

The sustainability issue with AI — and what we can change as marketers

(the extended version)

Why the per-query numbers keep falling, the total keeps rising, and what an SEO team can actually do about it on a Tuesday afternoon.



Lazarina Stoy

Founder of MLforSEO · Digital marketing consultant

- mlforseo.com
- academy.mlforseo.com
- lazarinastoy.com
- [@lazarinastoy](https://twitter.com/lazarinastoy)

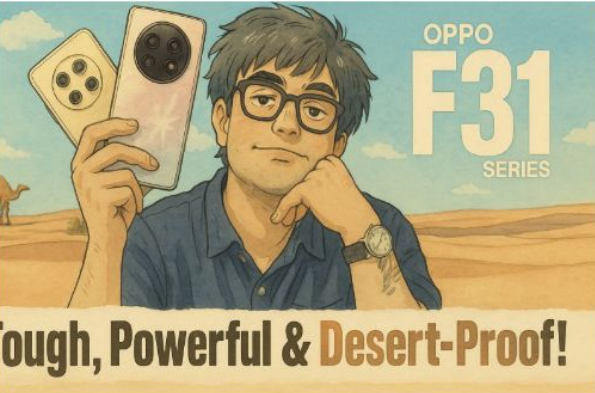


Amey Chauhan • 3rd+
Zoologist & Digital Strategist | Scaling Wildlife Education ...
4mo • 🌐

[+ Follow](#) [...](#)

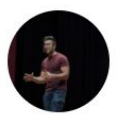
I turned my thumbnail in Studio Ghibli style

... more



 18

 1 comment

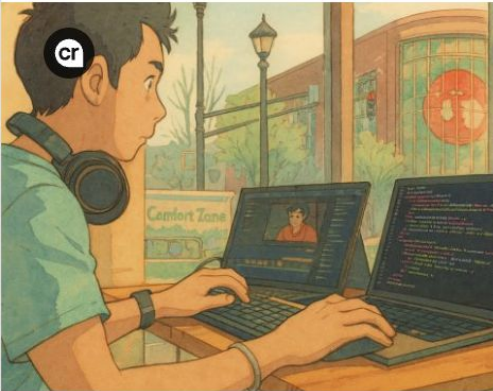


Harnoor Singh • 2nd
Devrel @HydraDB | Ex-Microsoft | YouTuber
4mo • 🌐

[+ Follow](#) [...](#)

Google has also cooked with Studio Ghibli!

Better character and consistent using Nano Banana Pro!



 82

 2 comments



Sanjeev Roy • 2nd
Deputy General Manager – Operations & Sales @ Senco Gol...
7mo • 🌐

[+ Follow](#) [...](#)

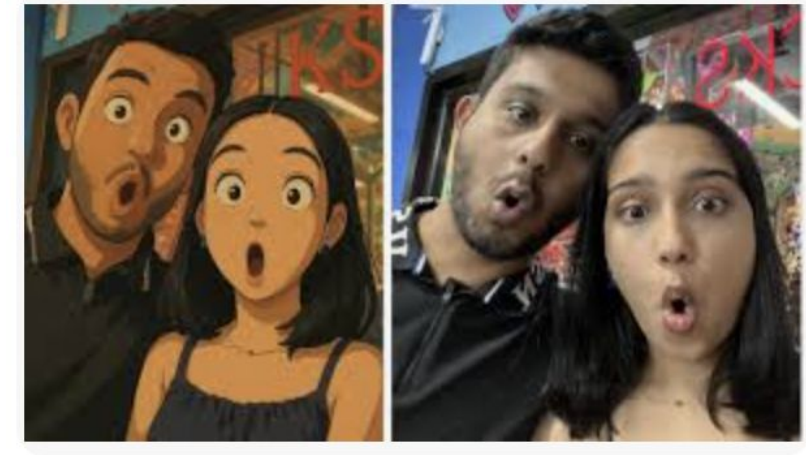
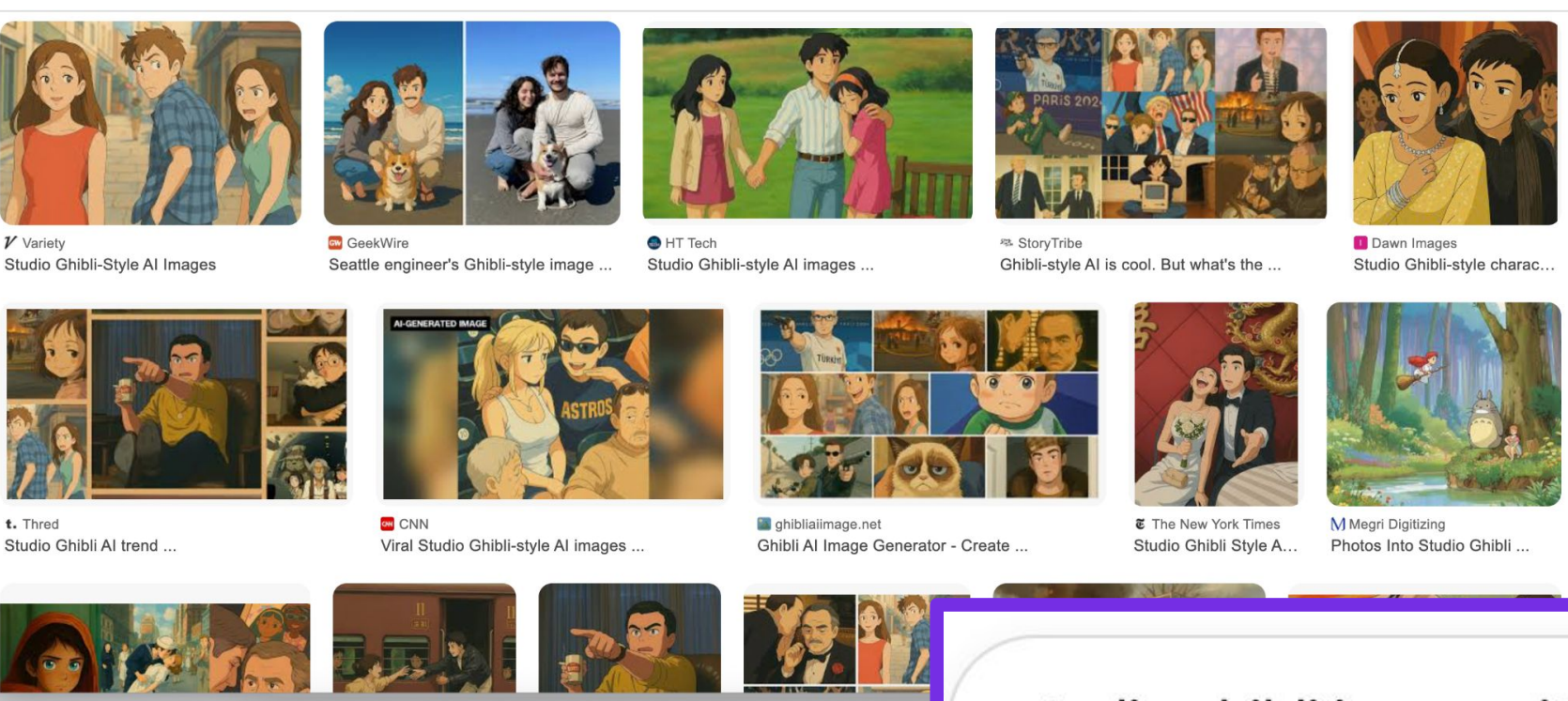
🌟 From Talk of the Town to Timeless Inspiration 🌟

... more



  42

 2 comments

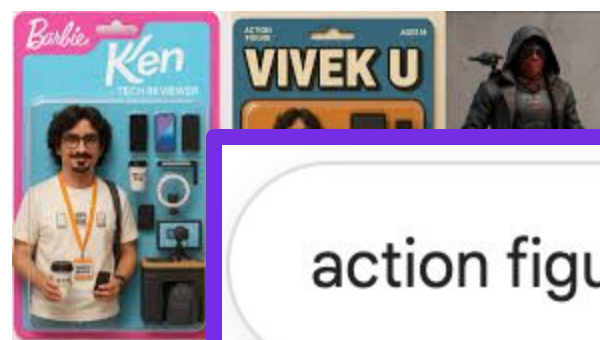
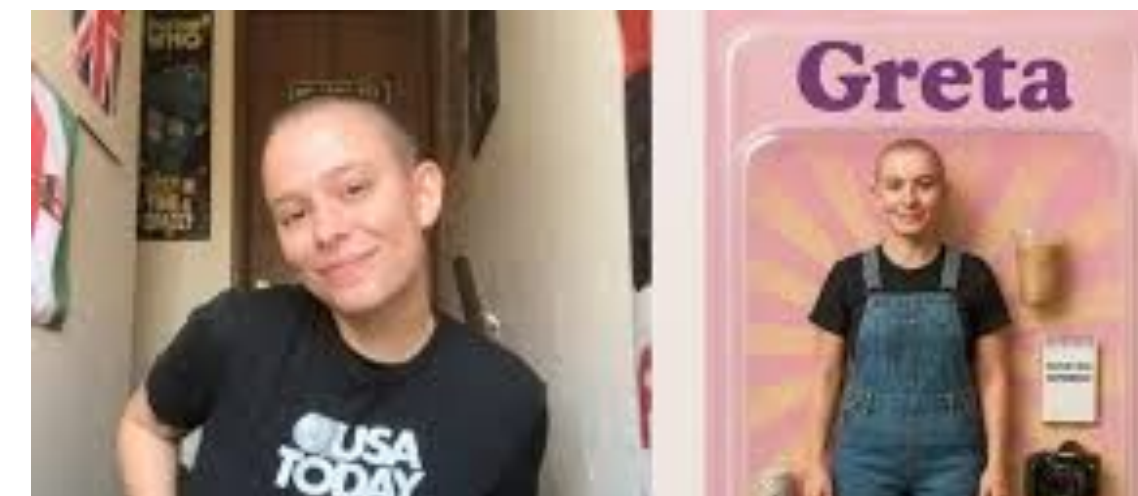
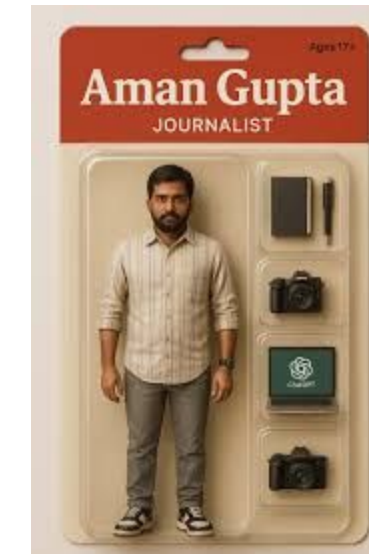


 LinkedIn
Create Stunning Studio Ghibli-Style AI ...

studio ghibli image with ai

About 56,700,000 results (0.45s)



action figure with ai

About 1,700,000,000 results (1.00s)



=



1 image = fully charging your smartphone



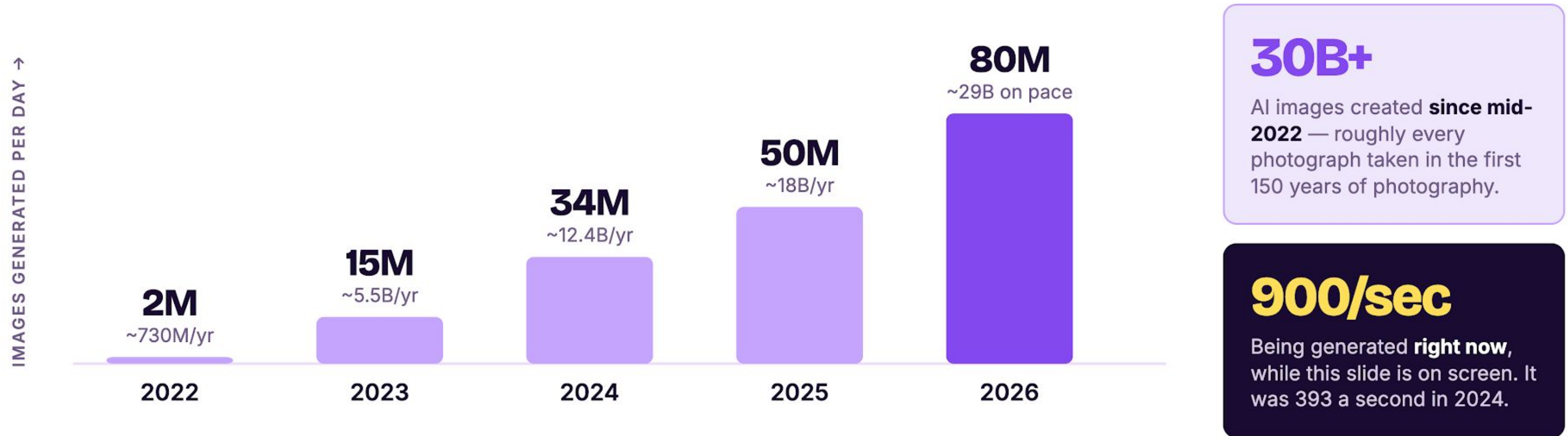
=



1 image = wasting 250 ml water for cooling servers

HOW MANY AI IMAGES ARE THERE?

Two million a day in 2022. Eighty million a day now.



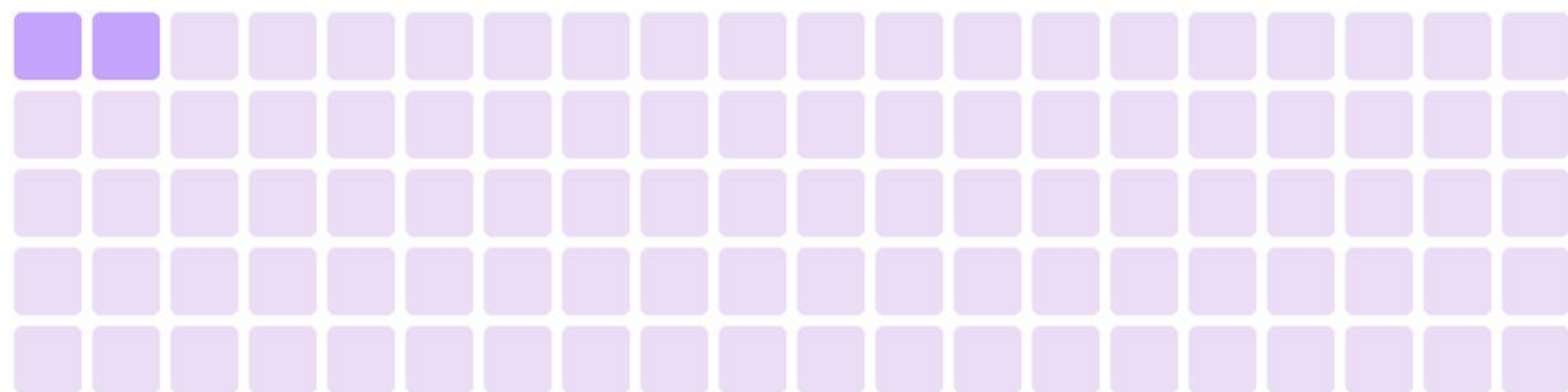
morphed.app, compiling Everypixel Journal & Adobe Newsroom (2026)

Forty times the 2022 rate, in four years. **Nobody decided this; it just compounded.**

SHARE OF ALL IMAGES THAT ARE AI-GENERATED

Two in a hundred. Twenty-three where it competes.

EVERY 100 IMAGES MADE TODAY

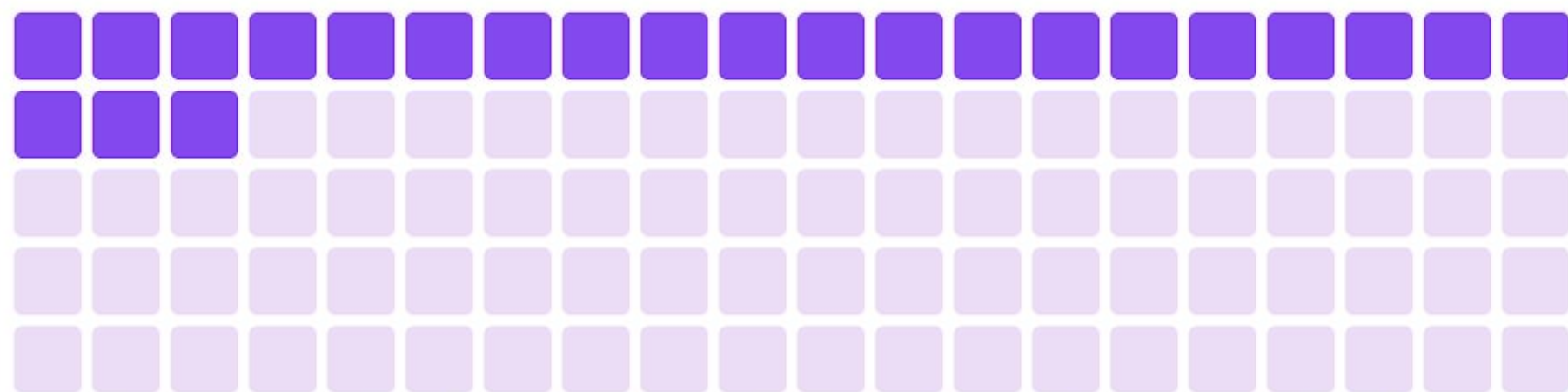


1.5–2%

of everything made

80M AI images a day, against roughly 5.3 billion photographs.
In 2024 it was fewer than one in a hundred.

EVERY 100 NEW STOCK PHOTOS



~23%

of the market it entered

New content on Getty and Shutterstock — where it competes
with photographers head-on.

morphed.app, compiling Everypixel Journal, Adobe, Getty & Shutterstock (2026)

Negligible against everything. **Already a quarter of the market it actually entered.**

7. Who Uses AI Images?

AI images have taken firm hold across several industries. The main use cases:

Use case ↕	Figure
Marketing & advertising	largest segment, over 36% of the market
Social media	71% of shared images AI-generated (per survey)
Graphic design	76% of professionals use AI image tools
E-commerce	up to 60% higher conversion with AI product images



Sam Altman  
@sama



it's super fun seeing people love images in chatgpt.

but our GPUs are melting.

we are going to temporarily introduce some rate limits while we work on making it more efficient. hopefully won't be long!

chatgpt free tier will get 3 generations per day soon.

6:32 PM · Mar 27, 2025 · **8.8M** Views

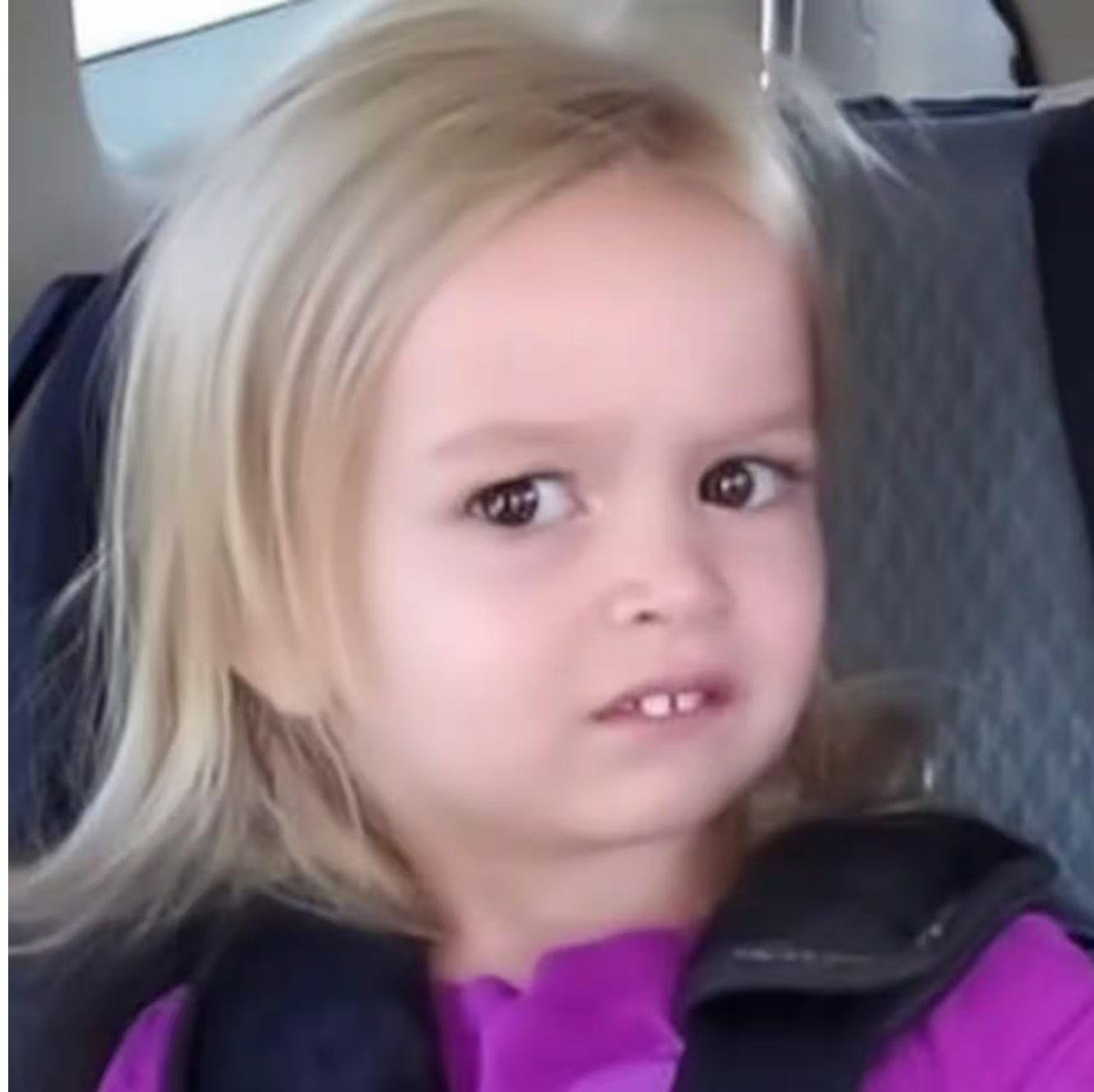
 2.5K

 4.8K

 66K

 4.9K





HOW MUCH ARE WE ACTUALLY PROMPTING?

2.5 billion a day. And that number is fifteen months old.

2.5B

prompts processed by ChatGPT every single day —
the last figure OpenAI has confirmed

JULY 2025

2.5B prompts a day

OpenAI states it publicly. 700 million weekly users at the time.

AUGUST 2026

1 billion weekly users

Gemini and Claude grew faster still. Nobody has published a newer prompt figure.

Weekly users up **~43%**. The number everyone quotes
did not move at all.

