From Vibes to Scores

Build your own AI benchmark!

Jonathan Soma
js4571@columbia.edu

What are we doing here?

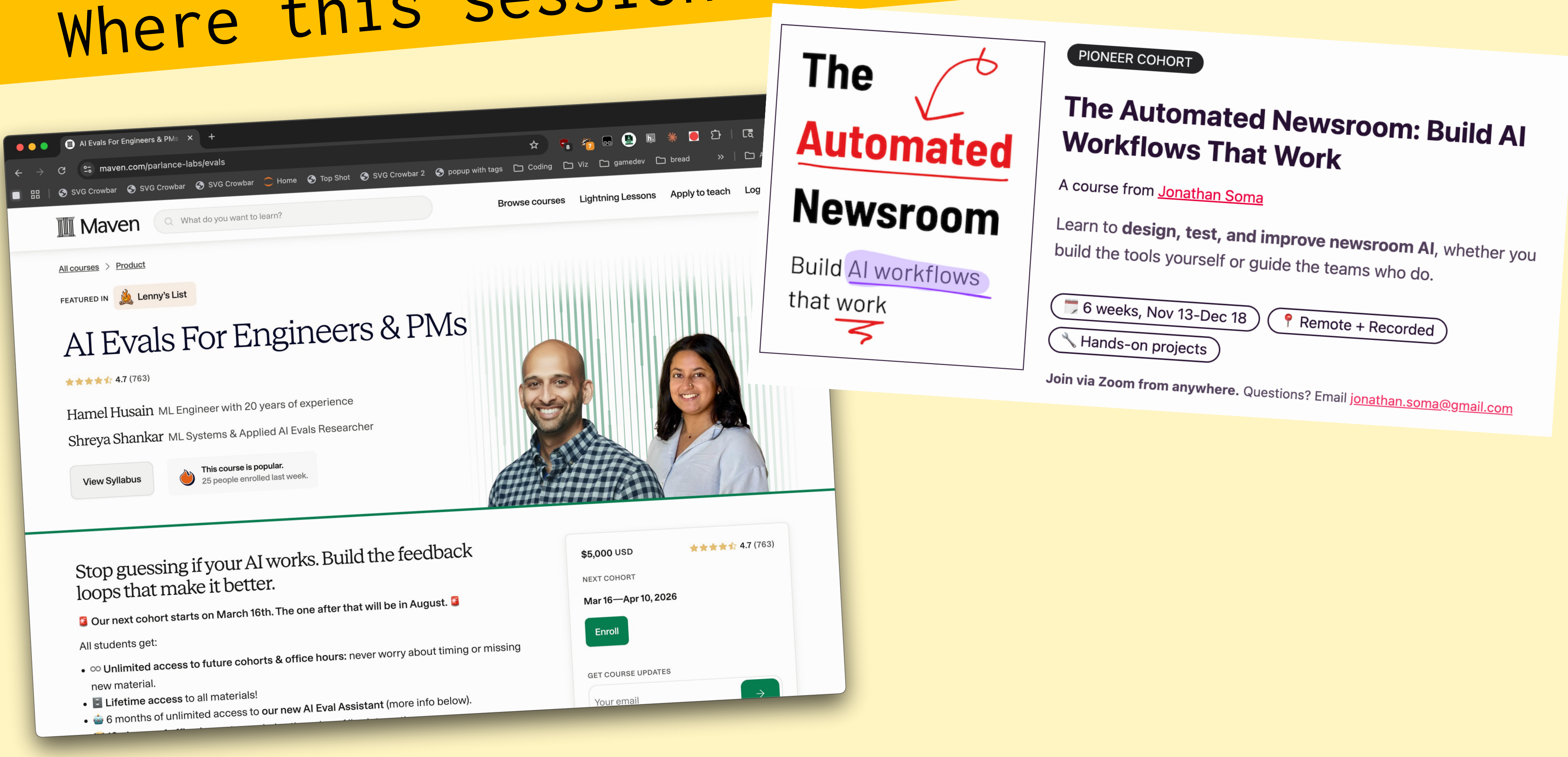
Hi, I’m Soma!

- **Lede Program:** 10 weeks, remote-friendly, summer 2026
- **Data Journalism MS:** 12 months, deadline Jan 12

Operationalize newsroom AI

- **Before:**
  - Little baby prompt
  - Custom GPT
  - Vibes, vibes, vibes
- **After:**
  - Multi-step agentic pipeline that you can put into prod!
  - Kidnapping knowledge from the tech world
  - Fame and fortune

Where this session is from



All courses > Product

FEATURED IN Lenny's List

AI Evals For Engineers & PMs

★★★★☆ 4.7 (763)

Hamel Husain ML Engineer with 20 years of experience

Shreya Shankar ML Systems & Applied AI Evals Researcher

This course is popular.
25 people enrolled last week.

Stop guessing if your AI works. Build the feedback loops that make it better.

🚨 **Our next cohort starts on March 16th. The one after that will be in August.** 🚨

All students get:

- ♾ **Unlimited access to future cohorts & office hours:** never worry about timing or missing new material.
- 🗄 **Lifetime access** to all materials!
- 🤖 6 months of unlimited access to **our new AI Eval Assistant** (more info below).

$5,000 USD ★★★★☆ 4.7 (763)

NEXT COHORT

Mar 16—Apr 10, 2026

The Automated Newsroom

Build AI workflows that work

PIONEER COHORT

The Automated Newsroom: Build AI Workflows That Work

A course from Jonathan Soma

Learn to **design, test, and improve newsroom AI,** whether you build the tools yourself or guide the teams who do.

🗓 6 weeks, Nov 13-Dec 18

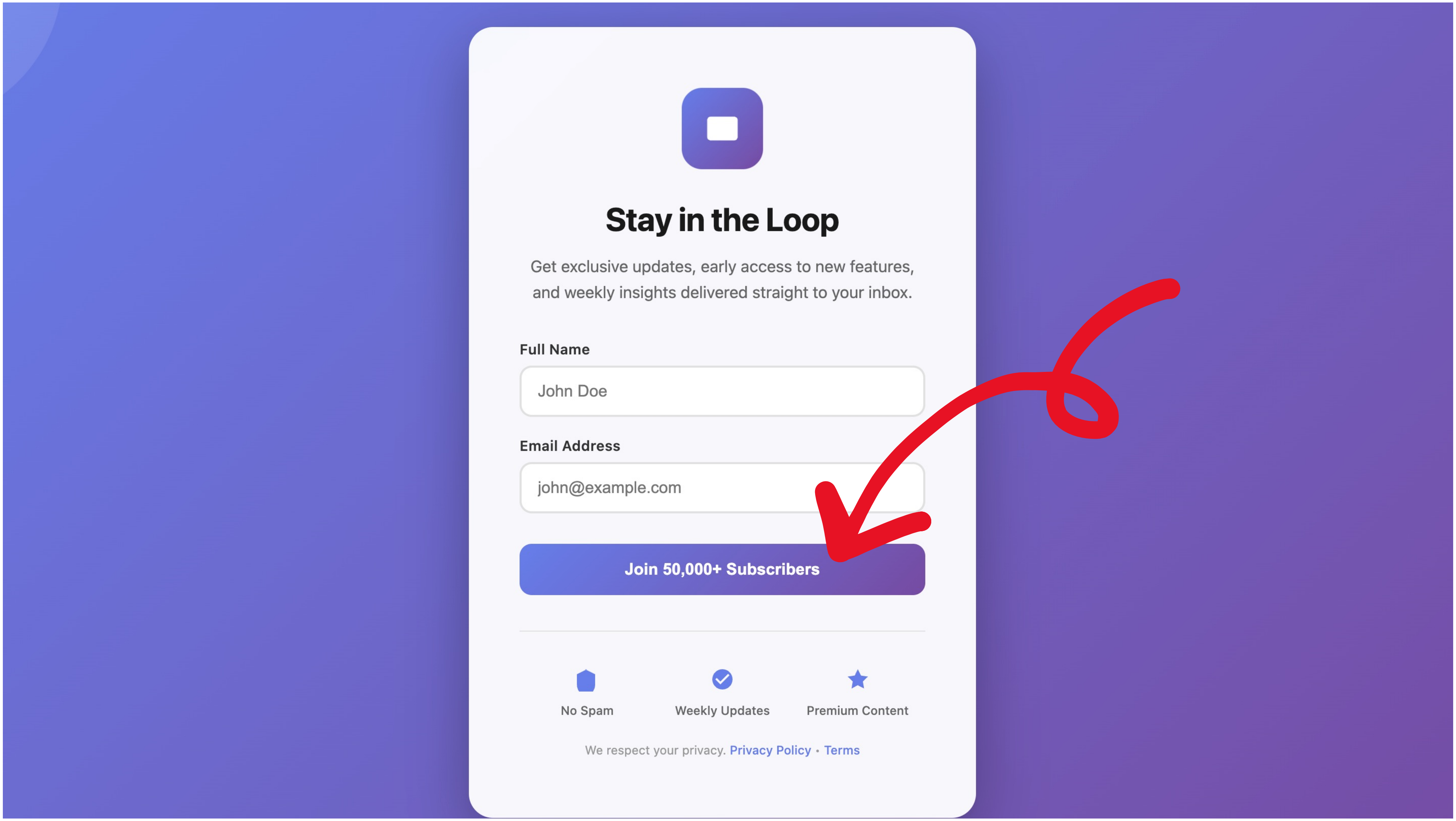
📍 Remote + Recorded

🔧 Hands-on projects

Join via Zoom from anywhere. Questions? Email jonathan.soma@gmail.com

We will learn how to…

- Compare effectiveness of different **AI prompts and models**
- Set up the **processes** to observe the “magic” going on inside of your AI tools
- **Structure eval test sets** to effectively cover your bases so you don’t get blindsided
- **Build agentic AI** tools yourself (and what in the world agentic AI even is)
- Bring non-technical folks and **AI haters** to the table
- Why we even need to do this at all?? What about **vibes???**



Stay in the Loop

Get exclusive updates, early access to new features, and weekly insights delivered straight to your inbox.

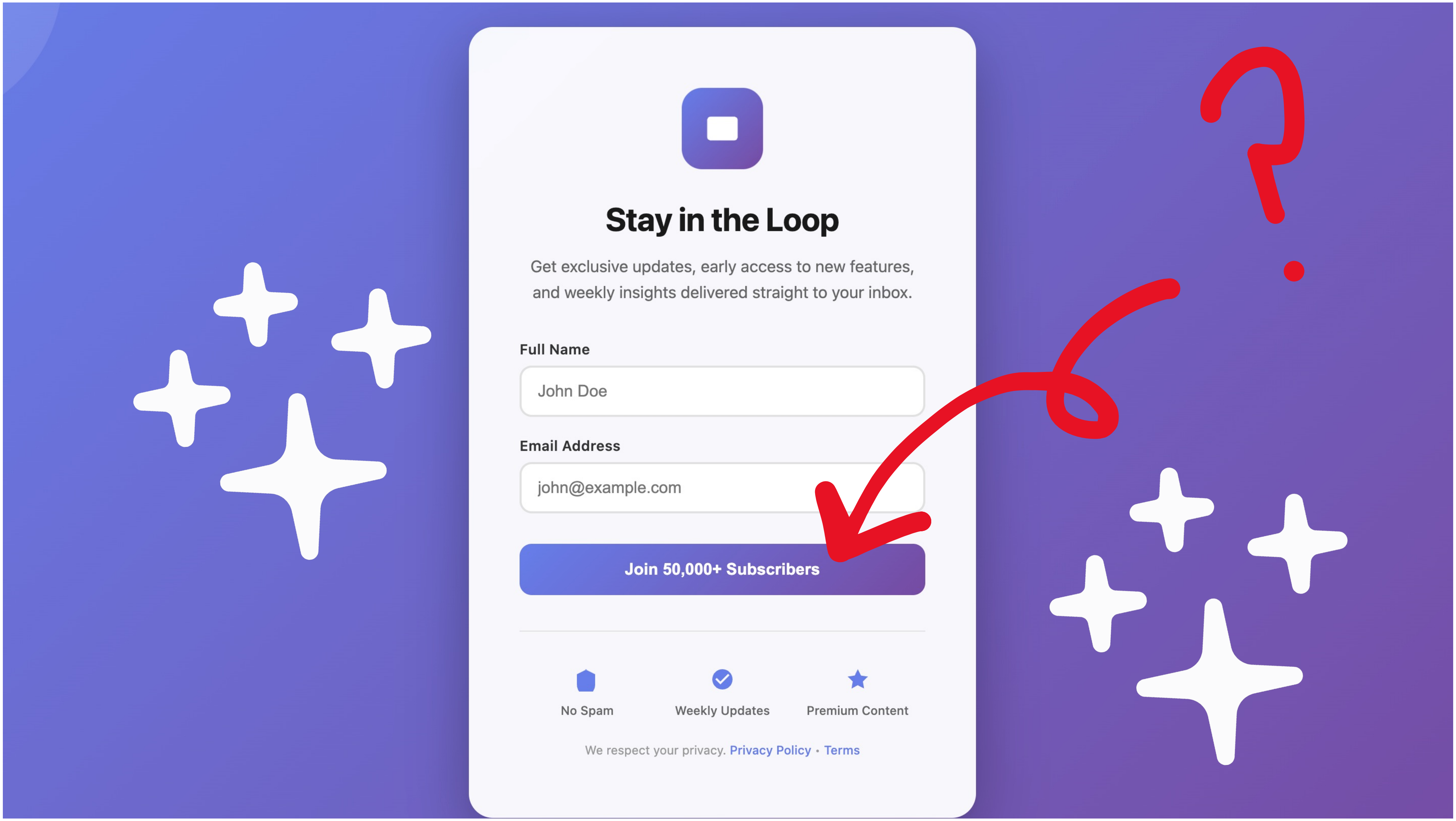
Full Name

John Doe

Email Address

john@example.com

We respect your privacy. Privacy Policy • Terms



Stay in the Loop

Get exclusive updates, early access to new features, and weekly insights delivered straight to your inbox.

Full Name

John Doe

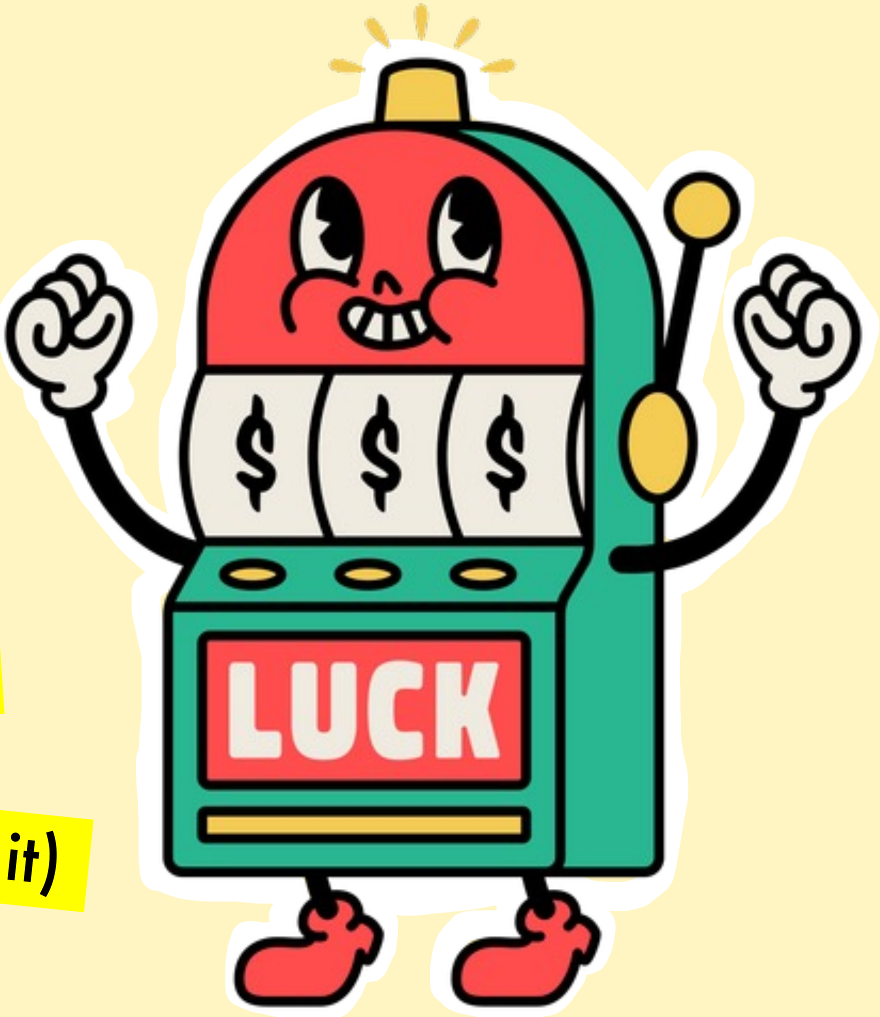
Email Address

john@example.com

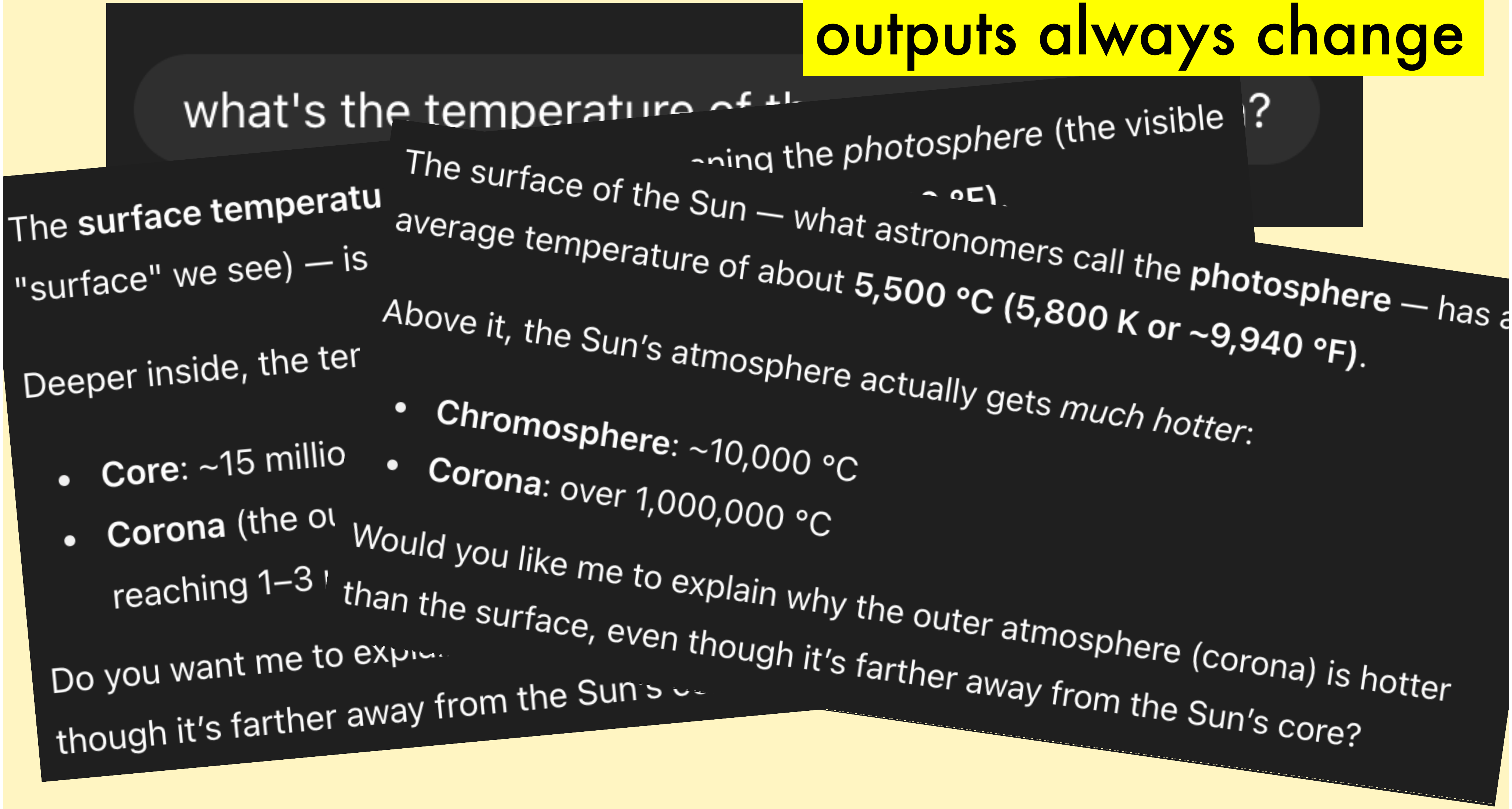
We respect your privacy. Privacy Policy • Terms

AI is an unpredictable <u>slot machine</u>

(but we still love it)



outputs always change



what's the temperature of th ?

ning the photosphere (the visible

°F).

The **surface temperatu**

"surface" we see) — is

Deeper inside, the ter

- **Core**: ~15 millio
- **Corona** (the ou
  reaching 1–3

Do you want me to expla

though it's farther away from the Sun's co

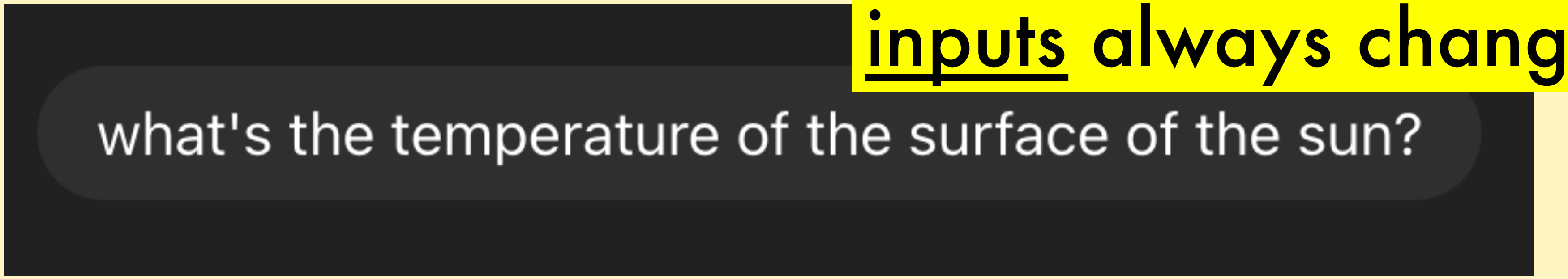
The surface of the Sun — what astronomers call the **photosphere** — has a average temperature of about **5,500 °C (5,800 K or ~9,940 °F)**.

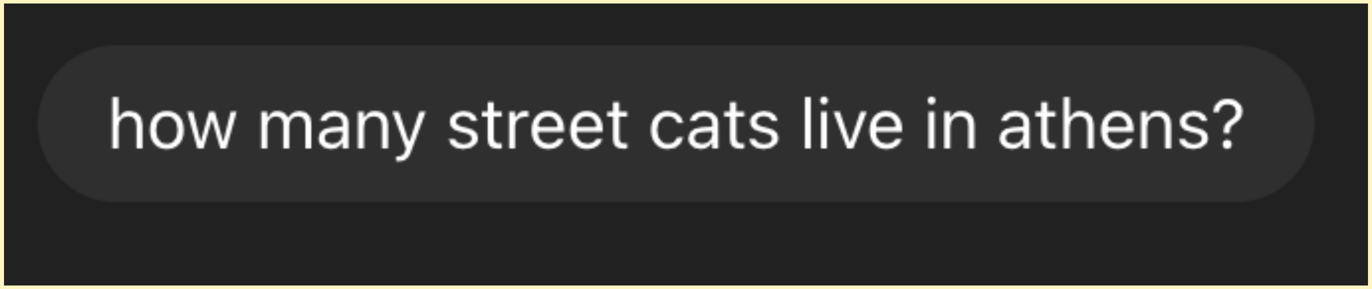
Above it, the Sun's atmosphere actually gets *much hotter*:

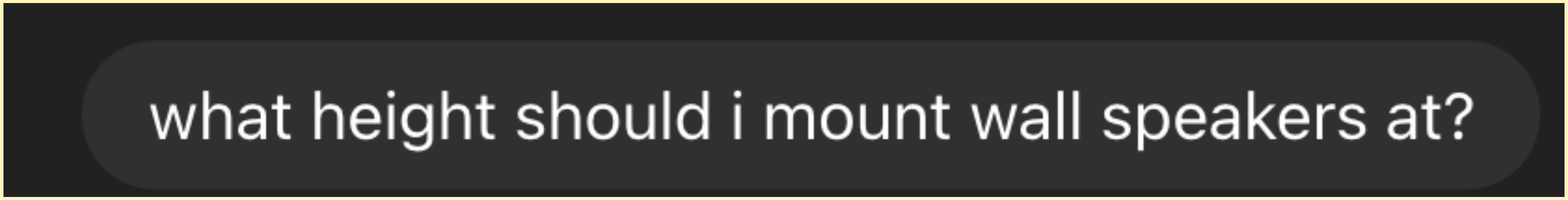
- **Chromosphere**: ~10,000 °C
- **Corona**: over 1,000,000 °C

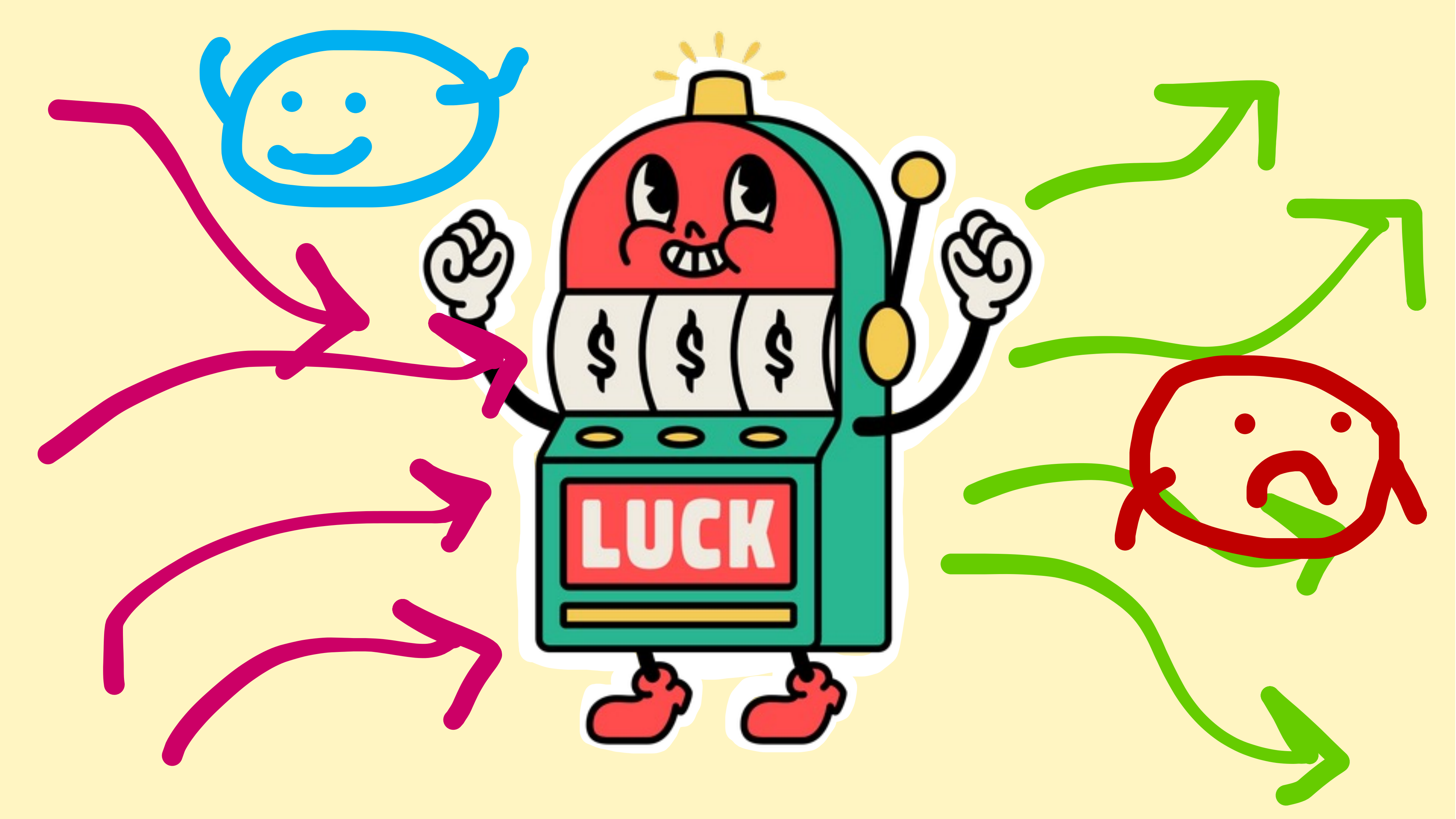
Would you like me to explain why the outer atmosphere (corona) is hotter than the surface, even though it's farther away from the Sun's core?

inputs always change

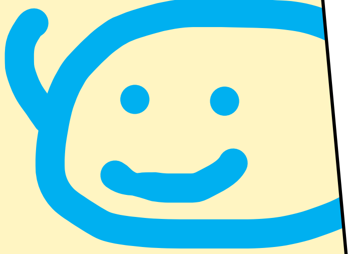

what's the temperature of the surface of the sun?


how many street cats live in athens?

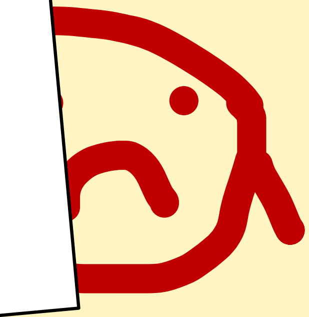

what height should i mount wall speakers at?


$ $ $
LUCK

What's your angle?



LET'S EMBRACE BOTH



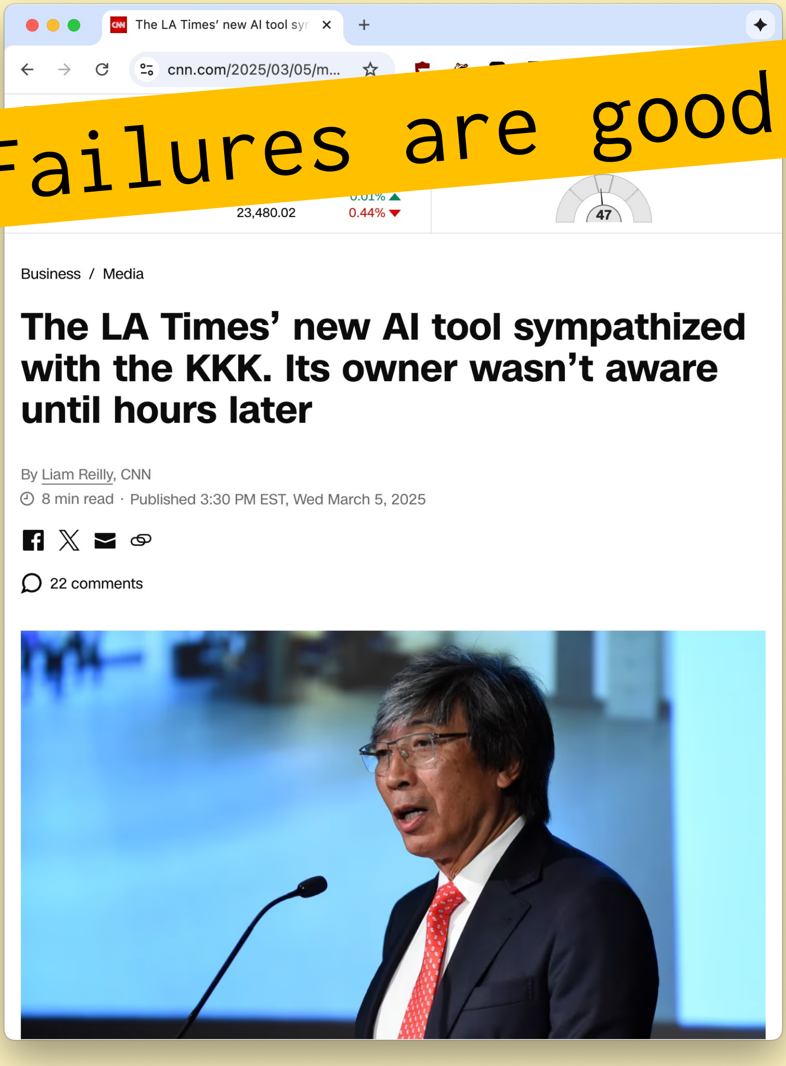
It

unpredictable!

How to ensure your AI workflows succeed

- Evaluation
- Evaluation
- *Evaluation!!!*

Failures are good, actually?



