Newsroom Infrastructure for AI Experimentation

Jonathan Soma
Columbia University
@dangerscarf・js4571@columbia.edu

https://bit.ly/dh26-infra

What we'll be looking at

- Python tools
  - Gradio
  - Streamlit
  - Ngrok/Cloudflare Tunnel/other tunnels (but not!!)
- Platforms
  - Prompt engineering and tests: Braintrust
  - *Doing it all themselves: n8n???*
- Staying flexible and model/vendor agnostic
  - Pydantic
  - LiteLLM
  - OpenRouter

What we'll be looking at

- ~~Python tools~~ Inviting feedback early
  - Gradio
  - Streamlit
  - Ngrok/Cloudflare Tunnel/other tunnels (but not!!)
- ~~Platforms~~ Embracing domain experts
  - Prompt engineering and tests: Braintrust
  - *Doing it all themselves: n8n???*
- Staying flexible and model/vendor agnostic
  - Pydantic
  - LiteLLM
  - OpenRouter

Inviting feedback early

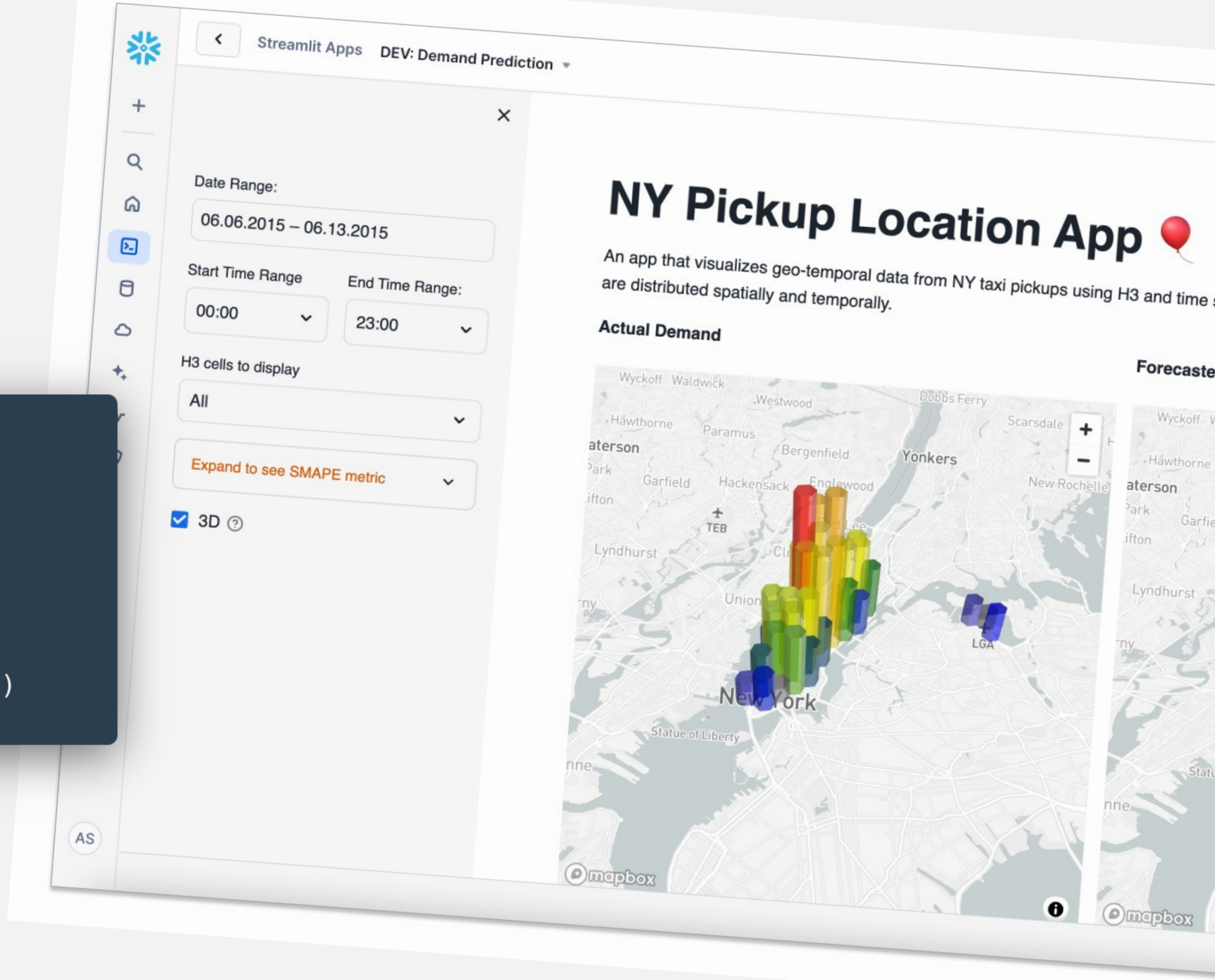
- You need something **shareable**
- Don’t make other people run your code!
- **The best approaches have maybe changed since early 2026**