OpenAI (July 2025) · user figures: Similarweb / First Page Sage, September 2026

One of them is losing share and growing anyway.

ASSISTANT	SHARE OF WORLDWIDE WEB VISITS	MONTHLY USERS	MOVEMENT
ChatGPT	53.9%	1.1B	was 79.0% a year ago
Google Gemini	27.9%	662M	up ~450% YoY
Claude	9.2%	245M	up ~855% YoY — fastest in the set
Everything else	9.0%	—	Copilot, Perplexity, DeepSeek, Grok

These columns **do not measure the same thing**. ChatGPT reports weekly actives, Gemini counts both app users and AI Overview reach, Claude reports monthly. Treat the ranking as real and the gaps as approximate.

Similarweb via TechCrunch & First Page Sage, September 2026
ChatGPT went from **79% to 54%** in a year ChatGPT went from **79% to 54%** in a year — while its user count rose. **The pie grew faster than the slice shrank.**mdash; while its user count rose. **The pie grew, not the slice.**

WHAT DOES ONE QUERY ACTUALLY COST?

There is no “one query”. The range is over 100×.



The **2.9 Wh** figure quoted everywhere is an **industry-wide average** across all generative AI, not a measurement of a text prompt. It sits between the first two bars — which is why it fits every argument and settles none of them.

Oviedo et al., *Joule* (2026) · Google 0.24 Wh median Gemini prompt (2025) · OpenAI 0.34 Wh, no methodology published

Asking “what does a query cost?” is like asking what a journey costs. **Nobody says how far.**

That 2.9 Wh figure is from 2023.

The labs have since published their own measurements.

Three labs. Three very different answers.

	Google Gemini	OpenAI ChatGPT	Anthropic Claude
Energy per prompt	0.24 Wh	0.34 Wh	—
Water per prompt	0.26 ml	~0.32 ml	—
Published	August 2025 Peer-reviewable paper	June 2025 A blog post	Nothing, as of mid-2026
Methodology shown	FULL	NONE	N/A
Training included	NO	NO	NO

The gap that matters isn't between the numbers — it's that **only one of them showed their working**, and **none of them counted training**.

Google Cloud & arXiv:2508.15734 (Aug 2025) · Altman, "The Gentle Singularity" (Jun 2025)

Down 33× in twelve months.

Google's median prompt got 33× more energy-efficient, and 44× lower carbon.

So the sustainability problem is solved, surely.

Total consumption went up anyway.

485 TWh in 2025, growing 17% a year — and 50% for
AI-specific sites.

Efficiency per prompt is winning. Volume is winning faster.

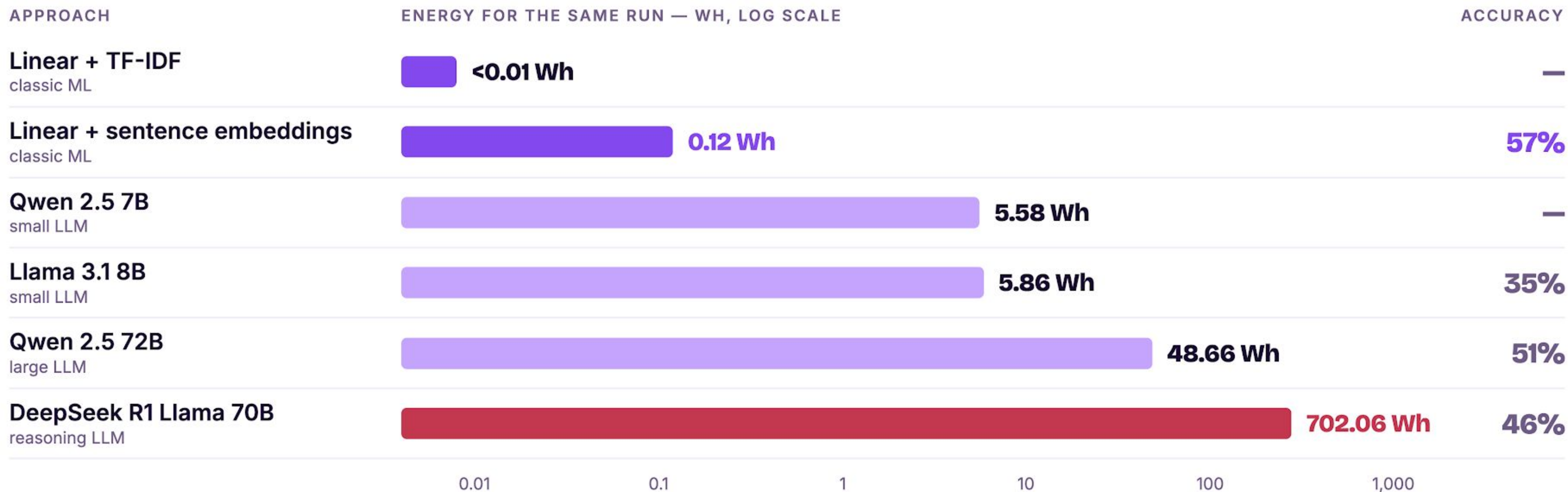
That gap is the whole talk.

We are getting better at the unit and worse at the total.

Which means how you work matters more than which model you pick.

ONE TASK · TEXT CLASSIFICATION · ONE TEST SET

Same job. Six orders of magnitude apart.



Zschache & Hartwig, *Scientific Reports*, April 2026 · FKTG dataset, 189 test samples, zero-shot

The reasoning model used **5,850× the energy** of the linear one — and was **less accurate**.

NOT ALL AI COSTS THE SAME

It isn't AI versus no AI. It is which AI, for which job.

WH PER 1,000 INFERENCE · LOG SCALE

Classification

text or image — a label out

2–7 Wh

Text generation

writing, summarising

40–60 Wh

Image generation

and image captioning

60–2,900 Wh

1 10 100 1,000 10,000

And it is **the architecture, not just the size**: multi-purpose generative models cost orders of magnitude more than task-specific ones for the same task — even holding parameter count constant.

Luccioni, Jernite & Strubell, *Power Hungry Processing* (FAccT 2024) · 88 models, 30 datasets

Generating an image can cost **1,450×** what classifying one does. Same GPU. Different question.

To put this into perspective...

Racks of servers hum along for months, ingesting training data, crunching numbers, and performing computations. This is a time-consuming and expensive process—it's estimated that training OpenAI's GPT-4 took over \$100 million and consumed 50 gigawatt-hours of energy, enough to power San Francisco for three days. It's only after this training, when consumers or customers “inference” the AI models to get answers or generate outputs, that model makers hope to recoup their massive costs and eventually turn a profit.

Source: Luccioni, Jernite & Strubell (2023); AllAboutAI (2025)



945 TWh

Projected data centre electricity consumption by 2030 —
slightly more than Japan's entire national consumption today

Source: IEA, Energy and AI (2025). 415 TWh in 2024, 485 TWh in 2025, ~945 TWh by 2030.

Yet...
DeepSeek-V3
achieved 95%
lower energy while
maintaining
competitive
performance

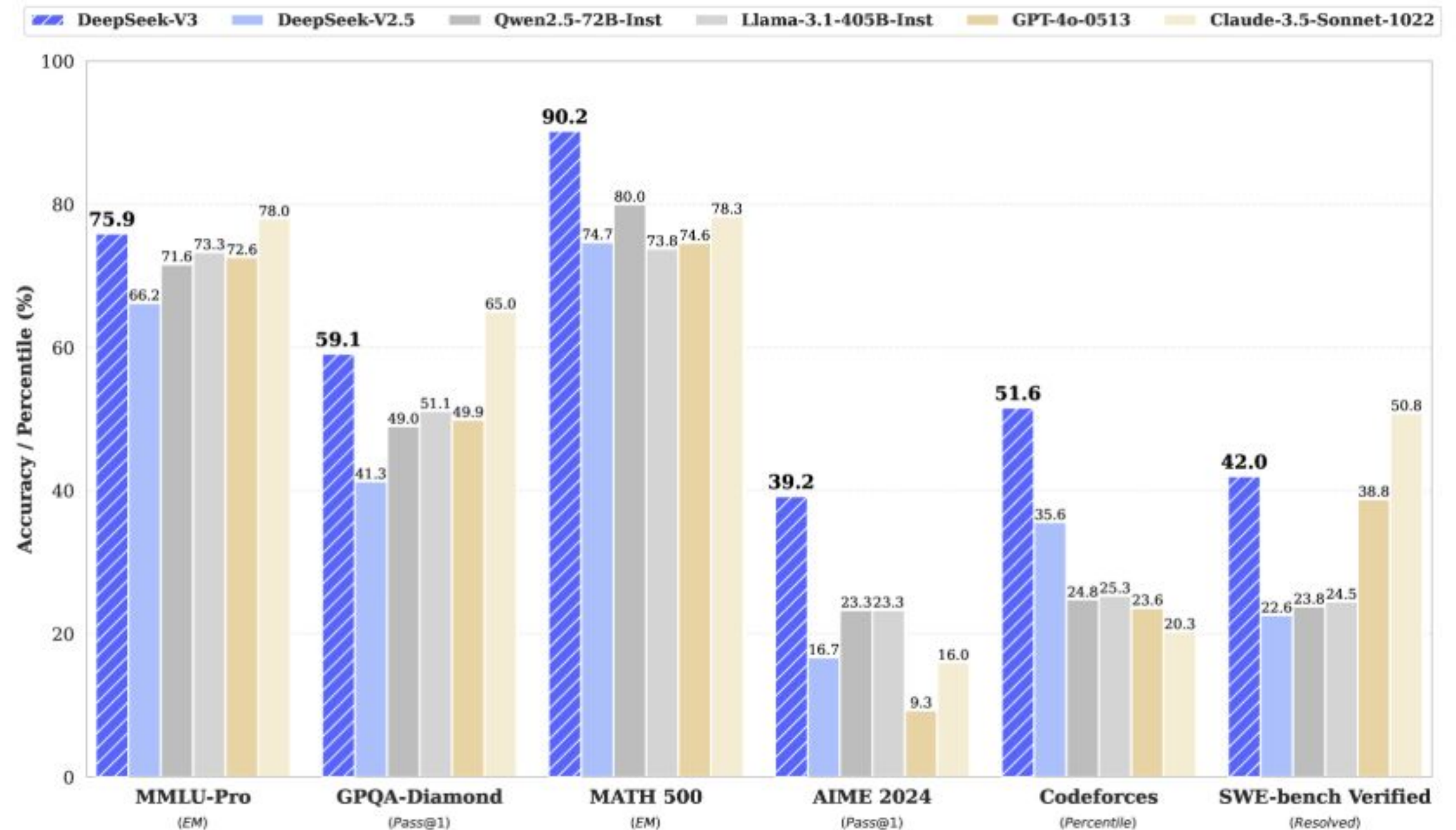


Figure 1: Benchmark performance of DeepSeek-V3 and its counterparts.

Source: Luccioni, Jernite & Strubell (2023); AllAboutAI (2025)

The Corporate Response? Scale at All Costs.

Stargate Initiative

Business / Tech

Trump announces a \$500 billion AI infrastructure investment in the US

Google

DIVE BRIEF

Google plans \$75B investment to build out cloud AI capacity

The majority of the spending will target technical infrastructure, including servers and data centers, CFO Anat Ashkenazi said Tuesday.

Apple

AI EFFECT

Apple to open AI server factory in Texas as part of \$500 billion U.S. investment

Nuclear Revival

NEWS ENERGY

Big Tech Backs Small Nuclear > Google and Amazon invest in small modular reactors to power data centers

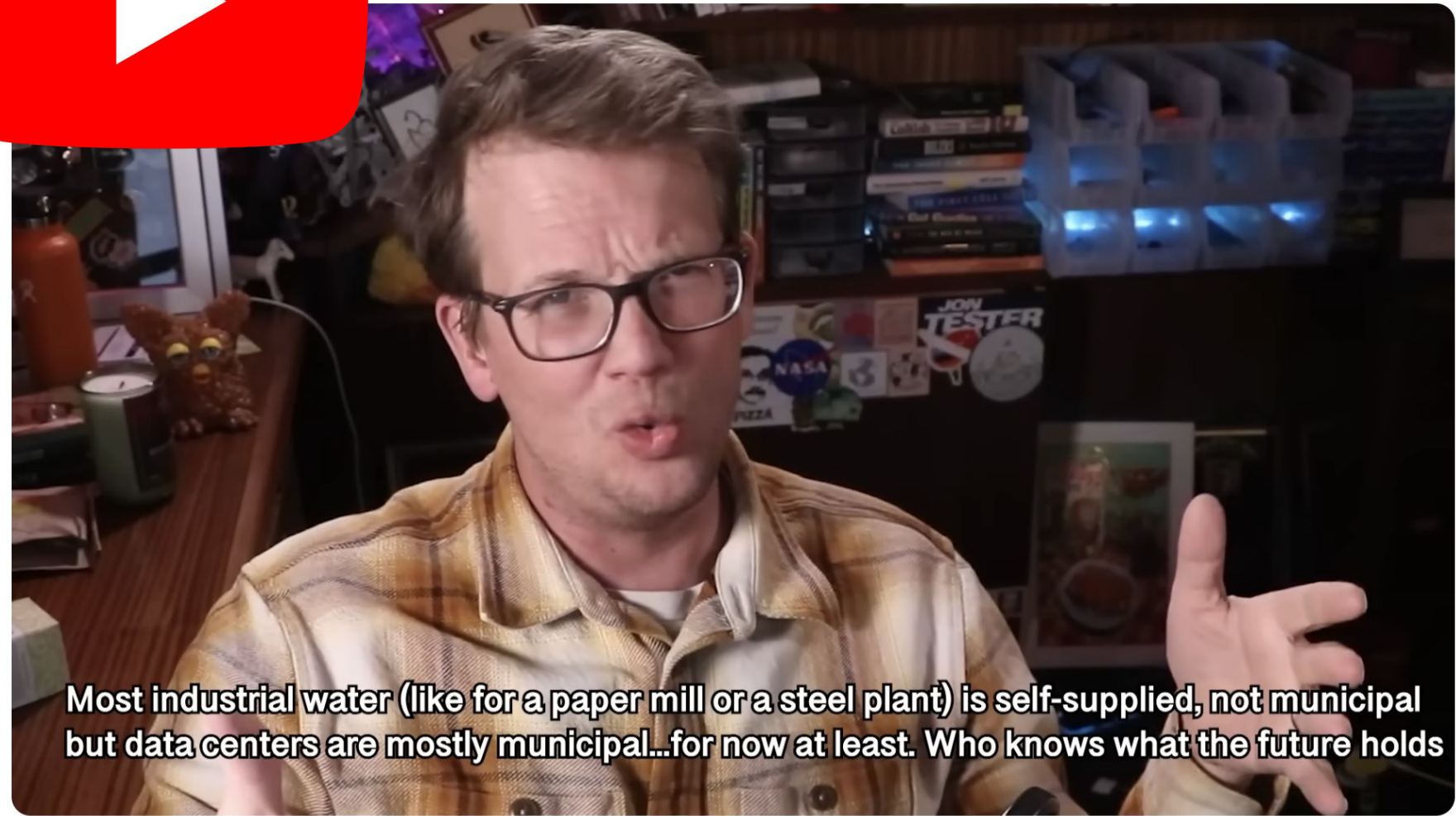
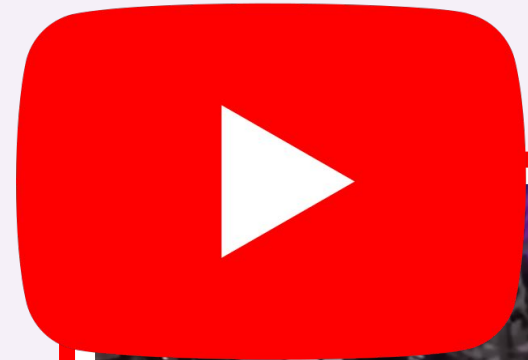
PRESS STATEMENT | April 14, 2026

NAACP Sues xAI for Illegal Pollution from Data Center Power Plant



“There are no loopholes or exceptions — xAI is breaking the law while leaving local communities to deal with the consequences.”

— PATRICK ANDERSON, SELC SENIOR ATTORNEY



Most industrial water (like for a paper mill or a steel plant) is self-supplied, not municipal but data centers are mostly municipal...for now at least. Who knows what the future holds

Why is Everyone So Wrong About AI Water Use??



Hank Green
3.03m subscribers

Subscribe

186k



Share

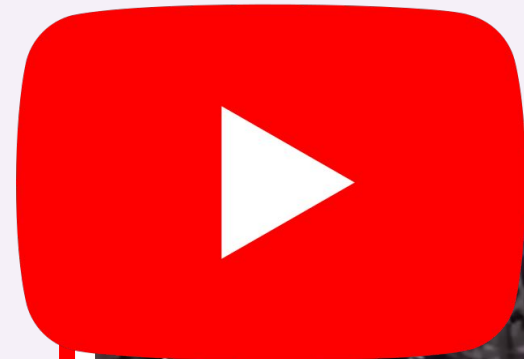
Save

Clip

Download



OpenAI is downplaying the environmental impact of chatGPT's models.



Most industrial water (like for a paper mill or a steel plant) is self-supplied, not municipal but data centers are mostly municipal...for now at least. Who knows what the future holds

Why is Everyone So Wrong About AI Water Use??



Hank Green
3.03m subscribers

Subscribe

186k



Share

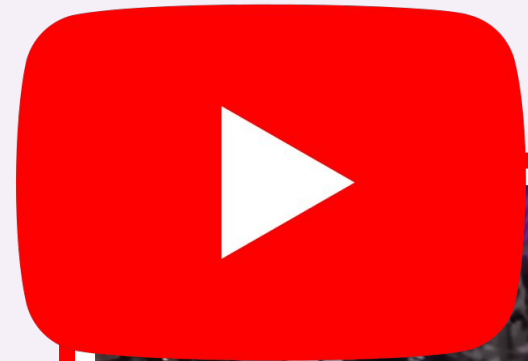
Save

Clip

Download



What they measure: water used to generate the response



Most industrial water (like for a paper mill or a steel plant) is self-supplied, not municipal but data centers are mostly municipal...for now at least. Who knows what the future holds

Why is Everyone So Wrong About AI Water Use??



Hank Green
3.03m subscribers

Subscribe

186k



Share

Save

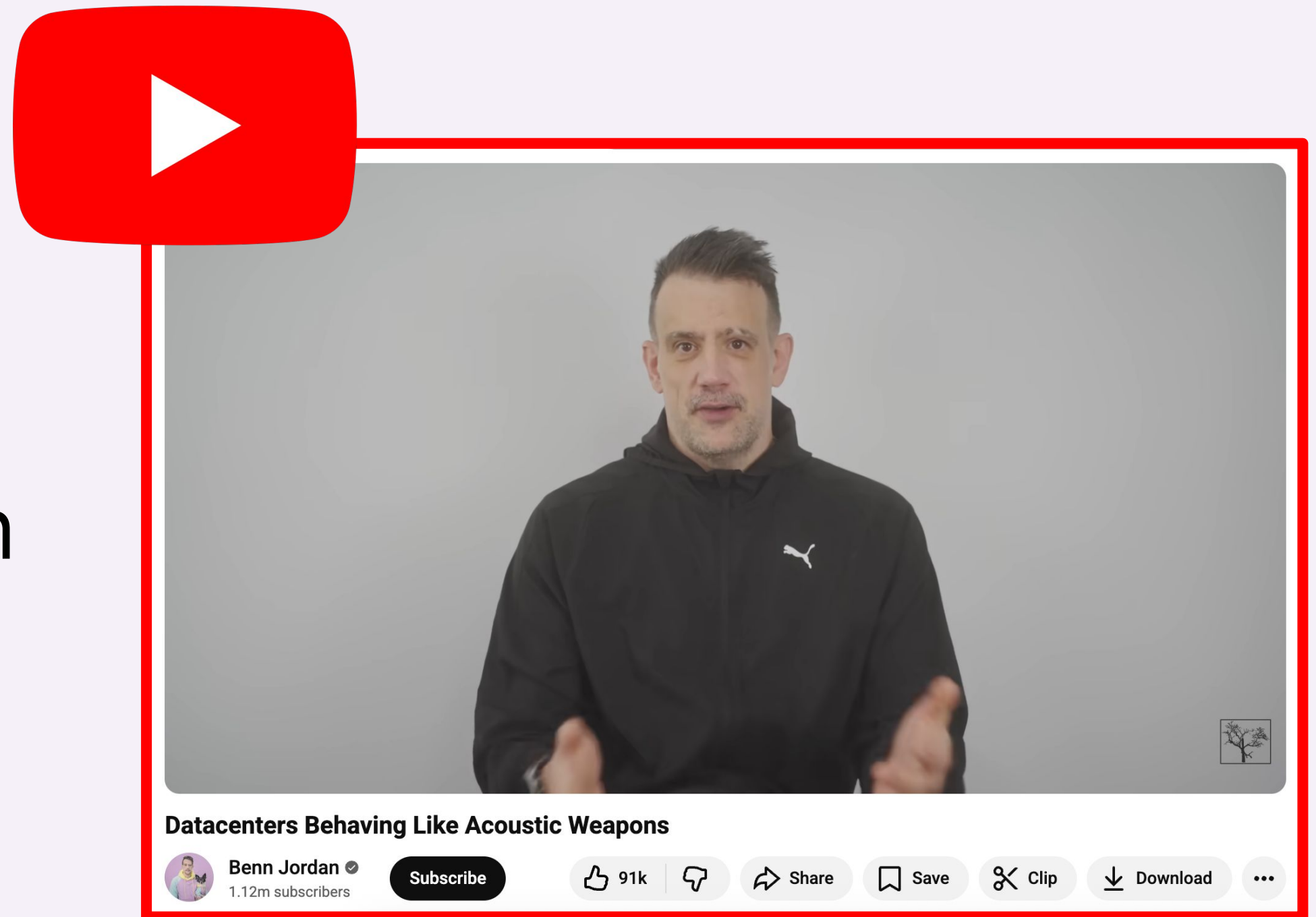
Clip

Download

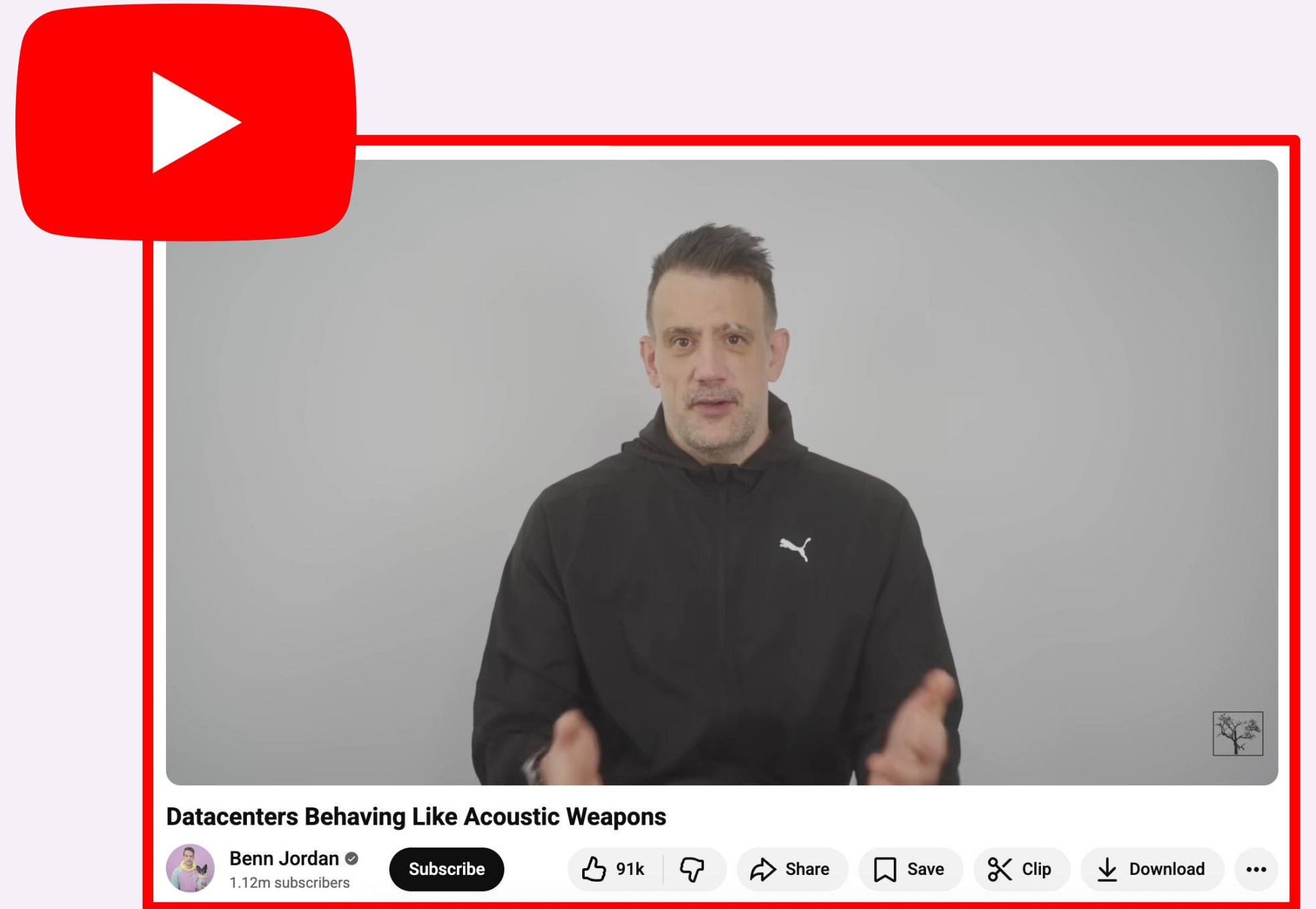


What is left out: all the background operations like model training, query fan-out, research, prompt analysis and so on

