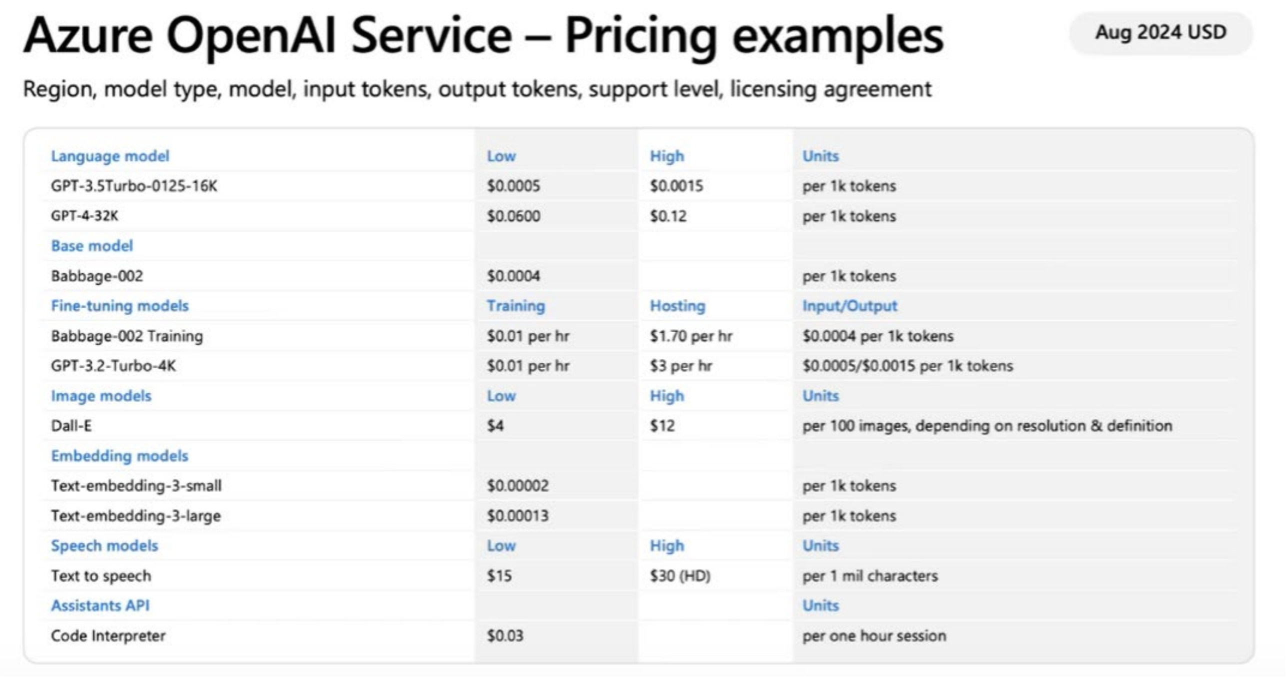
Series of videos (notes from costs optimization)

aka.ms/GenAIforDevelopers



Azure OpenAI Service – Pricing examples

Aug 2024 USD

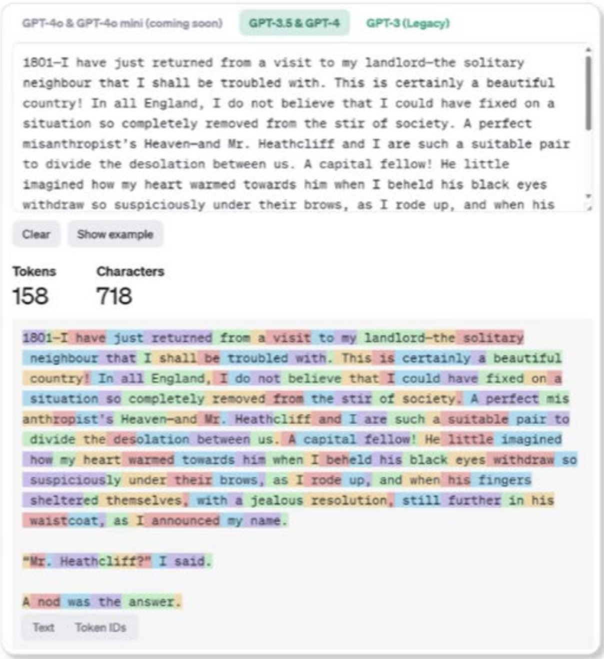
Region, model type, model, input tokens, output tokens, support level, licensing agreement

