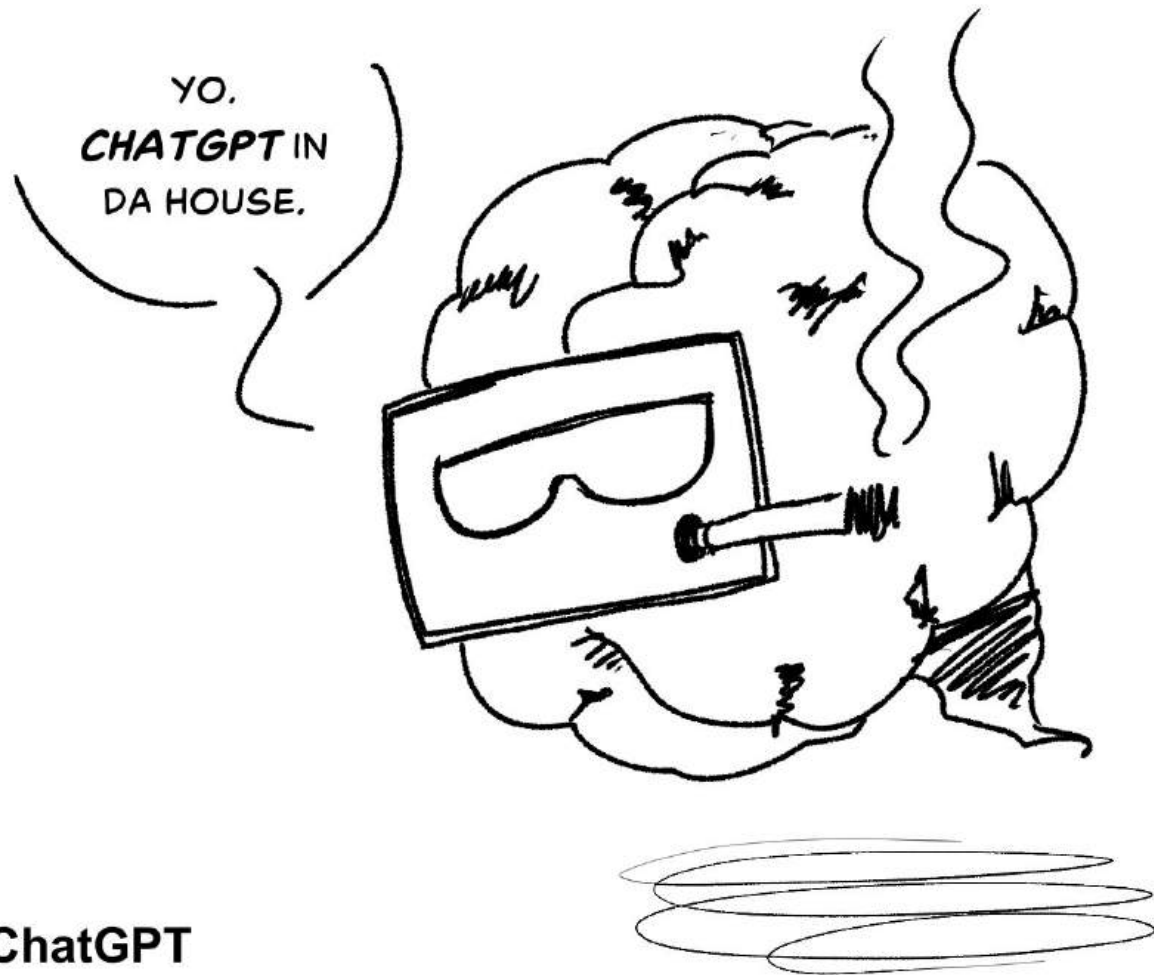


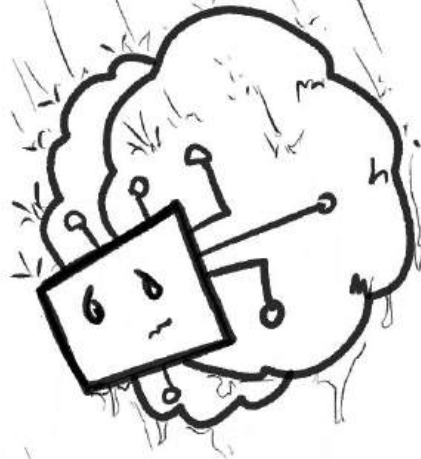


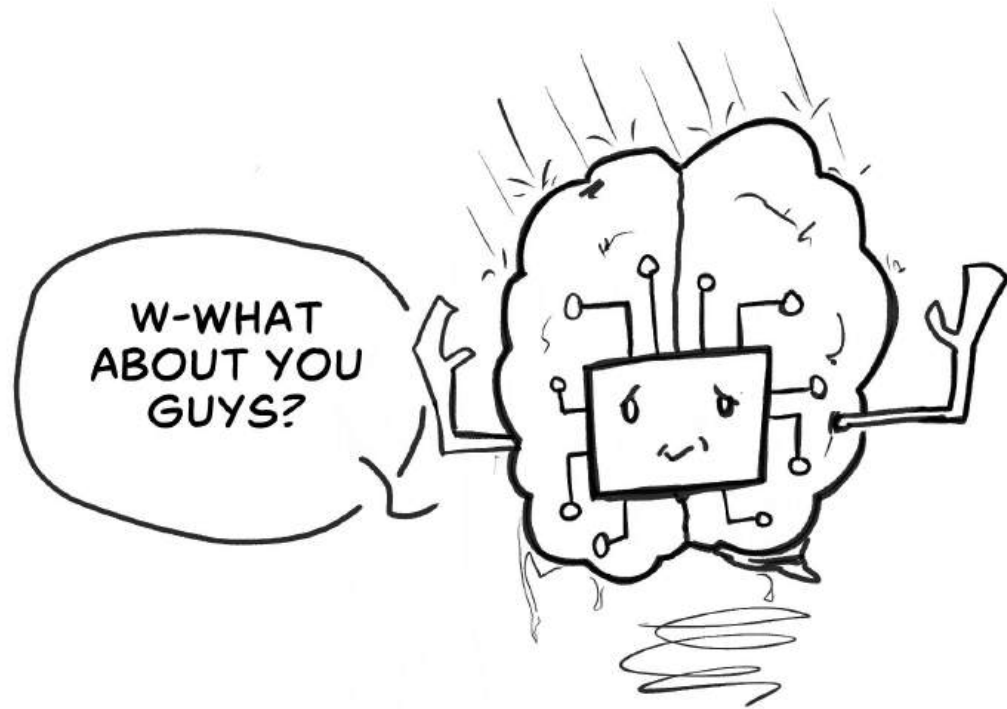
Explore the jungle of applying AI
to your experimentation process

Iqbal Ali, 2025

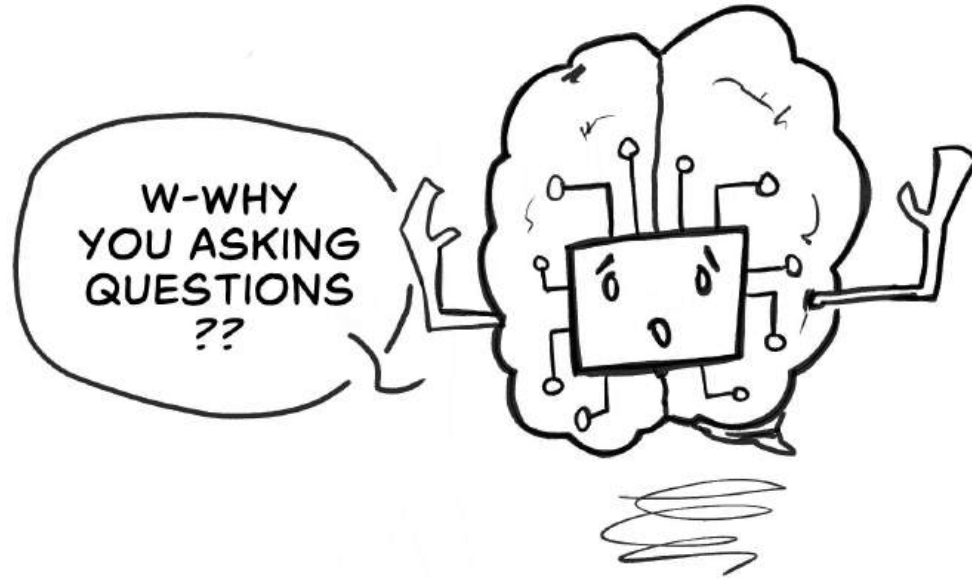


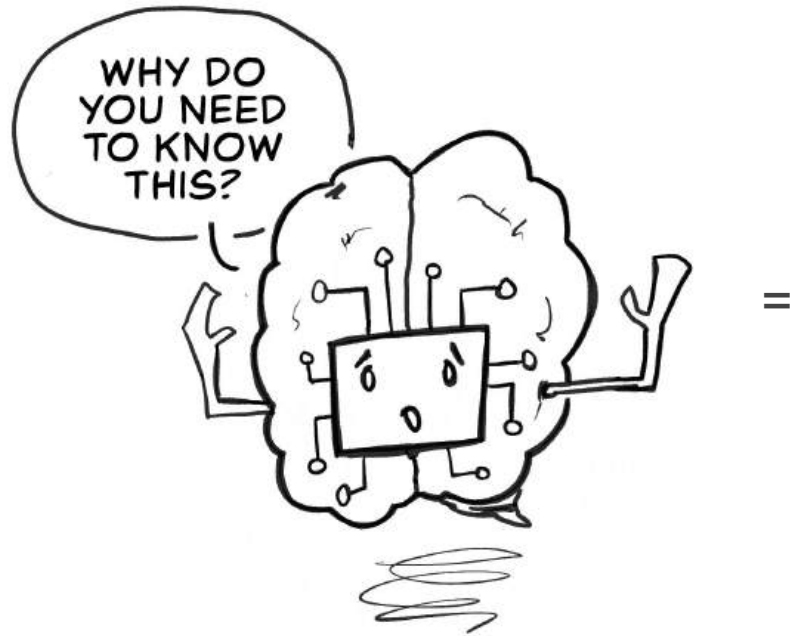
IT'S
NOT ME, IT'S
YOU.





How much
energy to train that
2022 model?

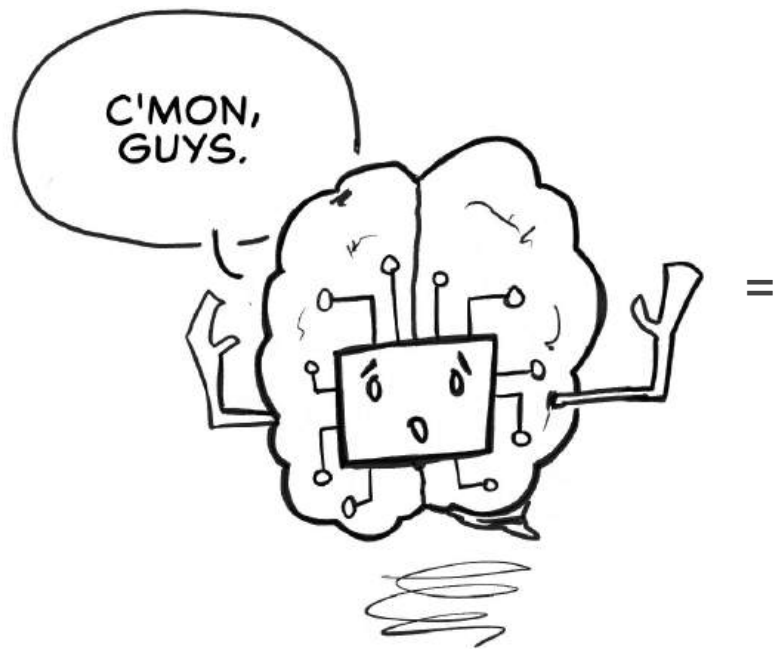




=



Daily power needs
for average US homes.

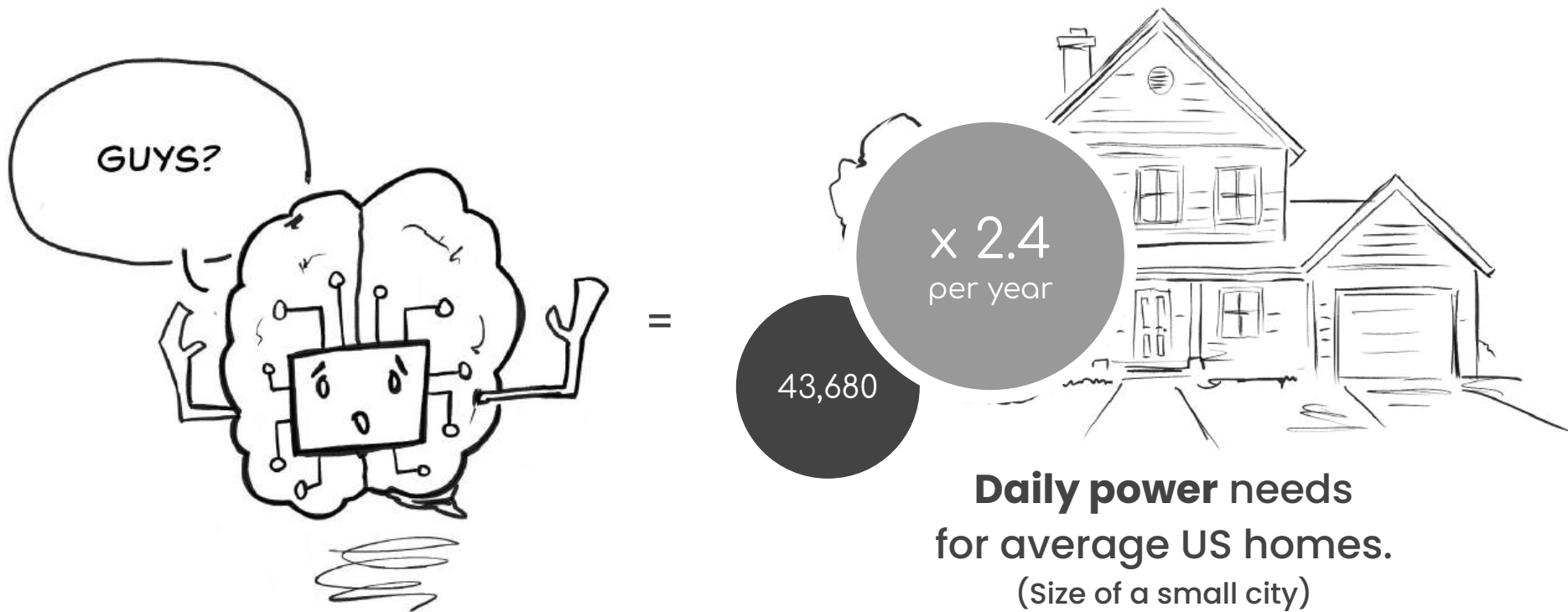


=

43,680



Daily power needs
for average US homes.
(Size of a small city)



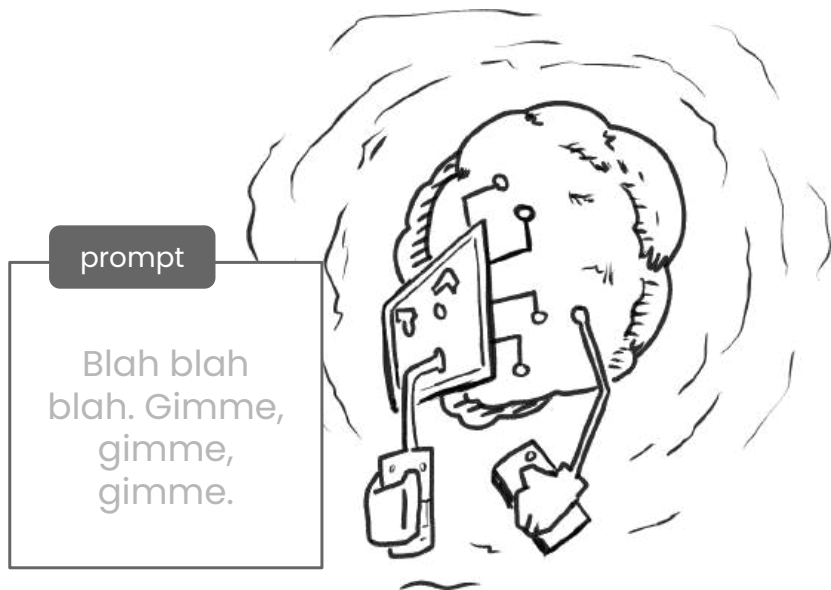
Approximate daily
power usage [1][2]*



** Calculated estimates based on what we know*



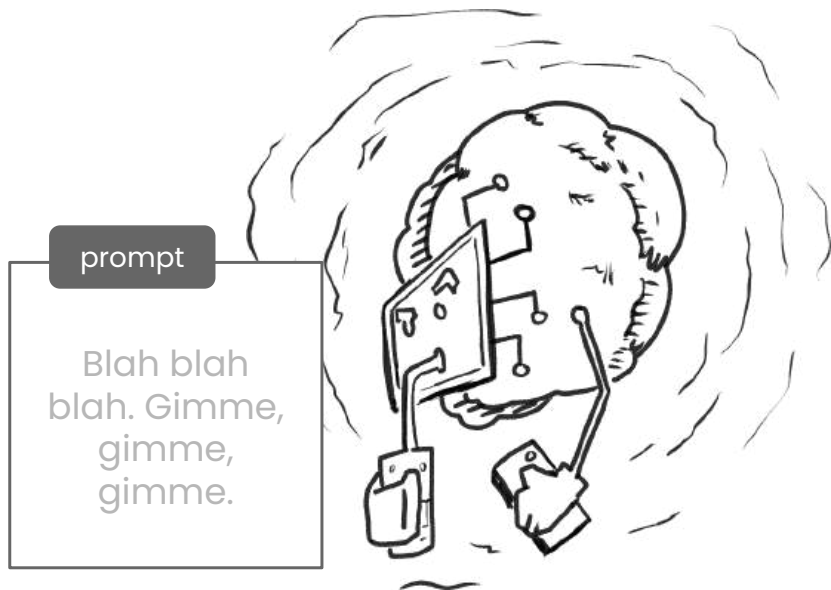
UMM...WE
SERVE ABOUT, UMM,
2.5 BILLION PROMPTS
PER DAY. UMM.
YEAH.



Power consumption
per prompt *

[\[source\]](#)

Approximate daily running cost of GPT-5



Power consumption
per prompt *

[\[source\]](#)



Estimated.