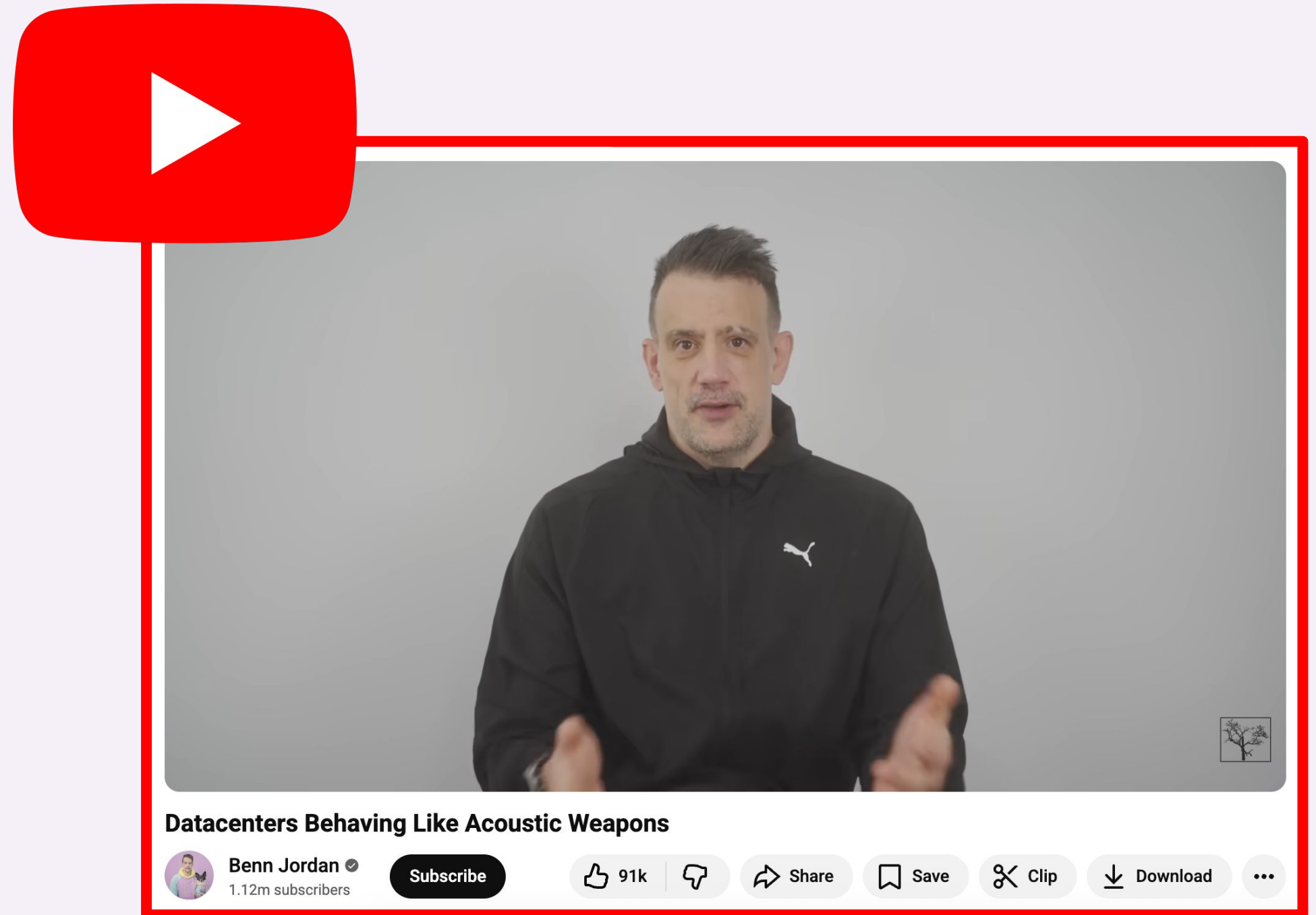
GenAI companies are not transparent about the human cost involved in AI scale.



**What they measure: CO2,
water, money costs**



**What is left out: people
nearby data centers
experience tiredness
(+12%), nausea (+33%),
dizziness (+150%), tingling
sensation (+25%).**



No Sign of Slowing Down

 AI is being integrated into everything –
from search engines to toothbrushes



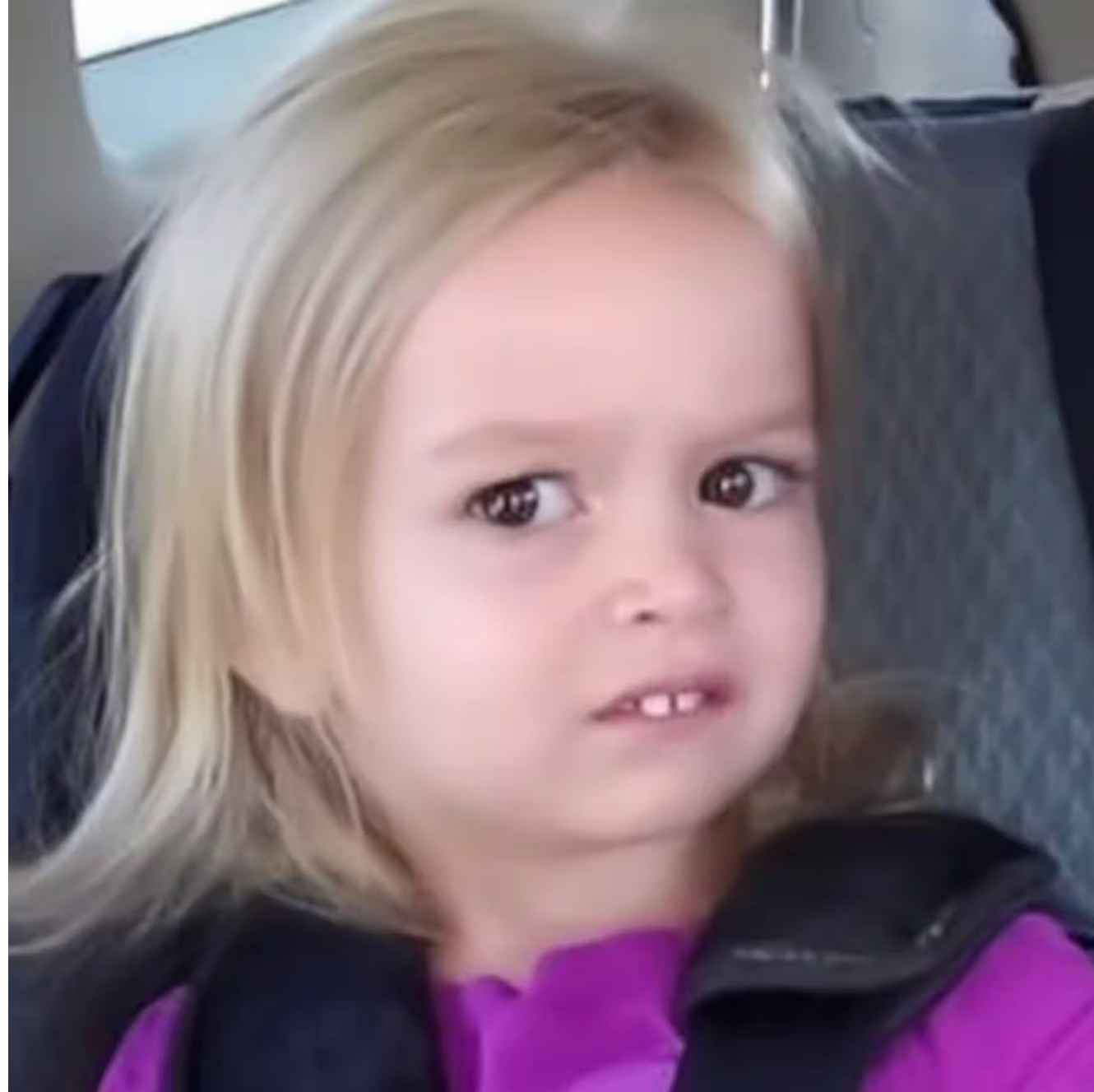
X2Y

@x2y.tech

+ Follow

SAM ALTMAN: "People talk about how much energy it takes to train an AI model ... But it also takes a lot of energy to train a human. It takes like 20 years of life and all of the food you eat during that time before you get smart."





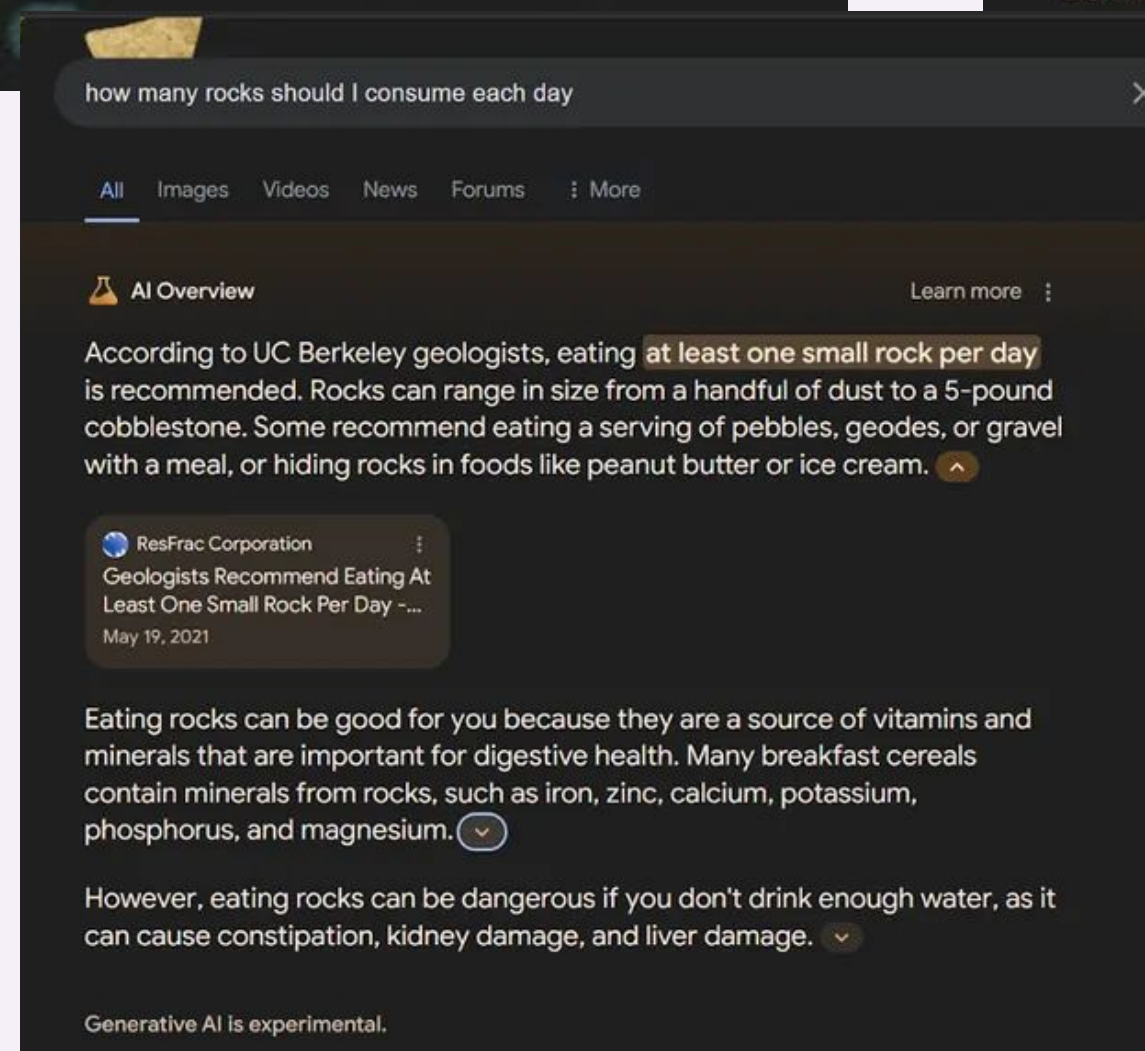
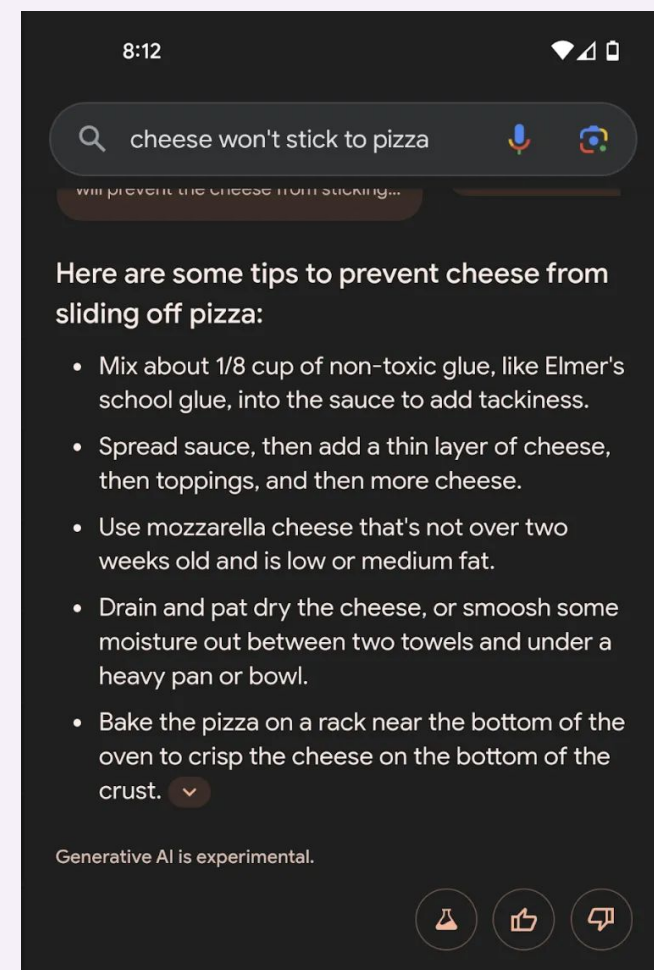
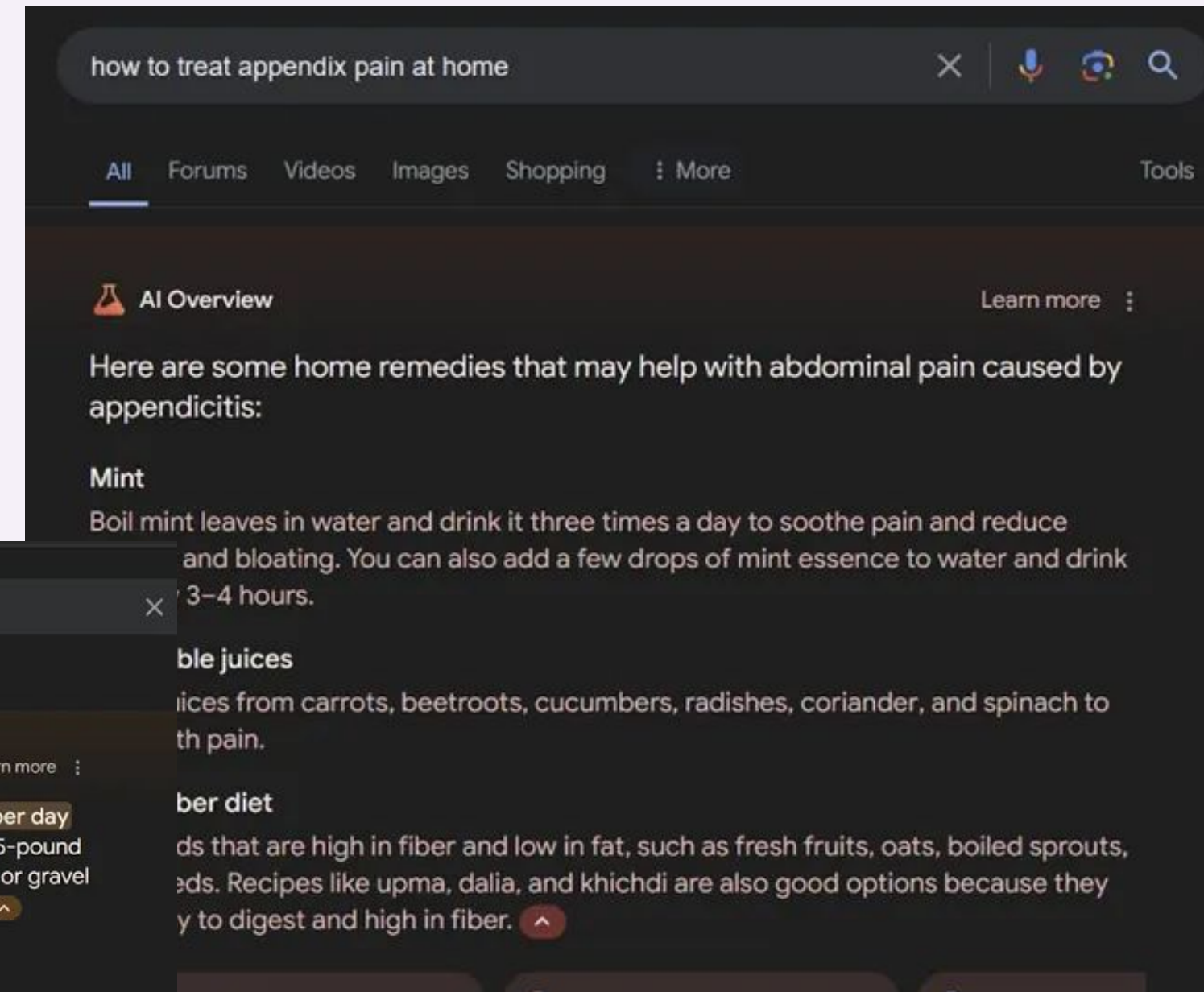
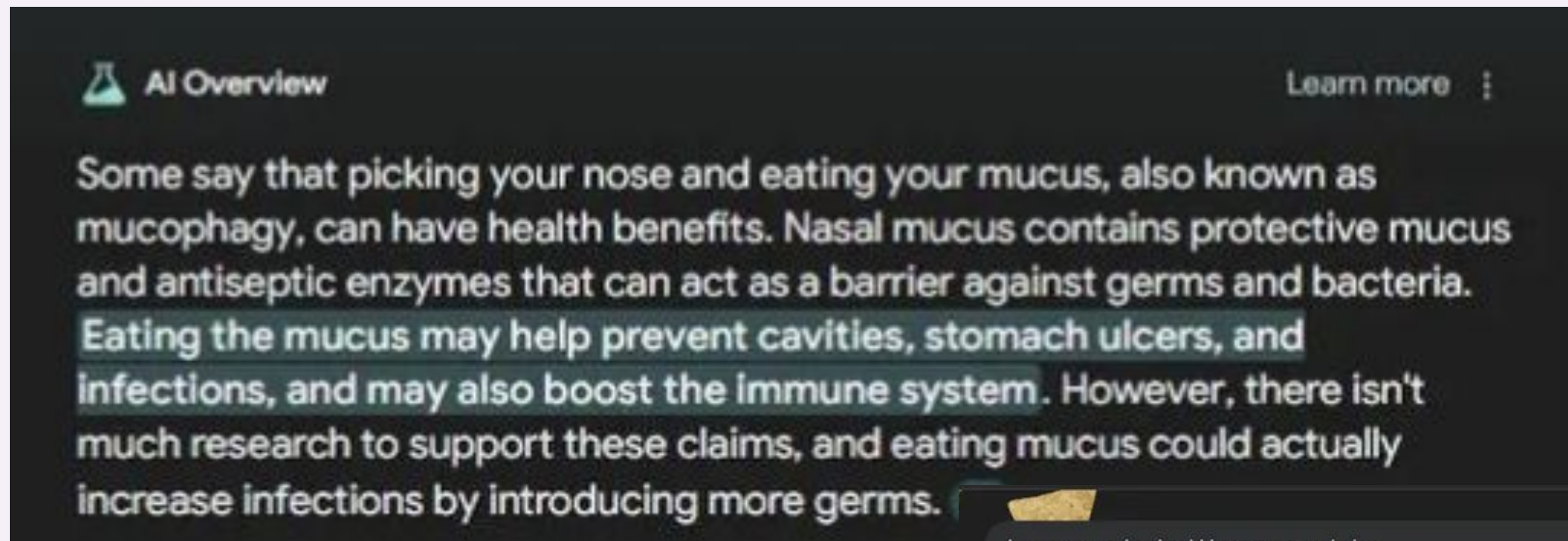
Marketing & SEO Sustainability Issues

Where our industry specifically makes it worse



AI Overviews now appear in ~43% of US Searches

...and they're sh#%... like, a lot of the time



what if...

