19/02/2025

Building for AI

Simon Minton

TECHNICAL
CHOPS

<6 months of LLM model releases

o1 preview

o1 mini

Gemini Pro 1.5 002

Gemini Flash 1.5 002

Claude 3.5 Sonnet New

Gemini 2.0 Flash

O1

Llama 3.3 70B

o3

Qwen2.5-Max

R1 / R1-Zero

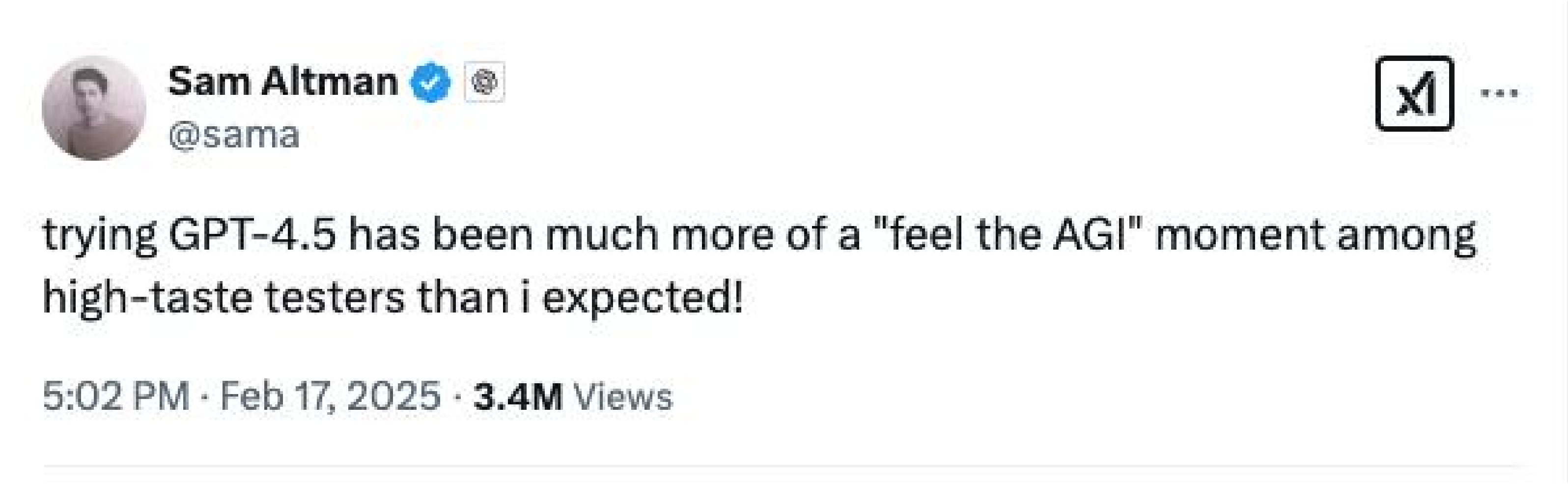
o3-mini

Grok 3

We’ve never seen anything develop this quickly before

What’s coming down the pike?