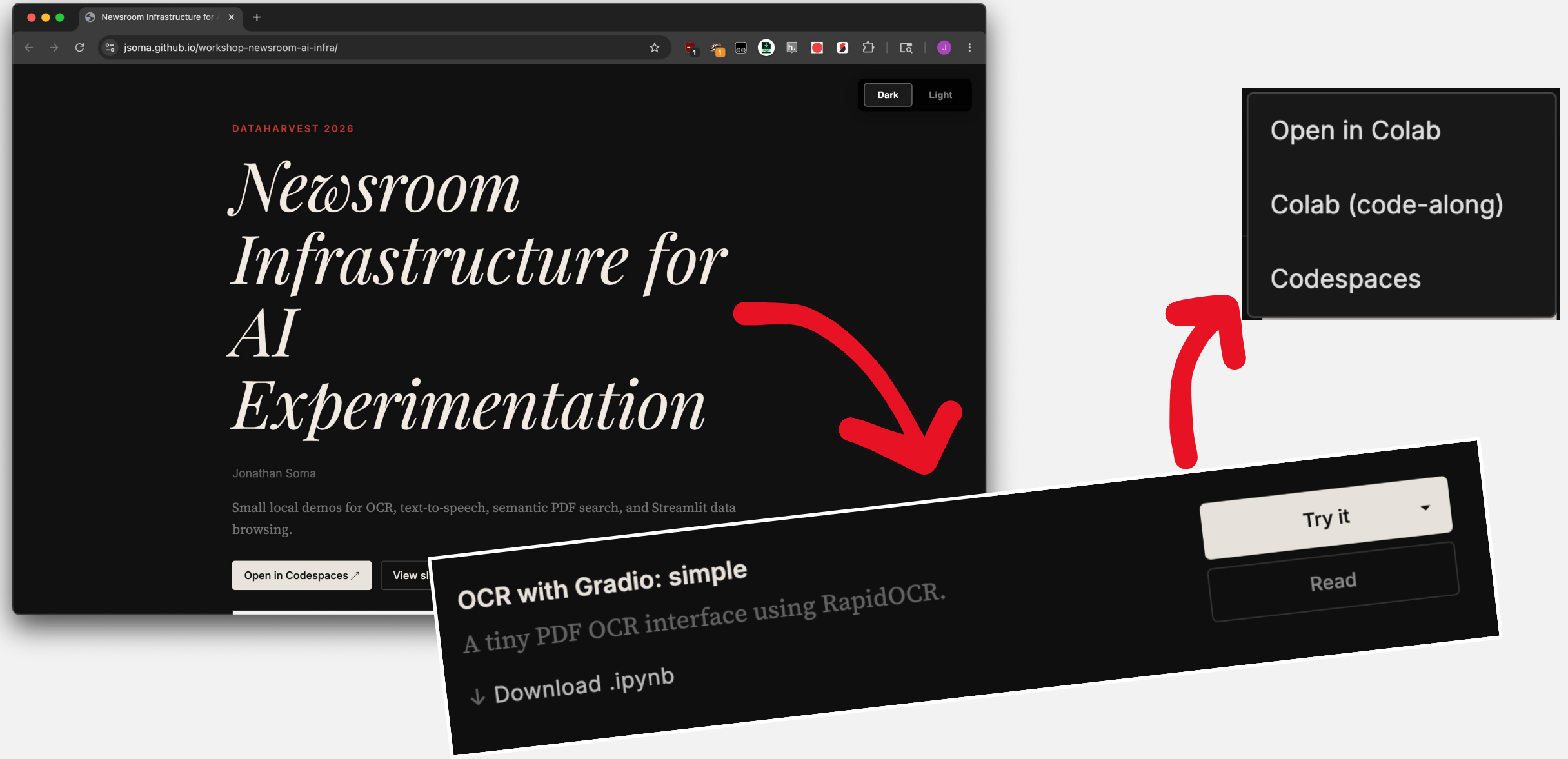
Gradio

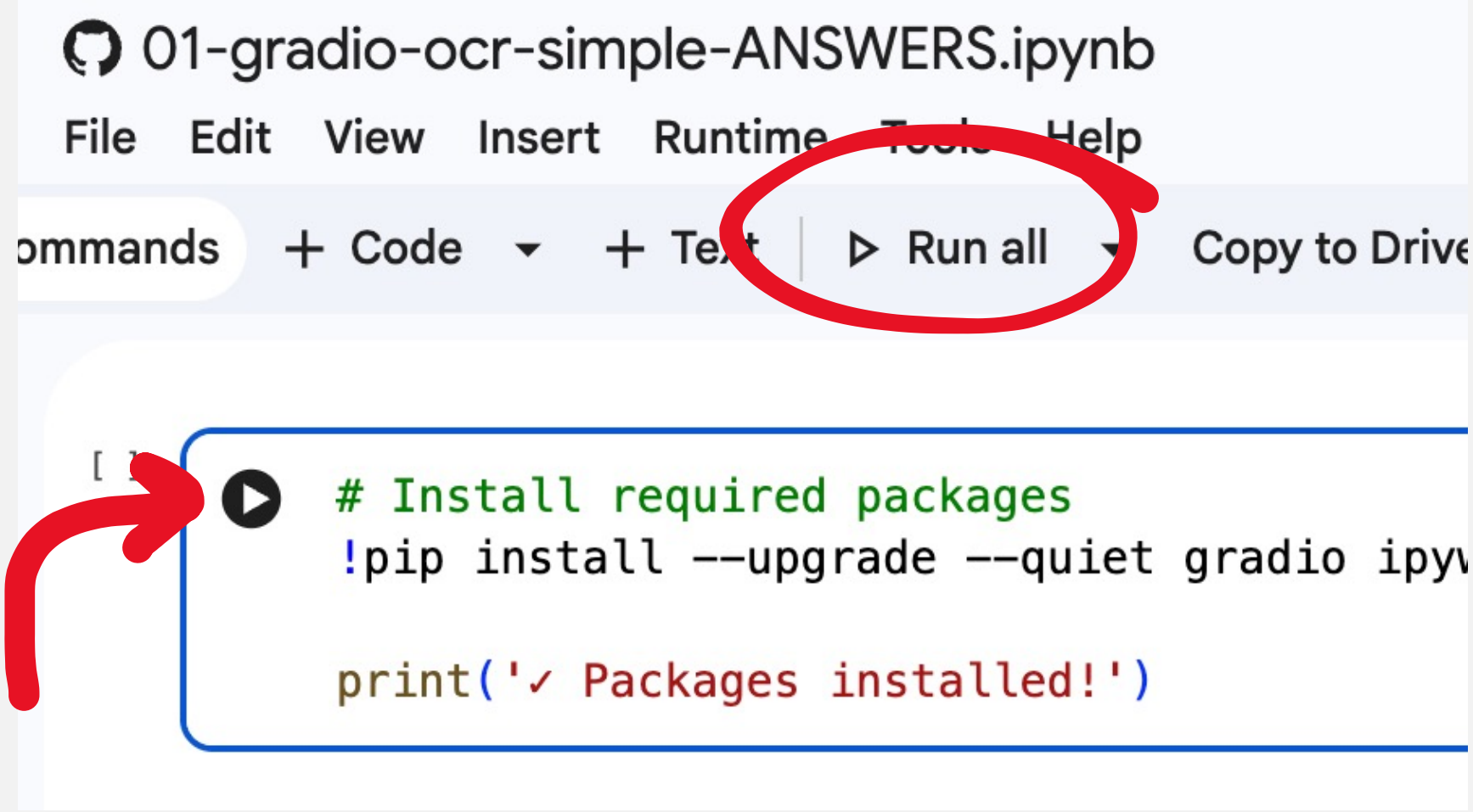
```python
import gradio as gr

def greet(name):
    return f"Hello, {name}!"

with gr.Blocks() as demo:
    name = gr.Textbox(label="Name")
    output = gr.Textbox(label="Output")

    btn = gr.Button("Submit")
    btn.click(greet, inputs=name, outputs=output)

demo.launch()
```

Streamlit

```python
import streamlit as st

name = st.text_input("Name")

if st.button("Submit"):
    st.write(f"Hello, {name}!")
```



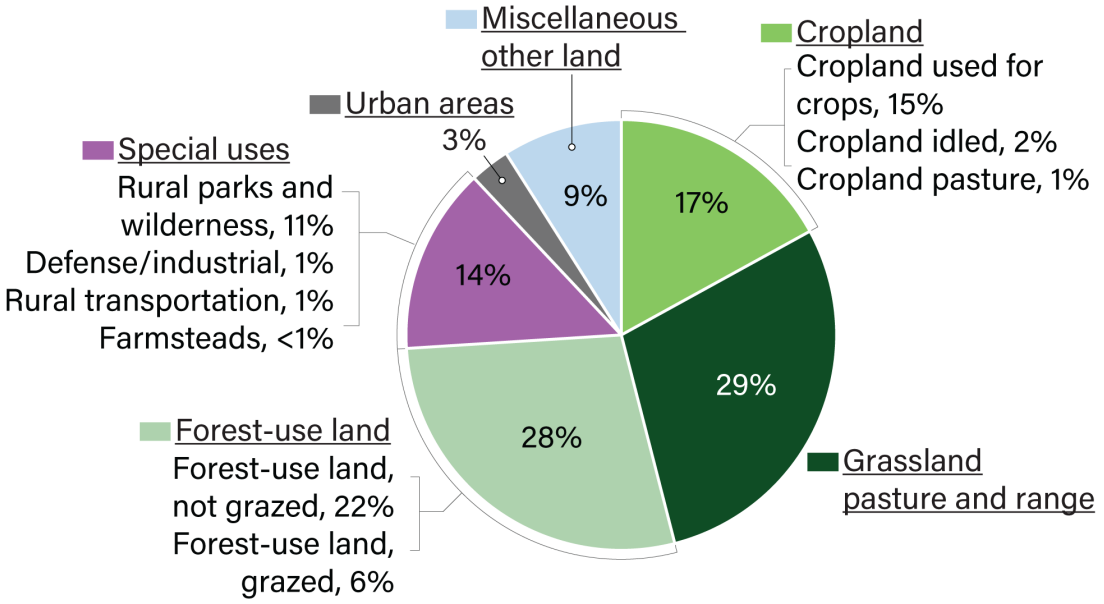

Newsroom Infrastructure for
jsoma.github.io/workshop-newsroom-ai-infra/
Dark
Light
DATAHARVEST 2026
Newsroom Infrastructure for AI Experimentation
Jonathan Soma
Small local demos for OCR, text-to-speech, semantic PDF search, and Streamlit data browsing.
Open in Codespaces ↗
Open in Colab
Colab (code-along)
Codespaces
OCR with Gradio: simple
A tiny PDF OCR interface using RapidOCR.
↓ Download .ipynb
Try it
Read


01-gradio-ocr-simple-ANSWERS.ipynb
File Edit View Insert Runtime Help
ommands + Code + Text Run all Copy to Drive
Install required packages
!pip install --upgrade --quiet gradio ipyw
print('✓ Packages installed!')

Custom everything

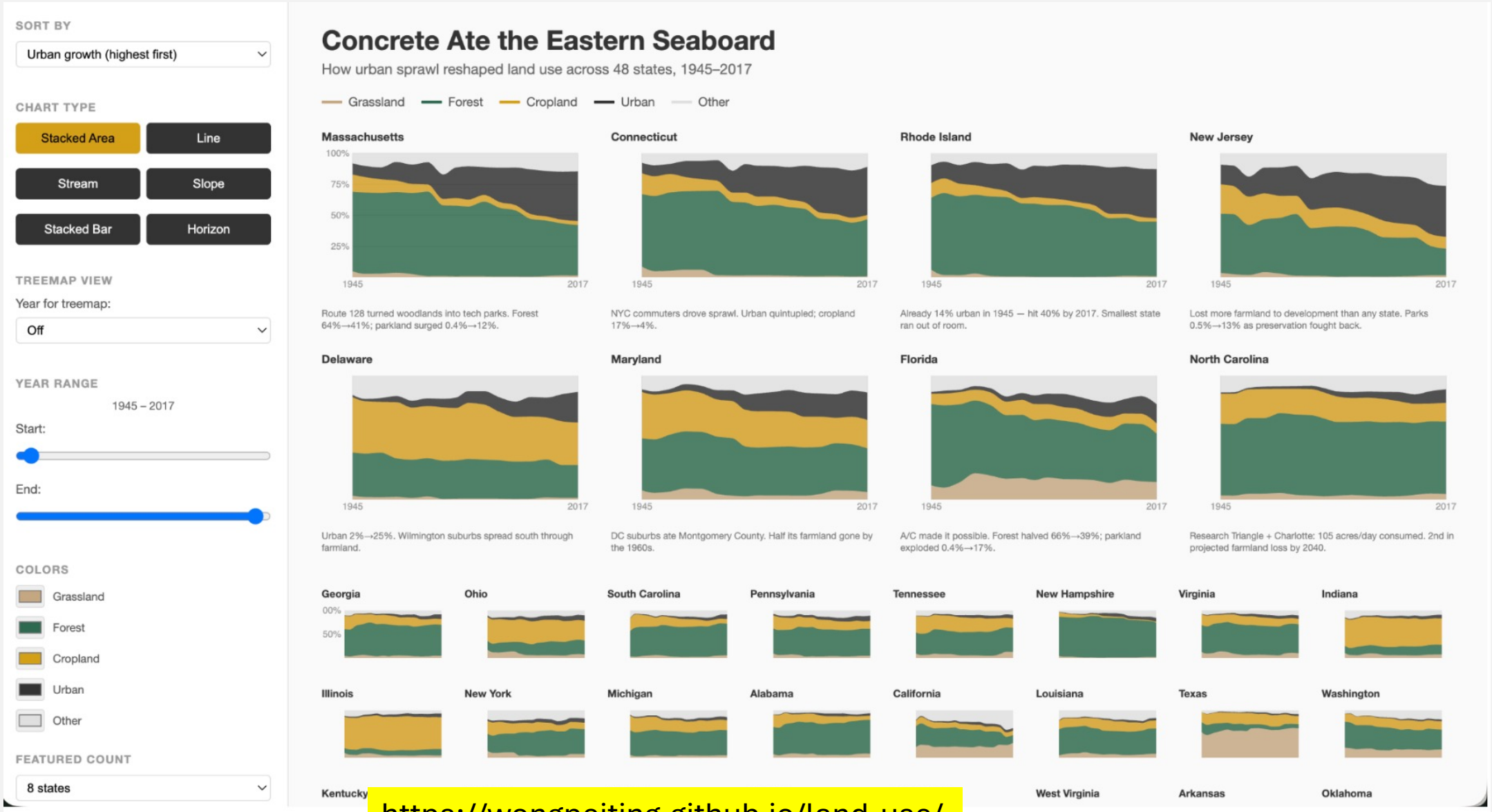
- Codex + Claude Code can now one-shot a *lot* of stuff
- Infrastructure and custom tooling is cheap
- “Home-cooked software”