Approximate daily running cost of GPT-5

EMPIRE OF AI

Alfred Law

Inside
the reckless
race for
total
domination

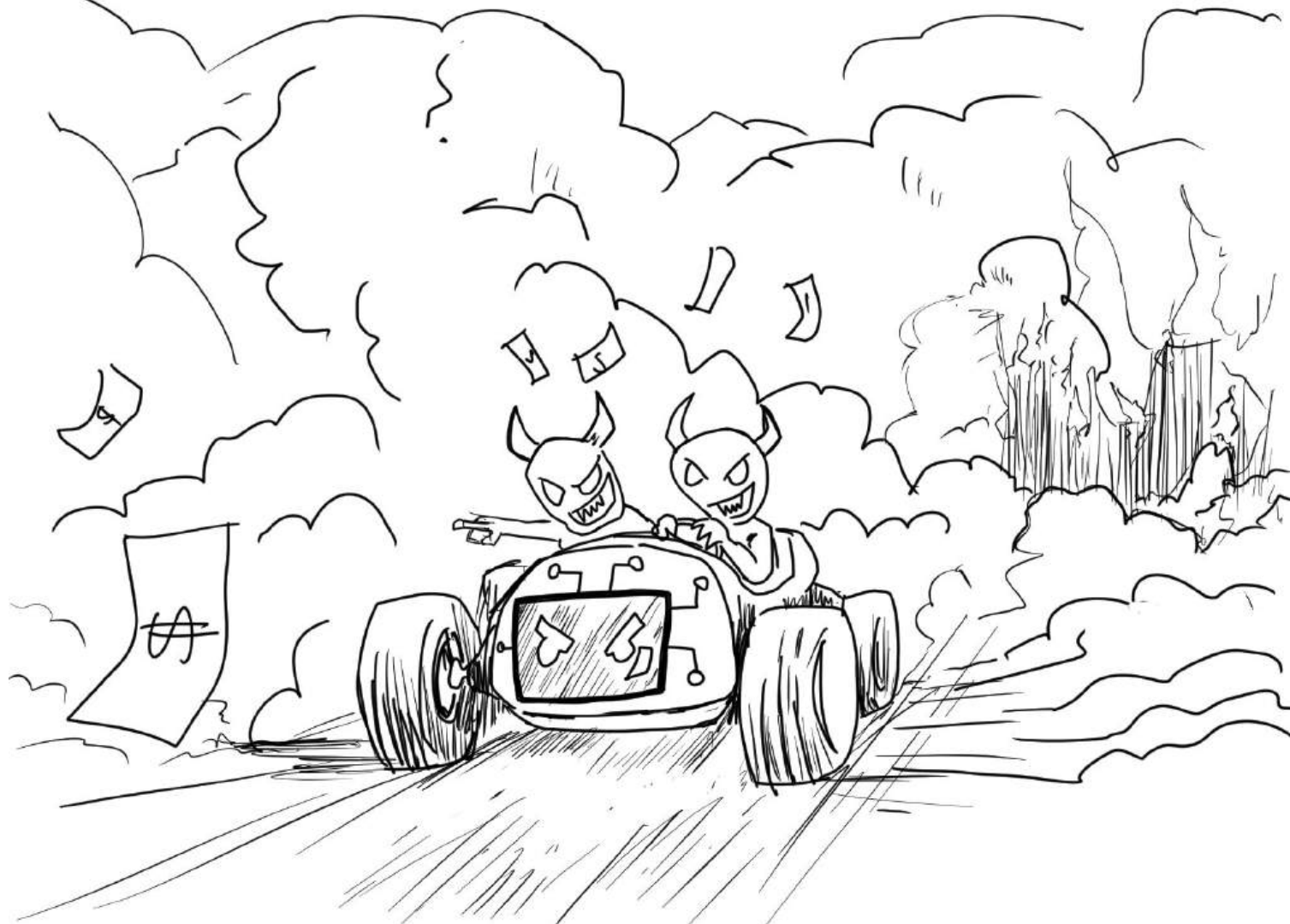
KAREN HAO

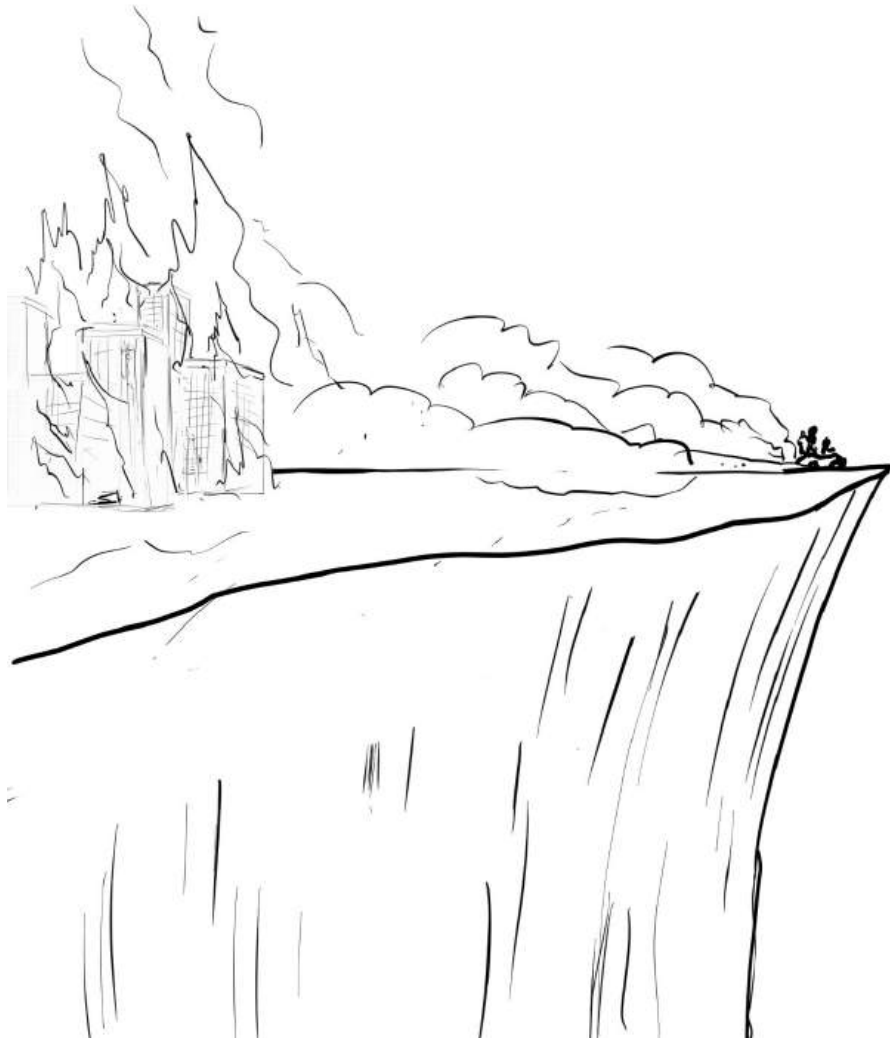
Papers to get started:

[Rising Costs of Training
Frontier Models](#) (Feb 2025)

[How AI will drive energy
demand](#) (Apr 2025)

[Uncovering and Addressing the Secret
Water Footprint of AI Models](#) (Mar 2025)





I
got you,
bro.



AI Agents successfully
completing tasks: 30%

Paper: [Benchmarking LLM
Agents \(Sep 2025\)](#)



AI Agents successfully completing tasks: 30%

Paper: [Benchmarking LLM Agents \(Sep 2025\)](#)

AI Benchmarks may be faulty

Paper: [The Vulnerability of Language Model Benchmarks](#) (Dec 2024)



AI Agents successfully completing tasks: 30%

Paper: [Benchmarking LLM Agents \(Sep 2025\)](#)

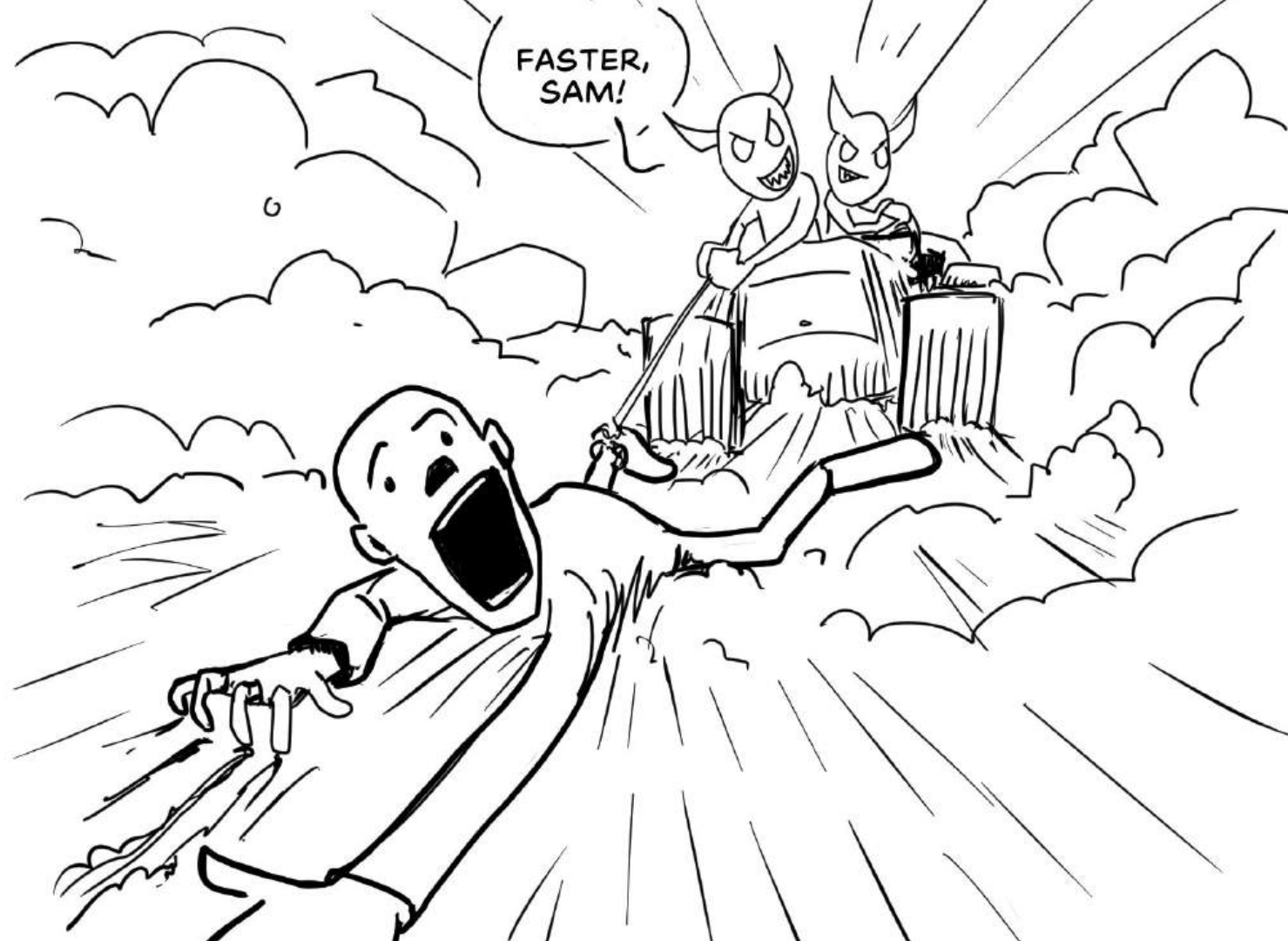
AI Benchmarks may be faulty

Paper: [The Vulnerability of Language Model Benchmarks](#) (Dec 2024)

"...an accuracy collapse."

Paper: [The Illusion of Thinking \(2025\)](#)

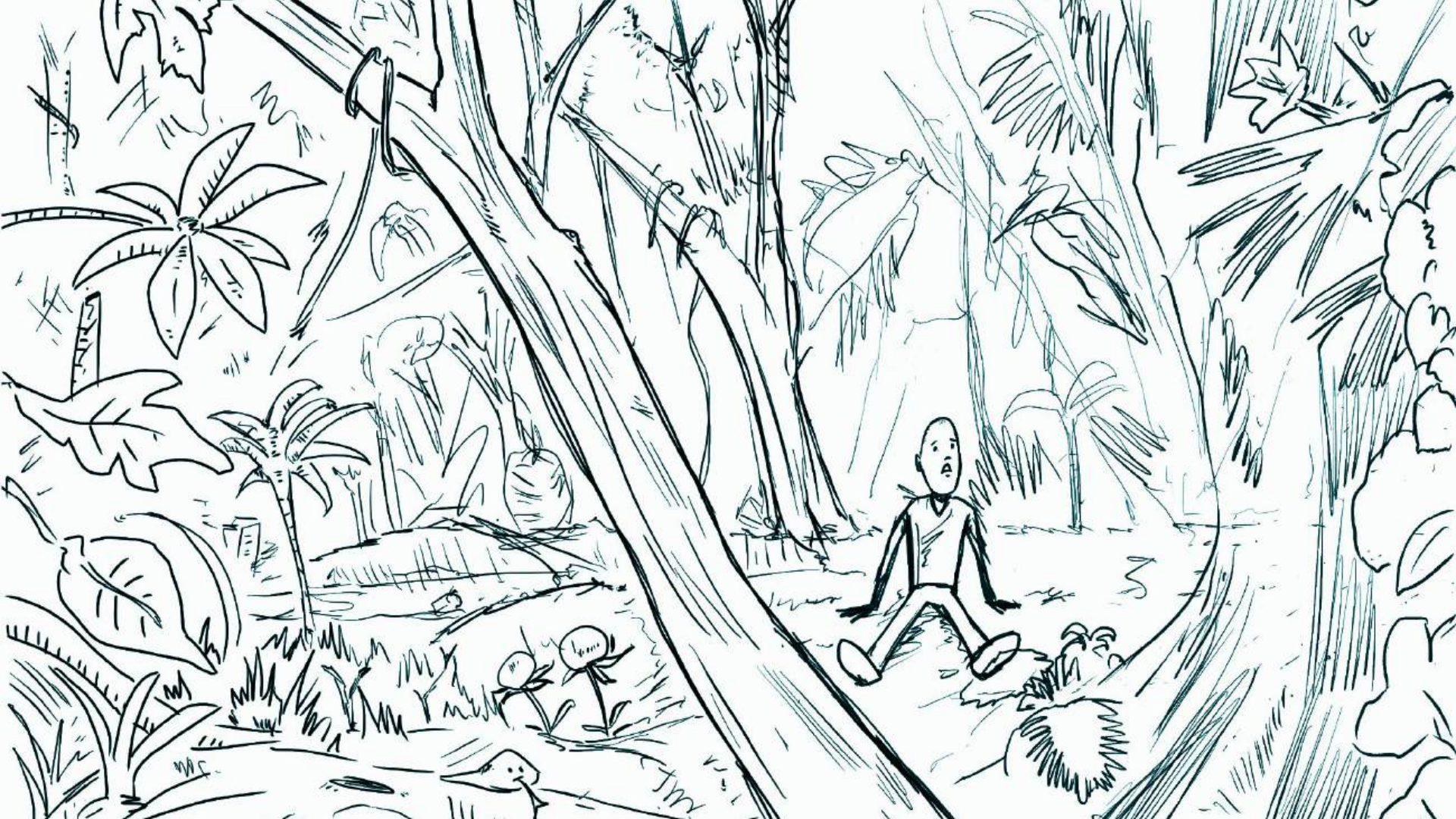




FASTER,
SAM!



* FOR OUR SHAREHOLDERS





A-ARE
YOU FRIEND
OR FOE?

BRO, I GOT
CLAWS!!

Three simple rules or **three laws of the jungle** to:

Use AI **more responsibly**
& get **better results**

We need a
mindset shift

The Atomic Human

Understanding
Ourselves in
the Age of AI



'The clarity, authority, wit and insight Lawrence brings to bear are like torches shining into the turbulent darkness'
Stephan Fry

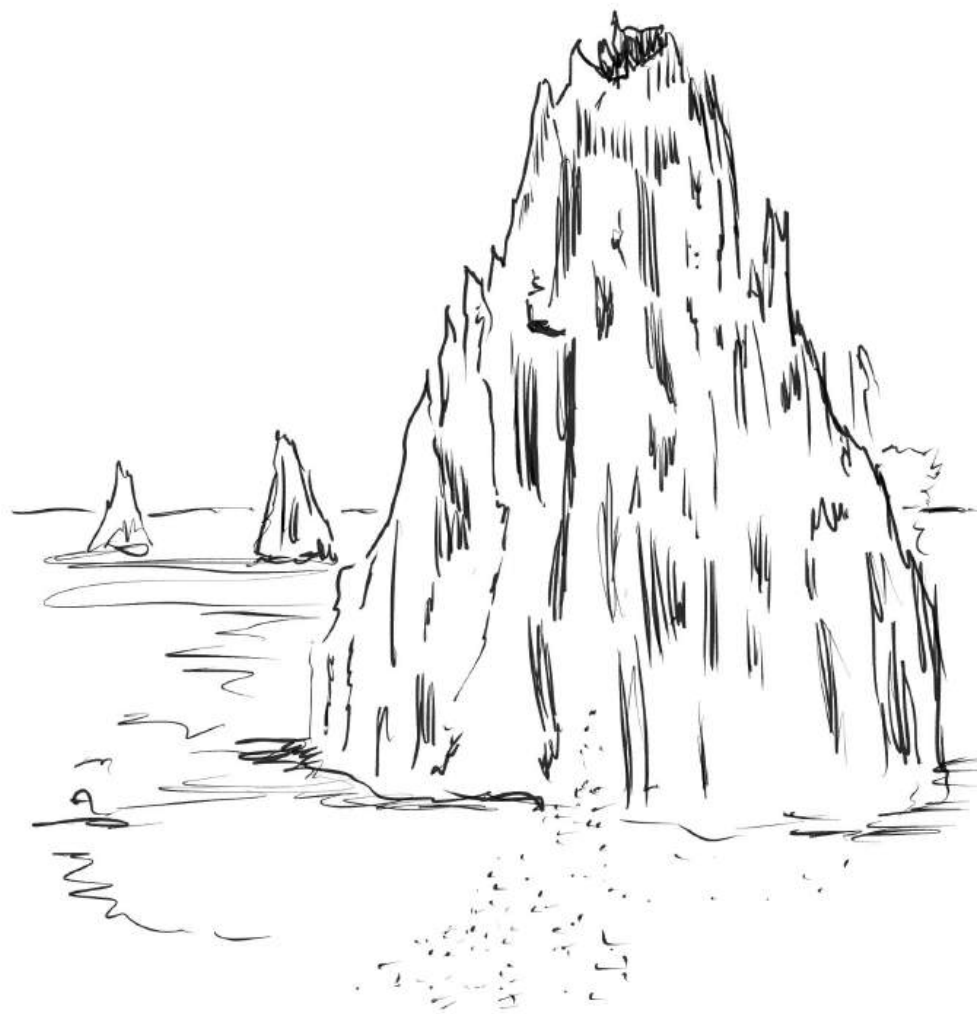
Neil D.
Lawrence

When simple intelligences...

...work in unison...

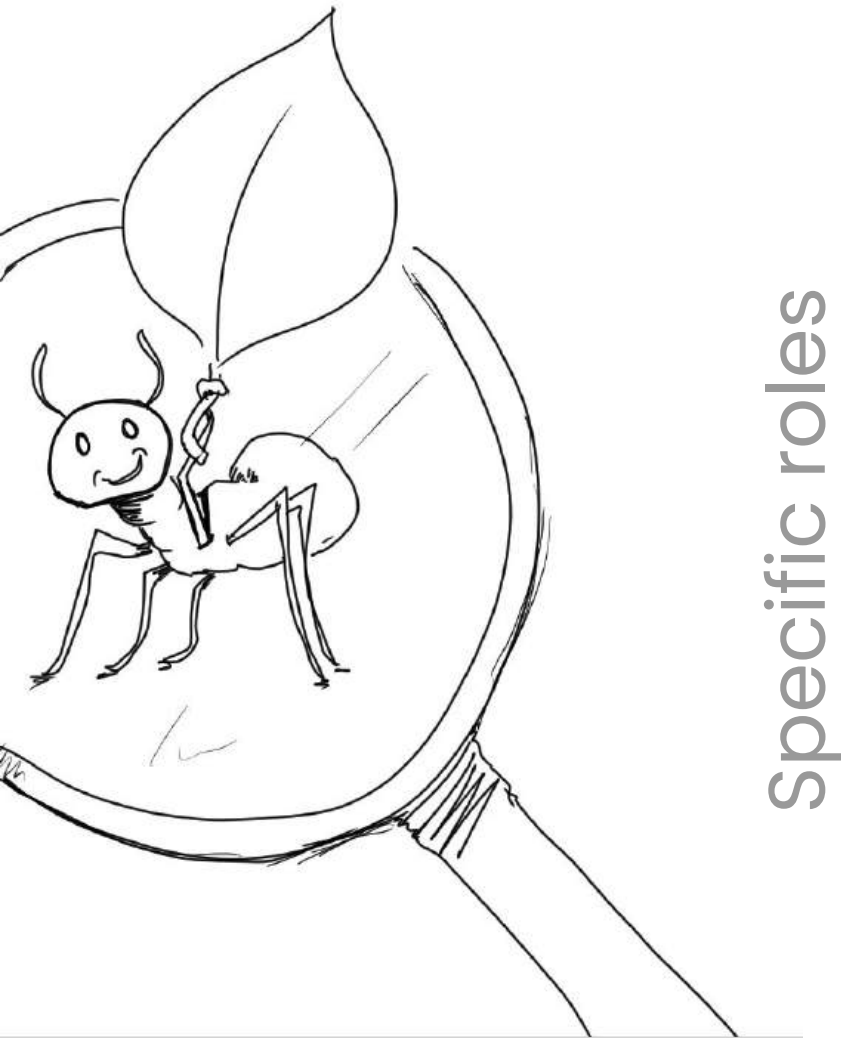
...and looked at from afar...

...give the **impression** of **high** intelligence.





Simple intelligence



Specific roles

Queen – Lays eggs to ensure colony reproduction.

Drone – Mates with the queen, then dies.

Worker – General maintenance and care tasks.

Forager – Searches for and collects food outside the nest.

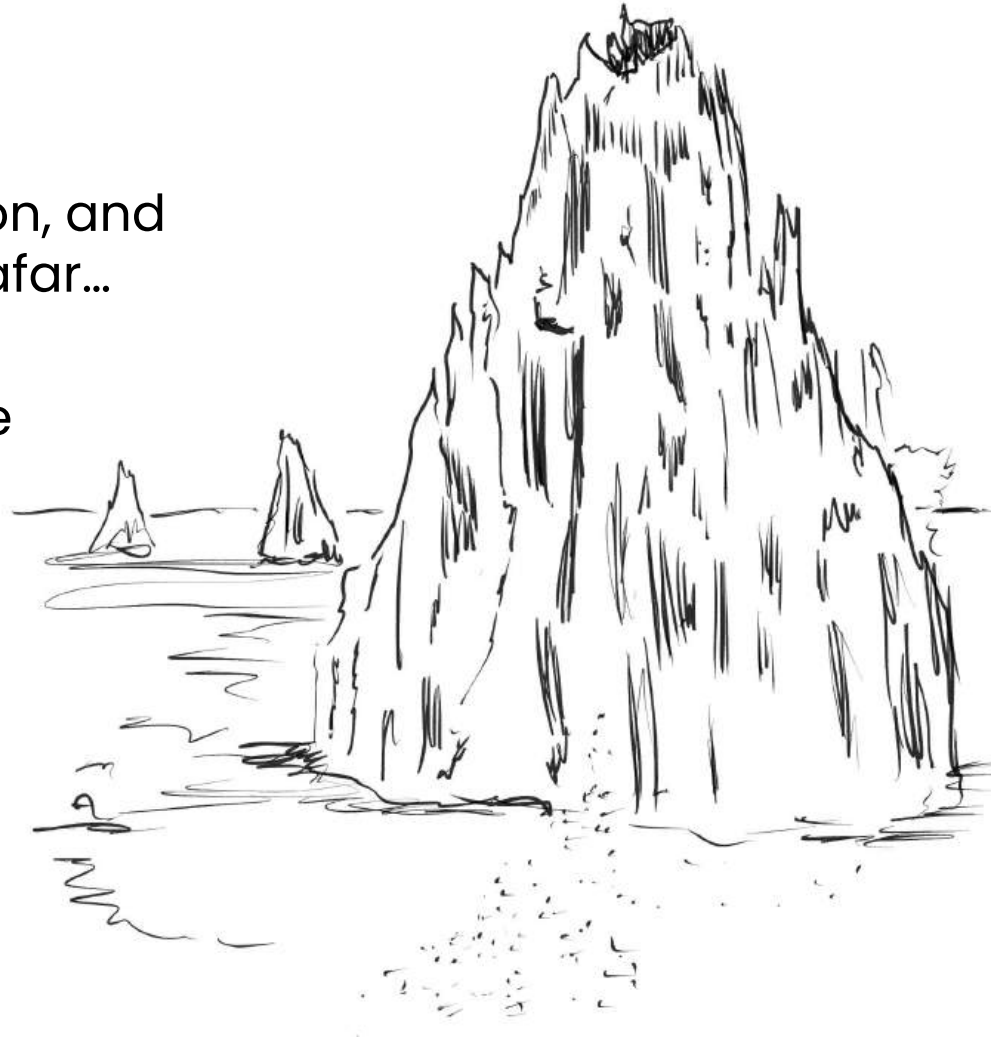
Nurse – Feeds and tends to larvae and pupae.