Language model	Low	High	Units
GPT-3.5Turbo-0125-16K	$0.0005	$0.0015	per 1k tokens
GPT-4-32K	$0.0600	$0.12	per 1k tokens
Base model			
Babbage-002	$0.0004		per 1k tokens
Fine-tuning models	Training	Hosting	Input/Output
Babbage-002 Training	$0.01 per hr	$1.70 per hr	$0.0004 per 1k tokens
GPT-3.2-Turbo-4K	$0.01 per hr	$3 per hr	$0.0005/$0.0015 per 1k tokens
Image models	Low	High	Units
Dall-E	$4	$12	per 100 images, depending on resolution & definition
Embedding models			
Text-embedding-3-small	$0.00002		per 1k tokens
Text-embedding-3-large	$0.00013		per 1k tokens
Speech models	Low	High	Units
Text to speech	$15	$30 (HD)	per 1 mil characters
Assistants API			Units
Code Interpreter	$0.03		per one hour session

Tokens

One token generally corresponds to ~4 characters of text for common English text.

This translates to roughly ¾ of a word (so 100 tokens ~= 75 words).



```
1801—I have just returned from a visit to my landlord—the solitary
neighbour that I shall be troubled with. This is certainly a beautiful
country! In all England, I do not believe that I could have fixed on a
situation so completely removed from the stir of society. A perfect
misanthropist's Heaven—and Mr. Heathcliff and I are such a suitable pair
to divide the desolation between us. A capital fellow! He little
imagined how my heart warmed towards him when I beheld his black eyes
withdraw so suspiciously under their brows, as I rode up, and when his
```

Tokens	Characters
158	718

```
1801—I have just returned from a visit to my landlord—the solitary
 neighbour that I shall be troubled with. This is certainly a beautiful
 country! In all England, I do not believe that I could have fixed on a
 situation so completely removed from the stir of society. A perfect mis
anthropist's Heaven—and Mr. Heathcliff and I are such a suitable pair to
 divide the desolation between us. A capital fellow! He little imagined
 how my heart warmed towards him when I beheld his black eyes withdraw so
 suspiciously under their brows, as I rode up, and when his fingers
 sheltered themselves, with a jealous resolution, still further in his
 waistcoat, as I announced my name.

"Mr. Heathcliff?" I said.

A nod was the answer.
```