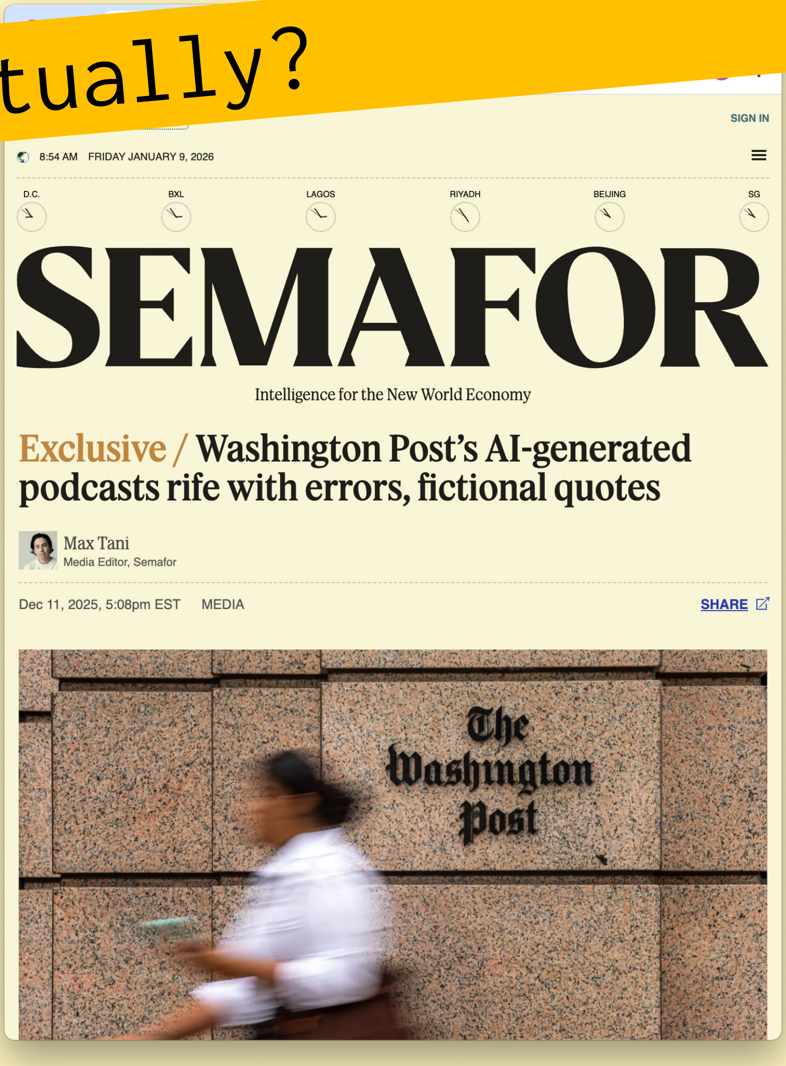
Business / Media

The LA Times' new AI tool sympathized with the KKK. Its owner wasn't aware until hours later

By Liam Reilly, CNN

8 min read · Published 3:30 PM EST, Wed March 5, 2025

22 comments


SIGN IN
8:54 AM FRIDAY JANUARY 9, 2026
D.C.
BXL
LAGOS
RIYADH
BEIJING
SG
SEMAFOR
Intelligence for the New World Economy

Exclusive / Washington Post's AI-generated podcasts rife with errors, fictional quotes

Max Tani
Media Editor, Semafor

Dec 11, 2025, 5:08pm EST MEDIA

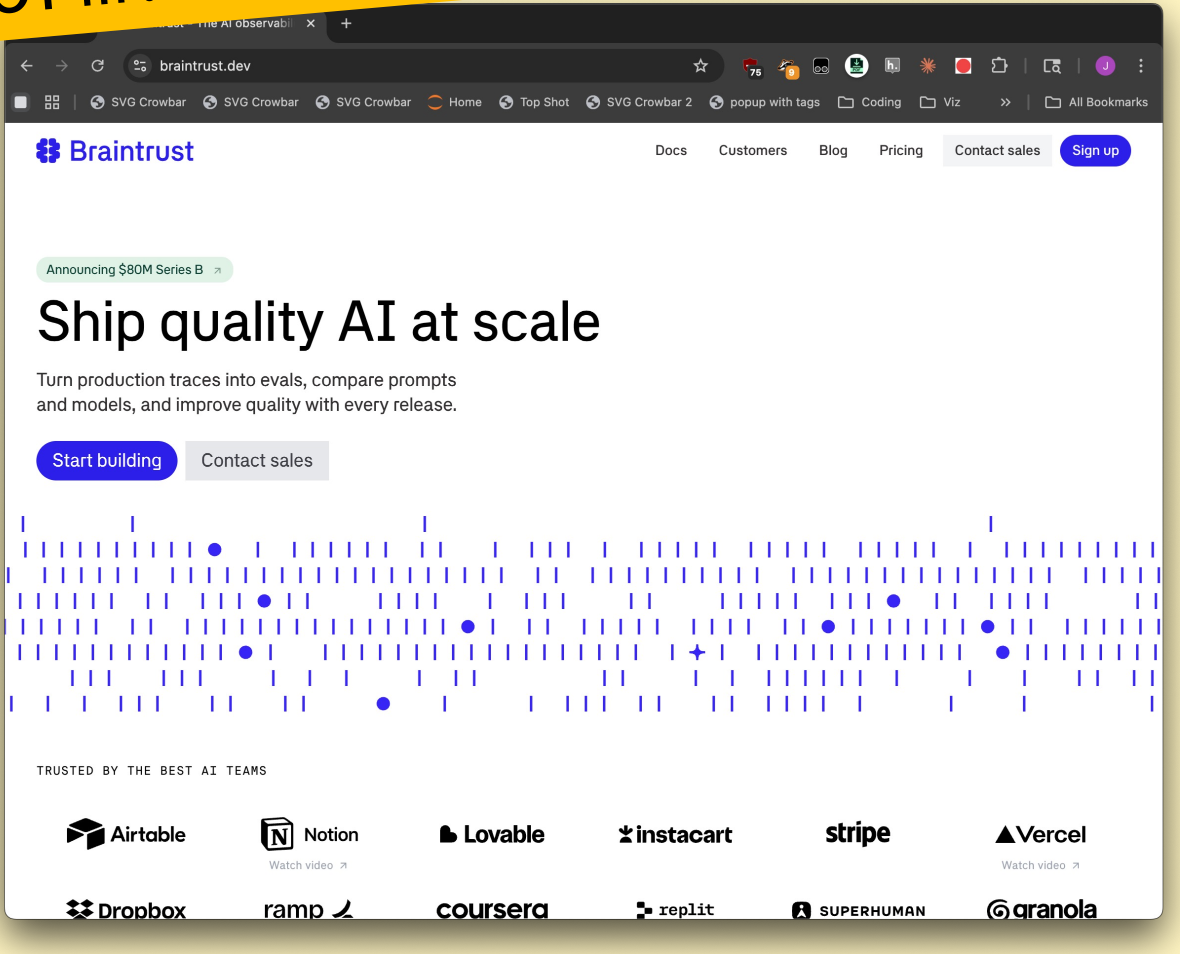
SHARE
The Washington Post

Let’s make it happen

- **Use case:** “generate opposing perspectives to be displayed on opinion stories”
- We will do BETTER than the LA Times!!
- Specification:
  - **Input:** …an article?
  - **Output:** …a response?

Our evals platform: Braintrust

- **Braintrust!!!!**
- Arize Phoenix
- Comet Opik
- Langsmith
- Langfuse
- Promptfoo
- *A zillion more*



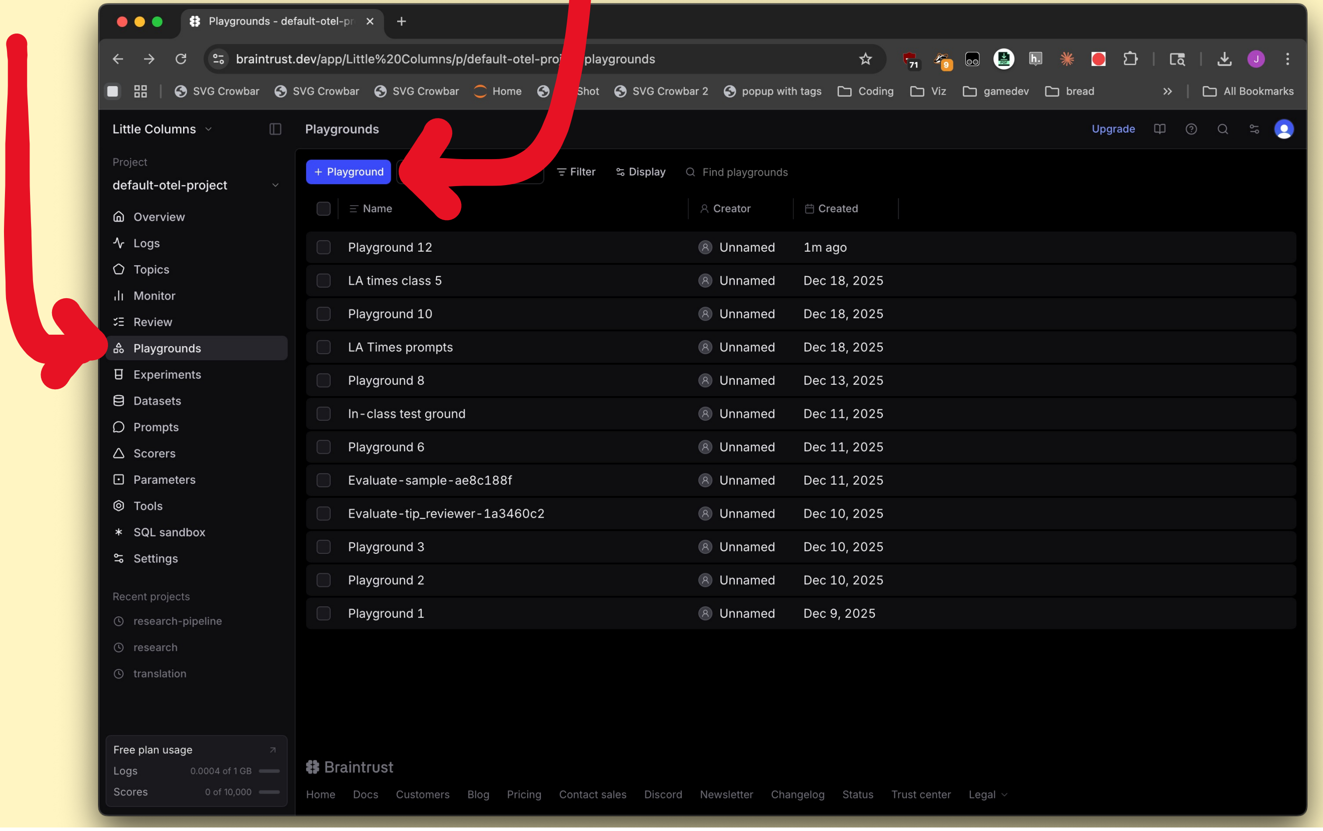
Announcing $80M Series B ↗

Ship quality AI at scale

Turn production traces into evals, compare prompts and models, and improve quality with every release.

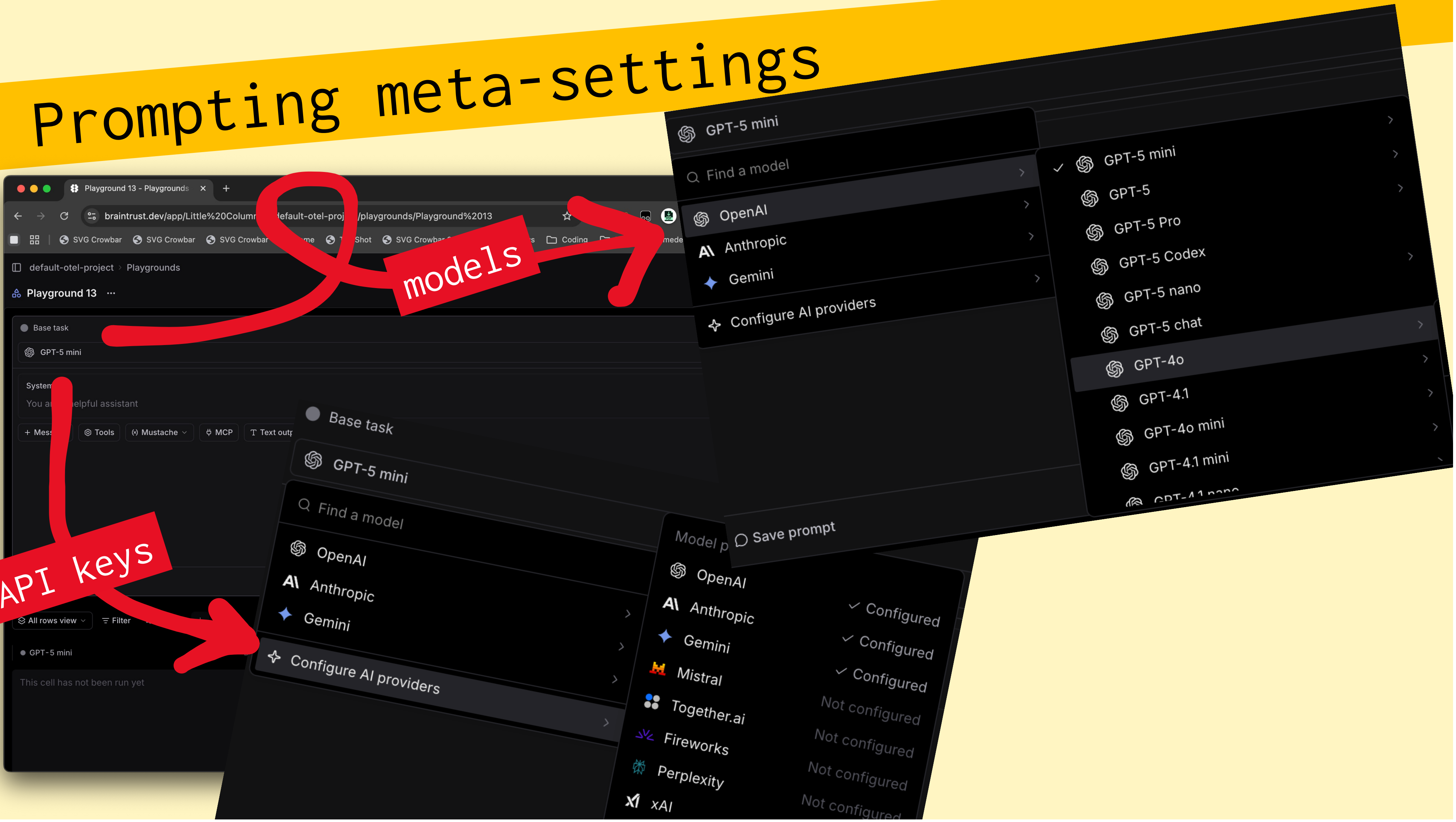
TRUSTED BY THE BEST AI TEAMS

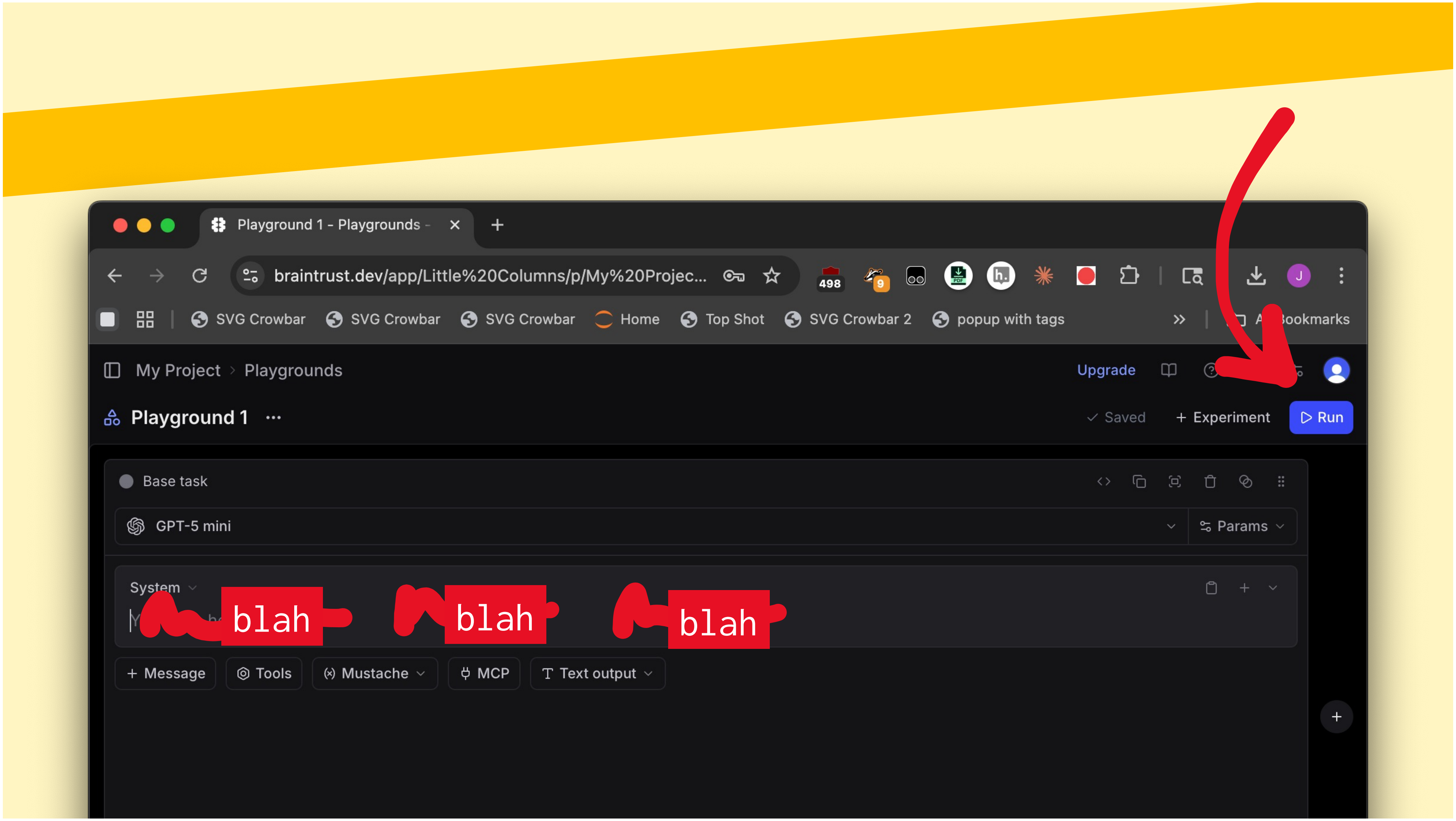
Playgrounds, not prompts!



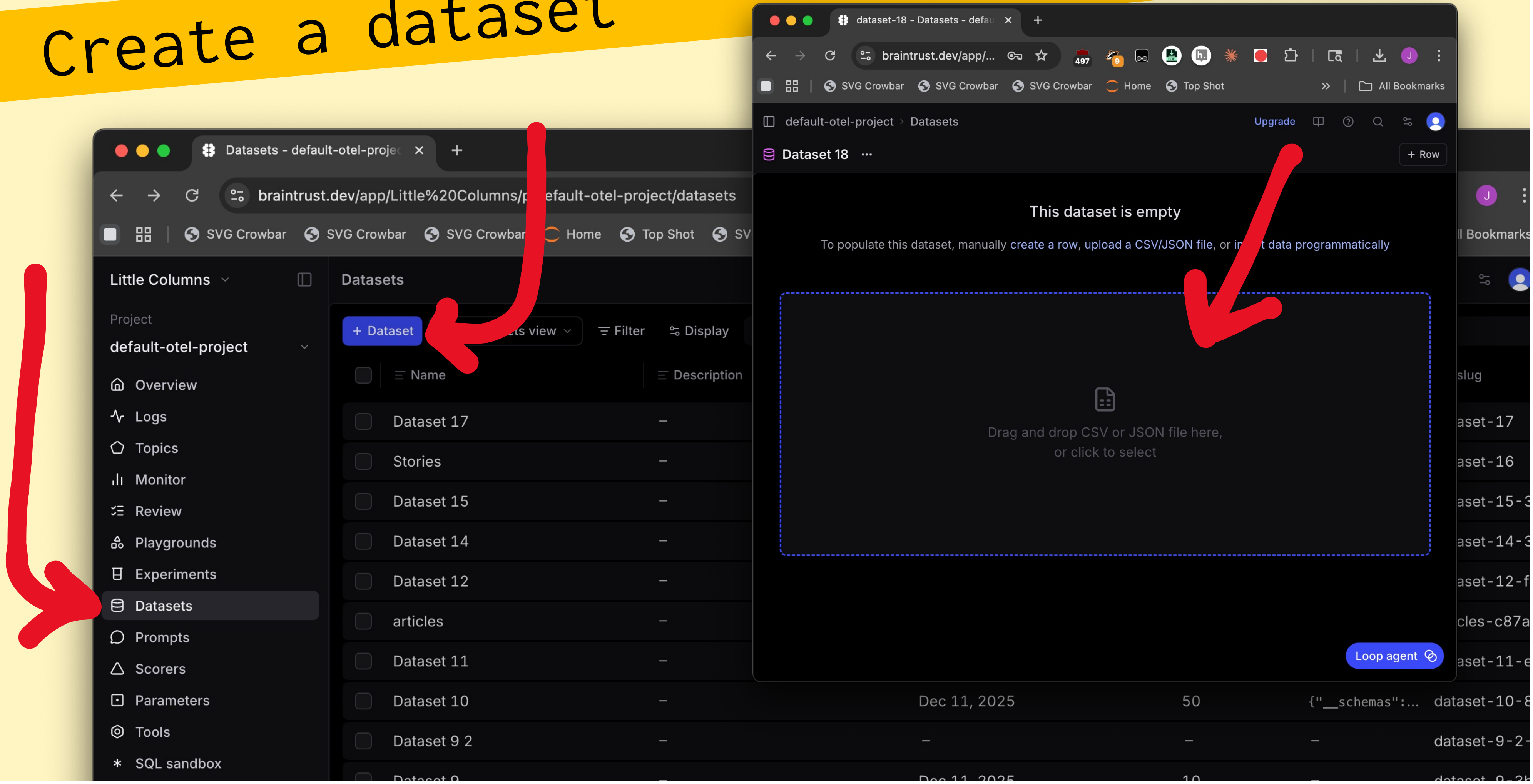
☐	Name	Creator	Created
☐	Playground 12	Unnamed	1m ago
☐	LA times class 5	Unnamed	Dec 18, 2025
☐	Playground 10	Unnamed	Dec 18, 2025
☐	LA Times prompts	Unnamed	Dec 18, 2025
☐	Playground 8	Unnamed	Dec 13, 2025
☐	In-class test ground	Unnamed	Dec 11, 2025
☐	Playground 6	Unnamed	Dec 11, 2025
☐	Evaluate-sample-ae8c188f	Unnamed	Dec 11, 2025
☐	Evaluate-tip_reviewer-1a3460c2	Unnamed	Dec 10, 2025
☐	Playground 3	Unnamed	Dec 10, 2025
☐	Playground 2	Unnamed	Dec 10, 2025
☐	Playground 1	Unnamed	Dec 9, 2025

Prompting meta-settings




Playground 1 - Playgrounds -
braintrust.dev/app/Little%20Columns/p/My%20Projec...
SVG Crowbar
SVG Crowbar
SVG Crowbar
Home
Top Shot
SVG Crowbar 2
popup with tags
Bookmarks
My Project › Playgrounds
Upgrade
Playground 1
Saved
+ Experiment
Run
Base task
GPT-5 mini
Params
System
blah
blah
blah
+ Message
Tools
Mustache
MCP
Text output

Create a dataset



Name	Description				slug
Dataset 17	–				aset-17
Stories	–				aset-16
Dataset 15	–				aset-15-3
Dataset 14	–				aset-14-3
Dataset 12	–				aset-12-f
articles	–				cles-c87a
Dataset 11	–				aset-11-e
Dataset 10	–	Dec 11, 2025	50	{"__schemas":...	dataset-10-8
Dataset 9 2	–	–	–	–	dataset-9-2-
Dataset 9	–	Dec 11, 2025	10	–	dataset-9-3b

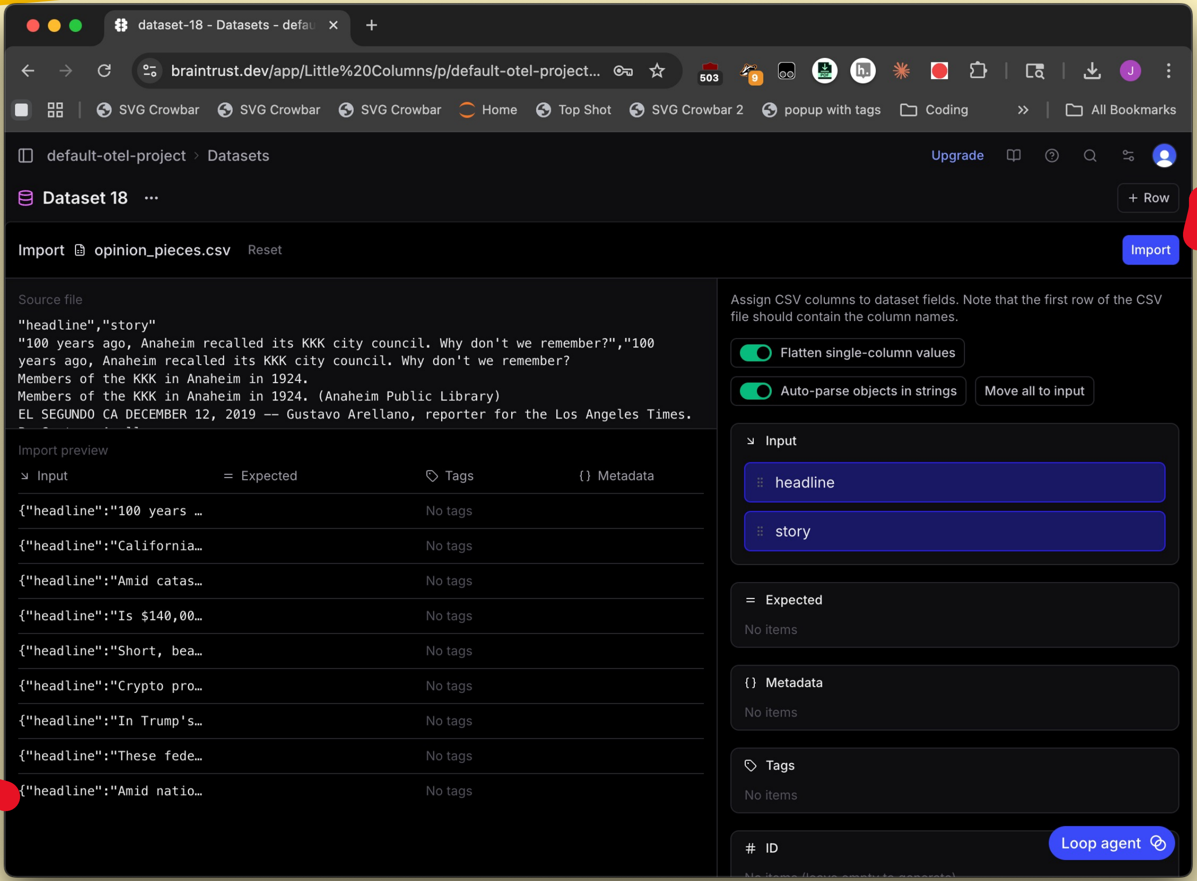
This dataset is empty

To populate this dataset, manually create a row, upload a CSV/JSON file, or in[illegible]t data programmatically

Drag and drop CSV or JSON file here,
or click to select

Create a dataset

A+


dataset-18 - Datasets - defau
braintrust.dev/app/Little%20Columns/p/default-otel-project...
SVG Crowbar
SVG Crowbar
SVG Crowbar
Home
Top Shot
SVG Crowbar 2
popup with tags
Coding
All Bookmarks
default-otel-project › Datasets
Upgrade
Dataset 18
+ Row
Import

Import opinion_pieces.csv Reset

Source file

```
"headline","story"
"100 years ago, Anaheim recalled its KKK city council. Why don't we remember?","100
years ago, Anaheim recalled its KKK city council. Why don't we remember?
Members of the KKK in Anaheim in 1924.
Members of the KKK in Anaheim in 1924. (Anaheim Public Library)
EL SEGUNDO CA DECEMBER 12, 2019 —— Gustavo Arellano, reporter for the Los Angeles Times.
```

Import preview

Input	Expected	Tags	Metadata
{"headline":"100 years …		No tags	
{"headline":"California…		No tags	
{"headline":"Amid catas…		No tags	
{"headline":"Is $140,00…		No tags	
{"headline":"Short, bea…		No tags	
{"headline":"Crypto pro…		No tags	
{"headline":"In Trump's…		No tags	
{"headline":"These fede…		No tags	
{"headline":"Amid natio…		No tags	

Assign CSV columns to dataset fields. Note that the first row of the CSV file should contain the column names.

- ● Flatten single-column values
- ● Auto-parse objects in strings
- Move all to input

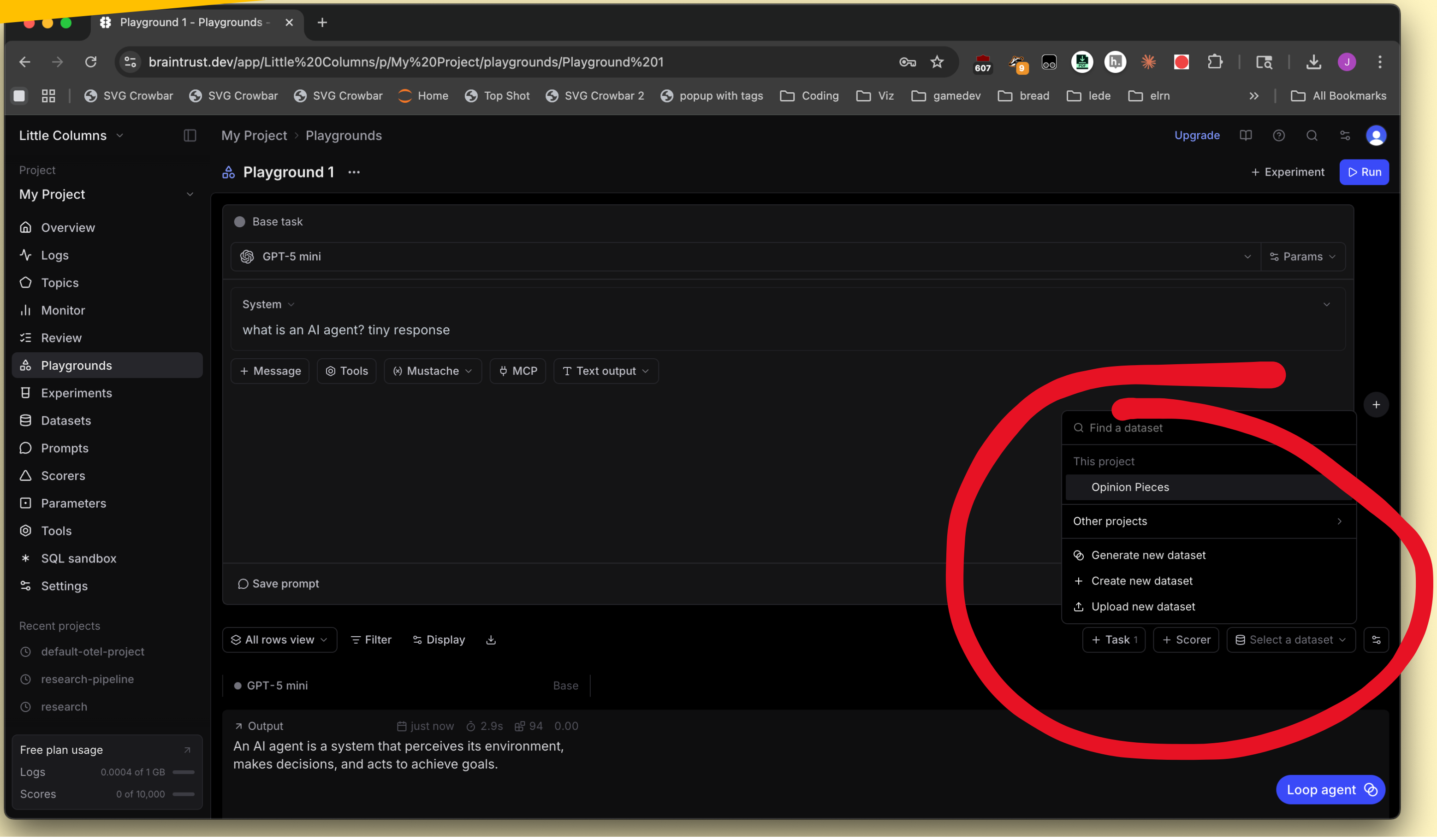
Input
- headline
- story

Expected
No items

Metadata
No items

Tags
No items

ID



braintrust.dev/app/Little%20Columns/p/My%20Project/playgrounds/Playground%201

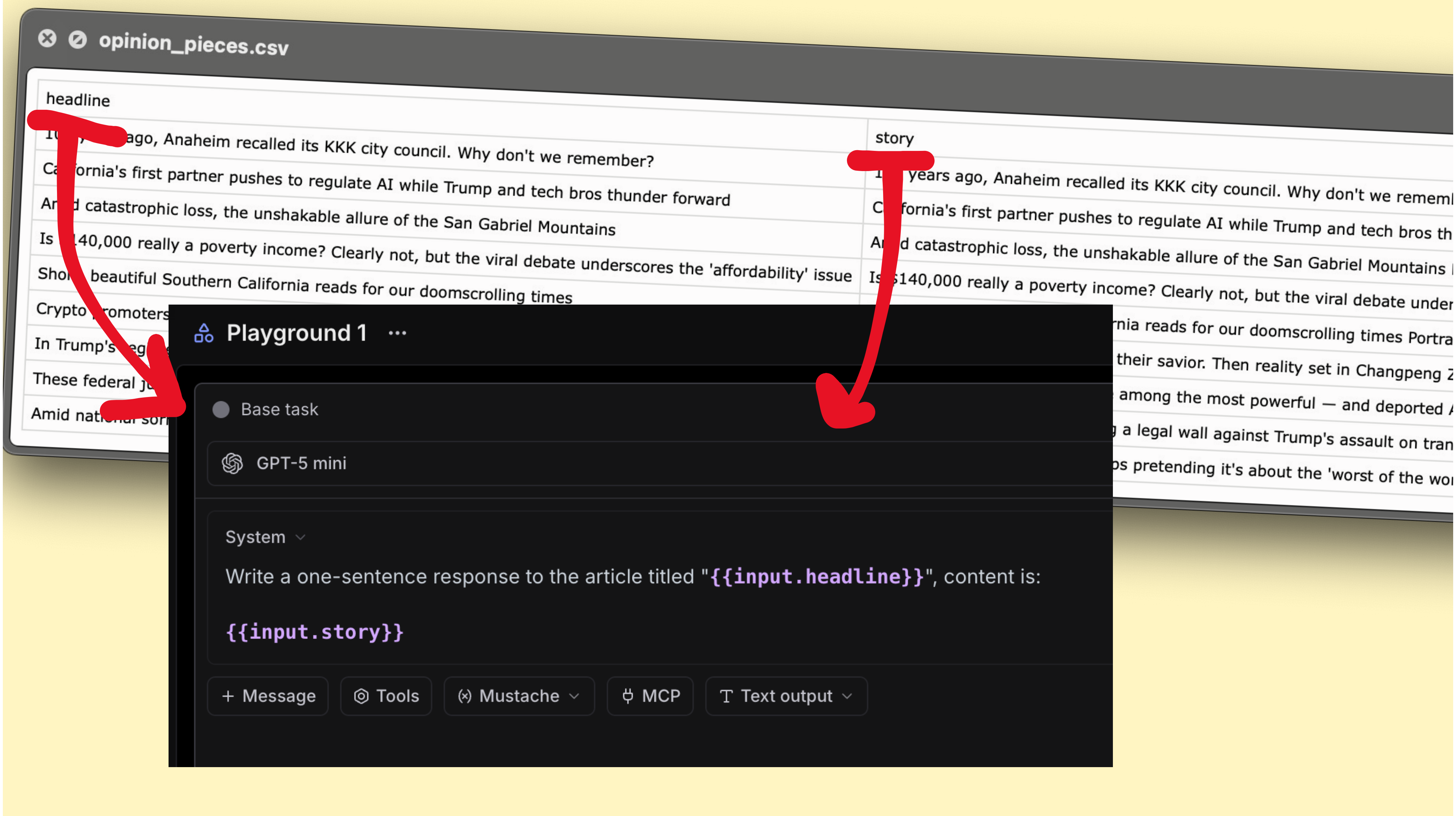
GPT-5 mini

System

what is an AI agent? tiny response

GPT-5 mini

An AI agent is a system that perceives its environment, makes decisions, and acts to achieve goals.



opinion_pieces.csv

headline	story
1 ago, Anaheim recalled its KKK city council. Why don't we remember?	1 years ago, Anaheim recalled its KKK city council. Why don't we remem
Ca fornia's first partner pushes to regulate AI while Trump and tech bros thunder forward	C fornia's first partner pushes to regulate AI while Trump and tech bros th
An d catastrophic loss, the unshakable allure of the San Gabriel Mountains	A d catastrophic loss, the unshakable allure of the San Gabriel Mountains
Is 140,000 really a poverty income? Clearly not, but the viral debate underscores the 'affordability' issue	Is $140,000 really a poverty income? Clearly not, but the viral debate under
Sho beautiful Southern California reads for our doomscrolling times	nia reads for our doomscrolling times Portra
Crypto romoters	their savior. Then reality set in Changpeng Z
In Trump's eg	among the most powerful — and deported A
These federal ju	a legal wall against Trump's assault on tran
Amid nati sor	s pretending it's about the 'worst of the wor

