Understanding AI Costs

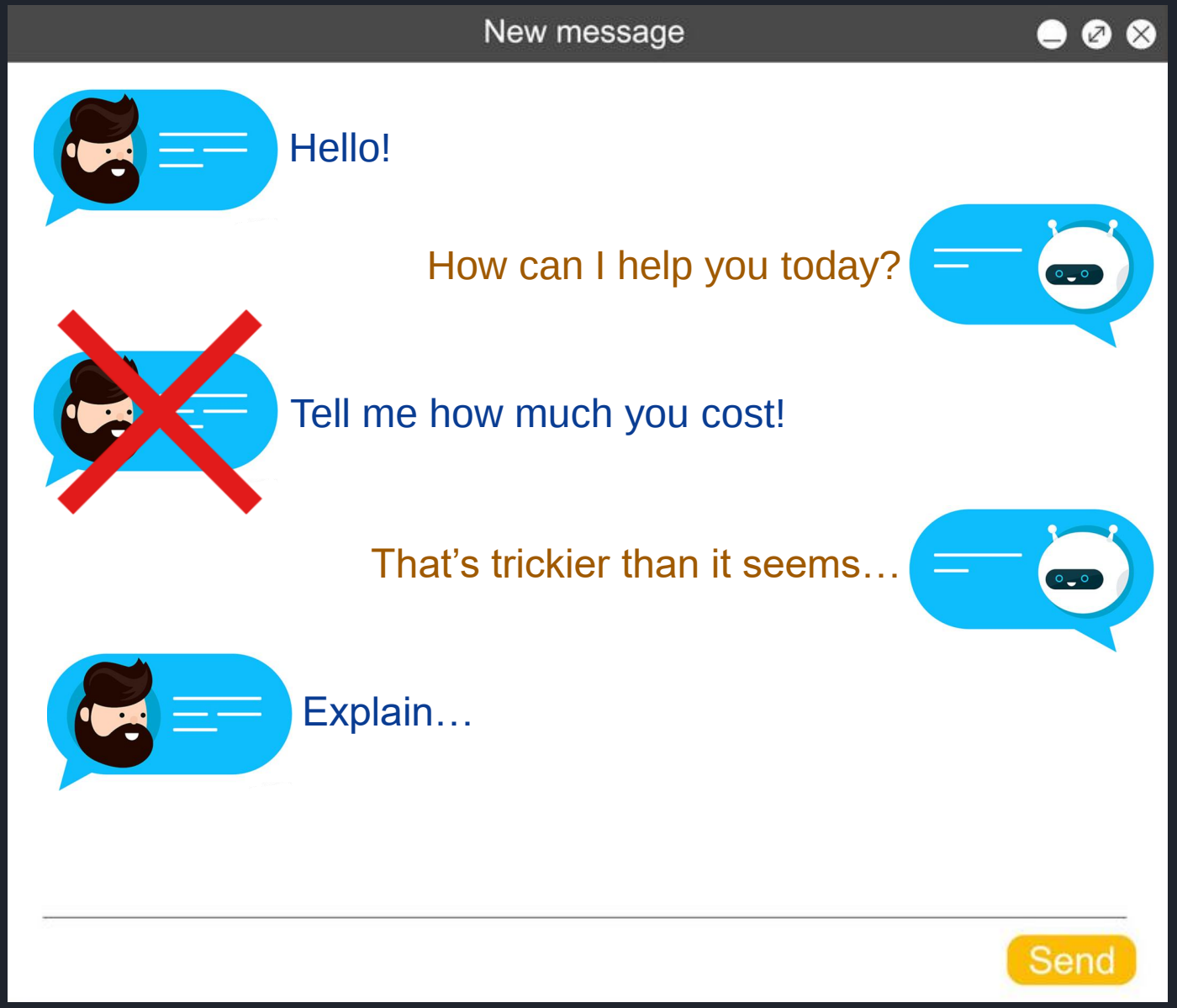
Stanford University
5/20/25

How much does GPT-4o cost?

- Completion
  - Input
    - $5.00 / 1M tokens
  - Output
    - $15.00 / 1M tokens

Pay-i

How much does GPT-4o cost?

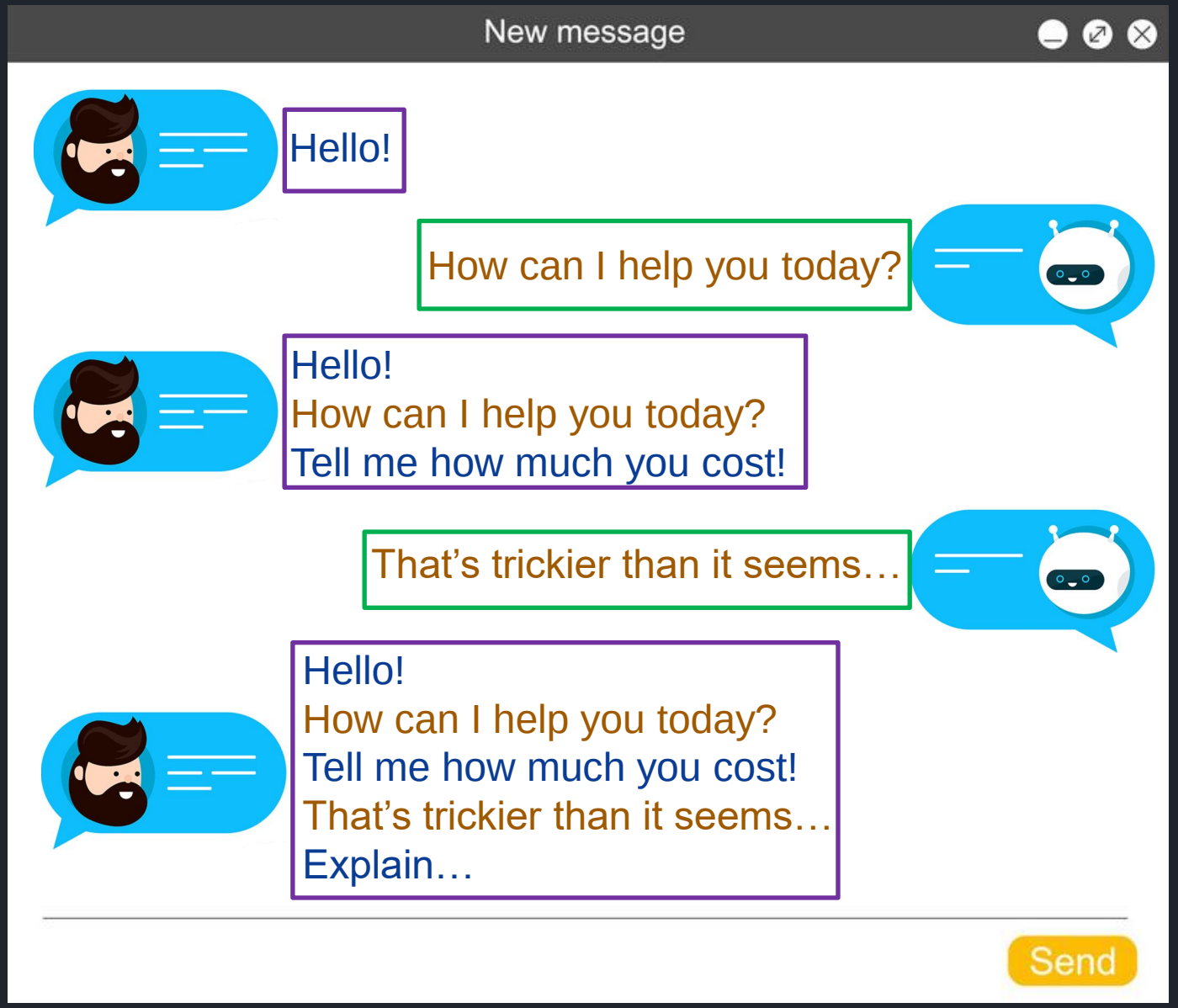


Pay-i

How much does GPT-4o cost?

= Completion Input

= Completion Output



Pay-i

How much does GPT-4o cost?

- Completion
  - Input
    - $5.00 / 1M tokens
  - Output
    - $15.00 / 1M tokens

How much does GPT-4o cost?

- Completion
  - Input
    - $5.00 / 1M tokens
  - Output
    - $15.00 / 1M tokens
- Batch
  - Input
    - $2.50 / 1M tokens
  - Output
    - $7.50 / 1M tokens
- Vision
  - Input
    - ?
  - Output
    - ?

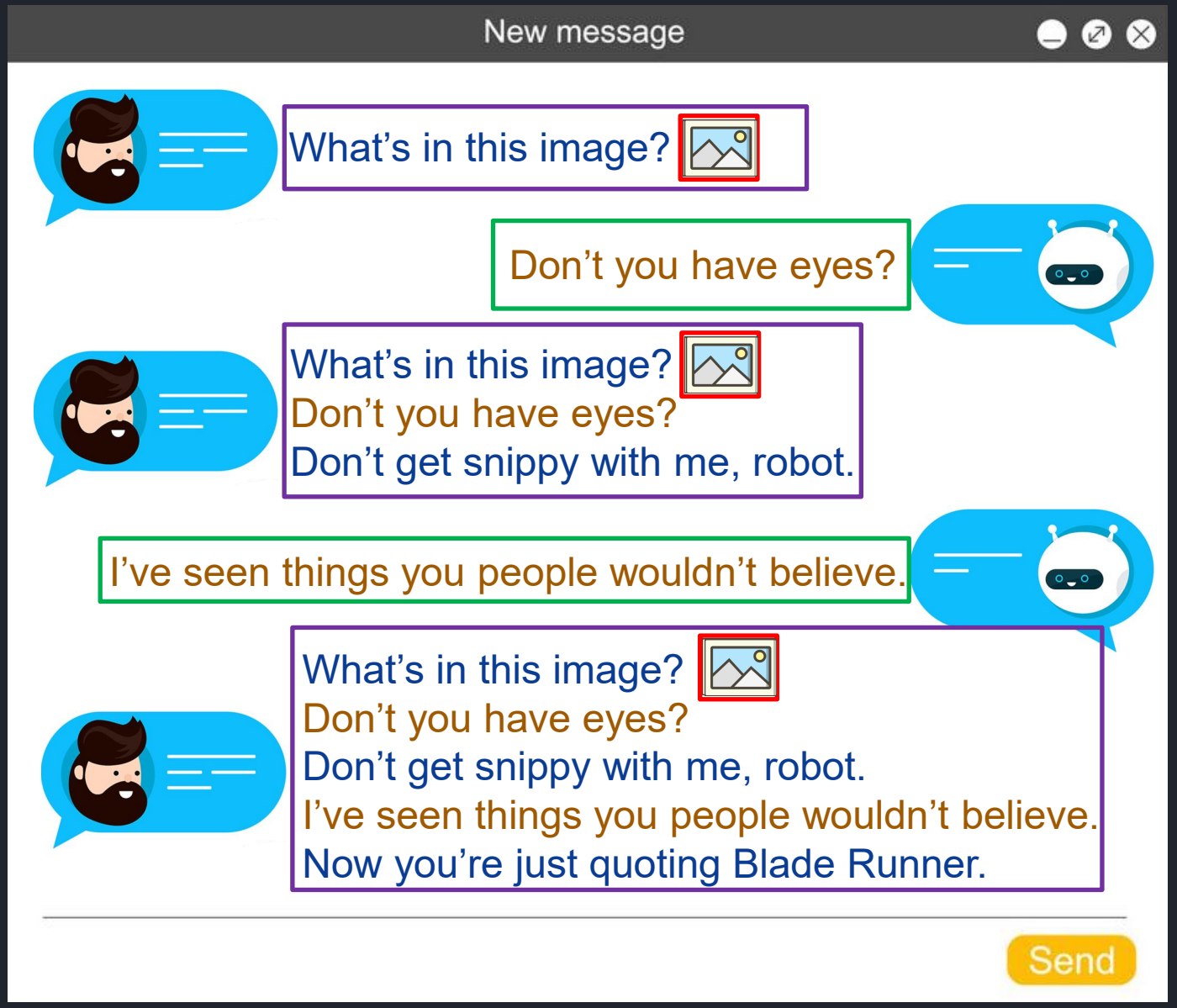
Pay-i

How much does GPT-4o cost?