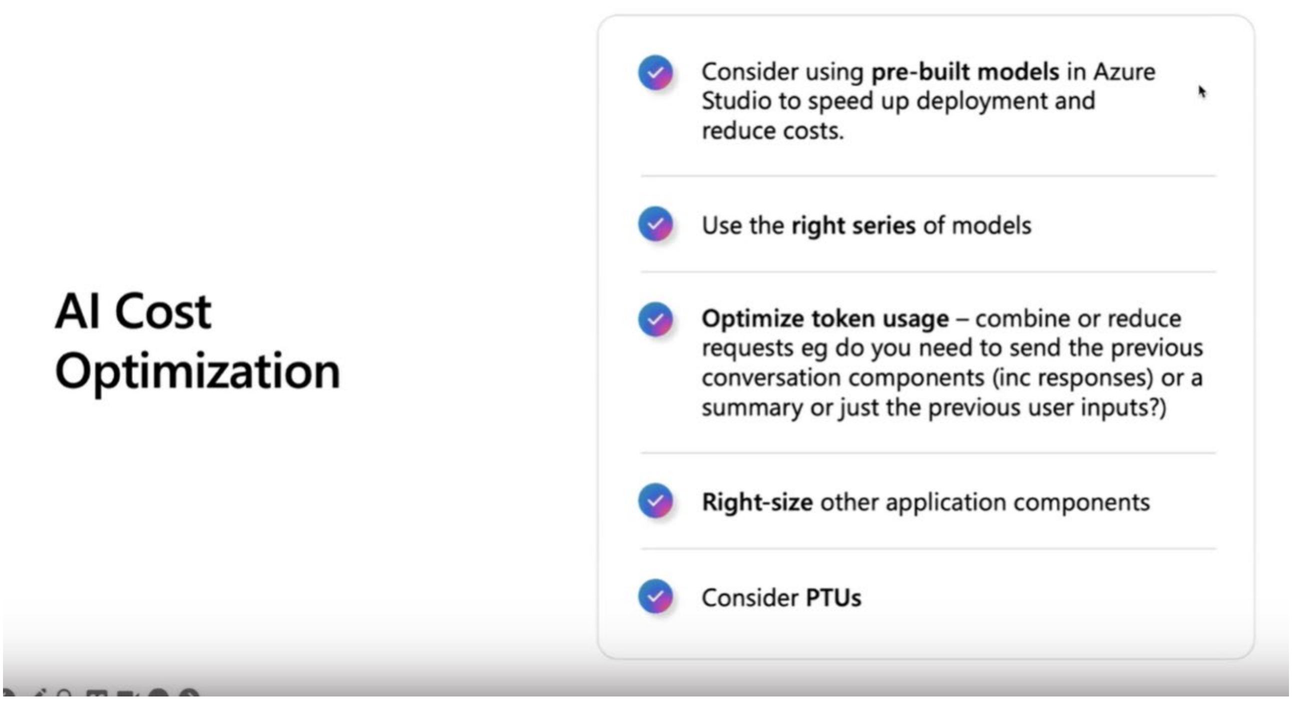
Azure AI Search – Pricing examples

Aug 2024 USD

Region, model type, model, input tokens, output tokens, support level, licensing agreement

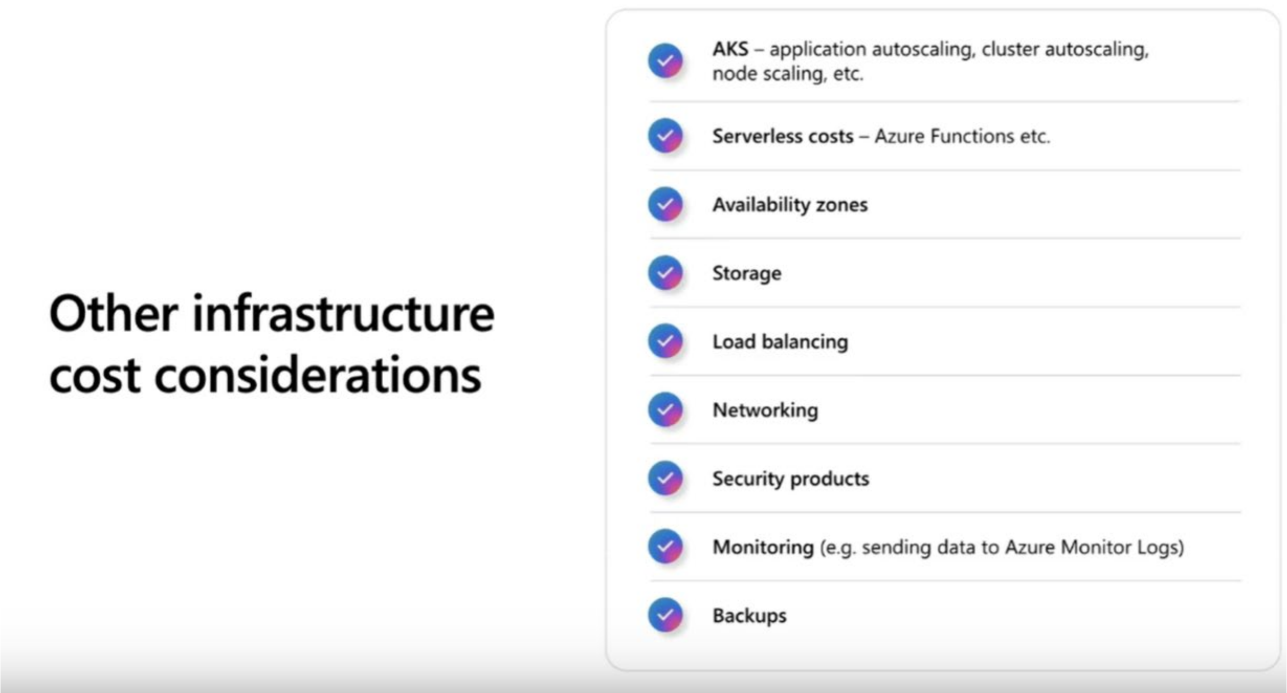
	Free	Basic	Standard S1	Standard S2	Standard S3	Storage Optimized L1	Storage Optimized L2
Storage[1]	50 MB	15 GB (max 45 GB per service)	160 GB (max 1.9 TB per service)	512 GB (max 6 TB per service)	1 TB (max 12 TB per service)	2 TB (max 24 TB per service)	4 TB (max 48 TB per service)
Max indexes per service	3	15	50	200	200 or 1,000/partition in high density mode	10	10
Scale out limits	N/A	Up to 9 units per service[2] (max 3 partition, max 3 replicas)	Up to 36 units per service (max 12 partition; max 12 replicas)	Up to 36 units per service (max 12 partition; max 12 replicas)	Up to 36 units per service (max 12 partition; max 12 replicas) up to 3 partitions in high density[3] mode	Up to 36 units per service (max 12 partition; max 12 replicas)	Up to 36 units per service (max 12 partition; max 12 replicas)
Price per SU (Scale Unit)	$0/hour	$0.11/hour	$0.34/hour	$1.35/hour	$2.69/hour	$3.84/hour	$7.68/hour