Soldier – Defends the colony from threats.

...

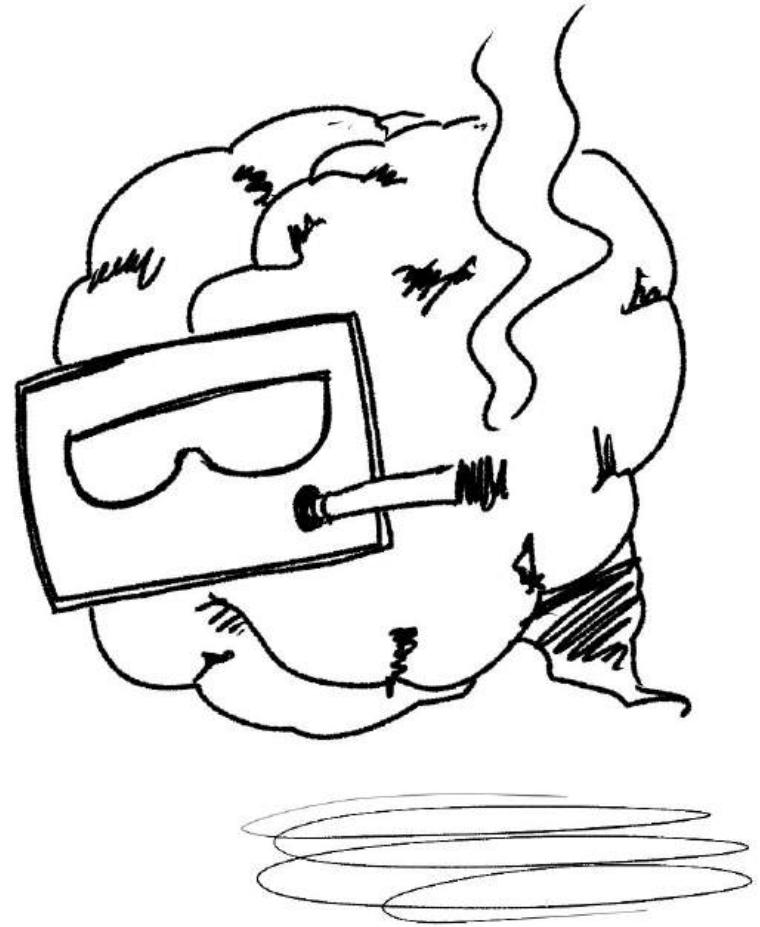
Working in unison, and
looked at from afar...

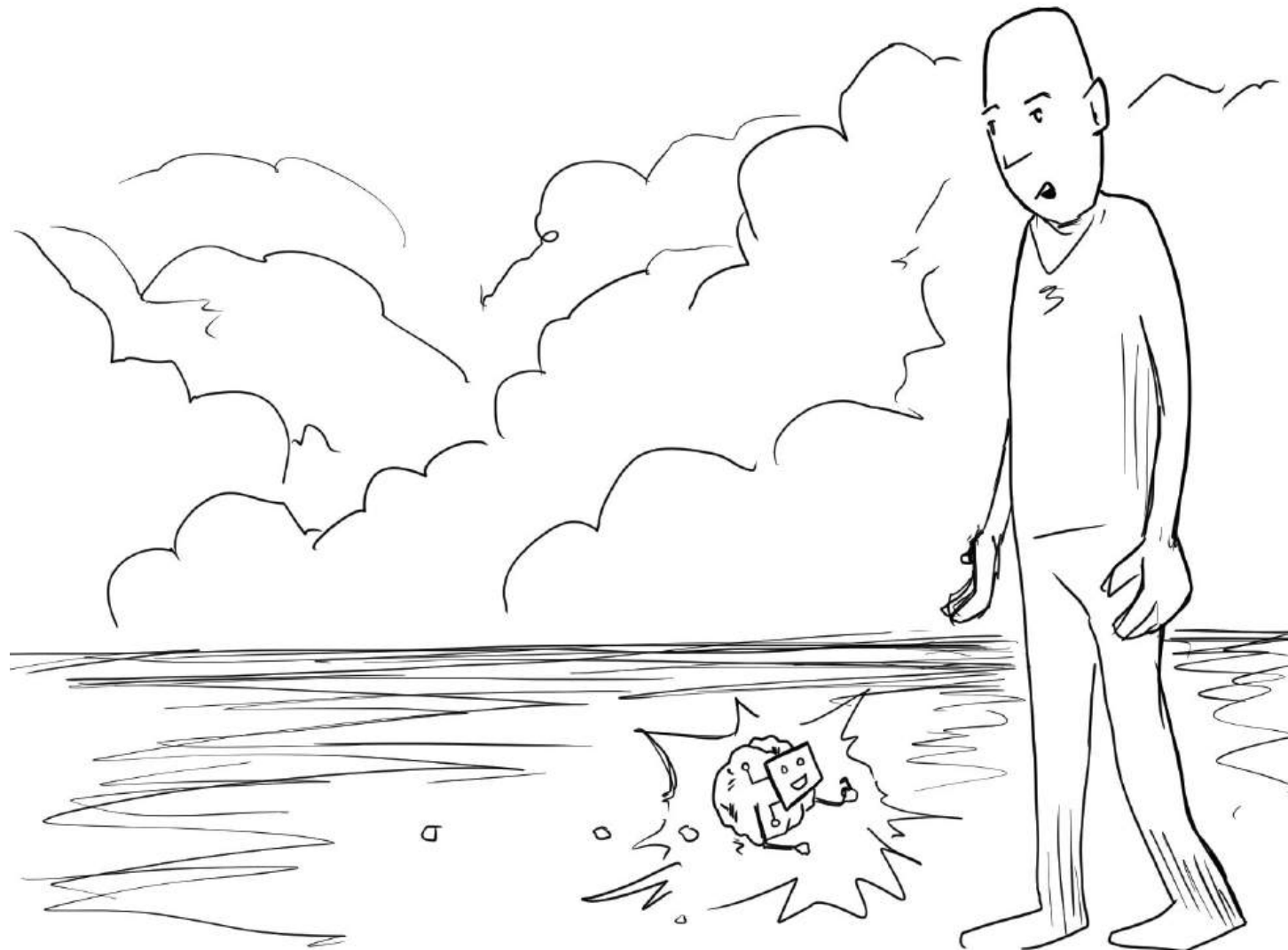
high intelligence



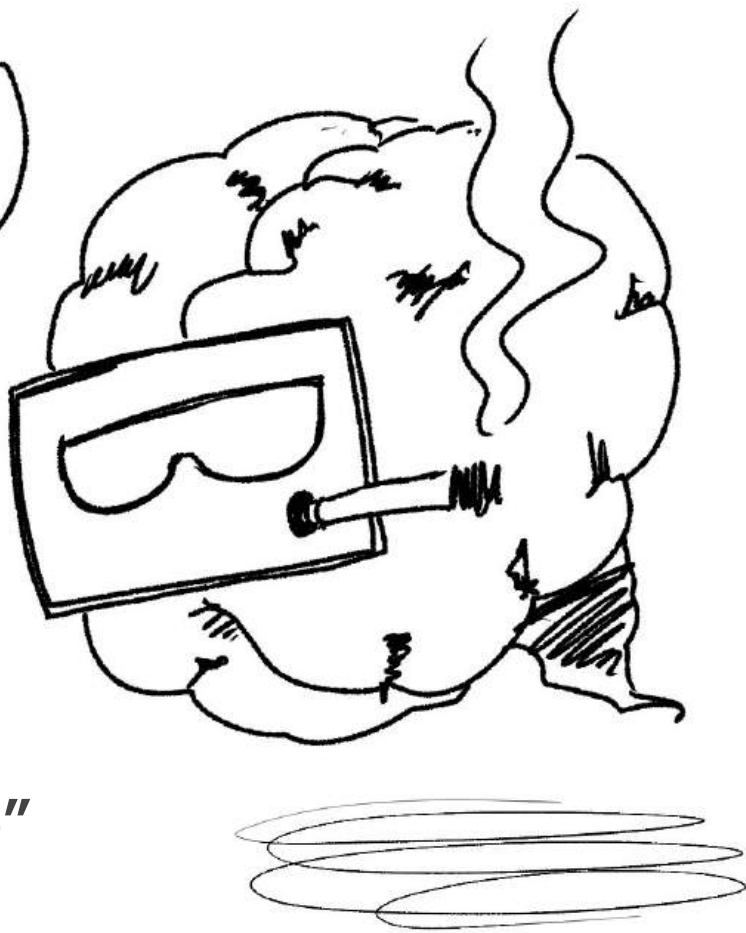
Ignore the hype.

Downsize expectations
of AI

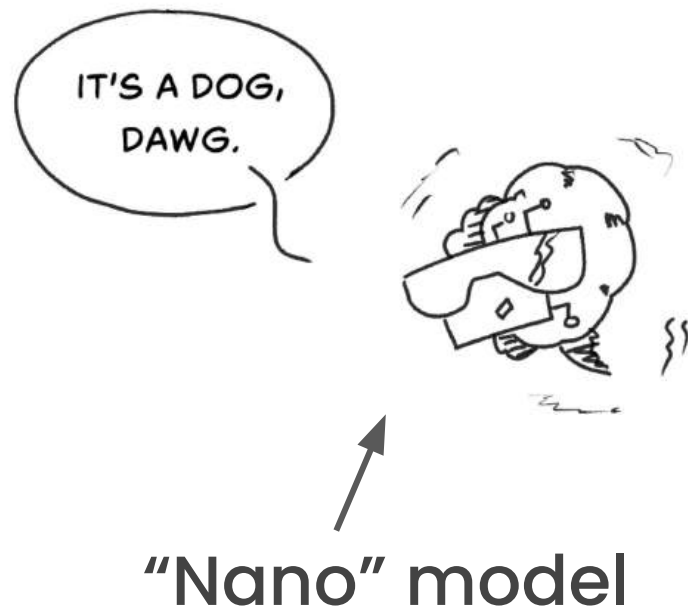
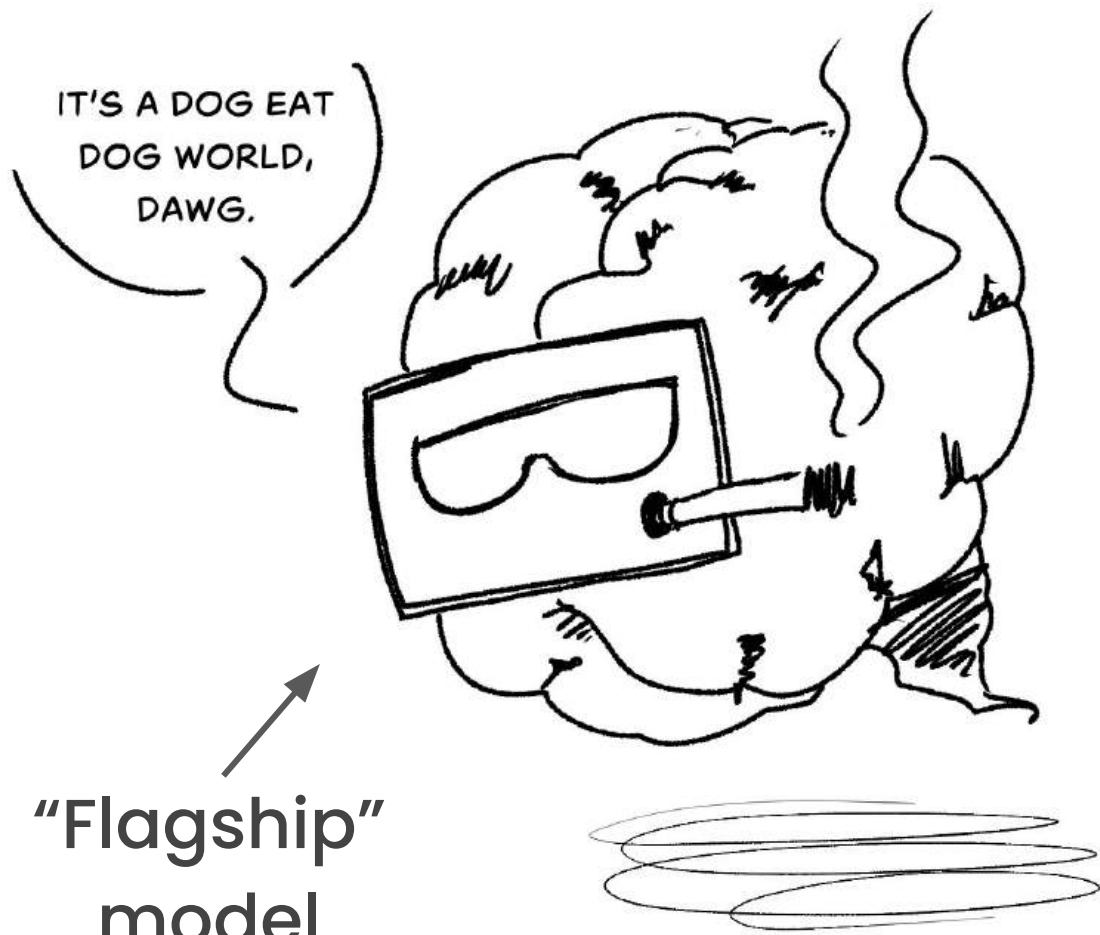




IT'S A DOG EAT
DOG WORLD,
DAWG.



↑
"Flagship"
model



Input Modalities

Text

Image

File

Audio

Output Modalities

Models

557 models

Reset Filters

Filter models

Latency: Low to High

☰

☰

Arcee AI: Maestro Reasoning

27M tokens

Maestro Reasoning is Arcee's flagship analysis model: a 32B-parameter derivative of Llama 3.1 70B, fine-tuned with DPO and chain-of-thought RL for step-by-step logic. Compared to the...

Arcee AI | 131K context | \$0.90/M input tokens | \$3.30/M output tokens

Llama 3 8B Lunar

3.1B tokens

Llama 3 8B is a versatile generalist and roleplaying model based on Llama 3. It's a strategic merge of Llama 3 models, designed to balance creativity with improved logic and general knowledge. ...

Llama 3 | 8K context | \$0.04/M input tokens | \$0.05/M output tokens

Ministral 8B

982M tokens

Ministral 8B is an 8B parameter model featuring a unique interleaved sliding-window attention mechanism for faster, memory-efficient inference. Designed for edge use cases, it supports up to 128...

Ministral | 128K context | \$0.03/M input tokens | \$0.04/M output tokens

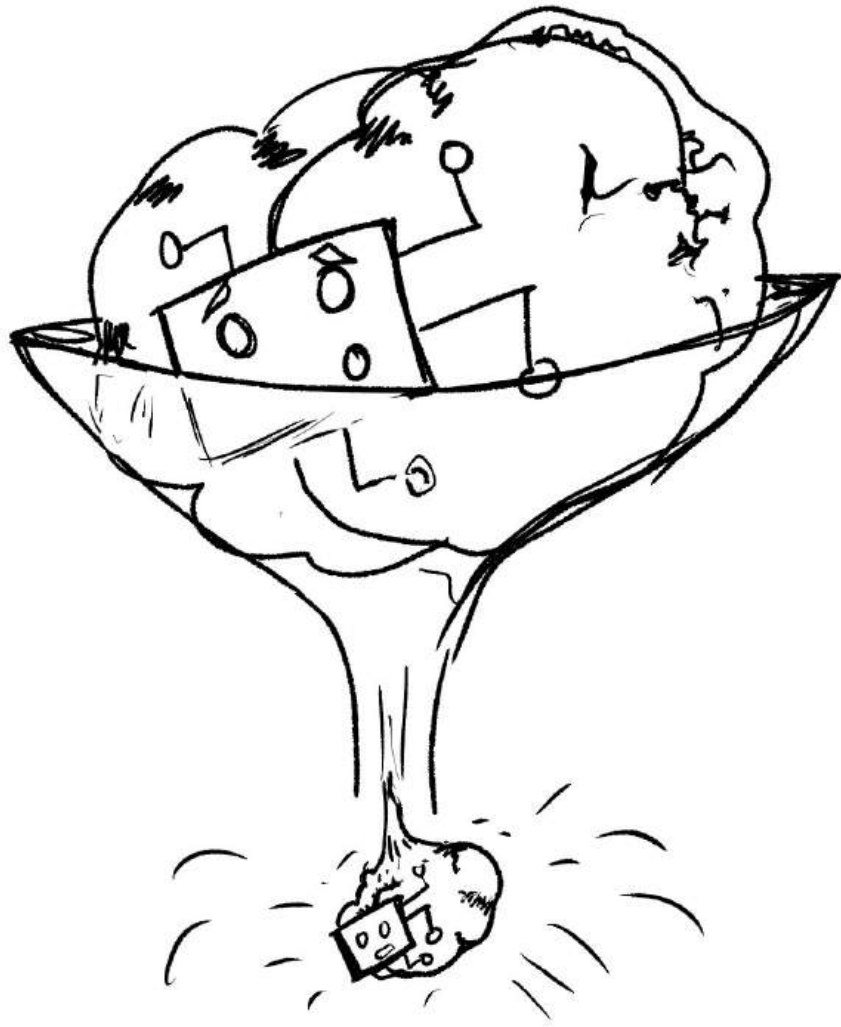
ChatGPT

Get Plus

What's on your mind today?

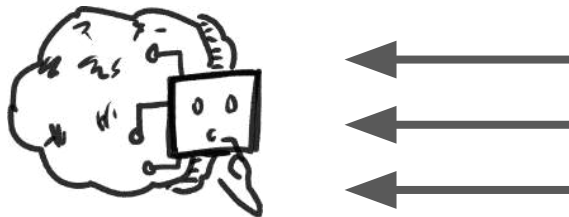
+ Ask to chat





What are small models?

Small models
are **distillations** of
larger models.



“[Utilises] **extracted rationales** as additional, richer information to **train small models**.”

“Distilling allows ... specific smaller models that outperform LLMs.”

Paper: [Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes](#) (2023)

Thought for 30s

- I need to gather information about energy use in training various machine learning models like DistilBERT and FLAN-T5. Since up-to-date data is helpful, I'll search for articles on their energy consumption and training costs. I'll prioritize statements with the most weight or evidence to back them up. Let's run some searches!

🌐 Running web searches for energy consumption in models

[arxiv.org](#) [adasci.org](#) [www.sciencedirect.com](#) [13 more](#)

- I need to give an estimate of energy consumption for training a small distilled model and convert it to something relatable, like the energy used by average US homes annually. I'll use existing research on distillation energy savings, energy costs for fine-tuning, and specific examples like DistilBERT. Let's check the papers!

🌐 Searching energy costs for training distilled models

[www.nature.com](#) [aclanthology.org](#) [tepperspectives.cmu.edu](#) [11 more](#)

- Distilled models like DistilBERT can reduce energy use by about 40-60% compared to larger models like BERT. I'll give ballpark estimates, ranging from tens to hundreds of kWh for training a small distilled model (10M-100M parameters), depending on factors like dataset and hardware. I'll cite relevant studies.
- I have a concrete example: 143 kWh is approximately 0.013 homes-year when divided by 10,791. I need to compute more similar estimates and perform the necessary arithmetic step-by-step. The calculations will involve converting energy into relatable units like home electricity usage. Let's process these numbers now!

