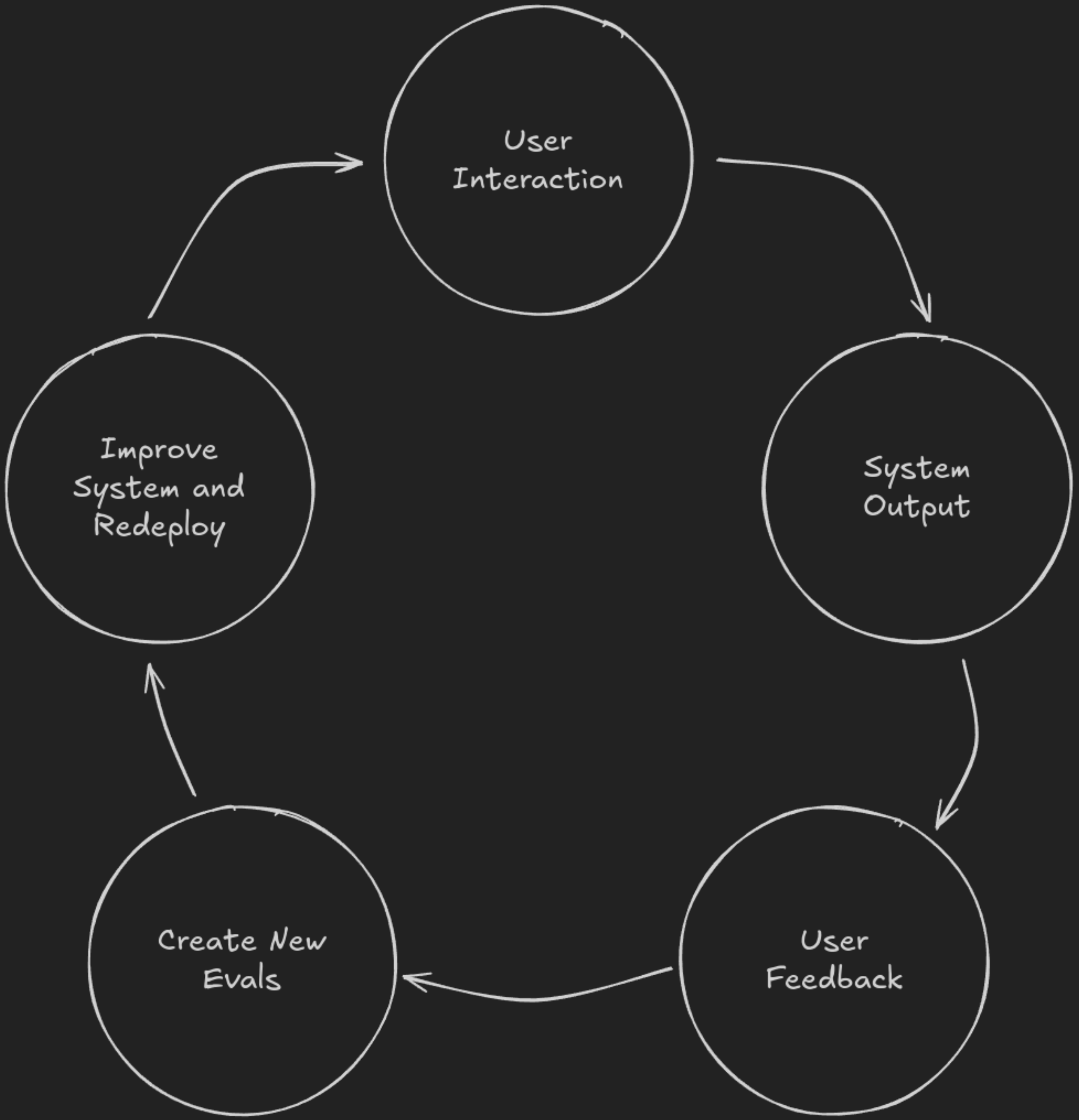
Search Eval

86% | 12.5s | Just now

Results

Input	Output	Expected	Contains Links	Factuality	Relevance
What is the best AI SDK for PHP?	Based on the search results, here's a breakdown of AI tools and SDKs that are relevant for PHP development as of July 2025: **AI SDKs and Clients:** - **OpenAI PHP Client:** This is a community-maintained PHP API client specifically designed for interacting with the OpenAI API. It seems to be a popular choice. - **AI SDK:** From the creators of Next.js, this free open-source library provides tools for building AI-powered products. **AI Tools & Assistants:** - **GitHub Copilot:** An AI-powered code completion tool that integrates well with many IDEs, including PHPStorm. - **JetBrains AI:** If you're using PHPStorm, this is a good option, offering both conversational AI and code suggestions. - **Tabnine:** Another AI code completion tool. - **ChatGPT:** A versatile AI assistant that can be helpful for various coding tasks. - **Other AI Models:** Gemini, DeepSeek,	'Best' is subjective, but some current popular options are: Neuron AI, Prism, LLPhant, PHP-ML, and OpenAI API	100%	60%	100%
]]>

CONTINUOUS IMPROVEMENT


User Interaction
System Output
User Feedback
Create New Evals
Improve System and Redeploy

SUMMARY

- AI can be powerful, but introduces new technical challenges
- Solutions follow well-known practices, applied in new ways
    - Version Control
    - Observability
    - Evaluation
    - Iteration

THANK YOU

slides: https://cameronpavey.me/talks/building-reliable-ai-features