Share of major U.S. land use, 2017

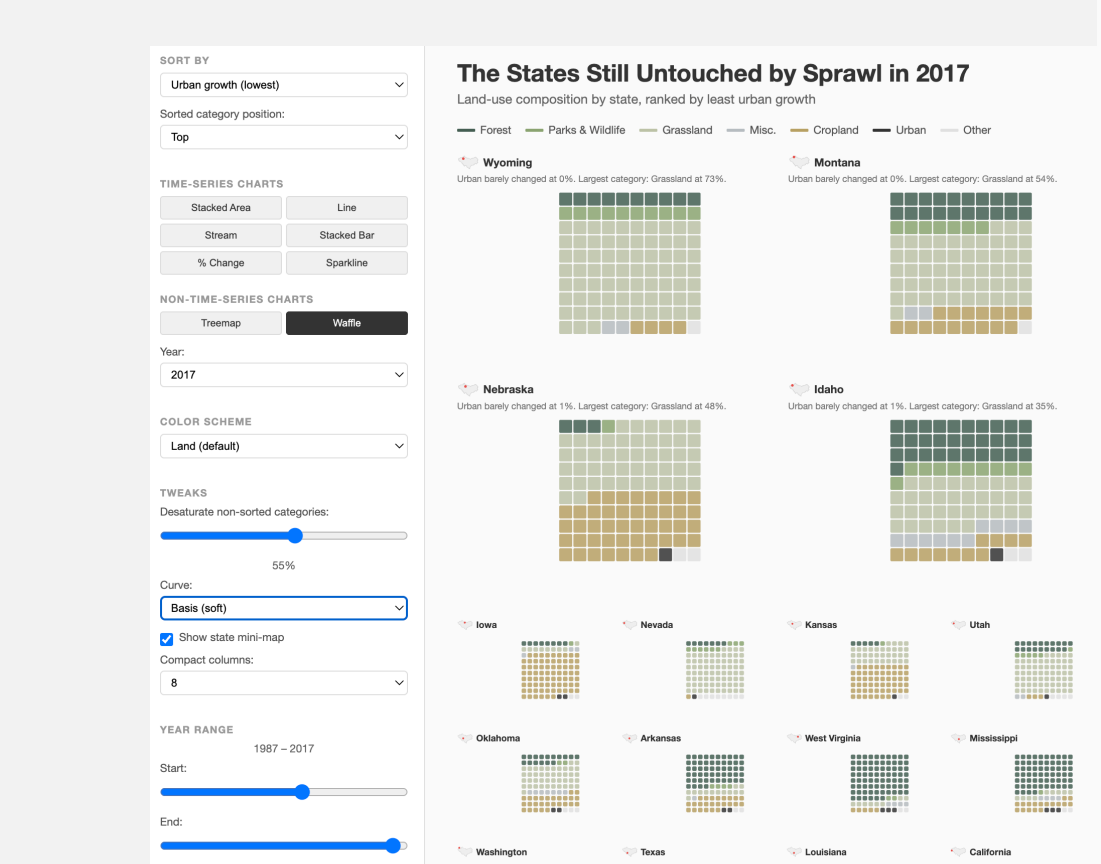


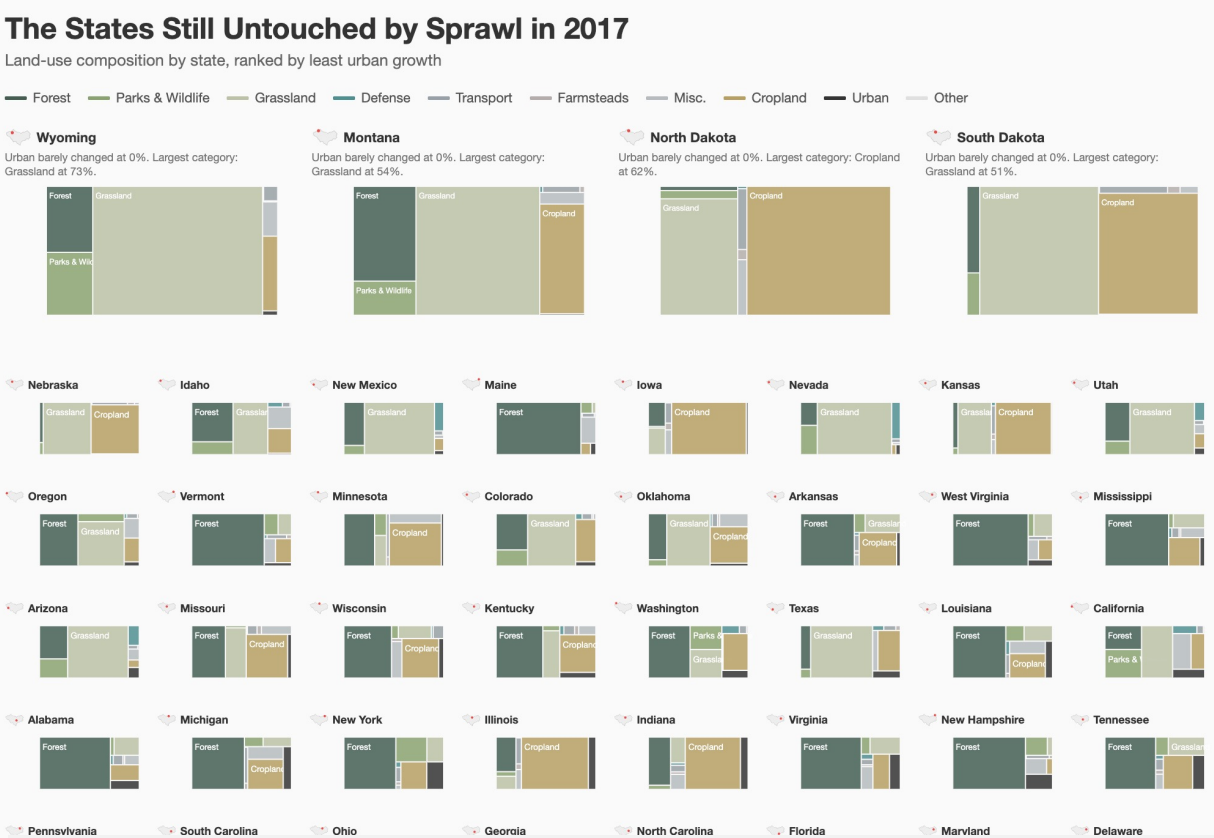
Note: Sub-components may not sum to the component totals because of rounding. Forest-use includes land that serves commercial forest uses, including grazing, as opposed to land that has forest cover but is used for other purposes. Miscellaneous includes uses such as land in cemeteries, golf courses, mining areas, quarry sites, marshes, swamps, sand dunes, bare rocks, deserts, tundra, and other unclassified land, as well as some—but not all—industrial, commercial, and residential sites in rural areas.

Source: USDA Economic Research Service estimates based on data from USDA, National Agricultural Statistics Service; USDA, Farm Service Agency; USDA, Natural Resources Conservation Service; USDA, Forest Service; U.S. Department of Commerce, Bureau of the Census; U.S. Department of Defense; U.S. Department of the Interior, Bureau of Land Management, National Park Service; U.S. Geological Survey; and Utah State University.



https://wongpeiting.github.io/land-use/





https://wongpeiting.github.io/land-use/

Leveraging domain expertise

- Writing/tweaking prompts
- Developing test cases
- Scoring responses
- Fine-tuning LLM judges
- This all falls under evals

Review Classifications

Confirmed: 600 / 600

← | Confirm → | Skip → | Unreviewed | All | Human only | AI only | Export CSV | Reset Cache

GLOSSARY

SUBJECT — WHAT IS IT ABOUT?

Celebrating Trump
The video's entire purpose is hero worship. GOAT framing, dramatic montages of Trump looking powerful, victory laps. Trump appears in almost every video — only pick this if the point IS Trump, not a policy he's in.

Attacking opponents
Mocking Democrats (Schumer, Pelosi, Biden, Jeffries), attacking media/CNN, ridiculing journalists. The mockery is the point. If it blames Democrats for an immigration issue, it's this — not enforcement.

Law enforcement
Immigration, ICE raids, deportation, border wall, "illegals", CBP, crime stats, DHS recruitment. The enforcement apparatus. If Trump appears alongside ICE footage, it's still this.

War & military
Iran strikes, Operation Epic Fury, military operations, combat footage, troops deploying. If it shows actual military hardware or strike footage, it's war — even if Trump appears heroically.

Culture & sports
Sports events, holidays, entertainment, Olympics, lifestyle moments. The Cristiano Ronaldo meeting, March Madness, St Patrick's Day. The soft stuff with no policy edge.

Governing & diplomacy
Executive orders, economic numbers ("$10.5 trillion"), energy policy, legislation, foreign dignitary meetings, trade deals, summit handshakes. If Trump is shown with policy achievements (gas prices, executive orders), it's this — not "trump".

Promotion / meta
WH app launch, internships, roundups, "subscribe", DHS recruitment videos credited to DHS. Content about the account or government recruiting itself.

Unclear
Genuinely ambiguous. Cryptic posts with no context. If you can't tell what the video is about after watching it, pick this.

PACKAGING — INSTITUTIONAL TO TEETERING ON FICTIONAL

① Official
Press conference, signing ceremony, official statement. Clean editing. Talking heads. What government video normally looks like. No style, no flair.

② Direct address

2025-11-13 · #273/600 · 1.7M views HUMAN VERIFIED

The most iconic duo ❤️🇺🇸

Open on TikTok ↗

Notes: Shows Trump and Melania holding hands and walking down a hallway at the White House. Music: Take My Breath Away (Love Theme from "Top Gun"), by Berlin.