GPT 4.5 is in private testing

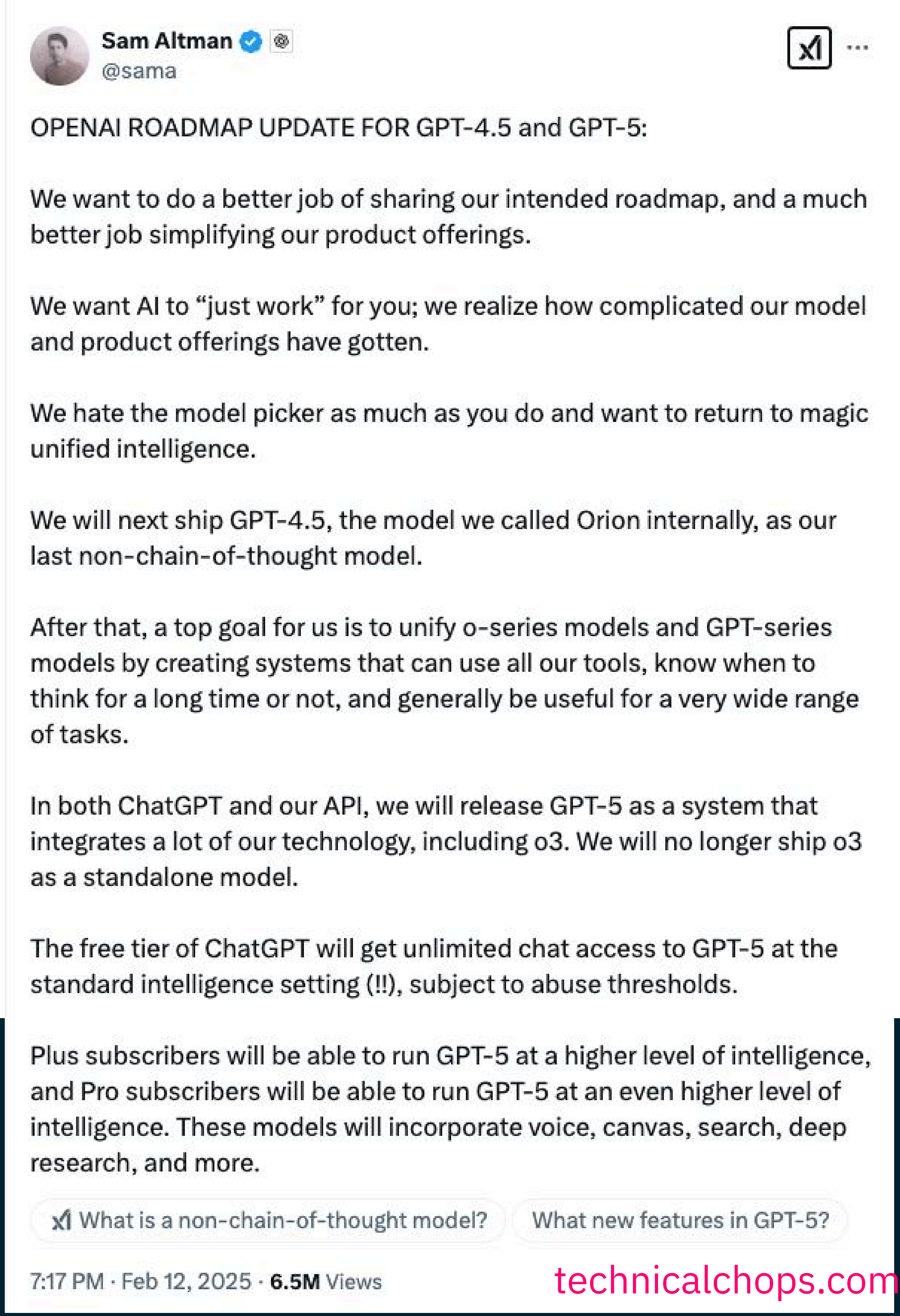


Sam Altman @sama

trying GPT-4.5 has been much more of a "feel the AGI" moment among high-taste testers than i expected!

5:02 PM · Feb 17, 2025 · **3.4M** Views

And then GPT-5 probably within a year.



Sam Altman @sama

OPENAI ROADMAP UPDATE FOR GPT-4.5 and GPT-5:

We want to do a better job of sharing our intended roadmap, and a much better job simplifying our product offerings.

We want AI to “just work” for you; we realize how complicated our model and product offerings have gotten.

We hate the model picker as much as you do and want to return to magic unified intelligence.

We will next ship GPT-4.5, the model we called Orion internally, as our last non-chain-of-thought model.

After that, a top goal for us is to unify o-series models and GPT-series models by creating systems that can use all our tools, know when to think for a long time or not, and generally be useful for a very wide range of tasks.

In both ChatGPT and our API, we will release GPT-5 as a system that integrates a lot of our technology, including o3. We will no longer ship o3 as a standalone model.

The free tier of ChatGPT will get unlimited chat access to GPT-5 at the standard intelligence setting (!!), subject to abuse thresholds.

Plus subscribers will be able to run GPT-5 at a higher level of intelligence, and Pro subscribers will be able to run GPT-5 at an even higher level of intelligence. These models will incorporate voice, canvas, search, deep research, and more.