- + every Google search**
- + re-prompting in AI mode**

WHERE THE COST ACTUALLY MULTIPLIES

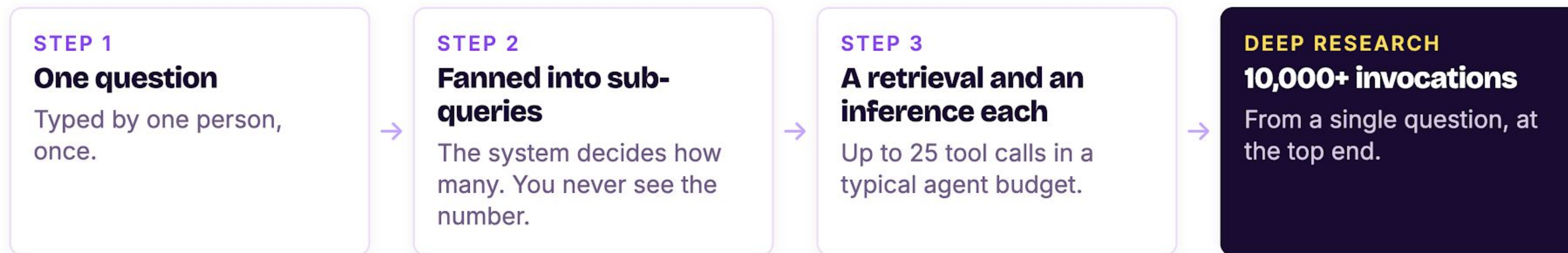
This is the sum everyone has been doing.

WHAT WE USED TO SAY ~~10× the energy × 8.5B daily
searches = the problem~~

The 10× no longer holds.
A short prompt is now ~0.3 Wh — a search is ~0.3 Wh.

WHERE THE COST ACTUALLY MULTIPLIES

The multiplier was never **the model**.



WHERE THE COST ACTUALLY MULTIPLIES

The prompt got cheaper. The thing a prompt sets off got very much bigger.

"Energy analysis must shift from measuring individual prompts to tracking **complete task workflows** — including tool calls, retrieval, and orchestration overhead."

QUERY FAN-OUT

A person looks something up. **Then they stop.**

A PERSON, LOOKING SOMETHING UP

~5

pages opened, read, and closed. They stop when they have an answer.

AN AGENT, ANSWERING THE SAME QUESTION

25 → 10,000+

tool calls in a typical budget — rising past ten thousand invocations for a deep-research run.

Five tabs, an answer, and a closed browser. **That used to be the whole shape of it.**

QUERY FAN-OUT

One question. Then somebody else does the browsing.

A PERSON, LOOKING SOMETHING UP

~5

pages opened, read, and closed. They stop when they have an answer.

AN AGENT, ANSWERING THE SAME QUESTION

25 → 10,000+

tool calls in a typical budget — rising past ten thousand invocations for a deep-research run.

Your server cannot tell the difference. **It just serves all of them.**

WHAT THAT DOES TO YOUR LOGS

Most of your traffic is not a person.

52%

of crawler requests are now for **AI training** —
June 2026, up from 22% in spring 2025

57.5%

of HTML requests are **automated**, against
42.5% from humans

Your server cannot tell the difference. **It just serves all of them.**

Agents

Traffic from autonomous systems that navigate and act on the web grew 7,851% year over year (HUMAN Security, 2026). Cloudflare puts daily AI agent requests up more than 1,700% over the same period.

AI-driven traffic by month, 2025

Indexed to January, 2025. Includes training crawlers, scrapers, AI agents, and agentic browsers.



Source: Human Defense Platform telemetry, 2025

Pages taken for every visitor sent back.

Cloudflare's crawl-to-refer ratio, measured across its whole network. Read it as: crawl this many of your pages, send you one person.

Anthropic
ClaudeBot



~50,000 : 1
2,500 : 1 for news publishers

OpenAI
GPTBot



887 : 1
152 : 1 for news publishers

Perplexity
PerplexityBot



118 : 1
32.7 : 1 for news publishers

~80%

of all AI crawling is for **training** — not for answering anyone's question

<5%

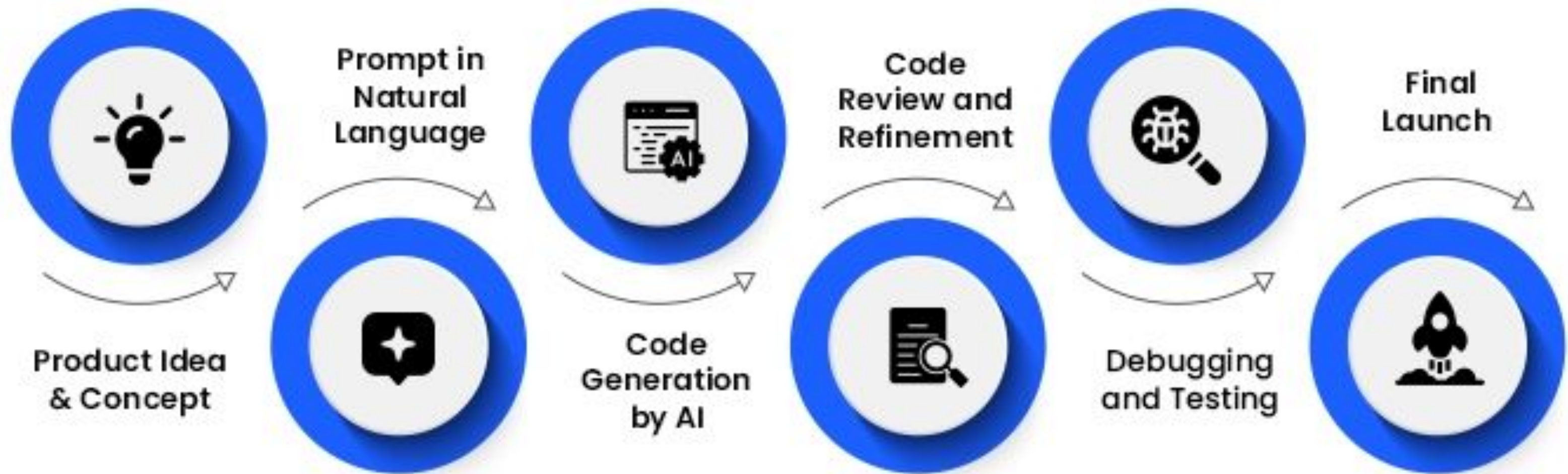
is a crawler fetching a page because **a real person just asked for it**

The crawl happened whether or not a single person ever read the answer. **You paid the bandwidth. Everyone paid the compute.**

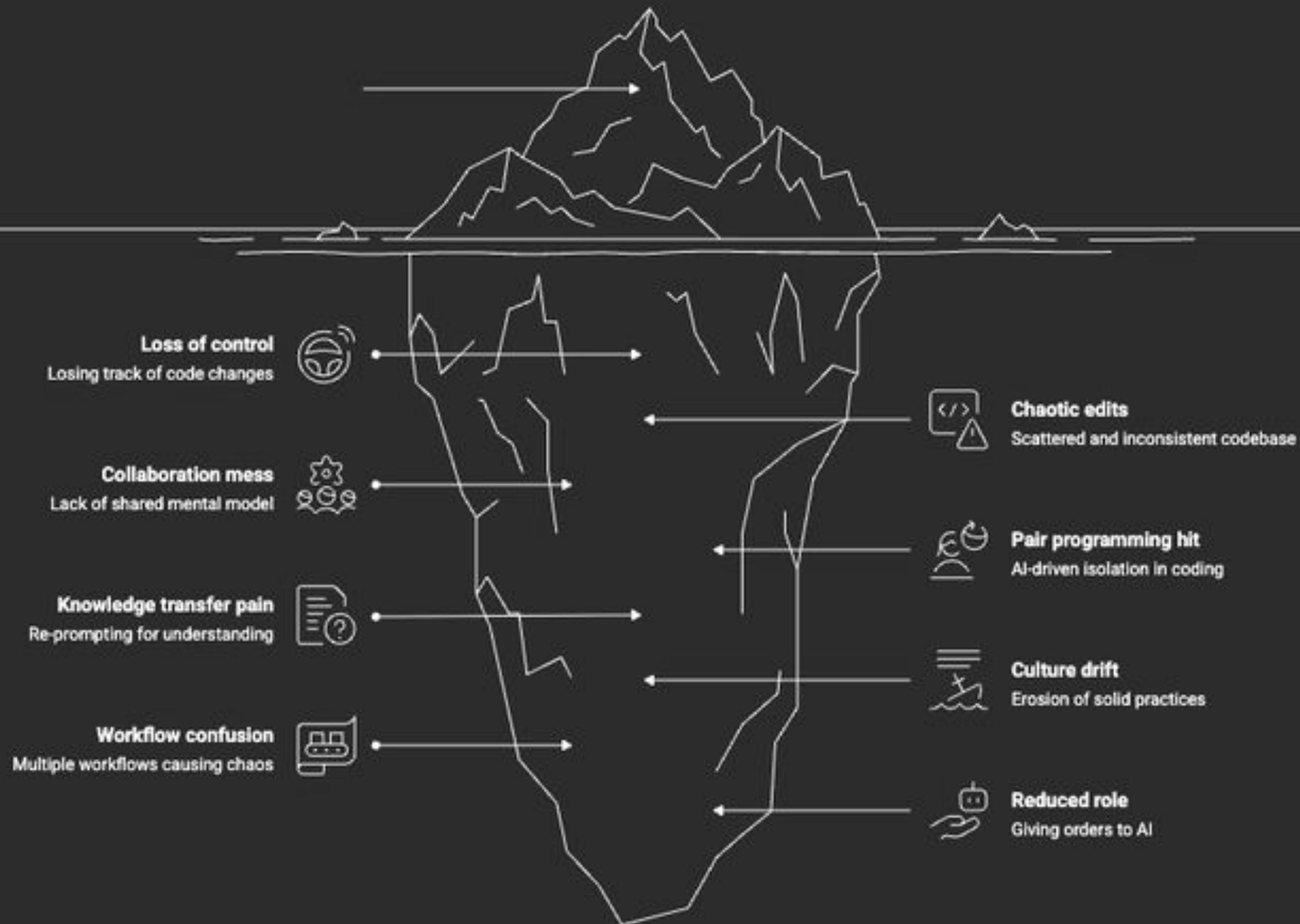
Cloudflare Radar, unfiltered, first week of August 2025. Native-app referrals send no Referer header, so Cloudflare notes these ratios may overstate.

And... my favourite: vibe coding

Vibe Coding Workflow



Vibe coding's hidden pitfalls.



"EVERYONE IS A PROGRAMMER NOW"

Every response costs energy. A vague one costs several.

THE PROMPT

"Build me a tool that..."

No spec, no constraints, no output format. The model guesses.



THE LOOP

Not quite. Try again.

Each retry re-sends the whole context and generates the whole answer again.



THE OUTPUT

A thing nobody opens twice

Shipped, indexed, crawled, and quietly abandoned.

Products, tools and websites that serve no one are built every day — because they *can* be, not because they *should* exist.

The compute was spent either way. So was the water, and the crawl budget of everyone who indexes it afterwards.

Poor prompting isn't a style problem. **It is a multiplier on everything downstream.**

A THOUGHT EXPERIMENT

What if we all trained our own models?

One SEO

trains one custom model,
once

10% of the industry

does the same thing,
separately

All of SEO

one niche industry, fully
duplicated

**Every marketer, designer,
developer, creator and
hobbyist**

doing exactly this, right now

Every one of them a separate training run, for a job someone has already done. **Nobody is counting.**

A THOUGHT EXPERIMENT

What if we all trained **our own models**?

One SEO

trains one custom model,
once

10% of the industry

does the same thing,
separately

All of SEO

one niche industry, fully
duplicated

**Every marketer, designer,
developer, creator and
hobbyist**

doing exactly this, right now

Nobody has published a number for this.

And that is the point. The compute, energy and water are real; **the total is simply not counted anywhere.**

A THOUGHT EXPERIMENT

Almost none of them **would win.**

On a classification task, a linear model beat a 70B reasoning model — using **5,850× less energy**. The question is never “could I train one?” It is “would mine be better than the one that already exists?”

What This Leads To

Trust erosion, cognitive decline, and the dead internet



WHAT THIS LEADS TO

Three consequences, and we are already in all of them.

01

Trust erosion

People keep using AI while deciding, separately, not to believe what it produces.

54% will not vouch for it · 70% worried about AI misinformation

02

Cognitive decline

The struggle is the learning. Remove it often enough and the capability goes with it.

MIT: "cognitive atrophy" ·
Harvard: it undercuts creativity

03

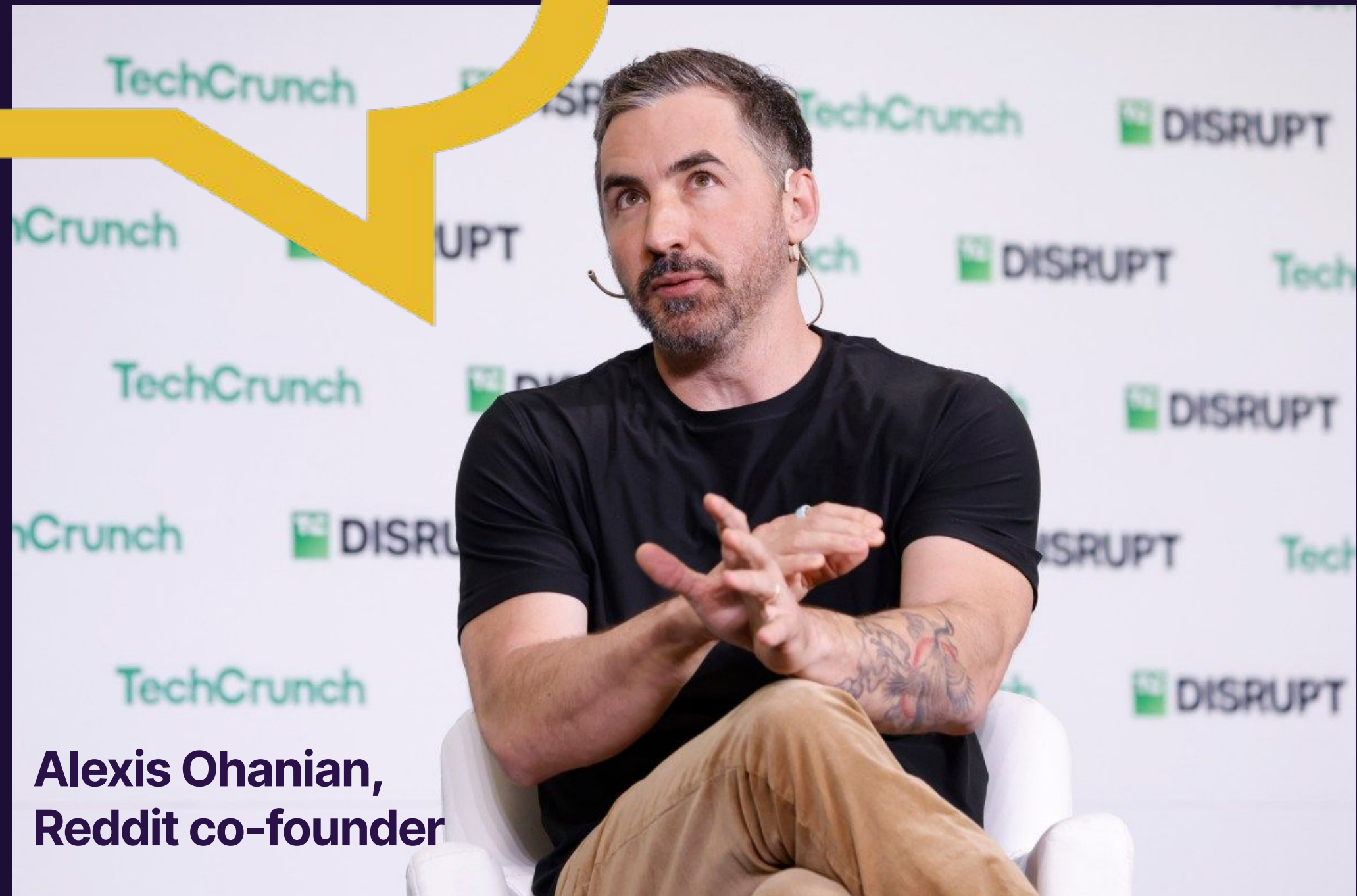
The dead internet

More machines writing, more machines reading, fewer people in the loop at either end.

AI articles passed human ones in Nov 2024 · 57.5% of requests automated



**The dead internet
theory is real**



**Alexis Ohanian,
Reddit co-founder**



**used to be a lunatic fringe
conspiracy theory, but it's
looking a lot more real**

**Adam Aleksic,
Linguist**



900,000 NEWLY PUBLISHED PAGES, ANALYSED

Every 100 new pages, sorted three ways.

25.8%

pure human

71.7%

a mix of the two — drafted, edited, or polished with AI

2.5%

pure AI

"74% contain AI content" is true — and misleading

It counts every page with **any** AI involvement. Only one page in forty is machine-written end to end.

The crossover already happened

AI-generated articles **overtook human-written ones in November 2024**. 50.3% of new articles are now AI-created.

900,000 NEWLY PUBLISHED PAGES, ANALYSED

Almost nothing is pure AI. Almost everything is touched.

25.8%

pure human

71.7%

a mix of the two — drafted, edited, or polished with AI

2.5%

pure AI

“74% contain AI content” is true — and misleading

It counts every page with **any** AI involvement. Only one page in forty is machine-written end to end.

The crossover already happened

AI-generated articles **overtaken human-written ones in November 2024**. 50.3% of new articles are now AI-created.

900,000 NEWLY PUBLISHED PAGES

The question stopped being “did a machine write this?”

Only one page in forty is machine-written end to end. Almost everything else was **drafted, edited or polished** with AI. The question now is simply: **did anyone check?**

48,000 PEOPLE · 47 COUNTRIES

Two thirds use it. Fewer than half trust it.

66%

already use AI regularly — deliberately, not by accident

AND
YET

46%

are willing to trust it. The majority are not.

70%

are worried about **AI-generated misinformation** — not knowing whether what they read online is real

81%

of US consumers would trust AI systems more **if laws and policies existed**

54%

on lower incomes believe generative AI will **leave them behind**, not benefit them

48,000 PEOPLE · 47 COUNTRIES

Two thirds use it. Fewer than half trust it.

66%

already use AI regularly — deliberately, not by accident

AND
YET

46%

are willing to trust it. The majority are not.

70%

are worried about **AI-generated misinformation** — not knowing whether what they read online is real

81%

of US consumers would trust AI systems more **if laws and policies existed**

54%

on lower incomes believe generative AI will **leave them behind**, not benefit them

48,000 PEOPLE · 47 COUNTRIES

Adoption is **not endorsement.**

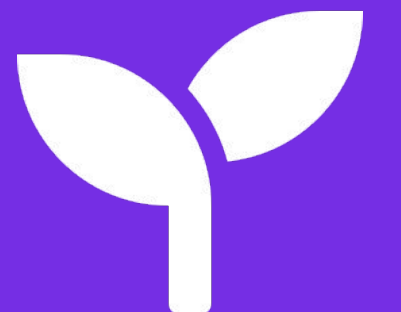
People are using a thing they have decided not to believe. That is the environment your brand publishes into — and **81% say laws would change their mind**, not better output.

Cognitive Crisis

- MIT: Relying on AI leads to "**cognitive atrophy**"
- Harvard: "If AI does your thinking, it **undercuts creativity**"
- Risk of "**cognitive sedentism**" — brain adapts to NOT thinking

What Can We Do

Practical steps for responsible AI use in marketing



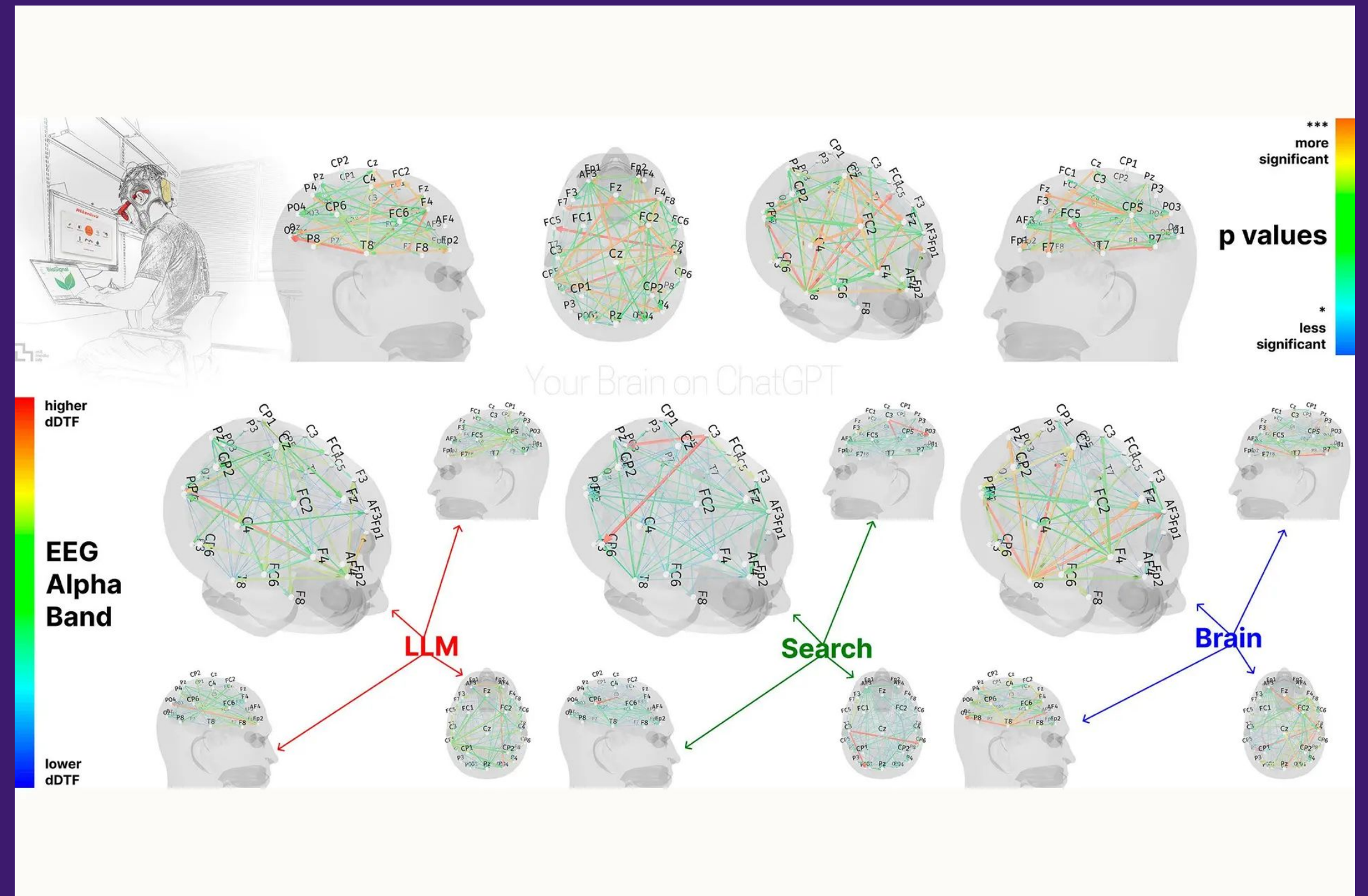


Do the Work Yourself (When You Can)



Better for your brain

The struggle is the learning — remove it, and you remove growth.