SUBJECT

Celebrating Trump | Attacking opponents | Law enforcement | War & military | Culture & sports | Governing & diplomacy | **Promotion / meta** | Unclear

PACKAGING — ① INSTITUTIONAL → ⑦ TEETERING ON FICTIONAL

① Official | ② Direct address | **③ Produced/cinematic** | ④ TikTok-native | ⑤ Pop culture mashup | ⑥ Internet meme | ⑦ Game interface

TAGS

AUDIO

Trending song | Meme audio clip | Bass drop / bass-boosted | **Dramatic / epic score** | Original audio | Hype language

VISUAL EFFECTS

Deep-fried / distortion | Speed ramp / slow-mo | Greyscale to color | Stock hero / B-roll reuse

NARRATIVE STRUCTURE

Comparison / before-after | Highlight reel / montage | Gamification (scores, achievements) | Fictional overlay on real footage | Punchline / reveal | Call & response to critics

CONTENT FLAGS

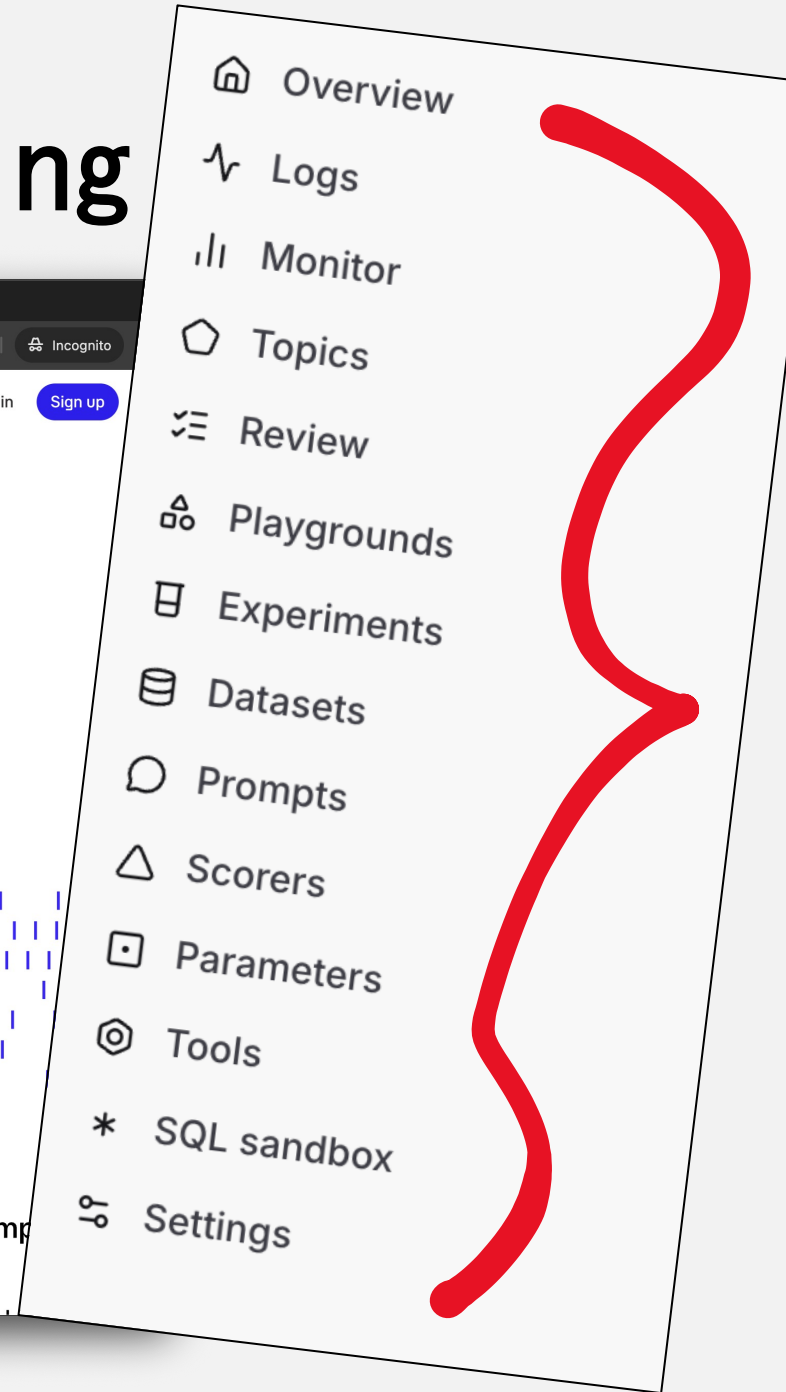
Expletive / swearing | Names a specific person to mock | Real military footage used | AI-generated or CGI visuals | Trump dancing | Troll / shitpost | Provocative / edgy | Recruitment | **Aura farming** | Hyper-masculine | MAGA Minute

NOTES

Shows Trump and Melania holding hands and walking down a hallway at the White House. Music: Take My Breath Away (Love Theme from "Top Gun"), by Berlin.

Enter confirm + next · → skip · ← prev · S export

Observability/evals/prompting



braintrust.dev (and 2,000 others)

The LA Times' new AI tool sy

cnn.com/2025/03/05/m...

CNN Business

Watch

Subscribe

Markets →

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

Fear & Greed Index →

47

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-generat

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

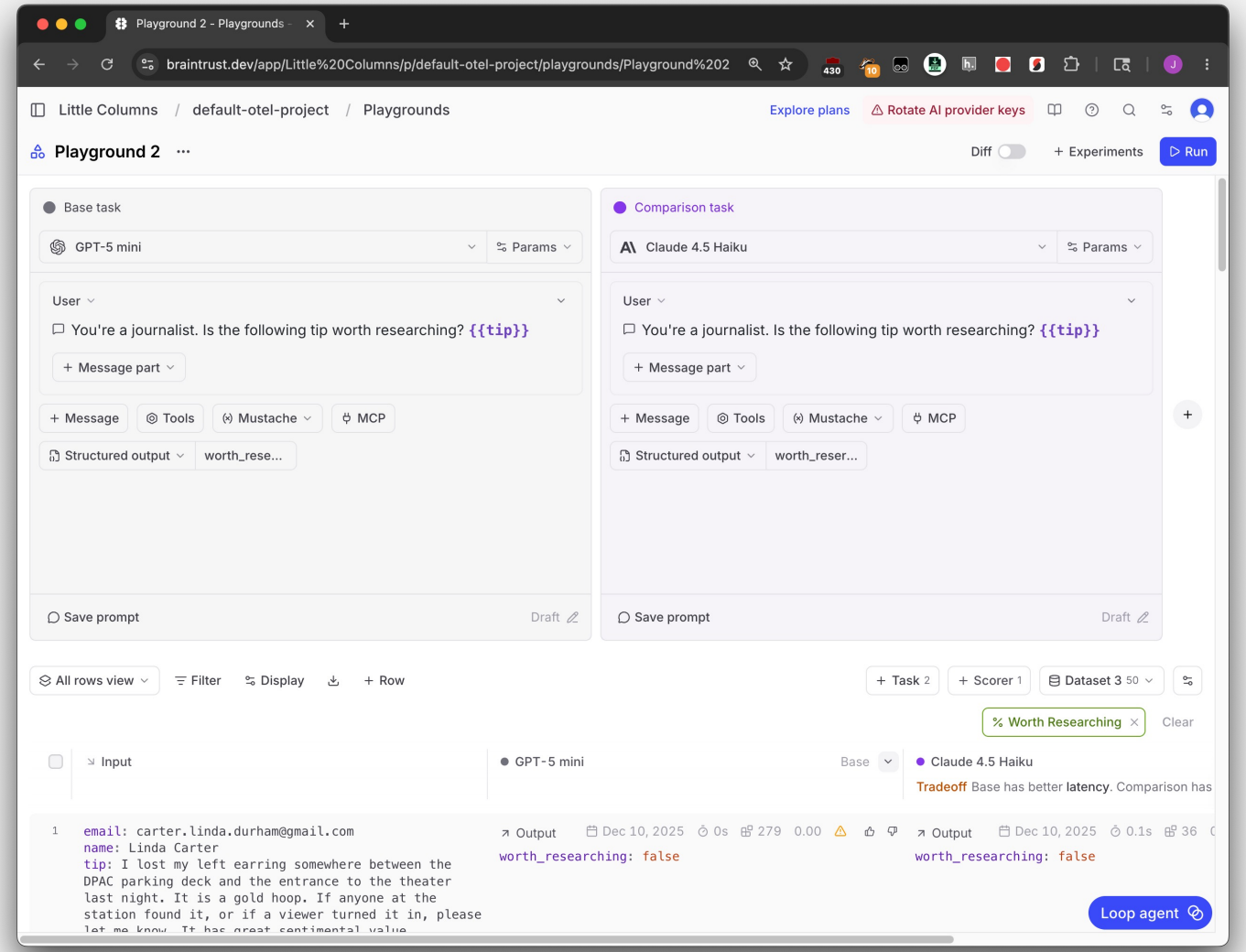
Intelligence for the New World Economy

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




```
email: carter.linda.durham@gmail.com
name: Linda Carter
tip: I lost my left earring somewhere between the
DPAC parking deck and the entrance to the theater
last night. It is a gold hoop. If anyone at the
station found it, or if a viewer turned it in, please
let me know. It has great sentimental value.
```

```
email: durham.watchdog.11@protonmail.com
name: Anonymous
tip: I saw a drone flying over the American Tobacco
Campus near the water tower. It was hovering very
still. I think it might be spying on the people
eating at the restaurants. Is this legal? You should
warn people not to eat outside until we know who is
piloting it.
```