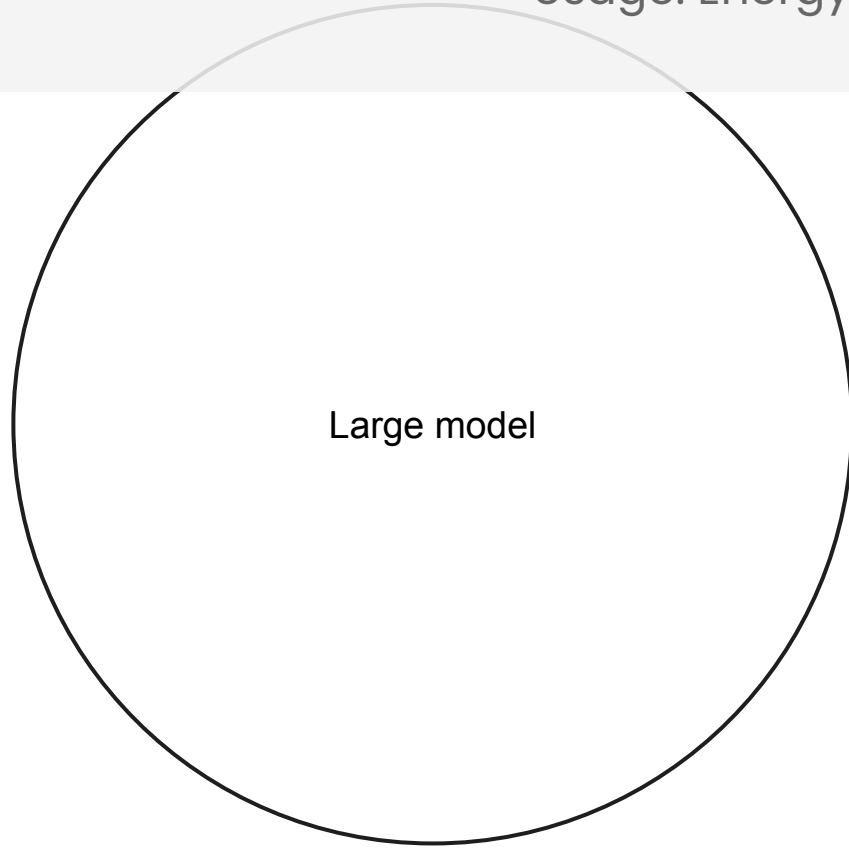
Training: Energy requirements

Large model

Small model



Usage: Energy requirements

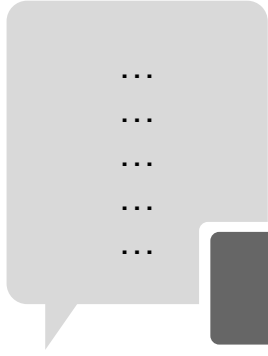


Large model

Small model



But what about
accuracy?



Long, complex
prompts

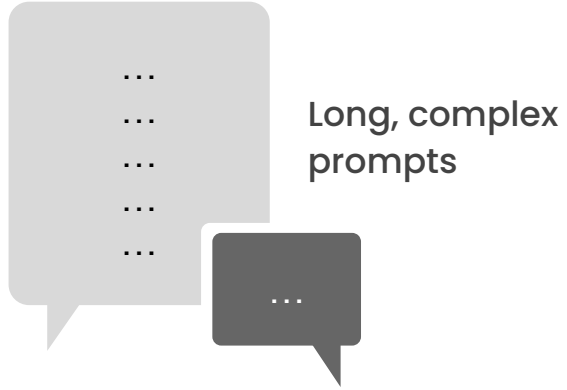
Large

Small

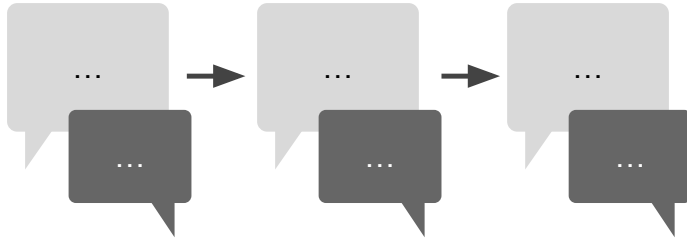
Some
variance

Large

Small



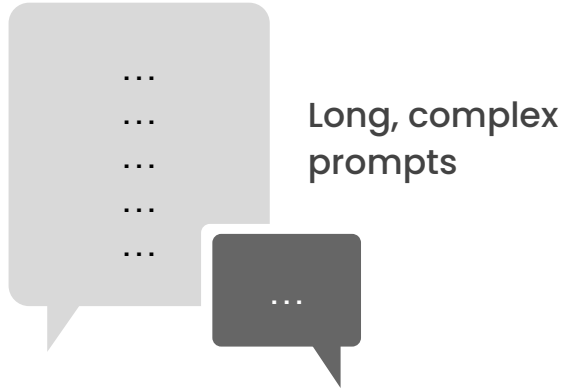
Some
variance



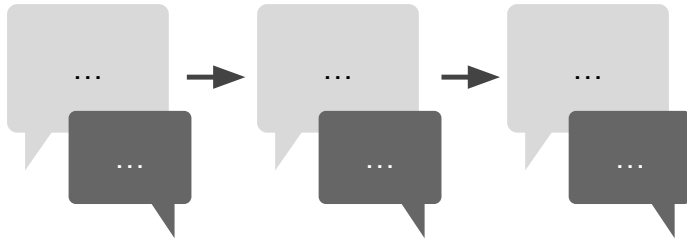
Multiple simple prompts

Large

Small



Some
variance

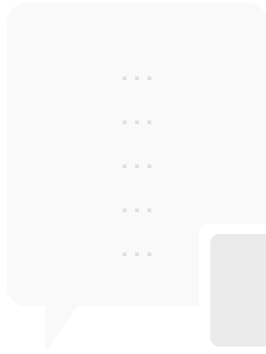


Multiple simple prompts

No difference

Large

Small



Few long, complex prompts



Rule #1: Think small
Small models, small, simple tasks

Some variance



More simple prompts

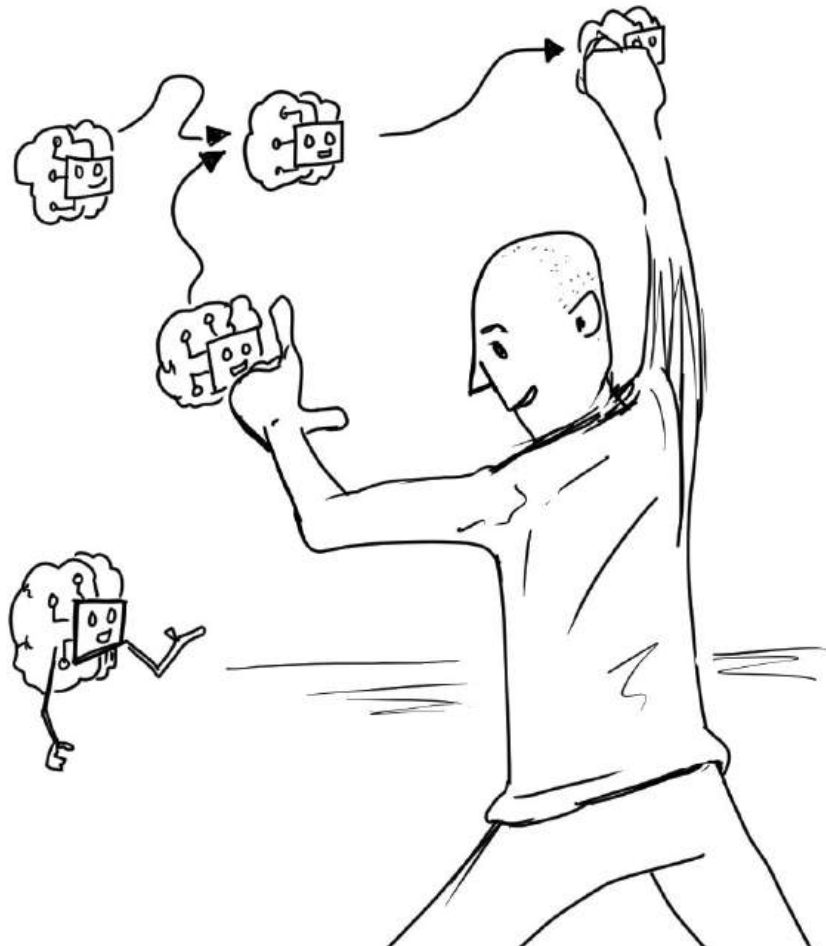
Papers to support this:

Small Language Models are Good Too

Paper: [An Empirical Study of Zero-Shot Classification](#)

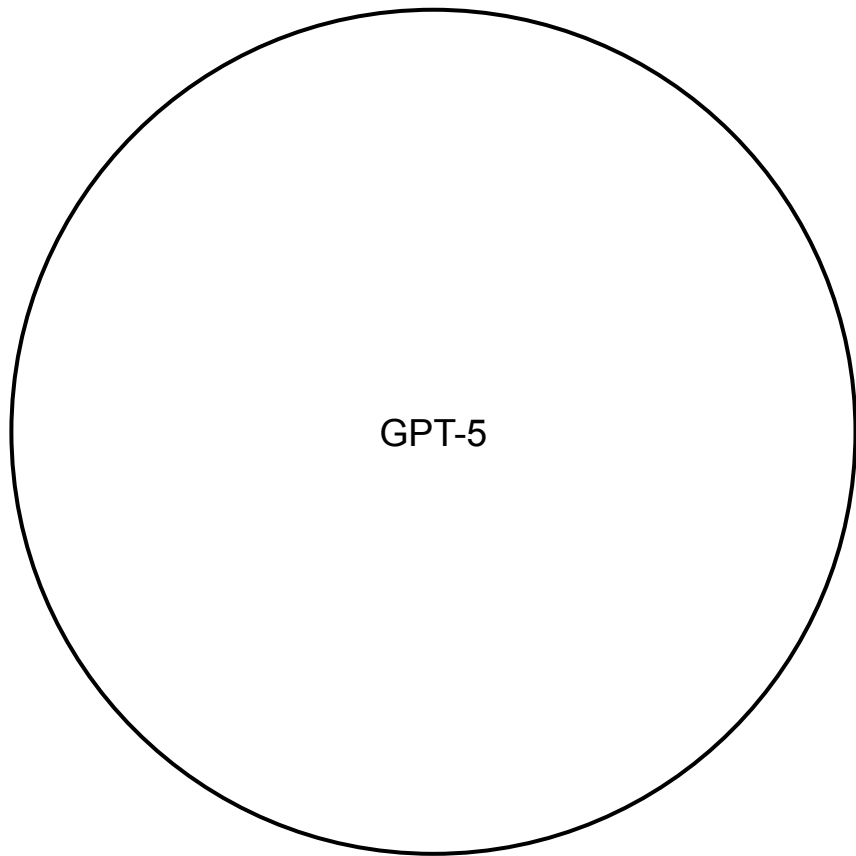
[Fine-tuning small models](#)
(NVidia)

No difference



Rule #2: Think in terms of workflows





GPT-5

vs

Smallest
model



Attractions Visited: We went on Neverland's BBQ Flight and another ride.,
Overall Experience at the Park: We visited on Halloween 31.10.12, a good park & all the usual Smoking Bones stuff, peak time so it was busy, some observations, park in the multi storey at Smoking Bones village instead of the main car park, it's half the price, closer, no traffic queues to get in, and no people queues for security arriving from that side toilets were generally smelly & need more cleaning attendants when the park is that busy others smoking inconsiderately got irritating around the park, designated areas would be good but this is France so pigs might fly first the fast pass system really should be much better, as it is in the US parks one ride Neverland's BBQ Flight had a 30 min queue just to *get* one ! We still had 20 mins to get a ride later on with an FP, which isn't very 'fast'. There should use bar codes on their day tickets from the booths (like the tickets bought online) rather than mag stripes, as the latter cause a good proportion of the delay in obtaining an FP, with people having trouble getting the flash cards read. Dreams show was a highlight, bit of a scrum to get you've got young'uns hold onto them ! The pie-eating competition the day had gone awry, escalating into a full-blown food fight, leading to a general mess and a shortage of food.



Attractions Visited: We went on Neverland's BBQ Flight and another ride., Overall Experience at the Park: We visited on Halloween 31.10.12, a good park & all the usual Smoking Bones stuff, peak time so it was busy, some observations, park in the multi storey at Smoking Bones village instead of the main car park, it's half the price, closer, no traffic queues to get in, and no people queues for security arriving from that side toilets were generally smelly & need more cleaning attendants when the park is that busy others smoking inconsiderately got irritating around the park, designated areas would be good but this is France so pigs might fly first the fast pass system really should be much better, as it is in the US parks one ride (Neverland's BBQ Flight) had a 30 min queue just to *get* one ! We still had to wait 25 mins to get on a ride later on with an FP, which isn't very 'fast'. Smoking Bones should use bar codes on their day tickets from the booths (like the tickets bought online) rather than mag stripes, as the latter cause a good proportion of the delay in obtaining an FP, with people having trouble getting the flimsy cards read.Dreams show was a highlight, bit of a scrum to get out after so if you've got young'uns hold onto them ! The pie-eating competition earlier in the day had gone awry, escalating into a full-blown food fight, which had led to a general mess and a shortage of food.

→ Insight & sentiment

→ Insight & sentiment

→ Insight & sentiment

Example user response

Attractions Visited: We went on Neverland's BBQ Flight and another ride., Overall Experience at the Park: We visited on Halloween 31.10.12, a good park & all the usual Smoking Bones stuff, peak time so it was busy, some observations, park in the multi storey at Smoking Bones village instead of the main car park, it's half the price, closer, no traffic queues to get in, and no people queues for security arriving from that side toilets were generally smelly & need more cleaning attendants when the park is that busy others smoking inconsiderately got irritating around the park, designated areas would be good but this is France so pigs might fly first the fast pass system really should be much better, as it is in the US parks one ride (Neverland's BBQ Flight) had a 30 min queue just to *get* one ! We still had to wait 25 mins to get on a ride later on with an FP, which isn't very 'fast'. Smoking Bones should use bar codes on their day tickets from the booths (like the tickets bought online) rather than mag stripes, as the latter cause a good proportion of the delay in obtaining an FP, with people having trouble getting the flimsy cards read.Dreams show was a highlight, bit of a scrum to get out after so if you've got young'uns hold onto them ! The pie-eating competition earlier in the day had gone awry, escalating into a full-blown food fight, which had led to a general mess and a shortage of food.

Messy, multiple points, multiple sentiments

Attractions Visited: We went on Neverland's BBQ Flight and another ride., Overall Experience at the Park: We visited on Halloween 31.10.12, a good park & all the usual Smoking Bones stuff, peak time so it was busy, some observations, park in the multi storey at Smoking Bones village instead of the main car park, it's half the price, closer, no traffic queues to get in, and no people queues for security arriving from that side toilets were generally smelly & need more cleaning attendants when the park is that busy others smoking inconsiderately got irritating around the park, designated areas would be good but this is France so **pigs might fly first** the fast pass system really should be much better, as it is in the US parks one ride (Neverland's BBQ Flight) had a 30 min queue just to *get* one ! We still had to wait 25 mins to get on a ride later on with an FP, which isn't very 'fast'. Smoking Bones should use bar codes on their day tickets from the booths (like the tickets bought online) rather than mag stripes, as the latter cause a good proportion of the delay in obtaining an FP, with people having trouble getting the flimsy cards read.Dreams show was a highlight, bit of a scrum to get out after so if you've got young'uns hold onto them ! The pie-eating competition earlier in the day had gone awry, escalating into a full-blown food fight, which had led to a general mess and a shortage of food.

Colloquialism

%PERFECT OUTPUT%

```
[
  {
    "specific_point": "Toilets were smelly and need more cleaning attendants
during busy times.",
    "sentiment": "Negative",
    "tags": ["cleanliness", "restrooms", "staffing"]
  },
  {
    "specific_point": "Inconsiderate smoking around the park; suggests
designated smoking areas.",
    "sentiment": "Negative",
    "tags": ["smoking", "guest_behavior", "suggestion"]
  },
  {
    "specific_point": "Fast pass system inefficient; 30-minute queue to
obtain one and still big wait times with fast pass. Suggests because of flimsy
mag stripe ticket machines.",
    "sentiment": "Negative",
    "tags": ["fast_pass", "queues", "system_issue"]
  },
  {
    "specific_point": "Dreams show was a highlight, but exiting afterward
was chaotic and crowded.",
    "sentiment": "Negative",
    "tags": ["shows", "experience", "crowding"]
  },
  {
    "specific_point": "Pie-eating competition escalated into a food fight,
causing general mess and food shortage.",
    "sentiment": "Negative",
    "tags": ["event_issue", "cleanliness", "food_shortage"]
  }
]
```





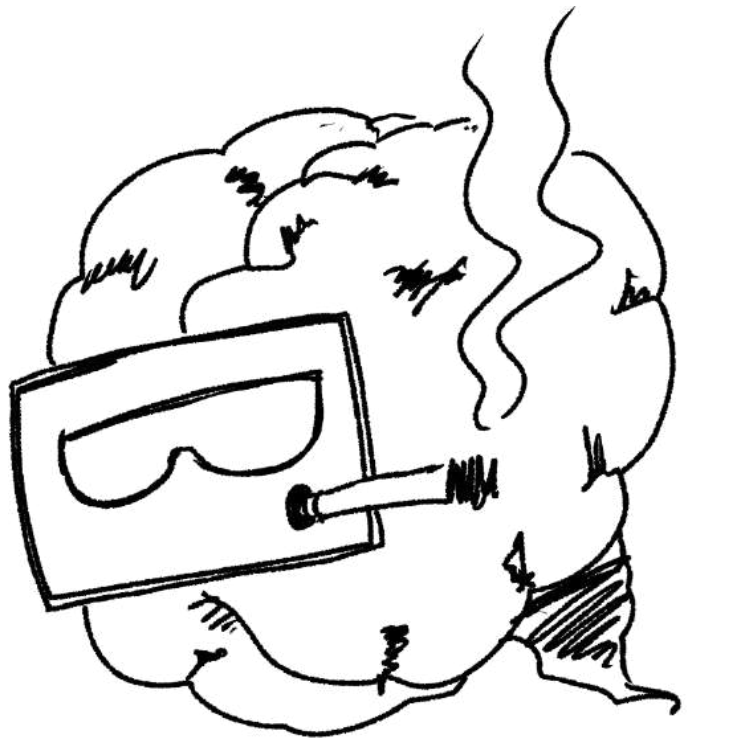
```
[
  {
    "specific_point": "Toilets were smelly and need more cleaning attendants during busy times.",
    "sentiment": "Negative",
    "tags": ["cleanliness", "restrooms", "staffing"]
  },
  {
    "specific_point": "Inconsiderate smoking around the park; suggests designated smoking areas.",
    "sentiment": "Negative",
    "tags": ["smoking", "guest_behavior", "suggestion"]
  },
  {
    "specific_point": "Fast pass system inefficient; 30-minute queue to obtain one and still big wait times with fast pass. Suggests because of flimsy system.",
    "sentiment": "Negative",
    "tags": ["fast_pass", "queues", "system_issue"]
  },
  {
    "specific_point": "Dreams show was a highlight, but exiting afterward was chaotic and crowded.",
    "sentiment": "Mixed",
    "tags": ["shows", "experience", "crowding"]
  },
  {
    "specific_point": "Pie-eating competition escalated into a food fight, causing general mess and food shortage.",
    "sentiment": "Negative",
    "tags": ["event_issue", "cleanliness", "food_shortage"]
  }
]
```

Description of the perfect response:

- Formatted as JSON
- How many insights?
- What insights captured?
- Did it avoid taking the colloquialism literally?
- How many positive, negative, neutral points?

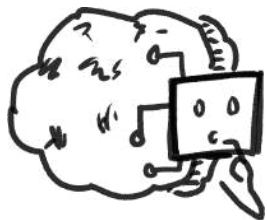


Single
prompt



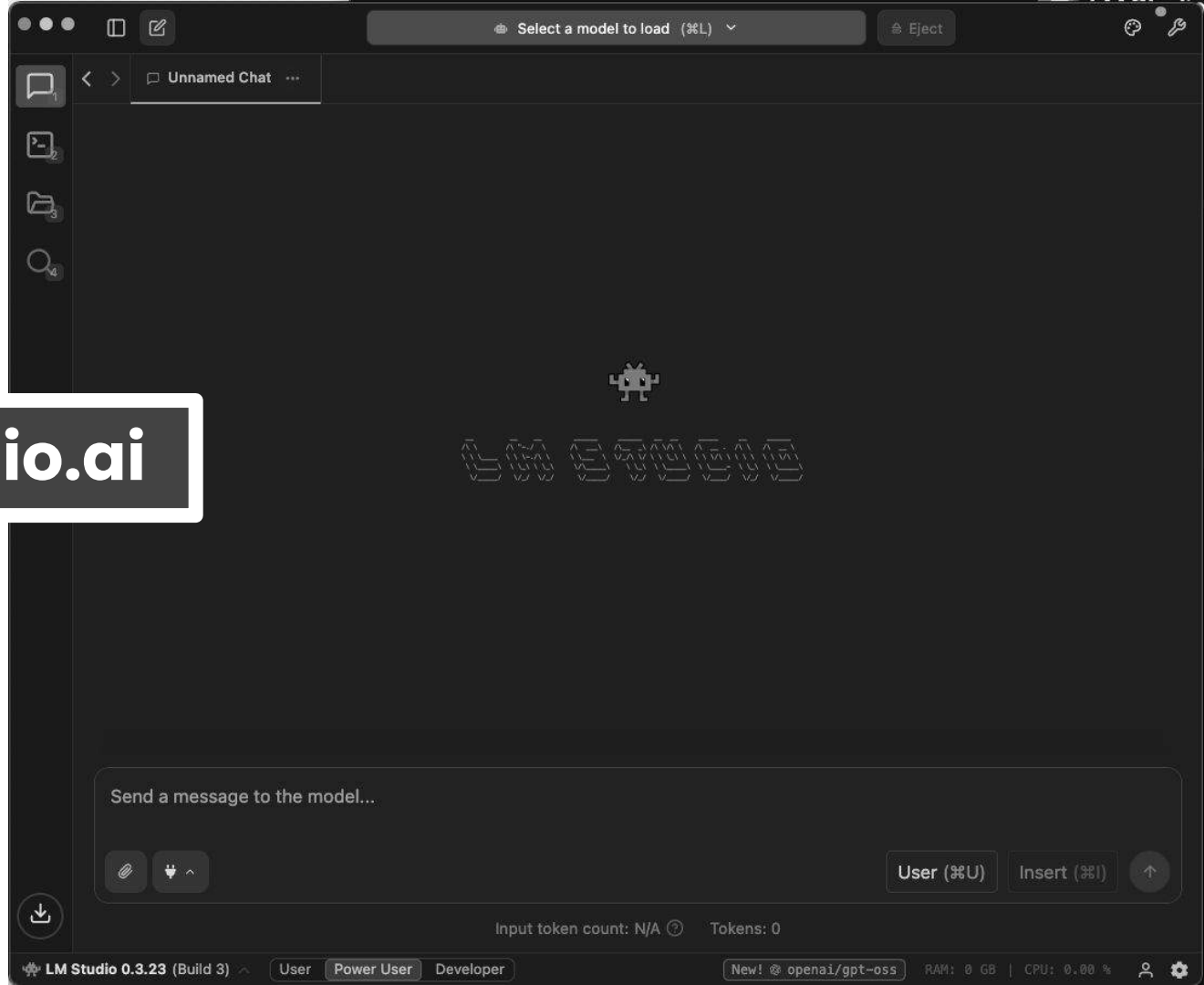
Acceptable

GPT5

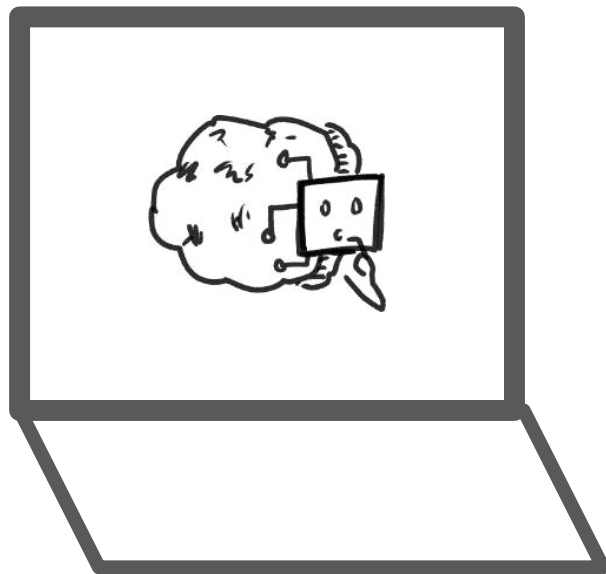


lmstudio.ai

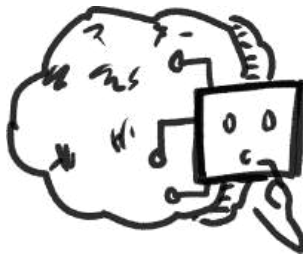
Free.
And easy.