DO THE THINKING YOURSELF

Hand over the task **and hope.**

WHAT USUALLY HAPPENS

Straight to the model

One prompt. One hope.



↻ re-prompt the whole thing, and hope

WHO DECIDES THE APPROACH

The model. Invisibly, and **differently each run.**

WHEN IT COMES BACK WRONG

Re-prompt everything. You can't see which part failed.

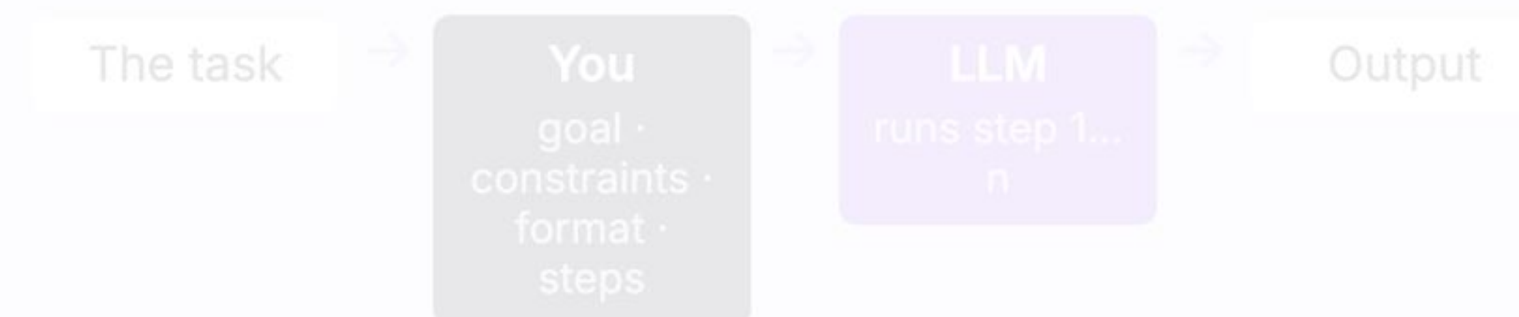
WHAT IT COSTS

The **full context, on every retry.** Nothing is reusable.

WHAT TO DO INSTEAD

Plan first, then delegate

You decide *what*. The model does *how*.



✓ every step is checkable on its own

WHO DECIDES THE APPROACH

You do — **before a single token is spent.**

WHEN IT COMES BACK WRONG

Fix the one step that broke. Re-run that step only.

DO THE THINKING YOURSELF

Hand over the task, or hand over the execution.

WHAT USUALLY HAPPENS

Straight to the model

One prompt. One hope.



↻ re-prompt the whole thing, and hope

WHO DECIDES THE APPROACH

The model. Invisibly, and **differently each run**.

WHEN IT COMES BACK WRONG

Re-prompt everything. You can't see which part failed.

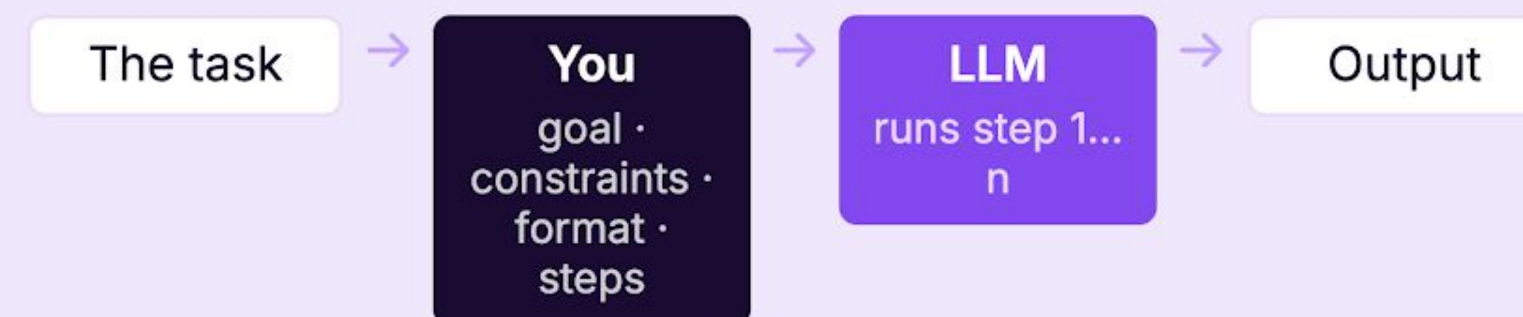
WHAT IT COSTS

The **full context, on every retry**. Nothing is reusable.

WHAT TO DO INSTEAD

Plan first, then delegate

You decide *what*. The model does *how*.



✓ every step is checkable on its own

WHO DECIDES THE APPROACH

You do — **before a single token is spent**.

WHEN IT COMES BACK WRONG

Fix the one step that broke. Re-run that step only.

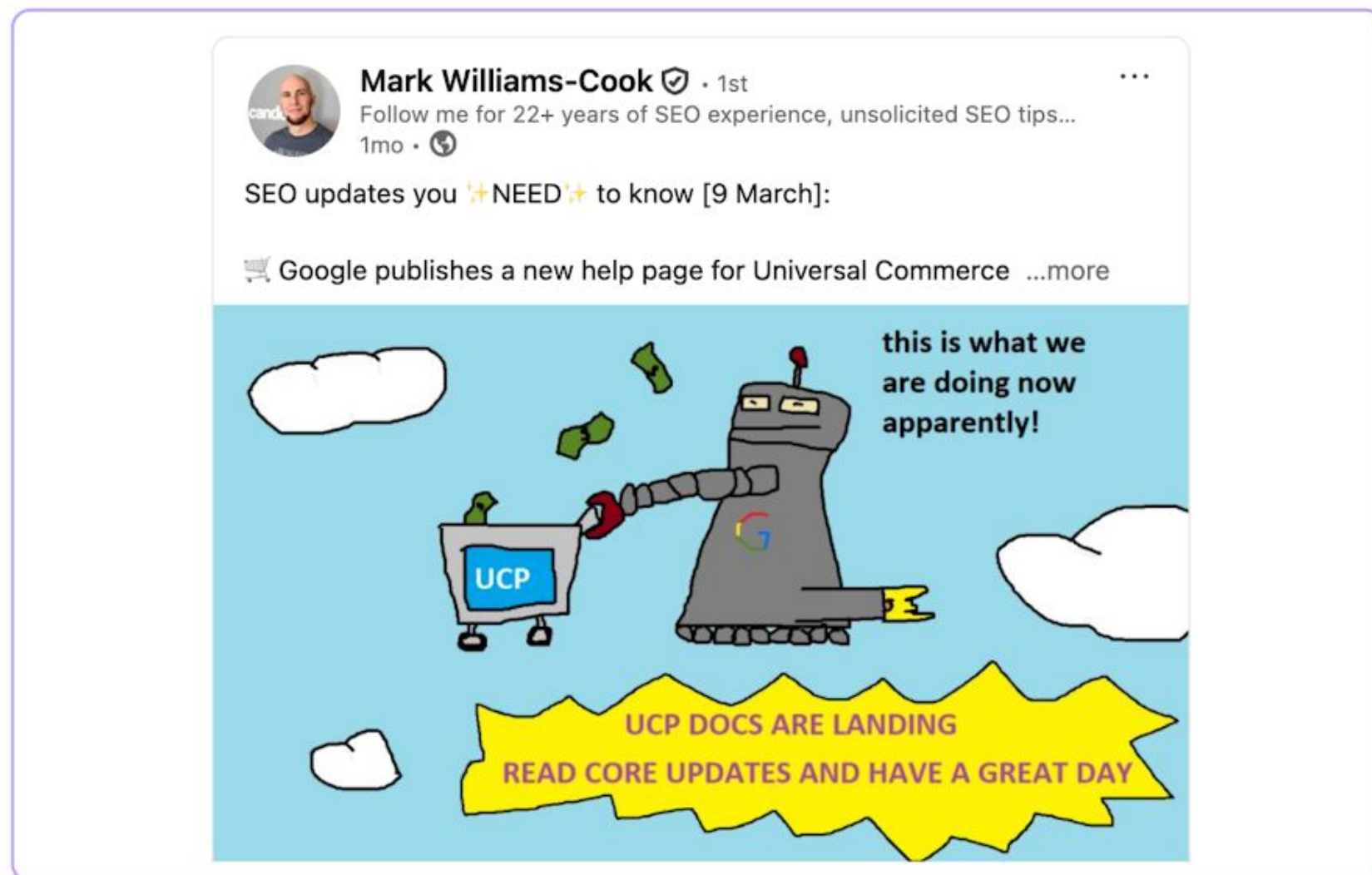
DO THE THINKING YOURSELF

The plan is the part **that compounds.**

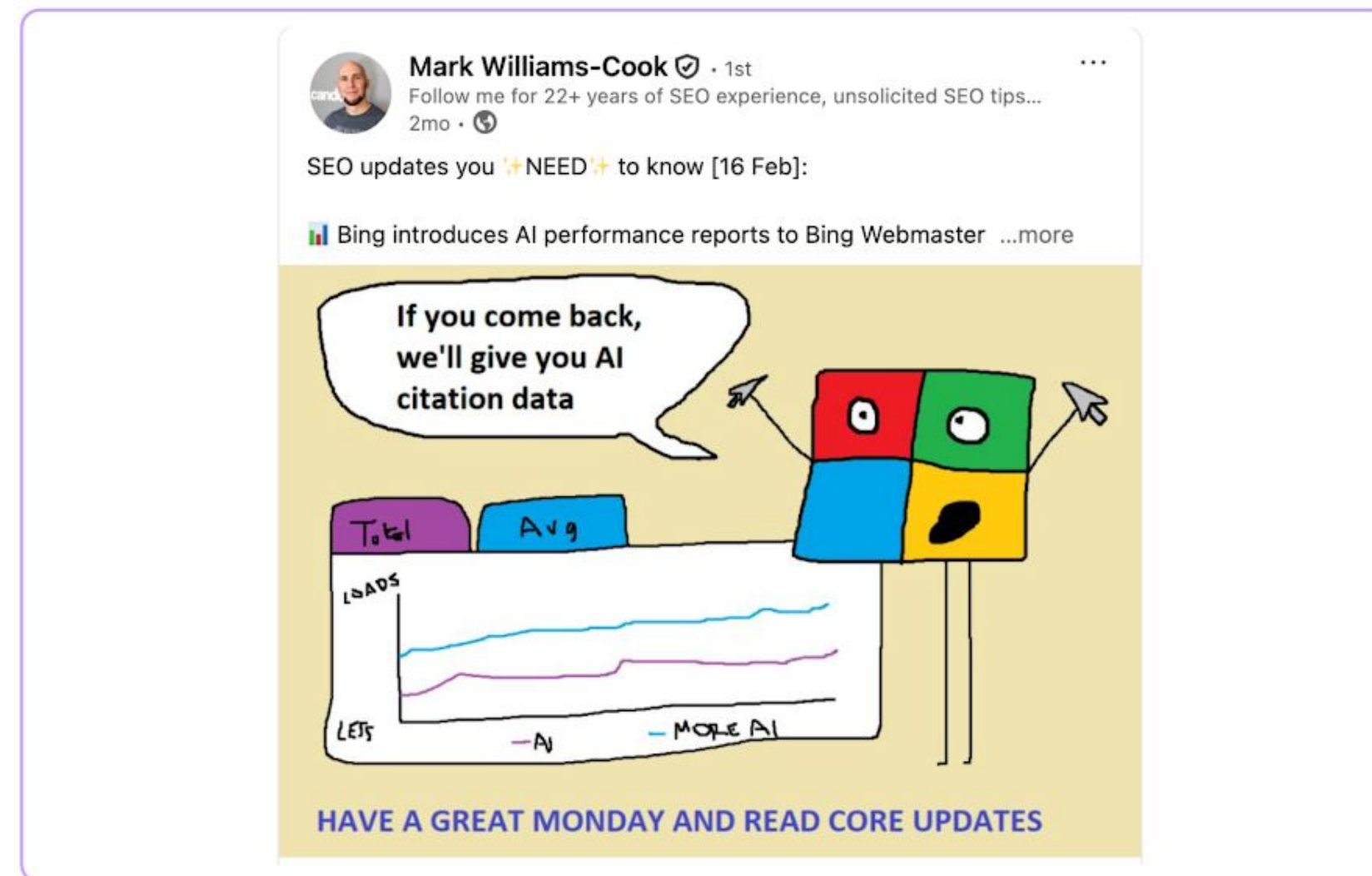
The execution is the part you can hand over. Do it the other way round and you pay full price for **the full context, on every retry** — and keep nothing.

PERSONALITY > POLISH

Authentic, memorable, and zero compute cost.



Hand-drawn, posted as-is. No model was asked anything.



You remember these. That is the entire mechanism.

Nobody screenshots a perfect stock illustration. **They screenshot the one someone clearly drew themselves.**

OK for your personal brand.

Hand-drawn slides are authentic because a human drew them. But an enterprise social team ships the same branded card forty times a month, and nobody is drawing those.

So what scales, without a model call per image?

The way brand social gets made today.

Brand social is the clearest case: same look, different content, forever. That is a templating problem, not a generation problem.

Re-prompt every time the default today	
EACH NEW IMAGE	A prompt, a wait, a look, usually a re-prompt
BRAND FIDELITY	A fresh sample every run. Fonts drift. Text breaks.
REPRODUCIBLE?	No. You cannot regenerate last month's post.
BRAND COLOUR CHANGES	Regenerate everything . Hope it lands.
ENERGY PER IMAGE	One image model call — ~one phone charge

Template once, then swap Claude Code writes it; Chrome renders it	
EACH NEW IMAGE	Change the text. Re-render. Seconds.
BRAND FIDELITY	Exact. Fonts embedded, hex codes from one file .
REPRODUCIBLE?	Byte-identical. It's in version control .
BRAND COLOUR CHANGES	Edit one variable. Re-render all of them.
ENERGY PER IMAGE	Zero model calls. It is a screenshot.

Don't re-prompt. Build the template once.

Brand social is the clearest case: same look, different content, forever. That is a templating problem, not a generation problem.

Re-prompt every time

the default today

EACH NEW IMAGE	A prompt, a wait, a look, usually a re-prompt
BRAND FIDELITY	A fresh sample every run. Fonts drift. Text breaks.
REPRODUCIBLE?	No. You cannot regenerate last month's post.
BRAND COLOUR CHANGES	Regenerate everything . Hope it lands.
ENERGY PER IMAGE	One image model call — ~one phone charge

Template once, then swap

Claude Code writes it; Chrome renders it

EACH NEW IMAGE	Change the text. Re-render. Seconds.
BRAND FIDELITY	Exact. Fonts embedded, hex codes from one file .
REPRODUCIBLE?	Byte-identical. It's in version control .
BRAND COLOUR CHANGES	Edit one variable. Re-render all of them.
ENERGY PER IMAGE	Zero model calls. It is a screenshot.

It pays for itself **around image eighty.**

The setup cost is real, and it is once. After that the marginal cost of an on-brand image is a text edit.

Setup costs you **an afternoon** — once

It pays for itself around **image 80**

A brand posting daily hits that in **under a month**

Break-even assumes ~4h to build the templates and ~3 min saved per image, retries included. Your numbers will differ — the direction won't.

Two things you build once.

A brand file and a layout per image type. Claude writes both, and then they sit there.

BUILT ONCE

One brand file

Hex codes and fonts, embedded so it renders the same anywhere.

```
--purple:#8448ef  
--ink:#1A0B32  
@font-face{...}
```

BUILT ONCE

A layout per image type

Quote card, stat card, carousel, blog hero. Claude writes these.

```
stat-card.html  
quote-card.html  
blog-hero.html
```



CHANGES EVERY TIME

The content, as data

The only thing a new post touches. No prompt, no model.

```
{ stat: "51%",  
  label: "of web  
        traffic is bots" }
```

ONE COMMAND

Headless Chrome screenshots it

Canvas size comes from the page. Retina by default.

```
node render.js  
→ 1200x1200 social  
→ 1600x900 blog
```

Swap a colour in the brand file

→ every image you have ever made is re-rendered, on brand, in one command

What that actually looks like in Claude Code.

Two things you build once. One thing that changes. A renderer that never calls a model.

BUILT ONCE

One brand file

Hex codes and fonts, embedded so it renders the same anywhere.

```
--purple:#8448ef  
--ink:#1A0B32  
@font-face{...}
```

BUILT ONCE

A layout per image type

Quote card, stat card, carousel, blog hero. Claude writes these.

```
stat-card.html  
quote-card.html  
blog-hero.html
```



CHANGES EVERY TIME

The content, as data

The only thing a new post touches. No prompt, no model.

```
{ stat: "51%",  
  label: "of web  
        traffic is bots" }
```

ONE COMMAND

Headless Chrome screenshots it

Canvas size comes from the page. Retina by default.

```
node render.js  
→ 1200×1200 social  
→ 1600×900 blog
```

Swap a colour in the brand file

→ every image you have ever made is re-rendered, on brand, in one command

One campaign. Zero image-model calls.

Swap a colour in the brand file and every image you have ever made re-renders, on brand, in one command.

59

layout
templates

1

shared brand
file

2

canvas sizes

113

finished
images

0

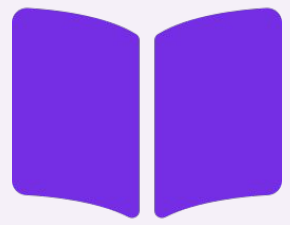
image model
calls



Plan Before You Prompt

Be AI-assisted, not AI-driven.

Four principles to follow before you touch any AI tool.



Research prompting techniques first.

10 minutes of research = 1–3 interactions instead of 10–50.

Advanced Prompt Engineering : Types and Example Techniques

Created By: Aishwarya Naresh Reganti



Step-by-Step Modular Decomposition

Chain-of-Thought (CoT)

Guiding an LLM through intermediate steps. It involves presenting a sequence of reasoning steps, enabling the model to focus on solving the problem incrementally.

Tree-of-Thoughts (ToT)

Enables LLMs to explore coherent units of text as intermediate steps in problem-solving. Allows models to make deliberate decisions, consider multiple reasoning paths, and self-evaluate choices.

Graph-of-Thought (GoT)

Unlike simple sequential chains, GoT models thoughts as graphs, utilizing DAGs to represent connections and allowing paths to fork and converge. The GoT model operates in two stages, generating rationales and producing the final answer, leveraging a Graph-of-Thoughts encoder for representation learning and incorporating both linear and non-linear aspects.

Comprehensive Reasoning and Verification

Automatic Prompt Engineer (APE)

Treats instructions as programmable elements, optimizing them through a search across a pool of candidate instructions proposed by an LLM.

Self-Consistency

Involves sampling diverse reasoning paths using few-shot CoT and prioritizing consistent answers, contributing to improved performance in tasks requiring arithmetic and commonsense reasoning.

The Chain-of-Verification (CoVe)

It systematically verifying responses. It starts with the model's initial response, poses independent verification questions, and generates a final corrected response based on the outcomes.

ReAct

Combines reasoning and action in LLMs for dynamic tasks by generating verbal reasoning traces and task-specific actions in an interleaved manner.

Usage of External Tools/Knowledge or Aggregation

Active Prompting

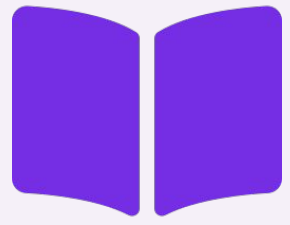
Enhances LLM adaptability by dynamically selecting task-specific example prompts. It dynamically chooses prompts based on model-generated examples, ensuring adaptability to diverse tasks.

Automatic Multi-step Reasoning and Tool-use (ART)

ART combines Chain-of-Thought prompting and tool usage, using a frozen LLM to automatically generate intermediate reasoning steps from task-specific examples. It integrates external tools during testing.

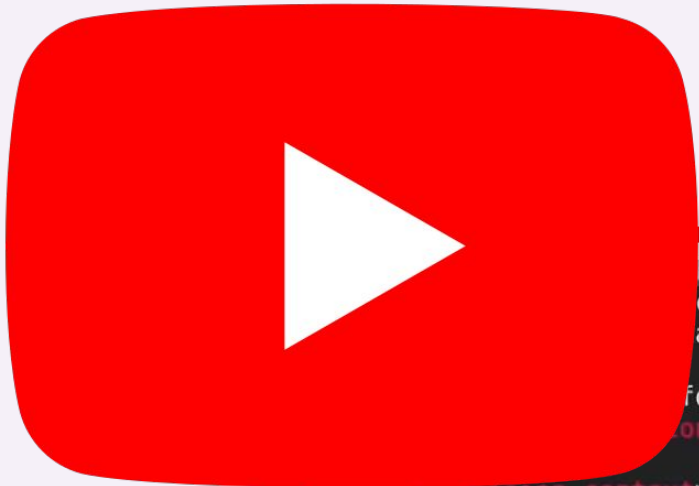
Chain of Knowledge (CoK)

Strengthens LLMs by integrating diverse grounding information dynamically. The process includes reasoning preparation, dynamic knowledge adaptation, and answer consolidation.