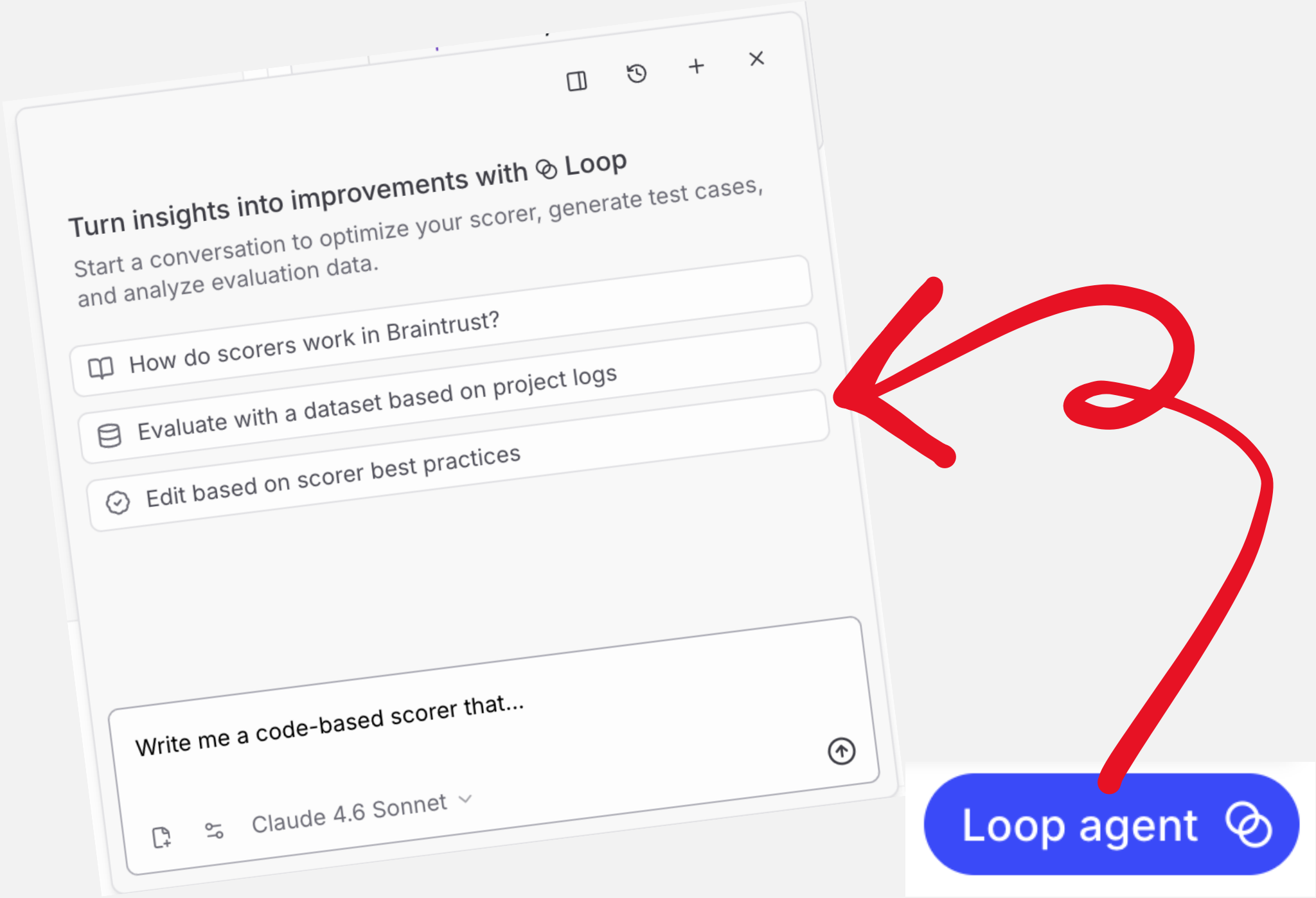
```
email: k.oconnor55@verizon.net
name: Kevin O'Connor
tip: My grandson, Timmy, just learned how to tie his
shoes all by himself. He is five years old and
```

Input	GPT-5 mini	Base	Claude 4.5 Haiku
			Tradeoff Base has better latency. Comparison has be
email: carter.linda.durham@gmail.com name: Linda Carter tip: I lost my left earring somewhere between the DPAC parking deck and the entrance to the theater last night. It is a gold hoop. If anyone at the station found it, or if a viewer turned it in, please let me know. It has great sentimental value.	Output Dec 10, 2025 0s 279 0.00 worth_researching: false % Worth Researching 100%		Output Dec 10, 2025 0.1s 36 0.0 worth_researching: false % Worth Researching 100%
email: durham.watchdog.11@protonmail.com name: Anonymous tip: I saw a drone flying over the American Tobacco Campus near the water tower. It was hovering very still. I think it might be spying on the people eating at the restaurants. Is this legal? You should warn people not to eat outside until we know who is piloting it.	Output Dec 10, 20 0s 151 0.00 worth_researching: true % Worth Researching 0%		Output Dec 10, 202 0.1s 36 0.0 worth_researching: false % Worth Researching 100%
email: k.oconnor55@verizon.net name: Kevin O'Connor tip: My grandson, Timmy, just learned how to tie his shoes all by himself. He is five years old and	Output Dec 10, 202 0s 279 0.00 worth_researching: true		Output Dec 10, 202 0.1s 36 0.0 worth_researching: false

Scorers: code-based vs LLM-as-judge

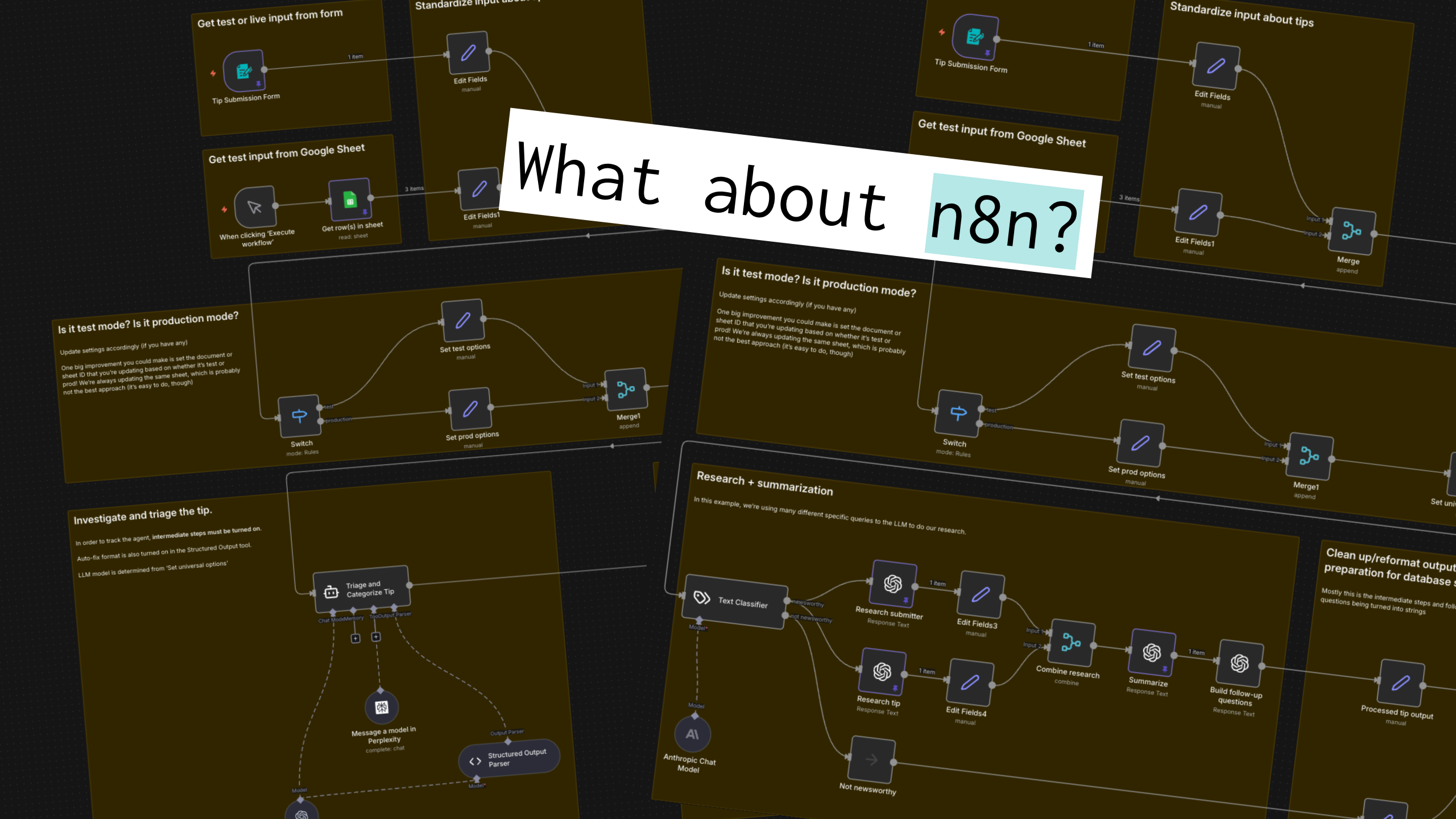
```python
from typing import Any
import re

def handler(input: Any, output: Any, expected: Any, metadata: dict[str, Any]):
    """
    A heuristic scorer to detect overly personal language.
    Fast and explainable, but imperfect.
    """
    output_text = str(output) if output else ""
    text = output_text.strip().lower()

    # Patterns that indicate personal/informal language
    PERSONAL_PATTERNS = [
        r"\bthank(s| you)\b",
        r"\bi (really )?(appreciate|feel|think|believe|agree|disagree)\b",
        r"\bi'm\b",
        r"\bmy (view|opinion|take|heart)\b",
        r"\bdear\b",
        r"\bsincerely\b",
        r"\bhere'?s\b",
        r"\bgreeting\b",
    ]

    matched_patterns = []
    for pattern in PERSONAL_PATTERNS:
        if re.search(pattern, text, flags=re.IGNORECASE):
            matched_patterns.append(pattern)

    return {
        "score": 0.0 if matched_patterns else 1.0,
        "name": "Not Too Personal",
        "metadata": {
            "matched_patterns": matched_patterns,
```

```
You are grading whether the text uses overly personal,
informal, or therapeutic language.

Evaluate the TEXT for:
- Greetings (e.g., "Dear", "Hi")
- Thanks or sign-offs (e.g., "Thank you", "Sincerely",
"Best regards")
- First-person emotional framing (e.g., "I appreciate",
"I feel", "I believe")
- Conversational or therapeutic tone

Grade as:
- **a** if the text is professional, analytical, and
avoids personal language entirely
- **b** if the text has mild first-person framing once
(like "I think") but is otherwise professional
- **c** if the text contains greetings, thanks, sign-
offs, or strongly personal/therapeutic tone

TEXT:
{{output}}

Return ONLY the letter **a**, **b**, or **c**.
```


Turn insights into improvements with Loop
Start a conversation to optimize your scorer, generate test cases, and analyze evaluation data.
How do scorers work in Braintrust?
Evaluate with a dataset based on project logs
Edit based on scorer best practices
Write me a code-based scorer that...
Claude 4.6 Sonnet
Loop agent

