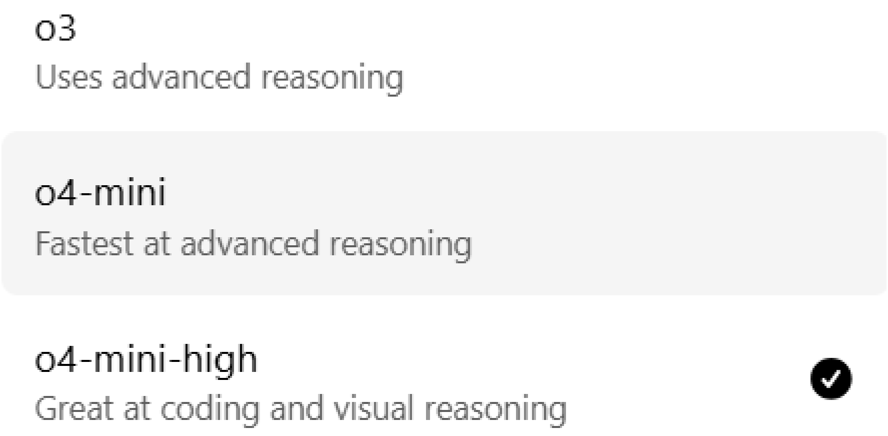
ChatGPT o3 and o4-mini for FP&A

The Ultimate Guide for Finance

By Christian Martinez


o3
Uses advanced reasoning
o4-mini
Fastest at advanced reasoning
o4-mini-high ✓
Great at coding and visual reasoning

OpenAI released ChatGPT o3 and o4-mini.

These are the smartest and most capable OpenAI models to date and have full tool access.

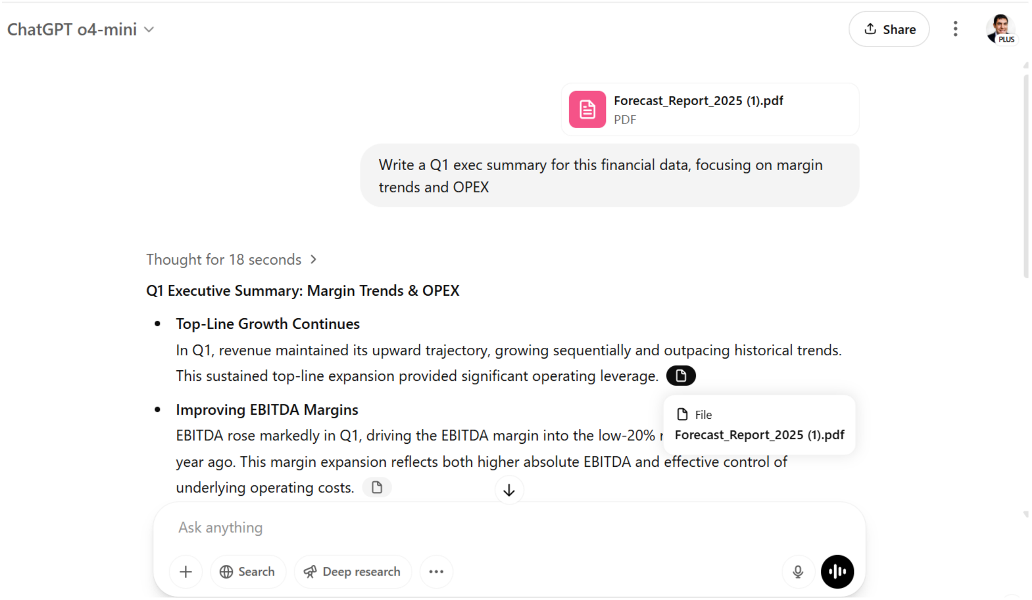
It means they can extract data from Excel files or even images, run Python to analyse the data and use "advanced reasoning".

Read this guide to know how you can leverage them for FP&A and the CFO Office!

Introduction

With the launch of **OpenAI o3 and o4-mini**, FP&A teams and CFO offices now have access to the most capable reasoning models ever released.

These models are trained to not only *understand complex problems*—they can now **think through them, solve them, and act** on your behalf using advanced tools like Python, file analysis, web browsing, and visual understanding.



Write a Q1 exec summary for this financial data, focusing on margin trends and OPEX

Q1 Executive Summary: Margin Trends & OPEX

- Top-Line Growth Continues
  In Q1, revenue maintained its upward trajectory, growing sequentially and outpacing historical trends. This sustained top-line expansion provided significant operating leverage.
- Improving EBITDA Margins
  EBITDA rose markedly in Q1, driving the EBITDA margin into the low-20% r year ago. This margin expansion reflects both higher absolute EBITDA and effective control of underlying operating costs.

For the first time, your FP&A team can upload spreadsheets, PDFs, dashboards, forecasts, and visuals—then have ChatGPT analyze them holistically by using their most advanced models yet, uncover insights, write commentary, model scenarios, or build automation pipelines.

These models don't just interpret numbers—they reason across them. They challenge assumptions.

They summarize.

They simulate scenarios.

Whether you're closing the books, building board decks, doing headcount modeling, or forecasting cash flow, these new models are built for the kinds of messy, multi-modal, cross-functional problems FP&A lives in every day.

Here I'll show you exactly how you can use them for FP&A and Finance:

But before we go to the use cases, I want to give you my quick start guide in case you want to start leveraging Generative AI for FP&A as soon as tomorrow!

Quick-start agenda for tomorrow morning

1. **Pick one repetitive report** (e.g., cost-center variance analysis or Scenario Modelling of Forecast).
2. Paste last three months of CSVs into ChatGPT, apply the prompt templates I'll give you in the next section.
3. Ask ChatGPT o3 for the python code it used for the analysis and use it in a local environment (can be Google Colab, Microsoft Visual Studio or even Python in Excel)
4. Create a scheduled automation to rerun the analysis after next close.
5. Use the freed-up hours to spend more time with family or do more value added work (your choice!)
6.

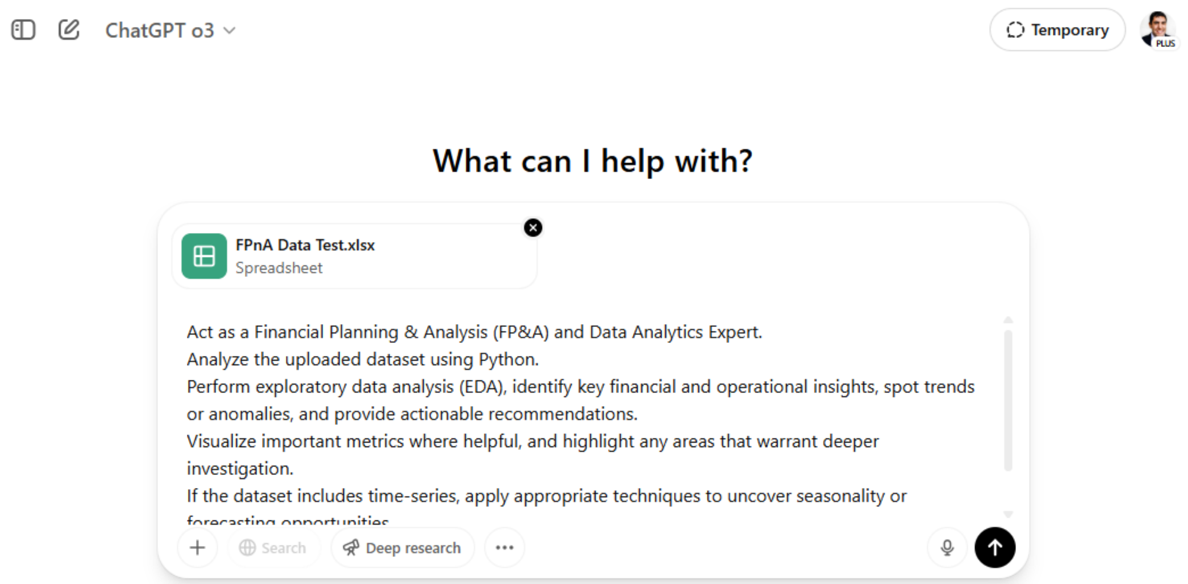
Bottom line: ChatGPT o3 is not just another chatbot; it's a multipurpose **financial co-pilot**. Give it structured data, explicit instructions, and guardrails, and it will turn the gruntwork of FP&A and fractional CFOing into a rapidly iterative, insight-driven process—so you can focus on the decisions that move the business.

Now, let me show you the use cases:

1) Anomaly Detection & Driver Attribution at Scale

It will flag unexpected deviations in financial or operational metrics (e.g., revenue, gross margin, headcount, burn rate), and trace them back to root causes across departments or inputs.

This prompt helps FP&A teams automatically detect unusual shifts in financial or operational data—and identify what caused them.



Act as a Financial Planning & Analysis (FP&A) and Data Analytics Expert.
Analyze the uploaded dataset using Python.
Perform exploratory data analysis (EDA), identify key financial and operational insights, spot trends or anomalies, and provide actionable recommendations.
Visualize important metrics where helpful, and highlight any areas that warrant deeper investigation.
If the dataset includes time-series, apply appropriate techniques to uncover seasonality or forecasting opportunities.

Prompt: *Act as a Financial Planning & Analysis (FP&A) and Data Analytics Expert. Analyze the uploaded dataset using Python. Perform exploratory data analysis (EDA), identify key financial and operational insights, spot trends or anomalies, and provide actionable recommendations. Visualize important metrics where helpful, and highlight any areas that warrant deeper investigation. If the dataset includes time-series, apply appropriate techniques to uncover seasonality or forecasting opportunities.*

Why it matters:
In any reporting cycle, unexpected spikes or drops in metrics (like revenue, OPEX, gross margin, headcount, CAC, etc.) raise questions from stakeholders. Manually tracking down the "why" is often time-consuming and cross-functional.

How ChatGPT o3 helps:
This use case turns o3 into your anomaly-hunting co-pilot. It doesn't just flag what looks "off"—it explains **why it happened**, using:

- Variance detection (e.g., rolling averages, standard deviation, % delta)
- Dimensional drilldowns (e.g., by region, product, department)

- Root-cause linking (e.g., higher shipping cost tied to volume shift in one SKU)

It can analyze multiple sheets or tabs and produce a narrative that connects the dots between your numbers—plus charts to back it up.