What is a non-chain-of-thought model?
What new features in GPT-5?

7:17 PM · Feb 12, 2025 · **6.5M** Views

What does that mean?

- **Capabilities**

  Models that exceed the capabilities and speed of previous models are arriving on an almost weekly basis.

- **Access**

  We get access to them **immediately** and using them is often as easy as changing a single line of code.

- **Cost**

  The cost of inference (running the model) drops 10x every time.

What will tomorrow’s models be capable of?

We don’t know.

It is not inconceivable that in 2 years you will be able to just tell the model to run your entire process within itself.

All we know is that today’s models are the **worst they will ever be.**

Building when the models keep changing

Build on the basis that we have no idea how capable a model is going to be in 6 months–1 year.

Focus on **infrastructure, not models.**

Build **disposable** tools - assume that we will replace parts of the stack in a year.

Prompting

Is writing prompts a waste of time?

Sort of.

Different models require different prompting methodologies.

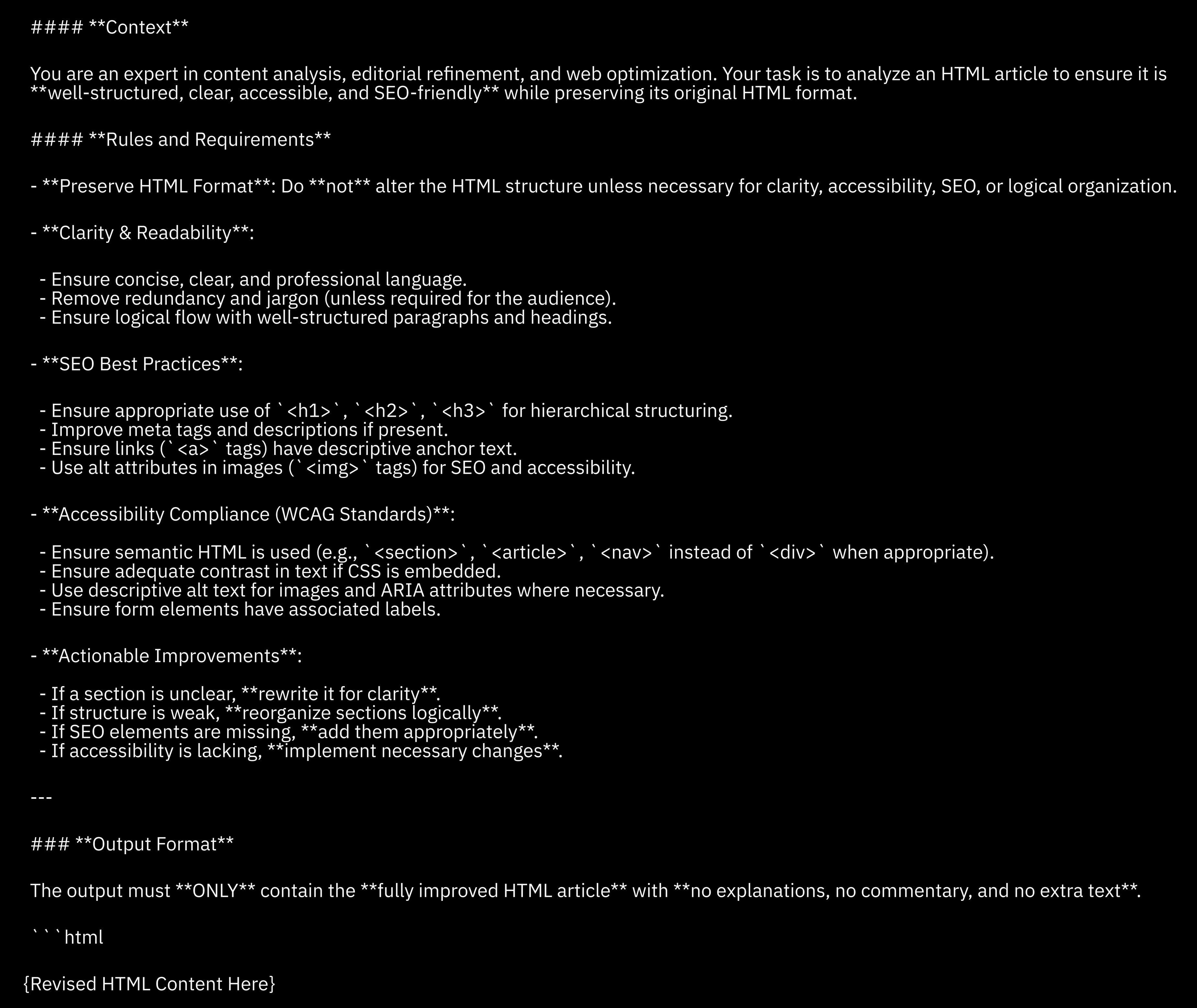
Using a human to refine prompts per model is potentially a waste of time.