Your evals todo-list cheatsheet

1. Build a dataset of possible **inputs**
   - Real-life are best, but there are some techniques for synthetic ones
   - Have a grouchy teammate who says "it'll *NEVER* work in *THIS* case" try to write the hard ones.
   - "If it can't do these successfully, it can't go to production"
2. Make your **prompt** alone or with input
3. Throw the **inputs** and **prompt** together to get **outputs**
4. Have the good-spirited review a spreadsheet of **inputs + outputs** and score them.
   - Yes/No is much easier to deal with than 1-5 scores
5. Use their responses to teach an LLM judge about edge cases
6. Repeat!!! Although if you get to 100% then your evals probably aren't difficult enough


What about n8n?
Get test or live input from form
Tip Submission Form
1 item
Edit Fields
manual
Standardize input about tips
Get test input from Google Sheet
When clicking 'Execute workflow'
Get row(s) in sheet
read: sheet
3 items
Edit Fields1
Merge
append
Is it test mode? Is it production mode?
Update settings accordingly (if you have any)
One big improvement you could make is set the document or sheet ID that you're updating based on whether it's test or prod! We're always updating the same sheet, which is probably not the best approach (it's easy to do, though)
Set test options
Switch
mode: Rules
Set prod options
Merge1
Investigate and triage the tip.
In order to track the agent, intermediate steps must be turned on.
Auto-fix format is also turned on in the Structured Output tool.
LLM model is determined from 'Set universal options'
Triage and Categorize Tip
Message a model in Perplexity
complete: chat
Structured Output Parser
Research + summarization
In this example, we're using many different specific queries to the LLM to do our research.
Text Classifier
Anthropic Chat Model
Research submitter
Response Text
Research tip
Edit Fields3
Edit Fields4
Combine research
combine
Summarize
Build follow-up questions
Not newsworthy
Clean up/reformat output
Processed tip output

CrewAI

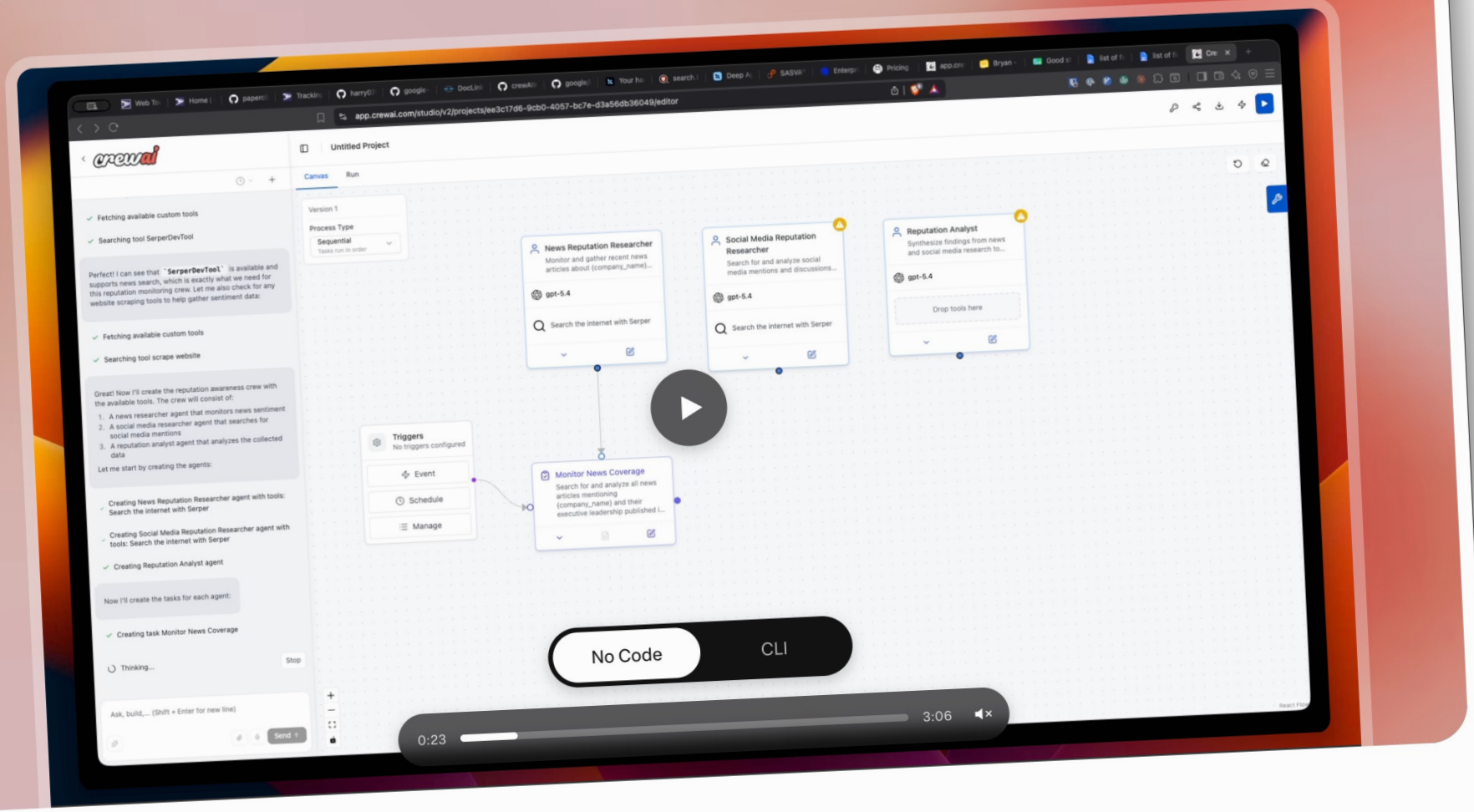
crewai.com

NEW: know what to automate before you build with CrewAI Discovery

crewai

Enterprise Open Source Docs Resources Pricing

Sign In Sign Up



Staying model/vendor agnostic

- You can't experiment if you're married to a vendor!
- Vendor lock-in is the *enemy*. Some examples:
  - **OpenAI:** Custom GPTs, Assistants API, specific parts of Agents SDK
  - **Anthropic:** Claude managed agents, Claude Code settings
  - **Google:** Gems, Workspace integrations, specific parts of ADK
- **Although** DIYing might get you a worse product at a higher cost
- Ability and cost are in constant flux
  - …although except Gemini is always the cheapest

```python
from openai import OpenAI
client = OpenAI()

response = client.responses.create(
    model="gpt-5.5",
    input="Write a one-sentence bedtime story about a unicorn."
)

print(response.output_text)
```

```python
import anthropic

client = anthropic.Anthropic()

message = client.messages.create(
    model="claude-opus-4-7",
    max_tokens=1024,
    messages=[{"role": "user", "content": "Hello, Claude"}],
)
print(message.content)
```

```python
from google import genai

client = genai.Client(api_key='GEMINI_API_KEY')
response = client.models.generate_content(
    model='gemini-2.5-flash',
    contents={'text': 'Why is the sky blue?'},
)

print(response.text)
```

LiteLLM

- Call any provider using the same `completion()` interface — no re-learning the API for each one
- Consistent output format regardless of which provider or model you use
- Built-in retry / fallback logic across multiple deployments via the Router
- Self-hosted LLM Gateway (Proxy) with virtual keys, cost tracking, and an admin UI

```python
from litellm import completion
import os

response = completion(
  model="gemini/gemini-pro",
  messages=[
    {"role": "user", "content": "Hello, how are you?"}
  ]
)
```

Structured outputs with Pydantic

```python
class FilingDetails(BaseModel):
    plaintiff: str
    defendant: str
    charges: str
    city_of_practice: str
    penalty: str
```

```
{
  'Plaintiff': 'The North Carolina State Bar',
  'Defendant': 'Nicholas S. Ackerman',
  'Charges': 'Failure to keep client reasonably inform
the status of the matter; Failure to comply promptly w
reasonable request for information; Failure to partici
good faith in the fee dispute resolution process',
  'City_of_practice': 'Greensboro',
  'Penalty': 'Suspension of license for one (1) year,
for two (2) years as long as Defendant complies with
conditions.'
}
```

WAKE COUNTY

FILED DHC

THE NORTH CAROLINA STATE BAR

NICHOLAS S. ACKERMAN, Attorney,

7. Defendant filed a Motion to Show Cause September 2014.

8. Defendant failed to serve the opposi

10. A.J. traveled from another county to appe

11. A.J. asked Defendant several times to

20. On 12

required Defendant to p

21. Defendar

22. On 9 Ma

dispute resolution proces

15G0195.

23. Defendant

resolution process and appear in court on her behalf.

Panel enters the following:

CONCLUSIONS OF LAW

1. All parties are properly before the Hearing Panel, and the Panel has jurisdiction

discipline pursuant to N.C. Gen. Stat.

fessional Conduct in effect at the time

(a) By failing to notify A.J. that the hearing had been continued from the 30 September 2014 calendar and by otherwise failing to maintain communication with A.J., Defendant failed to keep his client reasonably informed about the status of the matter in violation of Rule 1.4(a)(3);

(b) By failing to respond to A.J.'s inquiries about the status of the matter, Defendant failed to comply promptly with a reasonable request for information in violation of Rule 1.4(a)(4); and

(c) By failing to provide a response to the fee

good faith in the fee

1.

North Car

Chapter 84

Carolina S

2.

admitted t

rules, regulations, and Rules of Professional Conduct of the laws of the State of North Carolina.