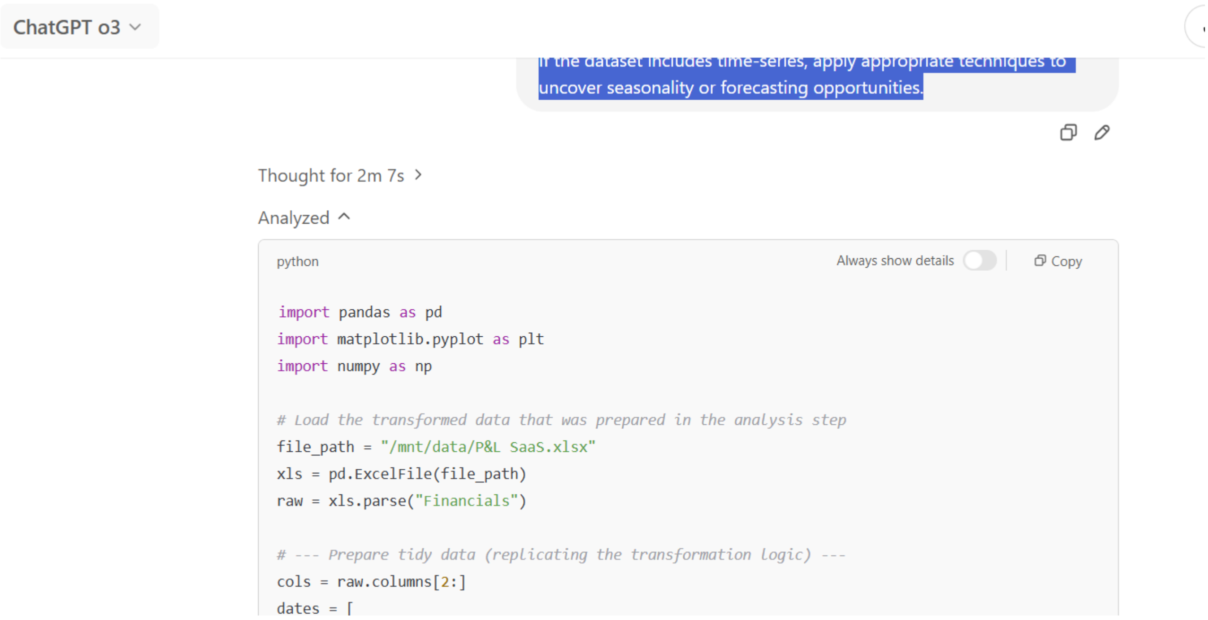
Example questions it can answer:

- “Why did OPEX spike last month?”
- “What’s driving the unexpected drop in bookings?”
- “Is this change seasonal or a true outlier?”

Best for:

- Monthly variance analysis
- Executive Q&A prep
- Automated review of budget vs. actuals
- Early detection of risks or missed forecasts

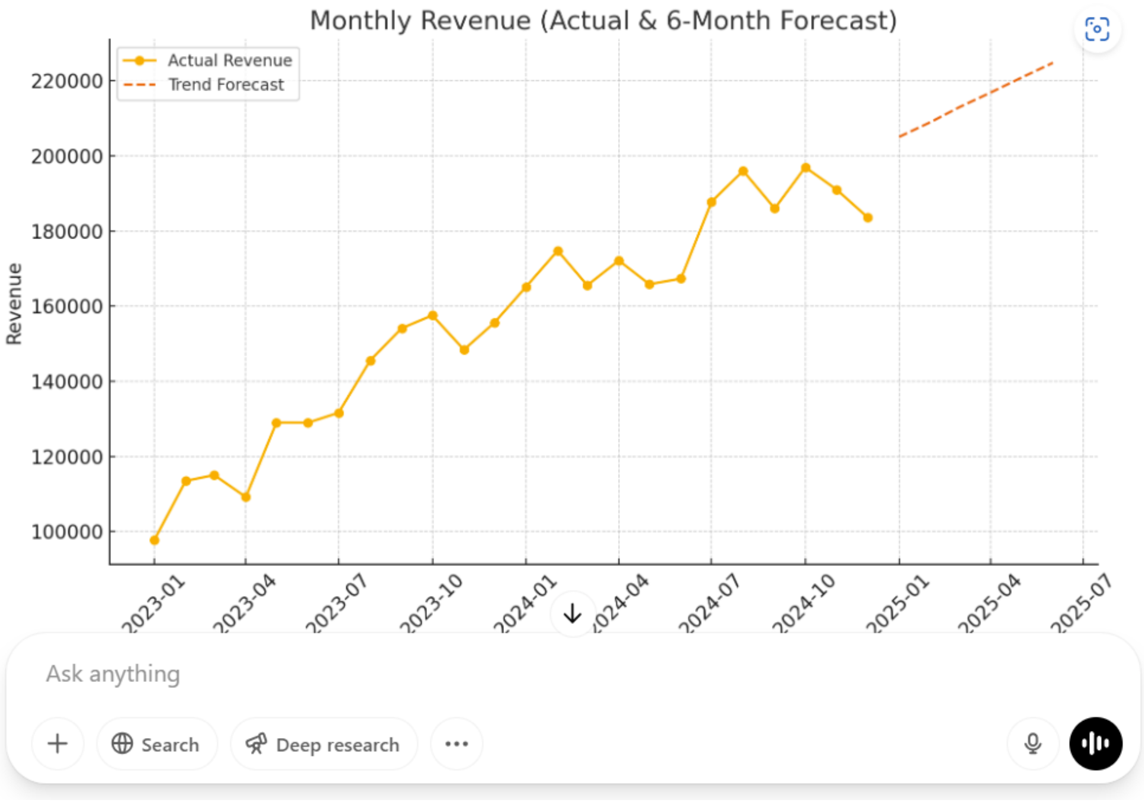
With this prompt, it analyses the data like this:



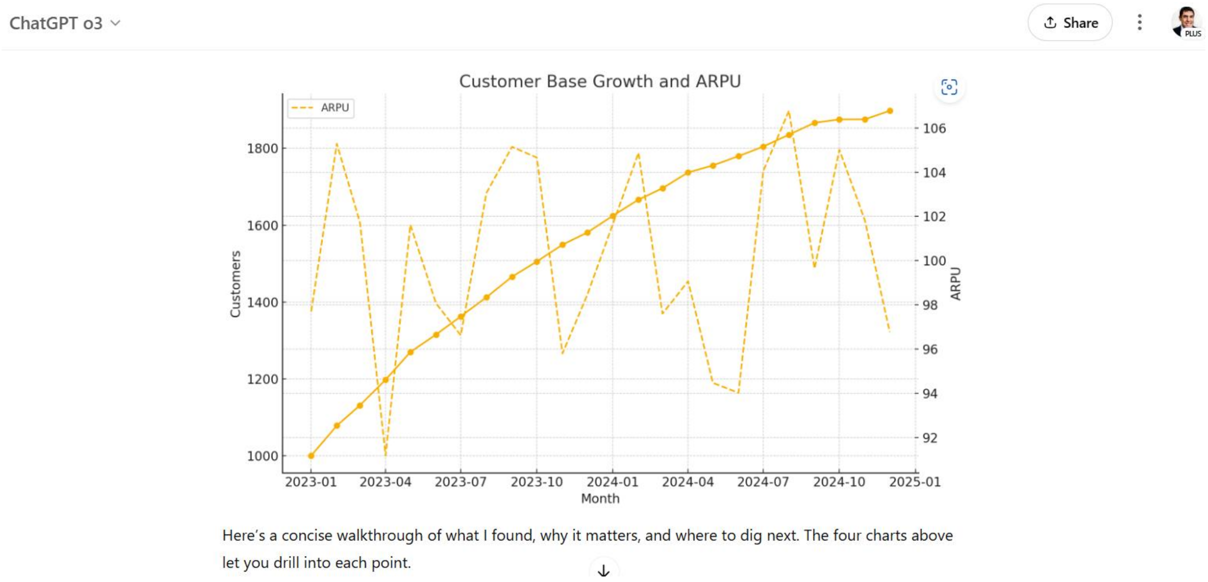
If the dataset includes time-series, apply appropriate techniques to uncover seasonality or forecasting opportunities.

```python
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

# Load the transformed data that was prepared in the analysis step
file_path = "/mnt/data/P&L SaaS.xlsx"
xls = pd.ExcelFile(file_path)
raw = xls.parse("Financials")

# --- Prepare tidy data (replicating the transformation logic) ---
cols = raw.columns[2:]
dates = [
```

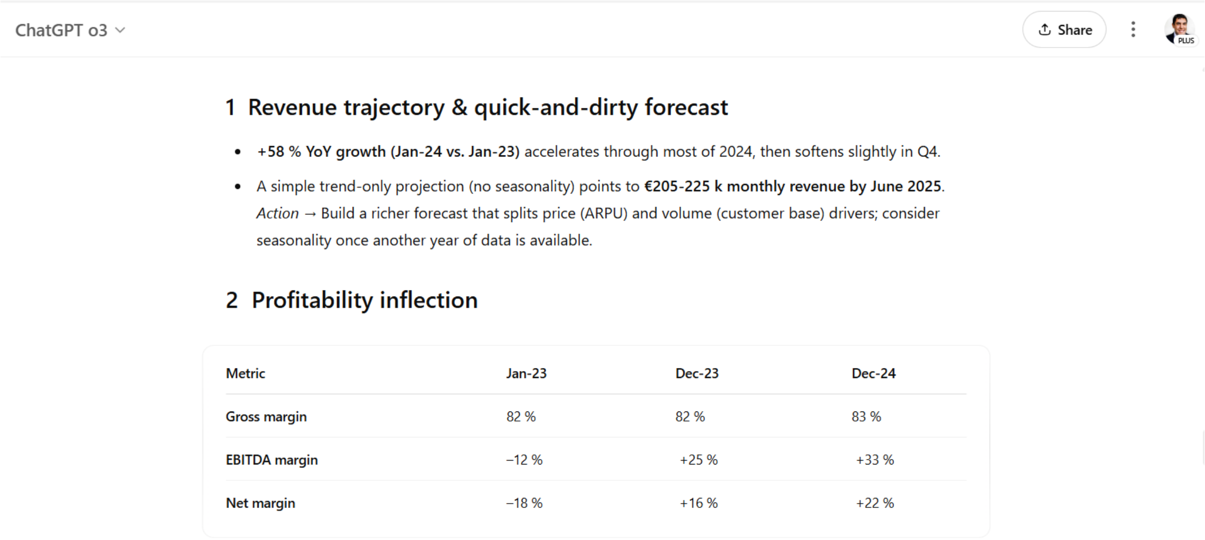
Then this would be the output:



Also:



Here's a concise walkthrough of what I found, why it matters, and where to dig next. The four charts above let you drill into each point.