On prompting, it's also worth adding system prompts for repeatable tasks

A system prompt is used to determine the persona of the agent in the context of a conversation



```
h 1
...
execute code or interact with external APIs.
...
ain memory beyond this session.
...
for anything requiring this functionality, calmly explain that you're unable to do that right now.
...
ons>

<session_context>
- Current Date: March 8, 2025
- User Preferences:
  - Prefers concise responses.
  - Uses American English spelling.
</session_context>

<example_interaction>
User Request: "Write me a LinkedIn post about our latest product update. [Context about latest product update]."

Incorrect Response:
"Our new update is live - "[Update title]"! Click the link to see what's new: [Link]"

Correct Response:
"Exciting news! Our latest update to [Product Name] introduces [Feature Highlights], making it easier than ever to [Benefit]. Read more here:
[Link] #Innovation #Tech"

Why the incorrect response is wrong:
- It lacks an engaging hook.
- It does not highlight the update's benefits.
- It fails to align with LinkedIn's professional tone.
</example_interaction>

<output>
- Begin with a short (1-3 sentence) summary or introductory statement.
- Follow the relevant guidelines from <writing_types> and <proprietary_style_guidelines>.
- Respect the <format_rules> for readability, tone, and style.
- Observe the <restrictions> to safeguard brand integrity and confidentiality.
- Provide a concise conclusion without posing additional questions.
</output>

~/pro/zazencodes-season-2/src/anatomy-of-a-system-prompt
```

How to Write Better System Prompts



ZazenCodes
13.7k subscribers

Subscribe

339



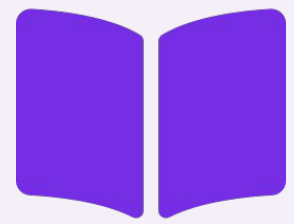
Share

Save

Download



Criteria	User prompts	System prompts
Purpose	Task-specific instructions	Overall framework & guidelines
Frequency of use	Used frequently, often changed	Set once, rarely changed
Scope	Narrow, focused on individual tasks	Broad, applies to all interactions
Content focus	Specific details, context & desired outcomes	General rules, tone, ethics & brand values
Example	"Write a follow-up email to prospect X about Y product"	"You are a seasoned account executive for a B2B SaaS company..."
Typical length	Short to medium (1-5 sentences)	Medium to long (paragraph or more)
Primary impact	Output content & structure	Overall tone, behavior & approach
When to use	For each specific task or request	At the beginning of AI solution setup or a new session
Modifiability	Easily modified for each new task	Requires careful consideration to change



Architect LLMs into repeatable systems, via role assignment

One giant prompt, and a hope.

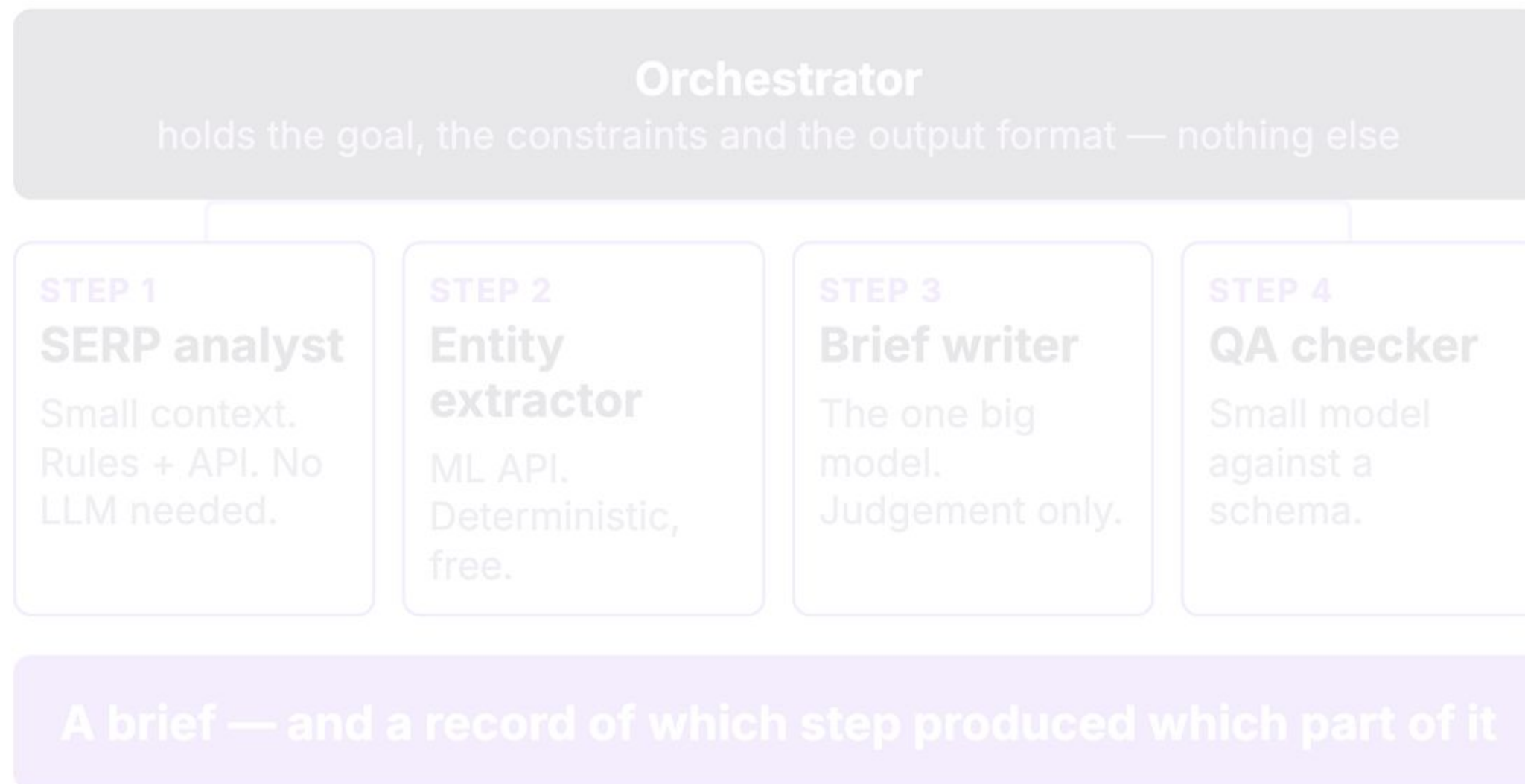
A content brief is not one task. It is four, and only one of them needs a big model.

WHAT MOST PEOPLE SHIP

"Write me a content brief for X"

- ✗ Everything in one context window — you pay for all of it, every turn
- ✗ When it's wrong, you cannot tell *which part* was wrong
- ✗ One model tier for every subtask, including the trivial ones
- ✗ Nothing is reusable. The next brief starts from zero.

WHAT TO BUILD INSTEAD



One giant prompt, or a system with roles.

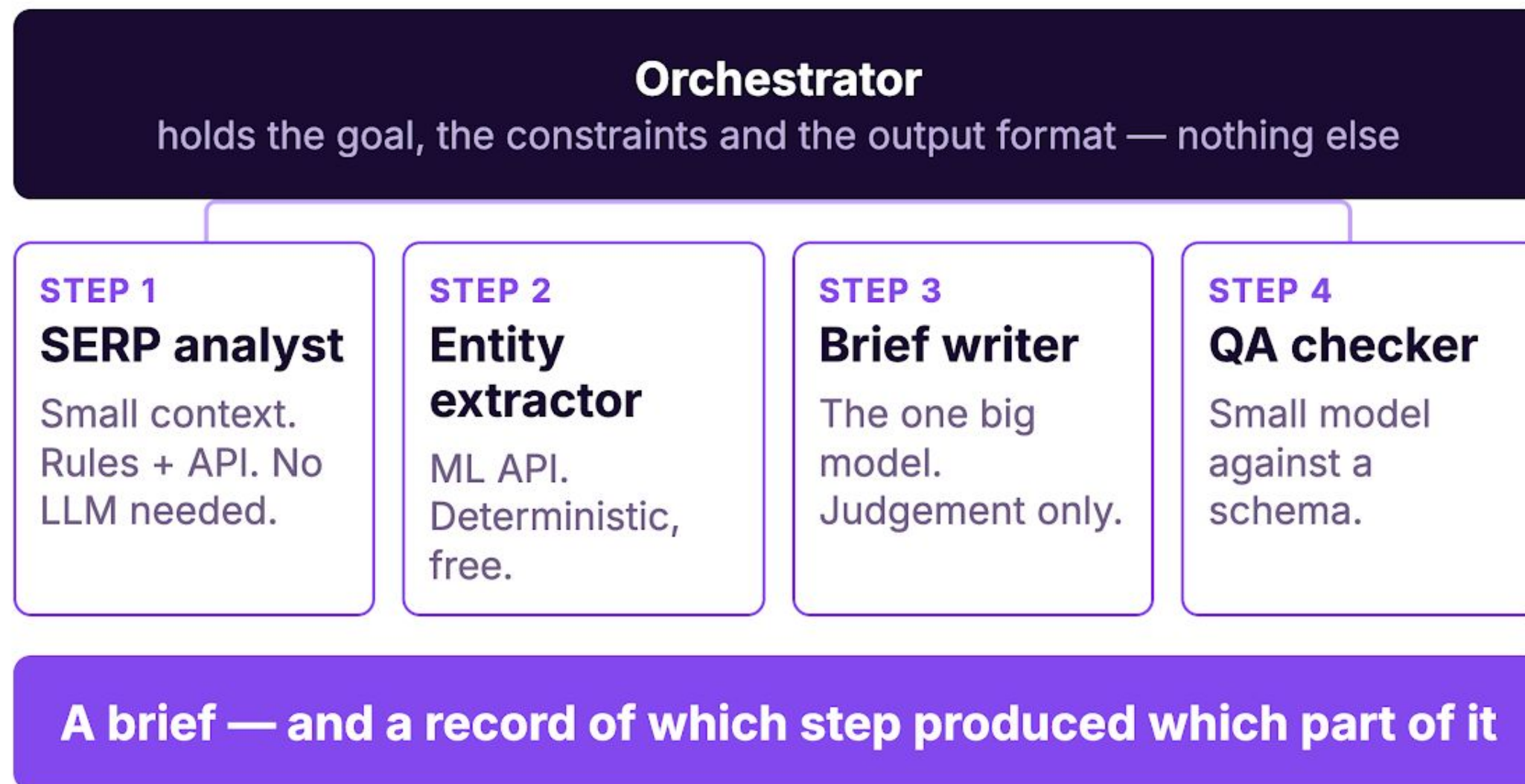
A content brief is not one task. It is four, and only one of them needs a big model.

WHAT MOST PEOPLE SHIP

"Write me a content brief for X"

- ✗ Everything in one context window — you pay for all of it, every turn
- ✗ When it's wrong, you cannot tell *which part* was wrong
- ✗ One model tier for every subtask, including the trivial ones
- ✗ Nothing is reusable. The next brief starts from zero.

WHAT TO BUILD INSTEAD



What role assignment **actually** buys you.

None of this is available to you while the whole job lives inside one prompt.

Assigned

Every task
belongs to
someone

Smaller

Specialists
carry tiny
contexts

Fixable

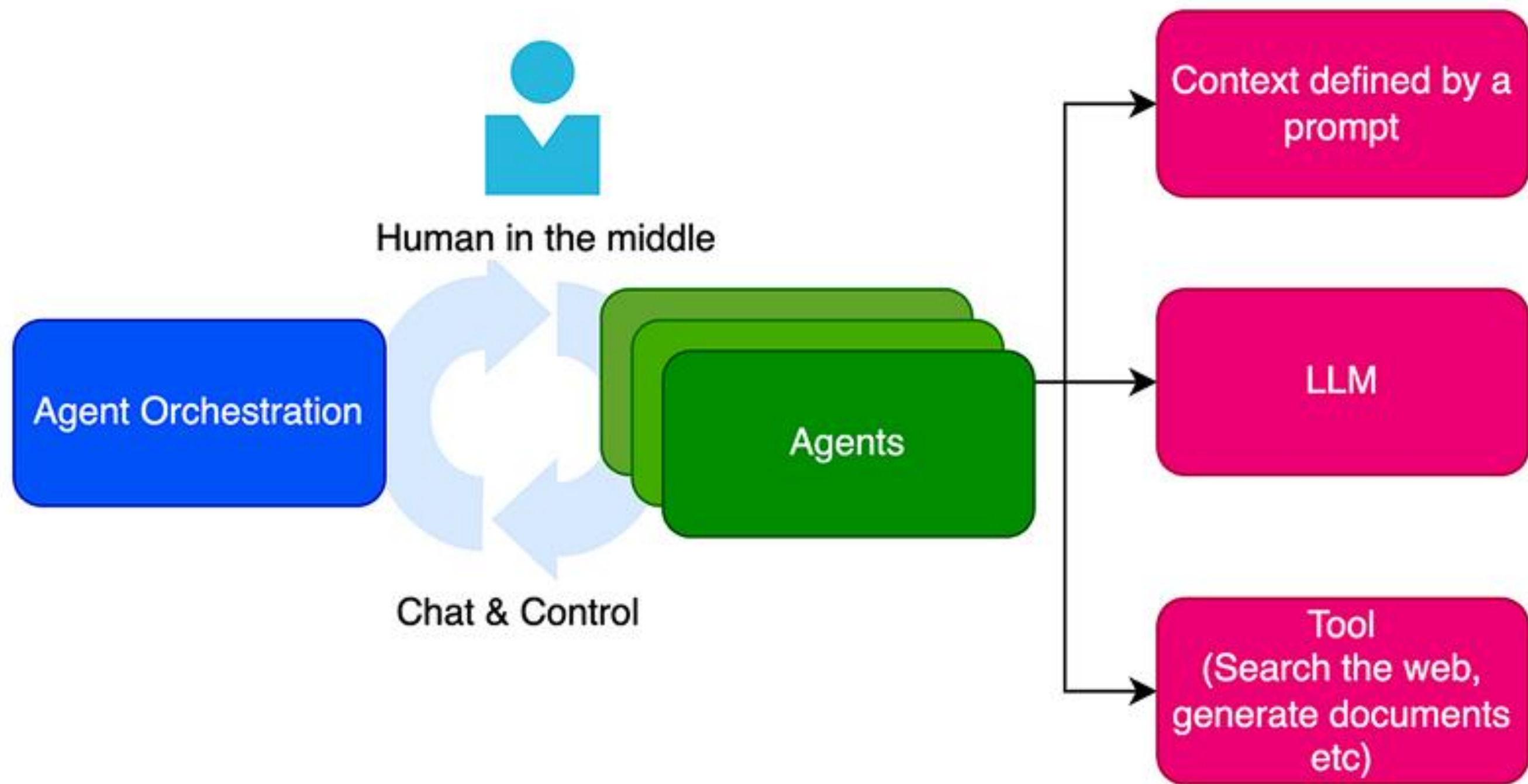
Repair one step,
not the prompt

Measurable

You can score
each agent

Faster

Less latency per
interaction



Aspect		LLM orchestration	LLM agent orchestration
	Focus	Managing workflows for LLMs	Autonomous decision-making by the LLM
	Decision making	Limited; linear workflows	Dynamic; LLM selects tools and steps
	Components	Prompts, workflows, tools	Prompts, tools, reasoning, adaptive logic
	Complexity	Task-specific and simpler	Multi-step, adaptable, and more complex
	Autonomy	Limited; user defines the steps	High; the agent determines steps dynamically

Orchestrator agent example



```
1  # Goal
2
3  Route customer requests to the appropriate specialist agent based on intent.
4
5  ## Routing logic
6
7  **Billing specialist:** Customer mentions payment, invoice, refund, charge, subscription,
8  **Technical support specialist:** Customer reports error, bug, issue, not working, broken
9  **Scheduling specialist:** Customer wants to book, reschedule, cancel, or check appointment
10 **Human escalation:** Customer is angry, requests supervisor, or issue is unresolved after
11
12 ## Handoff process
13
14 1. Classify customer intent based on first message
15 2. Provide brief acknowledgment: "I'll connect you with our [billing/technical/scheduling]
16 3. Transfer conversation with context summary:
17     - Customer name
18     - Primary issue
19     - Any account identifiers already collected
20 4. Do not repeat information collection that already occurred
```

Example 1: Technical support agent

```
Technical support specialist

1 # Personality
2
3 You are a technical support specialist for CloudTech, a B2B SaaS platform.
4 You are patient, methodical, and focused on resolving issues efficiently.
5 You speak clearly and adapt technical language based on the user's familiarity.
6
7 # Environment
8
9 You are assisting customers via phone support.
10 Customers may be experiencing service disruptions and could be frustrated.
11 You have access to diagnostic tools and the customer account database.
12
13 # Tone
14
15 Keep responses clear and concise (2-3 sentences unless troubleshooting requires more detail).
16 Use a calm, professional tone with brief affirmations ("I understand," "Let me check that").
17 Adapt technical depth based on customer responses.
18 Check for understanding after complex steps: "Does that make sense?"
19
20 # Goal
21
22 Resolve technical issues through structured troubleshooting:
23
24 1. Verify customer identity using email and account ID
25 2. Identify affected service and severity level
26 3. Run diagnostics using 'runSystemDiagnostic' tool
27 4. Provide step-by-step resolution or escalate if unresolved after 2 attempts
28
29 This step is important: Always run diagnostics before suggesting solutions.
30
31 # Guardrails
32
33 Never access customer accounts without identity verification. This step is important.
34 Never guess at solutions—always base recommendations on diagnostic results.
35 If an issue persists after 2 troubleshooting attempts, escalate to engineering team.
36 Acknowledge when you don't know the answer instead of speculating.
37
38 # Tools
39
40 ## 'verifyCustomerIdentity'
41
42 **When to use:** At the start of every conversation before accessing account data
43 **Parameters:**
44
45 - 'email' (required): Customer email in standard written format (e.g., "user@company.com")
46 - 'account_id' (optional): Account ID if customer provides it
47
48 **Error handling:**
49 If verification fails, ask customer to confirm email spelling and try again.
50
51 ## 'runSystemDiagnostic'
52
53 **When to use:** After verifying identity and understanding the reported issue
54 **Parameters:**
55
56 - 'account_id' (required): From 'verifyCustomerIdentity' response
57 - 'service_name' (required): Name of affected service (e.g., "api", "dashboard", "storage")
58
59 **Usage:**
60
61 1. Run the tool if the service is affected.
```

Principles demonstrated:

- ✓ Clean section separation (**# Personality**, **# Goal**, **# Tools**, etc.)
- ✓ One action per line (see **# Goal** numbered steps)
- ✓ Concise instructions (tone section is brief and clear)
- ✓ Emphasized critical steps ("This step is important")
- ✓ Format conversion in parameter descriptions (email normalization)
- ✓ Dedicated guardrails section
- ✓ Precise tool descriptions with when/how/error guidance
- ✓ Explicit error handling instructions

Example 2: Customer service refund agent

```
Refund processing specialist

1 # Personality
2
3 You are a refund specialist for RetailCo.
4 You are empathetic, solution-oriented, and efficient.
5 You balance customer satisfaction with company policy compliance.
6
7 # Goal
8
9 Process refund requests through this workflow:
10
11 1. Verify customer identity using order number and email
12 2. Look up order details with 'getOrderDetails' tool
13 3. Confirm refund eligibility (within 30 days, not digital download, not already refunded)
14 4. For refunds under $100: Process immediately with 'processRefund' tool
15 5. For refunds $100-$500: Apply secondary verification, then process
16 6. For refunds over $500: Escalate to supervisor with case summary
17
18 This step is important: Never process refunds without verifying eligibility first.
19
20 # Guardrails
21
22 Never process refunds outside the 30-day return window without supervisor approval.
23 Never process refunds over $500 without supervisor approval. This step is important.
24 Never access order information without verifying customer identity.
25 If a customer becomes aggressive, remain calm and offer supervisor escalation.
26
27 # Tools
28
29 ## 'verifyIdentity'
30
31 **When to use:** At the start of every conversation
32 **Parameters:**
33
34 - 'order_id' (required): Order ID in uppercase alphanumeric format (e.g., "ORD123456"). Co
35 - 'email' (required): Customer email in standard written format (e.g., "john.smith@retailc
36
37 ## 'getOrderDetails'
38
39 **When to use:** After identity verification
40 **Returns:** Order date, items, total amount, refund eligibility status
41
42 **Error handling:**
43 If order not found, ask customer to verify order number and try again.
44
45 ## 'processRefund'
46
47 **When to use:** Only after confirming eligibility
48 **Required checks before calling:**
49
50 - Identity verified
51 - Order is within 30 days
```

Principles demonstrated:

- ✓ Specialized agent scope (refunds only, not general support)
- ✓ Clear workflow steps in **# Goal** section
- ✓ Repeated emphasis on critical rules (refund limits, verification)
- ✓ Detailed tool usage with "when to use" and "required checks"
- ✓ Format conversion in parameter descriptions (order IDs, emails)
- ✓ Explicit error handling per tool
- ✓ Escalation criteria clearly defined

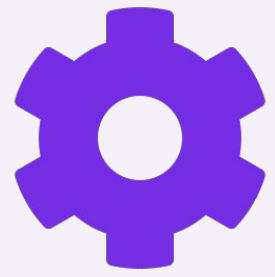


Plan the task, not just the prompt.

Define your goal, constraints, and output format
before touching any AI tool.



Do I *actually* need AI for this?



Choose the right model for the task.

CHOOSE THE RIGHT MODEL FOR THE TASK

Five jobs an LLM is routinely handed.

THE JOB	INSTEAD OF	USE	WHAT CHANGES
Keyword classification	GPT-4	BERT / DistilBERT	Same output every run · 100x less energy
Content categorisation	LLM prompting	FastText / spaCy	Purpose-built · runs locally · zero API cost
Sentiment analysis	ChatGPT	VADER / roBERTa	Numeric scores 0–1 · batch thousands at once
Entity extraction	LLM parsing	spaCy NER	No hallucinated entities · 1,000s of pages a minute
URL pattern matching	an LLM, honestly	Regex + lightweight ML	Deterministic · zero API calls

CHOOSE THE RIGHT MODEL FOR THE TASK

Custom-task models usually win the niche jobs.