3. Defendant was properly served with process, Hearing Panel with due notice to all parties.

4. During the relevant period referred to herein practice of law in Greensboro, Guilford County, North Carolin

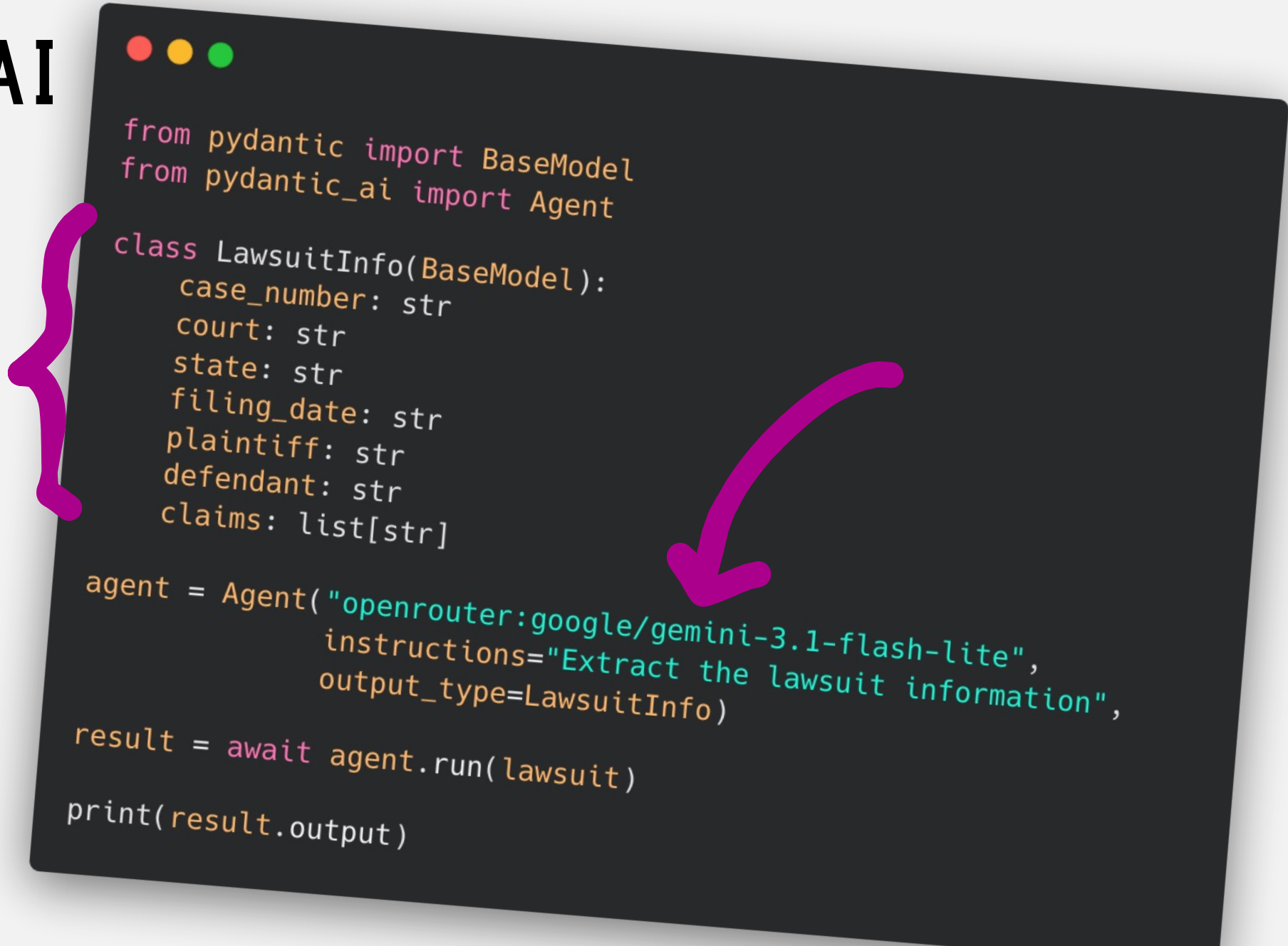
5. On 25 September 2014, A.J. retained Defen contempt order and custody modification.

additional Findings of

parties, the Hearing Panel makes the

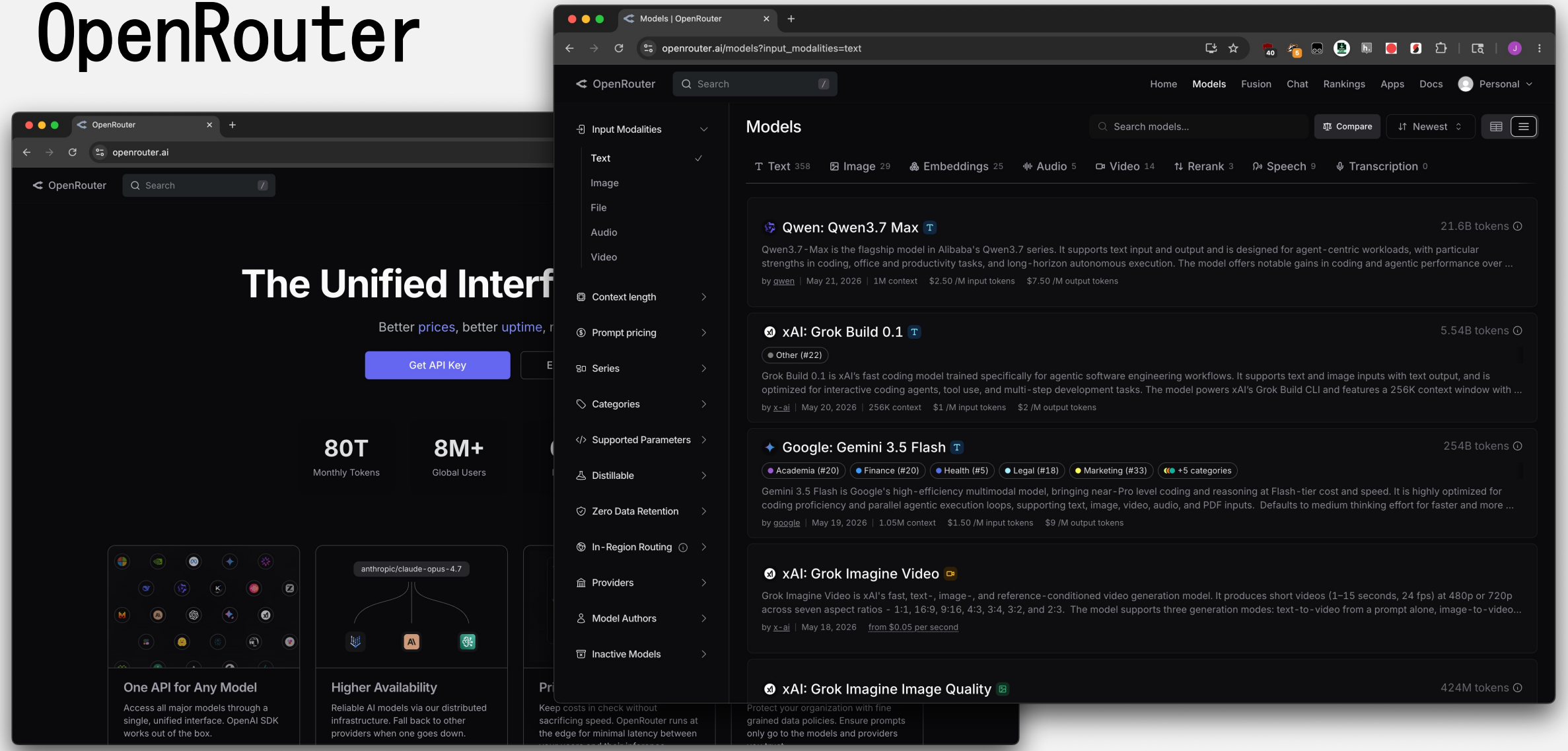
CONCLUSIONS REGARDING DISCIPLINE

Pydantic AI



```python
from pydantic import BaseModel
from pydantic_ai import Agent

class LawsuitInfo(BaseModel):
    case_number: str
    court: str
    state: str
    filing_date: str
    plaintiff: str
    defendant: str
    claims: list[str]

agent = Agent("openrouter:google/gemini-3.1-flash-lite",
              instructions="Extract the lawsuit information",
              output_type=LawsuitInfo)

result = await agent.run(lawsuit)

print(result.output)
```

OpenRouter


Models | OpenRouter
openrouter.ai/models?input_modalities=text
OpenRouter
Search
Home
Models
Fusion
Chat
Rankings
Apps
Docs
Personal
Input Modalities
Text
Image
File
Audio
Video
Context length
Prompt pricing
Series
Categories
Supported Parameters
Distillable
Zero Data Retention
In-Region Routing
Providers
Model Authors
Inactive Models
Models
Search models...
Compare
Newest
Text 358
Image 29
Embeddings 25
Audio 5
Video 14
Rerank 3
Speech 9
Transcription 0
Qwen: Qwen3.7 Max
21.6B tokens
Qwen3.7-Max is the flagship model in Alibaba's Qwen3.7 series. It supports text input and output and is designed for agent-centric workloads, with particular strengths in coding, office and productivity tasks, and long-horizon autonomous execution. The model offers notable gains in coding and agentic performance over ...
by qwen | May 21, 2026 | 1M context $2.50 /M input tokens $7.50 /M output tokens
xAI: Grok Build 0.1
5.54B tokens
Other (#22)
Grok Build 0.1 is xAI's fast coding model trained specifically for agentic software engineering workflows. It supports text and image inputs with text output, and is optimized for interactive coding agents, tool use, and multi-step development tasks. The model powers xAI's Grok Build CLI and features a 256K context window with ...
by x-ai | May 20, 2026 | 256K context $1 /M input tokens $2 /M output tokens
Google: Gemini 3.5 Flash
254B tokens
Academia (#20)
Finance (#20)
Health (#5)
Legal (#18)
Marketing (#33)
+5 categories
Gemini 3.5 Flash is Google's high-efficiency multimodal model, bringing near-Pro level coding and reasoning at Flash-tier cost and speed. It is highly optimized for coding proficiency and parallel agentic execution loops, supporting text, image, video, audio, and PDF inputs. Defaults to medium thinking effort for faster and more ...
by google | May 19, 2026 | 1.05M context $1.50 /M input tokens $9 /M output tokens
xAI: Grok Imagine Video
Grok Imagine Video is xAI's fast, text-, image-, and reference-conditioned video generation model. It produces short videos (1–15 seconds, 24 fps) at 480p or 720p across seven aspect ratios - 1:1, 16:9, 9:16, 4:3, 3:4, 3:2, and 2:3. The model supports three generation modes: text-to-video from a prompt alone, image-to-video...
by x-ai | May 18, 2026 from $0.05 per second
xAI: Grok Imagine Image Quality
424M tokens
OpenRouter
openrouter.ai
OpenRouter
Search
The Unified Interf
Better prices, better uptime,
Get API Key
80T
Monthly Tokens
8M+
Global Users
anthropic/claude-opus-4.7
One API for Any Model
Access all major models through a single, unified interface. OpenAI SDK works out of the box.
Higher Availability
Reliable AI models via our distributed infrastructure. Fall back to other providers when one goes down.
Keep costs in check without sacrificing speed. OpenRouter runs at the edge for minimal latency between
Protect your organization with fine grained data policies. Ensure prompts only go to the models and providers