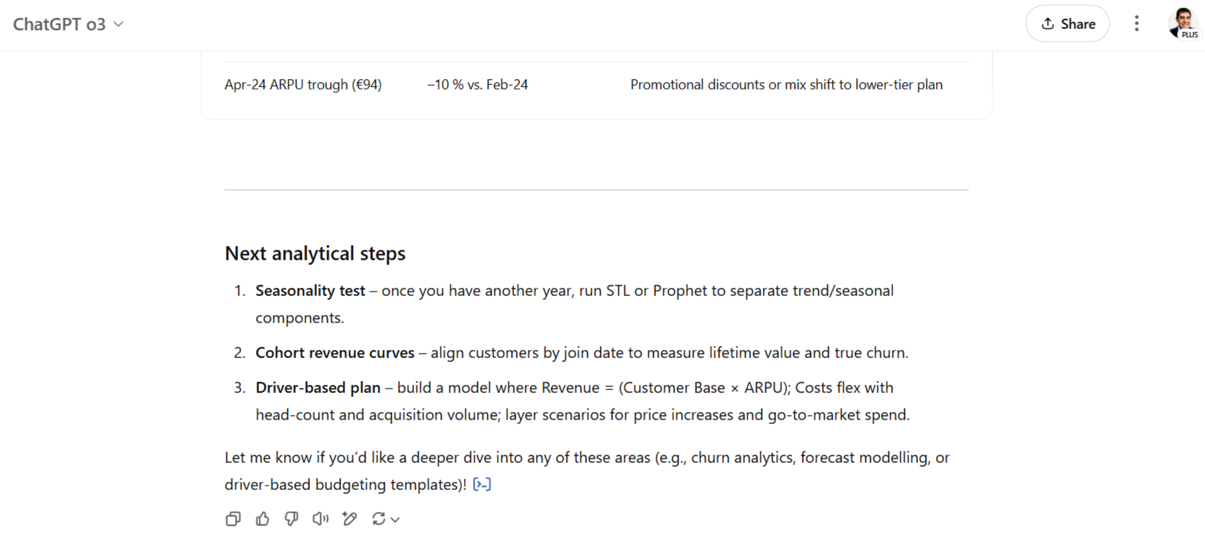
1 Revenue trajectory & quick-and-dirty forecast

- **+58 % YoY growth (Jan-24 vs. Jan-23)** accelerates through most of 2024, then softens slightly in Q4.
- A simple trend-only projection (no seasonality) points to **€205-225 k monthly revenue by June 2025**.
  Action → Build a richer forecast that splits price (ARPU) and volume (customer base) drivers; consider seasonality once another year of data is available.

2 Profitability inflection

Metric	Jan-23	Dec-23	Dec-24
Gross margin	82 %	82 %	83 %
EBITDA margin	−12 %	+25 %	+33 %
Net margin	−18 %	+16 %	+22 %

It also gives you next steps:



Apr-24 ARPU trough (€94)	−10 % vs. Feb-24	Promotional discounts or mix shift to lower-tier plan

---

Next analytical steps

1. **Seasonality test** – once you have another year, run STL or Prophet to separate trend/seasonal components.
2. **Cohort revenue curves** – align customers by join date to measure lifetime value and true churn.
3. **Driver-based plan** – build a model where Revenue = (Customer Base × ARPU); Costs flex with head-count and acquisition volume; layer scenarios for price increases and go-to-market spend.

Let me know if you'd like a deeper dive into any of these areas (e.g., churn analytics, forecast modelling, or driver-based budgeting templates)!

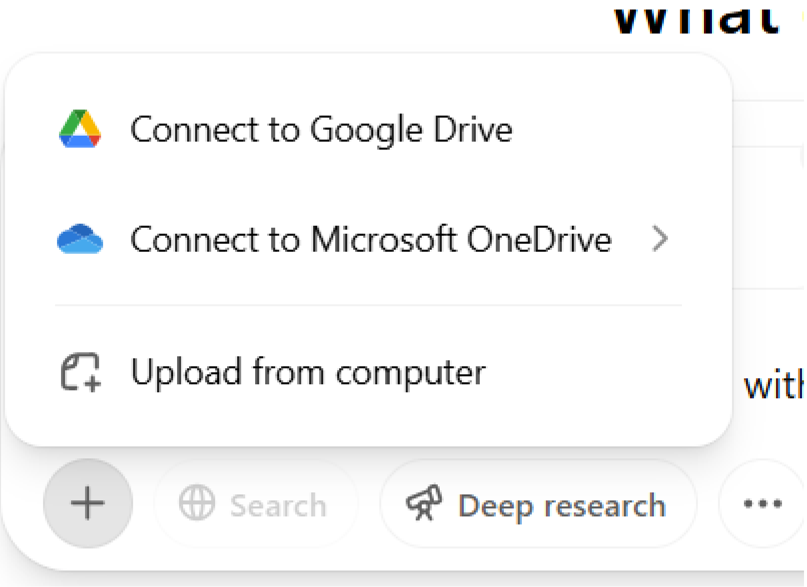
2) Multi-File Model Reconciliation

Use o3's deep context & file reasoning

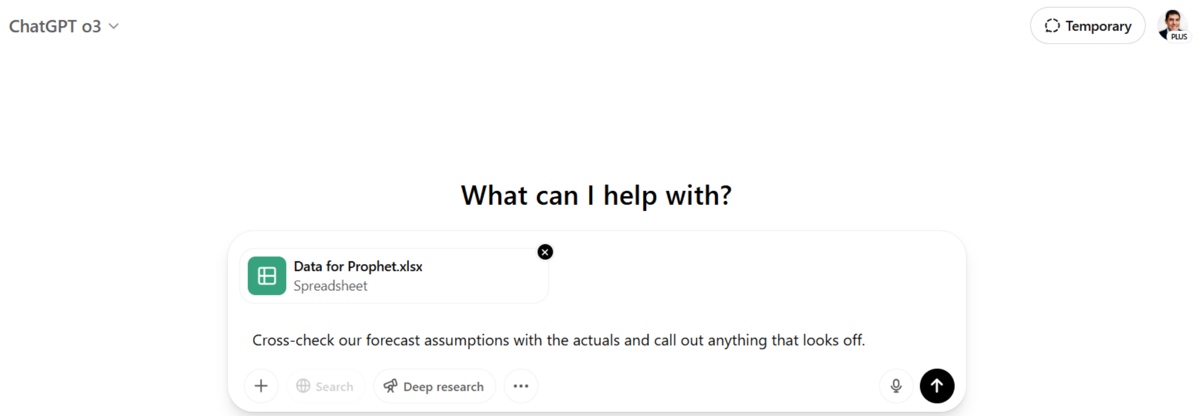
Upload actuals, forecasts, assumptions, and headcount files together. Ask o3 to reconcile them, trace formula logic, or flag inconsistencies.

Prompt: "Cross-check our forecast assumptions with the actuals and call out anything that looks off."

Remember you can add your data by clicking on the + sign

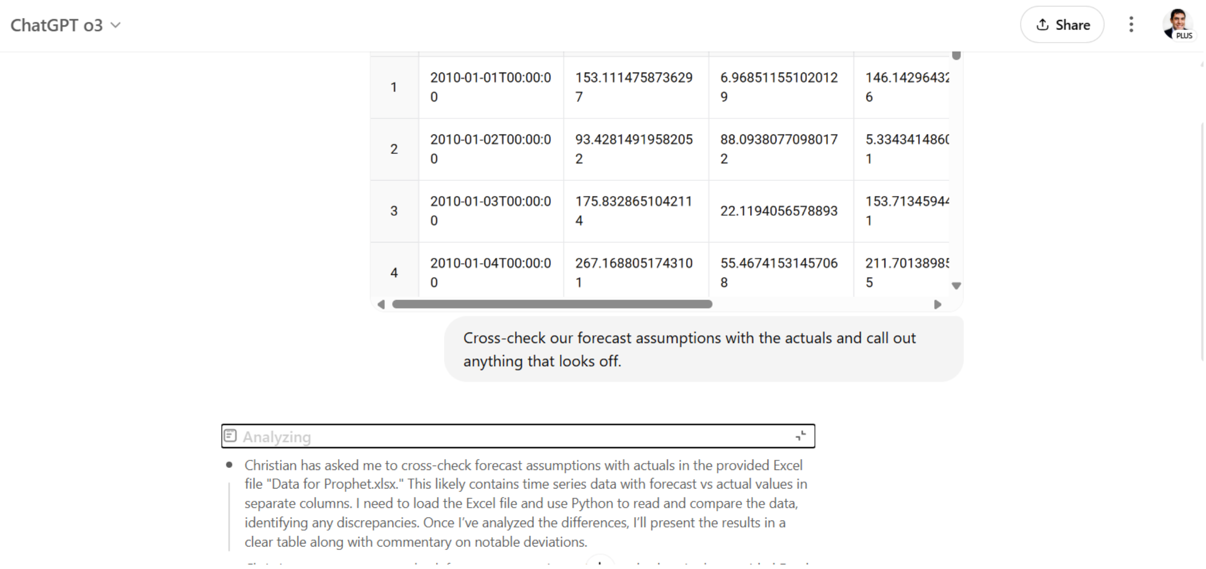


Then you'll prompt:



Cross-check our forecast assumptions with the actuals and call out anything that looks off.

ChatGPT o3 will start analyzing and telling you exactly what it is doing:

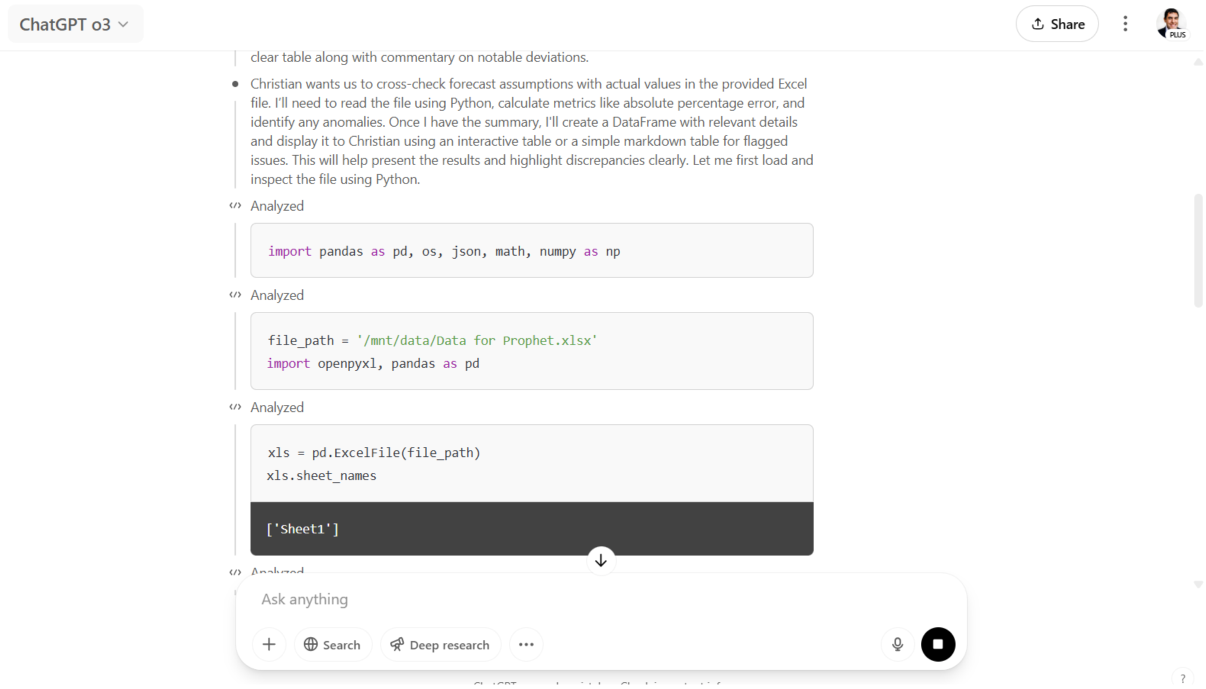


1	2010-01-01T00:00:00	153.1114758736297	6.968511551020129	146.1429643 6
2	2010-01-02T00:00:00	93.42814919582052	88.09380770980172	5.334341486 1
3	2010-01-03T00:00:00	175.8328651042114	22.1194056578893	153.7134594 1
4	2010-01-04T00:00:00	267.1688051743101	55.46741531457068	211.7013898 5

Cross-check our forecast assumptions with the actuals and call out anything that looks off.