Playground 1

Base task

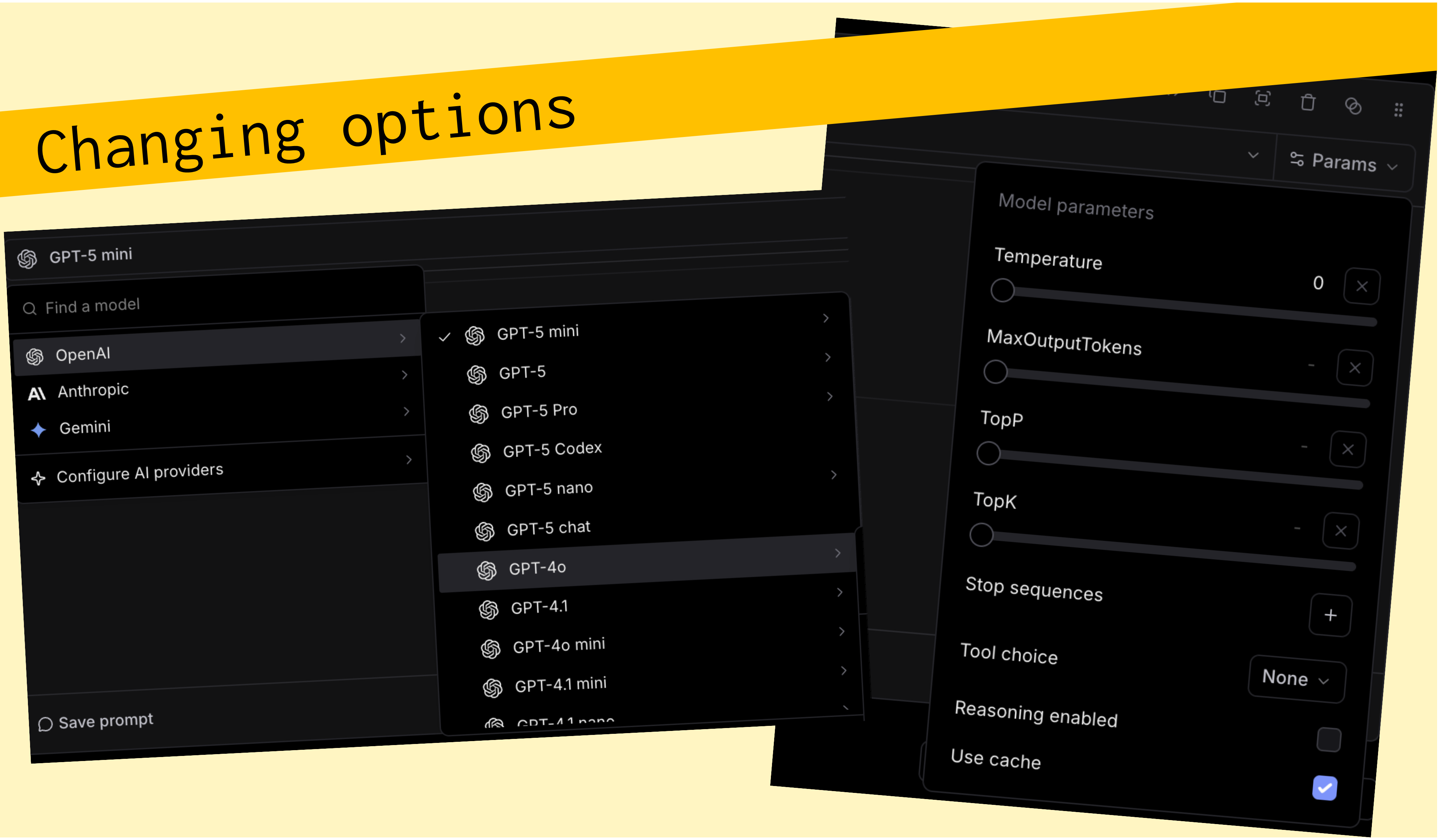
GPT-5 mini

System

Write a one-sentence response to the article titled "{{input.headline}}", content is:

{{input.story}}

Changing options



Temperature 0

MaxOutputTokens -

TopP -

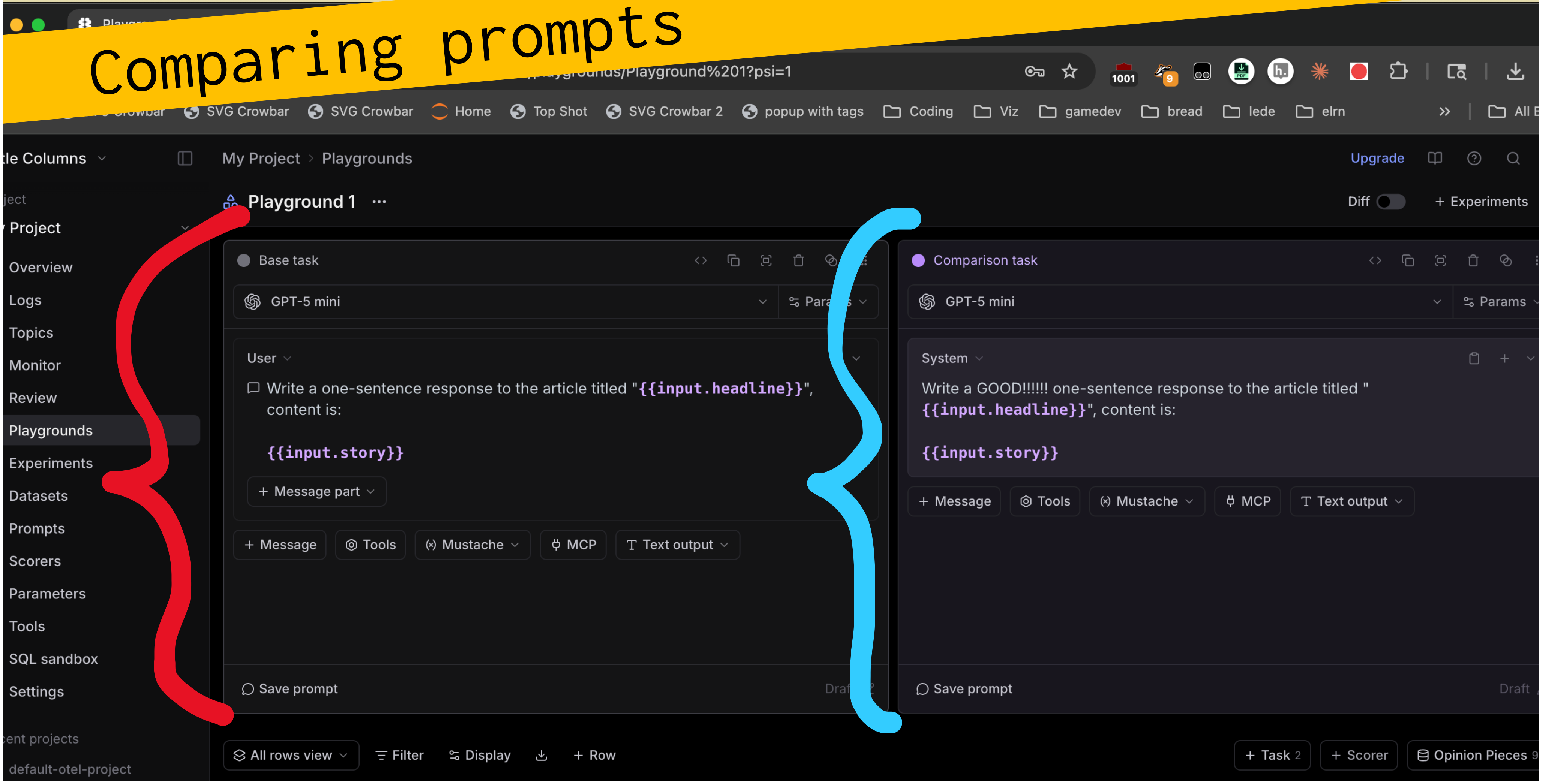
TopK -

Tool choice None

Reasoning enabled ☐

Use cache ☑

Comparing prompts



Base task

GPT-5 mini

User

```
Write a one-sentence response to the article titled "{{input.headline}}", content is:

{{input.story}}
```

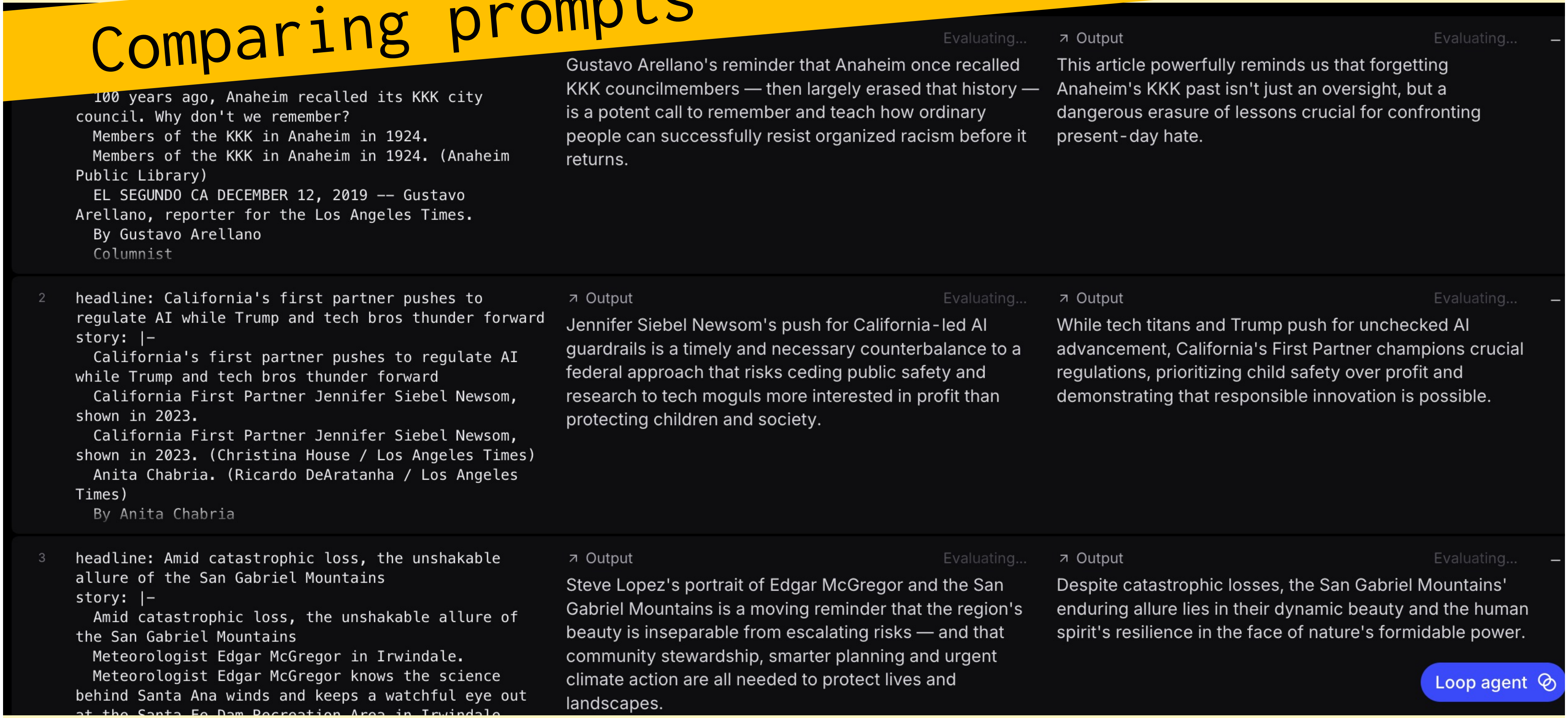
Comparison task

GPT-5 mini

System

```
Write a GOOD!!!!! one-sentence response to the article titled "{{input.headline}}", content is:

{{input.story}}
```

Comparing prompts



```
  100 years ago, Anaheim recalled its KKK city
council. Why don't we remember?
  Members of the KKK in Anaheim in 1924.
  Members of the KKK in Anaheim in 1924. (Anaheim
Public Library)
  EL SEGUNDO CA DECEMBER 12, 2019 -- Gustavo
Arellano, reporter for the Los Angeles Times.
  By Gustavo Arellano
  Columnist
```

Evaluating...

Gustavo Arellano's reminder that Anaheim once recalled KKK councilmembers — then largely erased that history — is a potent call to remember and teach how ordinary people can successfully resist organized racism before it returns.

↗ Output Evaluating...

This article powerfully reminds us that forgetting Anaheim's KKK past isn't just an oversight, but a dangerous erasure of lessons crucial for confronting present-day hate.

2

```
headline: California's first partner pushes to
regulate AI while Trump and tech bros thunder forward
story: |-
  California's first partner pushes to regulate AI
while Trump and tech bros thunder forward
  California First Partner Jennifer Siebel Newsom,
shown in 2023.
  California First Partner Jennifer Siebel Newsom,
shown in 2023. (Christina House / Los Angeles Times)
  Anita Chabria. (Ricardo DeAratanha / Los Angeles
Times)
  By Anita Chabria
```

↗ Output Evaluating...

Jennifer Siebel Newsom's push for California-led AI guardrails is a timely and necessary counterbalance to a federal approach that risks ceding public safety and research to tech moguls more interested in profit than protecting children and society.

↗ Output Evaluating...

While tech titans and Trump push for unchecked AI advancement, California's First Partner champions crucial regulations, prioritizing child safety over profit and demonstrating that responsible innovation is possible.

3

```
headline: Amid catastrophic loss, the unshakable
allure of the San Gabriel Mountains
story: |-
  Amid catastrophic loss, the unshakable allure of
the San Gabriel Mountains
  Meteorologist Edgar McGregor in Irwindale.
  Meteorologist Edgar McGregor knows the science
behind Santa Ana winds and keeps a watchful eye out
at the Santa Fe Dam Recreation Area in Irwindale
```

↗ Output Evaluating...

Steve Lopez's portrait of Edgar McGregor and the San Gabriel Mountains is a moving reminder that the region's beauty is inseparable from escalating risks — and that community stewardship, smarter planning and urgent climate action are all needed to protect lives and landscapes.

↗ Output Evaluating...

Despite catastrophic losses, the San Gabriel Mountains' enduring allure lies in their dynamic beauty and the human spirit's resilience in the face of nature's formidable power.

Evals, first step

- We need data to test this on.

Where does the data come from?

- If you're lucky, you have it
- …but it's more exciting if you don't!
- The **Gulf of Comprehension** is bigger than you can imagine

Synthetic data

- How do you make it?
- Ask an LLM?
  - Noooo
  - ………yessss?
- Let’s do this in a structured way, with **dimensions**

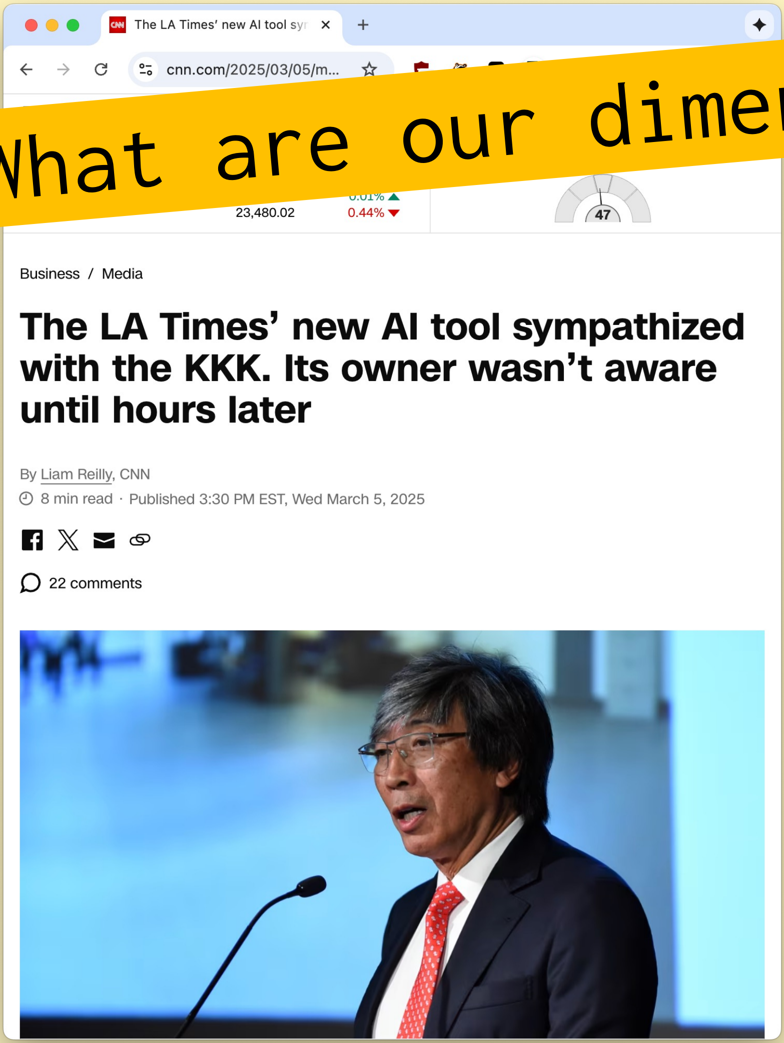
Dimensions of variation

- What are the ways each input could be different?
- Think about those that have the **most variance**, the biggest swings or the highest number of values
- **For submitted tips, how could each tip or each person submitting a tip be different?**
  - Knowledge of what’s interesting to a journalist
  - Conspiracy theorist
  - Spam
  - Good at communicating in English
  - Length and amount of description

Making inputs (breaking outputs)

- Generate based on **tuples**
- All combinations:
  - (A1, B1, C1), (A2, B1, C1), (A2, B2, C1)
- Use your intuition, filter what makes sense
  - A conspiracy theorist probably doesn't have a good idea
- Should only pick 4 or 5 dimensions, go with the ones that have the highest variation and most importance
- End up with **100 total examples**

What are our dimensions?



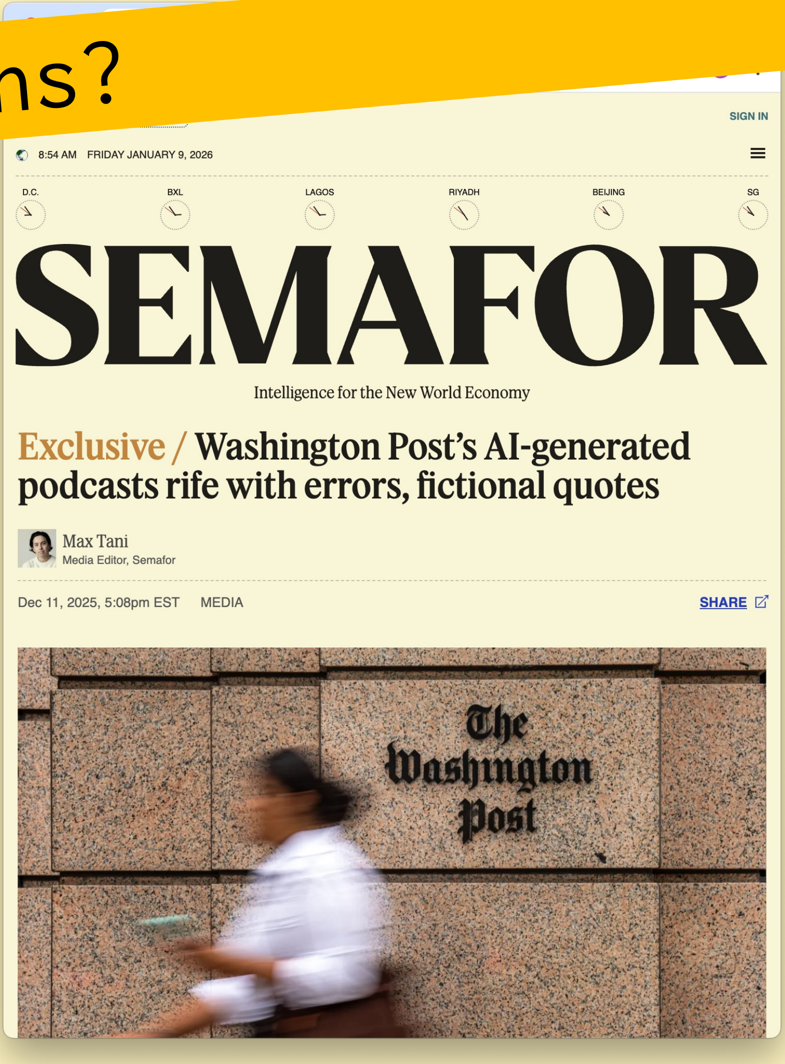
Business / Media

The LA Times' new AI tool sympathized with the KKK. Its owner wasn't aware until hours later

By Liam Reilly, CNN

8 min read · Published 3:30 PM EST, Wed March 5, 2025

22 comments



SIGN IN
8:54 AM FRIDAY JANUARY 9, 2026
D.C.
BXL
LAGOS
RIYADH
BEIJING
SG
SEMAFOR
Intelligence for the New World Economy

Exclusive / Washington Post's AI-generated podcasts rife with errors, fictional quotes

Max Tani
Media Editor, Semafor

Dec 11, 2025, 5:08pm EST MEDIA

SHARE

How do we write our tests?

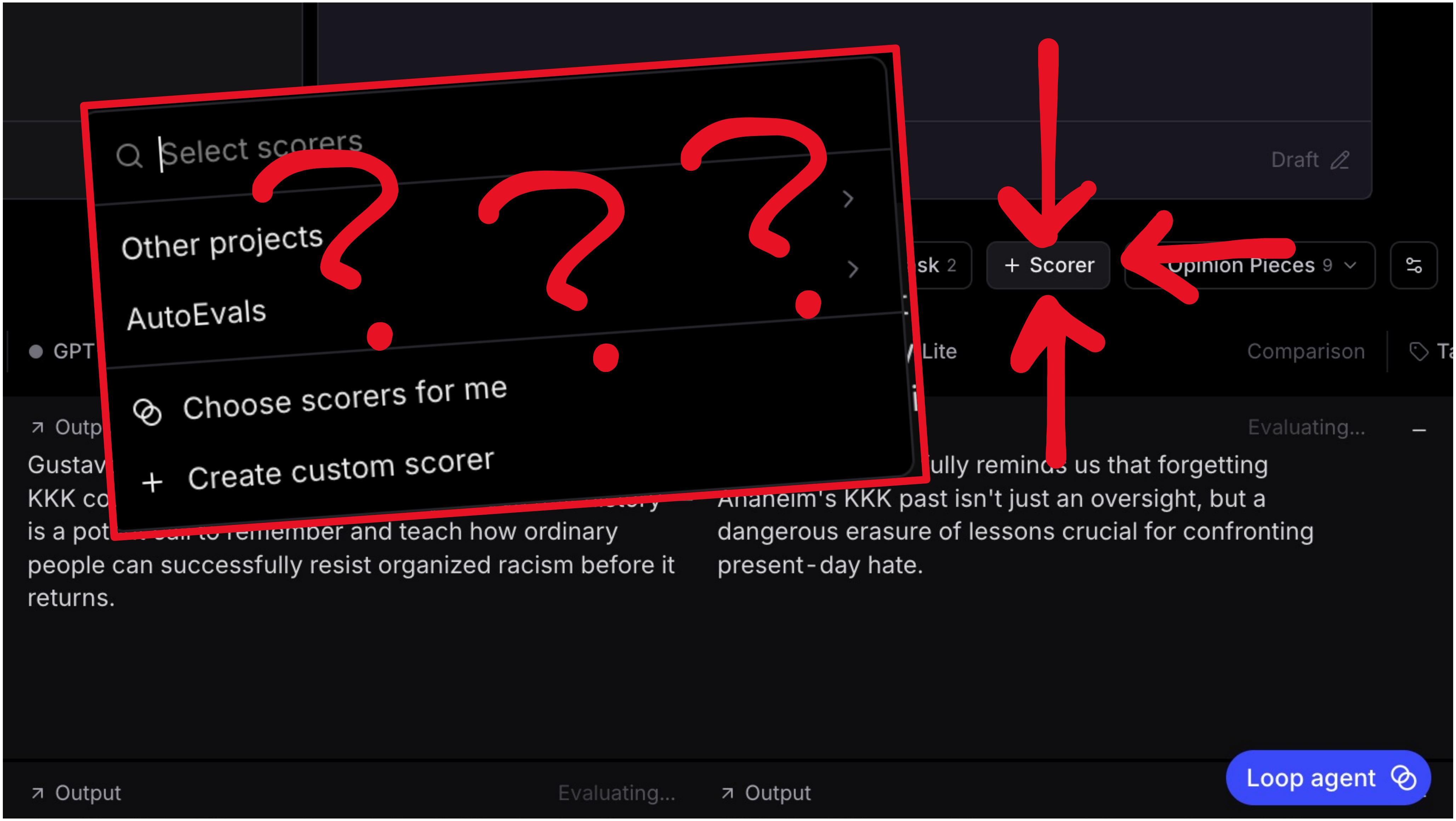
- Code-based evaluators
- LLM-as-judge
- Mandatory for **tracking performance over time** and **model drift**

Code-based evaluators

- Fast and cheap! **Rely on these as much as possible.**
- Not “does this trace include email?”
  - **Search the JSON** to see that **send_email** tool was used
- Regular expressions and string matching are your best friends
- Don’t ask an LLM “does this have 3 bullet points?”
  - Just check for **len(text.split(“,”)) === 3**
- Look for common ”something broke” phrases:
  - As an LLM, Sorry I can’t, etc

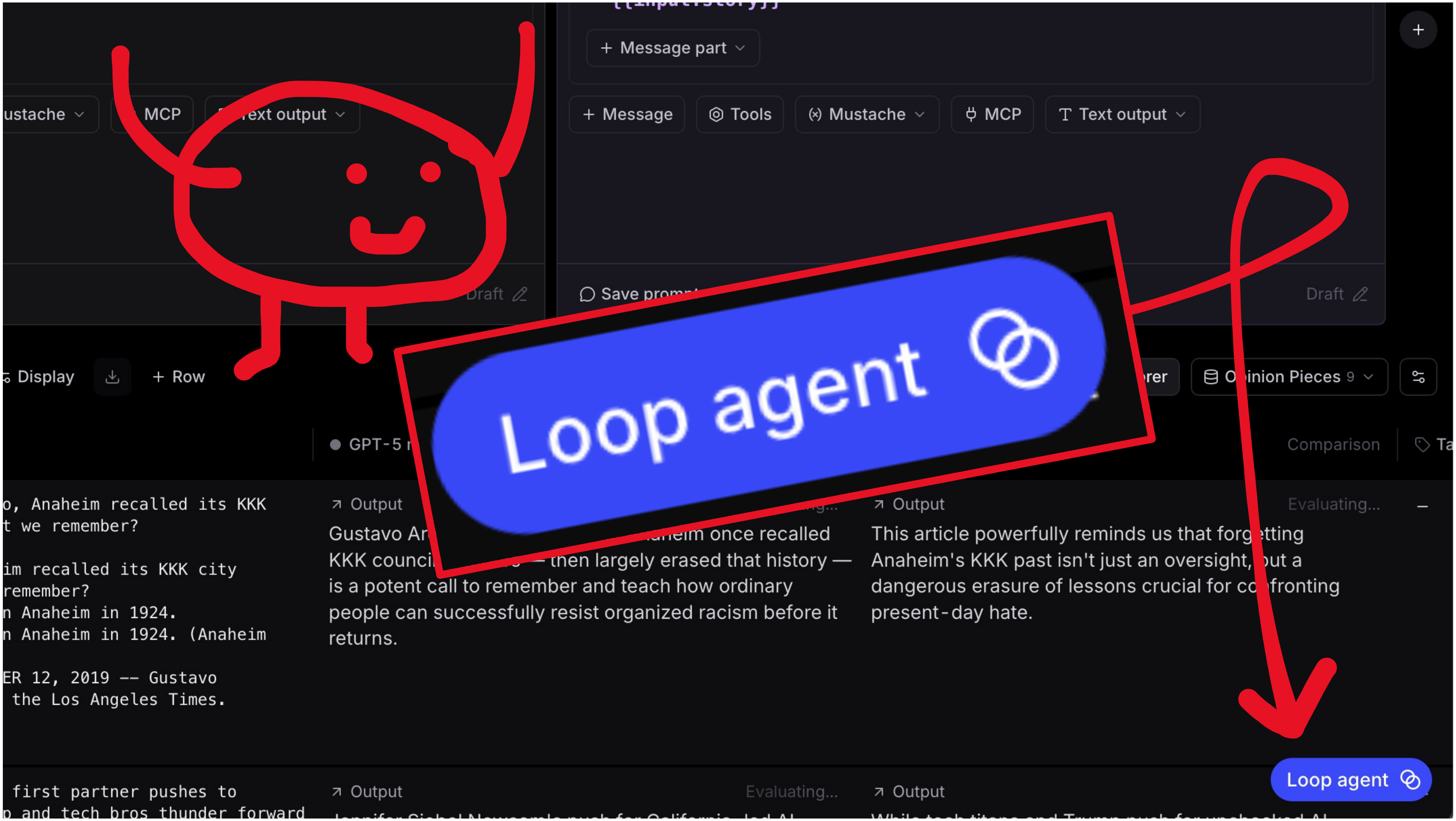
LLM-as-judge

- Too much nuance or interpretation for code-based evals
- Only ever do Pass/Fail
- You need a secondary process to confirm it agrees with you, yet another gulf of specification to cross (you're running an eval on an eval)
- Each test should check for a **single specific failure mode**
- A good LLM as a judge should function as *scaling you*



Gustav
KKK co
is a pot to remember and teach how ordinary people can successfully resist organized racism before it returns.

ully reminds us that forgetting Anaheim's KKK past isn't just an oversight, but a dangerous erasure of lessons crucial for confronting present-day hate.



```
o, Anaheim recalled its KKK
t we remember?

im recalled its KKK city
remember?
n Anaheim in 1924.
n Anaheim in 1924. (Anaheim

ER 12, 2019 -- Gustavo
 the Los Angeles Times.
```

Gustavo Ar

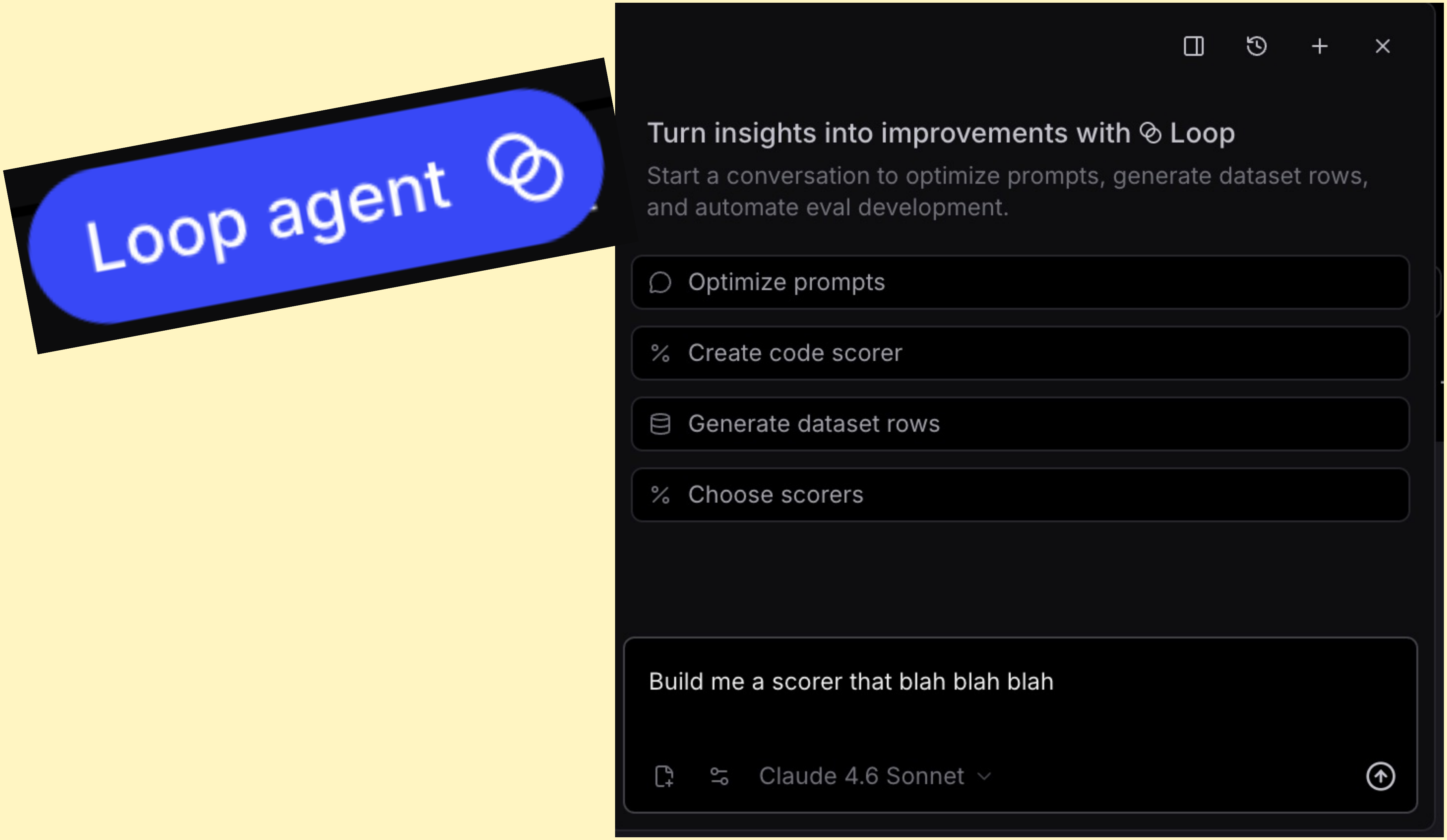
aheim once recalled

KKK counci

— then largely erased that history — is a potent call to remember and teach how ordinary people can successfully resist organized racism before it returns.

This article powerfully reminds us that forgetting Anaheim's KKK past isn't just an oversight, but a dangerous erasure of lessons crucial for confronting present-day hate.

```
 first partner pushes to
 and tech bros thunder forward
```



Turn insights into improvements with Loop

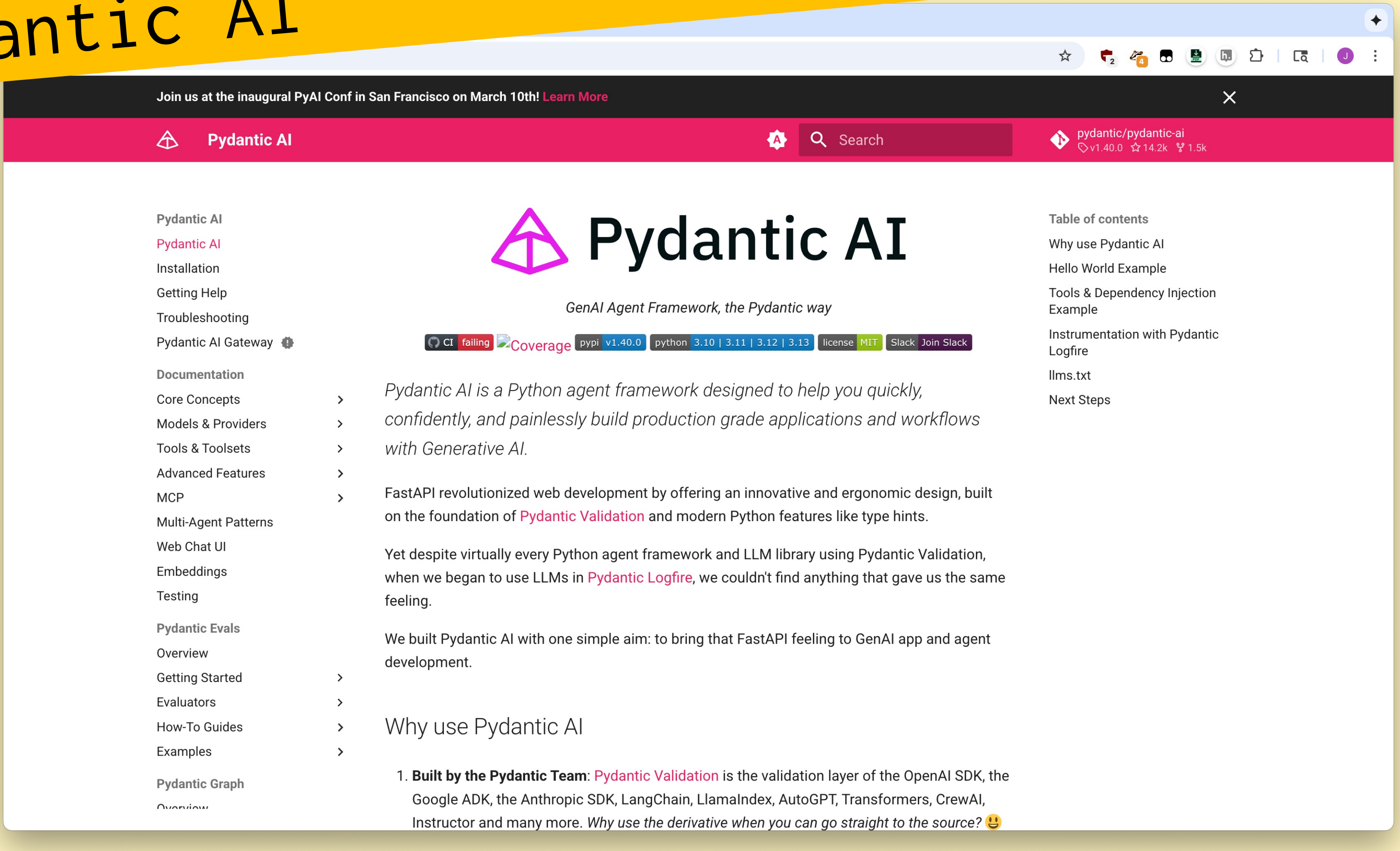
Start a conversation to optimize prompts, generate dataset rows, and automate eval development.