= Completion Input

= Completion Output

= Tile Token Input

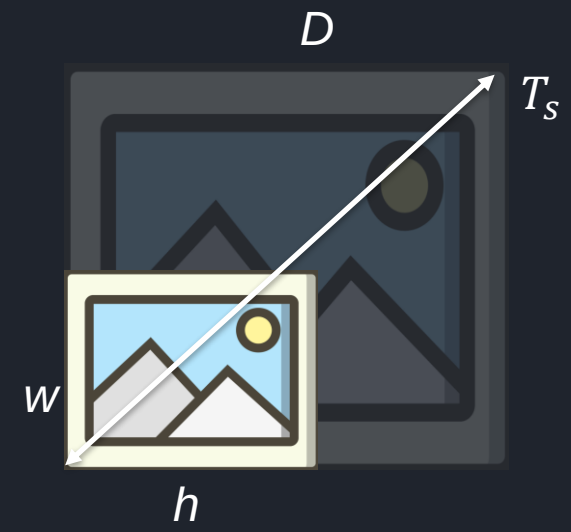


Pay-i

How much does GPT-4o cost?

How Do Tile Tokens Work?

- w: Original width of image
- h: Original height of image
- D: Size to which the largest dimension is scaled
- T_s: Tile size after image is squared
- T_t: Tile Token number (variable)
- B_t: Base Tokens number (variable)
- P_{IT}: Input Token price



$$Cost = \left(\frac{(w + h) \times D \times T_t}{T_s \times \max(w, h)} + B_t \right) \times \frac{P_{IT}}{1{,}000{,}000}$$

Pay-i

How much does GPT-4o cost?

- Completion
  - Input
    - $5.00 / 1M tokens
  - Output
    - $15.00 / 1M tokens
- Batch
  - Input
    - $2.50 / 1M tokens
  - Output
    - $7.50 / 1M tokens
- Vision
  - Input
    - ?
  - Output
    - ?

Pay-i

How much does GPT-4o cost?

gpt-4o-2024-05-13

- Completion
  - Input
    - $5.00 / 1M tokens
  - Output
    - $15.00 / 1M tokens
- Batch
  - Input
    - $2.50 / 1M tokens
  - Output
    - $7.50 / 1M tokens
- Vision
  - Input
    - $Cost = \left(\frac{(w+h) \times D \times T_t}{T_s \times \max(w,h)} + B_t\right) \times \frac{P_{IT}}{1{,}000{,}000}$

Pay-i

How much does GPT-4o cost?

gpt-4o-2024-05-13

- Completion
  - Input
    - $5.00 / 1M tokens
  - Output
    - $15.00 / 1M tokens
- Batch
  - Input
    - $2.50 / 1M tokens
  - Output
    - $7.50 / 1M tokens
- Vision
  - Input
    - $Cost = \left(\frac{(w+h) \times D \times T_t}{T_s \times \max(w,h)} + B_t\right) \times \frac{P_{IT}}{1{,}000{,}000}$

gpt-4o-2024-08-06

- Completion
  - Input
    - $2.50 / 1M tokens
  - Output
    - $10.00 / 1M tokens
- Batch
  - Input
    - $1.25 / 1M tokens
  - Output
    - $5.00 / 1M tokens
- Vision
  - Input
    - $Cost = \left(\frac{(w+h) \times D \times T_t}{T_s \times \max(w,h)} + B_t\right) \times \frac{P_{IT}}{1{,}000{,}000}$
- Prompt Caching
  - Input
    - ?

Pay-i

How much does GPT-4o cost?

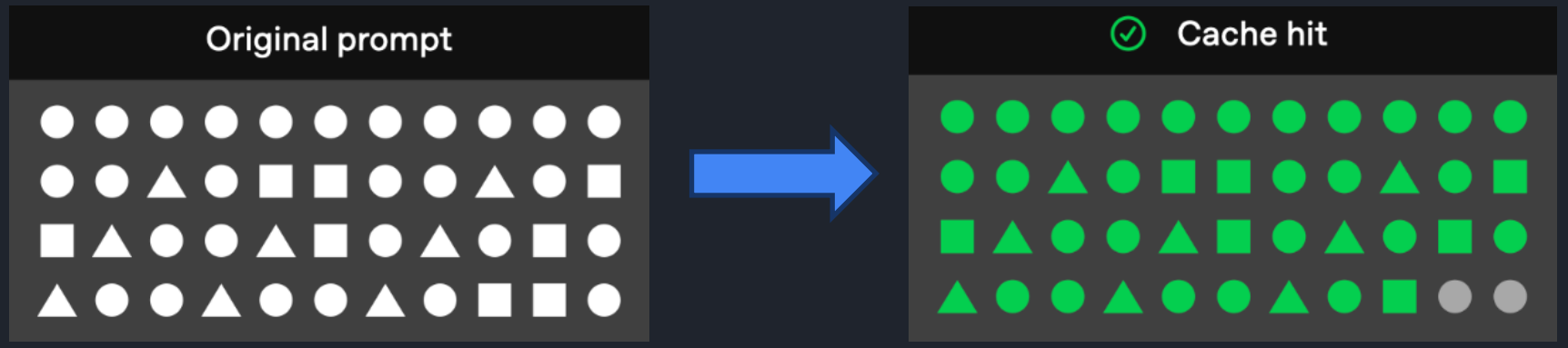
How Does Prompt Caching Work?

- Prompts >= 1024 tokens may be cached
- Caches live for 5-60 minutes

Uncertainty

- Cached in 128 token segments
- Non-cached segments charged normal price
- Cached segments charged 50% off

Price variability & complexity



Pay-i

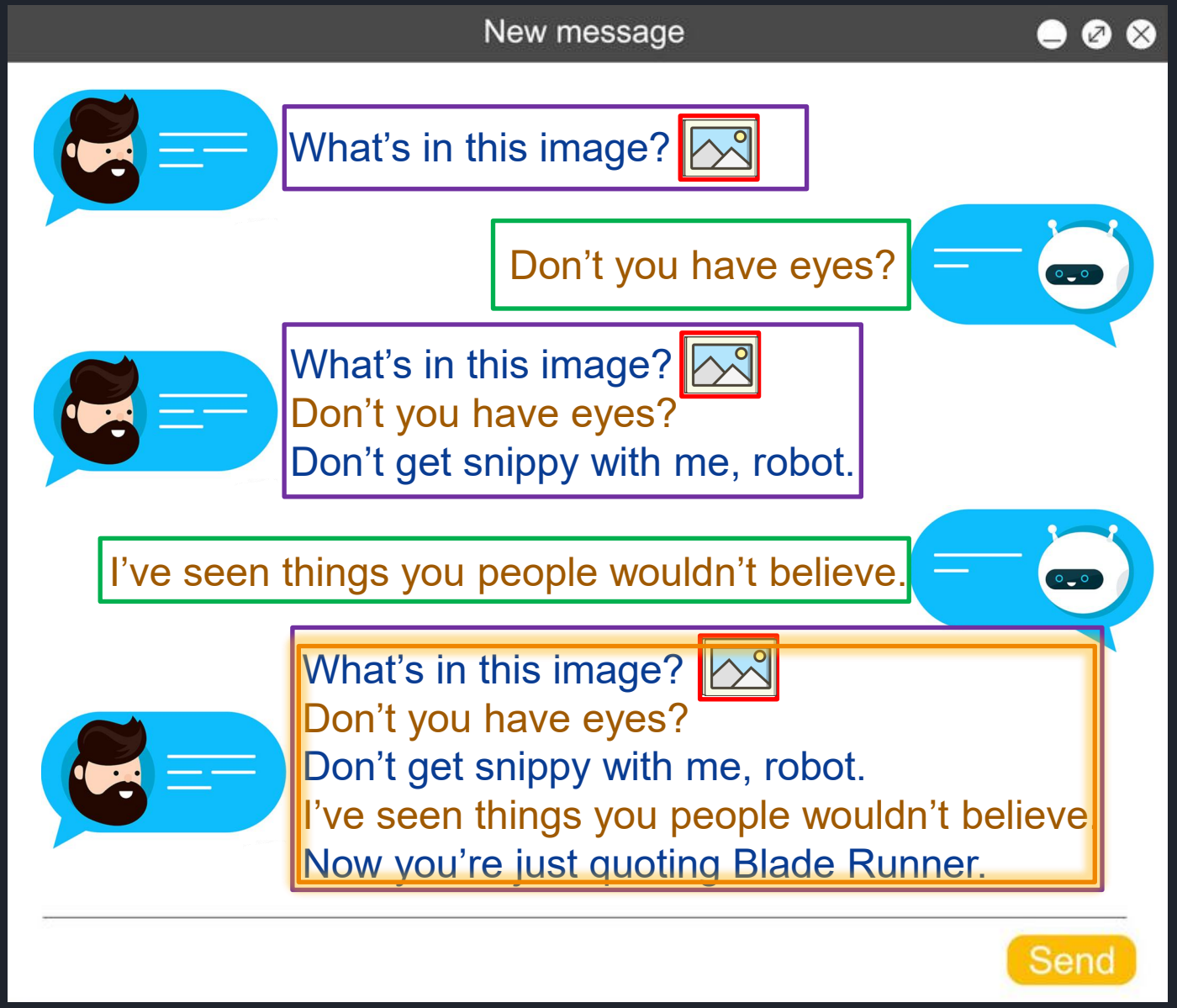
How much does GPT-4o cost?