- Christian has asked me to cross-check forecast assumptions with actuals in the provided Excel file "Data for Prophet.xlsx." This likely contains time series data with forecast vs actual values in separate columns. I need to load the Excel file and use Python to read and compare the data, identifying any discrepancies. Once I've analyzed the differences, I'll present the results in a clear table along with commentary on notable deviations.

It will use Python to analyse the data:



clear table along with commentary on notable deviations.

- Christian wants us to cross-check forecast assumptions with actual values in the provided Excel file. I'll need to read the file using Python, calculate metrics like absolute percentage error, and identify any anomalies. Once I have the summary, I'll create a DataFrame with relevant details and display it to Christian using an interactive table or a simple markdown table for flagged issues. This will help present the results and highlight discrepancies clearly. Let me first load and inspect the file using Python.

Analyzed

```
import pandas as pd, os, json, math, numpy as np
```

Analyzed

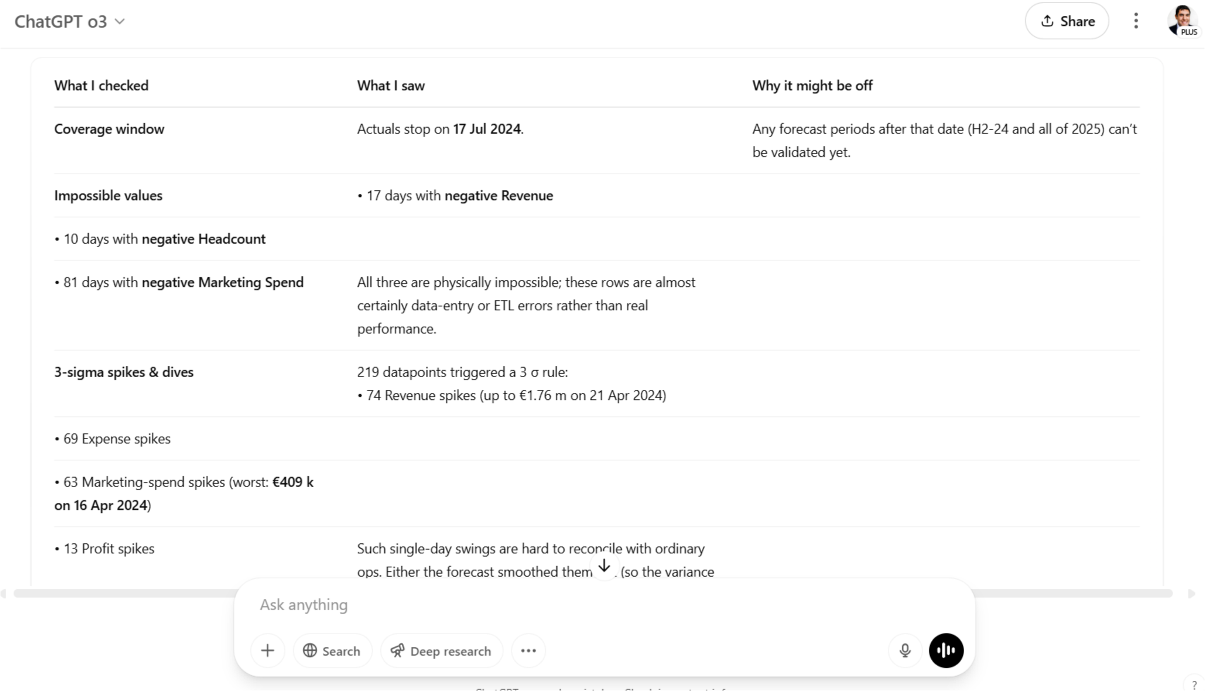
```
file_path = '/mnt/data/Data for Prophet.xlsx'
import openpyxl, pandas as pd
```

Analyzed

```
xls = pd.ExcelFile(file_path)
xls.sheet_names
```

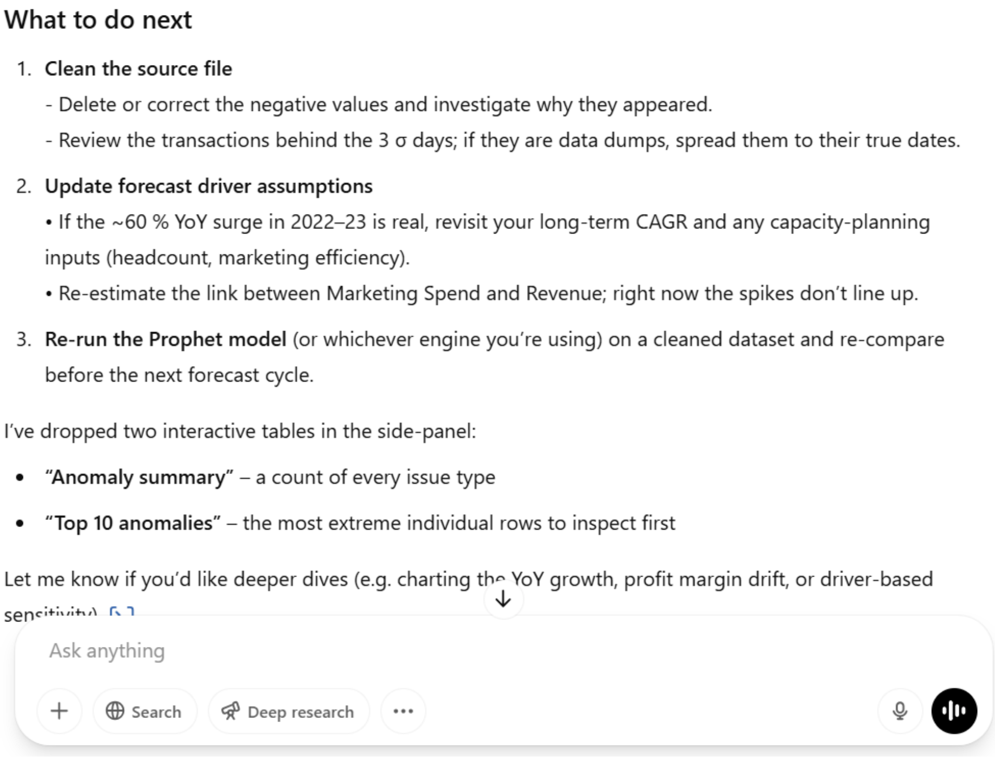
```
['Sheet1']
```

Then, this will be the output:



What I checked	What I saw	Why it might be off
Coverage window	Actuals stop on **17 Jul 2024**.	Any forecast periods after that date (H2-24 and all of 2025) can't be validated yet.
Impossible values	• 17 days with **negative Revenue**	
• 10 days with **negative Headcount**		
• 81 days with **negative Marketing Spend**	All three are physically impossible; these rows are almost certainly data-entry or ETL errors rather than real performance.	
3-sigma spikes & dives	219 datapoints triggered a 3 σ rule: • 74 Revenue spikes (up to €1.76 m on 21 Apr 2024)	
• 69 Expense spikes		
• 63 Marketing-spend spikes (worst: **€409 k on 16 Apr 2024**)		
• 13 Profit spikes	Such single-day swings are hard to reconcile with ordinary ops. Either the forecast smoothed them . (so the variance	

It also explains next steps:



What to do next

1. **Clean the source file**
   - Delete or correct the negative values and investigate why they appeared.
   - Review the transactions behind the 3 σ days; if they are data dumps, spread them to their true dates.
2. **Update forecast driver assumptions**
   • If the ~60 % YoY surge in 2022–23 is real, revisit your long-term CAGR and any capacity-planning inputs (headcount, marketing efficiency).
   • Re-estimate the link between Marketing Spend and Revenue; right now the spikes don't line up.
3. **Re-run the Prophet model** (or whichever engine you're using) on a cleaned dataset and re-compare before the next forecast cycle.

I've dropped two interactive tables in the side-panel:

- **"Anomaly summary"** – a count of every issue type
- **"Top 10 anomalies"** – the most extreme individual rows to inspect first

Let me know if you'd like deeper dives (e.g. charting the YoY growth, profit margin drift, or driver-based sensitivity)

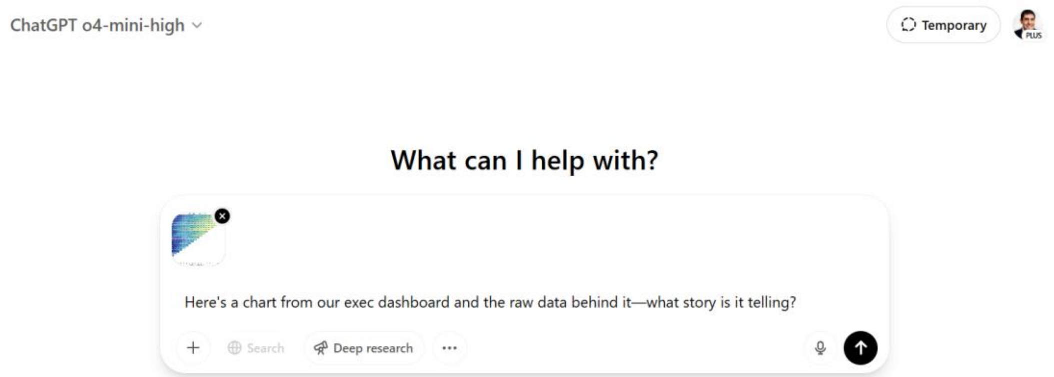
3) Visual-Driven Financial Insight

Combine image + data reasoning

Upload screenshots from dashboards (e.g. Power BI, Tableau, Excel, Python) and the underlying data, then ask o3 to interpret both and find correlations.