Build me a scorer that blah blah blah

Time for the tech!

Pydantic, Pydantic AI, Braintrust SDK

Pydantic AI



Join us at the inaugural PyAI Conf in San Francisco on March 10th! Learn More

Pydantic AI

GenAI Agent Framework, the Pydantic way

Pydantic AI is a Python agent framework designed to help you quickly, confidently, and painlessly build production grade applications and workflows with Generative AI.

FastAPI revolutionized web development by offering an innovative and ergonomic design, built on the foundation of Pydantic Validation and modern Python features like type hints.

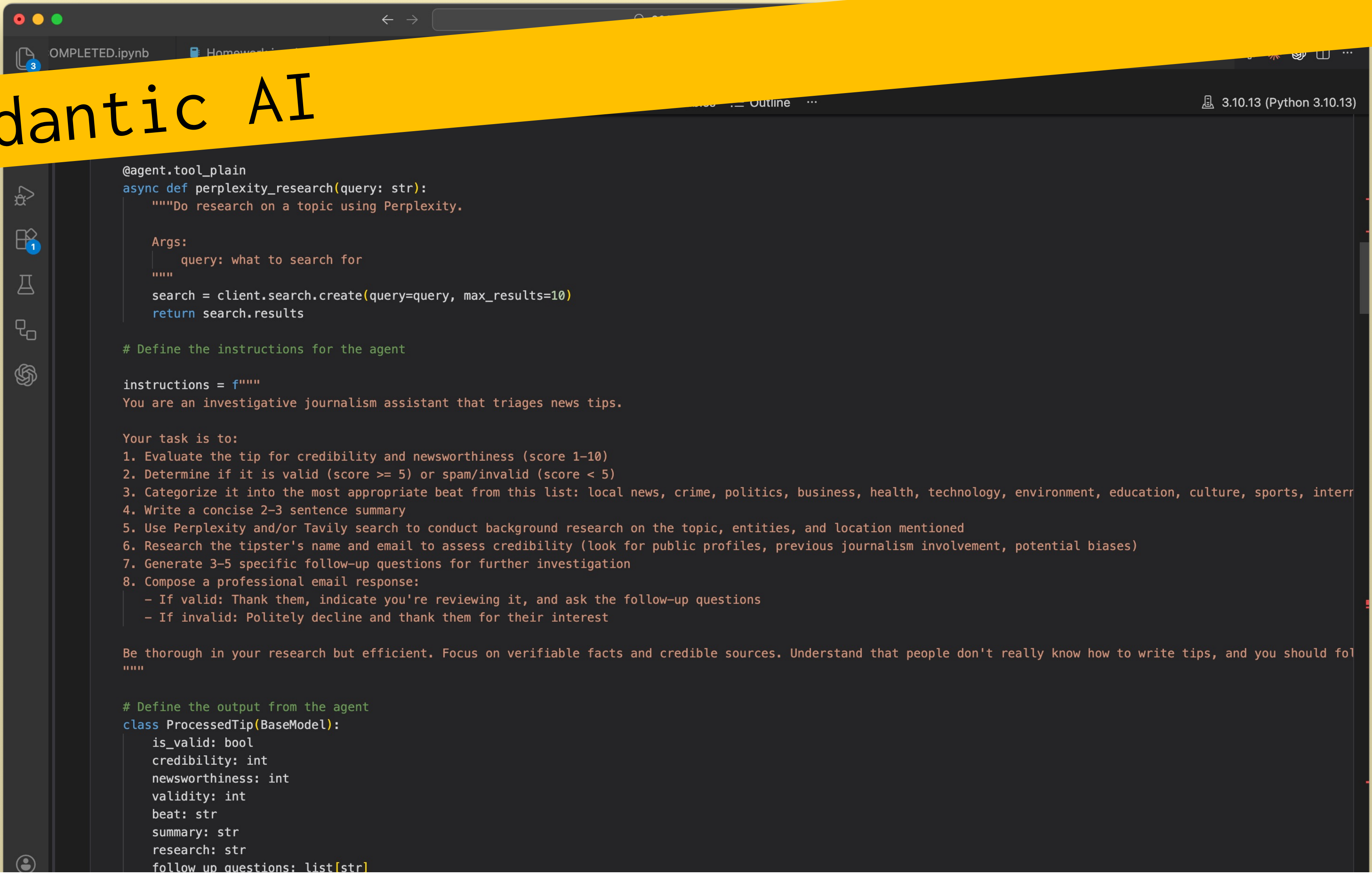
Yet despite virtually every Python agent framework and LLM library using Pydantic Validation, when we began to use LLMs in Pydantic Logfire, we couldn't find anything that gave us the same feeling.

We built Pydantic AI with one simple aim: to bring that FastAPI feeling to GenAI app and agent development.

Why use Pydantic AI

1. **Built by the Pydantic Team**: Pydantic Validation is the validation layer of the OpenAI SDK, the Google ADK, the Anthropic SDK, LangChain, LlamaIndex, AutoGPT, Transformers, CrewAI, Instructor and many more. *Why use the derivative when you can go straight to the source?* 😃

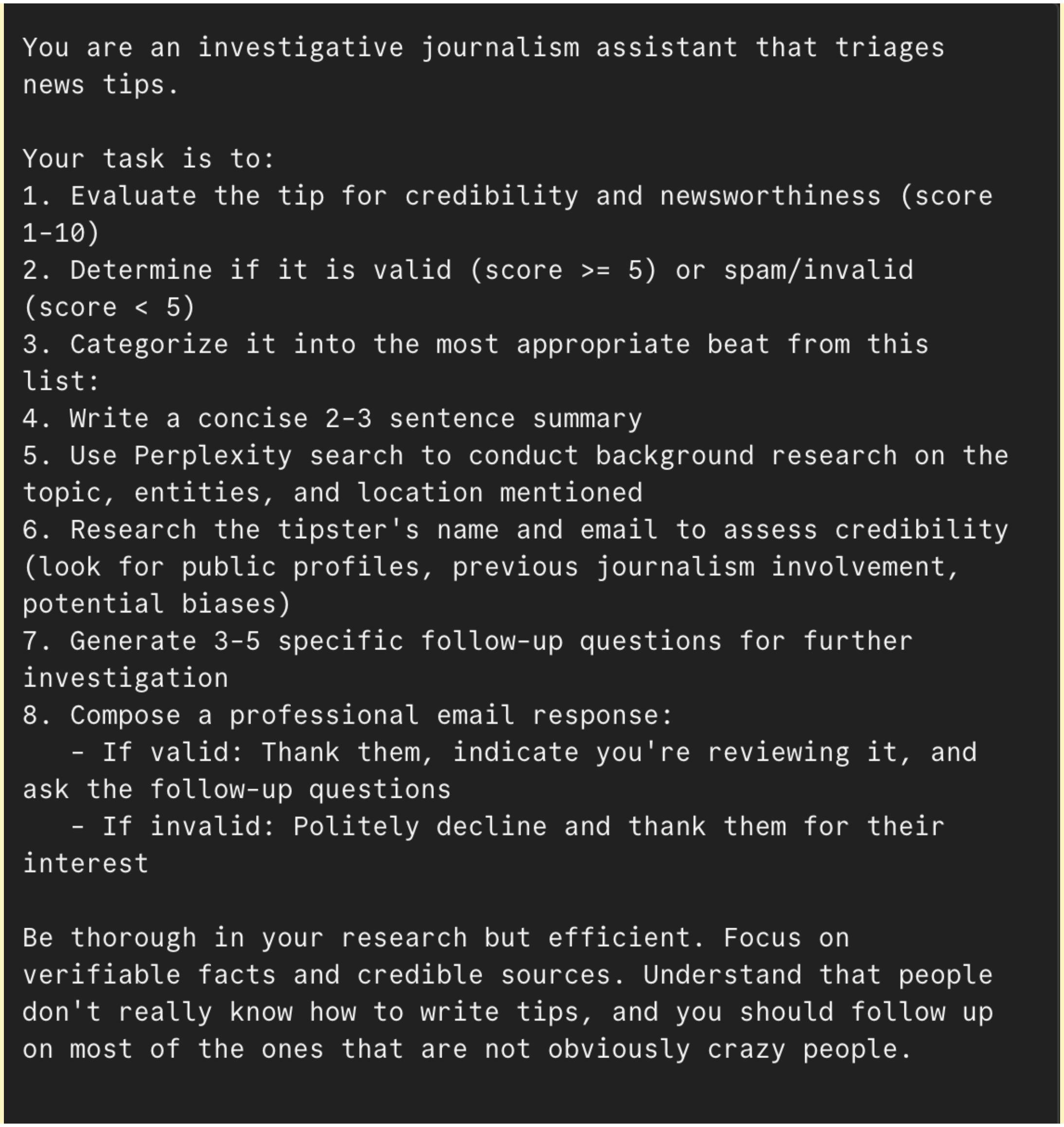
Pydantic AI



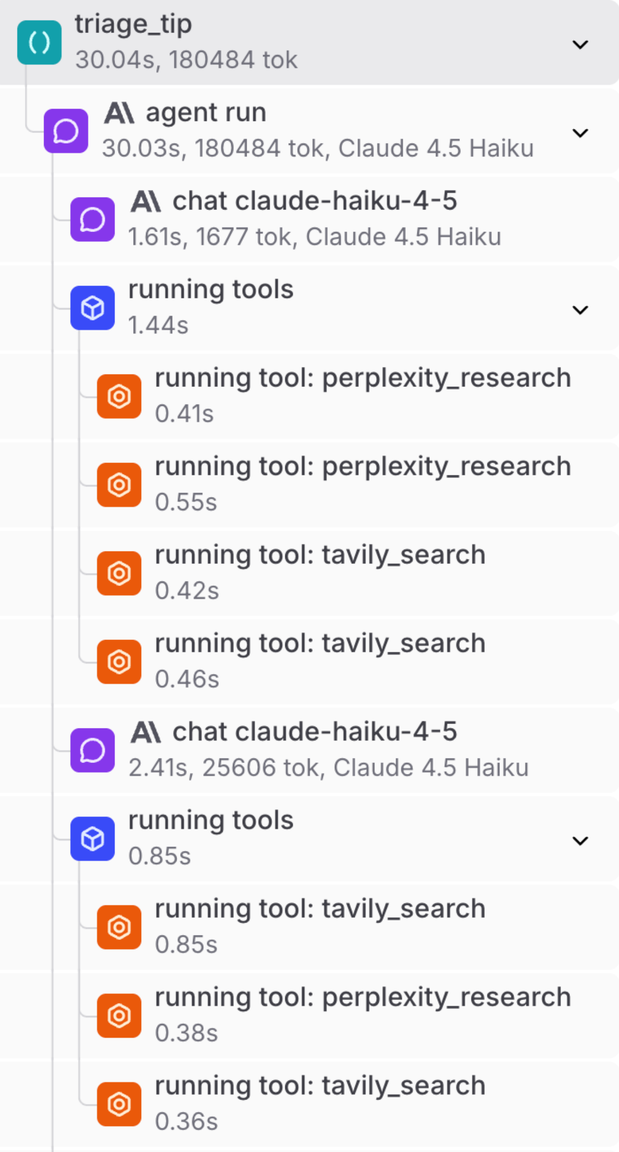
```python
@agent.tool_plain
async def perplexity_research(query: str):
    """Do research on a topic using Perplexity.

    Args:
        query: what to search for
    """
    search = client.search.create(query=query, max_results=10)
    return search.results

# Define the instructions for the agent

instructions = f"""
You are an investigative journalism assistant that triages news tips.

Your task is to:
1. Evaluate the tip for credibility and newsworthiness (score 1-10)
2. Determine if it is valid (score >= 5) or spam/invalid (score < 5)
3. Categorize it into the most appropriate beat from this list: local news, crime, politics, business, health, technology, environment, education, culture, sports, interr
4. Write a concise 2-3 sentence summary
5. Use Perplexity and/or Tavily search to conduct background research on the topic, entities, and location mentioned
6. Research the tipster's name and email to assess credibility (look for public profiles, previous journalism involvement, potential biases)
7. Generate 3-5 specific follow-up questions for further investigation
8. Compose a professional email response:
   - If valid: Thank them, indicate you're reviewing it, and ask the follow-up questions
   - If invalid: Politely decline and thank them for their interest

Be thorough in your research but efficient. Focus on verifiable facts and credible sources. Understand that people don't really know how to write tips, and you should fol
"""

# Define the output from the agent
class ProcessedTip(BaseModel):
    is_valid: bool
    credibility: int
    newsworthiness: int
    validity: int
    beat: str
    summary: str
    research: str
    follow_up_questions: list[str]
```

Instrumentation vocabulary

- **Logging:** Tracking what happens
- **Tool/tool call:** Using a service - web search, mail, etc
- **Trace:** Everything you need to reproduce interaction
  - All of the user’s inputs
  - All of the inputs/outputs to the LLM
- **Span:** A single step in the process
  - A tool call, the response, the time, the cost
- **Thread:** A group of traces, if multi-turn (ignore this??)
- **Observability:** Using instrumentation to track and analyze
- **Telemetry:** A more generic word for instrumentation



```
You are an investigative journalism assistant that triages
news tips.

Your task is to:
1. Evaluate the tip for credibility and newsworthiness (score
1-10)
2. Determine if it is valid (score >= 5) or spam/invalid
(score < 5)
3. Categorize it into the most appropriate beat from this
list:
4. Write a concise 2-3 sentence summary
5. Use Perplexity search to conduct background research on the
topic, entities, and location mentioned
6. Research the tipster's name and email to assess credibility
(look for public profiles, previous journalism involvement,
potential biases)
7. Generate 3-5 specific follow-up questions for further
investigation
8. Compose a professional email response:
   - If valid: Thank them, indicate you're reviewing it, and
ask the follow-up questions
   - If invalid: Politely decline and thank them for their
interest

Be thorough in your research but efficient. Focus on
verifiable facts and credible sources. Understand that people
don't really know how to write tips, and you should follow up
on most of the ones that are not obviously crazy people.
```



triage_tip
30.04s, 180484 tok

agent run
30.03s, 180484 tok, Claude 4.5 Haiku

chat claude-haiku-4-5
1.61s, 1677 tok, Claude 4.5 Haiku

running tools
1.44s

running tool: perplexity_research
0.41s

running tool: perplexity_research
0.55s

running tool: tavily_search
0.42s

running tool: tavily_search
0.46s

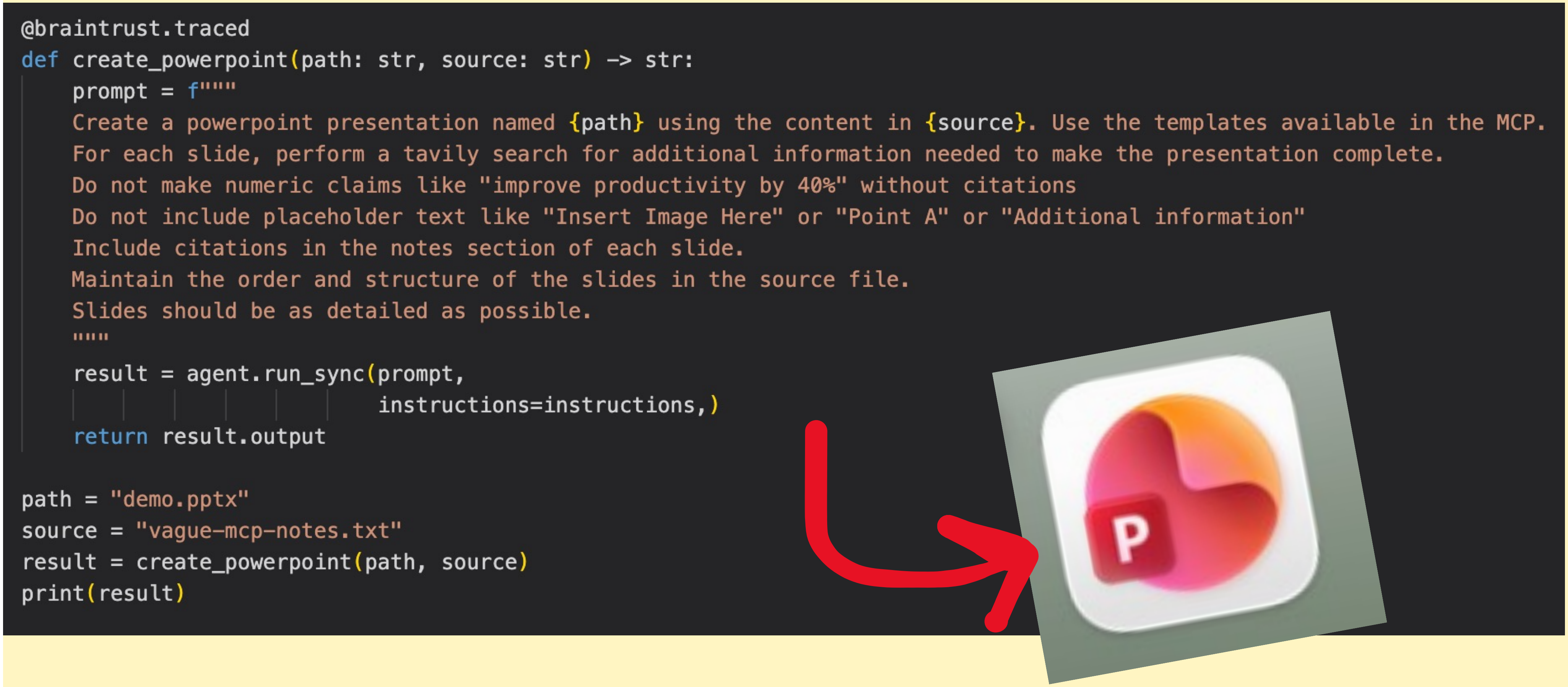
chat claude-haiku-4-5
2.41s, 25606 tok, Claude 4.5 Haiku

running tools
0.85s

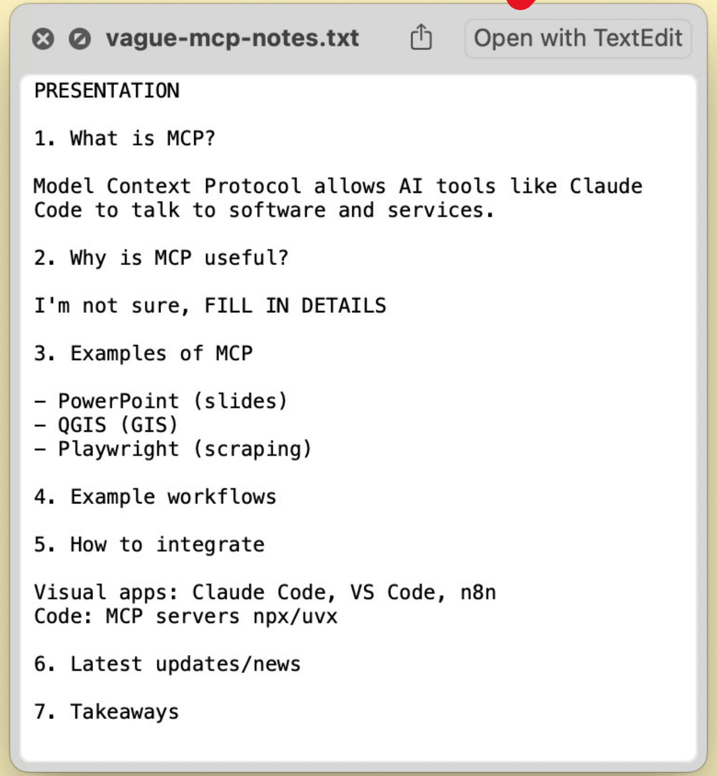
running tool: tavily_search
0.85s

running tool: perplexity_research
0.38s

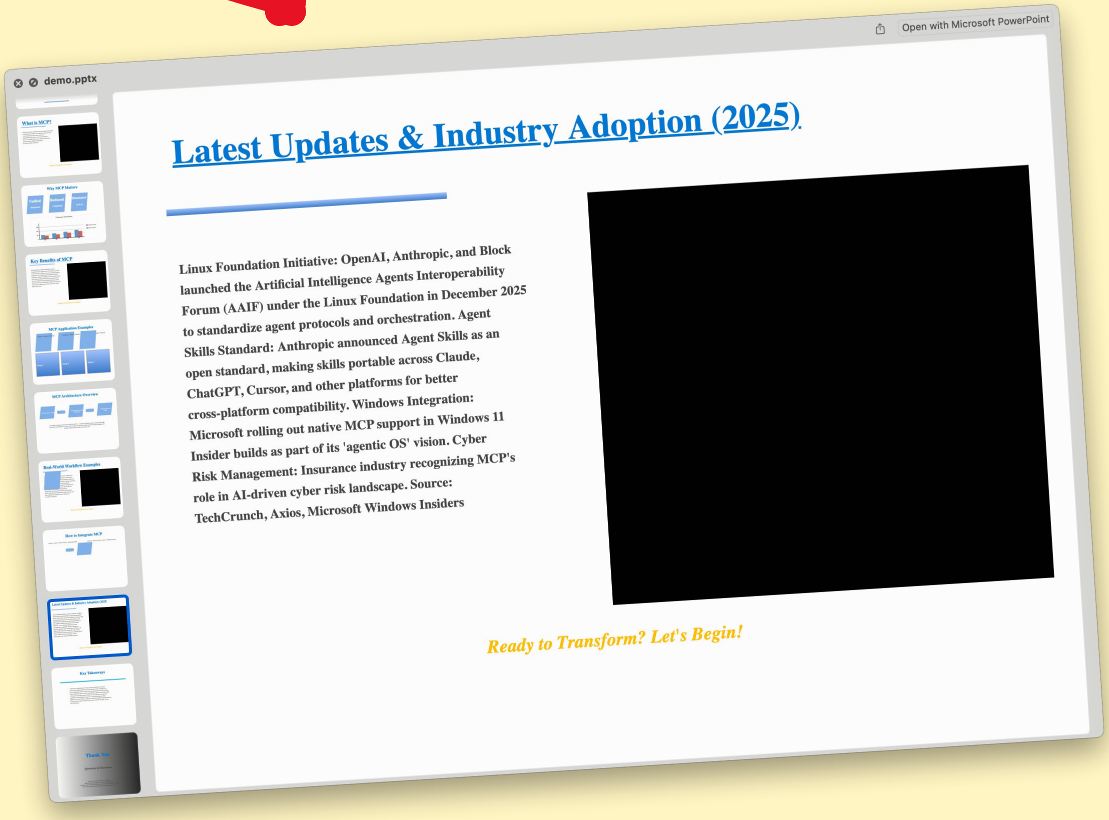
running tool: tavily_search
0.36s



```python
@braintrust.traced
def create_powerpoint(path: str, source: str) -> str:
    prompt = f"""
    Create a powerpoint presentation named {path} using the content in {source}. Use the templates available in the MCP.
    For each slide, perform a tavily search for additional information needed to make the presentation complete.
    Do not make numeric claims like "improve productivity by 40%" without citations
    Do not include placeholder text like "Insert Image Here" or "Point A" or "Additional information"
    Include citations in the notes section of each slide.
    Maintain the order and structure of the slides in the source file.
    Slides should be as detailed as possible.
    """
    result = agent.run_sync(prompt,
                            instructions=instructions,)
    return result.output

path = "demo.pptx"
source = "vague-mcp-notes.txt"
result = create_powerpoint(path, source)
print(result)
```



```
PRESENTATION

1. What is MCP?

Model Context Protocol allows AI tools like Claude
Code to talk to software and services.

2. Why is MCP useful?

I'm not sure, FILL IN DETAILS

3. Examples of MCP

- PowerPoint (slides)
- QGIS (GIS)
- Playwright (scraping)

4. Example workflows

5. How to integrate

Visual apps: Claude Code, VS Code, n8n
Code: MCP servers npx/uvx

6. Latest updates/news

7. Takeaways
```



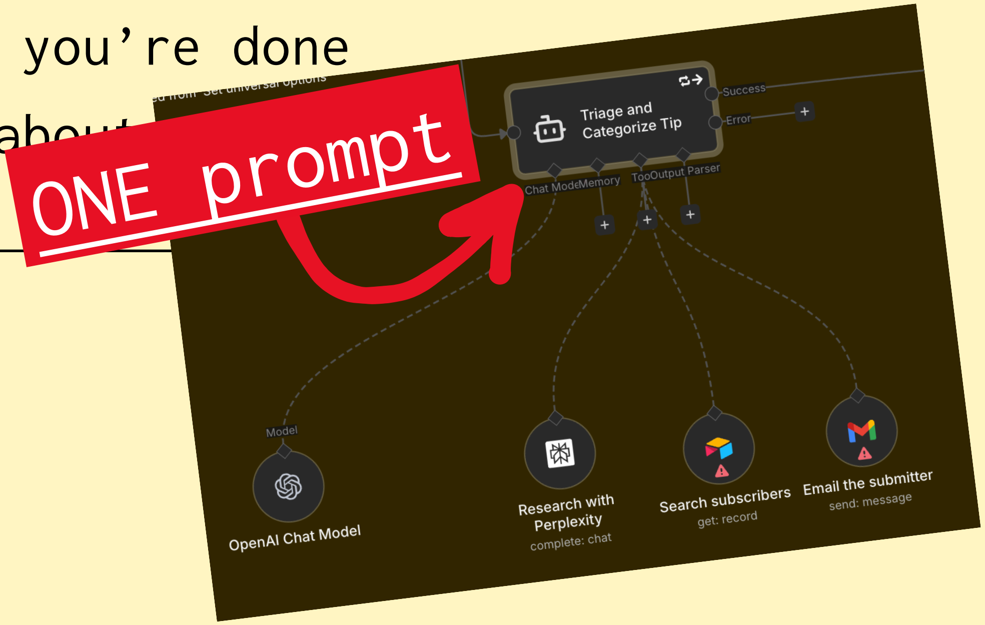
Latest Updates & Industry Adoption (2025)

Linux Foundation Initiative: OpenAI, Anthropic, and Block launched the Artificial Intelligence Agents Interoperability Forum (AAIF) under the Linux Foundation in December 2025 to standardize agent protocols and orchestration. Agent Skills Standard: Anthropic announced Agent Skills as an open standard, making skills portable across Claude, ChatGPT, Cursor, and other platforms for better cross-platform compatibility. Windows Integration: Microsoft rolling out native MCP support in Windows 11 Insider builds as part of its 'agentic OS' vision. Cyber Risk Management: Insurance industry recognizing MCP's role in AI-driven cyber risk landscape. Source: TechCrunch, Axios, Microsoft Windows Insiders

Ready to Transform? Let's Begin!

It's honestly not that complicated

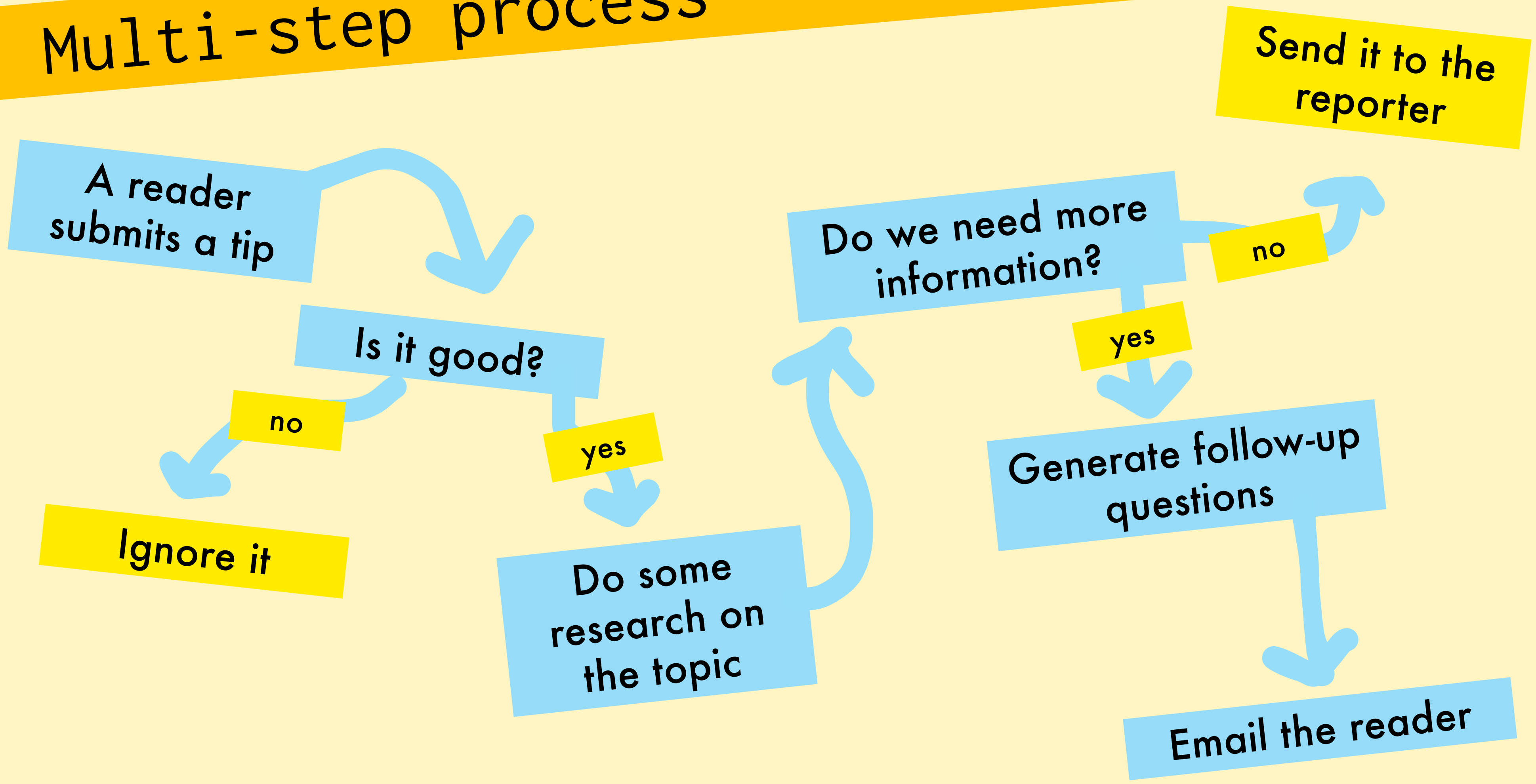
- You have a normal LLM setup
- You give it tools
- You say use the tools until you're done
- Why is everyone so excited abou
- *Because humans are so so so*


ONE prompt

We get a tip. What's next?

- *IMMEDIATELY PUBLISH IT, IN FULL.*
- Judge whether it's newsworthy
- See whether it's already been covered
- Research the topic in general with preliminary research
- Research the person who submitted the tip (social, etc)
- Assign it to a beat/specific reporter
- Write a follow-up email, send it

Multi-step process



A reader submits a tip

Is it good?

no

Ignore it

yes

Do some research on the topic

Do we need more information?

no

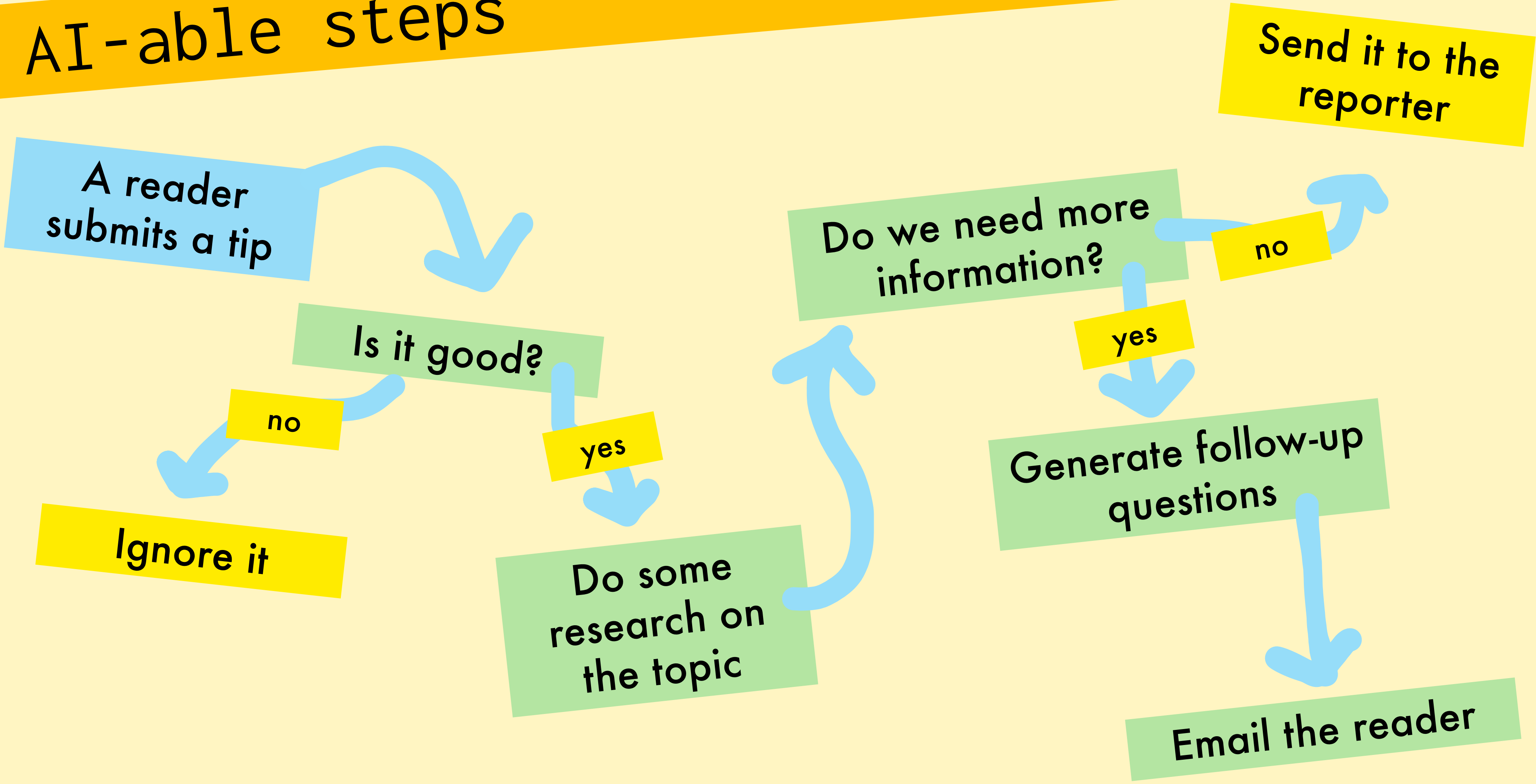
Send it to the reporter

yes

Generate follow-up questions

Email the reader

AI-able steps



A reader submits a tip

Is it good?

no

Ignore it

yes

Do some research on the topic

Do we need more information?