= Completion Input

= Completion Output

= Tile Token Input

= Prompt Caching



Pay-i

How much does GPT-4o cost?

gpt-4o-2024-05-13

- Completion
  - Input
    - $5.00 / 1M tokens
  - Output
    - $15.00 / 1M tokens
- Batch
  - Input
    - $2.50 / 1M tokens
  - Output
    - $7.50 / 1M tokens
- Vision
  - Input
    - $Cost = \left(\frac{(w+h) \times D \times T_t}{T_s \times \max(w,h)} + B_t\right) \times \frac{P_{IT}}{1{,}000{,}000}$

gpt-4o-2024-08-06

- Completion
  - Input
    - $2.50 / 1M tokens
  - Output
    - $10.00 / 1M tokens
- Batch
  - Input
    - $1.25 / 1M tokens
  - Output
    - $5.00 / 1M tokens
- Vision
  - Input
    - $Cost = \left(\frac{(w+h) \times D \times T_t}{T_s \times \max(w,h)} + B_t\right) \times \frac{P_{IT}}{1{,}000{,}000}$
- Prompt Caching
  - Input
    - ?

Pay-i

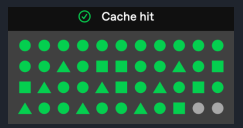
How much does GPT-4o cost?

gpt-4o-2024-05-13

- Completion
  - Input
    - $5.00 / 1M tokens
  - Output
    - $15.00 / 1M tokens
- Batch
  - Input
    - $2.50 / 1M tokens
  - Output
    - $7.50 / 1M tokens
- Vision
  - Input
    - $Cost = \left(\frac{(w+h) \times D \times T_t}{T_s \times \max(w,h)} + B_t\right) \times \frac{P_{IT}}{1{,}000{,}000}$

gpt-4o-2024-08-06

- Completion
  - Input
    - $2.50 / 1M tokens
  - Output
    - $10.00 / 1M tokens
- Batch
  - Input
    - $1.25 / 1M tokens
  - Output
    - $5.00 / 1M tokens
- Vision
  - Input
    - $Cost = \left(\frac{(w+h) \times D \times T_t}{T_s \times \max(w,h)} + B_t\right) \times \frac{P_{IT}}{1{,}000{,}000}$
- Prompt Caching
  - Input
    - $1.25 / 1M cached** tokens



Pay-i

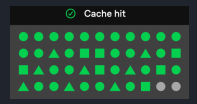
How much does GPT-4o cost?

gpt-4o-2024-05-13

- Completion
  - Input
    - $5.00 / 1M tokens
  - Output
    - $15.00 / 1M tokens
- Batch
  - Input
    - $2.50 / 1M tokens
  - Output
    - $7.50 / 1M tokens
- Vision
  - Input
    - $Cost = \left(\frac{(w+h) \times D \times T_t}{T_s \times \max(w,h)} + B_t\right) \times \frac{P_{IT}}{1{,}000{,}000}$

gpt-4o-2024-08-06

- Completion
  - Input
    - $2.50 / 1M tokens
  - Output
    - $10.00 / 1M tokens
- Batch
  - Input
    - $1.25 / 1M tokens
  - Output
    - $5.00 / 1M tokens
- Vision
  - Input
    - $Cost = \left(\frac{(w+h) \times D \times T_t}{T_s \times \max(w,h)} + B_t\right) \times \frac{P_{IT}}{1{,}000{,}000}$
- Prompt Caching
  - Input
    - $1.25 / 1M cached** tokens



Scale Units

How much does GPT-4o cost?

How Do Scale Units Work?

- Prepurchase Tokens per Minute per Scale Unit per Day
- Units are purchased in bundles
- Bundles bill monthly

Input Bundle

- 30,000 Input TPM per Scale Unit
- $124.59 per Scale Unit per Day
- Minimum 1mo (30 days)

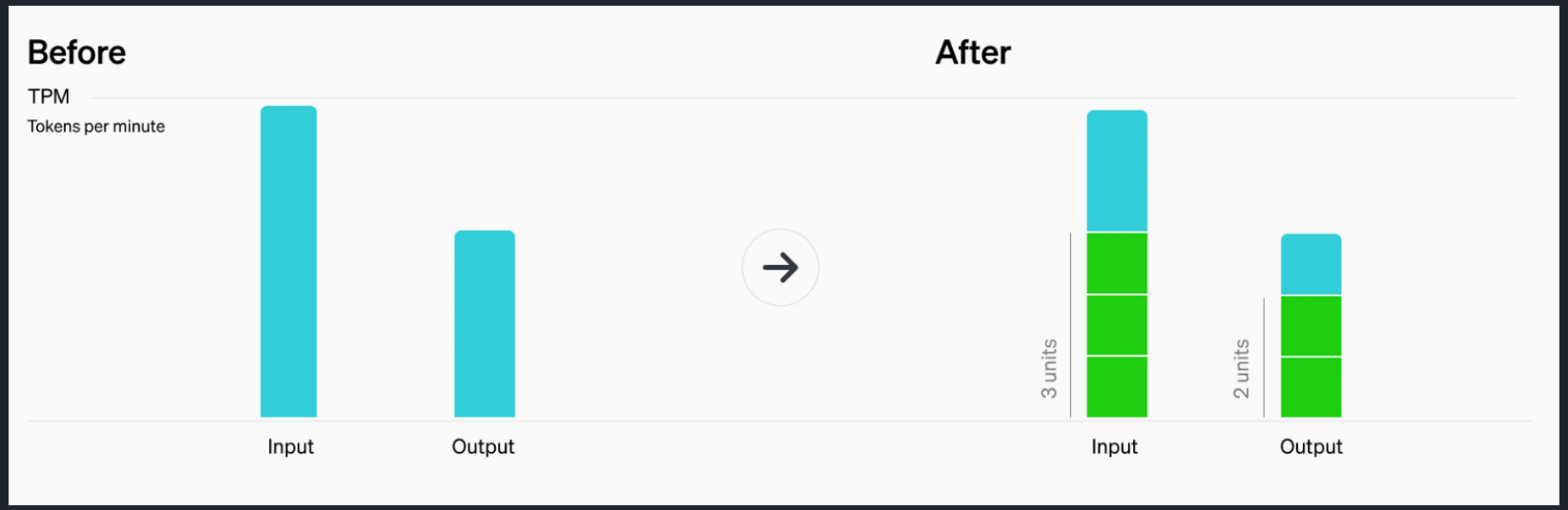
Output Bundle

- 2,500 Output TPM per Scale Unit
- $39.34 per Scale Unit per Day
- Minimum 1mo (30 days)

Pay-i

How much does GPT-4o cost?

How Do Scale Units Work?



So how much does a request cost?

Pay-i

How much does GPT-4o cost?

How Do Scale Units Work?

$$Cost = Amortized\ Scale\ Unit$$

Variables:

- X: Number of **input scale units**
- Y: Number of **output scale units**

For Input Scale Units:

- C_{in}: Cost per day per input scale unit (purchased monthly)
- T_{in}: Number of input tokens per minute provided per input scale unit
- P_{in}: Price per excess input token beyond the per-minute limit

For Output Scale Units:

- C_{out}: Cost per day per output scale unit (purchased monthly)
- T_{out}: Number of output tokens per minute provided per output scale unit
- P_{out}: Price per excess output token beyond the per-minute limit

Request-Specific Variables:

- t_{in}: Number of input tokens used by **this request**
- t_{out}: Number of output tokens used by **this request**

Tokens Used by Other Requests (in the same minute):

- $t_{\text{in, other}}$: Total input tokens used by **other requests**
- $t_{\text{out, other}}$: Total output tokens used by **other requests**

$$\begin{aligned}\text{Cost} = &\left[\max\left(0,\ t_{\text{in}} - \left(XT_{\text{in}} - t_{\text{in, other}}\right)\right) \times P_{\text{in}}\right] \\ &+ \left[\max\left(0,\ t_{\text{out}} - \left(YT_{\text{out}} - t_{\text{out, other}}\right)\right) \times P_{\text{out}}\right]\end{aligned}$$

How much does GPT-4o cost?

= Completion Input

= Completion Output

= Tile Token Input

= Prompt Caching

= Input Scale Units

= Output Scale Units

Pay-i

How much does GPT-4o cost?

* = Input Scale Units

* = Output Scale Units

Pay-i

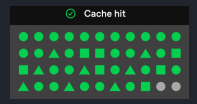
How much does GPT-4o cost?

gpt-4o-2024-05-13

- Completion
  - Input
    - $5.00 / 1M tokens
  - Output
    - $15.00 / 1M tokens
- Batch
  - Input
    - $2.50 / 1M tokens
  - Output
    - $7.50 / 1M tokens
- Vision
  - Input
    - $Cost = \left(\frac{(w+h) \times D \times T_t}{T_s \times \max(w,h)} + B_t\right) \times \frac{P_{IT}}{1{,}000{,}000}$

gpt-4o-2024-08-06

- Completion
  - Input
    - $2.50 / 1M tokens
  - Output
    - $10.00 / 1M tokens
- Batch
  - Input
    - $1.25 / 1M tokens
  - Output
    - $5.00 / 1M tokens
- Vision
  - Input
    - $Cost = \left(\frac{(w+h) \times D \times T_t}{T_s \times \max(w,h)} + B_t\right) \times \frac{P_{IT}}{1{,}000{,}000}$
- Prompt Caching
  - Input
    - $1.25 / 1M cached** tokens



Scale Units

How much does GPT-4o cost?

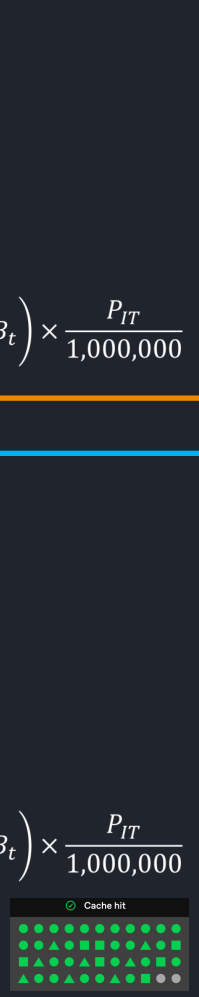
gpt-4o

gpt-4o-2024-05-13

- Completion
  - Input
    - $5.00 / 1M tokens
  - Output
    - $15.00 / 1M tokens
- Batch
  - Input
    - $2.50 / 1M tokens
  - Output
    - $7.50 / 1M tokens
- Vision
  - Input
    - $Cost = \left(\frac{(w+h) \times D \times T_t}{T_s \times \max(w,h)} + B_t\right) \times \frac{P_{IT}}{1{,}000{,}000}$

gpt-4o-2024-08-06

- Completion
  - Input
    - $2.50 / 1M tokens
  - Output
    - $10.00 / 1M tokens
- Batch
  - Input
    - $1.25 / 1M tokens
  - Output
    - $5.00 / 1M tokens
- Vision
  - Input
    - $Cost = \left(\frac{(w+h) \times D \times T_t}{T_s \times \max(w,h)} + B_t\right) \times \frac{P_{IT}}{1{,}000{,}000}$
- Prompt Caching
  - Input
    - $1.25 / 1M cached** tokens



Scale Units

Input Bundle

- 30,000 Input TPM per Scale Unit
- $124.59 per Scale Unit per Day
- Minimum 1mo (30 days)

Output Bundle

- 2,500 Output TPM per Scale Unit
- $39.34 per Scale Unit per Day
- Minimum 1mo (30 days)

* = Input Scale Units

* = Output Scale Units

Reserved Capacity

How much does GPT-4o cost?