500mb

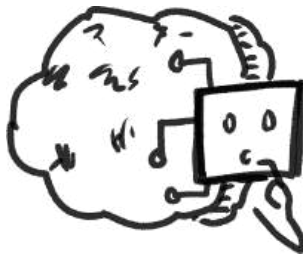


Single
prompt



500mb

Single
prompt

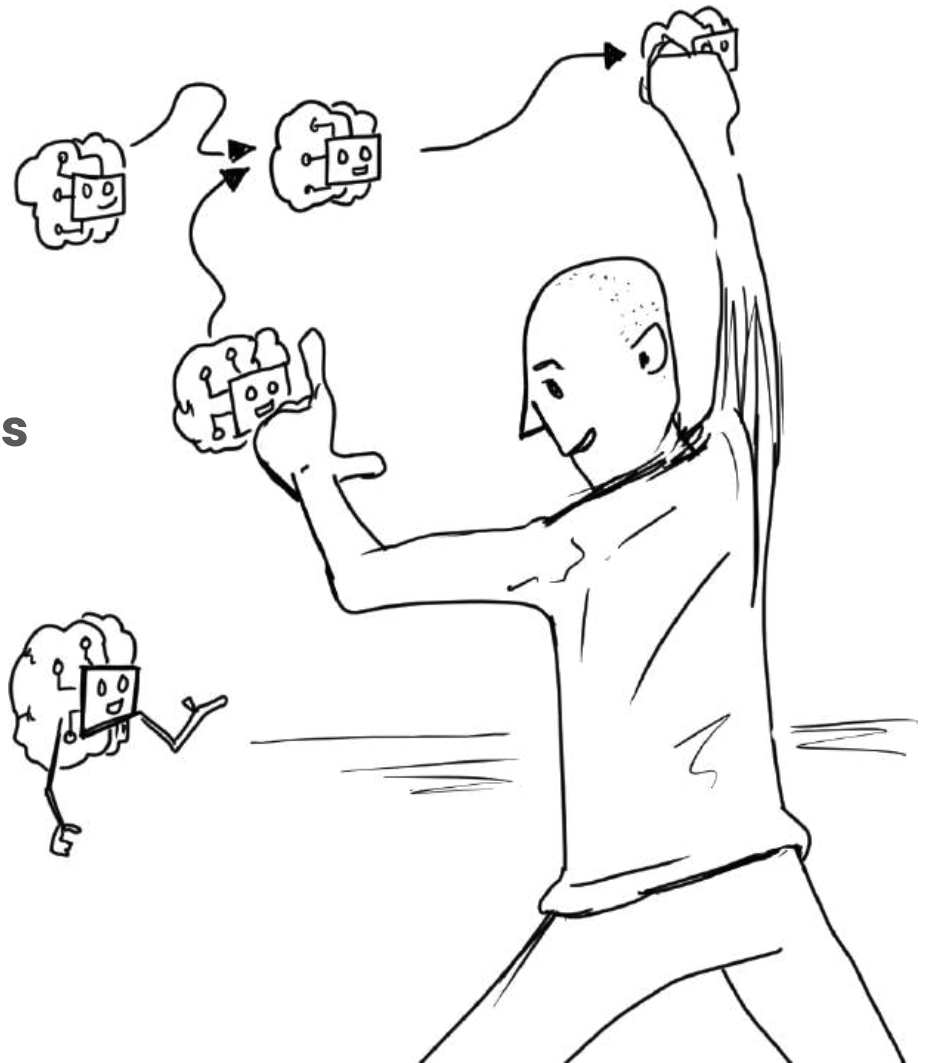


Bad

500mb

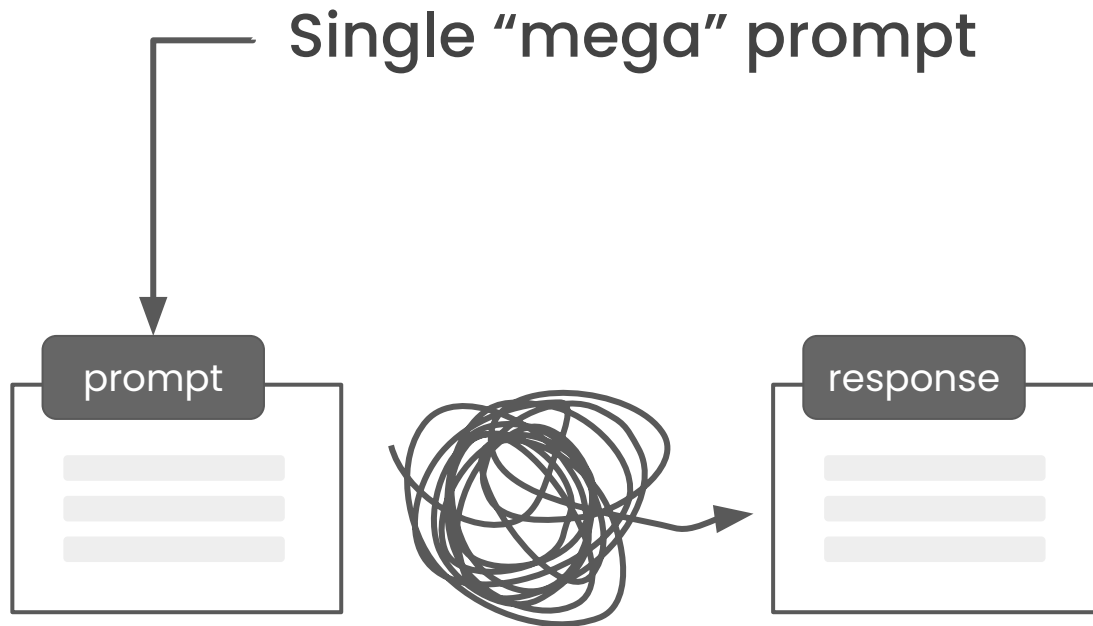
Rule #1: Think small
Small models, **small, simple tasks**

Rule #2: Think in terms of
workflows

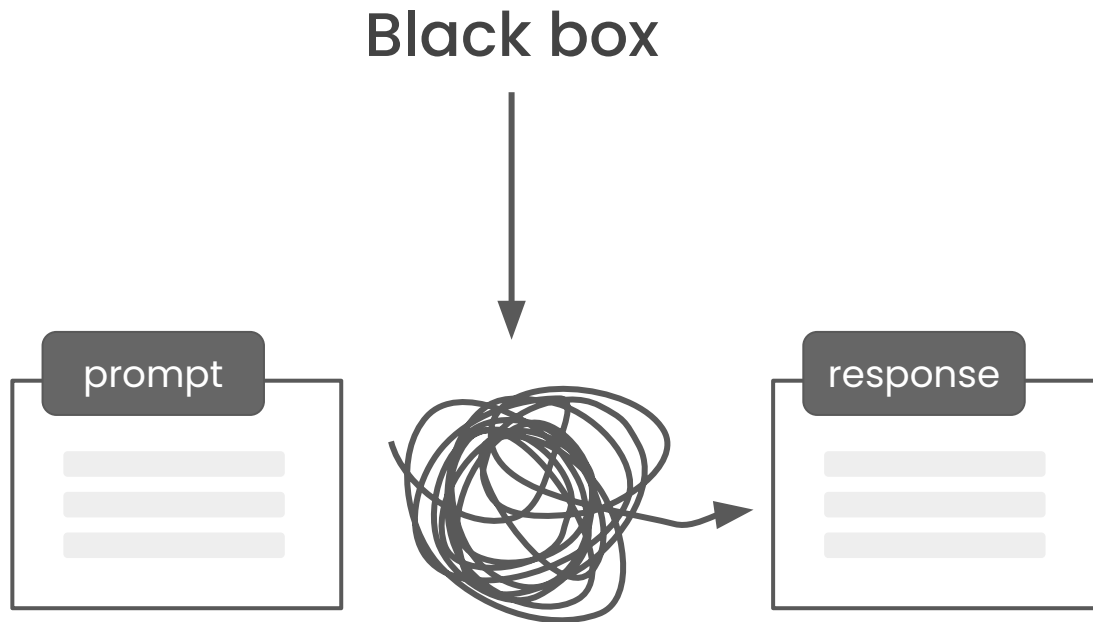


But how do you figure out
the **steps** of a workflow?

The current process...

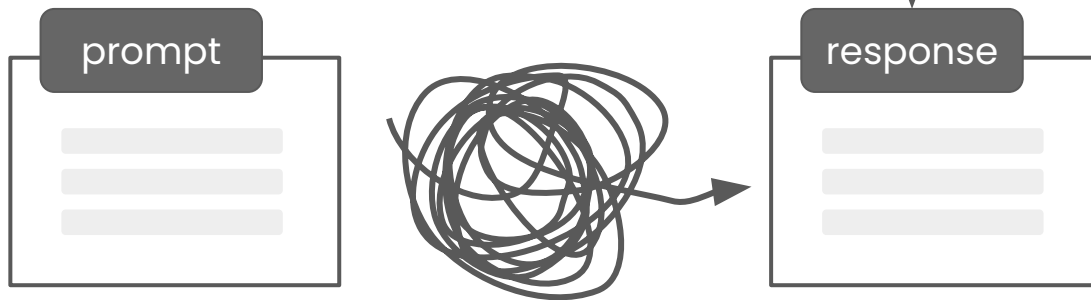


The current process...

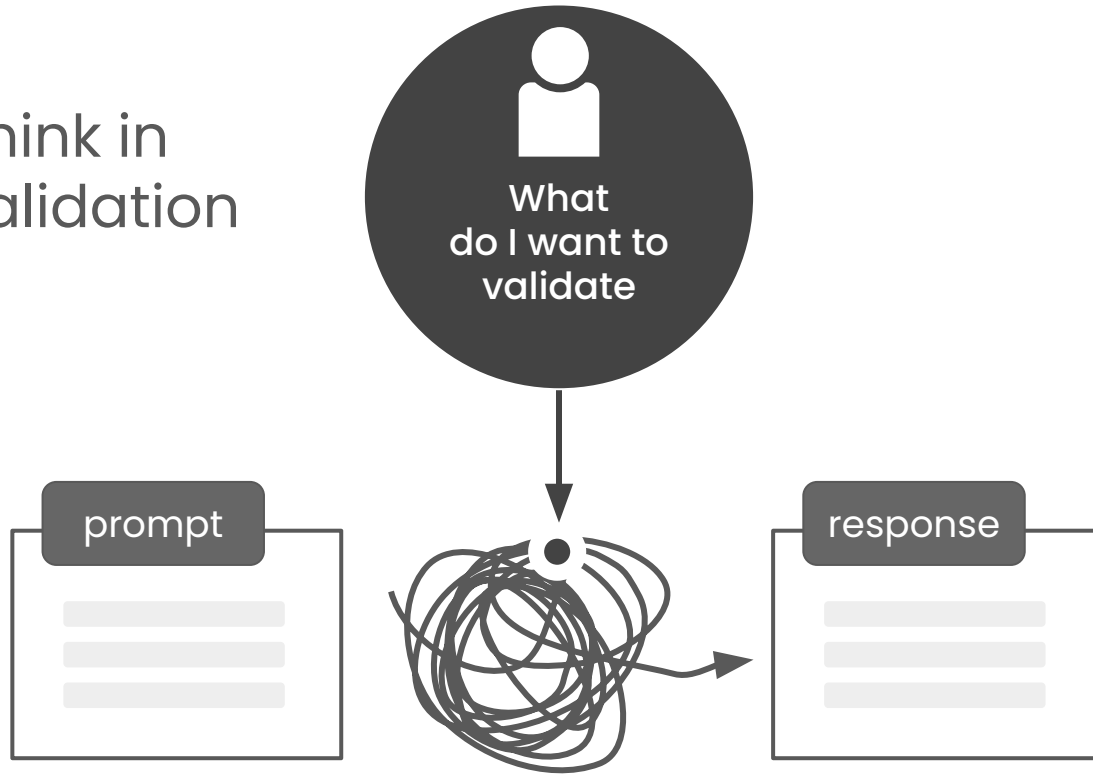


The current process...

Some output




Rule #3: Think in terms of validation





1 CODE { }
2 [] { }
3 } — ,
· — in
· []
· —
· — ()
· { — }
· — (-) :
· —
· —
· (—) — (—)
· — [] —
· — [—]
1000 — [—] :

Running
code



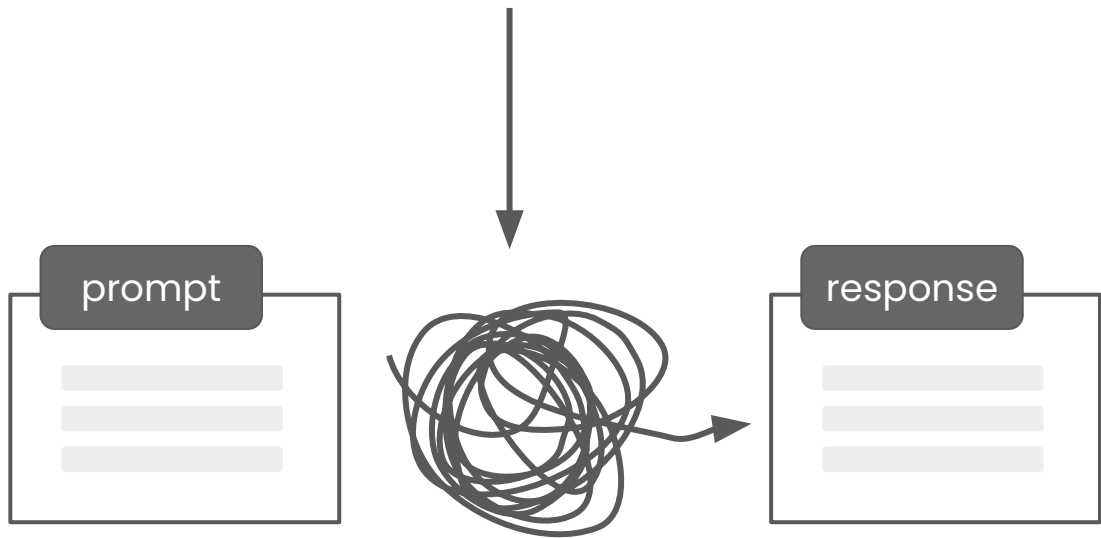
```
1 SET total TO 0
2 SET all_heights TO [1.87, 1.48, 1.57, 1.91, 2.01]
3 FOR EACH height FROM all_heights DO
4     IF height > 1.57 AND height <= 1.91 THEN
5         SET total TO total + 1
6     END IF
7 END FOR EACH
8
```

Breakpoint: insert a **pause** in execution,
based on **what you want to validate**.



Frozen in time, allows
to **validate** the state

Identify a
breakpoint



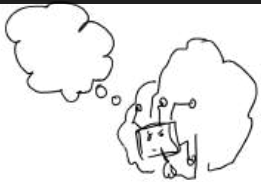
✓ *Thought for 5.53 seconds*

Okay, let's see. The user is talking about their experience at Smoking Bones Land. They mentioned visiting the BBQ Flight and another ride, with Halloween 10th in July. They note that the park was crowded, there were queues for food, kids areas were messy, and some people had trouble getting in. Also, they mention a fast pass system causing delays and suggest using bar codes instead of mag stripes.

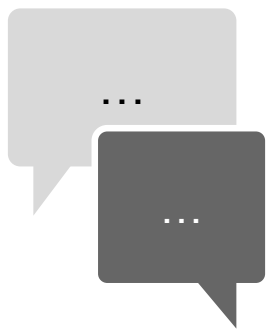
First, I need to identify specific points from their input. The user is highlighting several issues: overcrowding, food queue problems, mess from the pie-eating competition, and the fast pass system's delay. They also mention adding benches for waiting parents in kids areas.

Looking at the JSON structure provided, each specific point has a tag. So I need to map these points into the specified tags. The user mentioned "California park was very crowded" which is a negative sentiment. Then the food queues and mess from the competition are negatives. The fast pass system delay and bar code issue also contribute negatively. Finally, adding benches in kids areas is a suggestion but tagged as neutral.

Wait, the user's input has multiple points, so I need to list them all. Let me check again: overcrowding (negative), food queues (negative), mess from competition (negative), fast pass delay (negative), and bench suggestions (neutral). That covers all specific points.



Forget the formatting, just the **points**.



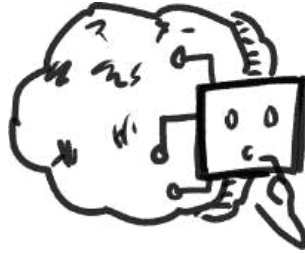
Insights
only



Review
accuracy

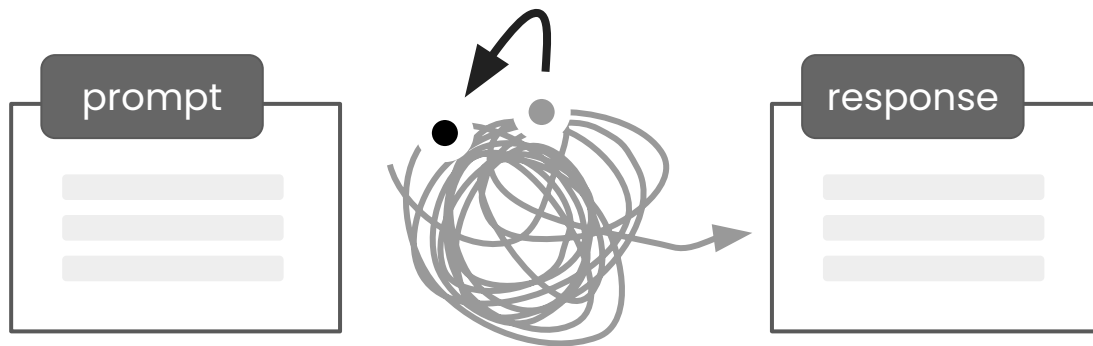


Format
only



Bad

Pick another
breakpoint



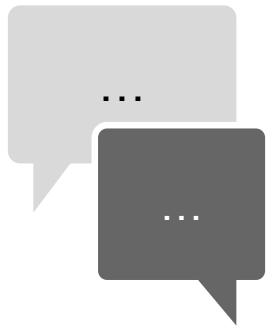
✓ *Thought for 5.53 seconds*

Okay, let's see. The user is talking about their experience at Smoking Bones Land. They mentioned visiting the BBQ Flight and another ride, with Halloween 10th in July. They note that the park was crowded, there were queues for food, kids areas were messy, and some people had trouble getting in. Also, they mention a fast pass system causing delays and suggest using bar codes instead of mag stripes.

First, I need to identify specific points from their input. The user is highlighting several issues: overcrowding, food queue problems, mess from the pie-eating competition, and the fast pass system's delay. They also mention adding benches for waiting parents in kids areas.