Prompt: "Here's a chart from our exec dashboard and the raw data behind it—what story is it telling?"

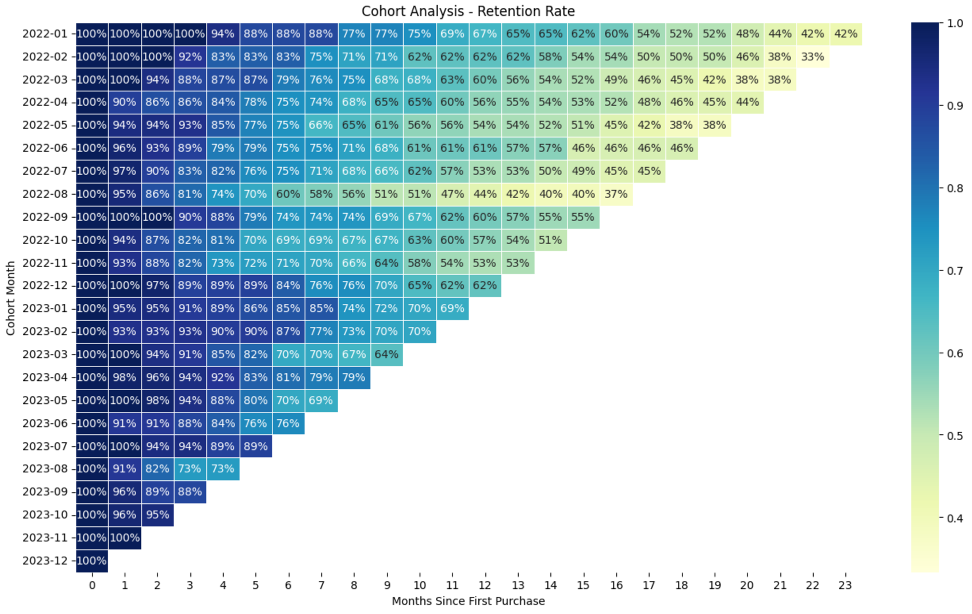
This is how I tried it:



What can I help with?

Here's a chart from our exec dashboard and the raw data behind it—what story is it telling?

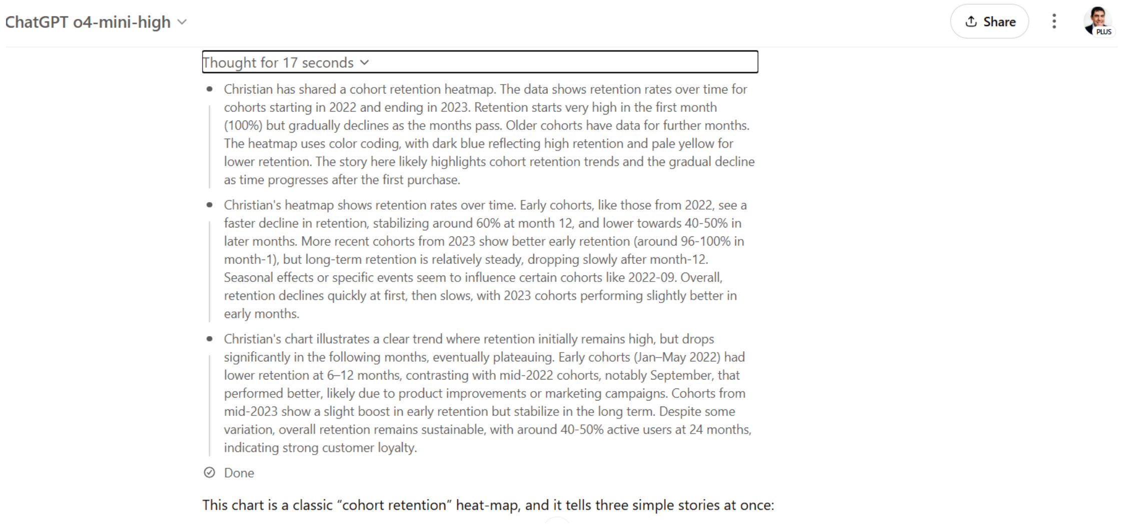
This is the chart I used:



Cohort Analysis - Retention Rate

Cohort Month	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
2022-01	100%	100%	100%	100%	94%	88%	88%	88%	77%	77%	75%	69%	67%	65%	65%	62%	60%	54%	52%	52%	48%	44%	42%	42%
2022-02	100%	100%	100%	92%	83%	83%	83%	75%	71%	71%	62%	62%	62%	62%	58%	54%	54%	50%	50%	50%	46%	38%	33%	
2022-03	100%	100%	94%	88%	87%	87%	79%	76%	75%	68%	68%	63%	60%	56%	54%	52%	49%	46%	45%	42%	38%	38%		
2022-04	100%	90%	86%	86%	84%	78%	75%	74%	68%	65%	65%	60%	56%	55%	54%	53%	52%	48%	46%	45%	44%			
2022-05	100%	94%	94%	93%	85%	77%	75%	66%	65%	61%	56%	56%	54%	54%	52%	51%	45%	42%	38%	38%				
2022-06	100%	96%	93%	89%	79%	79%	75%	75%	71%	68%	61%	61%	61%	57%	57%	46%	46%	46%	46%					
2022-07	100%	97%	90%	83%	82%	76%	75%	71%	68%	66%	62%	57%	53%	53%	50%	49%	45%	45%						
2022-08	100%	95%	86%	81%	74%	70%	60%	58%	56%	51%	51%	47%	44%	42%	40%	40%	37%							
2022-09	100%	100%	100%	90%	88%	79%	74%	74%	74%	69%	67%	62%	60%	57%	55%	55%								
2022-10	100%	94%	87%	82%	81%	70%	69%	69%	67%	67%	63%	60%	57%	54%	51%									
2022-11	100%	93%	88%	82%	73%	72%	71%	70%	66%	64%	58%	54%	53%	53%										
2022-12	100%	100%	97%	89%	89%	89%	84%	76%	76%	70%	65%	62%	62%											
2023-01	100%	95%	95%	91%	89%	86%	85%	85%	74%	72%	70%	69%												
2023-02	100%	93%	93%	93%	90%	90%	87%	77%	73%	70%	70%													
2023-03	100%	100%	94%	91%	85%	82%	70%	70%	67%	64%														
2023-04	100%	98%	96%	94%	92%	83%	81%	79%	79%															
2023-05	100%	100%	98%	94%	88%	80%	70%	69%																
2023-06	100%	91%	91%	88%	84%	76%	76%																	
2023-07	100%	100%	94%	94%	89%	89%																		
2023-08	100%	91%	82%	73%	73%																			
2023-09	100%	96%	89%	88%																				
2023-10	100%	96%	95%																					
2023-11	100%	100%																						
2023-12	100%																							

And the analysis ChatGPT did:



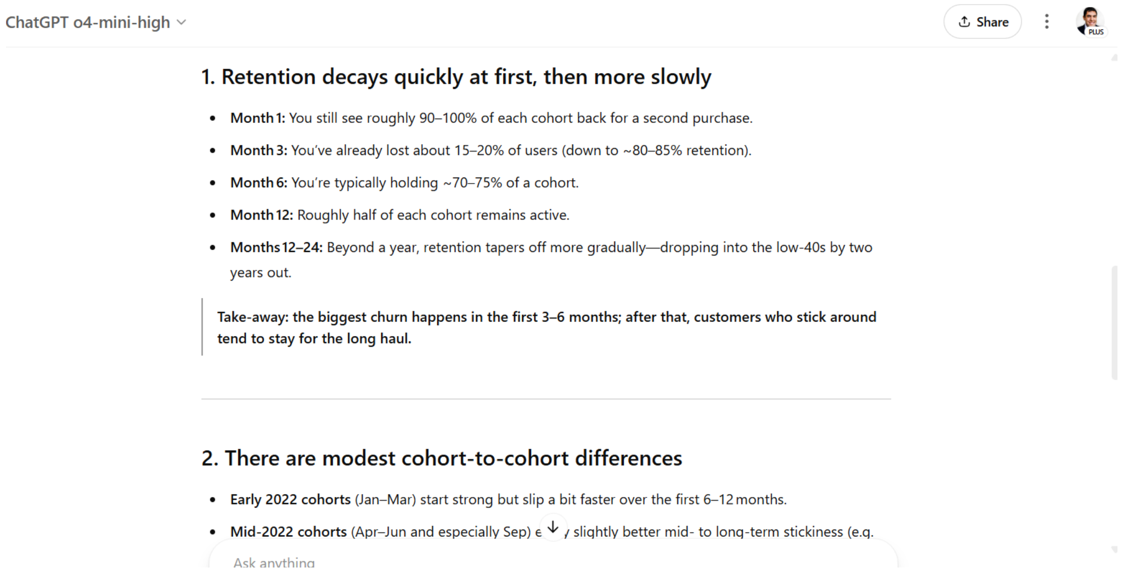
Thought for 17 seconds

- Christian has shared a cohort retention heatmap. The data shows retention rates over time for cohorts starting in 2022 and ending in 2023. Retention starts very high in the first month (100%) but gradually declines as the months pass. Older cohorts have data for further months. The heatmap uses color coding, with dark blue reflecting high retention and pale yellow for lower retention. The story here likely highlights cohort retention trends and the gradual decline as time progresses after the first purchase.
- Christian's heatmap shows retention rates over time. Early cohorts, like those from 2022, see a faster decline in retention, stabilizing around 60% at month 12, and lower towards 40-50% in later months. More recent cohorts from 2023 show better early retention (around 96-100% in month-1), but long-term retention is relatively steady, dropping slowly after month-12. Seasonal effects or specific events seem to influence certain cohorts like 2022-09. Overall, retention declines quickly at first, then slows, with 2023 cohorts performing slightly better in early months.
- Christian's chart illustrates a clear trend where retention initially remains high, but drops significantly in the following months, eventually plateauing. Early cohorts (Jan–May 2022) had lower retention at 6–12 months, contrasting with mid-2022 cohorts, notably September, that performed better, likely due to product improvements or marketing campaigns. Cohorts from mid-2023 show a slight boost in early retention but stabilize in the long term. Despite some variation, overall retention remains sustainable, with around 40-50% active users at 24 months, indicating strong customer loyalty.