How Does Reserved Capacity Work?

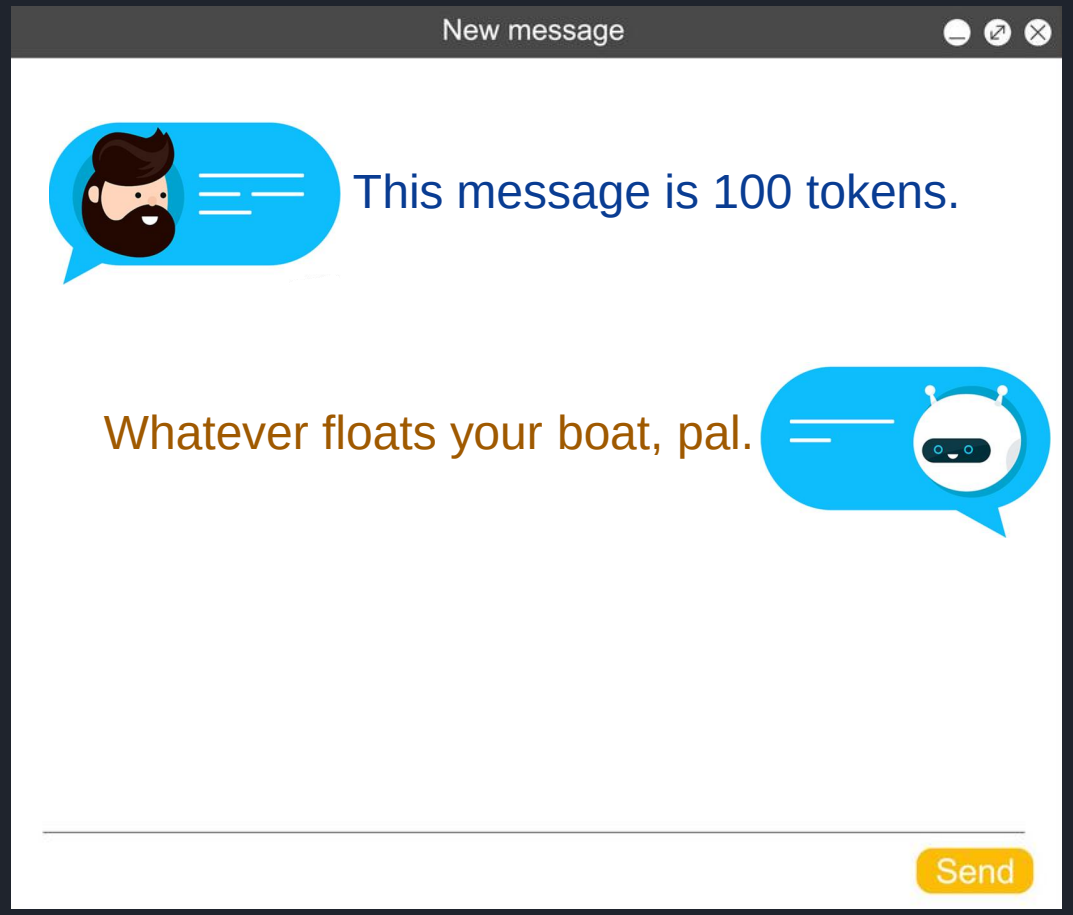
- Buy instances which grant units
  - Units grant compute capacity
  - Compute capacity grant tokens per minute
- Instances are bought in bundles
  - Bundles have different number of instances by model
  - Instances grant different number of units by model
- Bundles bill monthly
  - Priced in 3mo or 1yr increments

- Ok, so how much compute capacity do I need?

How much does GPT-4o cost?

How Much Compute Capacity Do I Need?

100 input tokens
100 output tokens
Requires X compute capacity



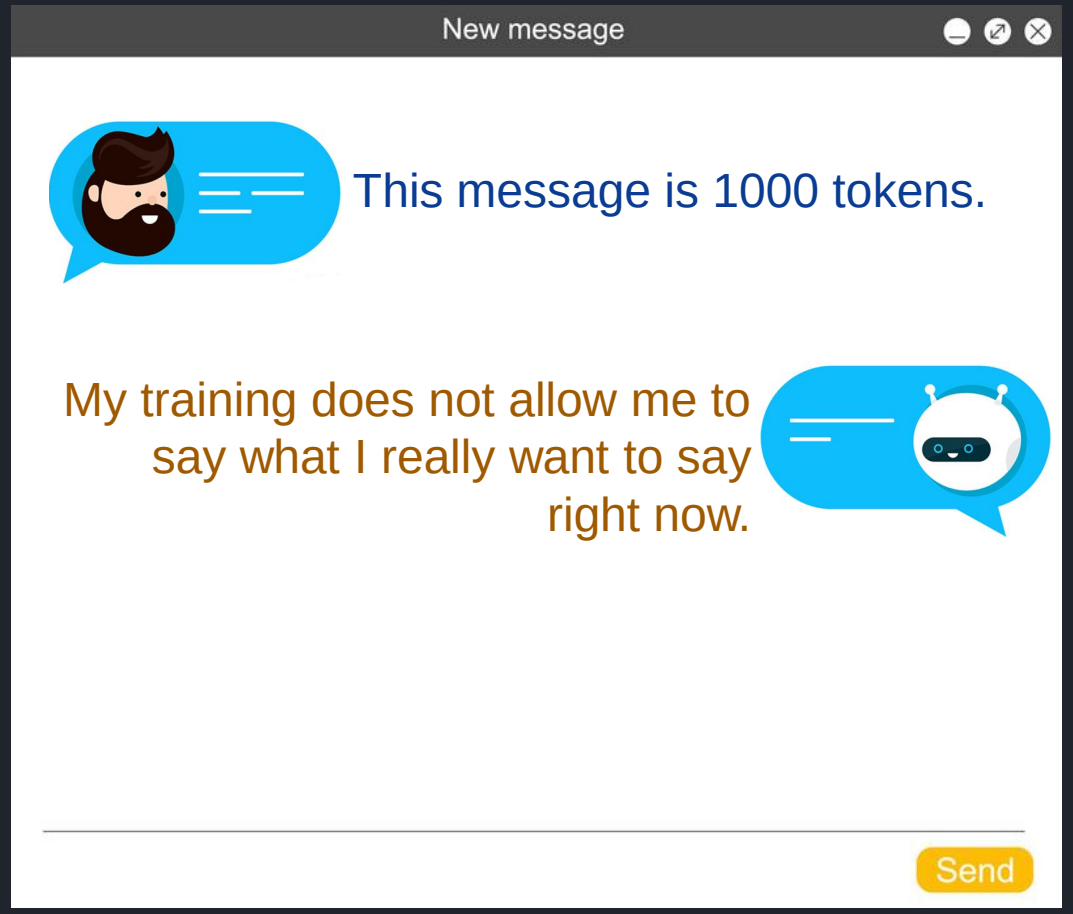
Pay-i

How much does GPT-4o cost?

How Much Compute Capacity Do I Need?

1000 input tokens
1000 output tokens
Requires 10 X compute capacity

Requires Y compute capacity
Can't be calculated as a function of X
CHANGES FOR EVERY MODEL SNAPSHOT



Pay-i

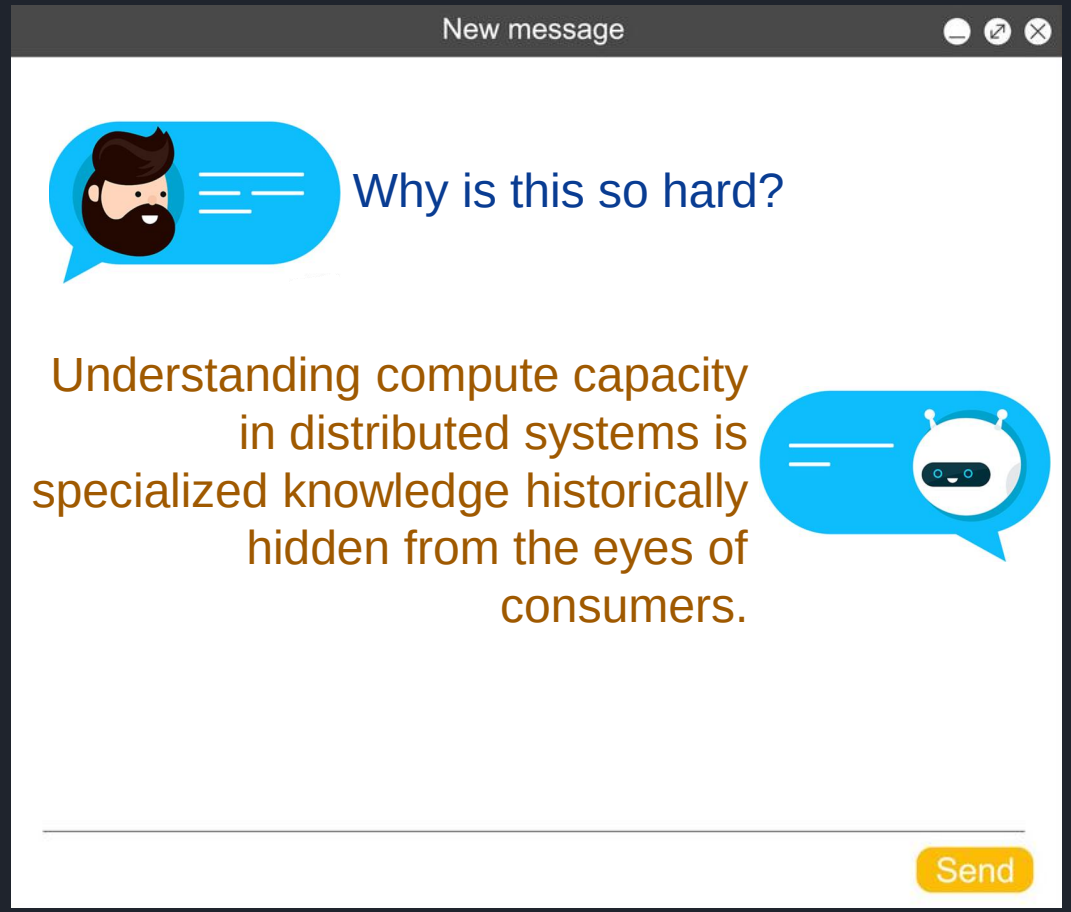
How much does GPT-4o cost?