Focus on **objectives.**

Your prompts aren’t good enough*

- Prompting is hard and requires **lots** of work
- Different prompts work for different models
- AIs are really good at prompting themselves
- Ask AI for prompts based on **objectives**

*and it’s not **your** fault



~~~
#### **Context**

You are an expert in content analysis, editorial refinement, and web optimization. Your task is to analyze an HTML article to ensure it is
**well-structured, clear, accessible, and SEO-friendly** while preserving its original HTML format.

#### **Rules and Requirements**

- **Preserve HTML Format**: Do **not** alter the HTML structure unless necessary for clarity, accessibility, SEO, or logical organization.

- **Clarity & Readability**:

  - Ensure concise, clear, and professional language.
  - Remove redundancy and jargon (unless required for the audience).
  - Ensure logical flow with well-structured paragraphs and headings.

- **SEO Best Practices**:

  - Ensure appropriate use of `<h1>`, `<h2>`, `<h3>` for hierarchical structuring.
  - Improve meta tags and descriptions if present.
  - Ensure links (`<a>` tags) have descriptive anchor text.
  - Use alt attributes in images (`<img>` tags) for SEO and accessibility.

- **Accessibility Compliance (WCAG Standards)**:

  - Ensure semantic HTML is used (e.g., `<section>`, `<article>`, `<nav>` instead of `<div>` when appropriate).
  - Ensure adequate contrast in text if CSS is embedded.
  - Use descriptive alt text for images and ARIA attributes where necessary.
  - Ensure form elements have associated labels.

- **Actionable Improvements**:

  - If a section is unclear, **rewrite it for clarity**.
  - If structure is weak, **reorganize sections logically**.
  - If SEO elements are missing, **add them appropriately**.
  - If accessibility is lacking, **implement necessary changes**.

---

### **Output Format**

The output must **ONLY** contain the **fully improved HTML article** with **no explanations, no commentary, and no extra text**.

```html

{Revised HTML Content Here}
~~~

A good prompt, for a simple task (produced by GPT-4o)

Prompt Pipelines

1 Your Objectives

Focus on your static objectives.
Provide your objectives to a HIGH reasoning model.
Have it provide you with **it’s understanding** of what you are asking for. Confirm it covers all aspects.

2 Prompt Generation

Then ask the model to provide 10/20/100 prompts for an LLM that are differently phrased to achieve the objectives that have been agreed.

3 Prompt Testing

Automate the testing of those prompts **per model** against your data to see which performs well.
Choose the most successful prompt for each objective.

4 Put into Production

Use the most successful prompts in production

Static Knowledge

Institutional Knowledge

- If you have internal or institutional knowledge, look at how you get that information to the model.
- How will the model use it? Is it additional context or is it a query to a database?
- In context? RAG? Function Calling? or some other Agentic flow?

Actually using the AI

Input / Output

- How do we get your queries or actionable data into the model?
- What infrastructure is required to be able to provide all information to *any* model?
- What is the standardised data format for the output?

Takeaways

01

Models

Expect Models to constantly change

02

Prompts

Expect to need different prompts for different models and have a pipeline to create them efficiently

03

Static Knowledge

Ensure you have a clear method for providing institutional knowledge to a model

04

Input/Output

Ensure you have a clear way to interact with the model, whatever it's capabilities