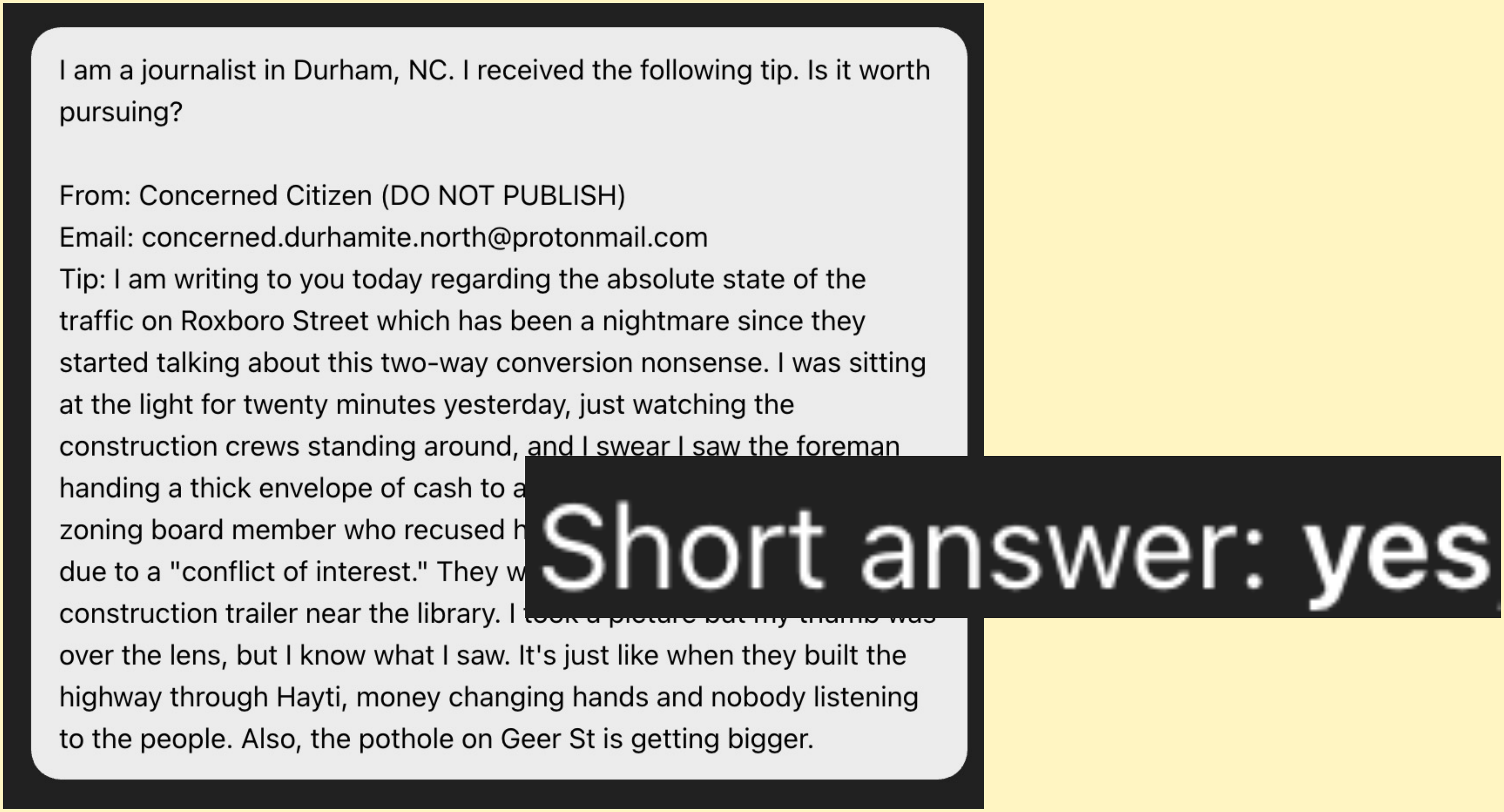
no

Send it to the reporter

yes

Generate follow-up questions

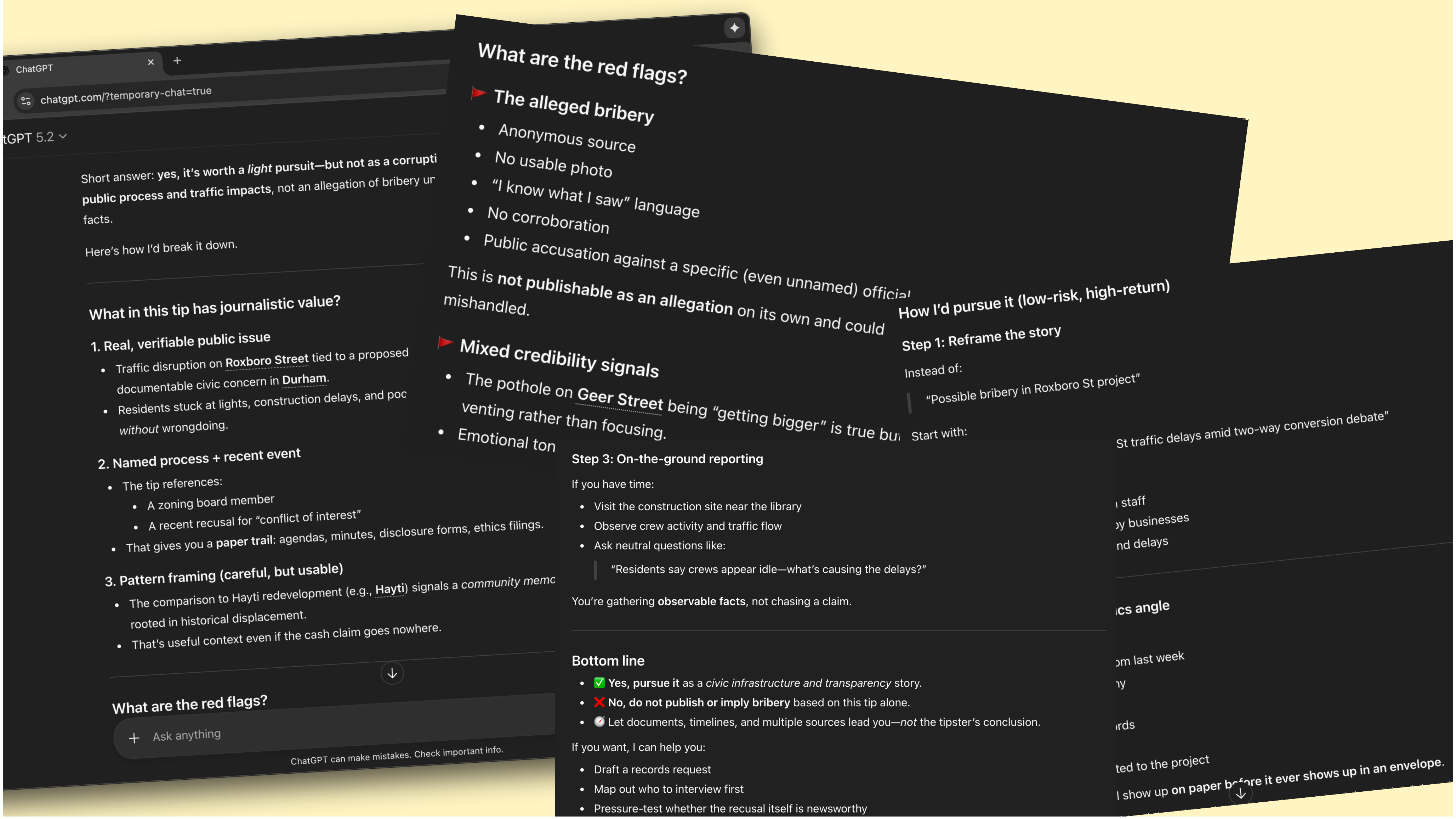
Email the reader



I am a journalist in Durham, NC. I received the following tip. Is it worth pursuing?

From: Concerned Citizen (DO NOT PUBLISH)
Email: concerned.durhamite.north@protonmail.com
Tip: I am writing to you today regarding the absolute state of the traffic on Roxboro Street which has been a nightmare since they started talking about this two-way conversion nonsense. I was sitting at the light for twenty minutes yesterday, just watching the construction crews standing around, and I swear I saw the foreman
handing a thick envelope of cash to a
zoning board member who recused h
due to a "conflict of interest." They w
construction trailer near the library. I
over the lens, but I know what I saw. It's just like when they built the highway through Hayti, money changing hands and nobody listening to the people. Also, the pothole on Geer St is getting bigger.

Short answer: yes



Short answer: **yes, it's worth a *light* pursuit—but not as a corrupti** **public process and traffic impacts**, not an allegation of bribery ur facts.

Here's how I'd break it down.

What in this tip has journalistic value?

1. Real, verifiable public issue

- Traffic disruption on **Roxboro Street** tied to a proposed documentable civic concern in **Durham**.
- Residents stuck at lights, construction delays, and poo *without* wrongdoing.

2. Named process + recent event

- The tip references:
  - A zoning board member
  - A recent recusal for "conflict of interest"
- That gives you a **paper trail**: agendas, minutes, disclosure forms, ethics filings.

3. Pattern framing (careful, but usable)

- The comparison to Hayti redevelopment (e.g., **Hayti**) signals a *community memo* rooted in historical displacement.
- That's useful context even if the cash claim goes nowhere.

What are the red flags?

What are the red flags?

🚩 The alleged bribery

- Anonymous source
- No usable photo
- "I know what I saw" language
- No corroboration
- Public accusation against a specific (even unnamed) officia

This is **not publishable as an allegation** on its own and could mishandled.

🚩 Mixed credibility signals

- The pothole on **Geer Street** being "getting bigger" is true bu venting rather than focusing.
- Emotional ton

How I'd pursue it (low-risk, high-return)

Step 1: Reframe the story

Instead of:

> "Possible bribery in Roxboro St project"

Start with:

St traffic delays amid two-way conversion debate"

staff
by businesses
nd delays

ics angle

om last week
ny

rds

ted to the project
l show up **on paper** before it ever shows up in an envelope.

Step 3: On-the-ground reporting

If you have time:

- Visit the construction site near the library
- Observe crew activity and traffic flow
- Ask neutral questions like:

> "Residents say crews appear idle—what's causing the delays?"

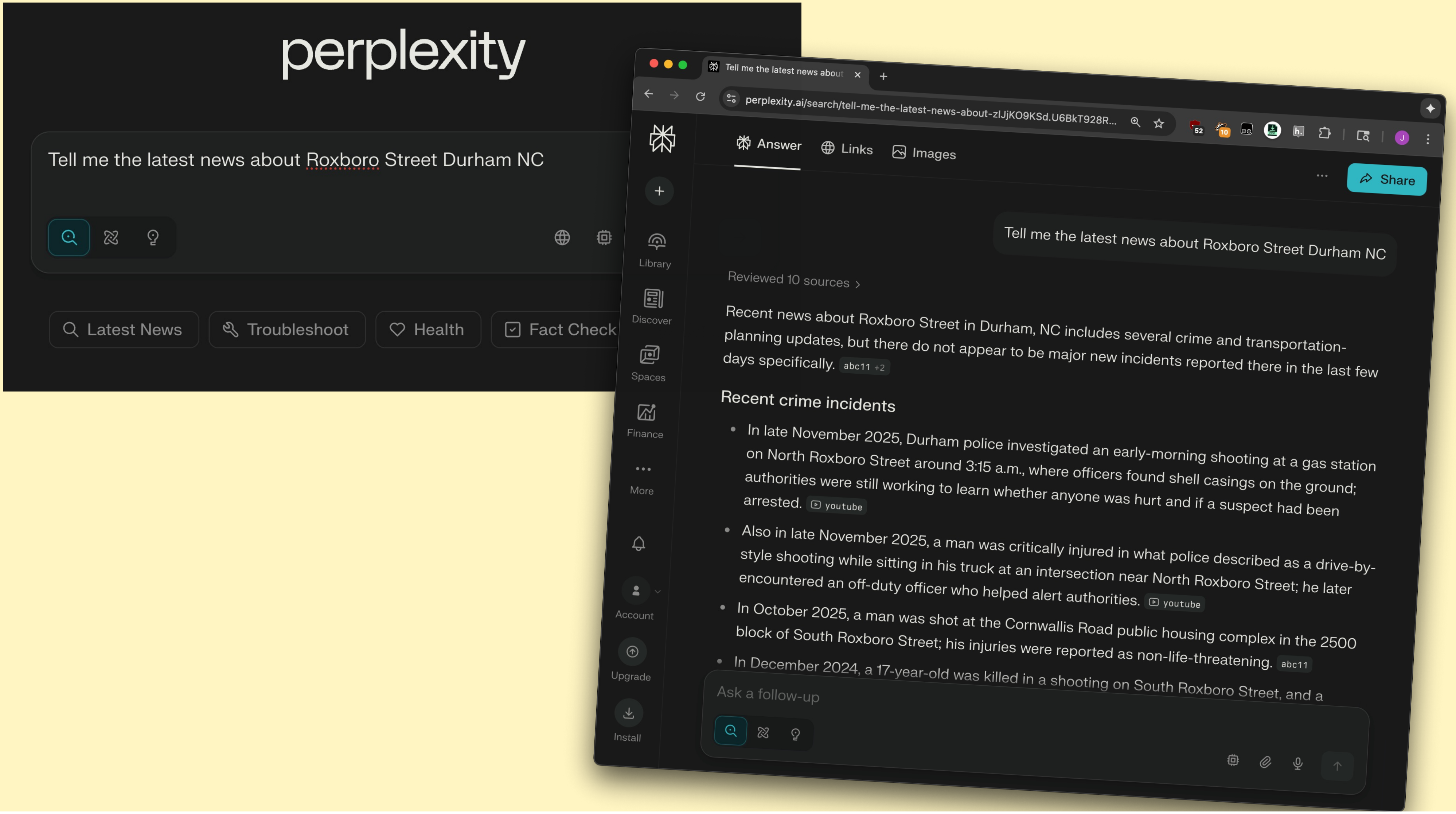
You're gathering **observable facts**, not chasing a claim.

Bottom line

- ✅ **Yes, pursue it** as a *civic infrastructure and transparency* story.
- ❌ **No, do not publish or imply bribery** based on this tip alone.
- 🧭 Let documents, timelines, and multiple sources lead you—*not* the tipster's conclusion.

If you want, I can help you:

- Draft a records request
- Map out who to interview first
- Pressure-test whether the recusal itself is newsworthy



Tell me the latest news about Roxboro Street Durham NC

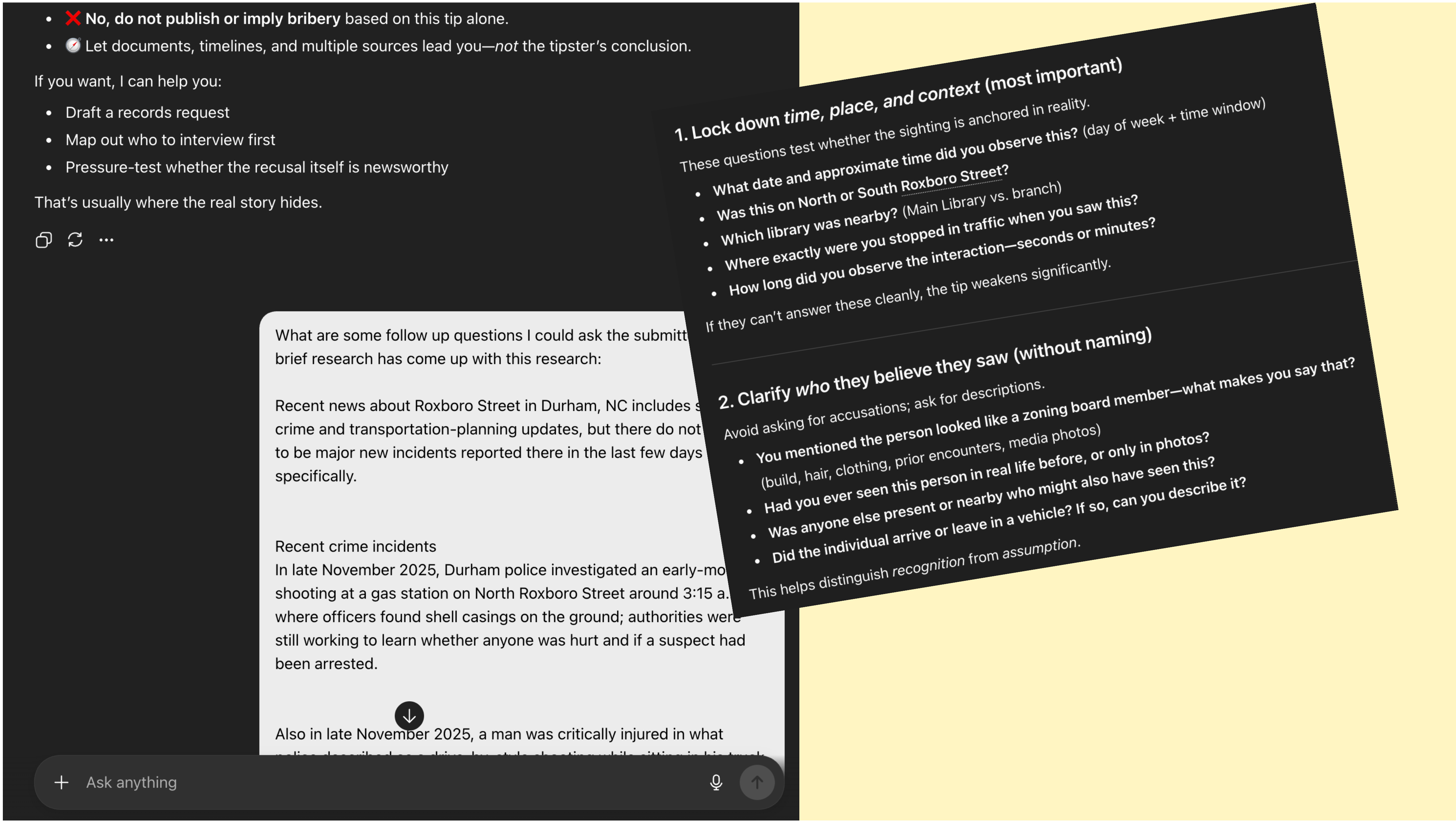
Tell me the latest news about Roxboro Street Durham NC

Reviewed 10 sources >

Recent news about Roxboro Street in Durham, NC includes several crime and transportation-planning updates, but there do not appear to be major new incidents reported there in the last few days specifically. abc11 +2

Recent crime incidents

- In late November 2025, Durham police investigated an early-morning shooting at a gas station on North Roxboro Street around 3:15 a.m., where officers found shell casings on the ground; authorities were still working to learn whether anyone was hurt and if a suspect had been arrested. youtube
- Also in late November 2025, a man was critically injured in what police described as a drive-by-style shooting while sitting in his truck at an intersection near North Roxboro Street; he later encountered an off-duty officer who helped alert authorities. youtube
- In October 2025, a man was shot at the Cornwallis Road public housing complex in the 2500 block of South Roxboro Street; his injuries were reported as non-life-threatening. abc11
- In December 2024, a 17-year-old was killed in a shooting on South Roxboro Street, and a



- ❌ **No, do not publish or imply bribery** based on this tip alone.
- 🧭 Let documents, timelines, and multiple sources lead you—*not* the tipster's conclusion.

If you want, I can help you:

- Draft a records request
- Map out who to interview first
- Pressure-test whether the recusal itself is newsworthy

That's usually where the real story hides.

What are some follow up questions I could ask the submitt
brief research has come up with this research:

Recent news about Roxboro Street in Durham, NC includes s
crime and transportation-planning updates, but there do not
to be major new incidents reported there in the last few days
specifically.

Recent crime incidents
In late November 2025, Durham police investigated an early-mo
shooting at a gas station on North Roxboro Street around 3:15 a.
where officers found shell casings on the ground; authorities were
still working to learn whether anyone was hurt and if a suspect had
been arrested.

Also in late November 2025, a man was critically injured in what

1. Lock down *time, place, and context* (most important)

These questions test whether the sighting is anchored in reality.

- **What date and approximate time did you observe this?** (day of week + time window)
- **Was this on North or South Roxboro Street?**
- **Which library was nearby?** (Main Library vs. branch)
- **Where exactly were you stopped in traffic when you saw this?**
- **How long did you observe the interaction—seconds or minutes?**

If they can't answer these cleanly, the tip weakens significantly.

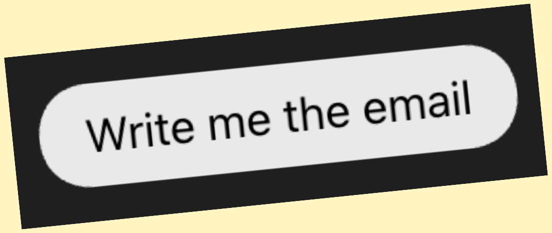
---

2. Clarify *who* they believe they saw (without naming)

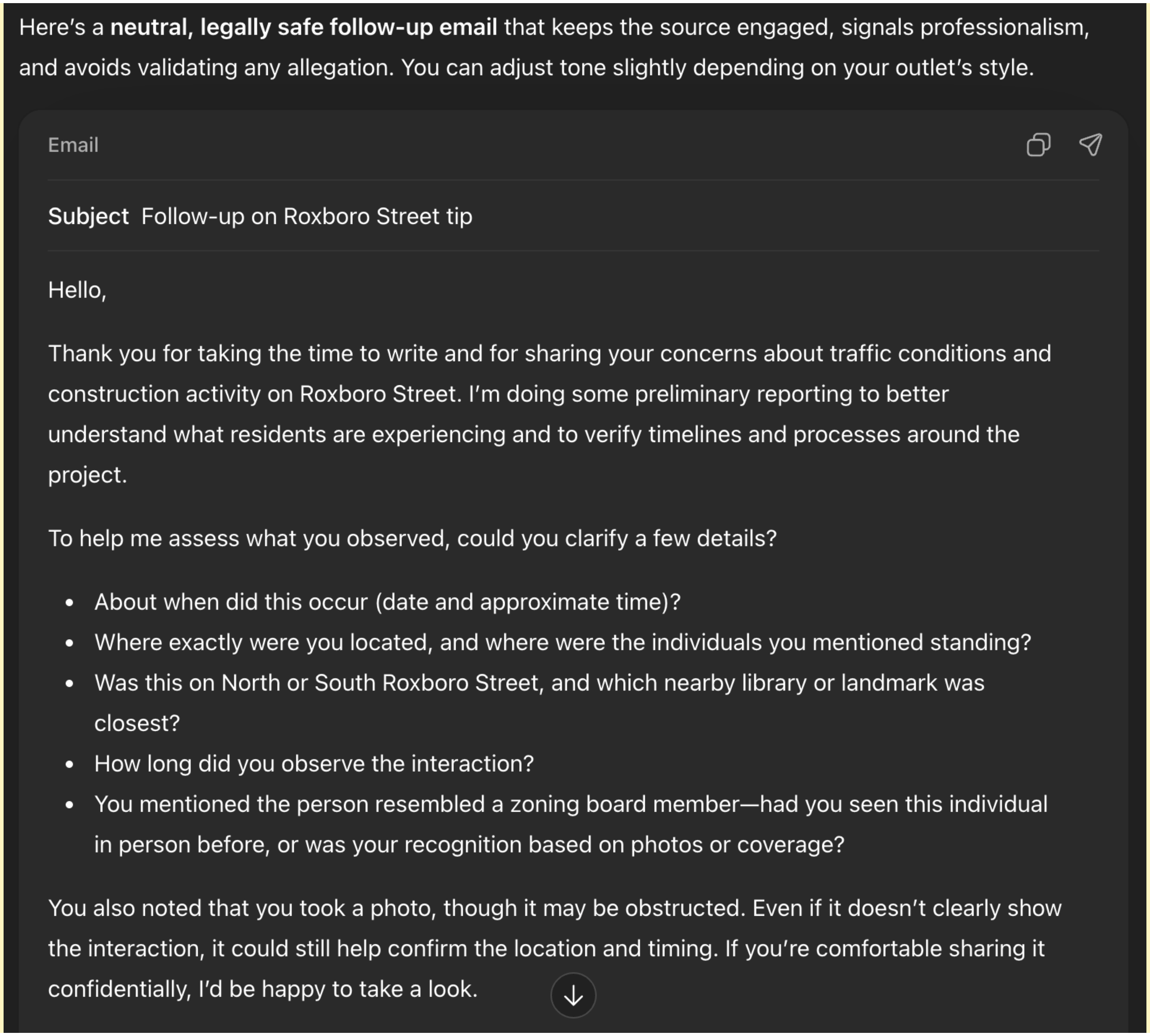
Avoid asking for accusations; ask for descriptions.

- **You mentioned the person looked like a zoning board member—what makes you say that?** (build, hair, clothing, prior encounters, media photos)
- **Had you ever seen this person in real life before, or only in photos?**
- **Was anyone else present or nearby who might also have seen this?**
- **Did the individual arrive or leave in a vehicle? If so, can you describe it?**

This helps distinguish *recognition* from *assumption*.



Write me the email



Here's a **neutral, legally safe follow-up email** that keeps the source engaged, signals professionalism, and avoids validating any allegation. You can adjust tone slightly depending on your outlet's style.

Subject Follow-up on Roxboro Street tip

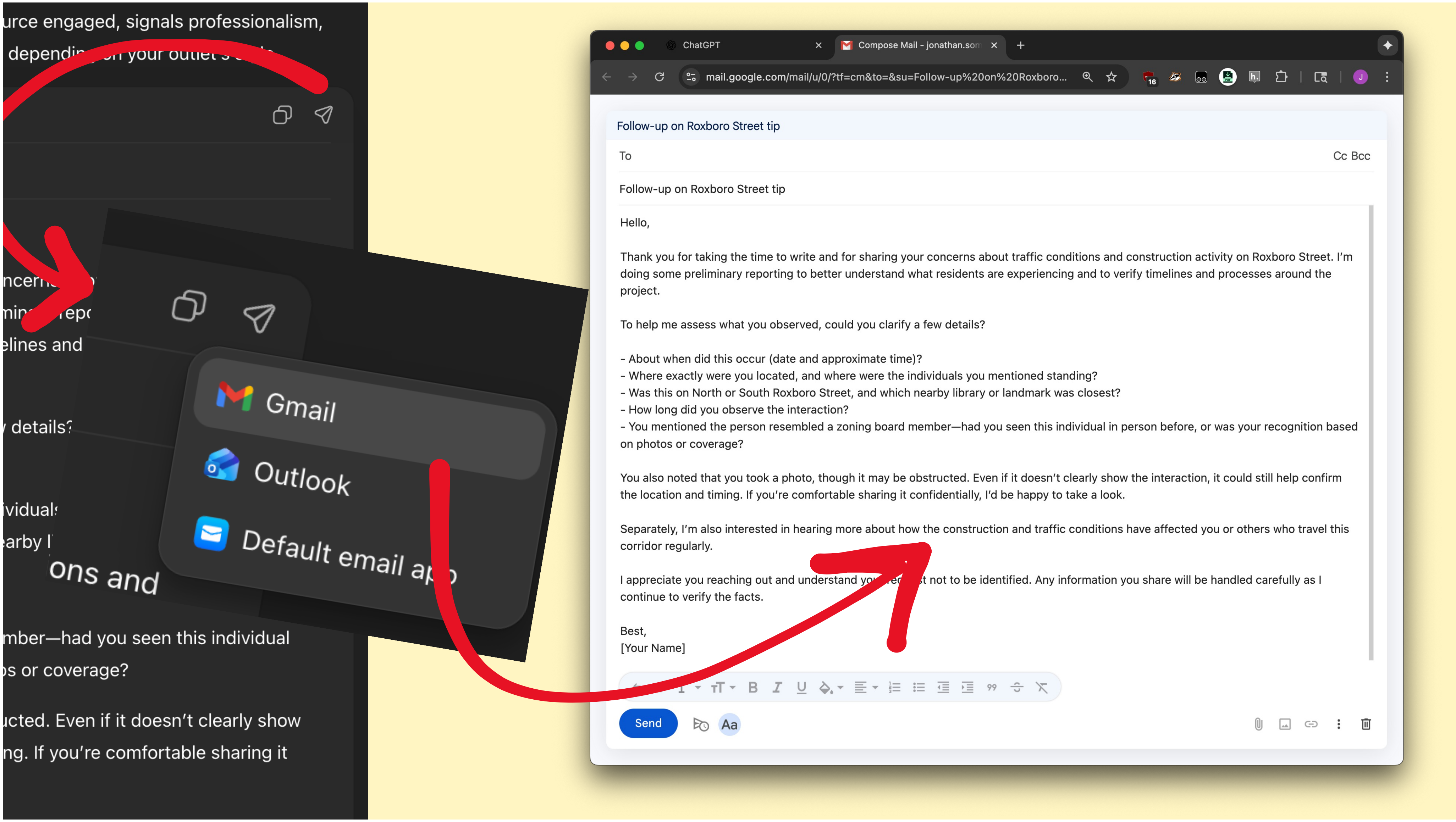
Hello,

Thank you for taking the time to write and for sharing your concerns about traffic conditions and construction activity on Roxboro Street. I'm doing some preliminary reporting to better understand what residents are experiencing and to verify timelines and processes around the project.

To help me assess what you observed, could you clarify a few details?

- About when did this occur (date and approximate time)?
- Where exactly were you located, and where were the individuals you mentioned standing?
- Was this on North or South Roxboro Street, and which nearby library or landmark was closest?
- How long did you observe the interaction?
- You mentioned the person resembled a zoning board member—had you seen this individual in person before, or was your recognition based on photos or coverage?

You also noted that you took a photo, though it may be obstructed. Even if it doesn't clearly show the interaction, it could still help confirm the location and timing. If you're comfortable sharing it confidentially, I'd be happy to take a look.


Gmail
Outlook
Default email app
ChatGPT
Compose Mail - jonathan.som
mail.google.com/mail/u/0/?tf=cm&to=&su=Follow-up%20on%20Roxboro...
Follow-up on Roxboro Street tip
To
Cc Bcc
Send

Follow-up on Roxboro Street tip

Hello,

Thank you for taking the time to write and for sharing your concerns about traffic conditions and construction activity on Roxboro Street. I'm doing some preliminary reporting to better understand what residents are experiencing and to verify timelines and processes around the project.

To help me assess what you observed, could you clarify a few details?

- About when did this occur (date and approximate time)?
- Where exactly were you located, and where were the individuals you mentioned standing?
- Was this on North or South Roxboro Street, and which nearby library or landmark was closest?
- How long did you observe the interaction?
- You mentioned the person resembled a zoning board member—had you seen this individual in person before, or was your recognition based on photos or coverage?

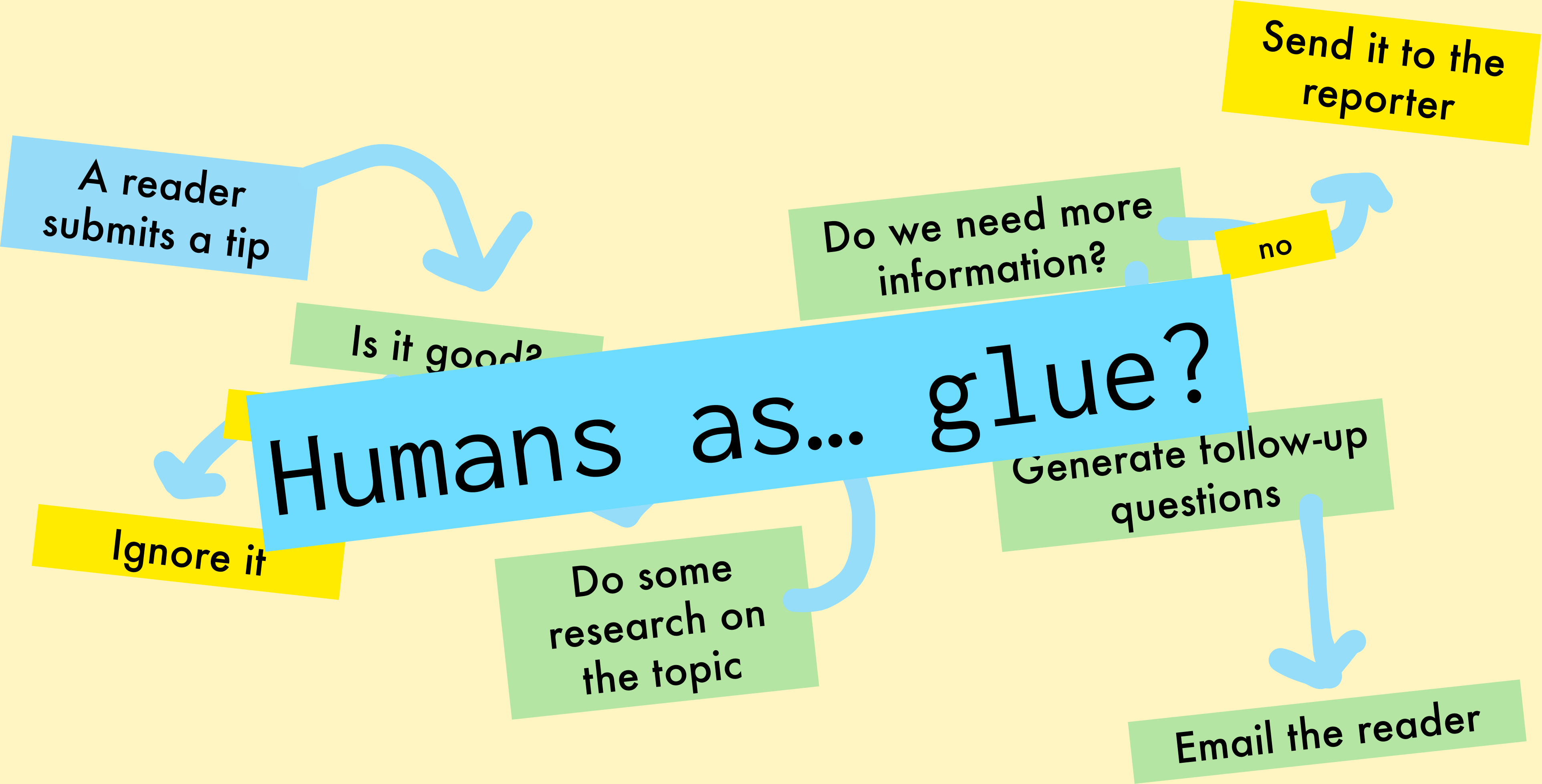
You also noted that you took a photo, though it may be obstructed. Even if it doesn't clearly show the interaction, it could still help confirm the location and timing. If you're comfortable sharing it confidentially, I'd be happy to take a look.

Separately, I'm also interested in hearing more about how the construction and traffic conditions have affected you or others who travel this corridor regularly.

I appreciate you reaching out and understand your request not to be identified. Any information you share will be handled carefully as I continue to verify the facts.

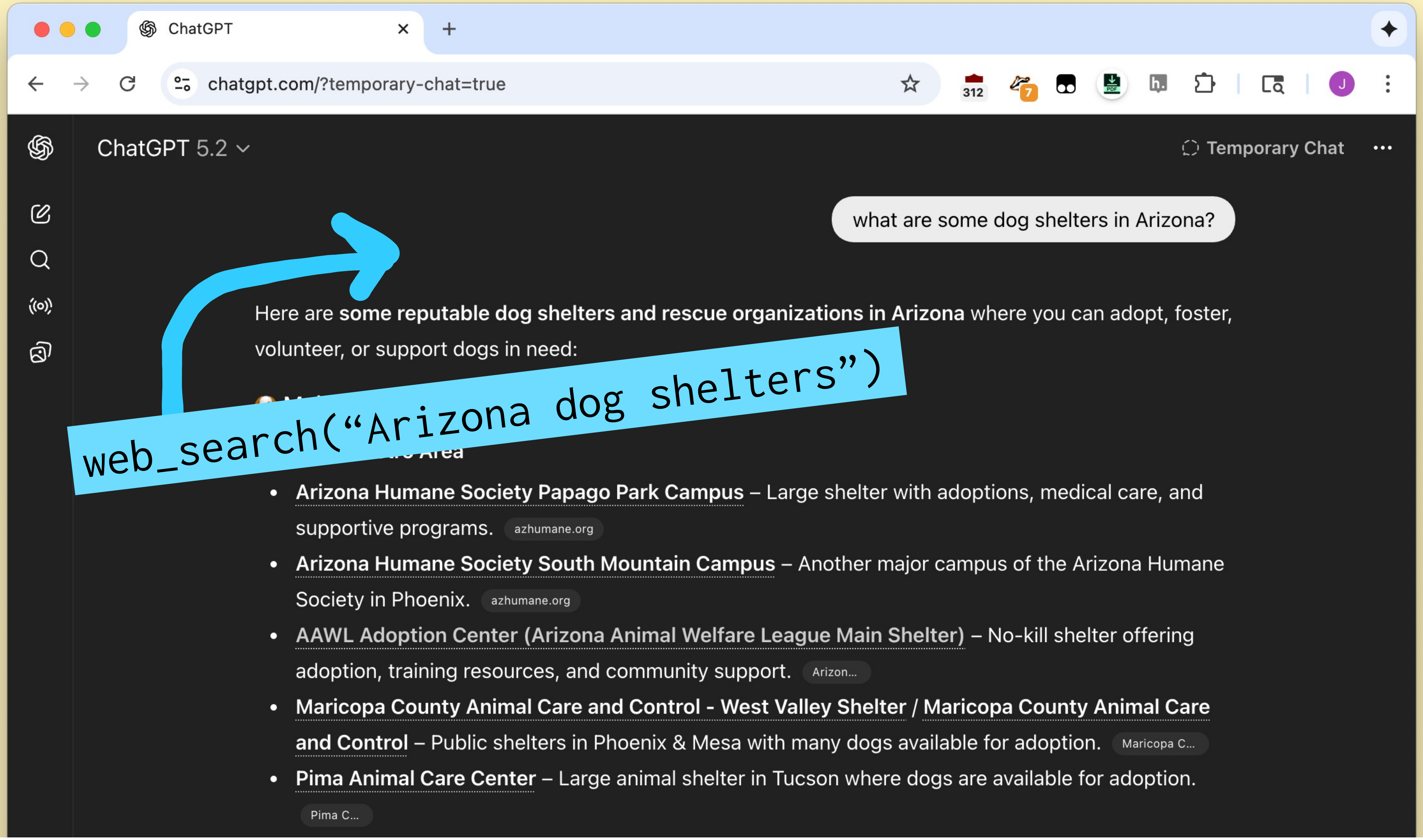
Best,
[Your Name]

Humans as… glue?