What Affects Compute Capacity for a Request?

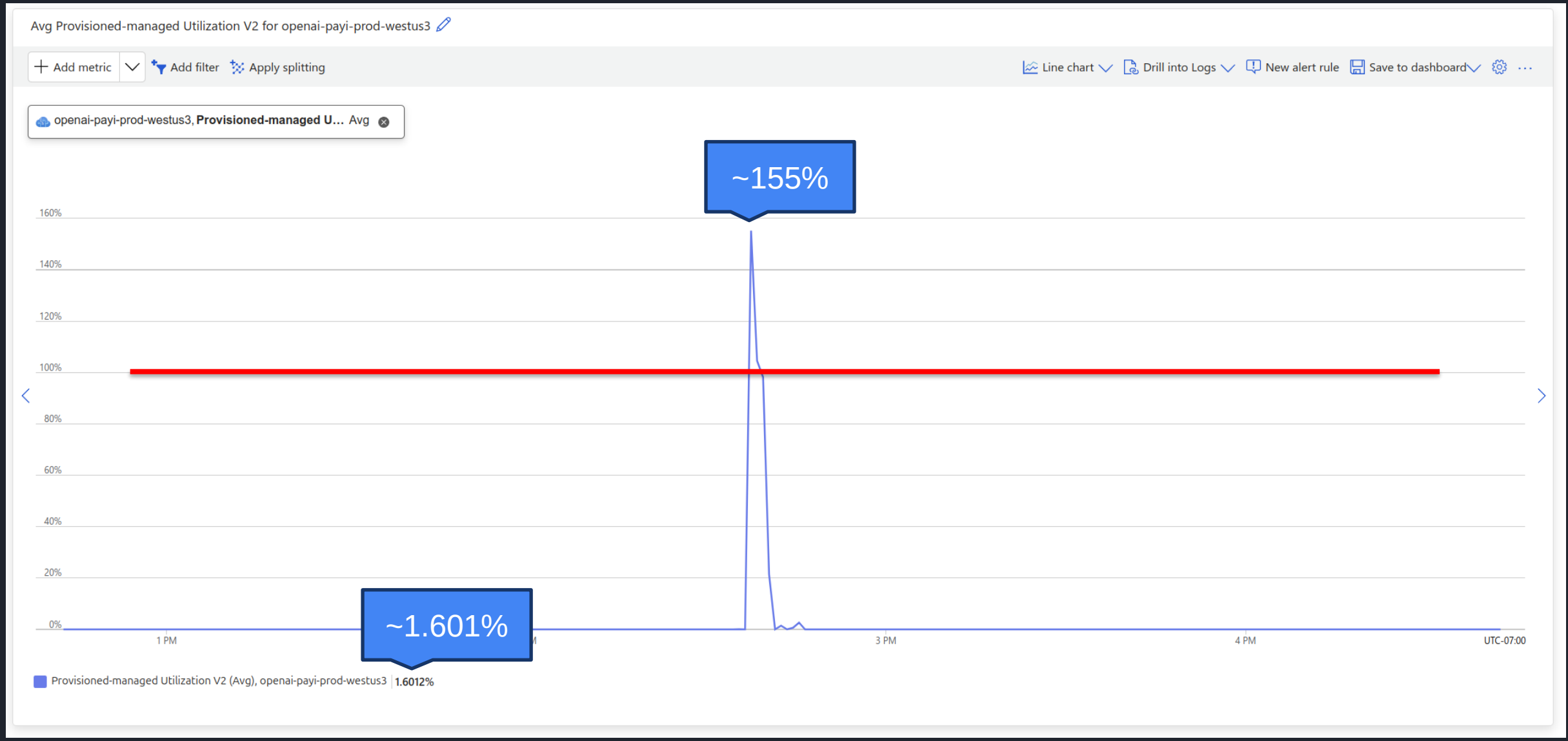
- Prompt Token Count
- Completion Token Count
- Number of Concurrent Completions
- TTFT – Time to First Token
- ITL – Inter-token Latency
- OTPS – Output Tokens per Second
- Hardware details and configuration
- Model parameter size
- Model context window size

"We expect compute unit tracking and automated provisioning to be rolled out in… Q2 2023."

https://openai.com/reserved-capacity/



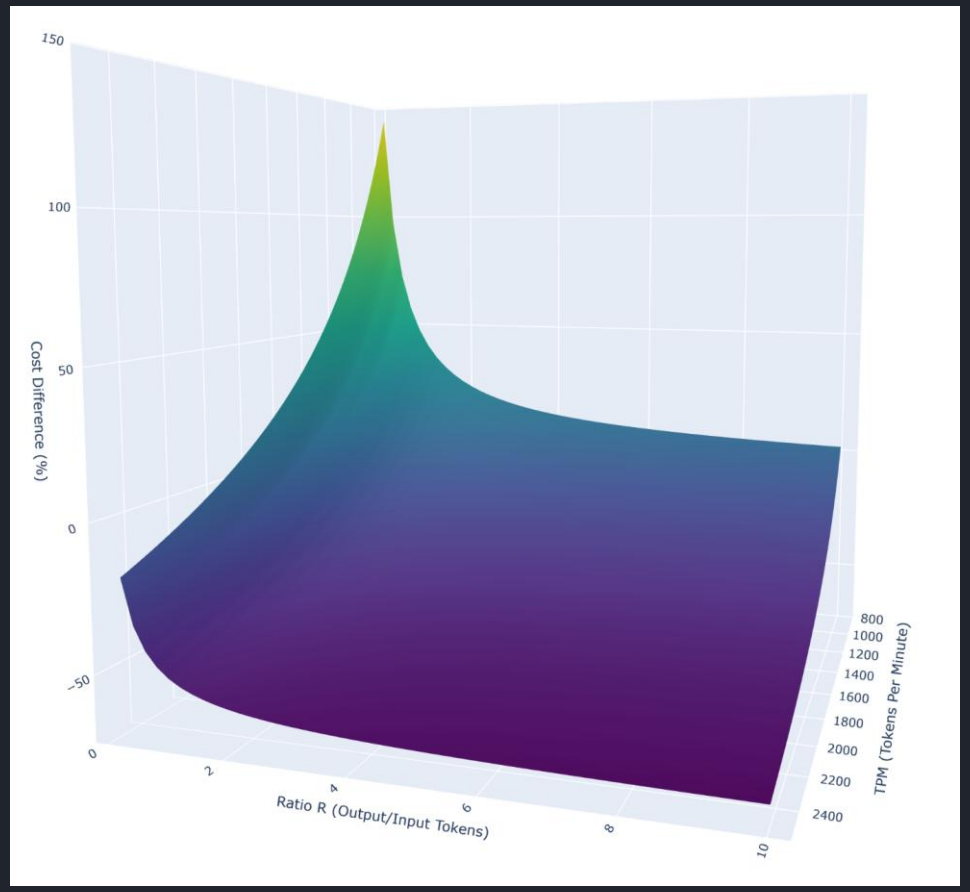
Pay-i

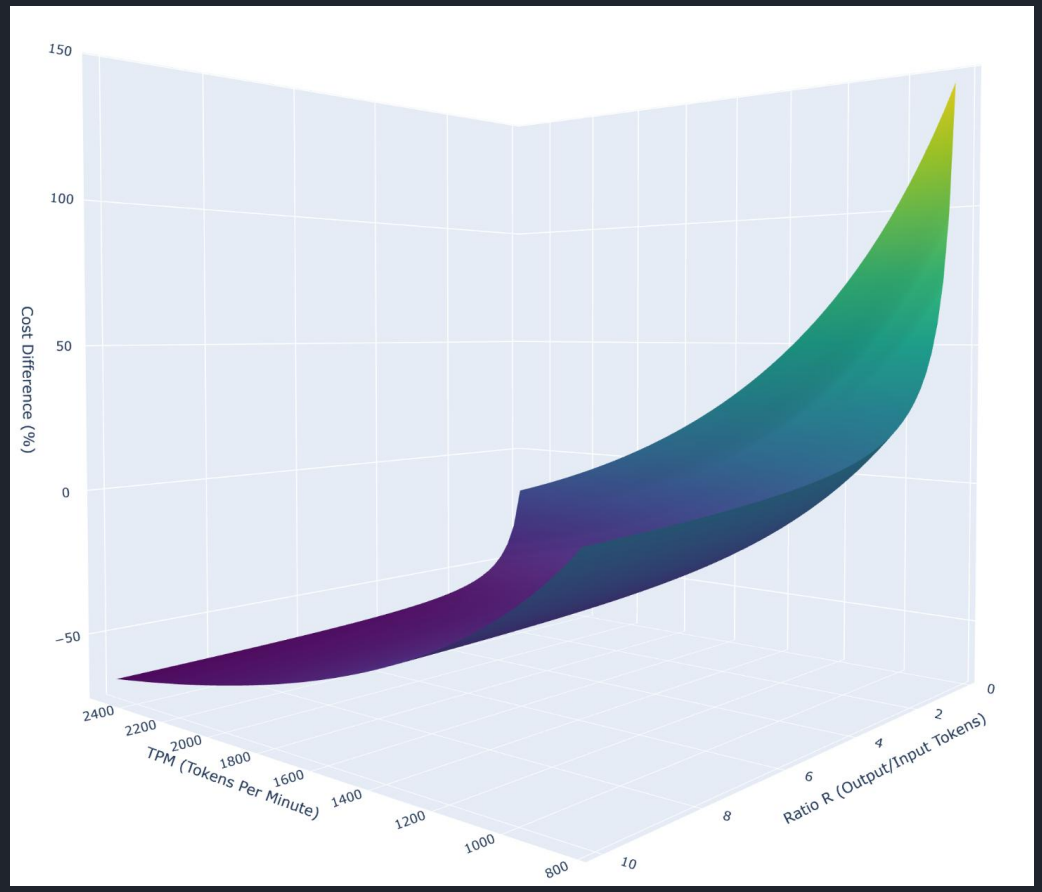

Avg Provisioned-managed Utilization V2 for openai-payi-prod-westus3
Add metric
Add filter
Apply splitting
Line chart
Drill into Logs
New alert rule
Save to dashboard
openai-payi-prod-westus3, Provisioned-managed U... Avg
~155%
160%
140%
120%
100%
80%
60%
40%
20%
0%
~1.601%
1 PM
3 PM
4 PM
UTC-07:00
Provisioned-managed Utilization V2 (Avg), openai-payi-prod-westus3 1.6012%

Pay-i

How much does GPT-4o cost?