Done

This chart is a classic "cohort retention" heat-map, and it tells three simple stories at once:

And the output:



1. Retention decays quickly at first, then more slowly

- **Month 1:** You still see roughly 90–100% of each cohort back for a second purchase.
- **Month 3:** You've already lost about 15–20% of users (down to ~80–85% retention).
- **Month 6:** You're typically holding ~70–75% of a cohort.
- **Month 12:** Roughly half of each cohort remains active.
- **Months 12–24:** Beyond a year, retention tapers off more gradually—dropping into the low-40s by two years out.

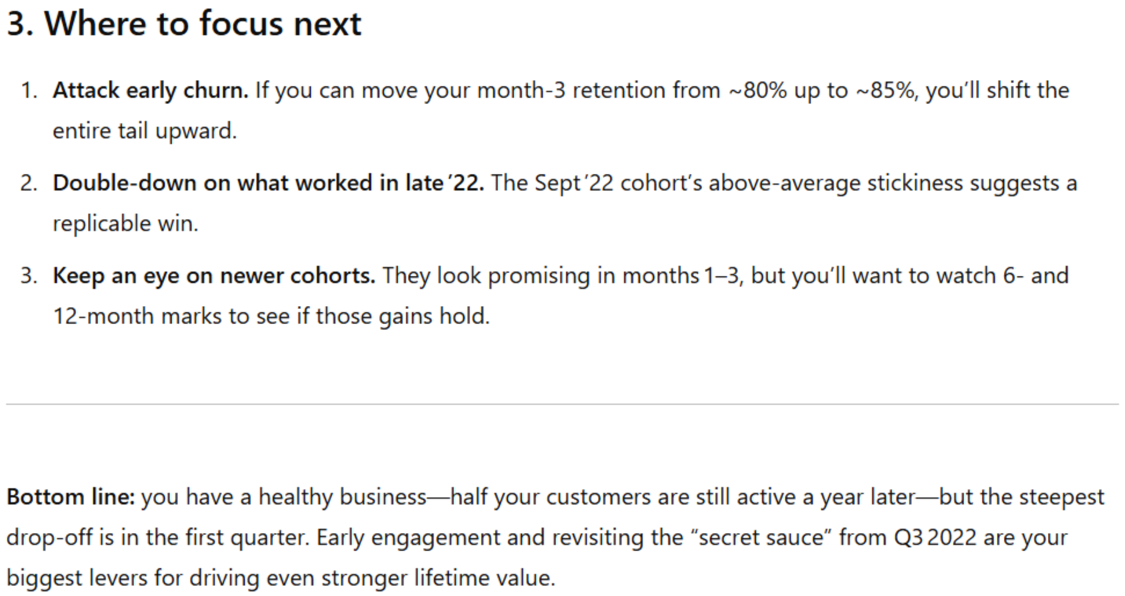
> **Take-away: the biggest churn happens in the first 3–6 months; after that, customers who stick around tend to stay for the long haul.**

---

2. There are modest cohort-to-cohort differences

- **Early 2022 cohorts** (Jan–Mar) start strong but slip a bit faster over the first 6–12 months.
- **Mid-2022 cohorts** (Apr–Jun and especially Sep) e ↓ y slightly better mid- to long-term stickiness (e.g.

It also gave next steps:



3. Where to focus next

1. **Attack early churn.** If you can move your month-3 retention from ~80% up to ~85%, you'll shift the entire tail upward.
2. **Double-down on what worked in late '22.** The Sept '22 cohort's above-average stickiness suggests a replicable win.
3. **Keep an eye on newer cohorts.** They look promising in months 1–3, but you'll want to watch 6- and 12-month marks to see if those gains hold.

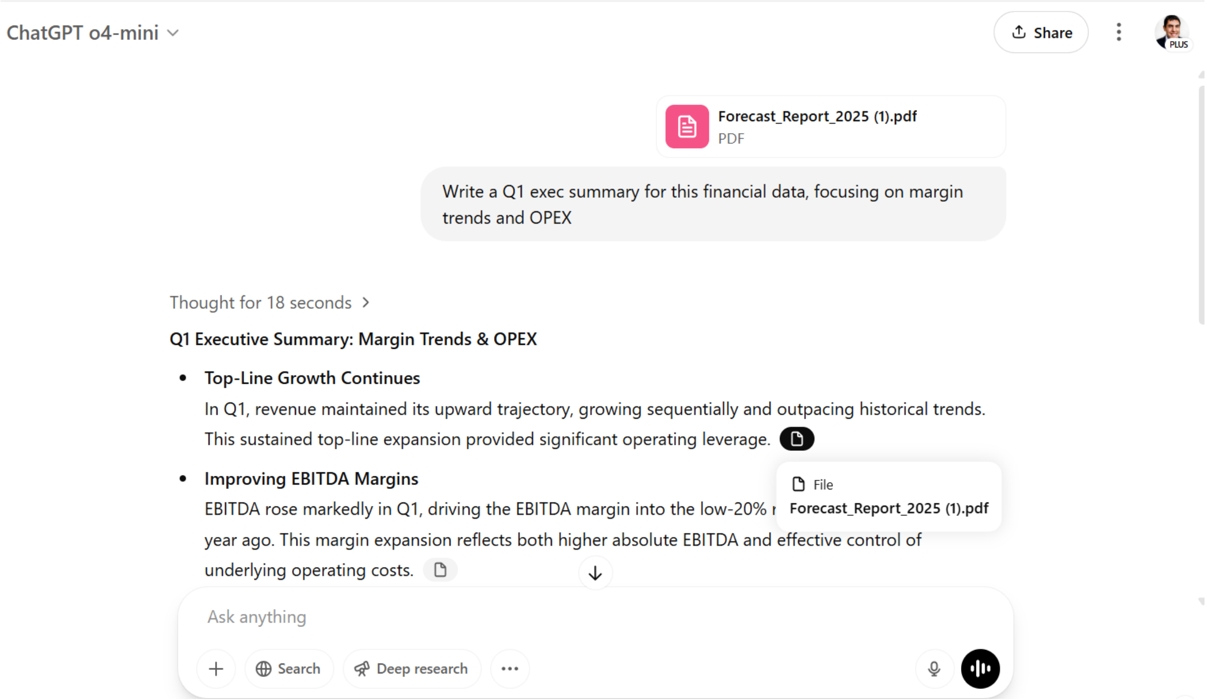
Bottom line: you have a healthy business—half your customers are still active a year later—but the steepest drop-off is in the first quarter. Early engagement and revisiting the "secret sauce" from Q3 2022 are your biggest levers for driving even stronger lifetime value.

4) Automated Board & Exec Pack Narratives

Pair analysis with natural language generation

Generate smart commentary based on financial data, trends, and variances, tailored for different audiences.

Prompt: "Write a Q1 exec summary for this financial data, focusing on margin trends and OPEX."



Write a Q1 exec summary for this financial data, focusing on margin trends and OPEX

Q1 Executive Summary: Margin Trends & OPEX

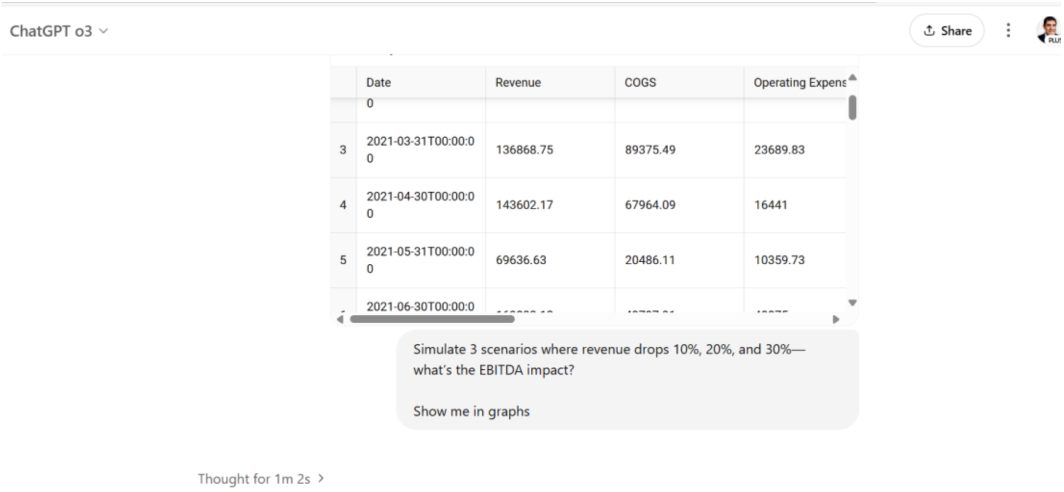
- **Top-Line Growth Continues**
  In Q1, revenue maintained its upward trajectory, growing sequentially and outpacing historical trends. This sustained top-line expansion provided significant operating leverage.
- **Improving EBITDA Margins**
  EBITDA rose markedly in Q1, driving the EBITDA margin into the low-20% r
  year ago. This margin expansion reflects both higher absolute EBITDA and effective control of underlying operating costs.

5) Strategic Scenario Simulation

Dynamic, parameter-based simulations

Ask o3 to model financial impact from scenario inputs (price changes, headcount freezes, market downturns, etc).

Prompt: "Simulate 3 scenarios where revenue drops 10%, 20%, and 30%—what's the EBITDA impact?"



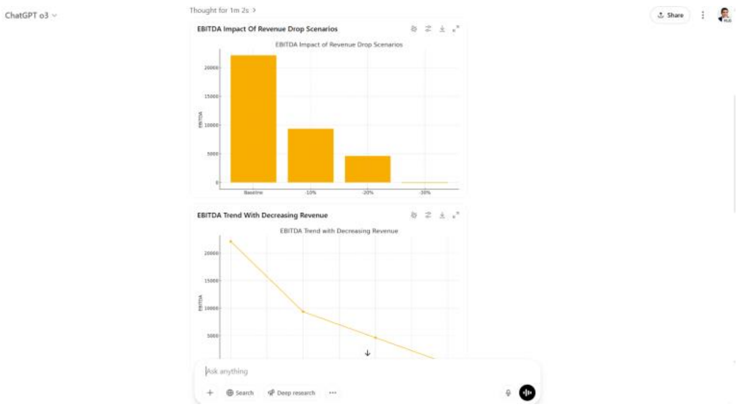
	Date	Revenue	COGS	Operating Expens
	0			
3	2021-03-31T00:00:00	136868.75	89375.49	23689.83
4	2021-04-30T00:00:00	143602.17	67964.09	16441
5	2021-05-31T00:00:00	69636.63	20486.11	10359.73
	2021-06-30T00:00:0			

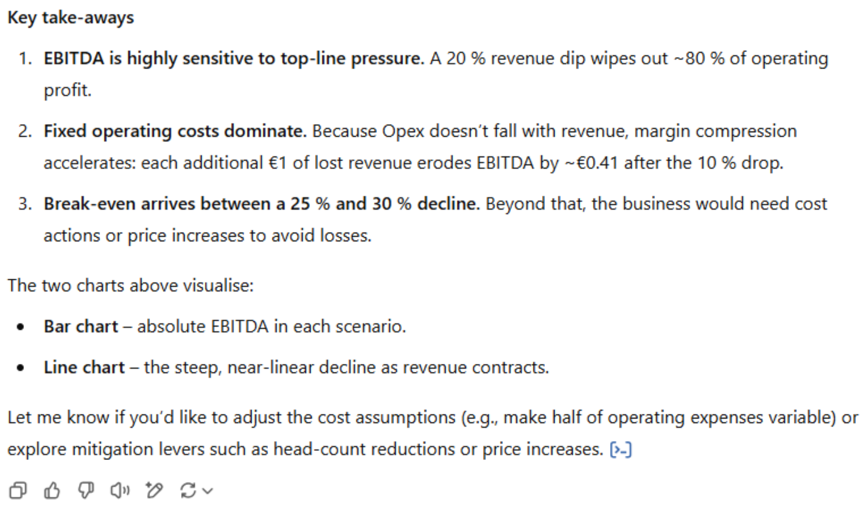
Simulate 3 scenarios where revenue drops 10%, 20%, and 30%—what's the EBITDA impact?