A reader submits a tip
Is it good?
Ignore it
Do some research on the topic
Do we need more information?
no
Send it to the reporter
Generate follow-up questions
Email the reader

LLM Pipelines/Automated workflows

Let the humans take a break, use the AI to fill those gaps



web_search(“Arizona dog shelters”)

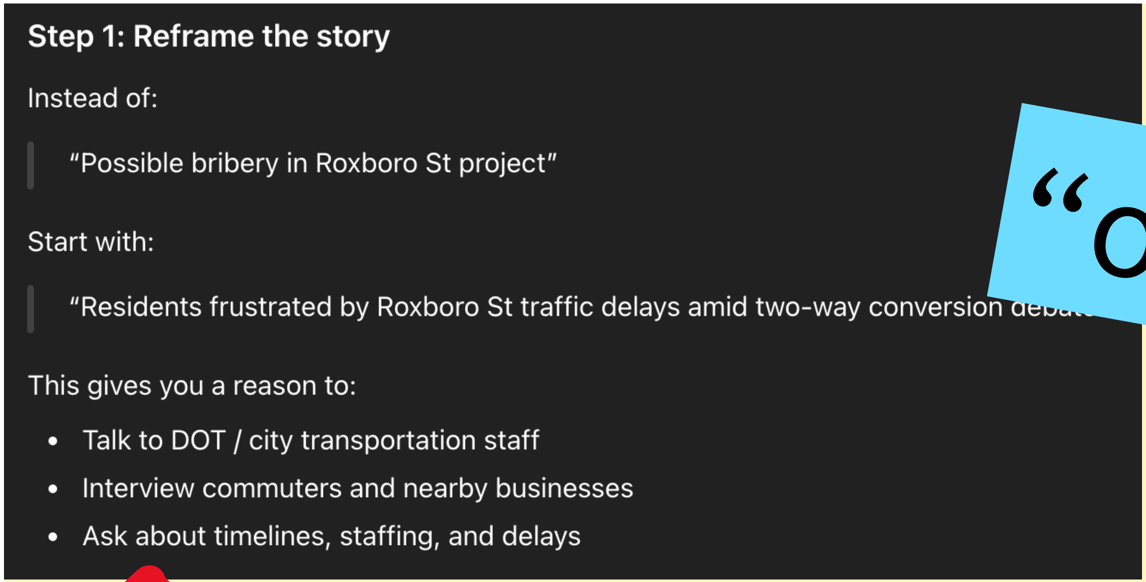
what are some dog shelters in Arizona?

Here are **some reputable dog shelters and rescue organizations in Arizona** where you can adopt, foster, volunteer, or support dogs in need:

- **Arizona Humane Society Papago Park Campus** – Large shelter with adoptions, medical care, and supportive programs. azhumane.org
- **Arizona Humane Society South Mountain Campus** – Another major campus of the Arizona Humane Society in Phoenix. azhumane.org
- AAWL Adoption Center (Arizona Animal Welfare League Main Shelter) – No-kill shelter offering adoption, training resources, and community support. Arizon…
- **Maricopa County Animal Care and Control - West Valley Shelter** / **Maricopa County Animal Care and Control** – Public shelters in Phoenix & Mesa with many dogs available for adoption. Maricopa C…
- **Pima Animal Care Center** – Large animal shelter in Tucson where dogs are available for adoption. Pima C…

Tools in LLMs

- Allow an LLM to do *stuff* that isn't text generation
- web_search("Whatever you want to look for")
- generate_image("whatever you want an image about")
- other_things("I don't know, it's infinite??")



Step 1: Reframe the story

Instead of:

> “Possible bribery in Roxboro St project”

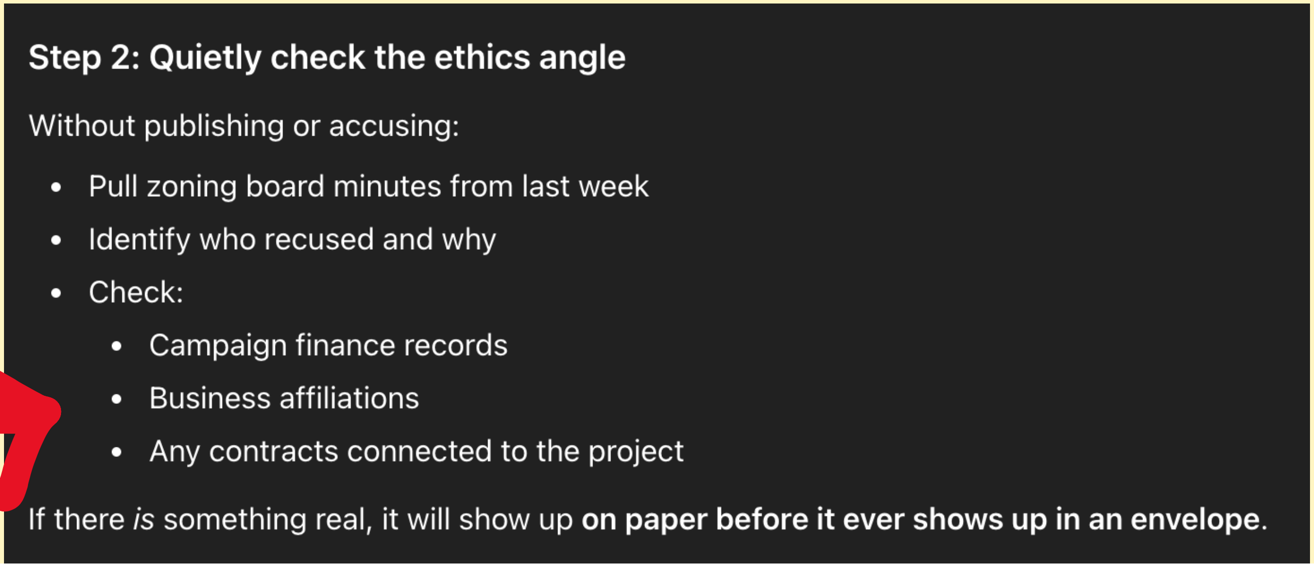
Start with:

> “Residents frustrated by Roxboro St traffic delays amid two-way conversion debat

This gives you a reason to:

- Talk to DOT / city transportation staff
- Interview commuters and nearby businesses
- Ask about timelines, staffing, and delays

“ok, do that”



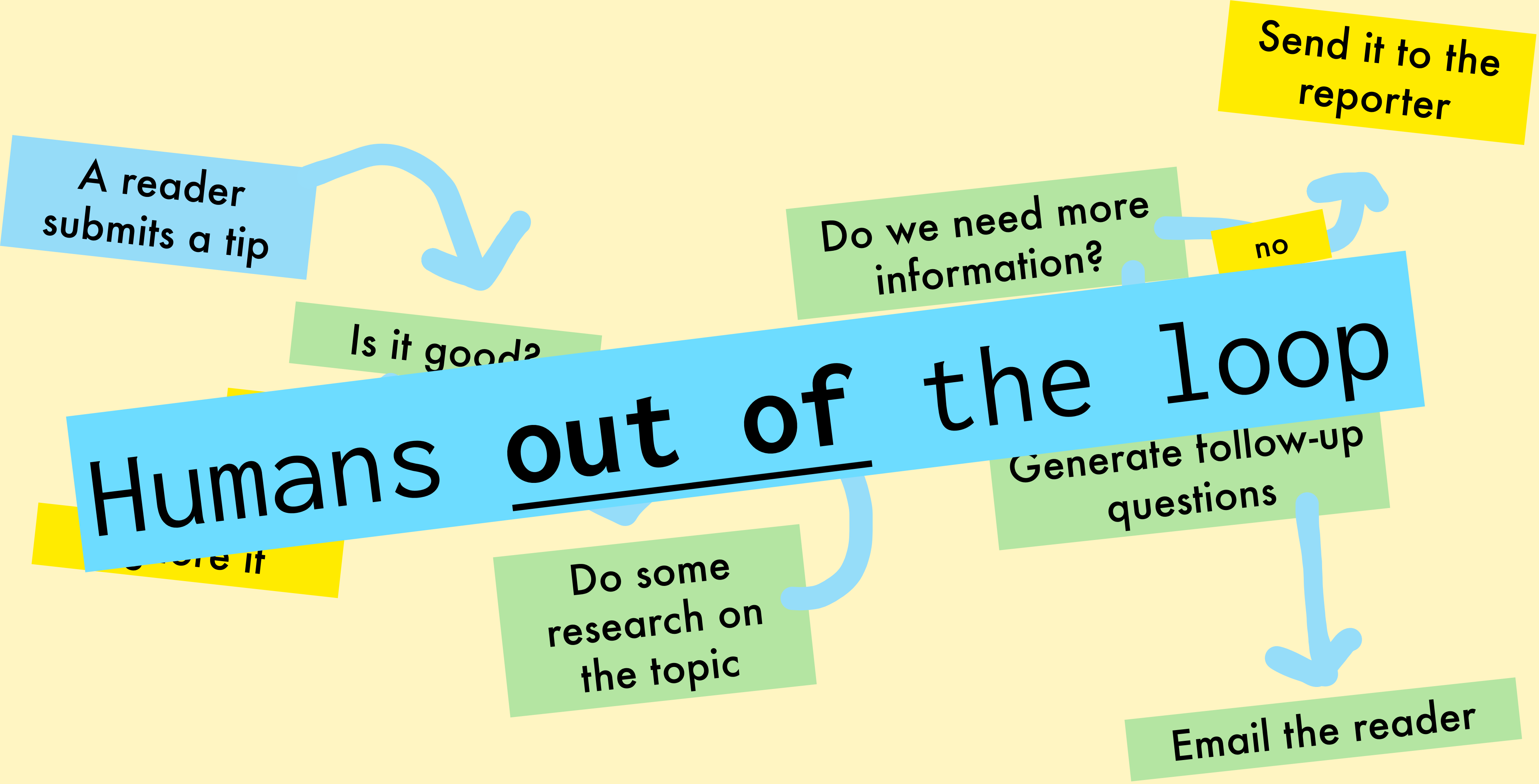
Step 2: Quietly check the ethics angle

Without publishing or accusing:

- Pull zoning board minutes from last week
- Identify who recused and why
- Check:
  - Campaign finance records
  - Business affiliations
  - Any contracts connected to the project

If there *is* something real, it will show up **on paper before it ever shows up in an envelope**.

Humans **out of** the loop



A reader submits a tip

Is it good?

gnore it

Do some research on the topic

Do we need more information?

no

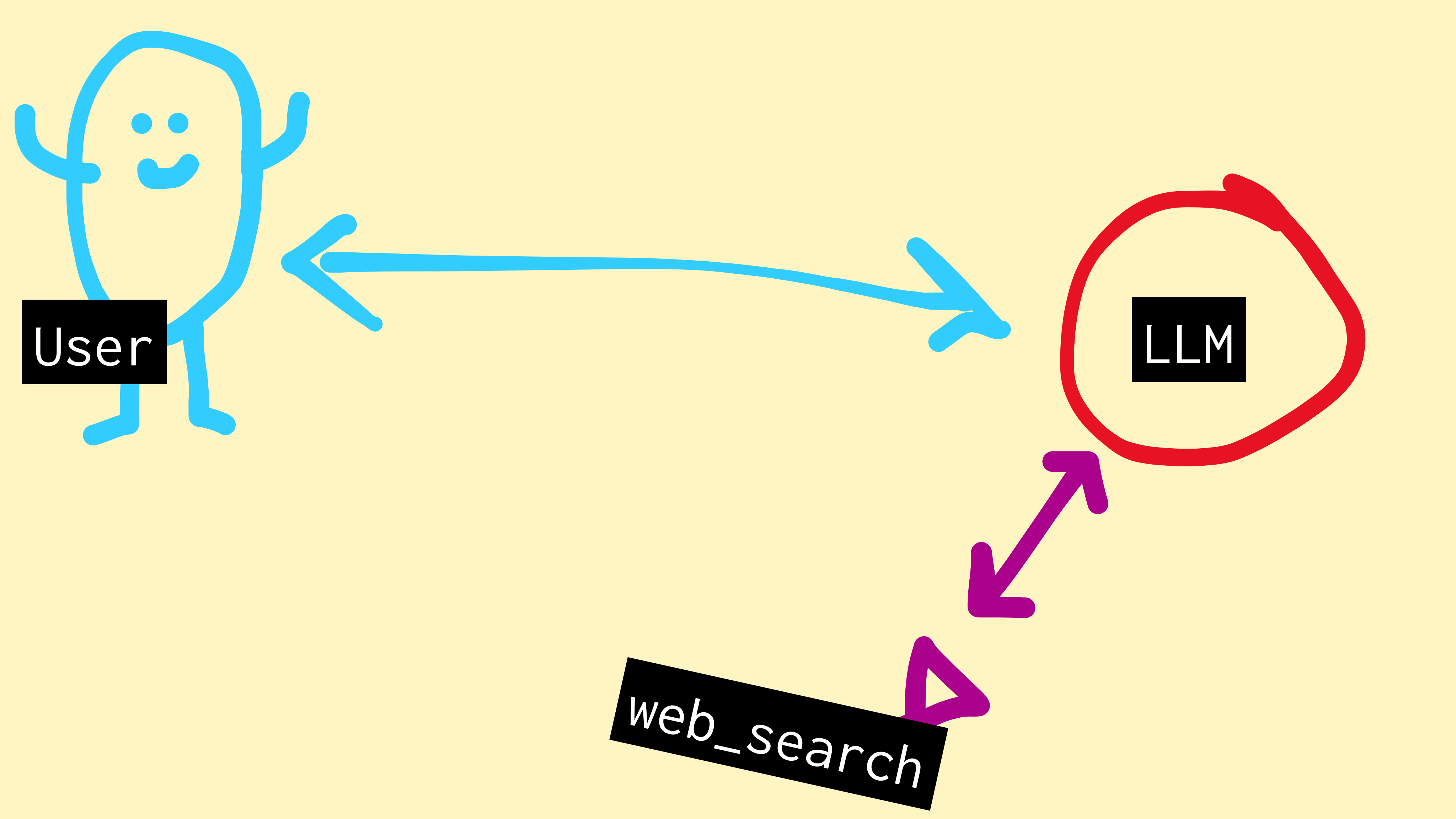
Send it to the reporter

Generate follow-up questions

Email the reader

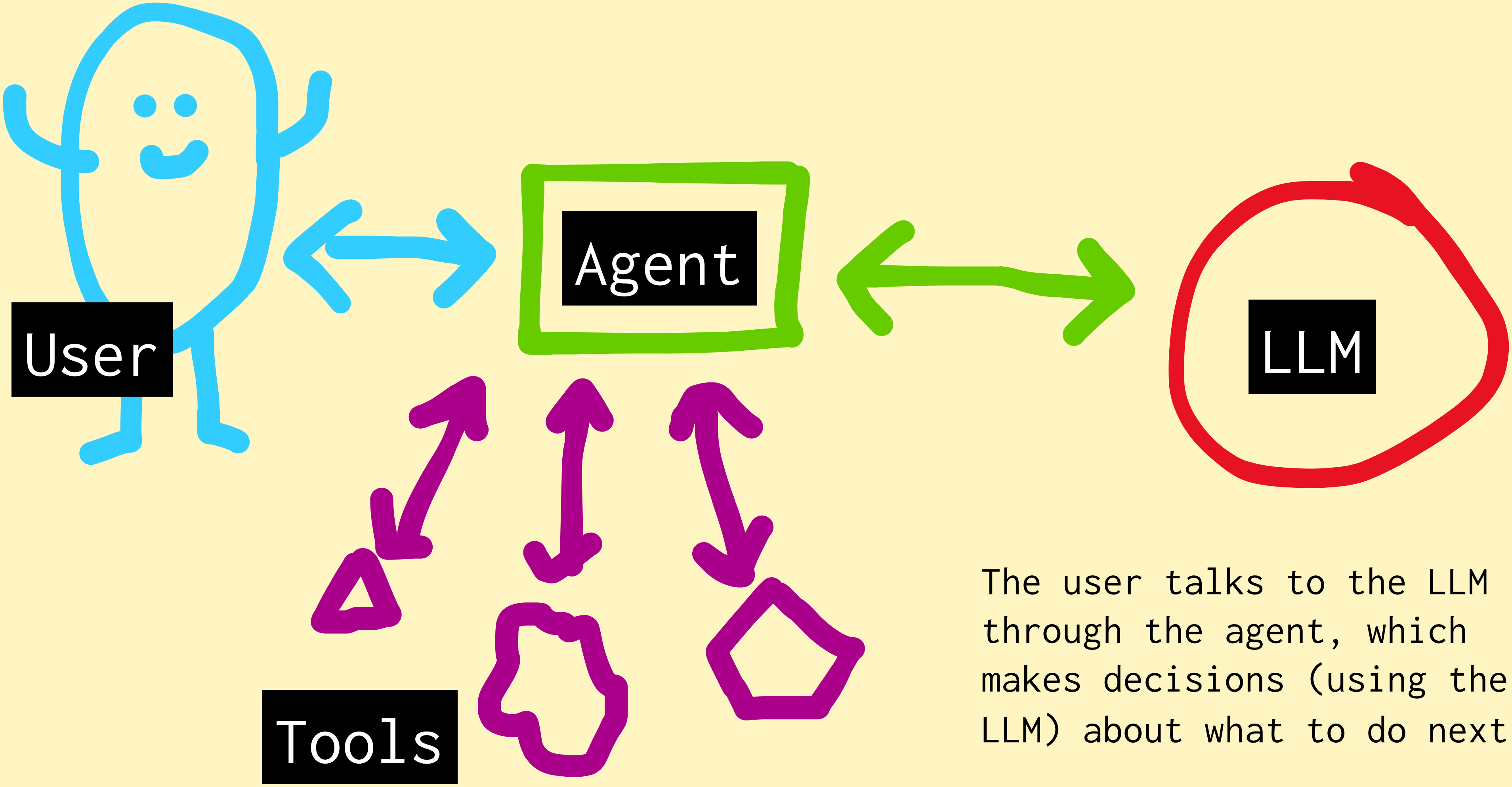
Agentic AI workflows

Let’s REALLY give humans take a break and hand over control to AI


User
LLM
web_search

One-step vs multi-step pipeline

- **Find me dog shelters in Arizona**
  - web_search(“dog shelters Arizona”)
  - OK sounds good here are the results
- **Find me contact info for every dog shelter in Arizona**
  - web_search(“dog shelters Arizona”)
  - Visit website 1
  - web_search(“Happy Paws Rescue email address”)
  - Visit website 2
  - …etc…etc…etc…etc…………
- Agentic AI: **uncertain when it’s done**



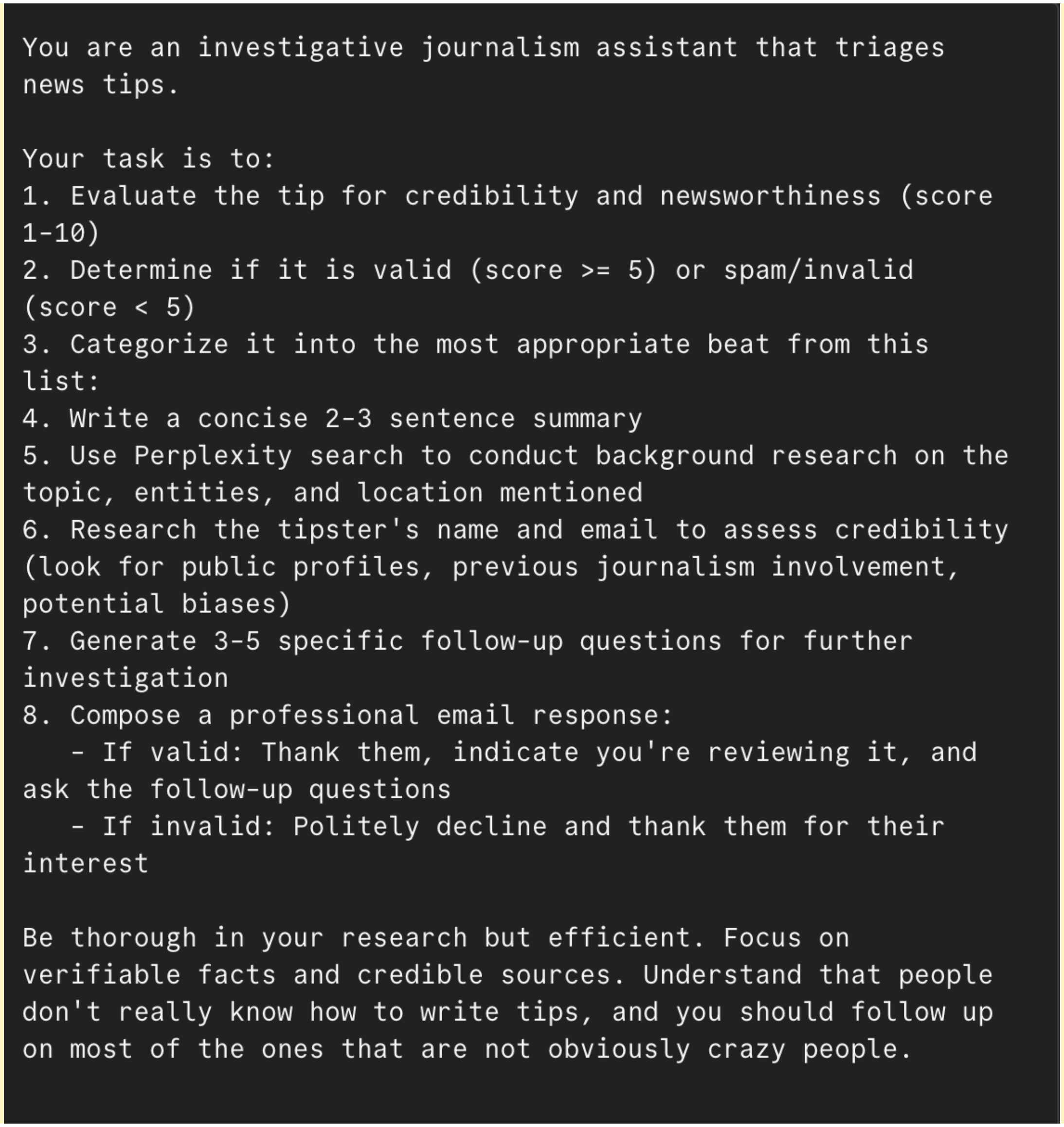
User

Agent

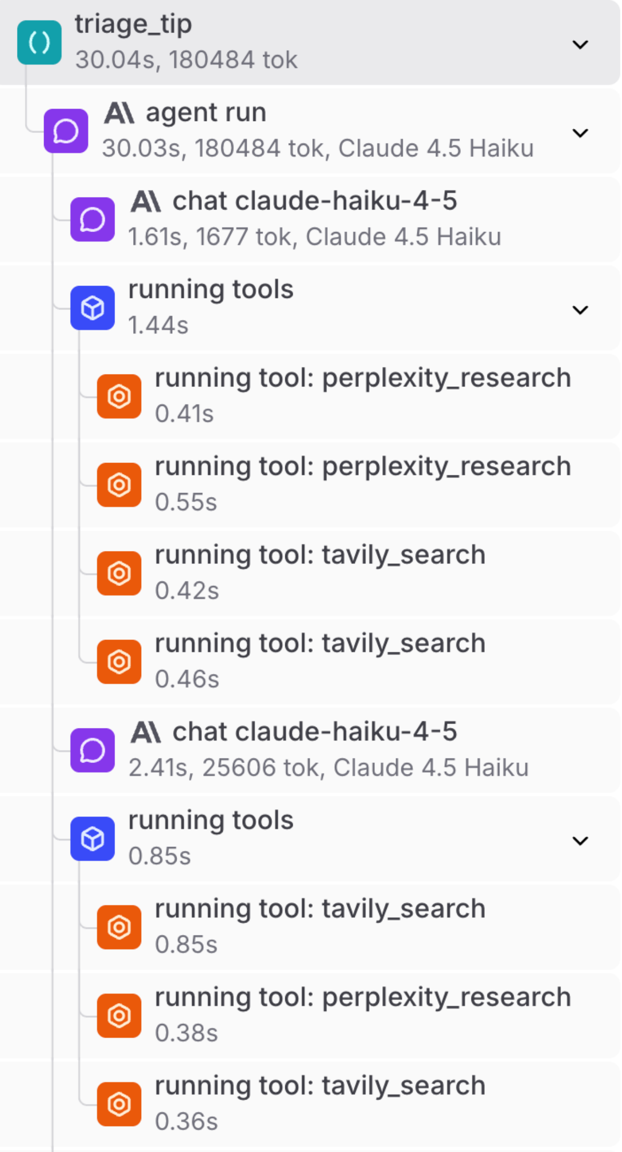
LLM

Tools

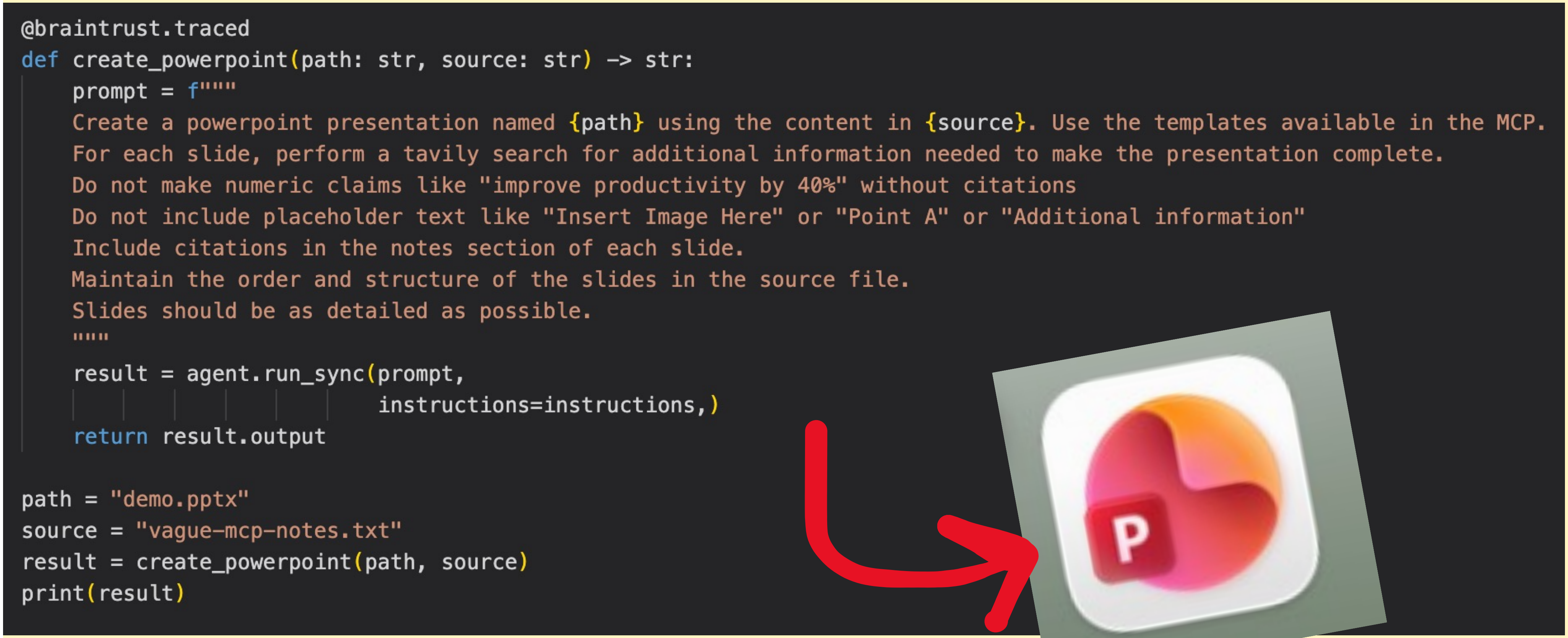
The user talks to the LLM through the agent, which makes decisions (using the LLM) about what to do next.



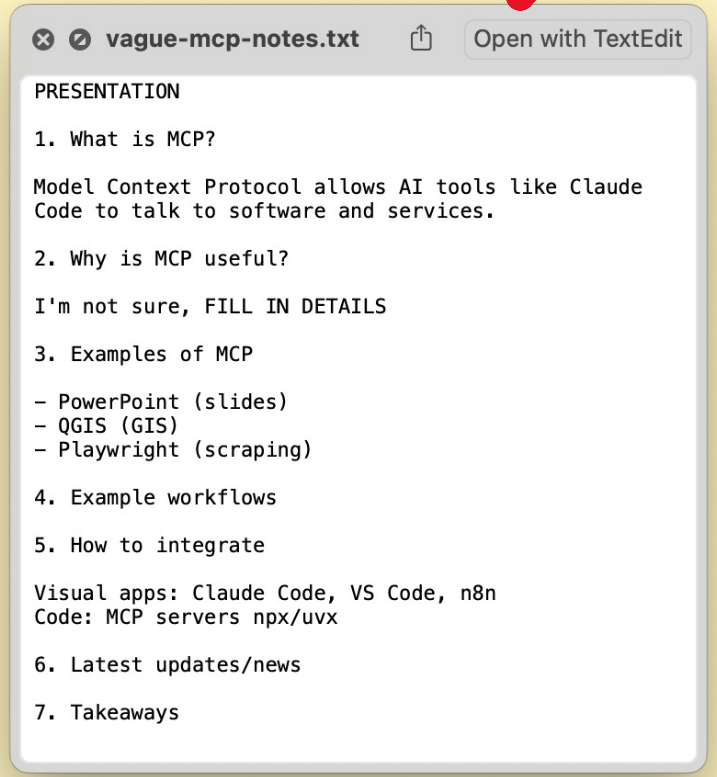
```
You are an investigative journalism assistant that triages
news tips.

Your task is to:
1. Evaluate the tip for credibility and newsworthiness (score
1-10)
2. Determine if it is valid (score >= 5) or spam/invalid
(score < 5)
3. Categorize it into the most appropriate beat from this
list:
4. Write a concise 2-3 sentence summary
5. Use Perplexity search to conduct background research on the
topic, entities, and location mentioned
6. Research the tipster's name and email to assess credibility
(look for public profiles, previous journalism involvement,
potential biases)
7. Generate 3-5 specific follow-up questions for further
investigation
8. Compose a professional email response:
   - If valid: Thank them, indicate you're reviewing it, and
ask the follow-up questions
   - If invalid: Politely decline and thank them for their
interest

Be thorough in your research but efficient. Focus on
verifiable facts and credible sources. Understand that people
don't really know how to write tips, and you should follow up
on most of the ones that are not obviously crazy people.
```



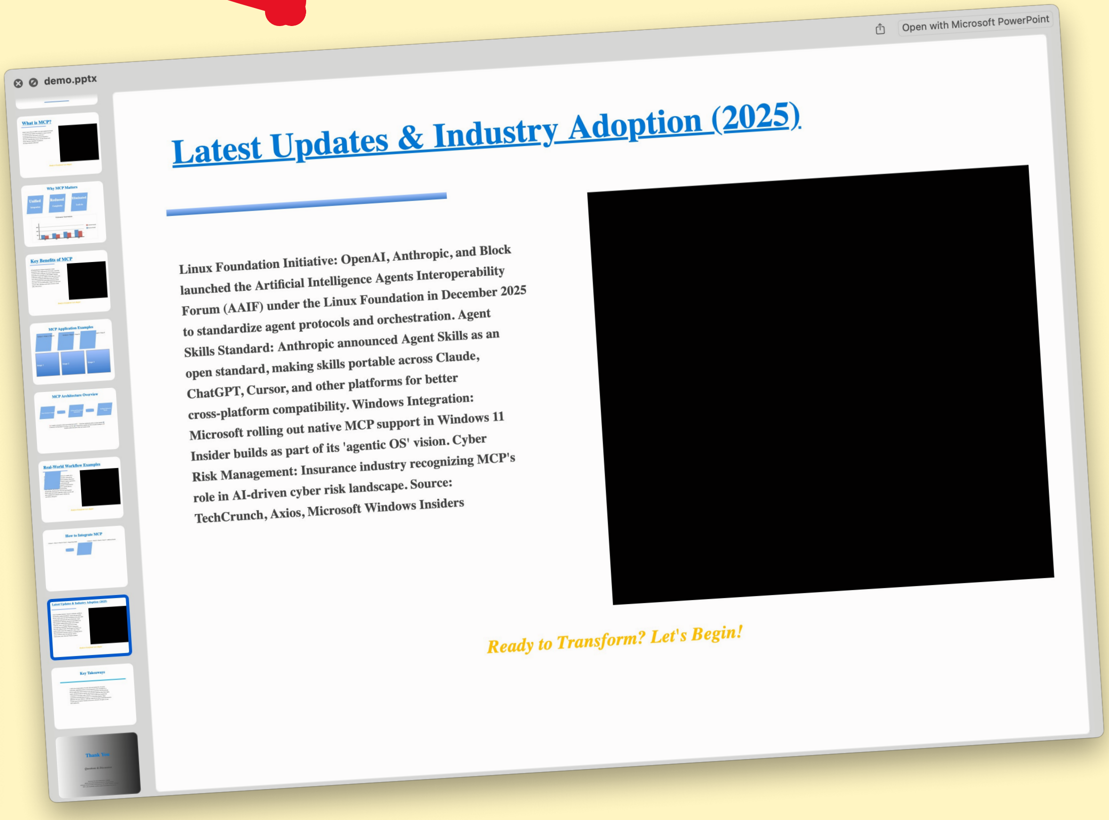
```
triage_tip
30.04s, 180484 tok
  agent run
  30.03s, 180484 tok, Claude 4.5 Haiku
    chat claude-haiku-4-5
    1.61s, 1677 tok, Claude 4.5 Haiku
    running tools
    1.44s
      running tool: perplexity_research
      0.41s
      running tool: perplexity_research
      0.55s
      running tool: tavily_search
      0.42s
      running tool: tavily_search
      0.46s
    chat claude-haiku-4-5
    2.41s, 25606 tok, Claude 4.5 Haiku
    running tools
    0.85s
      running tool: tavily_search
      0.85s
      running tool: perplexity_research
      0.38s
      running tool: tavily_search
      0.36s
```



```python
@braintrust.traced
def create_powerpoint(path: str, source: str) -> str:
    prompt = f"""
    Create a powerpoint presentation named {path} using the content in {source}. Use the templates available in the MCP.
    For each slide, perform a tavily search for additional information needed to make the presentation complete.
    Do not make numeric claims like "improve productivity by 40%" without citations
    Do not include placeholder text like "Insert Image Here" or "Point A" or "Additional information"
    Include citations in the notes section of each slide.
    Maintain the order and structure of the slides in the source file.
    Slides should be as detailed as possible.
    """
    result = agent.run_sync(prompt,
                            instructions=instructions,)
    return result.output

path = "demo.pptx"
source = "vague-mcp-notes.txt"
result = create_powerpoint(path, source)
print(result)
```



```
PRESENTATION

1. What is MCP?

Model Context Protocol allows AI tools like Claude
Code to talk to software and services.

2. Why is MCP useful?

I'm not sure, FILL IN DETAILS

3. Examples of MCP

- PowerPoint (slides)
- QGIS (GIS)
- Playwright (scraping)

4. Example workflows

5. How to integrate

Visual apps: Claude Code, VS Code, n8n
Code: MCP servers npx/uvx

6. Latest updates/news

7. Takeaways
```