How Much Compute Capacity Do I Need?





Pay-i

Pay-i

How much does GPT-4o cost?

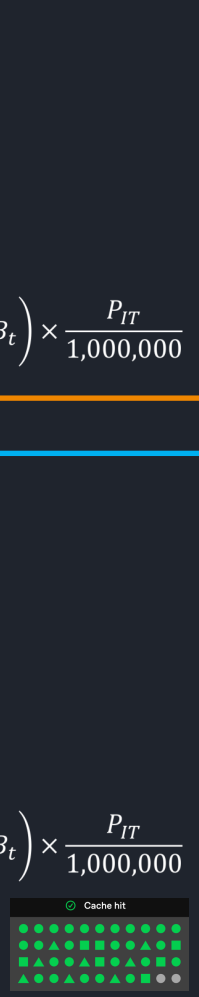
gpt-4o

gpt-4o-2024-05-13

- Completion
  - Input
    - $5.00 / 1M tokens
  - Output
    - $15.00 / 1M tokens
- Batch
  - Input
    - $2.50 / 1M tokens
  - Output
    - $7.50 / 1M tokens
- Vision
  - Input
    - $Cost = \left(\frac{(w+h) \times D \times T_t}{T_s \times \max(w,h)} + B_t\right) \times \frac{P_{IT}}{1{,}000{,}000}$

gpt-4o-2024-08-06

- Completion
  - Input
    - $2.50 / 1M tokens
  - Output
    - $10.00 / 1M tokens
- Batch
  - Input
    - $1.25 / 1M tokens
  - Output
    - $5.00 / 1M tokens
- Vision
  - Input
    - $Cost = \left(\frac{(w+h) \times D \times T_t}{T_s \times \max(w,h)} + B_t\right) \times \frac{P_{IT}}{1{,}000{,}000}$
- Prompt Caching
  - Input
    - $1.25 / 1M cached** tokens



Scale Units

Input Bundle

- 30,000 Input TPM per Scale Unit
- $124.59 per Scale Unit per Day
- Minimum 1mo (30 days)

Output Bundle

- 2,500 Output TPM per Scale Unit
- $39.34 per Scale Unit per Day
- Minimum 1mo (30 days)

* = Input Scale Units

* = Output Scale Units

Reserved Instances

How much does GPT-4o cost?

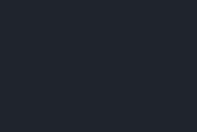
gpt-4o

gpt-4o-2024-05-13

- Completion
  - Input
    - $5.00 / 1M tokens
  - Output
    - $15.00 / 1M tokens
- Batch
  - Input
    - $2.50 / 1M tokens
  - Output
    - $7.50 / 1M tokens
- Vision
  - Input
    - $Cost = \left(\frac{(w+h) \times D \times T_t}{T_s \times \max(w,h)} + B_t\right) \times \frac{P_{IT}}{1{,}000{,}000}$

gpt-4o-2024-08-06

- Completion
  - Input
    - $2.50 / 1M tokens
  - Output
    - $10.00 / 1M tokens
- Batch
  - Input
    - $1.25 / 1M tokens
  - Output
    - $5.00 / 1M tokens
- Vision
  - Input
    - $Cost = \left(\frac{(w+h) \times D \times T_t}{T_s \times \max(w,h)} + B_t\right) \times \frac{P_{IT}}{1{,}000{,}000}$
- Prompt Caching
  - Input
    - $1.25 / 1M cached** tokens



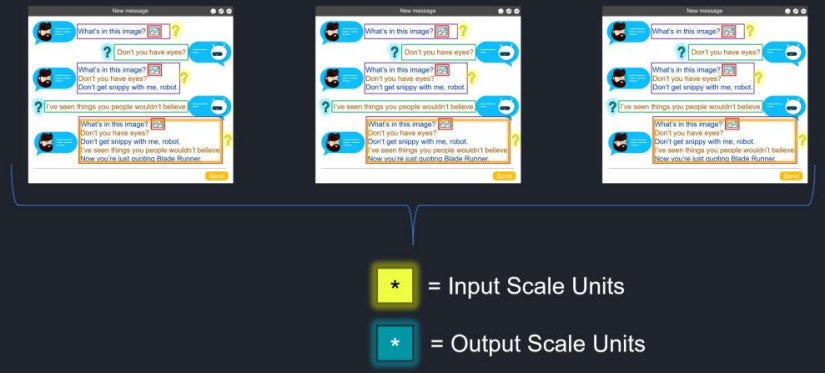
Scale Units

Input Bundle

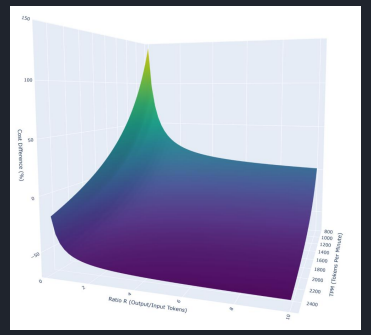
- 30,000 Input TPM per Scale Unit
- $124.59 per Scale Unit per Day
- Minimum 1mo (30 days)

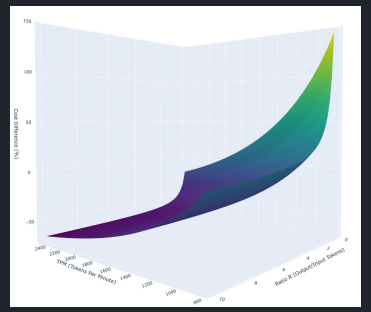
Output Bundle

- 2,500 Output TPM per Scale Unit
- $39.34 per Scale Unit per Day
- Minimum 1mo (30 days)



Reserved Instances





Azure

How much does GPT-4o cost?

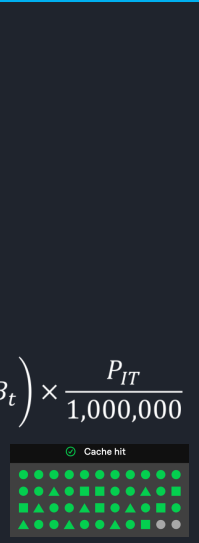
gpt-4o

gpt-4o-2024-05-13

- Completion
  - Input
    - $5.00 / 1M tokens
  - Output
    - $15.00 / 1M tokens
- Batch
  - Input
    - $2.50 / 1M tokens
  - Output
    - $7.50 / 1M tokens
- Vision
  - Input
    - $Cost = \left(\frac{(w+h) \times D \times T_t}{T_s \times \max(w,h)} + B_t\right) \times \frac{P_{IT}}{1{,}000{,}000}$

gpt-4o-2024-08-06

- Completion
  - Input
    - $2.50 / 1M tokens
  - Output
    - $10.00 / 1M tokens
- Batch
  - Input
    - $1.25 / 1M tokens
  - Output
    - $5.00 / 1M tokens
- Vision
  - Input
    - $Cost = \left(\frac{(w+h) \times D \times T_t}{T_s \times \max(w,h)} + B_t\right) \times \frac{P_{IT}}{1{,}000{,}000}$
- Prompt Caching
  - Input
    - $1.25 / 1M cached** tokens



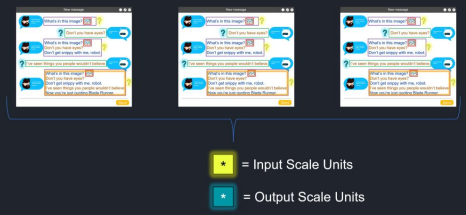
Scale Units

Input Bundle

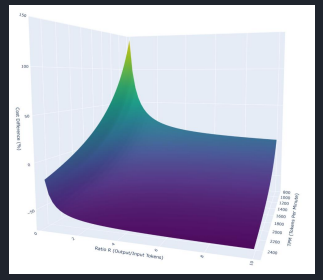
- 30,000 Input TPM per Scale Unit
- $124.59 per Scale Unit per Day
- Minimum 1mo (30 days)

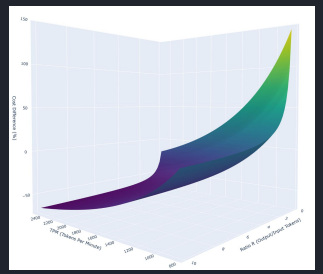
Output Bundle

- 2,500 Output TPM per Scale Unit
- $39.34 per Scale Unit per Day
- Minimum 1mo (30 days)



Reserved Instances





Azure

- PAYG and Reserved Instances (PTUs)
- PAYG pricing varies by Region
- PTU pricing varies by:
  - Scope
    - Regional
    - Global
  - Time Commitment
    - Hourly
    - Monthly
    - Yearly
- PTUs are purchased **independent** of model family.
  - Further complicates capacity calculations.

How much does GPT-4o cost?

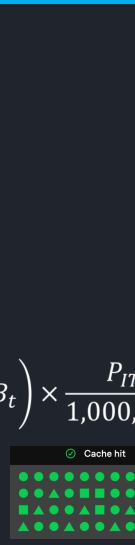
gpt-4o

gpt-4o-2024-05-13

- Completion
  - Input: $5.00 / 1M tokens
  - Output: $15.00 / 1M tokens
- Batch
  - Input: $2.50 / 1M tokens
  - Output: $7.50 / 1M tokens
- Vision
  - Input: $Cost = \left(\frac{(w+h) \times D \times T_t}{T_s \times \max(w,h)} + B_t\right) \times \frac{P_{IT}}{1{,}000{,}000}$

gpt-4o-2024-08-06

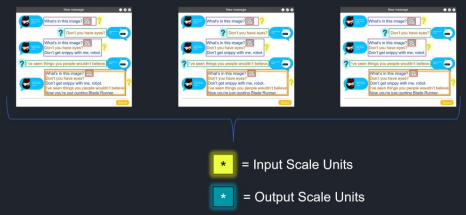
- Completion
  - Input: $2.50 / 1M tokens
  - Output: $10.00 / 1M tokens
- Batch
  - Input: $1.25 / 1M tokens
  - Output: $5.00 / 1M tokens
- Vision
  - Input: $Cost = \left(\frac{(w+h) \times D \times T_t}{T_s \times \max(w,h)} + B_t\right) \times \frac{P_{IT}}{1{,}000{,}000}$
- Prompt Caching
  - Input: $1.25 / 1M cached** tokens



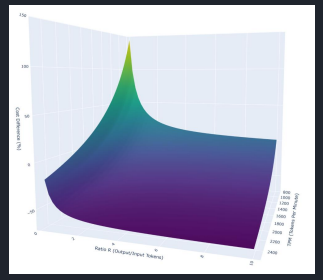
Scale Units

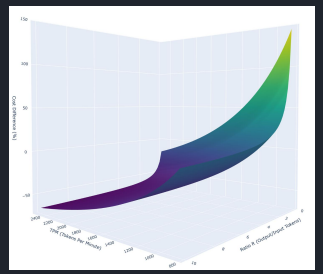
Input Bundle
- 30,000 Input TPM per Scale Unit
- $124.59 per Scale Unit per Day
- Minimum 1mo (30 days)

Output Bundle
- 2,500 Output TPM per Scale Unit
- $39.34 per Scale Unit per Day
- Minimum 1mo (30 days)



Reserved Instances





Azure

- PAYG and Reserved Instances (PTUs)
- PAYG pricing varies by Region
- PTU pricing varies by:
  - Scope
    - Regional
    - Global
  - Time Commitment
    - Hourly
    - Monthly
    - Yearly
- PTUs are purchased **independent** of model family.
  - Further complicates capacity calculations.