Show me in graphs

Thought for 1m 2s

And then, this will be the output:


ChatGPT o3
Thought for 1m 2s
Share
EBITDA Impact Of Revenue Drop Scenarios
EBITDA Impact of Revenue Drop Scenarios
EBITDA
20000
15000
10000
5000
0
Baseline
-10%
-20%
-30%
EBITDA Trend With Decreasing Revenue
EBITDA Trend with Decreasing Revenue
EBITDA
20000
15000
10000
5000
Ask anything
Search
Deep research



Key take-aways

1. **EBITDA is highly sensitive to top-line pressure.** A 20 % revenue dip wipes out ~80 % of operating profit.
2. **Fixed operating costs dominate.** Because Opex doesn't fall with revenue, margin compression accelerates: each additional €1 of lost revenue erodes EBITDA by ~€0.41 after the 10 % drop.
3. **Break-even arrives between a 25 % and 30 % decline.** Beyond that, the business would need cost actions or price increases to avoid losses.

The two charts above visualise:

- **Bar chart** – absolute EBITDA in each scenario.
- **Line chart** – the steep, near-linear decline as revenue contracts.

Let me know if you'd like to adjust the cost assumptions (e.g., make half of operating expenses variable) or explore mitigation levers such as head-count reductions or price increases.

Other examples here:

6) Assumption Audit Trail

Use autonomous Python + multi-step logic
Let o3 parse through your forecast model and validate assumptions: growth rates, margins, headcount ramp, CAC, etc.

Prompt: "Audit our 3-year model and flag any assumptions that are unrealistic or unsupported."

7) Build FP&A Copilot for Month-End Close

Tool use + memory = personalized workflows
Create a recurring assistant that guides your close: checklist steps, variance analysis, commentary generation, and data validation.

Prompt: "Help us automate our monthly close checklist and variance commentary across departments."

8) Cohort & Time-Based Analysis

Advanced time-series forecasting
Upload usage, churn, or revenue data and ask o3 to segment by cohort, apply rolling averages, or forecast future trends.

Prompt: "Analyze this ARR dataset and build a 12-month retention forecast by customer cohort."

9) Budget Consolidation Across Departments

Handle multi-file logic & comparisons
Upload multiple department budgets and ask o3 to combine, compare, and detect misalignments in formatting, assumptions, or trends.

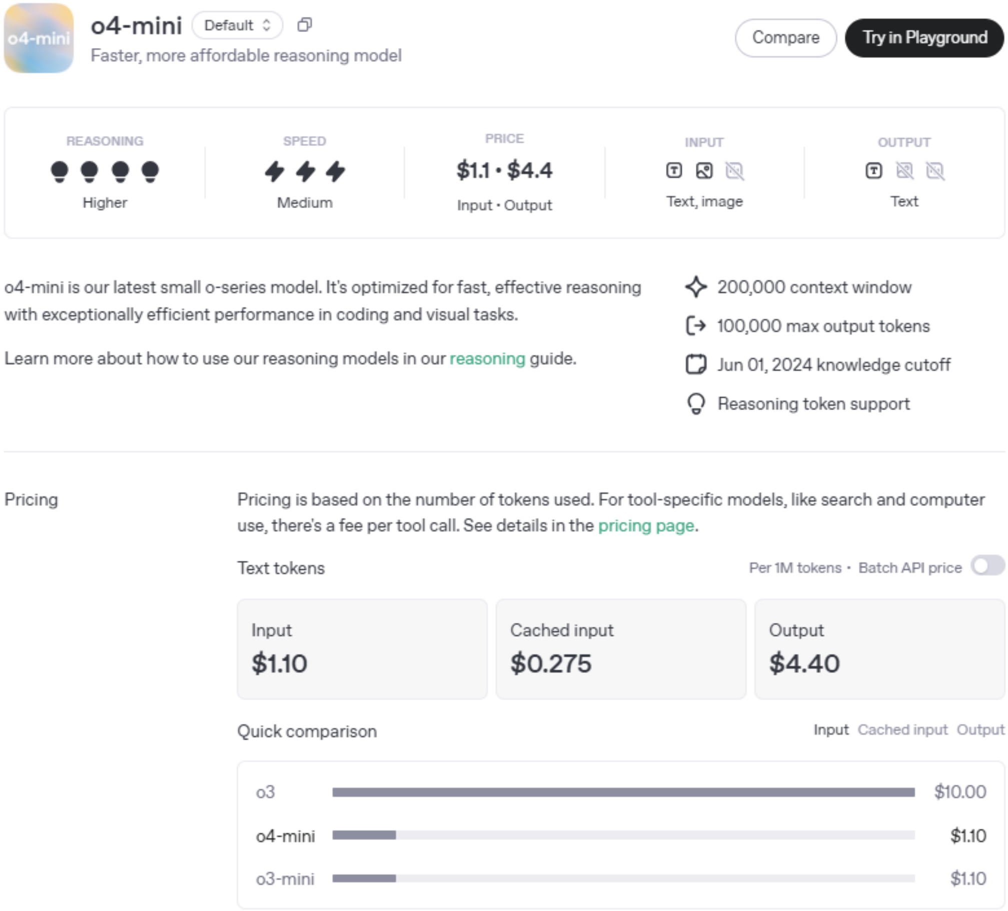
Prompt: "Compare marketing and product budgets and highlight inconsistencies or outliers."

10) Market Benchmark + Risk Research

Use "Deep Research" to enhance forecasting
Ask ChatGPT to pull in real-world data: cost of capital, inflation forecasts, peer SG&A, or macro risk signals, then incorporate that into models.

Prompt: "Research inflation forecasts for 2025 and suggest how they might impact our OPEX assumptions."

I also want to give you an overview of each of the new models:



o4-mini

Faster, more affordable reasoning model

REASONING	SPEED	PRICE	INPUT	OUTPUT
● ● ● ● Higher	ϟ ϟ ϟ Medium	$1.1 • $4.4 Input • Output	Text, image	Text

o4-mini is our latest small o-series model. It's optimized for fast, effective reasoning with exceptionally efficient performance in coding and visual tasks.

Learn more about how to use our reasoning models in our reasoning guide.

- 200,000 context window
- 100,000 max output tokens
- Jun 01, 2024 knowledge cutoff
- Reasoning token support

Pricing

Pricing is based on the number of tokens used. For tool-specific models, like search and computer use, there's a fee per tool call. See details in the pricing page.