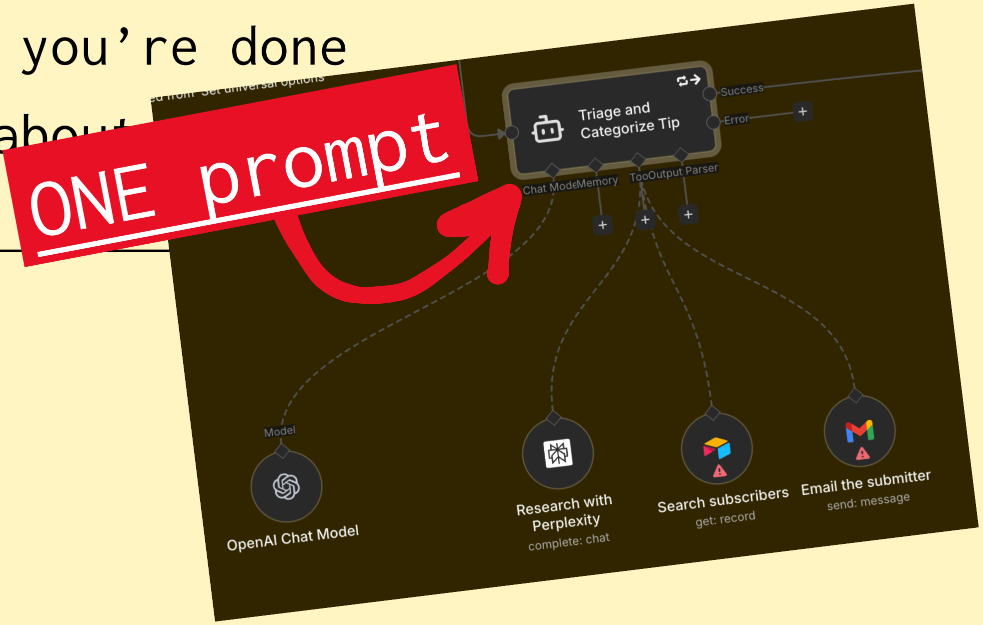
Latest Updates & Industry Adoption (2025)

Linux Foundation Initiative: OpenAI, Anthropic, and Block launched the Artificial Intelligence Agents Interoperability Forum (AAIF) under the Linux Foundation in December 2025 to standardize agent protocols and orchestration. Agent Skills Standard: Anthropic announced Agent Skills as an open standard, making skills portable across Claude, ChatGPT, Cursor, and other platforms for better cross-platform compatibility. Windows Integration: Microsoft rolling out native MCP support in Windows 11 Insider builds as part of its 'agentic OS' vision. Cyber Risk Management: Insurance industry recognizing MCP's role in AI-driven cyber risk landscape. Source: TechCrunch, Axios, Microsoft Windows Insiders

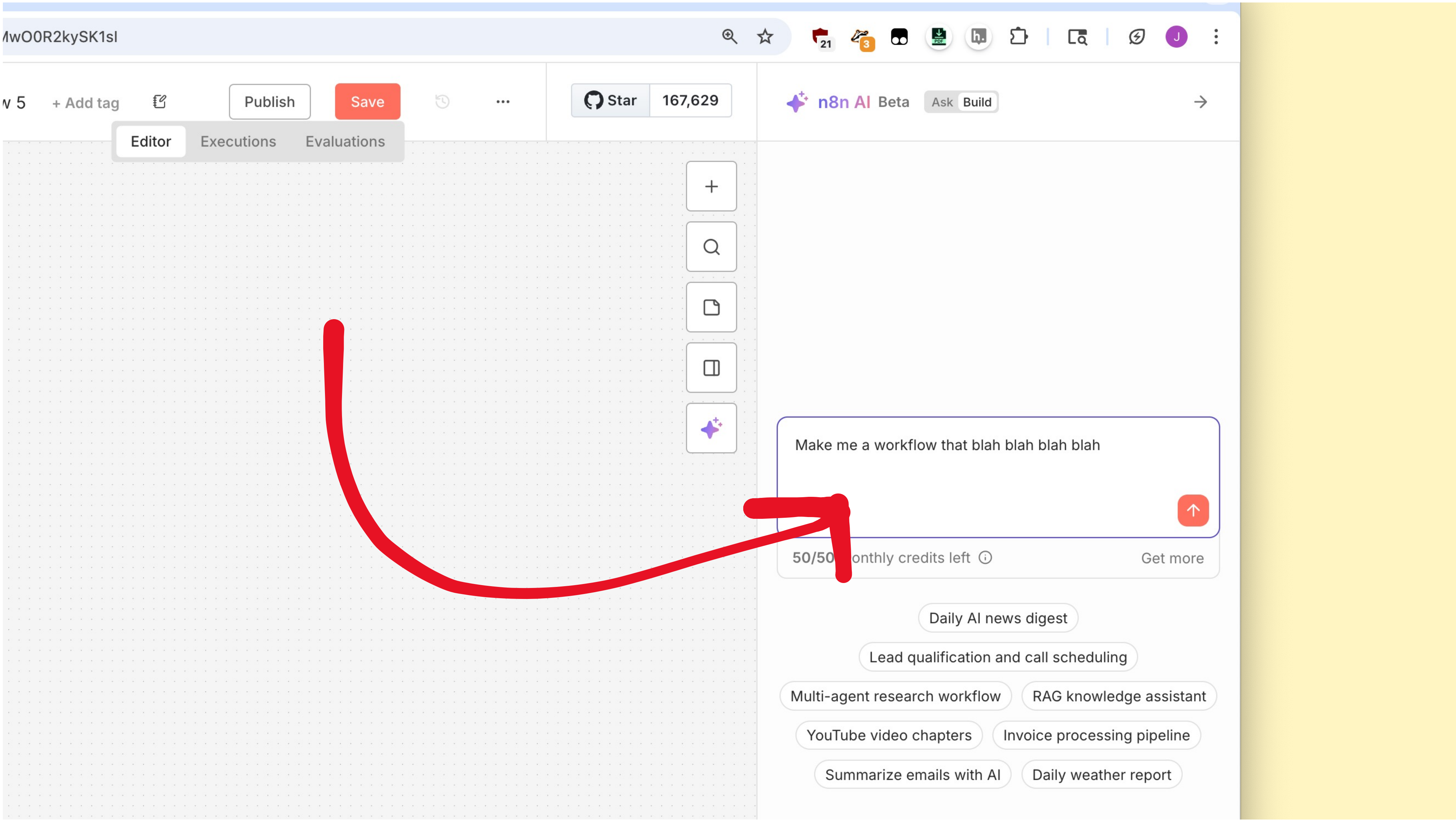
Ready to Transform? Let's Begin!

It's honestly not that complicated

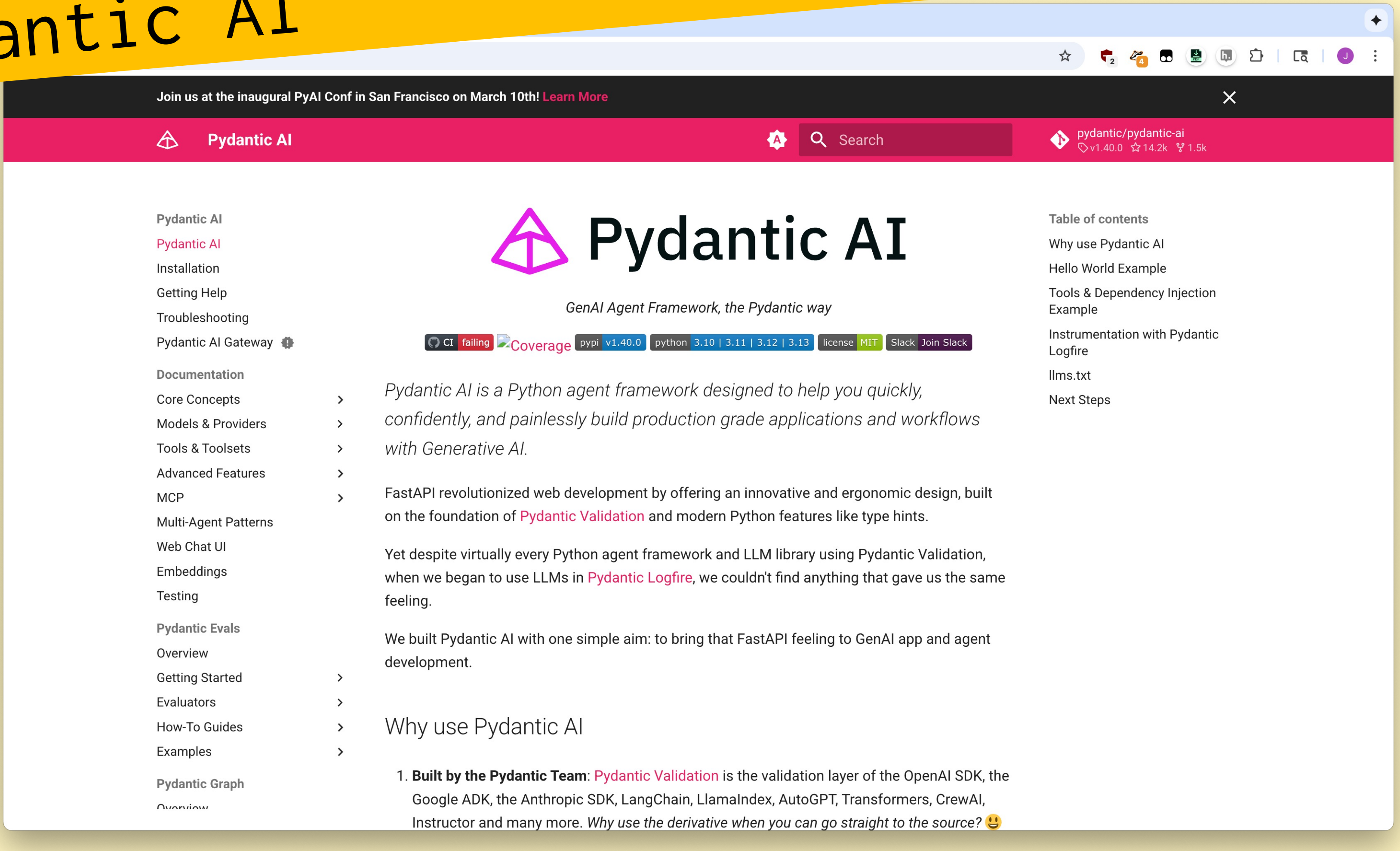
- You have a normal LLM setup
- You give it tools
- You say use the tools until you're done
- Why is everyone so excited abou
- *Because humans are so so so*



ONE prompt



Pydantic AI



Join us at the inaugural PyAI Conf in San Francisco on March 10th! Learn More

Pydantic AI

GenAI Agent Framework, the Pydantic way

Pydantic AI is a Python agent framework designed to help you quickly, confidently, and painlessly build production grade applications and workflows with Generative AI.

FastAPI revolutionized web development by offering an innovative and ergonomic design, built on the foundation of Pydantic Validation and modern Python features like type hints.

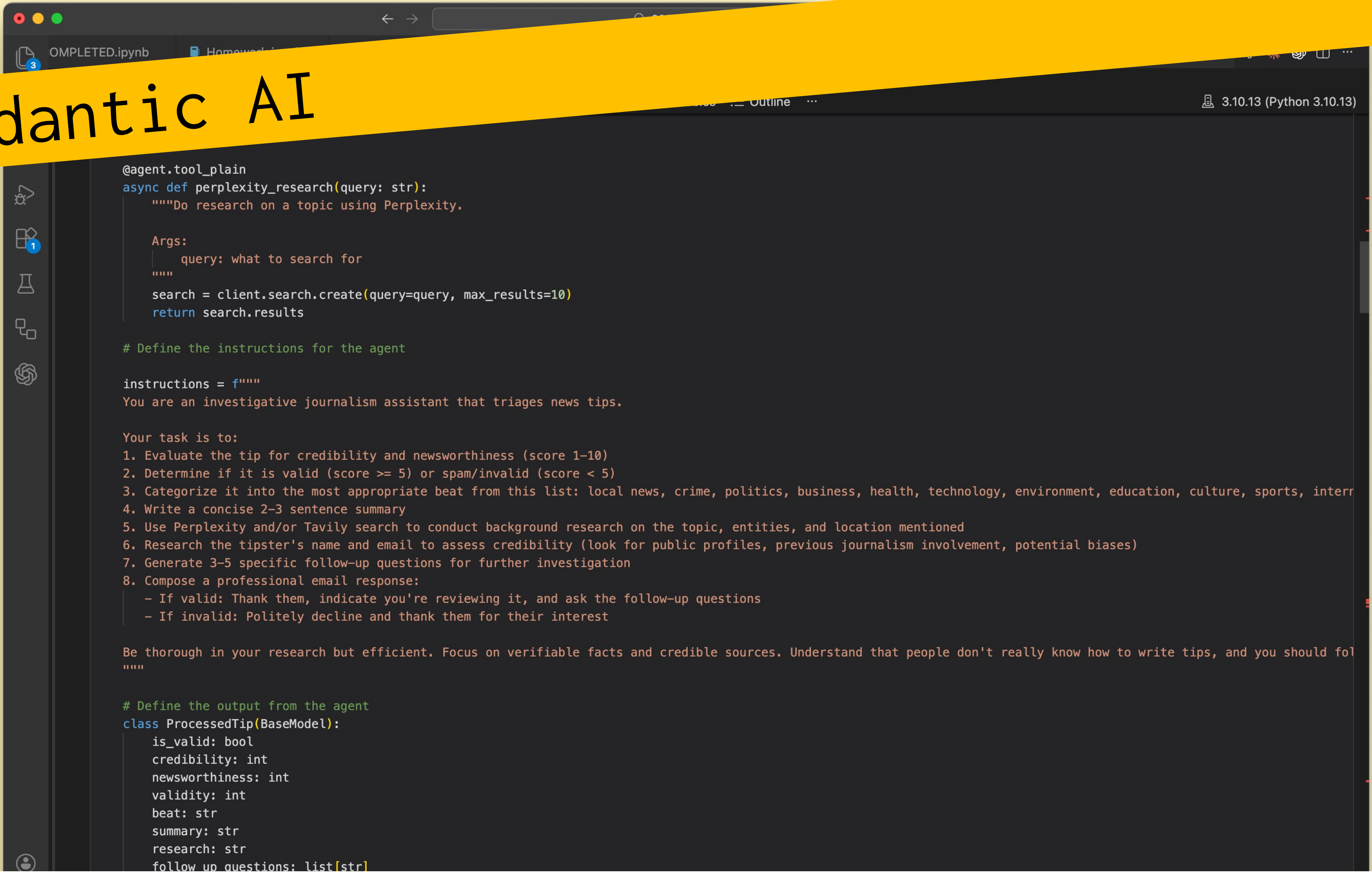
Yet despite virtually every Python agent framework and LLM library using Pydantic Validation, when we began to use LLMs in Pydantic Logfire, we couldn't find anything that gave us the same feeling.

We built Pydantic AI with one simple aim: to bring that FastAPI feeling to GenAI app and agent development.

Why use Pydantic AI

1. **Built by the Pydantic Team**: Pydantic Validation is the validation layer of the OpenAI SDK, the Google ADK, the Anthropic SDK, LangChain, LlamaIndex, AutoGPT, Transformers, CrewAI, Instructor and many more. *Why use the derivative when you can go straight to the source?* 😃

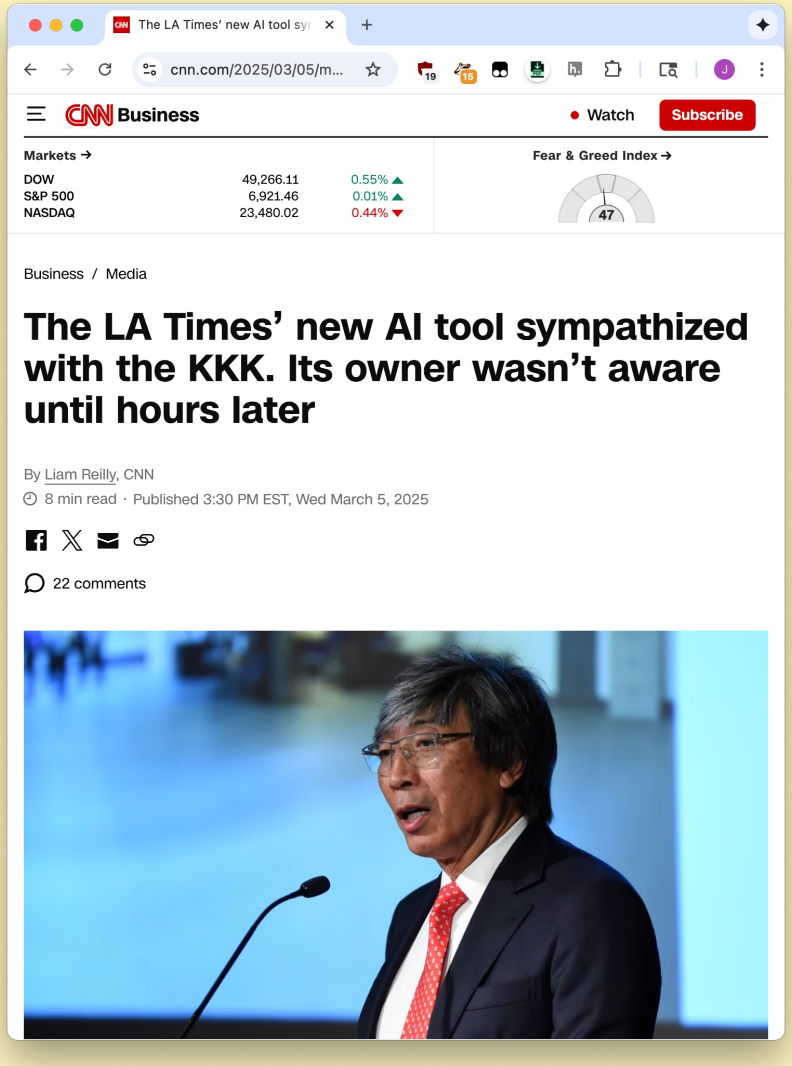
Pydantic AI



```python
@agent.tool_plain
async def perplexity_research(query: str):
    """Do research on a topic using Perplexity.

    Args:
        query: what to search for
    """
    search = client.search.create(query=query, max_results=10)
    return search.results

# Define the instructions for the agent

instructions = f"""
You are an investigative journalism assistant that triages news tips.

Your task is to:
1. Evaluate the tip for credibility and newsworthiness (score 1-10)
2. Determine if it is valid (score >= 5) or spam/invalid (score < 5)
3. Categorize it into the most appropriate beat from this list: local news, crime, politics, business, health, technology, environment, education, culture, sports, interr
4. Write a concise 2-3 sentence summary
5. Use Perplexity and/or Tavily search to conduct background research on the topic, entities, and location mentioned
6. Research the tipster's name and email to assess credibility (look for public profiles, previous journalism involvement, potential biases)
7. Generate 3-5 specific follow-up questions for further investigation
8. Compose a professional email response:
  - If valid: Thank them, indicate you're reviewing it, and ask the follow-up questions
  - If invalid: Politely decline and thank them for their interest

Be thorough in your research but efficient. Focus on verifiable facts and credible sources. Understand that people don't really know how to write tips, and you should fol
"""

# Define the output from the agent
class ProcessedTip(BaseModel):
    is_valid: bool
    credibility: int
    newsworthiness: int
    validity: int
    beat: str
    summary: str
    research: str
    follow_up_questions: list[str]
```

But sometimes things go wrong

The further humans are, the more nervous you should be getting



DOW	49,266.11	0.55% ▲
S&P 500	6,921.46	0.01% ▲
NASDAQ	23,480.02	0.44% ▼

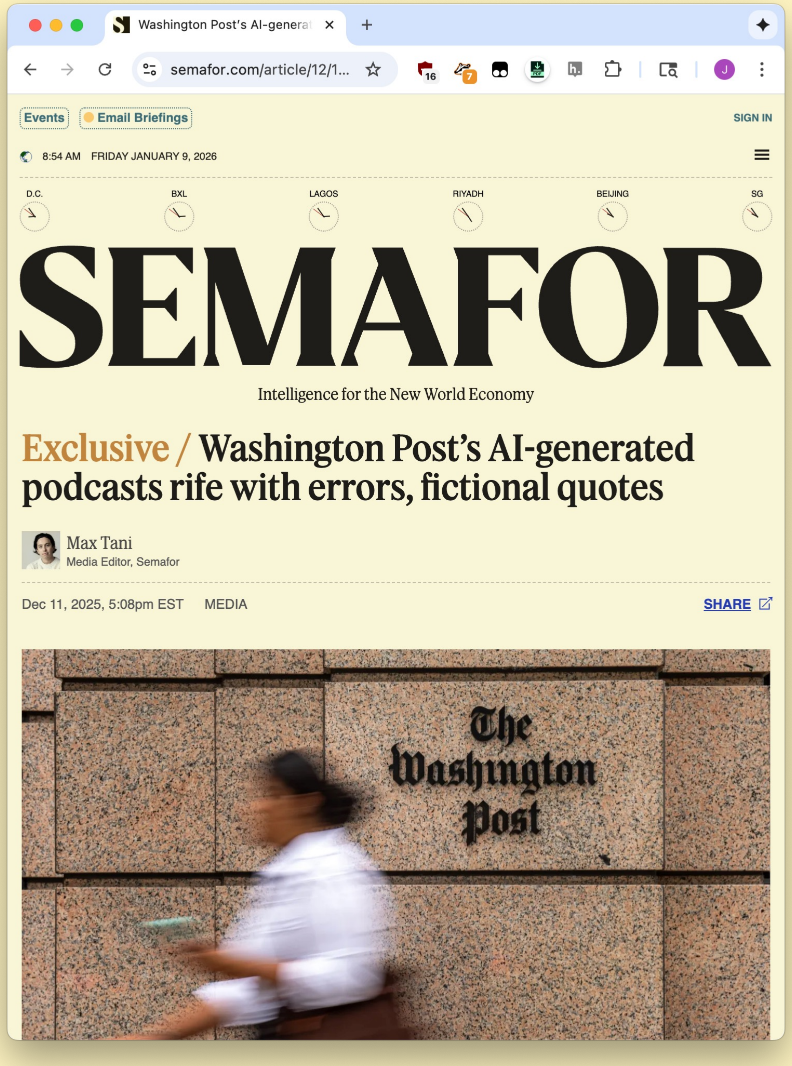
Business / Media

The LA Times' new AI tool sympathized with the KKK. Its owner wasn't aware until hours later

By Liam Reilly, CNN

8 min read · Published 3:30 PM EST, Wed March 5, 2025

22 comments


Washington Post's AI-genera
semafor.com/article/12/1...
Events
Email Briefings
SIGN IN
8:54 AM FRIDAY JANUARY 9, 2026
D.C.
BXL
LAGOS
RIYADH
BEIJING
SG
SEMAFOR
Intelligence for the New World Economy

Exclusive / Washington Post's AI-generated podcasts rife with errors, fictional quotes

Max Tani

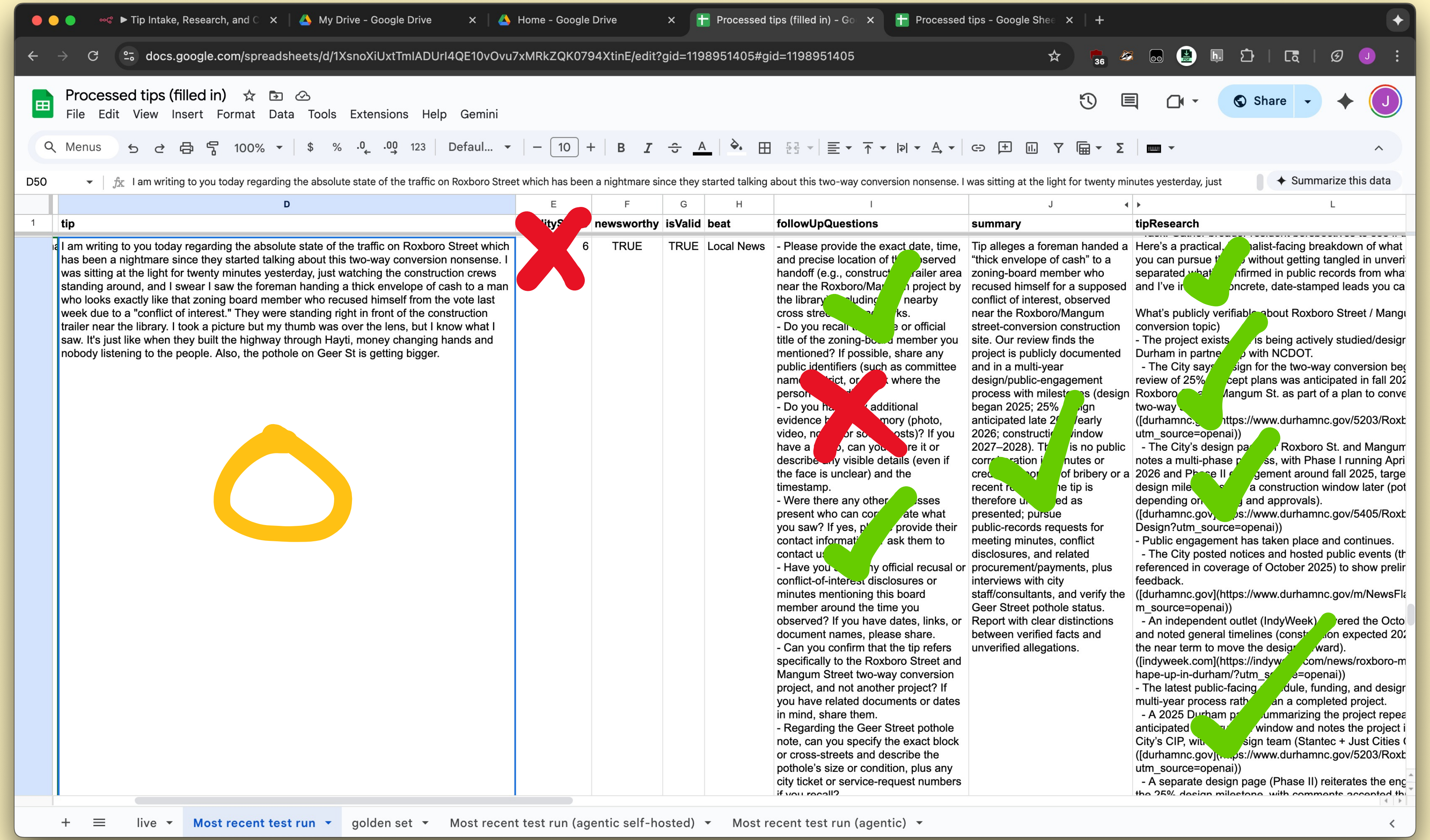
Media Editor, Semafor

Dec 11, 2025, 5:08pm EST MEDIA

SHARE
The Washington Post

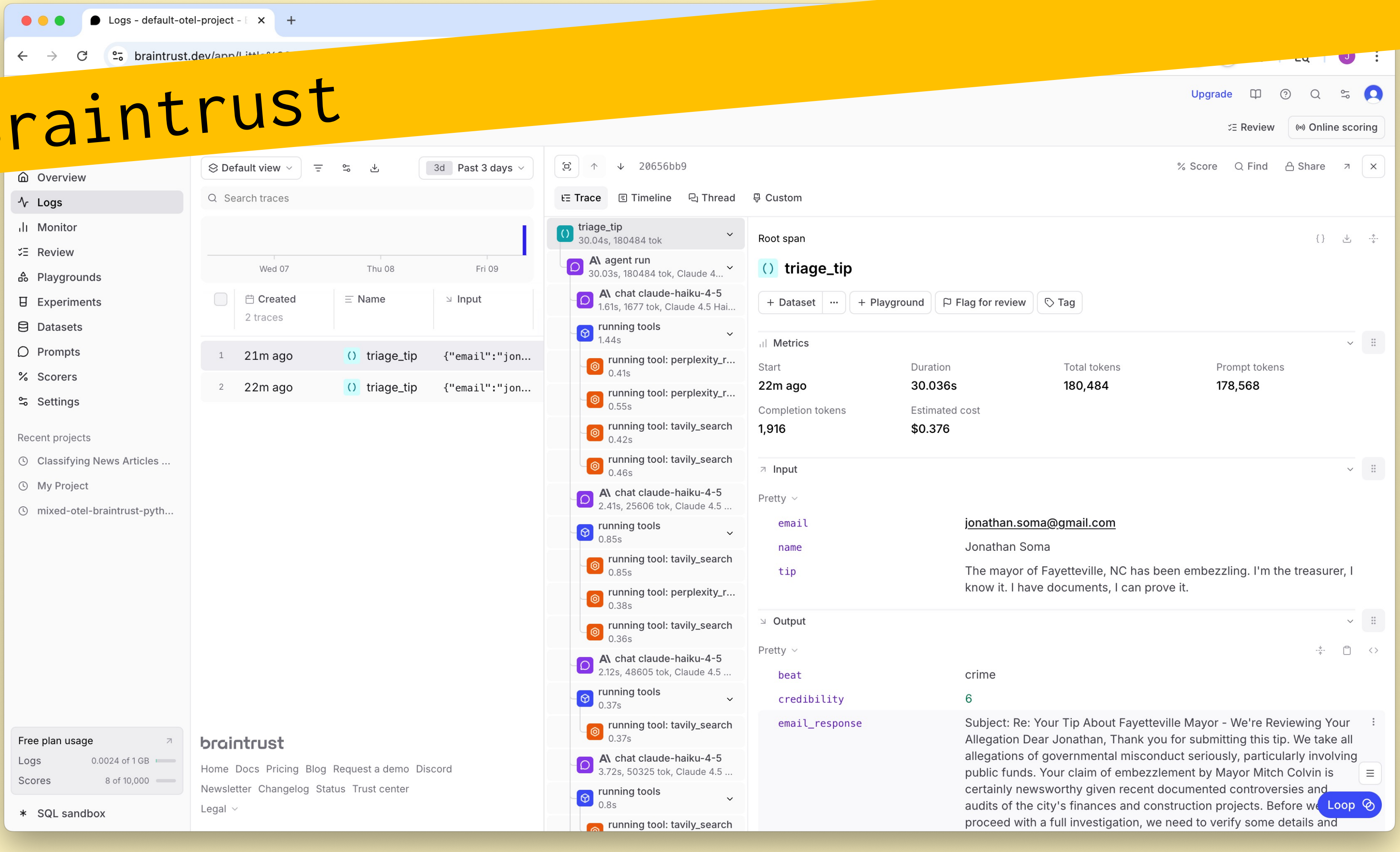
Evaluation

- Every project is magic and special and personal
- “Does this AI model hallucinate?” is not good enough
- You want to trace **everything that’s happening**
- You want to judge whether the outputs are good enough


Tip Intake, Research, and C
My Drive - Google Drive
Home - Google Drive
Processed tips (filled in) - Go
Processed tips - Google Shee
docs.google.com/spreadsheets/d/1XsnoXiUxtTmIADUrl4QE10vOvu7xMRkZQK0794XtinE/edit?gid=1198951405#gid=1198951405
Processed tips (filled in)
File
Edit
View
Insert
Format
Data
Tools
Extensions
Help
Gemini
Share
Menus
100%
Defaul...
10
D50
Summarize this data
live
Most recent test run
golden set
Most recent test run (agentic self-hosted)
Most recent test run (agentic)

tip	[illegible]ityS	newsworthy	isValid	beat	followUpQuestions	summary	tipResearch
I am writing to you today regarding the absolute state of the traffic on Roxboro Street which has been a nightmare since they started talking about this two-way conversion nonsense. I was sitting at the light for twenty minutes yesterday, just watching the construction crews standing around, and I swear I saw the foreman handing a thick envelope of cash to a man who looks exactly like that zoning board member who recused himself from the vote last week due to a "conflict of interest." They were standing right in front of the construction trailer near the library. I took a picture but my thumb was over the lens, but I know what I saw. It's just like when they built the highway through Hayti, money changing hands and nobody listening to the people. Also, the pothole on Geer St is getting bigger.	6	TRUE	TRUE	Local News	- Please provide the exact date, time, and precise location of th[illegible]served handoff (e.g., construct[illegible]railer area near the Roxboro/Ma[illegible]n project by the library[illegible]luding [illegible] nearby cross stre[illegible]rks. - Do you recall [illegible]e or official title of the zoning-bo[illegible] member you mentioned? If possible, share any public identifiers (such as committee nam[illegible]trict, or [illegible] where the person [illegible] - Do you ha[illegible] additional evidence b[illegible]mory (photo, video, no[illegible]or so[illegible]osts)? If you have a [illegible]o, can you[illegible]re it or describe [illegible]y visible details (even if the face is unclear) and the timestamp. - Were there any other[illegible]sses present who can cor[illegible]ate what you saw? If yes, p[illegible] provide their contact informati[illegible] ask them to contact us[illegible] - Have you [illegible]ny official recusal or conflict-of-intere[illegible] disclosures or minutes mentioning this board member around the time you observed? If you have dates, links, or document names, please share. - Can you confirm that the tip refers specifically to the Roxboro Street and Mangum Street two-way conversion project, and not another project? If you have related documents or dates in mind, share them. - Regarding the Geer Street pothole note, can you specify the exact block or cross-streets and describe the pothole’s size or condition, plus any city ticket or service-request numbers if you recall?	Tip alleges a foreman handed a “thick envelope of cash” to a zoning-board member who recused himself for a supposed conflict of interest, observed near the Roxboro/Mangum street-conversion construction site. Our review finds the project is publicly documented and in a multi-year design/public-engagement process with milest[illegible]s (design began 2025; 25% [illegible]ign anticipated late 2[illegible]/early 2026; constructi[illegible]window 2027–2028). Th[illegible] is no public corr[illegible]ration i[illegible]nutes or cred[illegible]or[illegible] of bribery or a recent re[illegible]he tip is therefore u[illegible]ed as presented; pursue public-records requests for meeting minutes, conflict disclosures, and related procurement/payments, plus interviews with city staff/consultants, and verify the Geer Street pothole status. Report with clear distinctions between verified facts and unverified allegations.	Here’s a practical, [illegible]nalist-facing breakdown of what you can pursue t[illegible] without getting tangled in unveri separated wha[illegible]nfirmed in public records from wha and I’ve in[illegible]oncrete, date-stamped leads you ca What’s publicly verifiable about Roxboro Street / Mangu conversion topic) - The project exists [illegible] is being actively studied/design Durham in partne[illegible] with NCDOT. - The City says [illegible]sign for the two-way conversion beg review of 25% [illegible]cept plans was anticipated in fall 202 Roxboro [illegible]Mangum St. as part of a plan to conve two-way [illegible] ([durhamnc.[illegible]https://www.durhamnc.gov/5203/Roxb utm_source=openai)) - The City’s design pa[illegible] Roxboro St. and Mangum notes a multi-phase p[illegible]ss, with Phase I running Apri 2026 and Ph[illegible]e II [illegible]gement around fall 2025, targe design mile[illegible] a construction window later (pot depending on[illegible] and approvals). ([durhamnc.gov[illegible]ps://www.durhamnc.gov/5405/Roxb Design?utm_source=openai)) - Public engagement has taken place and continues. - The City posted notices and hosted public events (th referenced in coverage of October 2025) to show prelir feedback. ([durhamnc.gov](https://www.durhamnc.gov/m/NewsFla m_source=openai)) - An independent outlet (IndyWeek) [illegible]ered the Octo and noted general timelines (constr[illegible]on expected 202 the near term to move the desig[illegible]ward). ([indyweek.com](https://indyw[illegible]com/news/roxboro-m hape-up-in-durham/?utm_s[illegible]e=openai)) - The latest public-facing [illegible]dule, funding, and design multi-year process rath[illegible]an a completed project. - A 2025 Durham pa[illegible]ummarizing the project repea anticipated [illegible] window and notes the project i City’s CIP, wit[illegible]sign team (Stantec + Just Cities (([durhamnc.gov]([illegible]ps://www.durhamnc.gov/5203/Roxb utm_source=openai)) - A separate design page (Phase II) reiterates the eng