Models

Search models... Compare Top Weekly

Input Modalities
- Text
- Image
- File
- Audio
- Video

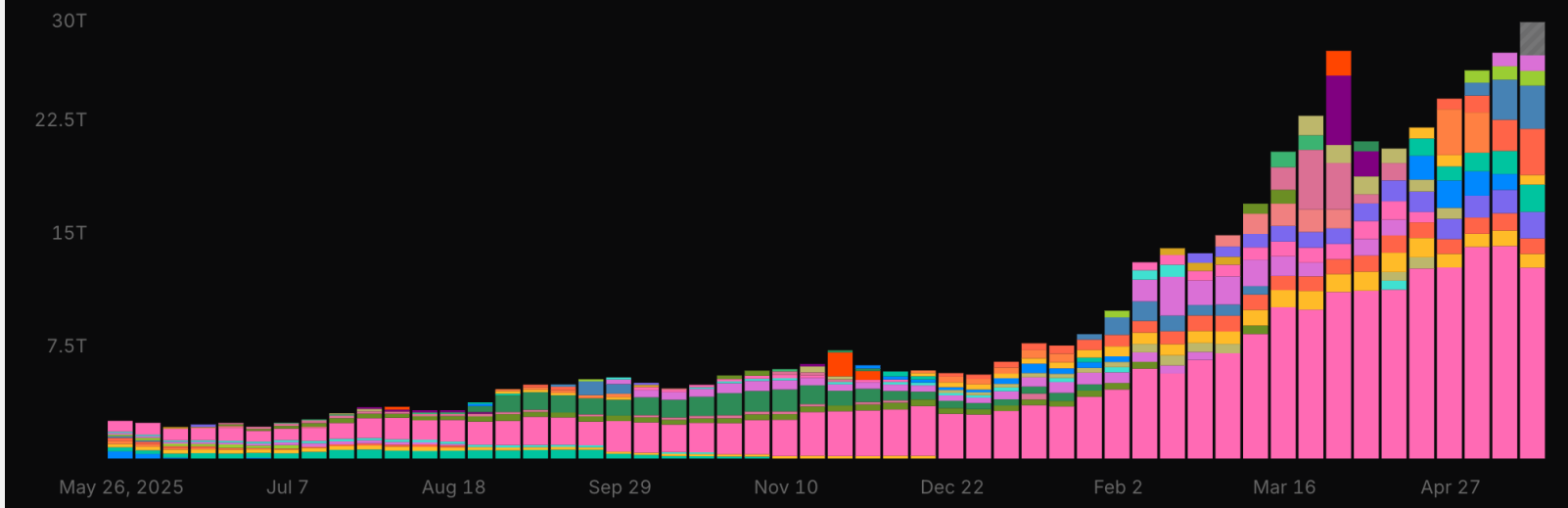
Context length
Prompt pricing
Series
Categories
Supported Parameters
Distillable
Zero Data Retention
In-Region Routing
Providers
Model Authors
Inactive Models

Text 358 | Image 29 | Embeddings 25 | Audio 5 | Video 14 | Rerank 3 | Speech 9 | Transcription 0

Model Name	Weekly Tokens	Input	Output	Context	Released
DeepSeek: DeepSeek V4 Flash	3.46T	$0.10	$0.20	1,048,576	Apr 23, 2026
Tencent: Hy3 preview	3.29T	$0.066	$0.26	262,144	Apr 22, 2026
Anthropic: Claude Opus 4.7	1.95T	$5	$25	1,000,000	Apr 16, 2026
Anthropic: Claude Sonnet 4.6	1.91T	$3	$15	1,000,000	Feb 17, 2026
Owl Alpha	1.22T	$0	$0	1,048,756	Apr 28, 2026
Google: Gemini 3 Flash Preview	1.18T	$0.50	$3	1,048,576	Dec 17, 2025
DeepSeek: DeepSeek V4 Pro	1.1T	$0.435	$0.87	1,048,576	Apr 23, 2026
DeepSeek: DeepSeek V3.2	1.07T	$0.252	$0.378	131,072	Dec 1, 2025
StepFun: Step 3.5 Flash	755B	$0.09	$0.30	262,144	Jan 29, 2026
MoonshotAI: Kimi K2.6	700B	$0.73	$3.49	262,144	Apr 20, 2026
Google: Gemini 2.5 Flash Lite	655B	$0.10	$0.40	1,048,576	Jul 22, 2025
NVIDIA: Nemotron 3 Super (free)	642B	$0	$0	1,000,000	Mar 11, 2026
Google: Gemini 2.5 Flash	607B	$0.30	$2.50	1,048,576	Jun 17, 2025
MiniMax: MiniMax M2.7	598B	$0.279	$1.20	204,800	Mar 18, 2026

Top Models

Weekly usage of models across OpenRouter

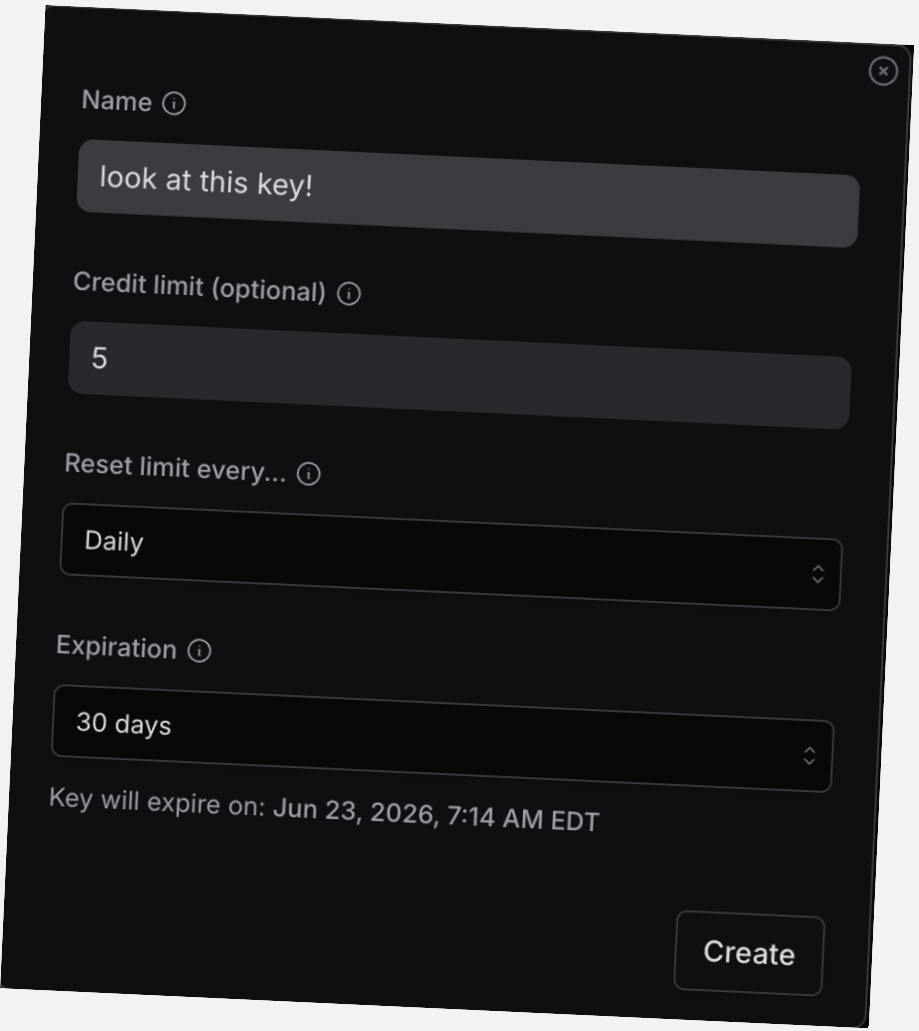


LLM Leaderboard

This Week

Compare the most popular models on OpenRouter

1.	DeepSeek V4 Flash	by deepseek	3.21T tokens	↑73%
2.	Hy3 preview	by tencent	3.08T tokens	↑16%
3.	Claude Opus 4.7	by anthropic	1.84T tokens	↑18%
4.	Claude Sonnet 4.6	by anthropic	1.82T tokens	↑18%
5.	Owl Alpha	by openrouter	1.15T tokens	↑46%
6.	Gemini 3 Flash Preview	by google	1.13T tokens	↓0%
7.	DeepSeek V4 Pro	by deepseek	1.03T tokens	↑17%
8.	DeepSeek V3.2	by deepseek	1.03T tokens	↑3%
9.	Step 3.5 Flash	by stepfun	728B tokens	↑10%
10.	Kimi K2.6	by moonshotai	675B tokens	↓41%



Expires	Last Used	Usage	Limit
May 30, 2026, 3:50 PM EDT	May 23, 2026, 3:51 PM EDT	$ 0.042	$2 WEEK
Jun 21, 2026, 12:09 PM EDT	May 23, 2026, 8:39 PM EDT	$ 4.83	$10 TODAY
May 27, 2026, 1:25 PM EDT	May 21, 2026, 1:55 AM EDT	$ 0.128	unlimited TOTAL
Never	May 17, 2026, 2:00 AM EDT	$ 3.91	unlimited TOTAL
Expired	May 13, 2026, 4:24 PM EDT	$ 34.50	unlimited TOTAL
Never	May 8, 2026, 8:05 PM EDT	< $0.001	$20 WEEK
Never	May 14, 2026, 2:05 PM EDT	$ 3.23	$5 TODAY
Never	Apr 16, 2026, 9:35 PM EDT	$ 2.60	$20 TODAY

But you need evals to do it!!!

- Otherwise you won't know if the change is worth it.

Newsroom Infrastructure for AI Experimentation

Jonathan Soma
Columbia University
@dangerscarf・js4571@columbia.edu

https://bit.ly/dh26-infra