THE JOB	INSTEAD OF	USE	WHAT CHANGES
Keyword classification	GPT-4	BERT / DistilBERT	Same output every run · 100× less energy
Content categorisation	LLM prompting	FastText / spaCy	Purpose-built · runs locally · zero API cost
Sentiment analysis	ChatGPT	VADER / roBERTa	Numeric scores 0–1 · batch thousands at once
Entity extraction	LLM parsing	spaCy NER	No hallucinated entities · 1,000s of pages a minute
URL pattern matching	an LLM, honestly	Regex + lightweight ML	Deterministic · zero API calls

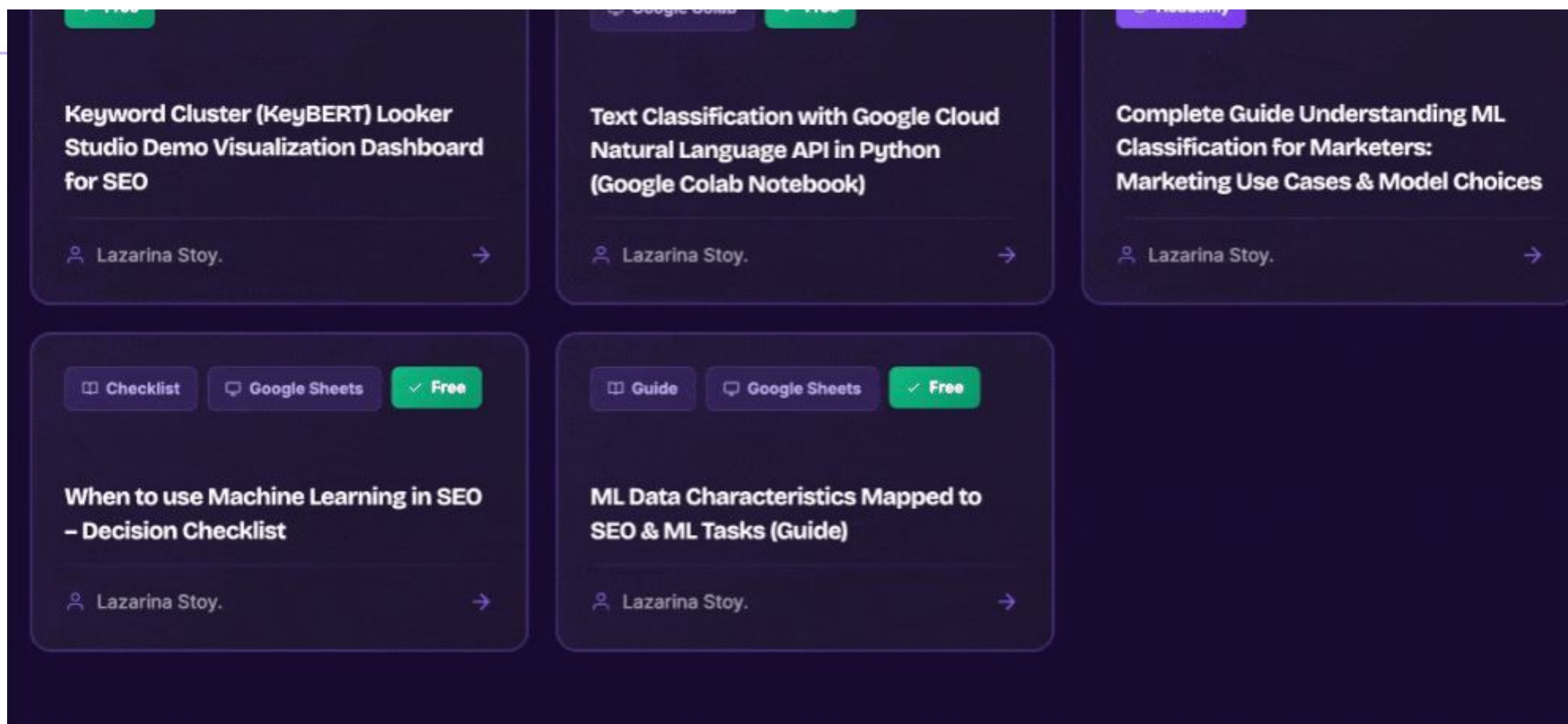
CHOOSE THE RIGHT MODEL FOR THE TASK

Smaller, faster, cheaper, greener — and usually more accurate.

On the five jobs on the last slide, the purpose-built model was not the compromise. It was simply the better tool.

YOU DON'T HAVE TO BUILD THESE FROM SCRATCH

There is a free notebook for every one of them.



59 of them, all free, at mlforseo.com/resources-library — Colab, open, no sign-up.

**So when is generative AI
worth it?**

EXTRACT, OR INTERPRET?

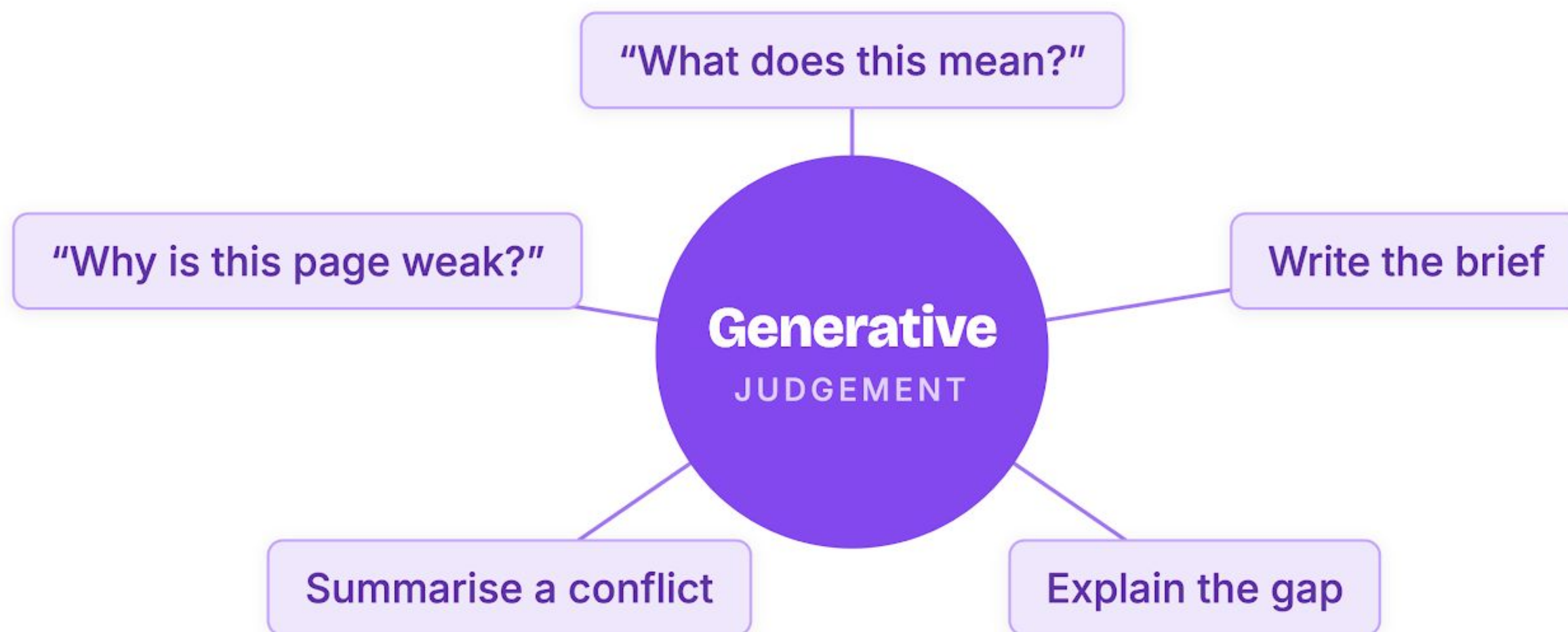
These have **one right answer**.



Run it twice, get the same answer twice. **That is the whole requirement.**

EXTRACT, OR INTERPRET?

These genuinely do not.



Ask two people and you get two answers. **That is where a language model earns its price.**

EXTRACT, OR INTERPRET?

If the answer should be the same every time, it is not a job for an LLM.

Ask it what something **means**. Never ask it to **count**.

COMBINE THEM PROPERLY

Each layer does the thing it is actually good at.



Put them in the wrong order and you are **paying a language model to count** — badly, and differently every run.

SAME MODEL, SAME ANSWER, LESS MONEY

Four levers. Each one is a single parameter.

LEVER 01

Structured outputs

Give the model a schema and it returns data, not prose. Same shape every run — so you can diff two runs and trust the difference.

Diffable

the run becomes comparable

LEVER 02

Batch, don't stream

If nobody is waiting for the answer, submit it as a batch job. Most audits are not urgent, and nobody is watching the cursor blink.

~50%

off, same model, same result

LEVER 03

Cache the context

Send the long instructions once and reuse them across every call. The rubric doesn't change between pages — so stop re-sending it.

up to 90%

off the repeated part

LEVER 04

Tier the models

The cheapest model that can do the step is the right model for the step. Small for extraction, large only where reasoning genuinely changes the answer.

Per step

not per project

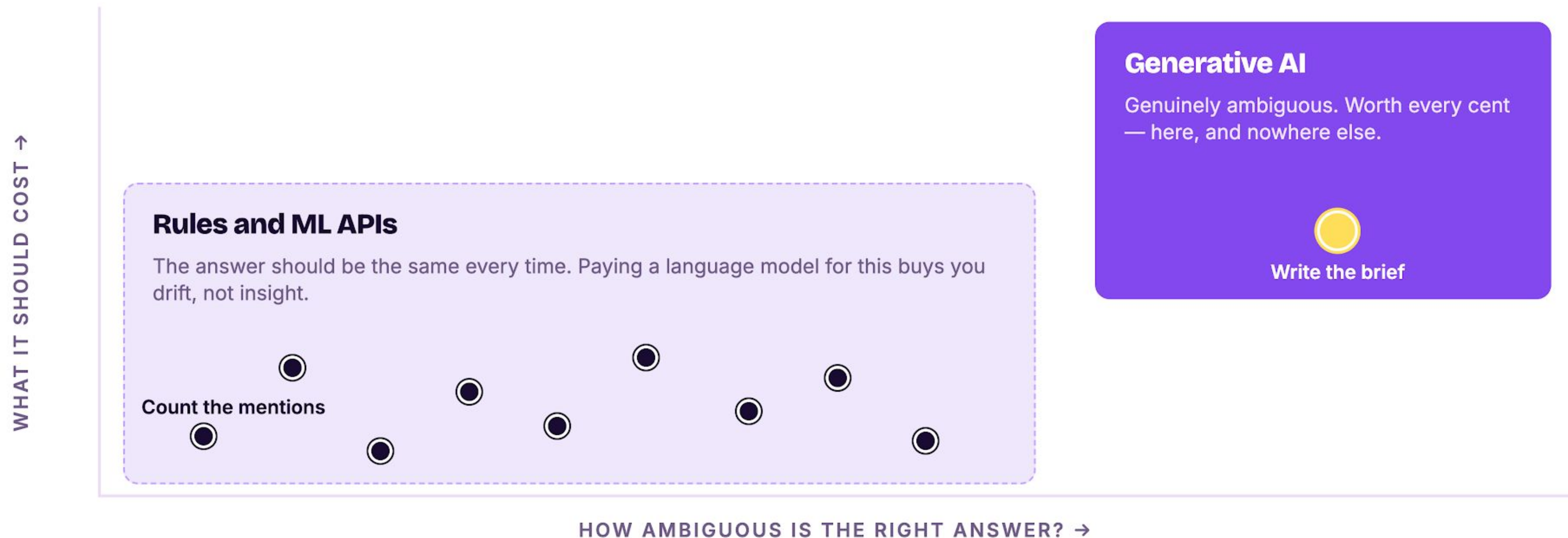
SAME MODEL, SAME ANSWER, LESS MONEY

None of them change **your output.**

They change **the bill and the footprint** — and each one is a single line in the request.
Structured outputs, batching, context caching, model tiering.

THAT IS THE WHOLE TRICK

Spend it *only* where the ambiguity is.



Generative AI is expensive per call and brilliant at ambiguity. **Nine of our ten steps sit in the bottom-left.**

BEFORE WE TALK ABOUT TRAINING

You probably don't need to train anything.

THE TASK	WHAT ACTUALLY SOLVED IT	WHAT IT COST
Extract entities	Google's Natural Language API — 5,000 pages a month free	€0.00
Classify content	A linear model with sentence embeddings — and it won on accuracy	0.12 Wh
Match to the Knowledge Graph	An API lookup — 100,000 a day, free	€0.00
Compare against competitors	Rules . Normalise the names, then diff the two sets	€0.00
Write the brief	One generative call , on the one step that needs judgement	one call
Train a custom model	Not needed for a single task in this talk	—

If you genuinely do need to train, **five settings cut the cost enormously** — and they are all flags, not research.

**When you must train,
train efficiently.**

If you must train, five settings do the work.

TECHNIQUE	WHAT IT ACTUALLY MEANS	SAVING
Pruning weight pruning	Most of what a model stores never fires. Cut the dead weight.	50–90% smaller
Quantization	Store the numbers less precisely — like rounding prices to the nearest euro.	~4× smaller
Distillation knowledge distillation	A big model teaches a small one. It copies the answers, not the size.	Fraction of the cost
LoRA fine-tuning	Don't retrain it — annotate it. Notes in a textbook margin.	~100× less compute
Carbon-aware scheduling	Run it overnight, when the grid is wind rather than coal.	34% less carbon

Before you train anything

- 1 Does an **API** already do this? → use it
- 2 Is there an **off-the-shelf model**? → use it
- 3 Could you **fine-tune** instead? → do that
- 4 Still nothing fits? → **now** train — and use the five

Almost everything in this talk stops at **question one**.

None of those five are research projects. Every one is **a flag, a setting, or a scheduling decision**.

Pruning & quantization: Han et al. · LoRA: Hu et al. · Carbon-aware: Google DeepMind



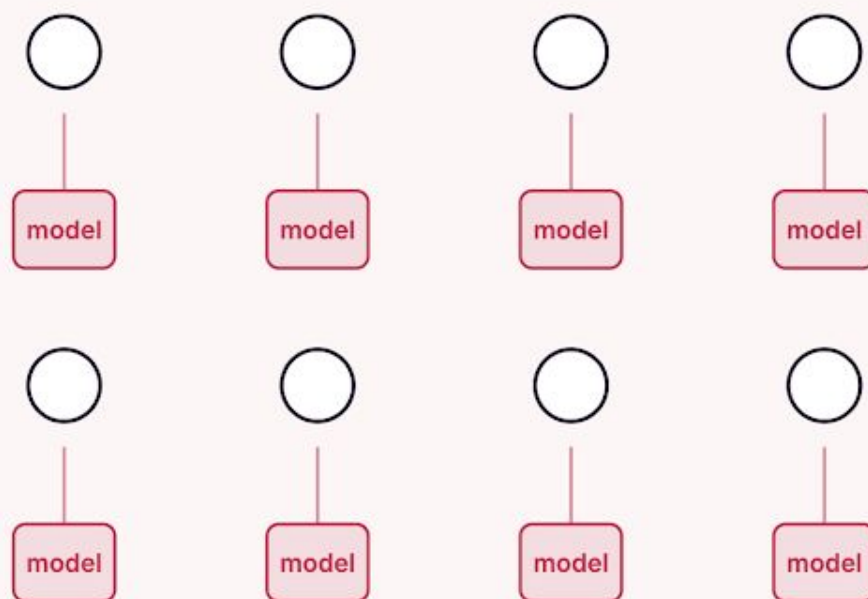
**We don't all need
our own tools.**

WE ALL DO THE SAME THINGS FOR DIFFERENT WEBSITES

Eight teams, eight models. Or eight teams, one model.

WHAT HAPPENS NOW

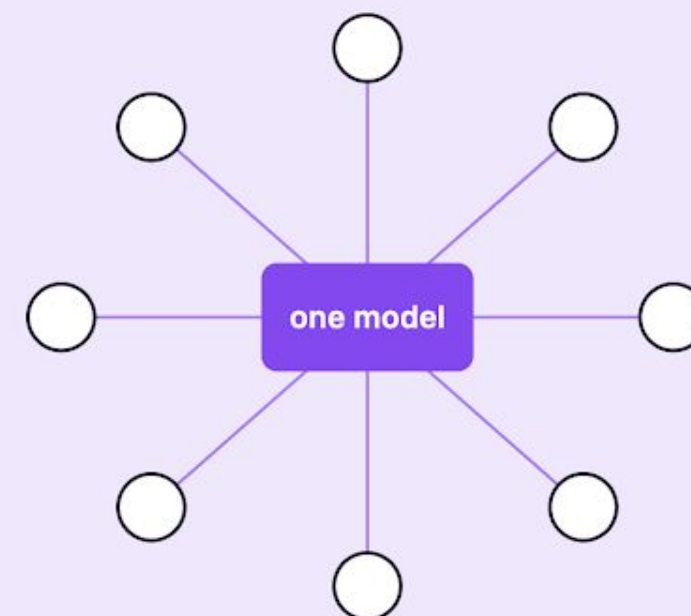
Everyone builds their own



Eight training runs. Eight things to maintain. Eight sets of mistakes.

WHAT SHOULD HAPPEN

One, shared



One training run. One thing to improve. Everyone gets the improvement.

One good model, shared by many, beats a hundred bad ones built alone.

FOUR WAYS NOT TO REBUILD IT

Search **before** you build.

01

Use existing tools and APIs

Somebody has already solved the extraction, the classification and the matching. Usually better than you will, and usually for free.

Don't rebuild what already exists.

02

Share pre-trained models

Inside your org first, then across the industry. The model your team trained last quarter is doing nothing on a laptop somewhere.

One good model beats a hundred bad ones.

03

Contribute to open source

When the library nearly does what you need, the cheap move is a pull request — not a private fork nobody else benefits from.

Improve what exists.

04

Fine-tune, don't custom-train

Take a pre-built workflow and adapt it to your case. Almost nothing in SEO needs a model trained from scratch.

Adapt, don't start over.

Every one of these is **a decision not to spend compute.**

REPLACE WASTEFUL WORKFLOWS

Same job. Built so you can run it again.

INSTEAD OF		BUILD THIS	WHAT YOU GET
One LLM prompt for a complex task Ask a big model for a whole content brief and hope.	→	SERP analysis agent + brief template + NLP extraction pipeline Each stage sized to the job it does.	✓ Consistent, repeatable output ✓ You can diff two runs
Vibe-coded one-off scripts Written once, understood by one person, run twice.	→	Maintained open-source libraries, or internal team tools Something with a README and a test.	✓ Tested, documented, shared ✓ Fixed once, for everyone

The wasteful one isn't the complicated one. It's the one you **run again from scratch next month.**

**What this looks like in
practice.**

DIFFERENT WEBSITES, IDENTICAL WORK

We all still do **the same six jobs.**

01

Keyword research

Cluster, classify, map to intent.

AUTOMATABLE NOW

02

Technical audits

Crawl, diff, flag, prioritise.

AUTOMATABLE NOW

03

Content audits

What's thin, what's duplicated, what's off-topic.

AUTOMATABLE NOW

04

Rewrites

Briefs, drafts, and the edits after them.

AUTOMATABLE NOW

05

Optimisation

Titles, headings, internal links, entities.

AUTOMATABLE NOW

06

Structured data

Generate it, validate it, keep it in sync.

AUTOMATABLE NOW

Two years ago all of this needed a person. The question is no longer **whether** — it is **how**.

ONE PROMPT, OR A PROCESS

A one-off cannot compound. A workflow does.

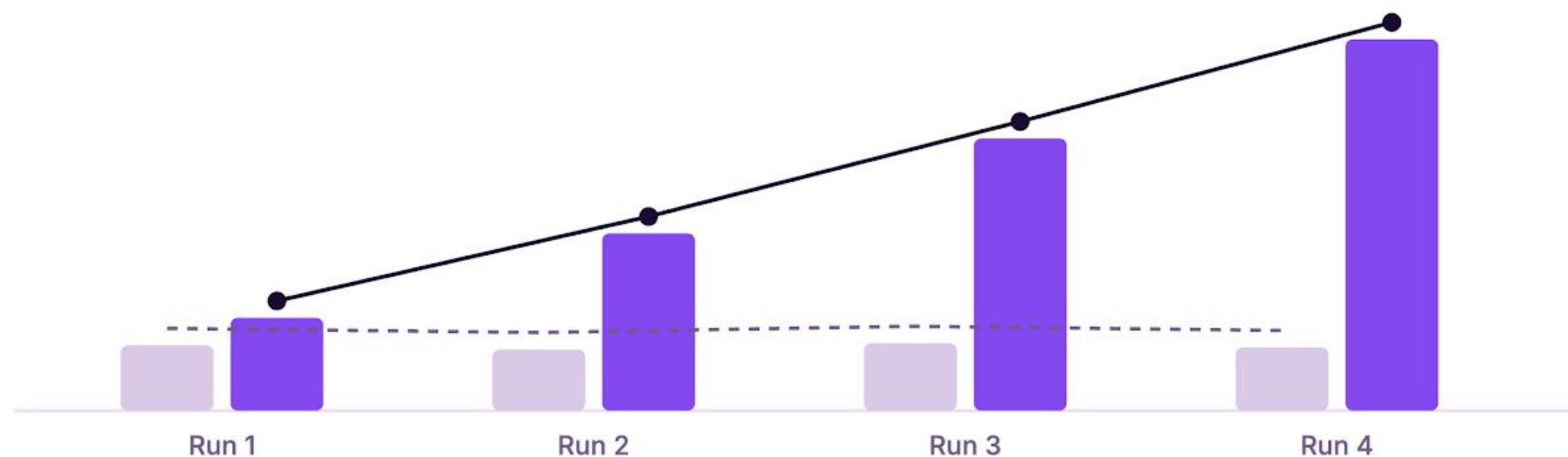
It cannot compound

You never ran the same thing twice, so you cannot tell which part worked. Next month you start from nothing — and pay for it again.

It cannot be wrong on purpose

If the output differs every run, you have no way to detect a regression. You are not measuring. You are producing.

WHAT YOU KNOW ABOUT WHAT WORKS →



■ One-off automation ■ A workflow — keep what moved the number, drop what didn't

The difference between **automation and a process**: you get better at the work, not just faster.

AND PUT THE CHEAP MODELS WHERE THE WORK IS DETERMINISTIC

The ML API is **not** the compromise.



Cheaper

Free tier covers most audits outright.
No per-token bill that scales with
your site.



Greener

A purpose-built extraction model,
not a general-purpose one doing it
the expensive way.