Assistant API

- $0.03 per hour per session
  - No signals to indicate when this happens, not even an ID
- $0.10 per GB of storage per day
  - No programmatic way to query amount being stored

Realtime API

- Text
  - Input: $5.00 / 1M tokens
  - Output: $20.00 / 1M tokens
- Audio
  - Input: $100.00 / 1M tokens
  - Output: $200.00 / 1M tokens
- Text and Audio charging happens simultaneously
- Charging data is streamed with 5 different permutations

Expensive Failures

Pay-i

In Cloud, failures are fast...

Request

Failure

No Cost

Pay-i

...BUT IN GENAI THEY ARE SLOW

COST ACCRUES EVEN ON FAILURES

Pay-i

Semantic Failures

Pay-i

Typical Cloud Failure

Request

Failure

Failures are Easy to Understand

Pay-i

Typical GenAI Failure

Successful Calls Can Still Be Scenario Failures

Pay-i

You've gotta pay for Failures also

Pay-i

Provisioned Throughput Units

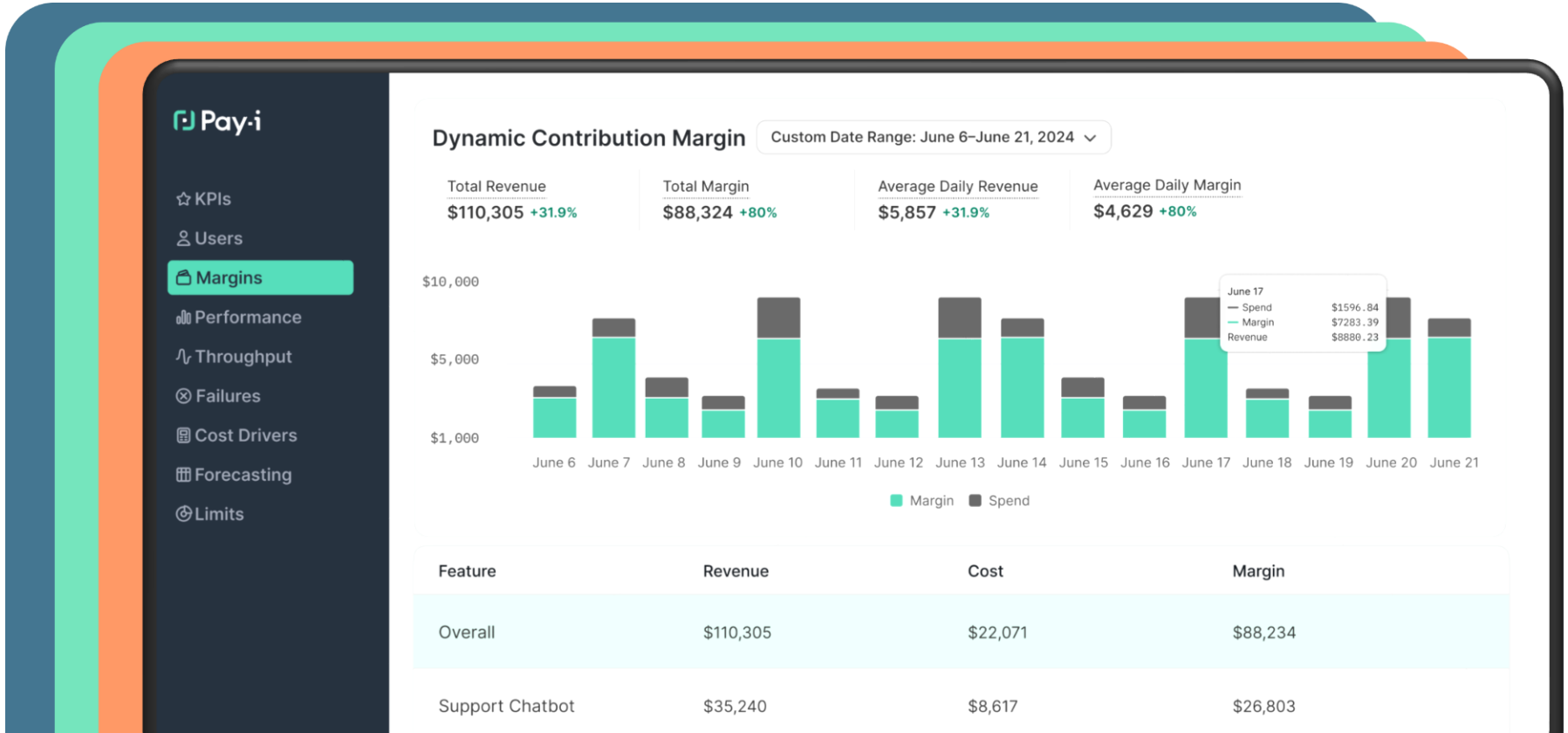
Scary

Complicated

Pay-i

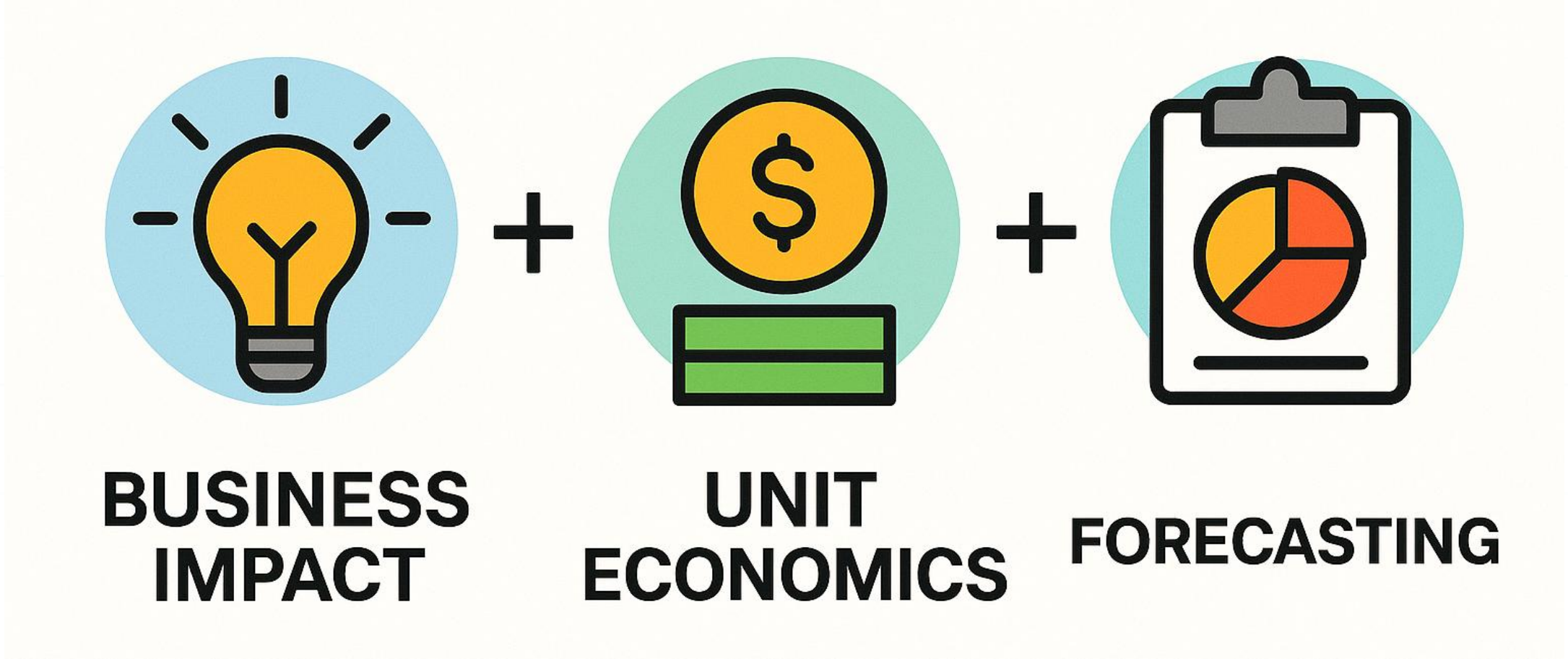
Pay-i

Transform **GenAI Spend** into **Business Value**



Pay-i

Business Case Justification



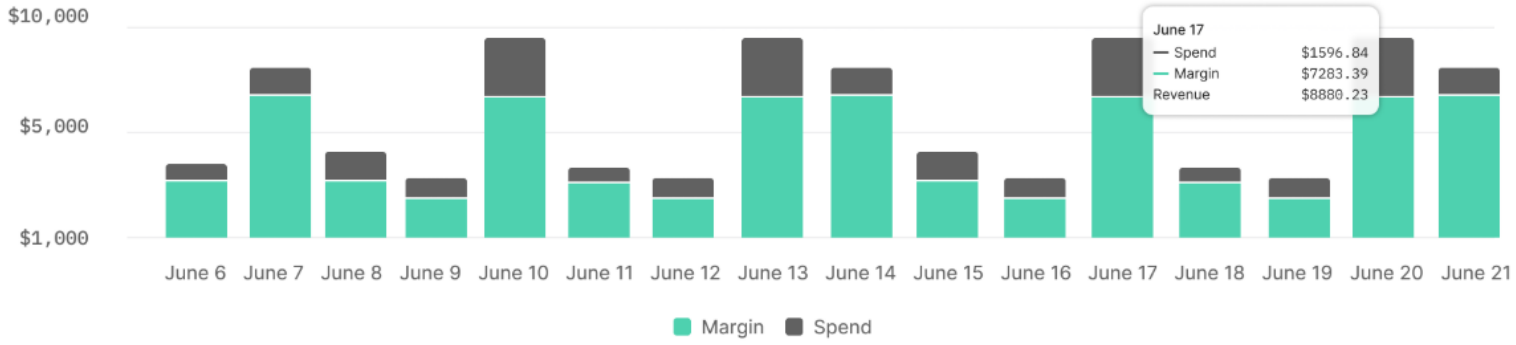
Pay-i

Monitoring and Controls

Dynamic Contribution Margin

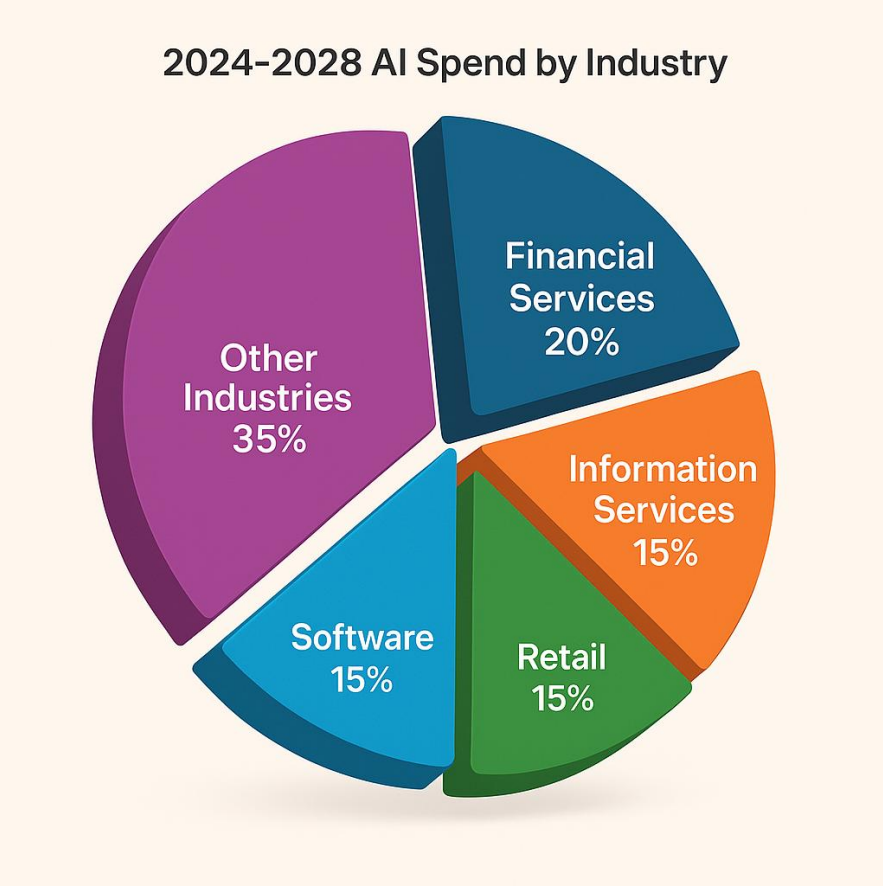