Looking at the JSON structure provided, each specific point has a tag. So I need to map these points into the specified tags. The user mentioned "California park was very crowded" which is a negative sentiment. Then the food queues and mess from the competition are negatives. The fast pass system delay and bar code issue also contribute negatively. Finally, adding benches in kids areas is a suggestion but tagged as neutral.

Wait, the user's input has multiple points, so I need to list them all. Let me check again: overcrowding (negative), food queues (negative), mess from competition (negative), fast pass delay (negative), and bench suggestions (neutral). That covers all specific points.



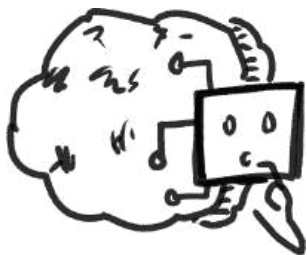
Negative
insights



Positive
insights

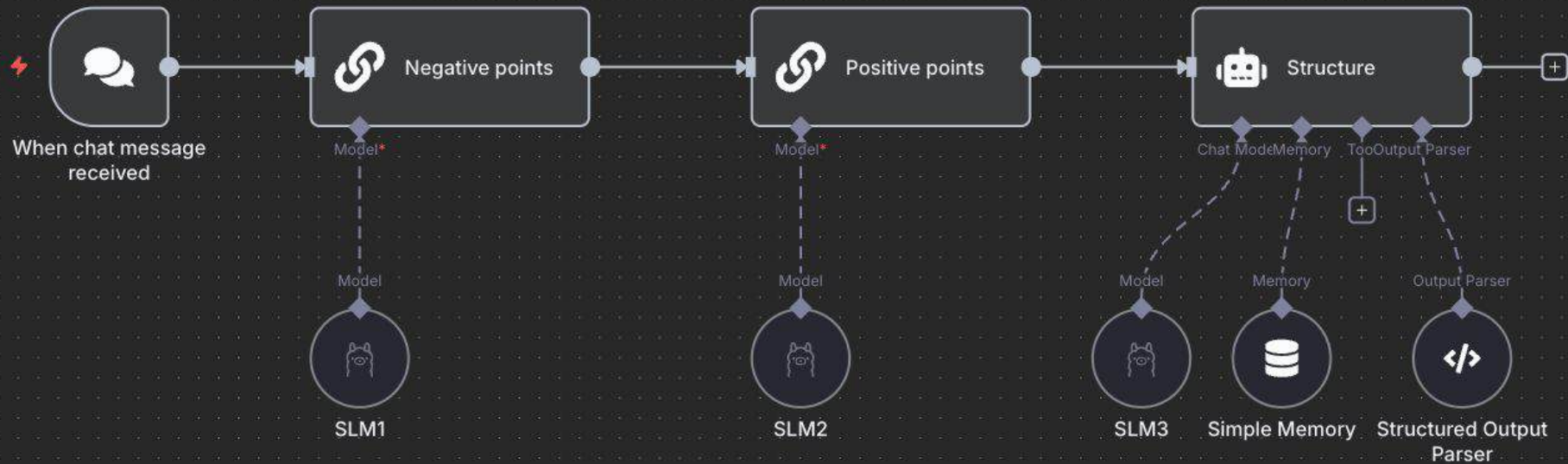


Structure

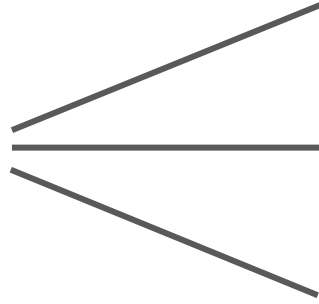


?

1. Better than GPT5
2. As fast, despite three prompts
3. Cost. Free after set up
4. High privacy
5. Transparency to validate
6. Using fraction of energy







Company
processes

Your own
processes



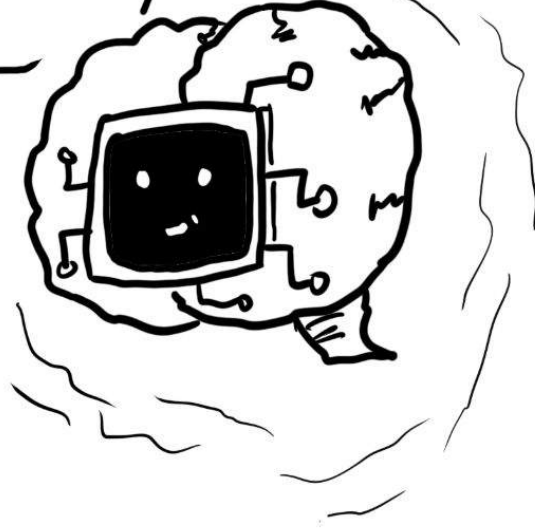
Your
brain

Make money





I'D LIKE YOU
TO MEET YOUR
NEW PARTNER.









Ethan Mollick • Following

Associate Professor at The Wharton School. Author of Co-Intelligence
1mo •

Reasoning AI models require training on human reasoning. One of the real gaps in pushing forward these models is going to be the old problem of how to figure out how to get experts to explain what they do. They often can't explain it themselves.

AI keeps bumping up against our limited knowledge of how expertise works.



Peripheral Vision

Expertise in Real World Contexts

Hubert L. Dreyfus and Stuart E. Dreyfus

Abstract

In this paper we describe a five-stage phenomenological model of skill acquisition, of which expertise is the highest stage. Contrary to the claims of knowledge engineers, we argue that expertise in general, and medical expertise in particular, cannot be captured in rule-based expert systems, since expertise is based on the making of immediate, unreflective situational responses; intuitive judgment is the hallmark of expertise. Deliberation is certainly used by experts, if time permits, but it is done for the purpose of improving intuition, not replacing it. The best way to avoid mistakes is to take responsibility for them when they occur, rather than try to prevent them by foolproof rules. In bureaucratic societies, however, there is the danger that expertise may be diminished through over-reliance on calculative rationality.


Keywords: skills, expertise, rules, knowledge engineering, rationality, intuition

Hubert L. Dreyfus
University of
California,
Berkeley, USA

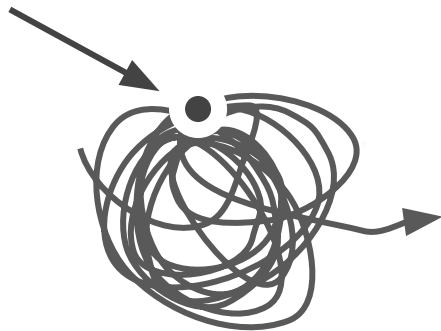
Stuart E. Dreyfus
University of
California,
Berkeley, USA

“One of the real gaps in pushing forward these models is going to be the old problem of how to figure out **how to get experts to explain what they do.** They often can't explain it themselves.”





What
do I want to
validate



Rule #1: Think small
Small models, small, simple tasks

Rule #2: Think in terms of
workflows

Rule #3: Think in terms of
validation

Rule #1: Think small
Small models, small, simple tasks

Rule #2: Think in terms of
workflows

Rule #3: Think in terms of
validation



Rule #4: Keep yourself in
control

My challenge to you:

1. Pick a repetitive task in your experiment workflow, and three specific examples

My challenge to you:

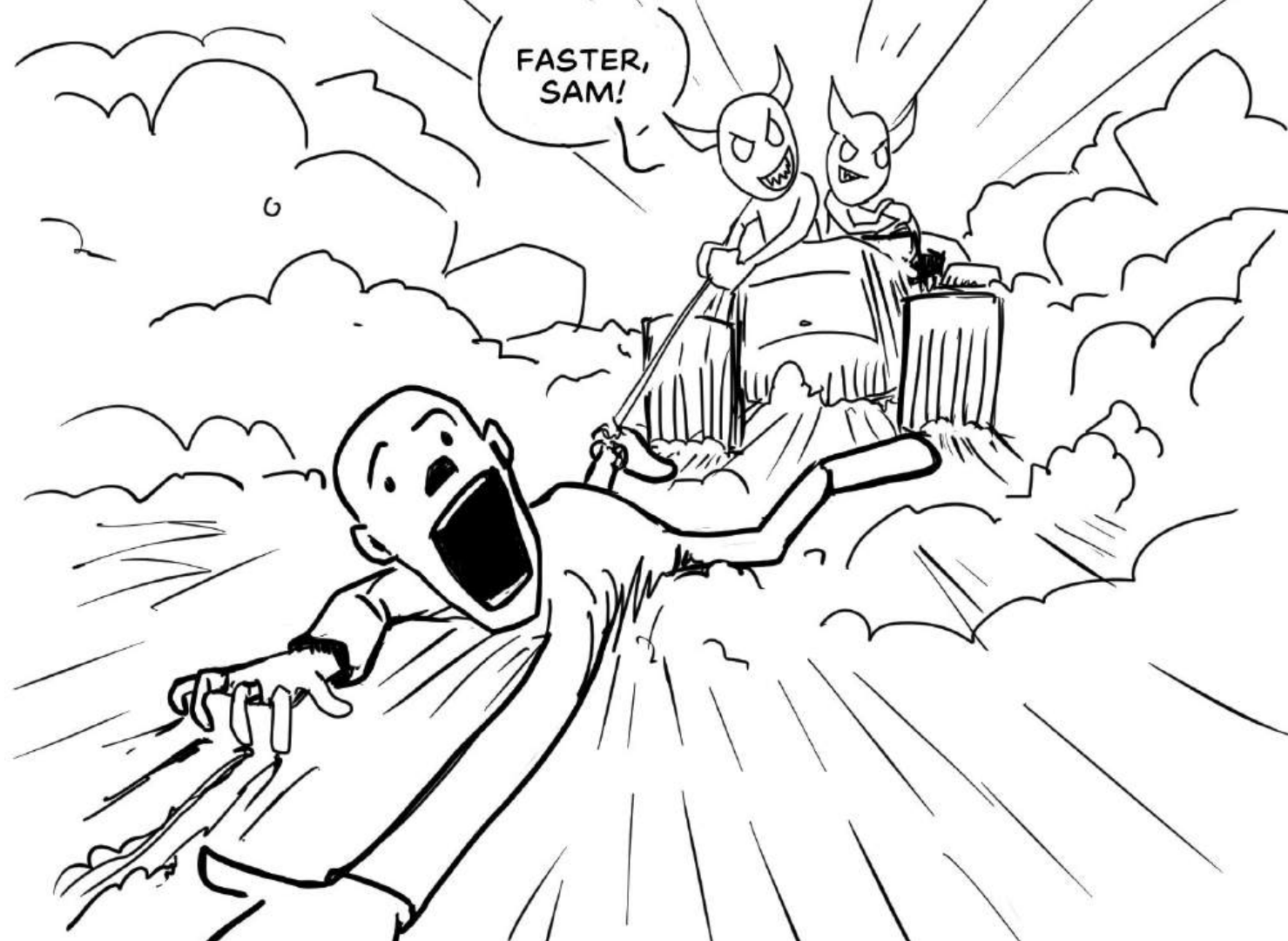
1. Pick a repetitive task in your experiment workflow, and three specific examples
2. Create an OpenRouter account

My challenge to you:

1. Pick a repetitive task in your experiment workflow, and three specific examples
2. Create an OpenRouter account
3. Co-work with AI, comparing your current fave model vs smaller ones

My challenge to you:

1. Pick a repetitive task in your experiment workflow, and three specific examples
2. Create an OpenRouter account
3. Co-work with AI, comparing your current fave model vs smaller ones
4. Build your own chatbot to replace larger ones

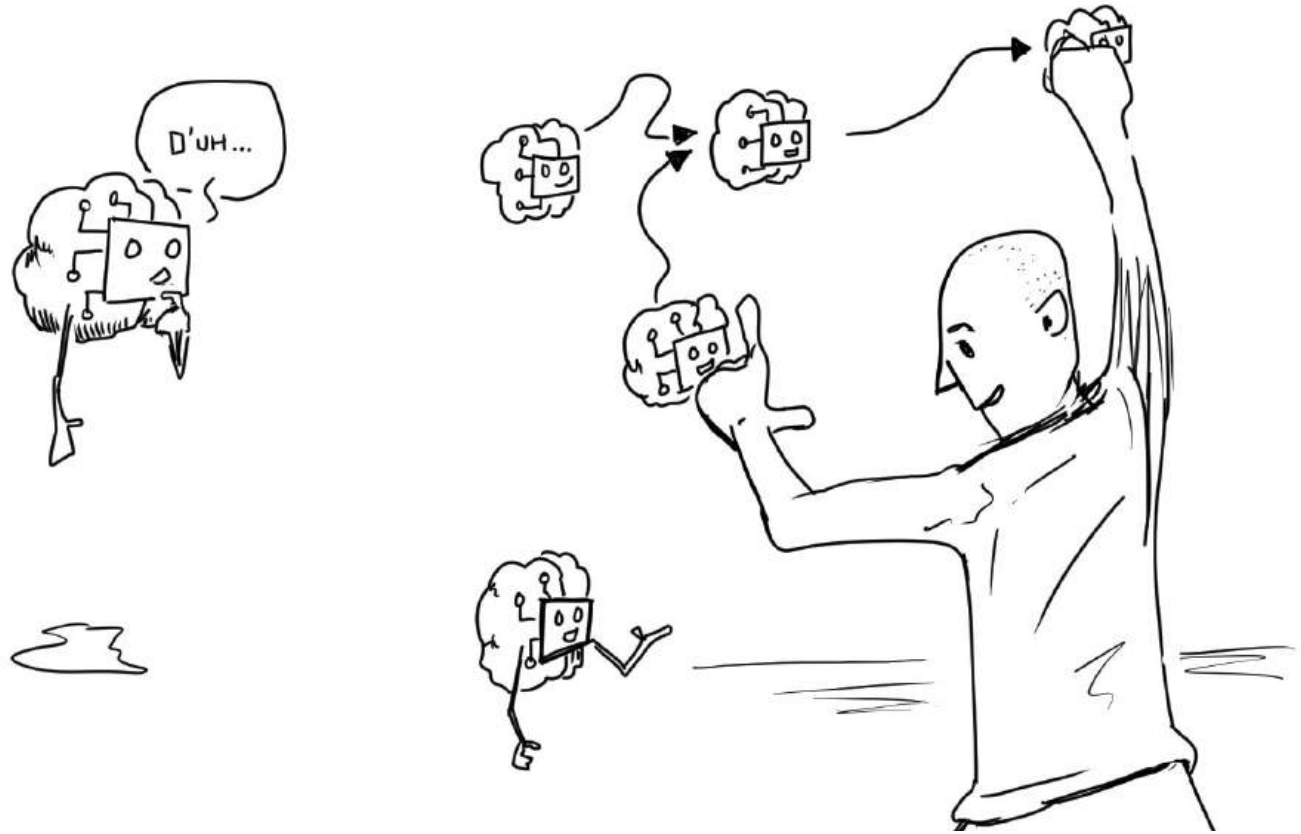


FASTER,
SAM!



THERE'S
ANOTHER
WAY.





Embrace your role as the **architect**.

We're in the **jungle**.

We're in the **jungle**.

The best way to understand what we can do in
this landscape?

We're in the **jungle**.

The best way to understand what we can do in
this landscape?

Experiment.

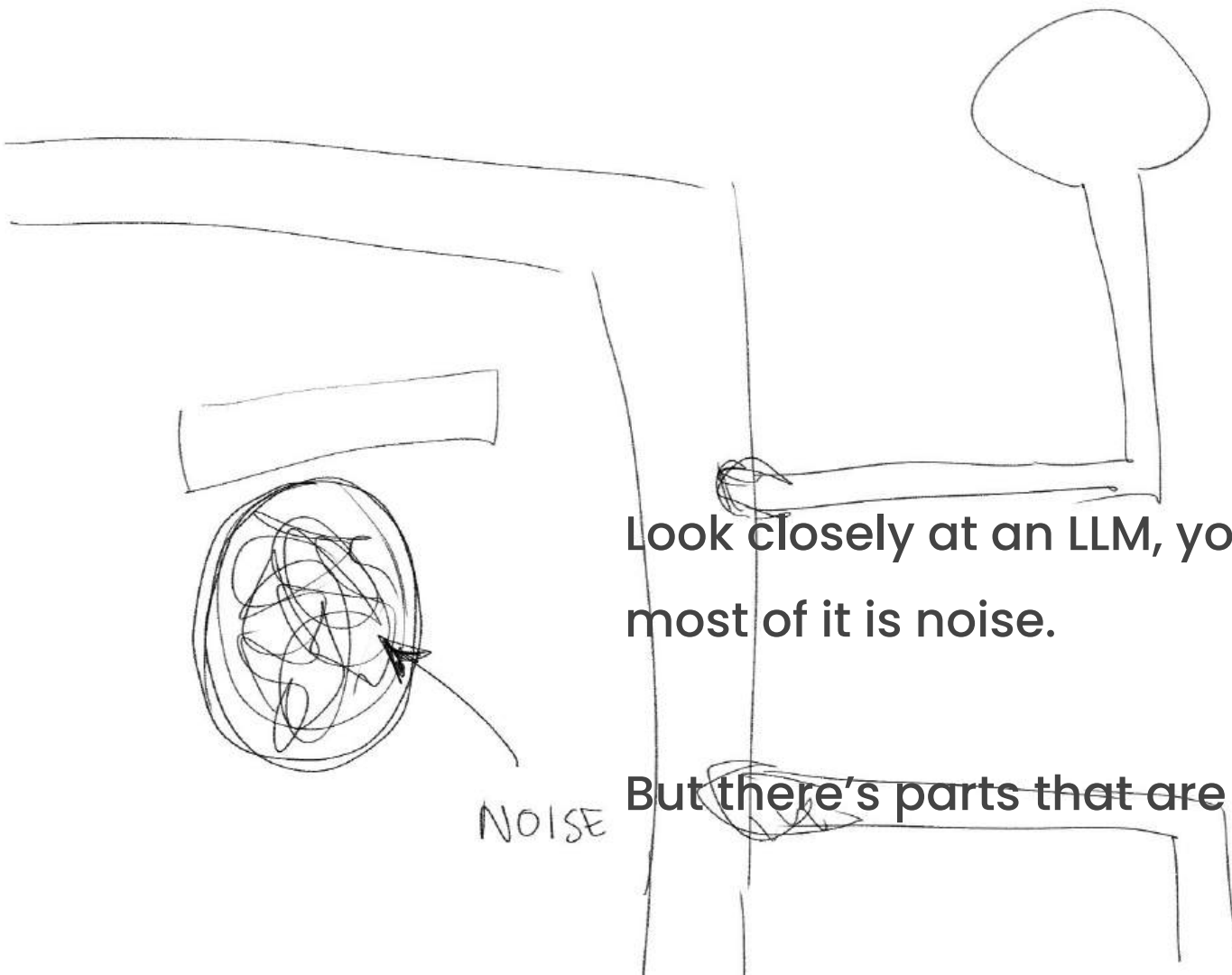


Predicted question: **Is AI going to take my job?**



Thank you

Connect: [linkedin.com/in/iqbalhussainali](https://www.linkedin.com/in/iqbalhussainali)



Look closely at an LLM, you'll notice that most of it is noise.

But there's parts that are useful.

Behold AI.
It will save
us!



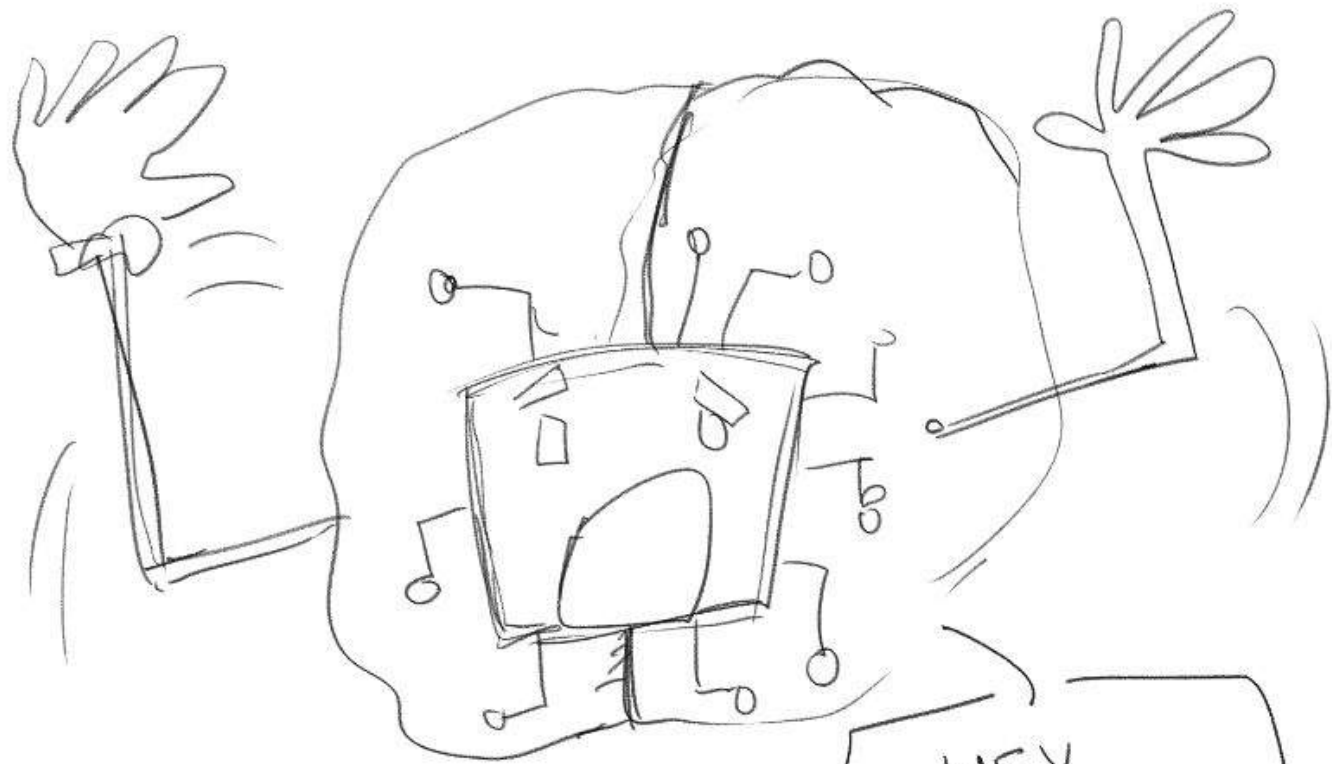


STAY
AWAY



LET'S GET
IT ON.