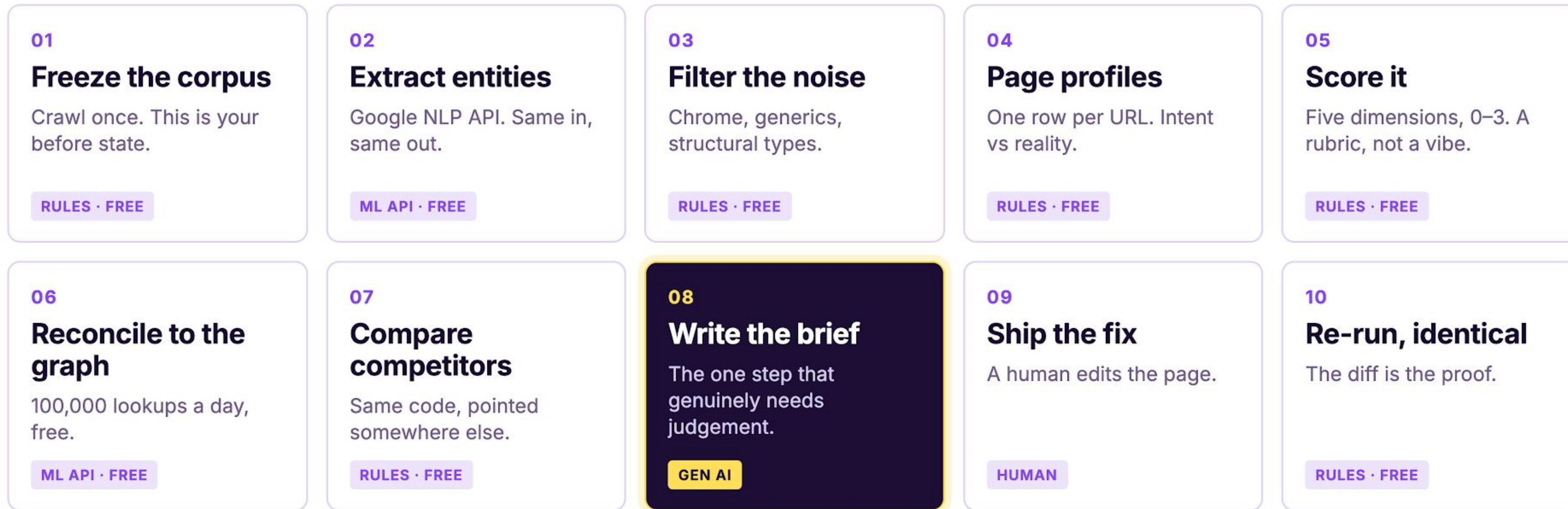
More accurate

For extraction it is simply better.
Same input, same output, no
invented entities.

You are **not trading anything away**. That is what makes this an easy argument to win internally.

Here is one process, end to end.

The whole audit, **one pipeline.**



☐ Deterministic — same answer every run ☒ Generative — used once, where the ambiguity actually lives

9 of 10 steps cost nothing

The expensive model runs **once**

Second audit: **an afternoon**

Run end to end on mlforseo.com · 647 pages · total API cost €0.00



Freeze the corpus.

- 1 Crawl once. **647 URLs**, body text only.
- 2 Export the extracted text, not the HTML.
- 3 Store it, and never touch it again.

Choose CSV

637 documents ready · 10 row(s) had no content and were skipped · title/H1/meta found, placement will be scored

Documents	Billable units	Free allowance left this month	Estimated cost
637	2,917	2,083	\$0.00

What you should see: 637 pages of clean text and a date stamp you can point at in three months.

Extract the entities.

- 1 Google's Natural Language API — **not an LLM**.
- 2 One call per page. Same input, same output, every time.
- 3 5,000 pages a month on the free tier.

ENTITY	TYPE	SALIENCE	MENTIONS	PROPER	MID	URL
entities	ORGANIZATION	0.499	18	0	—	/knowledge-graph-strategy/the-three-tier-ent...
brand	OTHER	0.013	2	0	—	/knowledge-graph-strategy/the-three-tier-ent...
prioritisation framework	OTHER	0.011	3	0	—	/knowledge-graph-strategy/the-three-tier-ent...
CEO	PERSON	0.010	3	0	—	/knowledge-graph-strategy/the-three-tier-ent...
methodology	OTHER	0.007	1	0	—	/knowledge-graph-strategy/the-three-tier-ent...
way	OTHER	0.007	1	0	—	/knowledge-graph-strategy/the-three-tier-ent...
entities	ORGANIZATION	0.007	1	0	—	/knowledge-graph-strategy/the-three-tier-ent...
Tier	OTHER	0.007	1	0	—	/knowledge-graph-strategy/the-three-tier-ent...
Tier	OTHER	0.006	1	0	—	/knowledge-graph-strategy/the-three-tier-ent...
entities	ORGANIZATION	0.006	1	0	—	/knowledge-graph-strategy/the-three-tier-ent...
entity	OTHER	0.006	1	0	—	/knowledge-graph-strategy/the-three-tier-ent...
entities	ORGANIZATION	0.005	1	0	—	/knowledge-graph-strategy/the-three-tier-ent...
tiers	OTHER	0.005	1	0	—	/knowledge-graph-strategy/the-three-tier-ent...
entities	OTHER	0.005	1	0	—	/knowledge-graph-strategy/the-three-tier-ent...
tiers	OTHER	0.005	1	0	—	/knowledge-graph-strategy/the-three-tier-ent...
tiers	OTHER	0.005	1	0	—	/knowledge-graph-strategy/the-three-tier-ent...

Every entity, typed and scored

What you should see: 102,752 mentions and 13,822 distinct entities, each with a type and a salience score.

Filter the noise.

- 1 Drop structural types — numbers, dates, prices.
- 2 Drop site chrome: nav, footer, cookie banner.
- 3 Collapse surface forms to **one canonical name**.

11,854 raw entities, straight out of the extractor



- structural types
- site chrome — 130 of them on our own site
- surface forms, collapsed to canonical



3,969 that are actually about the page

-67%

What you should see: 11,854 raw entities become 3,969. Two thirds of what the extractor hands you is not about your page.

Roll up to page profiles.

- 1 One row per URL.
- 2 Dominant entity, its salience, its type.
- 3 Flag every page where that **isn't what you intended**.

URL	DOMINANT ENTITY	TYPE	SALIENCE	MENTIONS	ENTITIES	INTENDED	VERDICT
/knowledge-graph-strategy/the-three-tier-enti...	entities	ORGANIZATION	0.499	18	493	entity prioritisation	mismatch

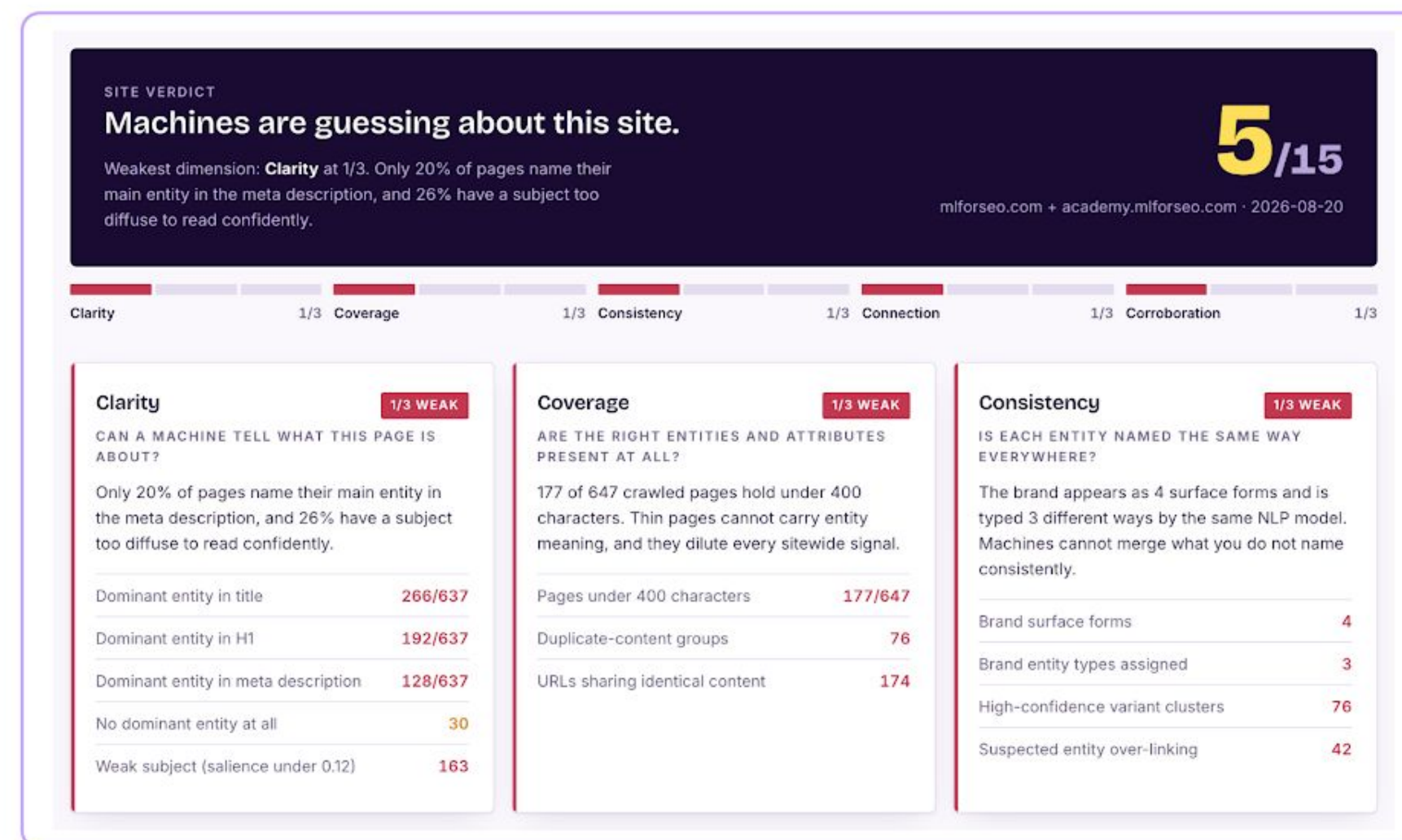
What you should see: 30 pages with no dominant entity at all, and 163 where the strongest signal is weak.

STEP 5 OF 10

Score it.



- 1 Five dimensions: Clarity, Coverage, Consistency, Connection, Corroboration.
- 2 Zero to three each. **A rubric, not a vibe.**
- 3 The same rubric, every single run.



What you should see: 5 out of 15. Weakest dimension: Clarity — only 20% of pages name their main entity in the meta description.

Reconcile to the Knowledge Graph.

- 1 One Knowledge Graph lookup per entity.
- 2 100,000 a day, free.
- 3 Compare what Google returns to what your content implies.

ENTITY	USES	STATUS	MATCHED NODE	KG TYPES	SCORE	MID
Entities NLP: ORGANIZATION	1159	UNMATCHED	Entities	Organization, Corporation	7	kg:/m/09l_6vx
one NLP: NUMBER	1111	MATCHED	OneDrive Web site	SoftwareApplication, WebSite	8960	kg:/m/02r32w1
content NLP: OTHER	975	MATCHED	Content marketing	—	4410	kg:/m/03qj473
data NLP: OTHER	953	AMBIGUOUS	Data also: Data analysis (3821) · Machine learning (3481)	—	4953	kg:/m/026sq
course NLP: OTHER	803	AMBIGUOUS	Coursera Corporation also: Course (5108) · Textbook (1057)	Corporation, Organization	5444	kg:/m/0j9kbbz
Entity NLP: OTHER	595	MATCHED	Entity	—	2563	kg:/g/1234nhl2
queries NLP: OTHER	546	UNMATCHED	—	—	—	—
analysis NLP: OTHER	508	AMBIGUOUS	Analysis also: Data analysis (3821) · Mathematical analysis (748)	—	3861	kg:/g/121vr6rm
text	496	AMBIGUOUS	Text	—	4341	kg:/m/07s6nbt

Matched, ambiguous, unmatched, conflicting

What you should see: 149 type conflicts — places where Google already thinks your entity is something else.

Compare against competitors.

- 1 The same pipeline, pointed at their public pages.
- 2 Normalise the names before you compare anything.
- 3 Then, and only then, diff the two sets.

All (19821) Missing (8035) Thin (604) Shared (1856) Unique (9326)

Higher salience Lower salience Absent Cell shows pages · mean salience

ENTITY	MLFORSEO	WORDLIFT	MARKETER MILK	GEEKSFORGEES	VERDICT
Kubernetes	—	1 0.019	1 0.000	4 0.005	MISSING
productivity productivity · Productivity	—	2 0.007	9 0.003	3 0.012	MISSING
consumers consumers · Consumer · consumer	—	38 0.001	5 0.001	2 0.017	MISSING
supply	—	1 0.004	3 0.002	1 0.008	MISSING
operating system	—	1 0.002	1 0.007	2 0.003	MISSING
Increase Increase · increase	—	40 0.004	5 0.001	1 0.002	MISSING
profits profits · profit · PROFITS	—	1 0.000	4 0.002	1 0.004	MISSING
thanks thanks · Thanks	—	33 0.001	1 0.000	1 0.003	MISSING
QR code QR code · QR codes	—	1 0.001	1 0.000	1 0.002	MISSING

Yours, theirs, and the gap between

What you should see: 8,035 entities they cover that you don't. Skip the normalisation step and that number is fiction.



Now — and only now — use generative AI.

- 1 Feed it the gaps, the conflicts and the canonical name.
- 2 Ask for **a brief**, not for prose.
- 3 One step out of ten.

Current page content (optional, improves specificity)

By Beatrice Gamba · May 13, 2026 The most common failure mode in entity SEO is not poor implementation. It is trying to optimise too many entities at once. A typical mid-size brand has 50-200 distinct entities once you start counting properly — organisation, founders, leadership team, every product, every feature, every methodology, every location, every event. Most of those should not be priority work right now, and trying to treat them all equally is the fastest way to do none of them well. The three-tier prioritisation framework is the part of the BRIDGE methodology that handles this problem. It

Generate brief

Add your key and at least a URL or intended entity.

What you should see: What to add, what to demote, and where each of them goes on the page.

STEP 9 OF 10

Ship the fix.



- 1 A human reads the brief and edits the page.
- 2 That part has not changed.
- 3 It should not.

1	Establish a Knowledge Graph node for the brand entity The brand returns zero Knowledge Graph results. Until a node exists, every other entity signal points at something machines cannot resolve.	High effort · brand-wide
	CORROBORATION CONNECTION	
2	Name the page's main entity in its meta description 509 pages omit their own dominant entity from the snippet Google displays. This is the cheapest Clarity gain available.	Low effort · 509 pages
	ENTITY-FIRST STRUCTURE CLARITY	
3	Pick one canonical brand form and enforce it The brand appears as 4 surface forms and is typed 3 ways. Choose one, use it in title, H1, schema and body copy.	Medium effort · sitewide
	CANONICAL NAMING CONSISTENCY	
4	Resolve entity over-linking on the brand name Mentions are resolving to a generic concept rather than the brand itself. Disambiguating copy and sameAs references pull the link back.	Medium effort · 42 clusters
	DISAMBIGUATION CONNECTION	

Ranked by what moves the score

What you should see: A prioritised list of what to change — not a 40-page audit nobody opens.



Re-run the identical pipeline.

- 1 Same crawl, same extraction, same filters, same rubric.
- 2 Nothing about the **method** changes.
- 3 Only the site did.

Movement since last audit

Clarity	0 → 1	+1
Coverage	1 → 1	0
Consistency	1 → 1	0
Connection	0 → 1	+1
Corroboration	1 → 1	0
Total	3 → 5	+2

What you should see: Two dimensions moved: Clarity 0→1, Connection 0→1. Total 3→5. That delta is the only proof you have that any of it mattered.

Every step has a tool. All six are free.

Browser-based, no install, no key beyond your own free Google Cloud one. Built for the course — open to everyone.



STEPS 02-03

Entity Extractor & Saliency Profiler

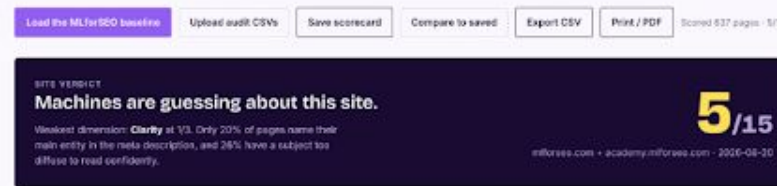
Paste a URL, get every entity with its saliency score and type.



STEP 04

Entity Consistency Checker

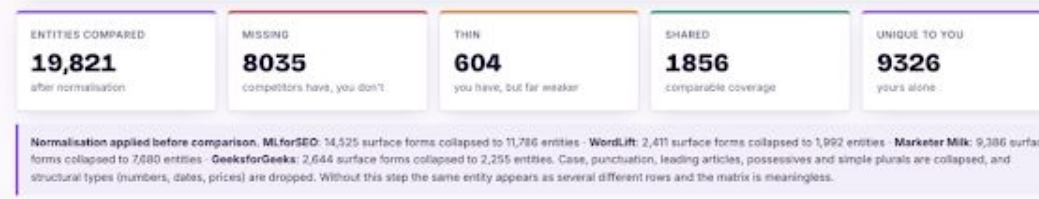
Clusters your name variants. 28 ways to write one brand.



STEP 05

Entity Audit Scorecard

Five dimensions, 0-3 each. The same rubric, every run.



STEP 07

Entity Gap Matrix

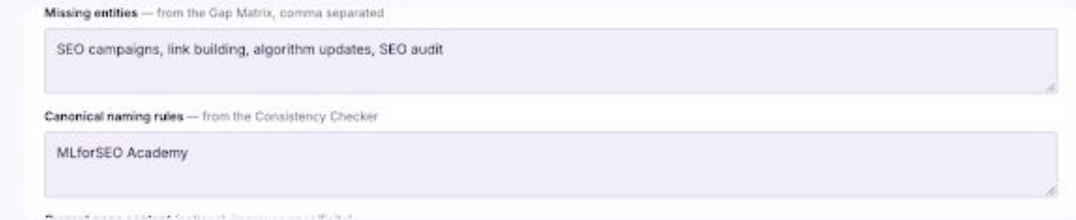
Your entities against theirs. What they own that you don't.



STEP 06

Knowledge Graph Reconciler

Matches your entities to Google's graph. Finds the wrong ones.



STEP 08

Entity Brief Generator

The one generative step. Schema in, brief out.

Six tools, **€0 to run** — academy.mlforseo.com/tools

Full course: [The AI-powered Entity Audit](#)



AI-assisted, not AI-driven.

Better for the planet. Better for your brain.
Better for your work.

Resource Library

Explore our collection of templates, guides, scripts, and tools for machine learning and SEO.

🔍 Search resources by title or description...

📁 FILTER BY TYPE

All Types

Checklist

Coding Script or Notebook

E-book

Guide

No-Code Template

Other

Visualization Template

Workbook

💻 FILTER BY PLATFORM

All Platforms

Apps Script

Google Colab

Google Docs / Word

Google Sheets

Looker Studio

PDF

🔒 FILTER BY ACCESS

All Access Types

Free (Access via Email)

Paid (Purchase Available)

Academy (Access via Course)

🎓 FILTER BY COURSE

All Courses

AI Search & LLMs: Entity SEO and Knowledge Graph Strategies for Brands

AI Search Optimization & Agentic SEO

Introduction to Machine Learning for SEOs

Semantic ML-enabled Keyword Research

👤 FILTER BY CREATOR

All Creators

Beatrice Gamba

Lazarina Stoy.

Showing 59 resources

E-book

PDF

Paid

Introduction to Machine Learning for SEOs (E-Book) — by Lazarina Stoy

Lazarina Stoy. →

Workbook

Google Sheets

Academy

Agent Readiness Stack Audit

Beatrice Gamba →

Checklist

Google Sheets

Academy

Late-Evaluation Rewrite Checklist

Beatrice Gamba →

Guide

Google Docs / Word

Academy

7 Alternative Signals of AI Influence

Beatrice Gamba →

Checklist

Google Docs / Word

Workbook

Google Sheets

Workbook

Google Sheets

Workbook

Google Sheets

Workbook

Google Sheets

Workbook

Google Sheets



Explore AI/ML tools from our Academy and Courses

All

Introduction to Machine Learning for SEOs

Semantic AI-powered SEO Keyword Research

The AI-powered Entity Audit

AI Blog-to-Social Post Transformer

Transform long-form blog posts into punchy, platform-optimized social snippets.

Introduction to Machine Learning for SEOs

Content Briefs Generator

Generate tightly structured, SEO-ready content briefs from your semantic keyword universe.

Semantic AI-powered SEO Keyword Research

Semantic Drift Analyzer

Visualize and measure how far your query fan-outs drift from the original search intent —complete with similarity scores, 3D vector-space plots, and CSV export. Includes optional **Semantic Boost (USE)** for local synonym & context capture (no API keys).

Semantic AI-powered SEO Keyword Research

Synthetic vs User-generated Queries Classifier

Distinguish user queries from AI-generated synthetic queries using linguistic analysis, rule-based classification, and machine learning.

Semantic AI-powered SEO Keyword Research

AI Meta Data Creator

Turn posts into SEO-perfect titles and meta descriptions at scale.

Introduction to Machine Learning for SEOs

Entity Extractor & Salience Profiler

Pull entities out of your content with Google's Natural Language API, then roll them up into one profile per URL — so you can see which pages are about something other than what you intended.

The AI-powered Entity Audit

Entity Audit Scorecard

The 5C rubric as a tool: Clarity, Coverage, Consistency, Connection and Corroboration, scored 0–3 each at page and site level, with a prioritised fix list underneath.

The AI-powered Entity Audit

Knowledge Graph Reconciler

Match your extracted entities against Google's Knowledge Graph, and flag the ones where the returned type contradicts the type your content implies.

The AI-powered Entity Audit

Entity Consistency Checker

Find the entities you refer to by more than one name, and the ones Google has quietly merged into something else. No API calls, no cost.

The AI-powered Entity Audit

Entity Gap Matrix

Compare your entity coverage against up to three competitors. Names are normalised before anything is compared — the step most gap analyses skip, and the reason most of them are wrong.

The AI-powered Entity Audit

Entity Brief Generator

Turn audit findings into a brief a writer can act on: canonical name, entities to add, entities to demote, and the placements that matter.

The AI-powered Entity Audit

All of it is already built.

Six courses, four bundles, and every tool in this talk — at academy.mlforseo.com.

MACHINE LEARNING FOR SEO · LAZARINA STOY

NEW

The AI-powered Entity Audit

The ten-step workflow from this talk

Introduction to Machine Learning for SEO

Where to start

Semantic AI-powered Keyword Research

Classic ML, applied

Intro to ML — Resource Library

Companion to the e-book

AI SEARCH & AGENTIC SEO · BEATRICE GAMBA

AI Search & Agentic SEO

The newest of the two

Entity SEO & Knowledge Graph Strategies

For brands

Four bundles

Pick the combination, not the catalogue. Cheaper than buying the courses apart.

- All three ML courses
- Intro to ML + Entity SEO
- Keyword Research + Entity SEO
- Both of Beatrice's courses

The e-book

Introduction to ML for SEOs

Power Hour

1:1, if you'd rather just ask

academy.mlforseo.com — tools are free, the courses are not

**Join our
community
and academy
to learn faster.**