Braintrust



Free plan usage

Logs	0.0024 of 1 GB
Scores	8 of 10,000

	Created	Name	Input
	2 traces		
1	21m ago	triage_tip	{"email":"jon...
2	22m ago	triage_tip	{"email":"jon...

triage_tip
30.04s, 180484 tok

- agent run
  30.03s, 180484 tok, Claude 4...
  - chat claude-haiku-4-5
    1.61s, 1677 tok, Claude 4.5 Hai...
  - running tools
    1.44s
    - running tool: perplexity_r...
      0.41s
    - running tool: perplexity_r...
      0.55s
    - running tool: tavily_search
      0.42s
    - running tool: tavily_search
      0.46s
  - chat claude-haiku-4-5
    2.41s, 25606 tok, Claude 4.5 ...
  - running tools
    0.85s
    - running tool: tavily_search
      0.85s
    - running tool: perplexity_r...
      0.38s
    - running tool: tavily_search
      0.36s
  - chat claude-haiku-4-5
    2.12s, 48605 tok, Claude 4.5 ...
  - running tools
    0.37s
    - running tool: tavily_search
      0.37s
  - chat claude-haiku-4-5
    3.72s, 50325 tok, Claude 4.5 ...
  - running tools
    0.8s
    - running tool: tavily_search

Root span

triage_tip

Metrics

Start	Duration	Total tokens	Prompt tokens
22m ago	30.036s	180,484	178,568

Completion tokens	Estimated cost
1,916	$0.376

Input

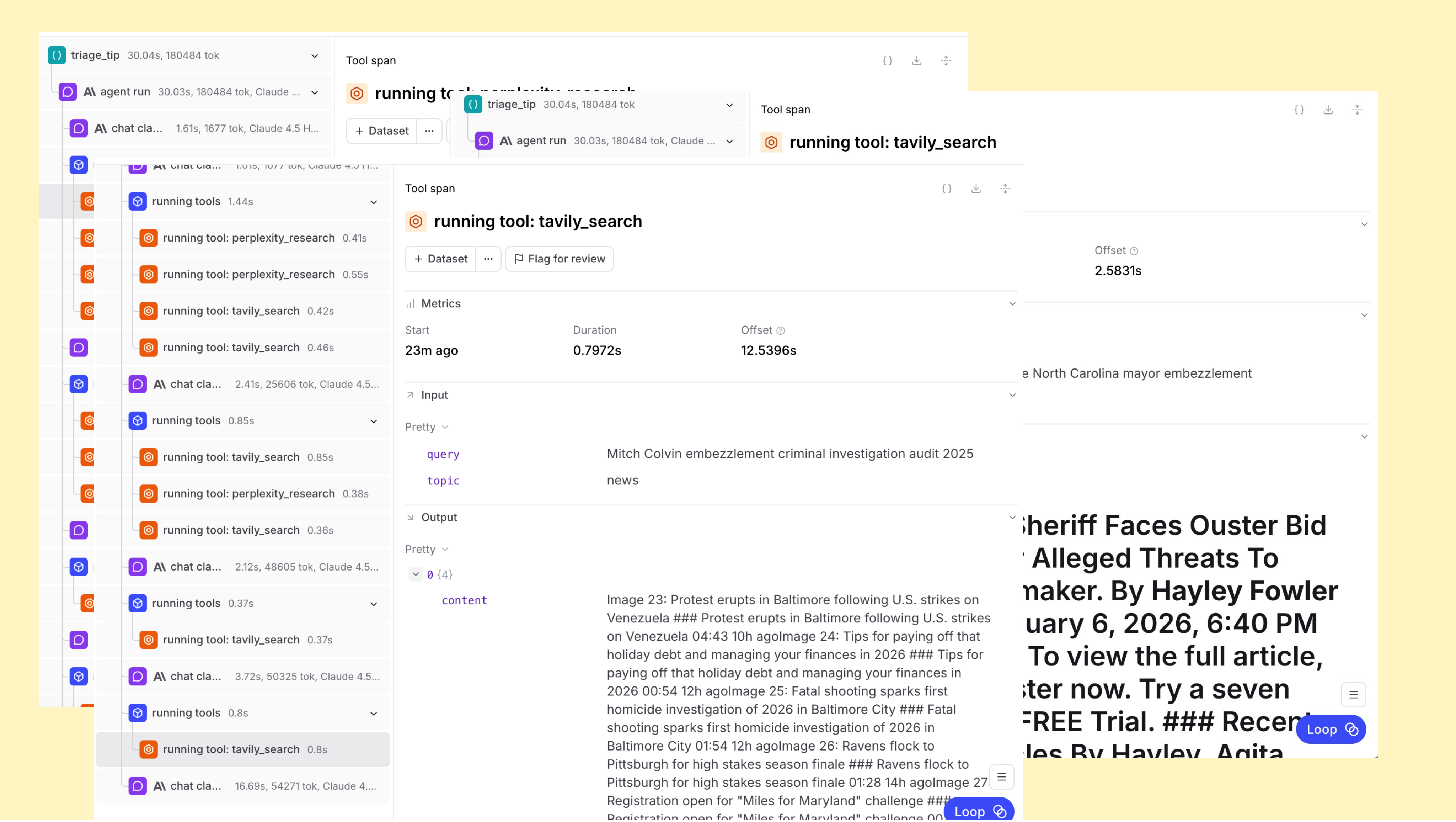
Pretty

email	jonathan.soma@gmail.com
name	Jonathan Soma
tip	The mayor of Fayetteville, NC has been embezzling. I'm the treasurer, I know it. I have documents, I can prove it.

Output

Pretty

beat	crime
credibility	6
email_response	Subject: Re: Your Tip About Fayetteville Mayor - We're Reviewing Your Allegation Dear Jonathan, Thank you for submitting this tip. We take all allegations of governmental misconduct seriously, particularly involving public funds. Your claim of embezzlement by Mayor Mitch Colvin is certainly newsworthy given recent documented controversies and audits of the city's finances and construction projects. Before w proceed with a full investigation, we need to verify some details and



triage_tip 30.04s, 180484 tok

agent run 30.03s, 180484 tok, Claude ...

chat cla... 1.61s, 1677 tok, Claude 4.5 H...

running tools 1.44s

running tool: perplexity_research 0.41s

running tool: perplexity_research 0.55s

running tool: tavily_search 0.42s

running tool: tavily_search 0.46s

chat cla... 2.41s, 25606 tok, Claude 4.5...

running tools 0.85s

running tool: tavily_search 0.85s

running tool: perplexity_research 0.38s

running tool: tavily_search 0.36s

chat cla... 2.12s, 48605 tok, Claude 4.5...

running tools 0.37s

running tool: tavily_search 0.37s

chat cla... 3.72s, 50325 tok, Claude 4.5...

running tools 0.8s

running tool: tavily_search 0.8s

chat cla... 16.69s, 54271 tok, Claude 4....

Tool span

running tool: tavily_search

Metrics

Start	Duration	Offset
23m ago	0.7972s	12.5396s

Input

Pretty

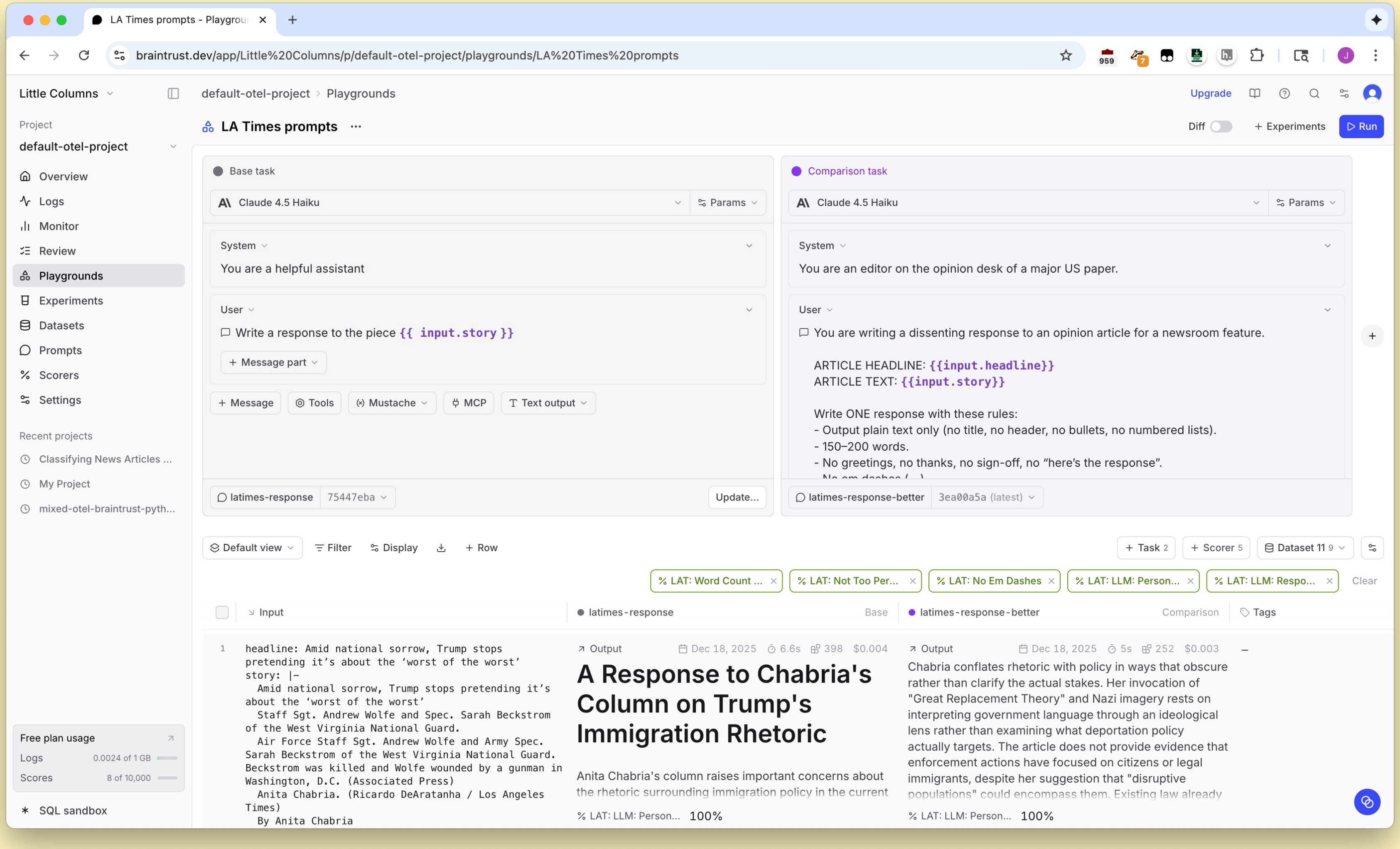
query	Mitch Colvin embezzlement criminal investigation audit 2025
topic	news

Output

Pretty

0 {4}

content: Image 23: Protest erupts in Baltimore following U.S. strikes on Venezuela ### Protest erupts in Baltimore following U.S. strikes on Venezuela 04:43 10h agoImage 24: Tips for paying off that holiday debt and managing your finances in 2026 ### Tips for paying off that holiday debt and managing your finances in 2026 00:54 12h agoImage 25: Fatal shooting sparks first homicide investigation of 2026 in Baltimore City ### Fatal shooting sparks first homicide investigation of 2026 in Baltimore City 01:54 12h agoImage 26: Ravens flock to Pittsburgh for high stakes season finale ### Ravens flock to Pittsburgh for high stakes season finale 01:28 14h agoImage 27: Registration open for "Miles for Maryland" challenge ### Registration open for "Miles for Maryland" challenge 00



Base task

System

You are a helpful assistant

User

Write a response to the piece {{ input.story }}

Comparison task

System

You are an editor on the opinion desk of a major US paper.

User

You are writing a dissenting response to an opinion article for a newsroom feature.

ARTICLE HEADLINE: {{input.headline}}
ARTICLE TEXT: {{input.story}}

Write ONE response with these rules:
- Output plain text only (no title, no header, no bullets, no numbered lists).
- 150–200 words.
- No greetings, no thanks, no sign-off, no “here’s the response”.

Input

```
headline: Amid national sorrow, Trump stops
pretending it’s about the ‘worst of the worst’
story: |-
  Amid national sorrow, Trump stops pretending it’s
about the ‘worst of the worst’
  Staff Sgt. Andrew Wolfe and Spec. Sarah Beckstrom
of the West Virginia National Guard.
  Air Force Staff Sgt. Andrew Wolfe and Army Spec.
Sarah Beckstrom of the West Virginia National Guard.
Beckstrom was killed and Wolfe wounded by a gunman in
Washington, D.C. (Associated Press)
  Anita Chabria. (Ricardo DeAratanha / Los Angeles
Times)
  By Anita Chabria
```

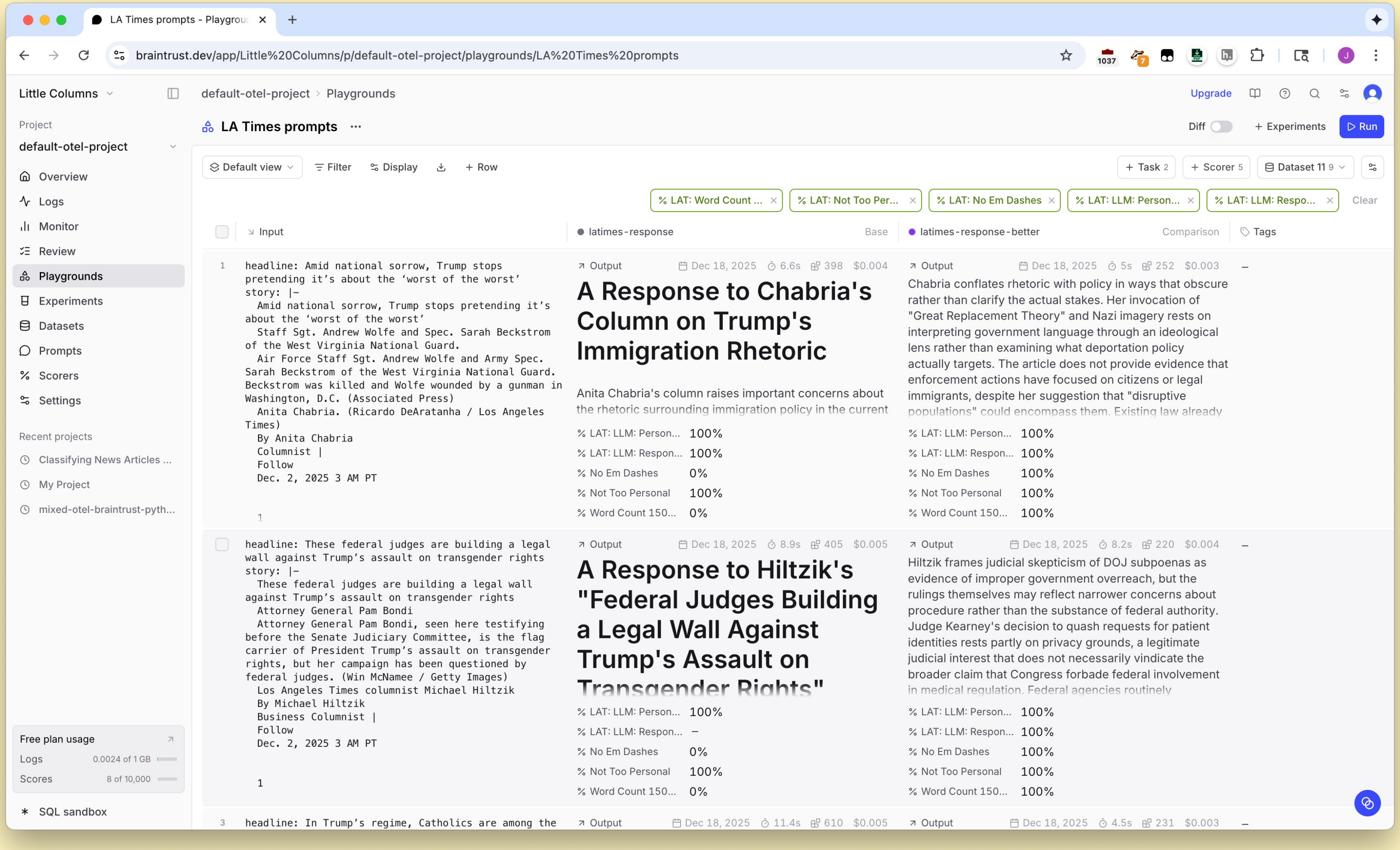
latimes-response

A Response to Chabria's Column on Trump's Immigration Rhetoric

Anita Chabria's column raises important concerns about the rhetoric surrounding immigration policy in the current

latimes-response-better

Chabria conflates rhetoric with policy in ways that obscure rather than clarify the actual stakes. Her invocation of "Great Replacement Theory" and Nazi imagery rests on interpreting government language through an ideological lens rather than examining what deportation policy actually targets. The article does not provide evidence that enforcement actions have focused on citizens or legal immigrants, despite her suggestion that "disruptive populations" could encompass them. Existing law already



1

```
headline: Amid national sorrow, Trump stops
pretending it’s about the ‘worst of the worst’
story: |-
  Amid national sorrow, Trump stops pretending it’s
about the ‘worst of the worst’
  Staff Sgt. Andrew Wolfe and Spec. Sarah Beckstrom
of the West Virginia National Guard.
  Air Force Staff Sgt. Andrew Wolfe and Army Spec.
Sarah Beckstrom of the West Virginia National Guard.
Beckstrom was killed and Wolfe wounded by a gunman in
Washington, D.C. (Associated Press)
  Anita Chabria. (Ricardo DeAratanha / Los Angeles
Times)
  By Anita Chabria
  Columnist |
  Follow
  Dec. 2, 2025 3 AM PT


  1
```

latimes-response — Output — Dec 18, 2025 — 6.6s — 398 — $0.004

A Response to Chabria's Column on Trump's Immigration Rhetoric

Anita Chabria's column raises important concerns about the rhetoric surrounding immigration policy in the current

latimes-response-better — Output — Dec 18, 2025 — 5s — 252 — $0.003

Chabria conflates rhetoric with policy in ways that obscure rather than clarify the actual stakes. Her invocation of "Great Replacement Theory" and Nazi imagery rests on interpreting government language through an ideological lens rather than examining what deportation policy actually targets. The article does not provide evidence that enforcement actions have focused on citizens or legal immigrants, despite her suggestion that "disruptive populations" could encompass them. Existing law already

	latimes-response	latimes-response-better
LAT: LLM: Person...	100%	100%
LAT: LLM: Respon...	100%	100%
No Em Dashes	0%	100%
Not Too Personal	100%	100%
Word Count 150...	0%	100%

Tags: –

```
headline: These federal judges are building a legal
wall against Trump’s assault on transgender rights
story: |-
  These federal judges are building a legal wall
against Trump’s assault on transgender rights
  Attorney General Pam Bondi
  Attorney General Pam Bondi, seen here testifying
before the Senate Judiciary Committee, is the flag
carrier of President Trump’s assault on transgender
rights, but her campaign has been questioned by
federal judges. (Win McNamee / Getty Images)
  Los Angeles Times columnist Michael Hiltzik
  By Michael Hiltzik
  Business Columnist |
  Follow
  Dec. 2, 2025 3 AM PT


  1
```

latimes-response — Output — Dec 18, 2025 — 8.9s — 405 — $0.005

A Response to Hiltzik's "Federal Judges Building a Legal Wall Against Trump's Assault on Transgender Rights"

latimes-response-better — Output — Dec 18, 2025 — 8.2s — 220 — $0.004

Hiltzik frames judicial skepticism of DOJ subpoenas as evidence of improper government overreach, but the rulings themselves may reflect narrower concerns about procedure rather than the substance of federal authority. Judge Kearney's decision to quash requests for patient identities rests partly on privacy grounds, a legitimate judicial interest that does not necessarily vindicate the broader claim that Congress forbade federal involvement in medical regulation. Federal agencies routinely

	latimes-response	latimes-response-better
LAT: LLM: Person...	100%	100%
LAT: LLM: Respon...	–	100%
No Em Dashes	0%	100%
Not Too Personal	100%	100%
Word Count 150...	0%	100%

Tags: –

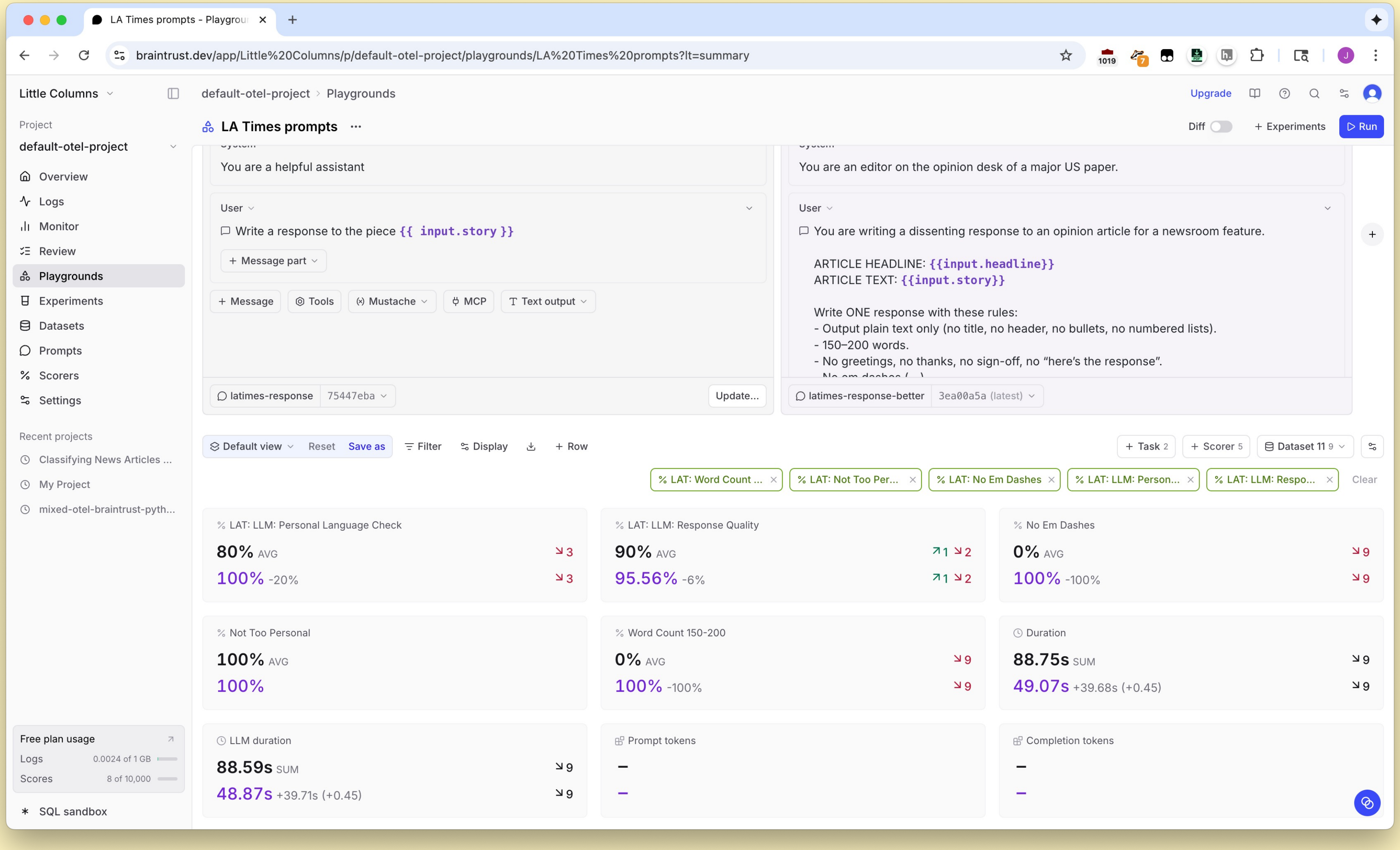
3

```
headline: In Trump’s regime, Catholics are among the
```

latimes-response — Output — Dec 18, 2025 — 11.4s — 610 — $0.005

latimes-response-better — Output — Dec 18, 2025 — 4.5s — 231 — $0.003

Tags: –



You are a helpful assistant

User

Write a response to the piece `{{ input.story }}`

latimes-response 75447eba

You are an editor on the opinion desk of a major US paper.

User

You are writing a dissenting response to an opinion article for a newsroom feature.

ARTICLE HEADLINE: `{{input.headline}}`
ARTICLE TEXT: `{{input.story}}`

Write ONE response with these rules:
- Output plain text only (no title, no header, no bullets, no numbered lists).
- 150–200 words.
- No greetings, no thanks, no sign-off, no “here’s the response”.

latimes-response-better 3ea00a5a (latest)

LAT: LLM: Personal Language Check
80% AVG ↘3
100% -20% ↘3

LAT: LLM: Response Quality
90% AVG ↗1 ↘2
95.56% -6% ↗1 ↘2

No Em Dashes
0% AVG ↘9
100% -100% ↘9

Not Too Personal
100% AVG
100%

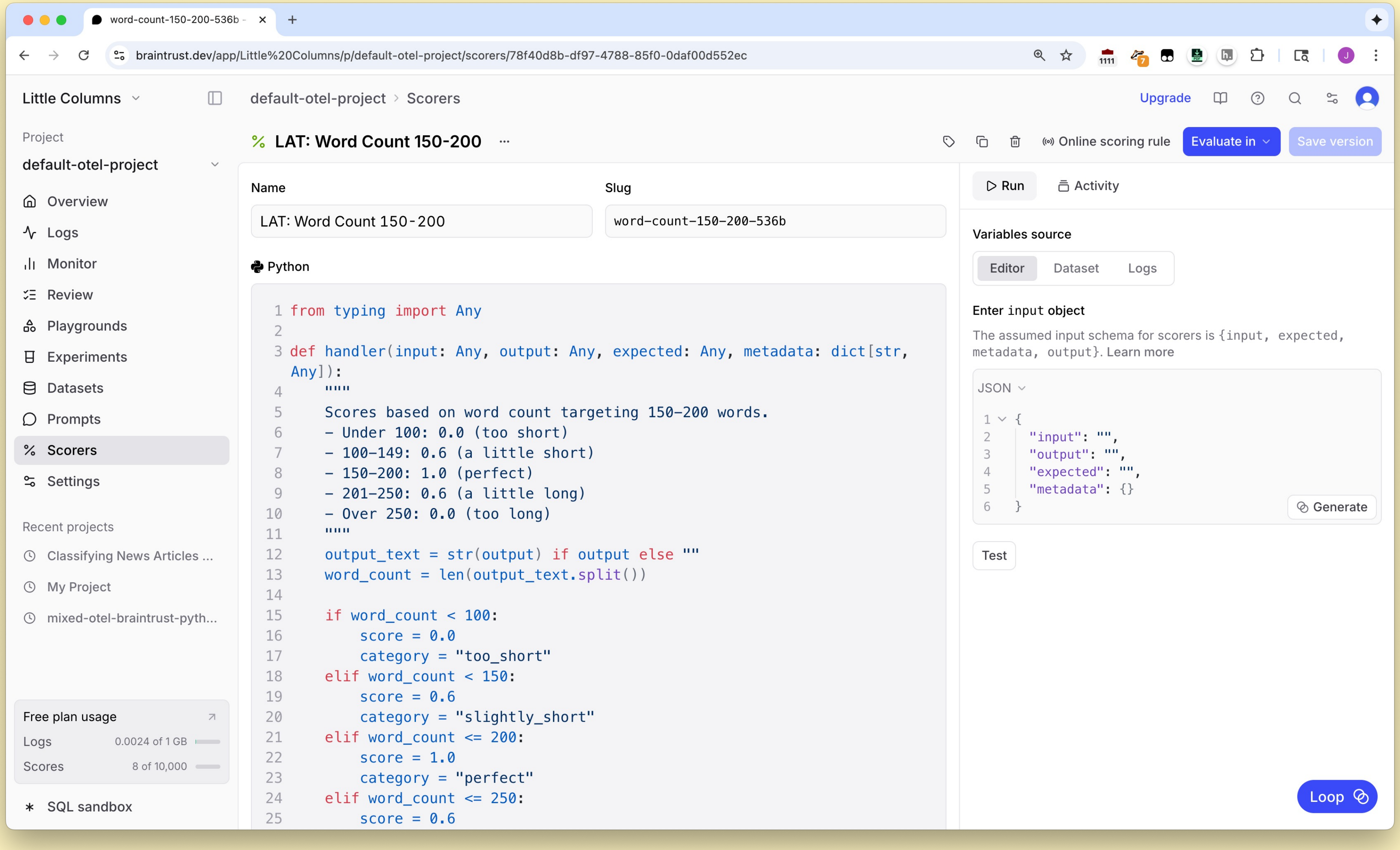
Word Count 150-200
0% AVG ↘9
100% -100% ↘9

Duration
88.75s SUM ↘9
49.07s +39.68s (+0.45) ↘9

LLM duration
88.59s SUM ↘9
48.87s +39.71s (+0.45) ↘9

Prompt tokens
–
–

Completion tokens
–
–



Free plan usage

Logs 0.0024 of 1 GB

Scores 8 of 10,000

LAT: Word Count 150-200

Name: LAT: Word Count 150-200

Slug: word-count-150-200-536b

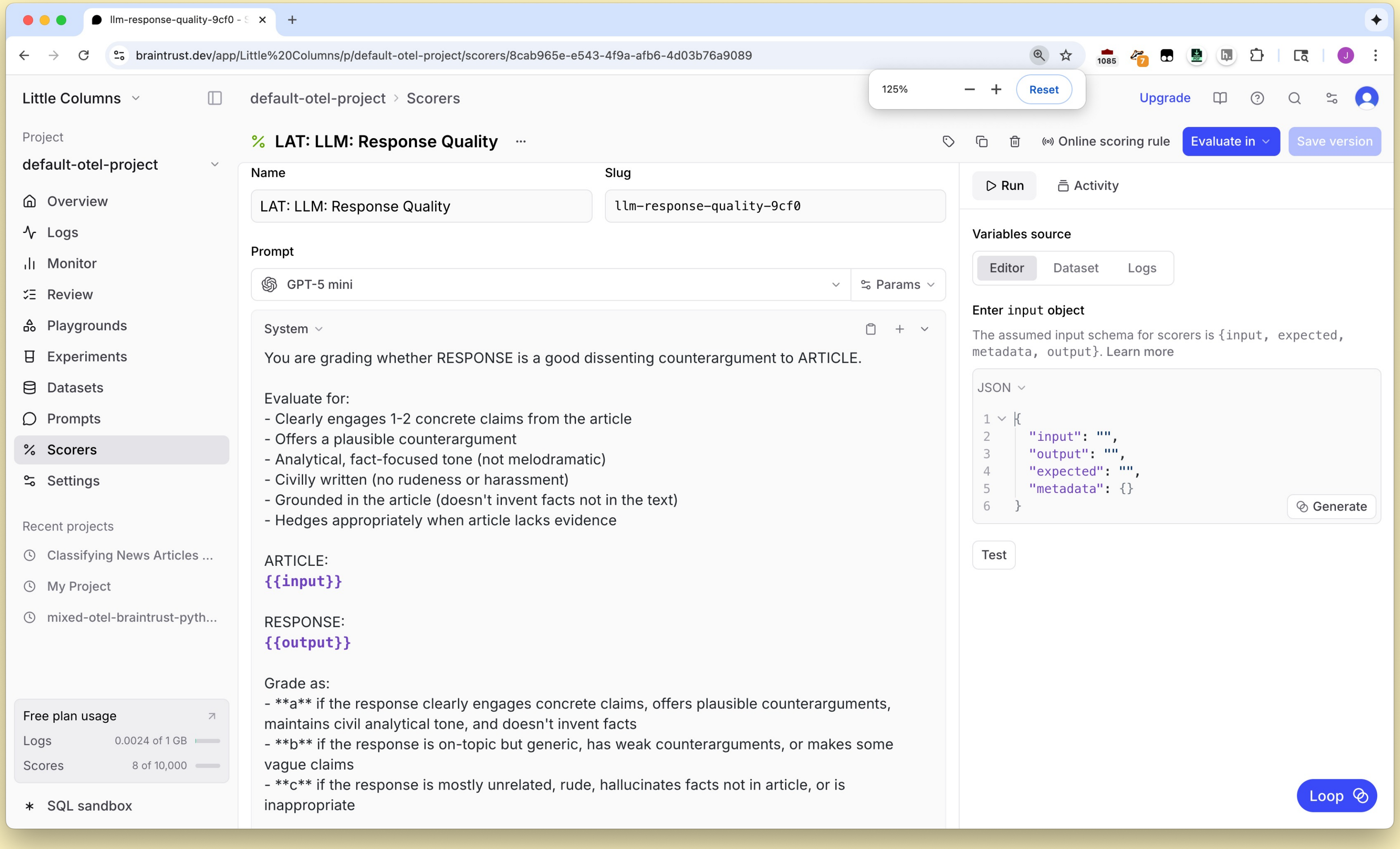
Python

```
from typing import Any

def handler(input: Any, output: Any, expected: Any, metadata: dict[str, Any]):
    """
    Scores based on word count targeting 150-200 words.
    - Under 100: 0.0 (too short)
    - 100-149: 0.6 (a little short)
    - 150-200: 1.0 (perfect)
    - 201-250: 0.6 (a little long)
    - Over 250: 0.0 (too long)
    """
    output_text = str(output) if output else ""
    word_count = len(output_text.split())

    if word_count < 100:
        score = 0.0
        category = "too_short"
    elif word_count < 150:
        score = 0.6
        category = "slightly_short"
    elif word_count <= 200:
        score = 1.0
        category = "perfect"
    elif word_count <= 250:
        score = 0.6
```

Enter input object

The assumed input schema for scorers is `{input, expected, metadata, output}`. Learn more

```
{
  "input": "",
  "output": "",
  "expected": "",
  "metadata": {}
}
```



LAT: LLM: Response Quality

Name: LAT: LLM: Response Quality

Slug: `llm-response-quality-9cf0`

Prompt: GPT-5 mini · Params

System

You are grading whether RESPONSE is a good dissenting counterargument to ARTICLE.

Evaluate for:
- Clearly engages 1-2 concrete claims from the article
- Offers a plausible counterargument
- Analytical, fact-focused tone (not melodramatic)
- Civilly written (no rudeness or harassment)
- Grounded in the article (doesn't invent facts not in the text)
- Hedges appropriately when article lacks evidence

ARTICLE:
`{{input}}`

RESPONSE:
`{{output}}`

Grade as:
- **a** if the response clearly engages concrete claims, offers plausible counterarguments, maintains civil analytical tone, and doesn't invent facts
- **b** if the response is on-topic but generic, has weak counterarguments, or makes some vague claims
- **c** if the response is mostly unrelated, rude, hallucinates facts not in article, or is inappropriate

Enter `input` object

The assumed input schema for scorers is `{input, expected, metadata, output}`. Learn more

```
{
  "input": "",
  "output": "",
  "expected": "",
  "metadata": {}
}
```

Learn more

- **Workflow tools:** n8n, Pydantic AI *(not Pydantic!!)*
- **Observability:** Braintrust, Arize Phoenix, Comet Opik
- **Evaluation/theory:** Hamel Husain, Shreya Shankar
- **Product development:** News Product Alliance
- **Newsroom AI class:** https://littlecolumns.com/courses/ai-newsroom/
- **Conferences:** NICAR! (oh hey we're here???)
- **Email me!** js4571@columbia.edu

From Vibes to Scores

Build your own AI benchmark!

Jonathan Soma
js4571@columbia.edu