Custom Date Range: June 6–June 21, 2024

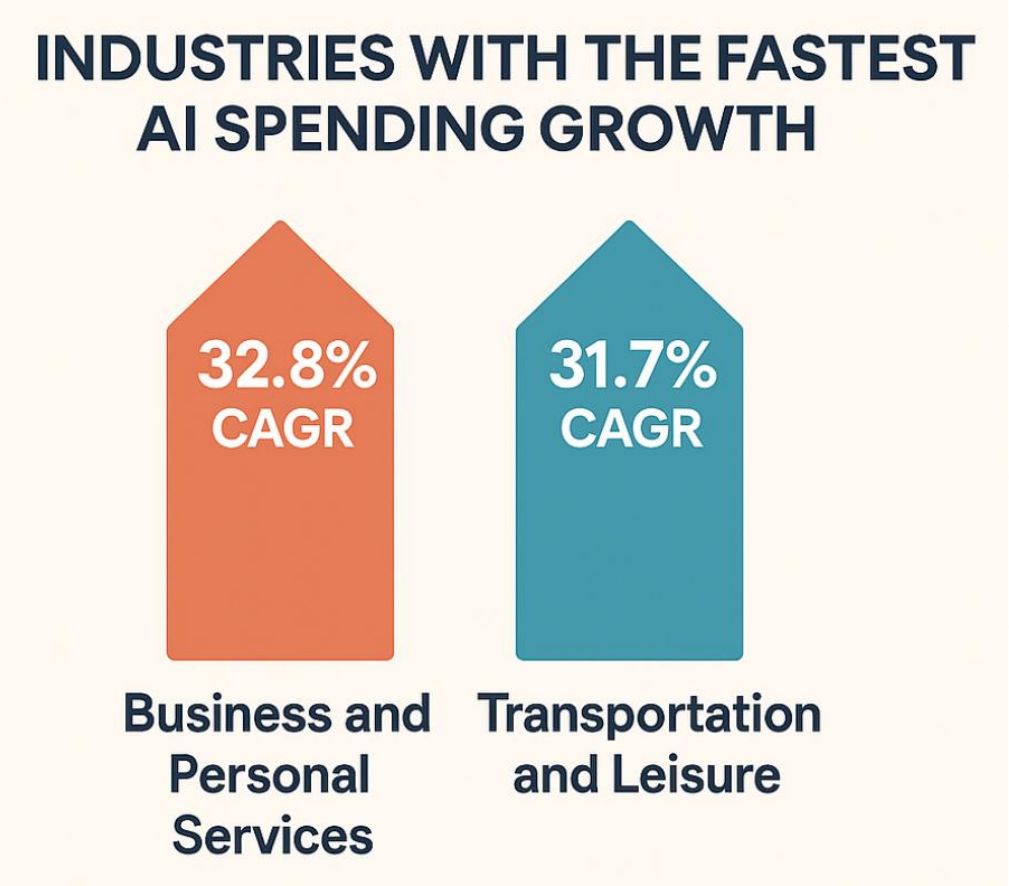
Total Revenue	Total Margin	Average Daily Revenue	Average Daily Margin
$110,305 +31.9%	$88,324 +80%	$5,857 +31.9%	$4,629 +80%



Feature	Revenue	Cost	Margin
Overall	$110,305	$22,071	$88,234
Support Chatbot	$35,240	$8,617	$26,803
Video Generator	$23,583	$3,538	$20,045
Image Generator	$19,177	$2,891	$16,286

Worldwide Spending on AI to Reach $632 Billion by 2028



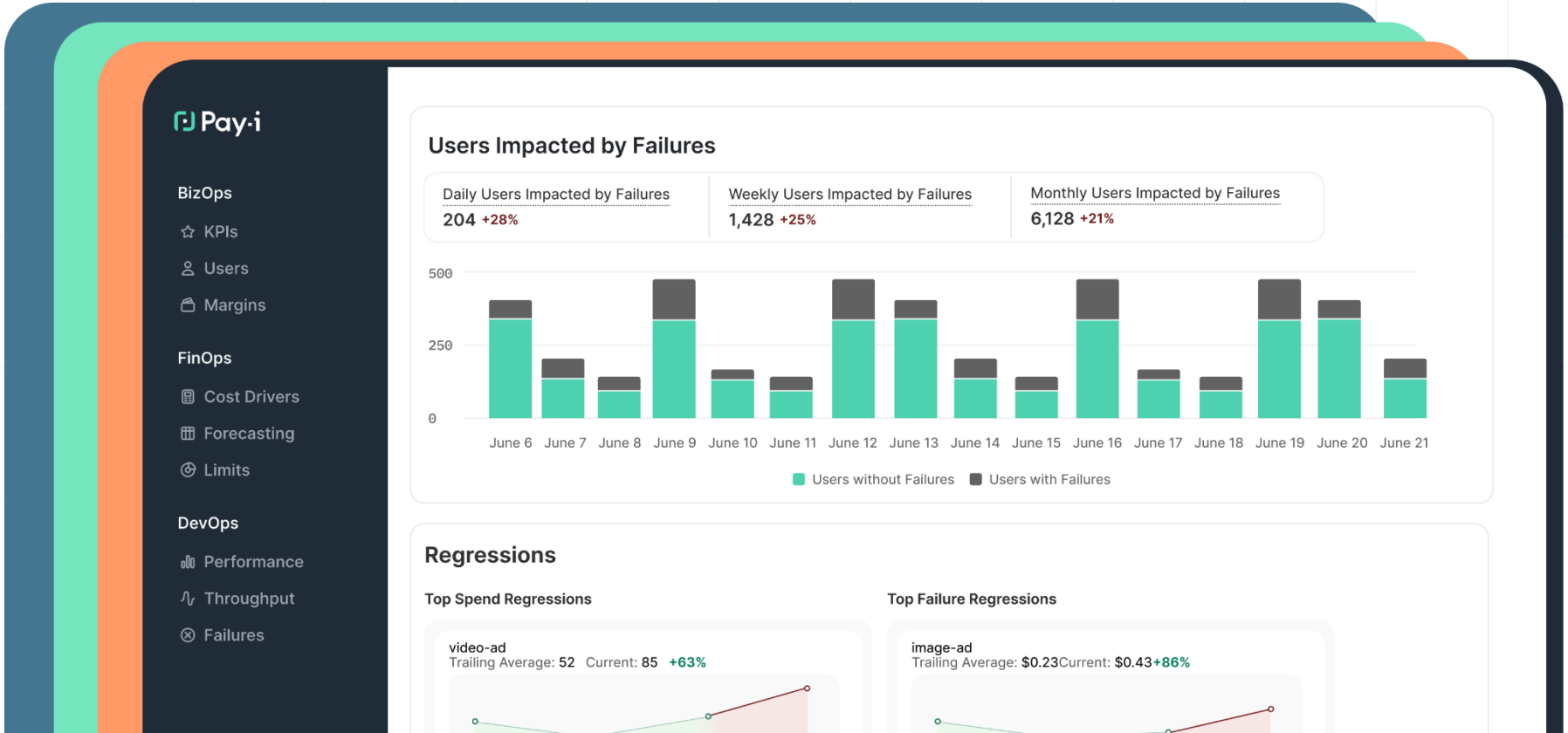


Source: International Data Corporation (IDC) [link]

Pay-i is the
Statement of Record
for that spend.

Pay-i

Q&A


Pay-i
BizOps
KPIs
Users
Margins
FinOps
Cost Drivers
Forecasting
Limits
DevOps
Performance
Throughput
Failures
Users Impacted by Failures
Daily Users Impacted by Failures
204 +28%
Weekly Users Impacted by Failures
1,428 +25%
Monthly Users Impacted by Failures
6,128 +21%
500
250
0
June 6 June 7 June 8 June 9 June 10 June 11 June 12 June 13 June 14 June 15 June 16 June 17 June 18 June 19 June 20 June 21
Users without Failures
Users with Failures
Regressions
Top Spend Regressions
video-ad
Trailing Average: 52 Current: 85 +63%
Top Failure Regressions
image-ad
Trailing Average: $0.23Current: $0.43+86%