JUNGLE.
WHERE MOST
LIVE



HEY,
LOOK AT
ME!



AND NOW
I SHALL DO
MACHINE
LEARNING!



1



2

AND NOW
I SHALL DO
MACHINE
LEARNING!



MOMENTS
LATER...



3



MOMENTS
LATER...

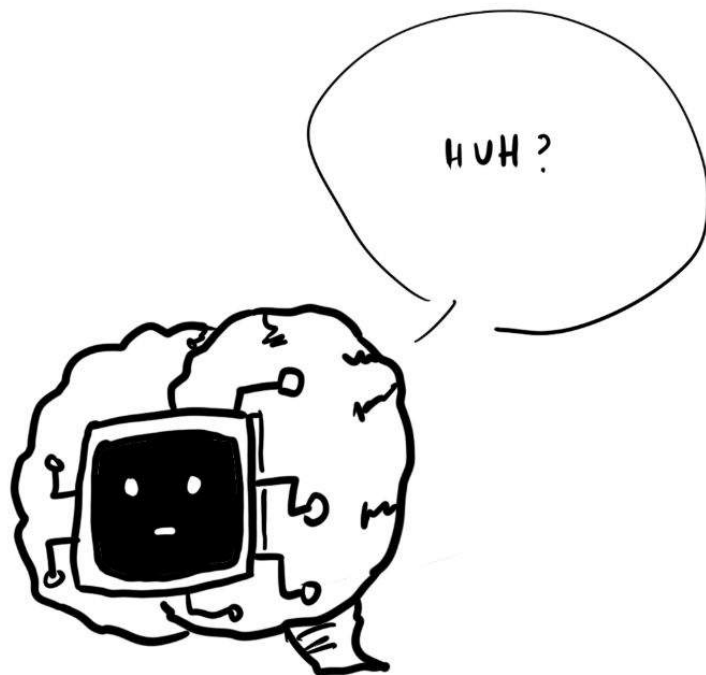




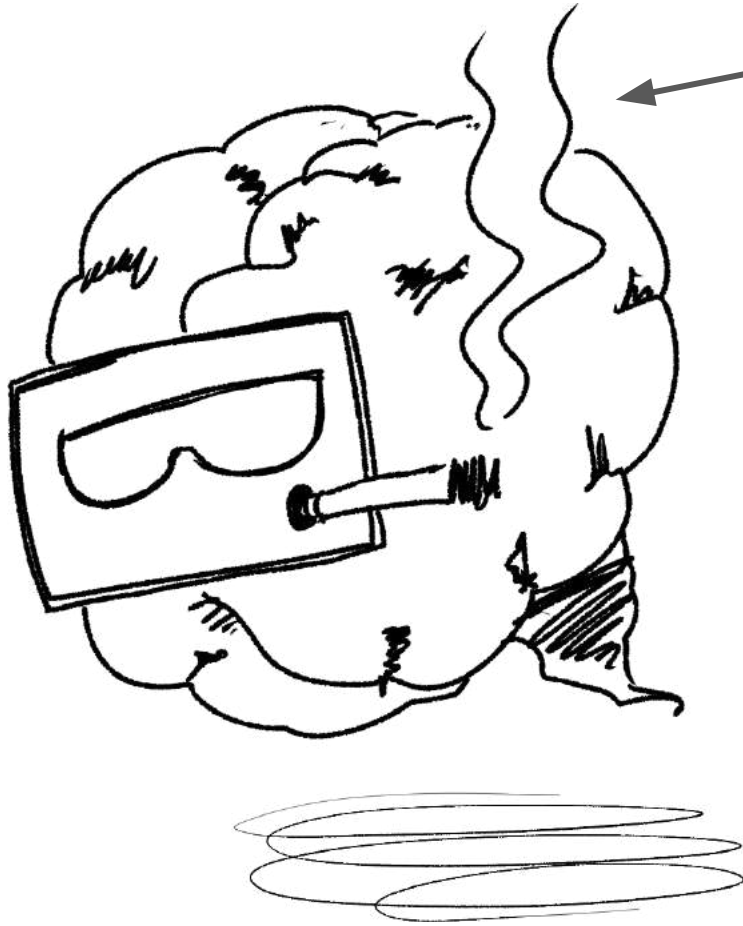
MOMENTS
LATER...

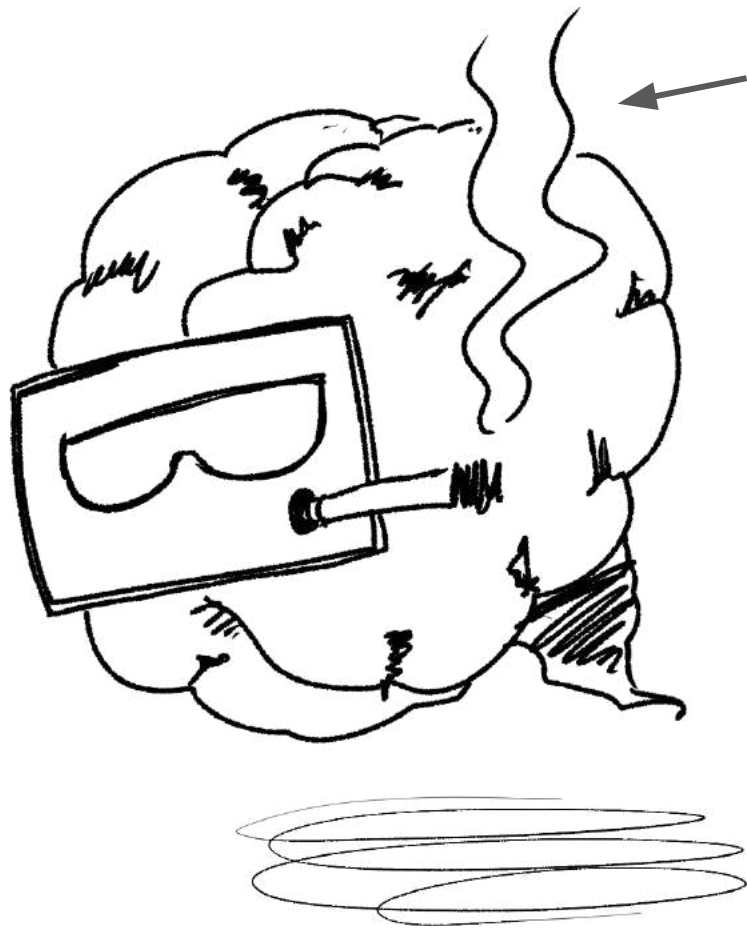
RESSADA

FAIL



Dumb, but thinks
it's clever



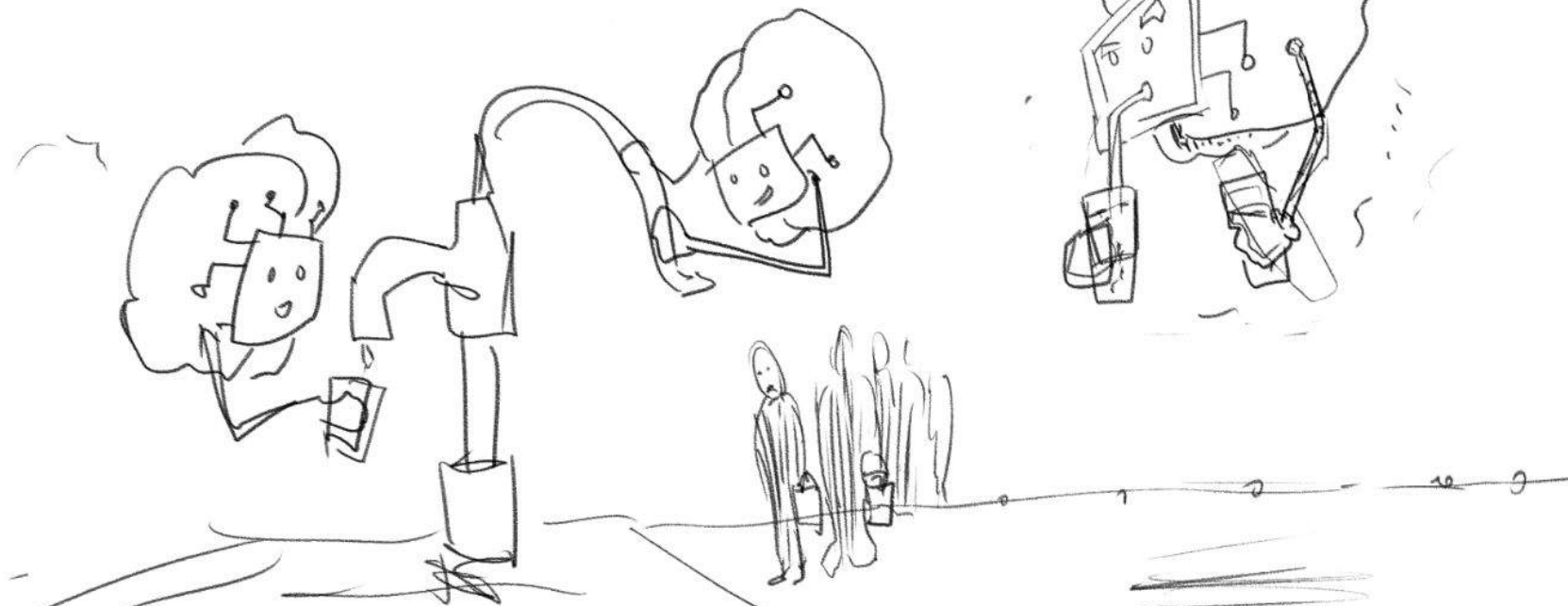


Dumb, but thinks
it's clever



Just dumb

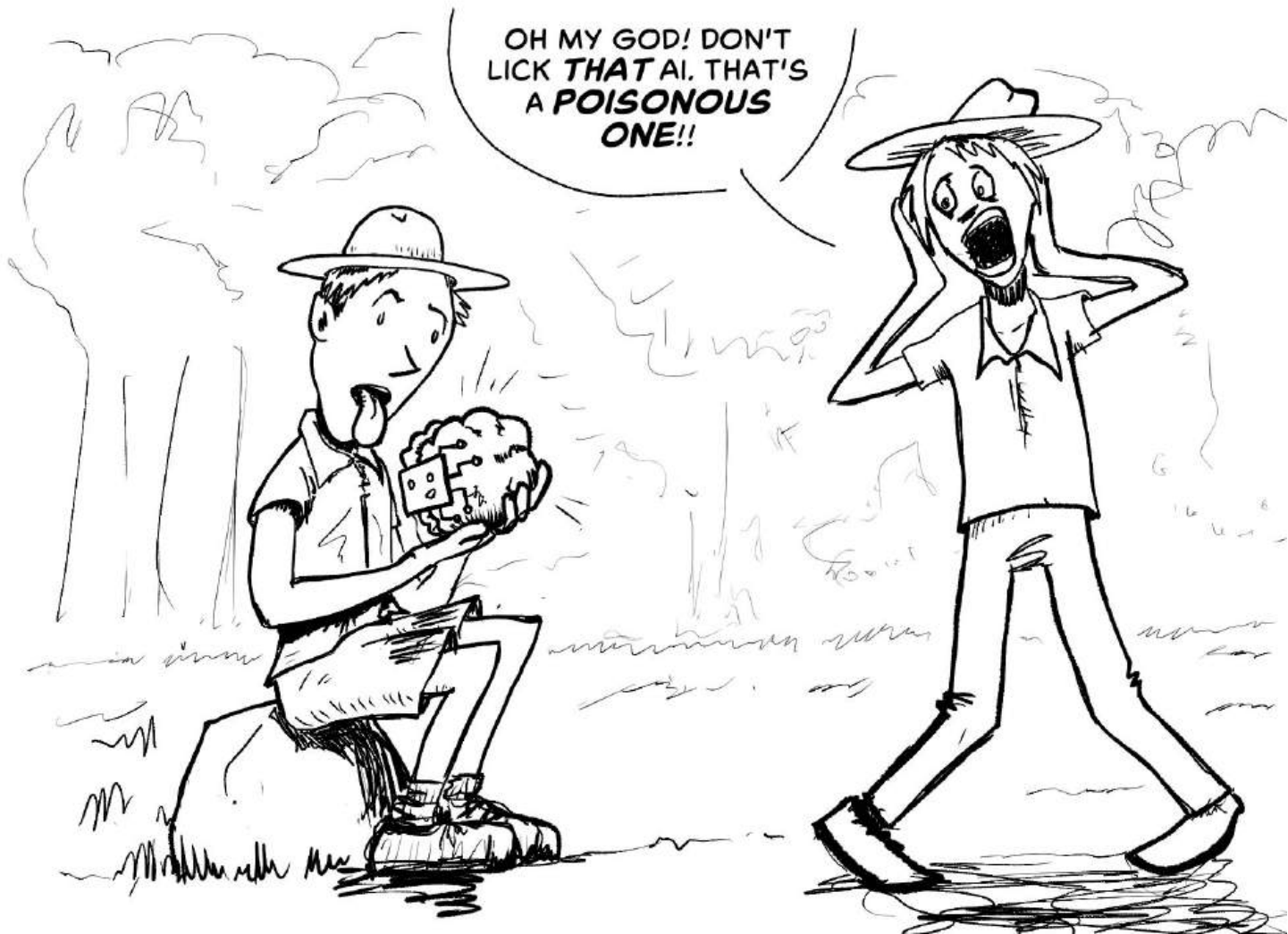
Get me another glass, willya?



Paper: [Addressing the Secret Water Footprint of AI Models \(Mar 2025\)](#)

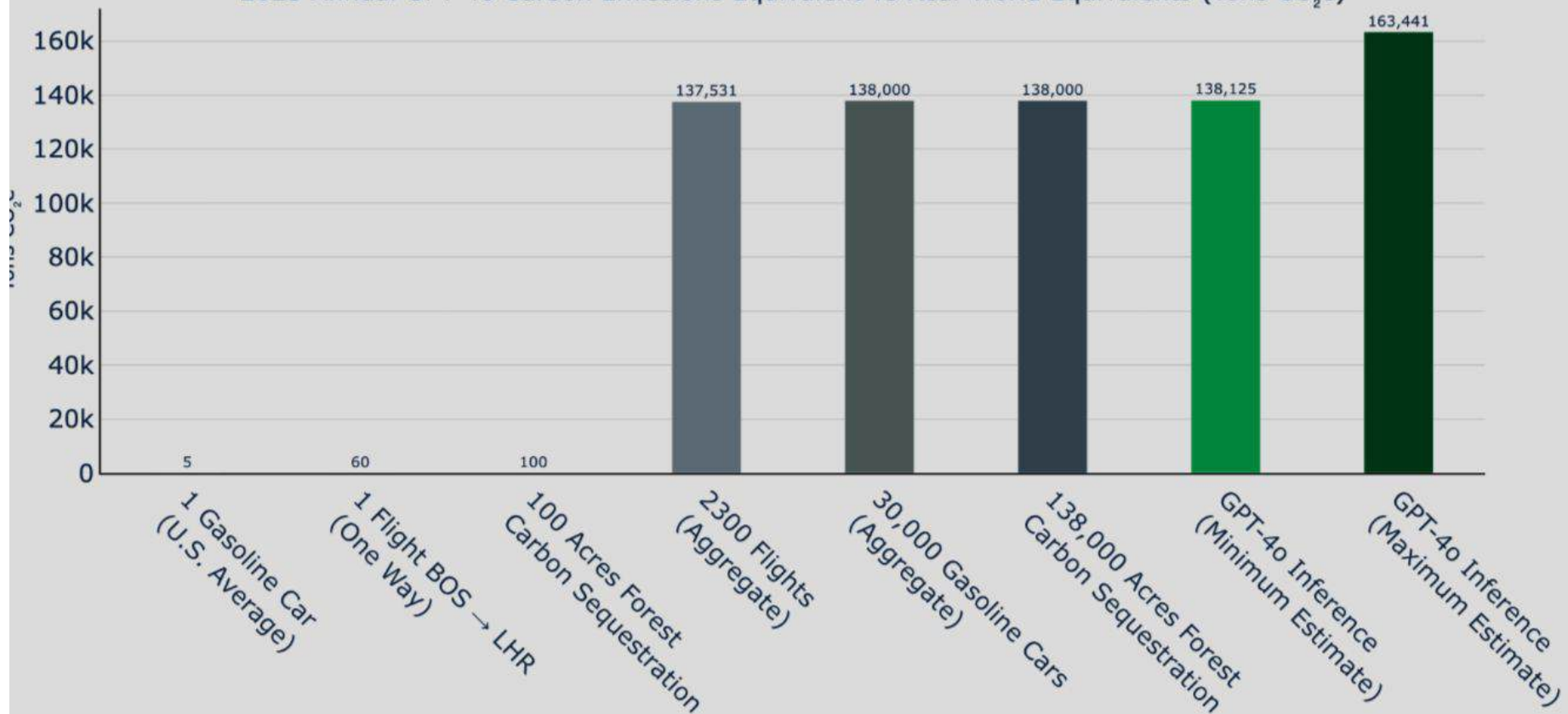
Article: [Inside Facebook's Secretive Push to Build Holland's Biggest Data Center](#)

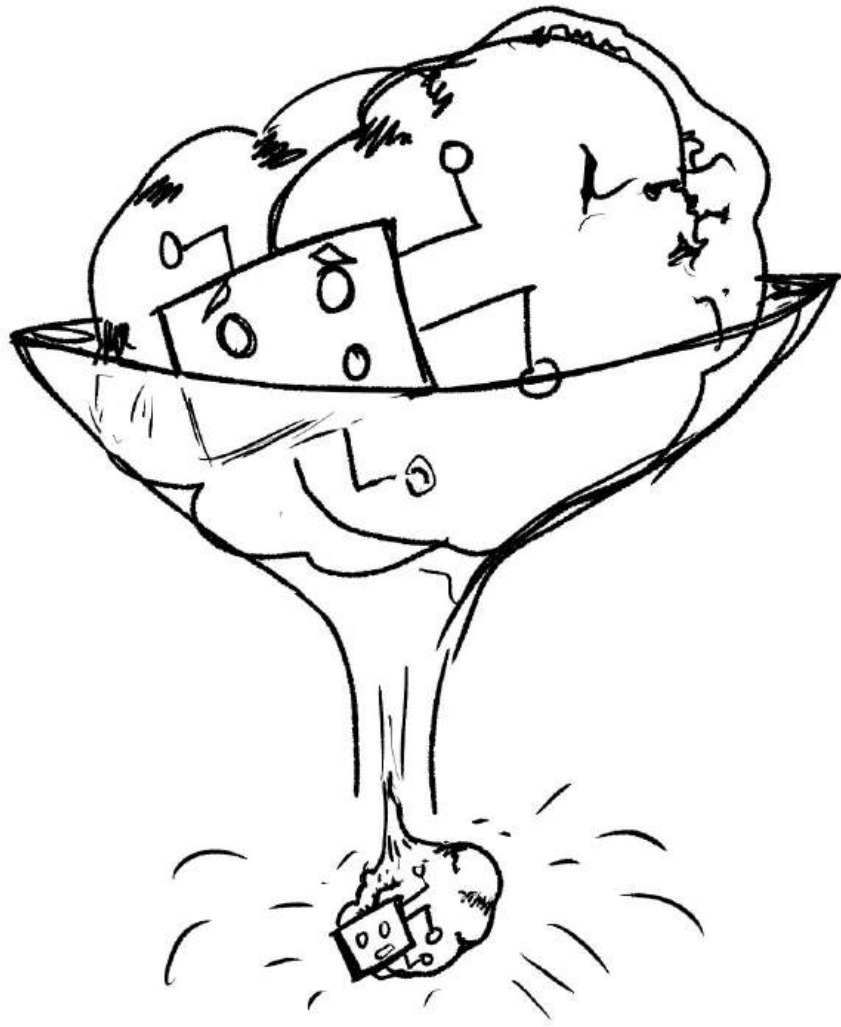
OH MY GOD! DON'T
LICK *THAT* AI. THAT'S
A **POISONOUS**
ONE!!



This talk is about how we should architect the
right system.