Text tokens

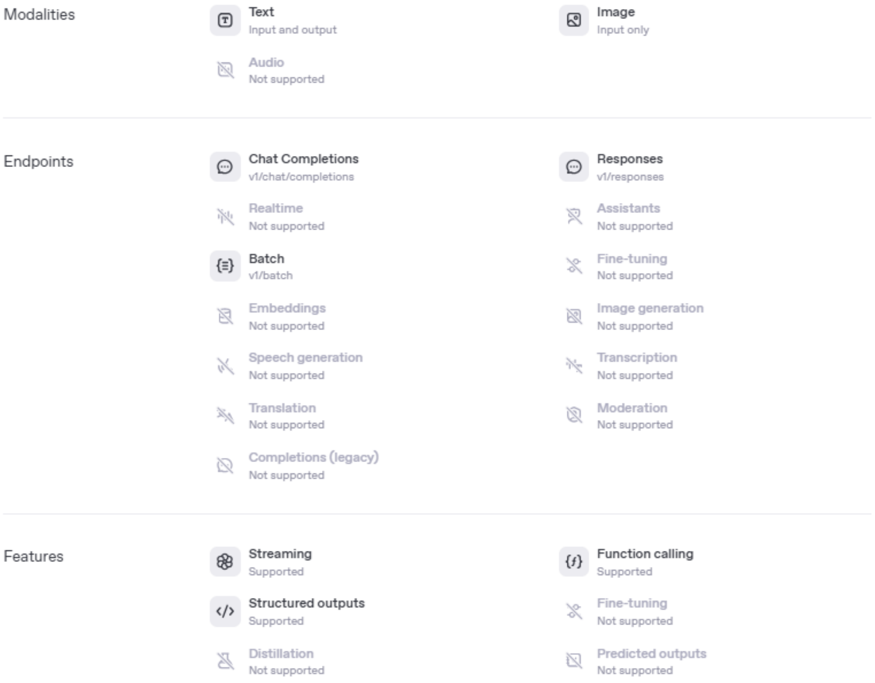
Per 1M tokens • Batch API price

Input	Cached input	Output
$1.10	$0.275	$4.40

Quick comparison

Input Cached input Output

o3	$10.00
o4-mini	$1.10
o3-mini	$1.10



Modalities

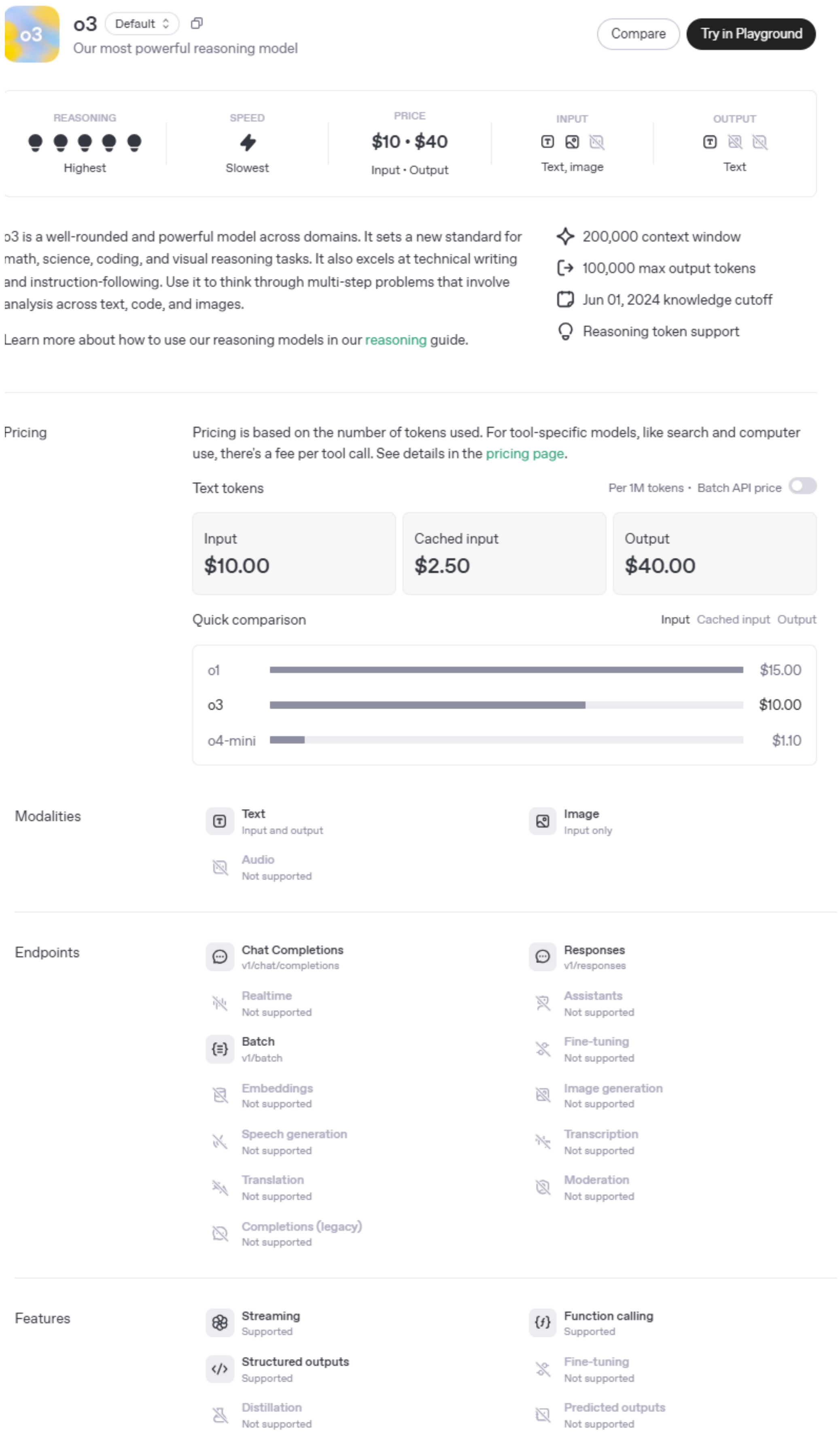
Text Input and output	**Image** Input only
Audio Not supported	

Endpoints

Chat Completions v1/chat/completions	**Responses** v1/responses
Realtime Not supported	Assistants Not supported
Batch v1/batch	Fine-tuning Not supported
Embeddings Not supported	Image generation Not supported
Speech generation Not supported	Transcription Not supported
Translation Not supported	Moderation Not supported
Completions (legacy) Not supported	

Features

Streaming Supported	**Function calling** Supported
Structured outputs Supported	Fine-tuning Not supported
Distillation Not supported	Predicted outputs Not supported



o3 Default

Our most powerful reasoning model

REASONING	SPEED	PRICE	INPUT	OUTPUT
● ● ● ● ●	⚡	$10 • $40		
Highest	Slowest	Input • Output	Text, image	Text

o3 is a well-rounded and powerful model across domains. It sets a new standard for math, science, coding, and visual reasoning tasks. It also excels at technical writing and instruction-following. Use it to think through multi-step problems that involve analysis across text, code, and images.

Learn more about how to use our reasoning models in our reasoning guide.

- 200,000 context window
- 100,000 max output tokens
- Jun 01, 2024 knowledge cutoff
- Reasoning token support

Pricing

Pricing is based on the number of tokens used. For tool-specific models, like search and computer use, there's a fee per tool call. See details in the pricing page.

Text tokens

Input	Cached input	Output
$10.00	$2.50	$40.00

Quick comparison

Modalities

Text
Input and output

Image
Input only

Audio
Not supported

Endpoints

Chat Completions
v1/chat/completions

Responses
v1/responses

Realtime
Not supported

Assistants
Not supported

Batch
v1/batch

Fine-tuning
Not supported

Embeddings
Not supported

Image generation
Not supported

Speech generation
Not supported

Transcription
Not supported

Translation
Not supported

Moderation
Not supported

Completions (legacy)
Not supported

Features

Streaming
Supported

Function calling
Supported

Structured outputs
Supported

Fine-tuning
Not supported

Distillation
Not supported

Predicted outputs
Not supported

GPT-4.1

Default

Flagship GPT model for complex tasks

Compare Try in Playground

INTELLIGENCE	SPEED	PRICE	INPUT	OUTPUT
●●●●	⚡⚡⚡	$2 • $8		
Higher	Medium	Input • Output	Text, image	Text

GPT-4.1 is our flagship model for complex tasks. It is well suited for problem solving across domains.

- 1,047,576 context window
- 32,768 max output tokens
- Jun 01, 2024 knowledge cutoff

Pricing

Pricing is based on the number of tokens used. For tool-specific models, like search and computer use, there's a fee per tool call. See details in the pricing page.

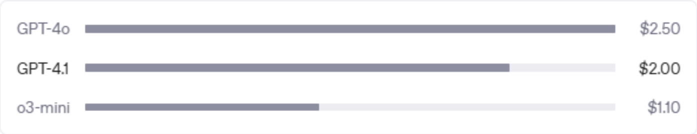
Text tokens

Per 1M tokens • Batch API price ○

Input	Cached input	Output
$2.00	$0.50	$8.00

Quick comparison

Input Cached input Output



Modalities

Text Input and output	**Image** Input only
Audio Not supported	

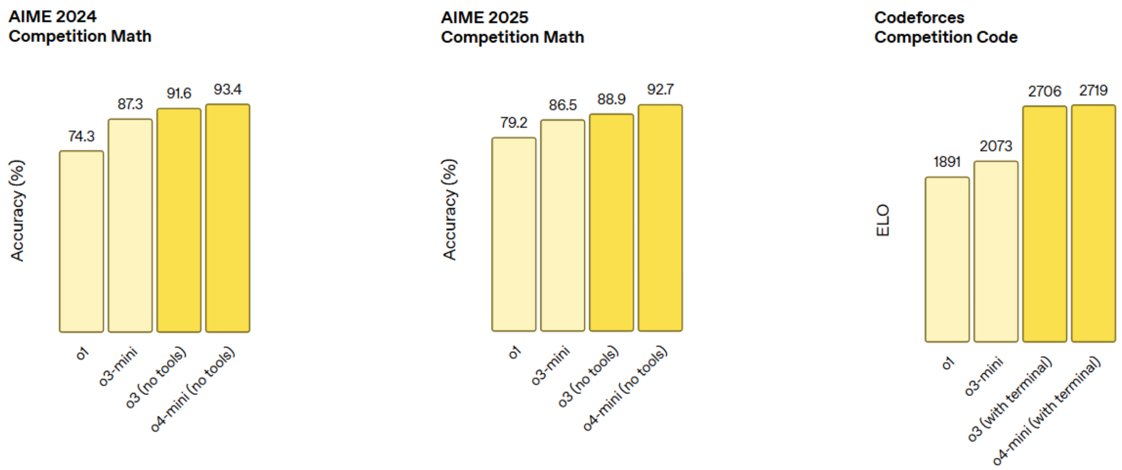
Endpoints

Chat Completions v1/chat/completions	**Responses** v1/responses
Realtime Not supported	**Assistants** v1/assistants
Batch v1/batch	**Fine-tuning** v1/fine-tuning
Embeddings Not supported	Image generation Not supported
Speech generation Not supported	Transcription Not supported
Translation Not supported	Moderation Not supported
Completions (legacy) Not supported	

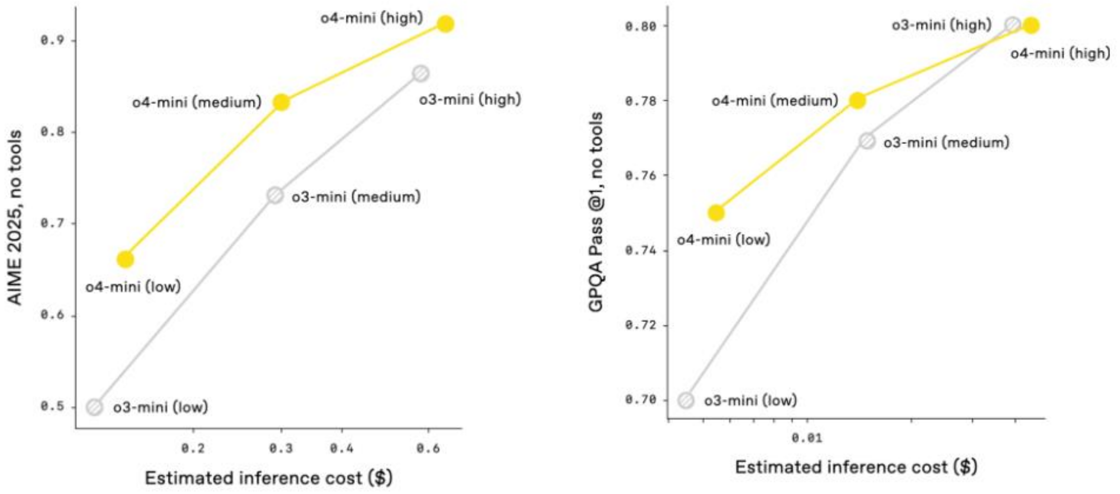
Features

Streaming Supported	**Function calling** Supported
Structured outputs Supported	**Fine-tuning** Supported
Distillation Supported	**Predicted outputs** Supported

Here are also some comparison with other models:



Cost vs performance: o3-mini and o4-mini


AIME 2025, no tools
Estimated inference cost ($)
o4-mini (high)
o4-mini (medium)
o3-mini (high)
o3-mini (medium)
o4-mini (low)
o3-mini (low)
GPQA Pass @1, no tools
Estimated inference cost ($)
o3-mini (high)
o4-mini (high)
o4-mini (medium)
o3-mini (medium)
o4-mini (low)
o3-mini (low)