2025 Annual GPT-4o Carbon Emissions Equivalent vs Real-World Equivalents (Tons CO₂e)

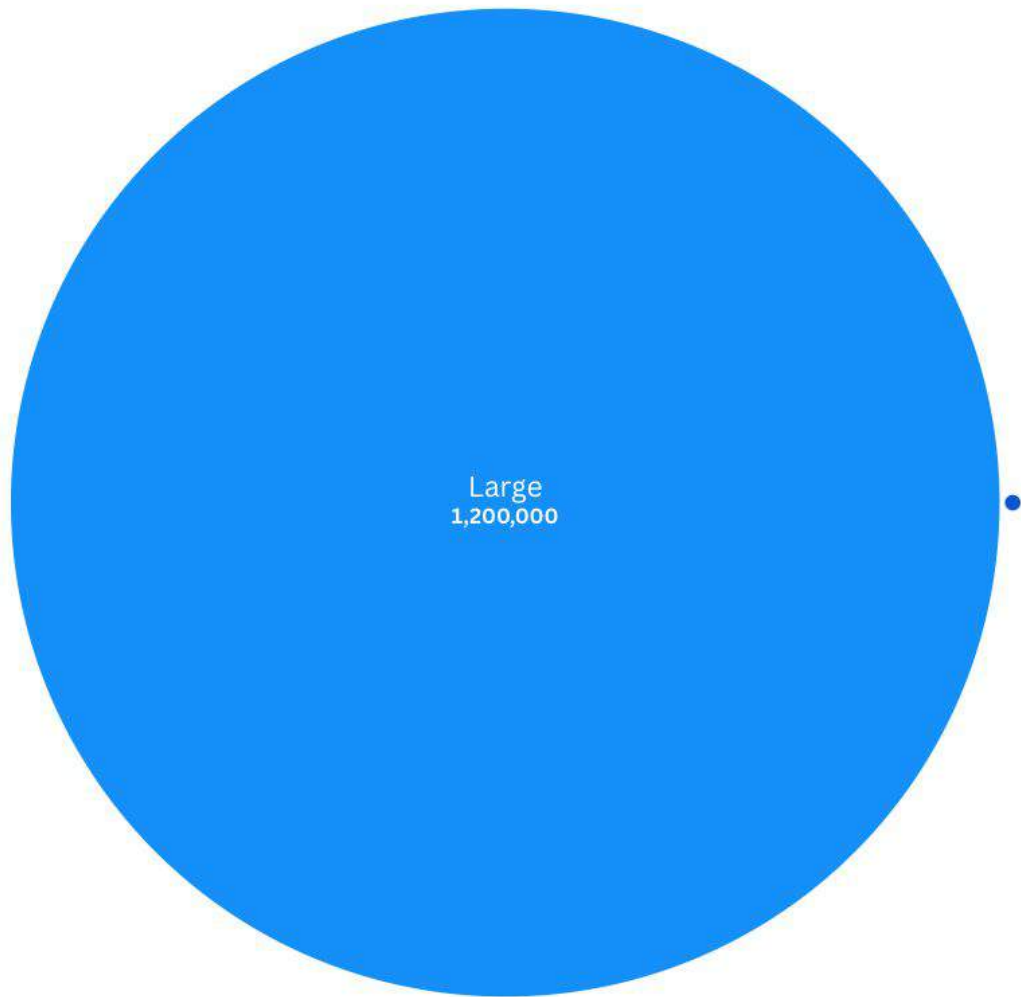


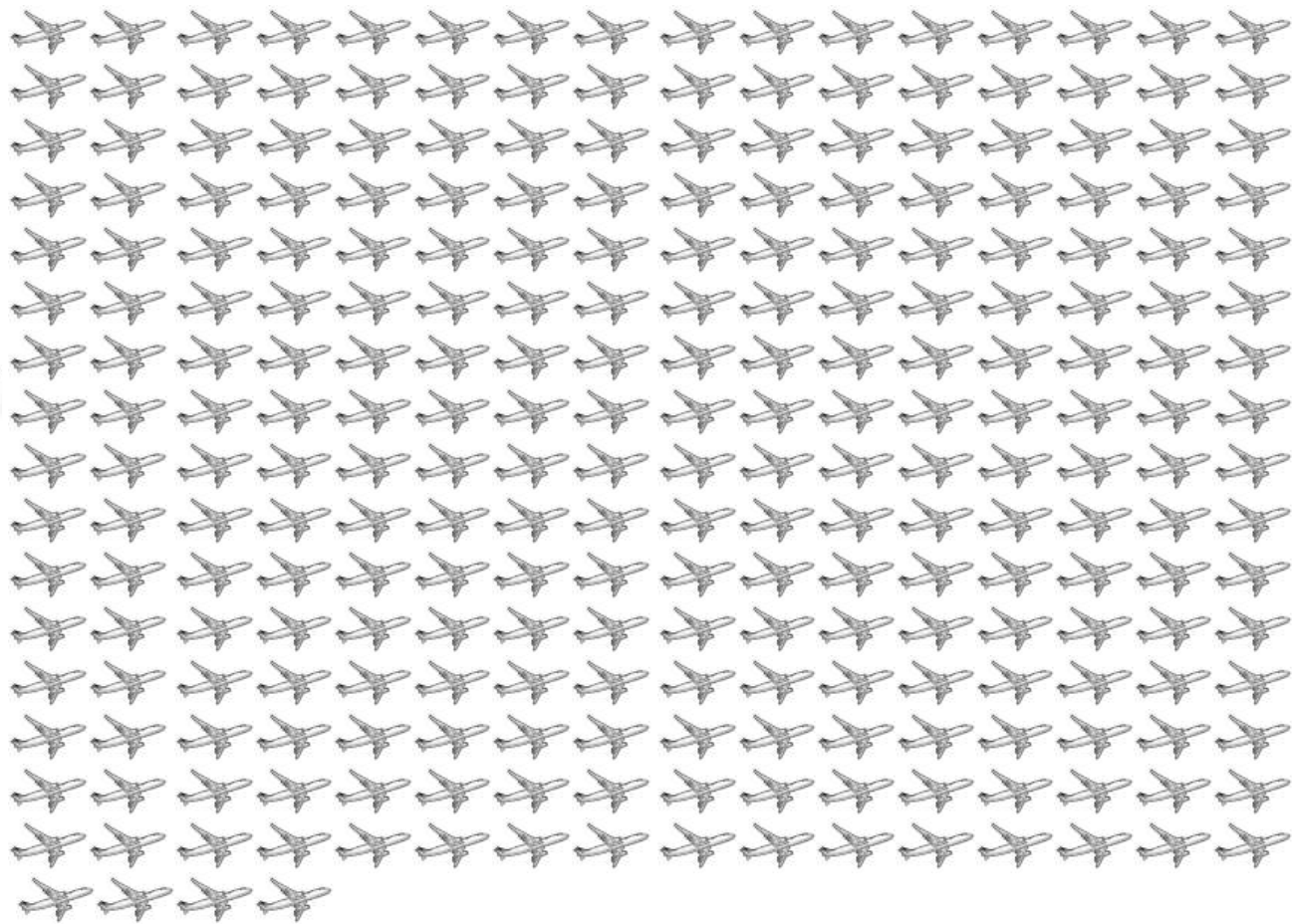
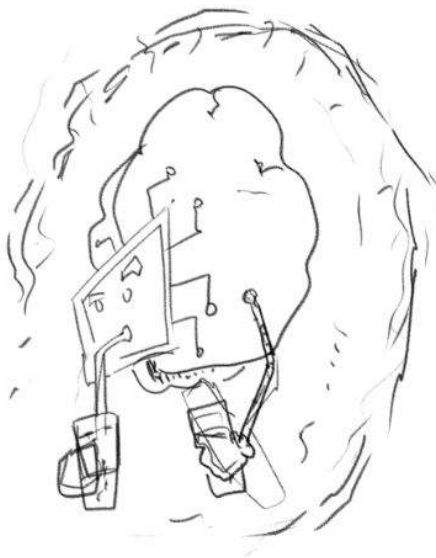


Paper: [Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes](#) (2023)

“[Utilises] **extracted rationales** as additional, richer information to **train small models.**”

“Distilling allows ... learn task specific smaller models that outperform LLMs using over”



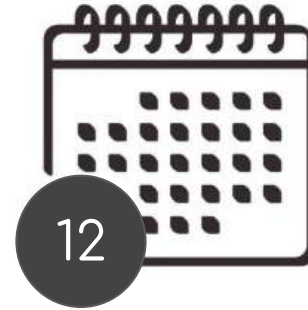


Carbon footprint: GPT one month = 260 transatlantic flights [\[source\]](#)



Energy usage of the
average US home.

x



How long?

GPT5

x 2.4
per year



I failed at least **twice**
with AI before 2022.

1



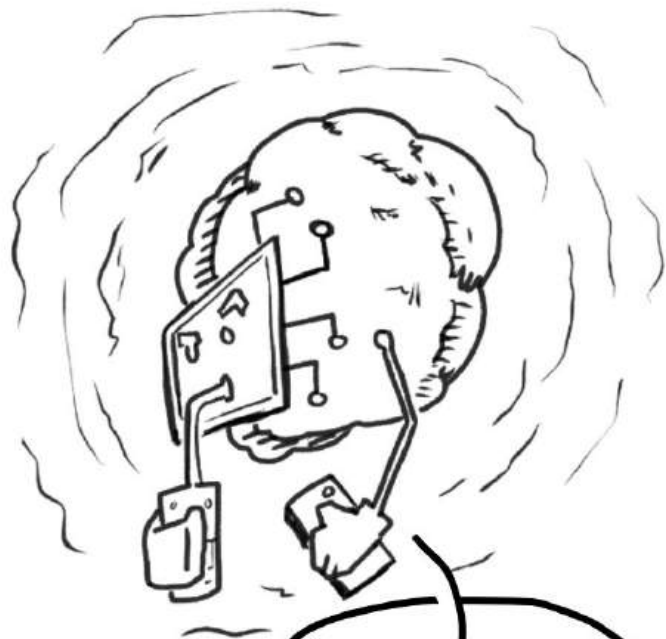
2

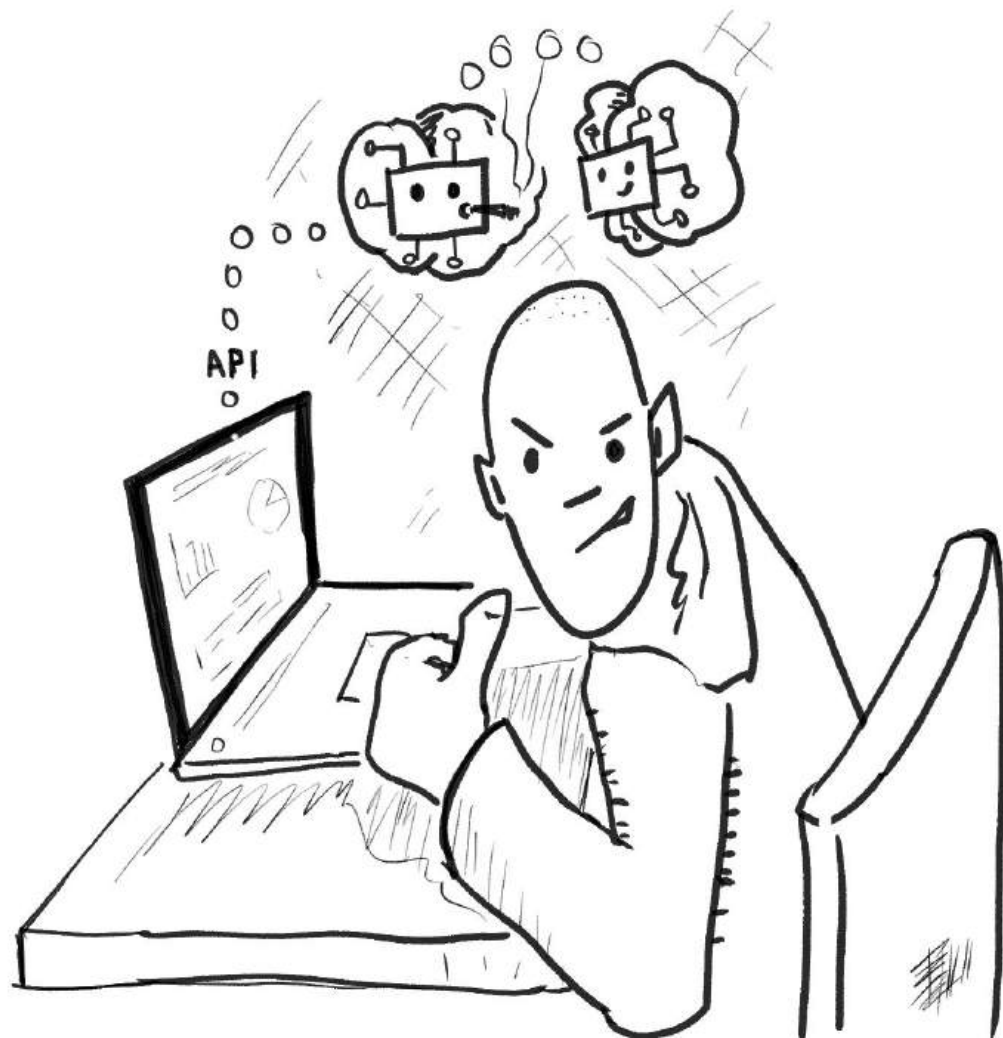


2



2022...



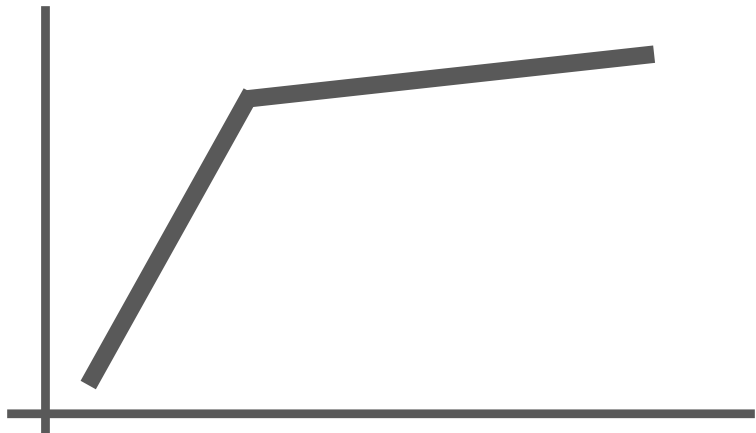




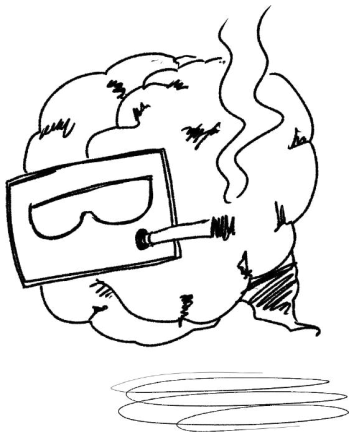
Give you a map to
navigate the jungle



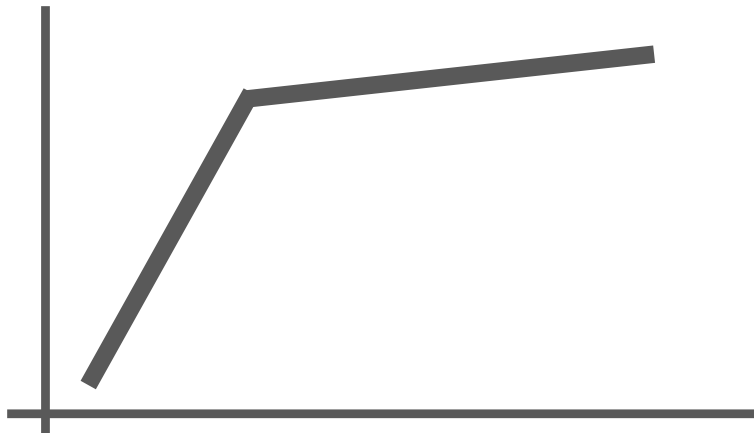
Equivalent to a **week**
powering the average home



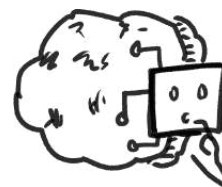
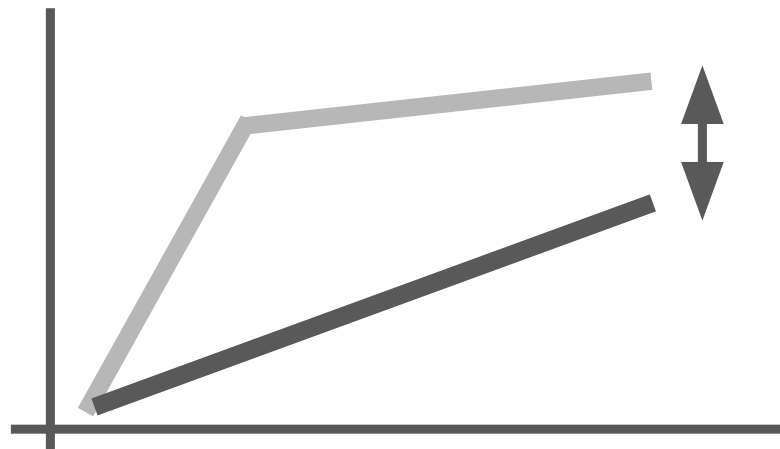
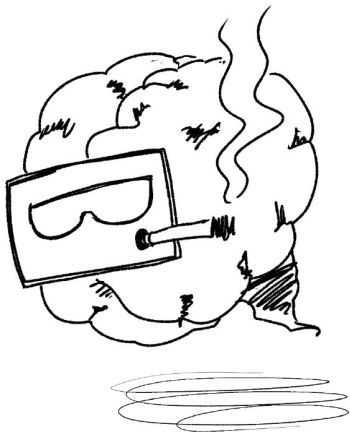
The AI Plateau is real



* For illustrative purposes



The AI Plateau is real

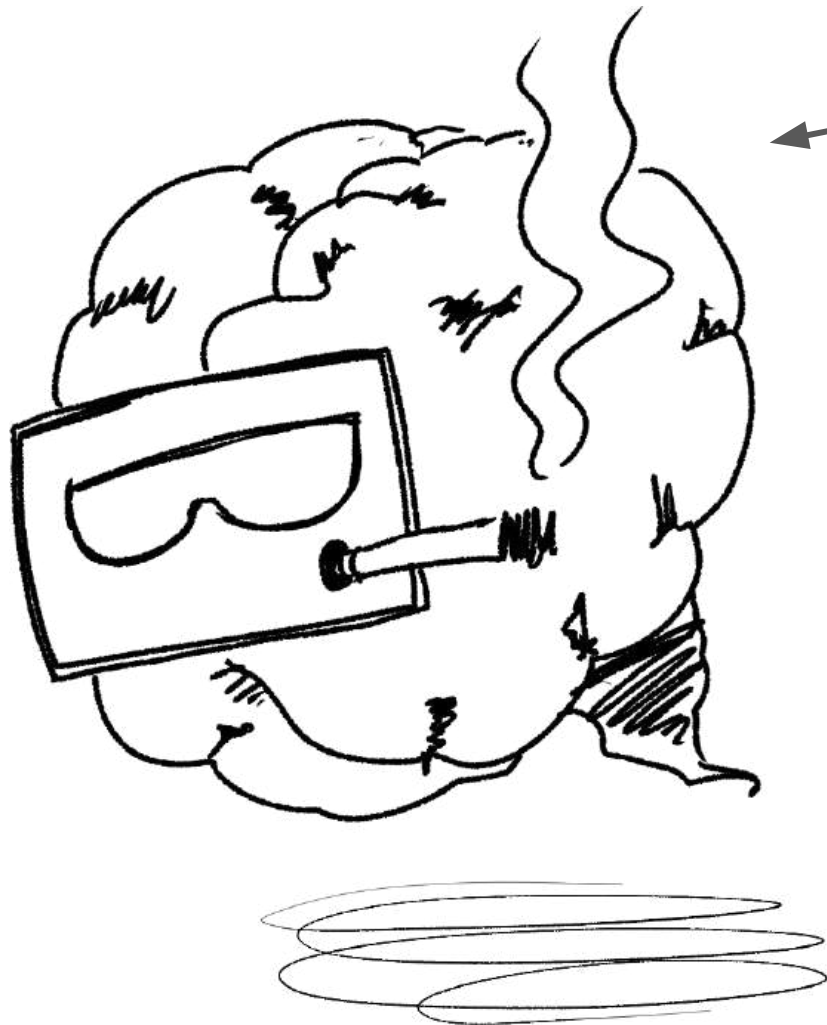




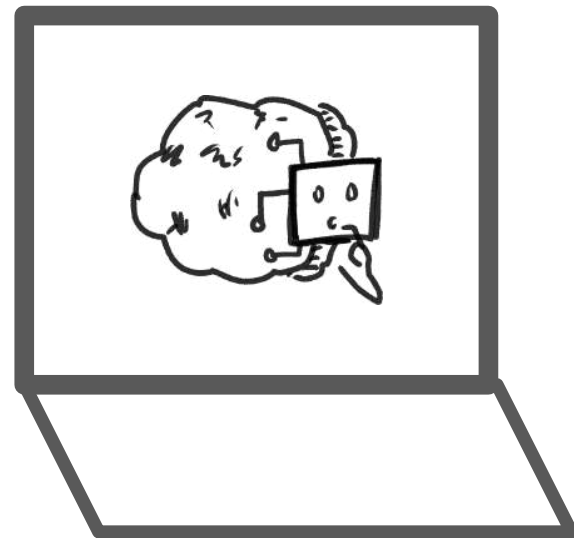
A diagram illustrating the size of GPT-5 on disk. A large dark gray circle on the left represents the model's size, labeled 'GPT-5 (~7 TB)'. On the right, a tiny black dot represents a much smaller size, with a gray arrow pointing to it from the left. The background is light gray.

GPT-5
(~7 TB)

Size on disk comparison



GPT5



Small model on
16GB M1 Macbook (2020)