Additional pricing for Custom entity lookup skill, Document cracking (image extraction), Semantic ranker



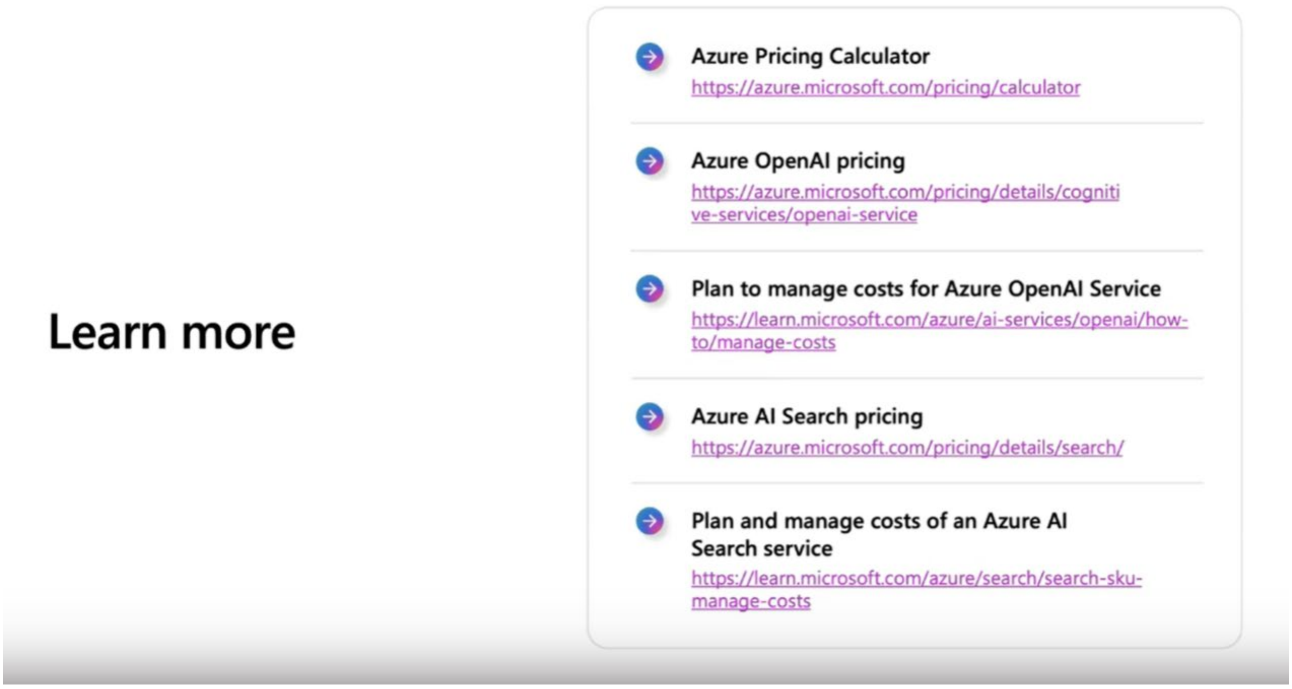
AI Cost Optimization

- ✓ Consider using **pre-built models** in Azure Studio to speed up deployment and reduce costs.
- ✓ Use the **right series** of models
- ✓ **Optimize token usage** – combine or reduce requests eg do you need to send the previous conversation components (inc responses) or a summary or just the previous user inputs?)
- ✓ **Right-size** other application components
- ✓ Consider **PTUs**



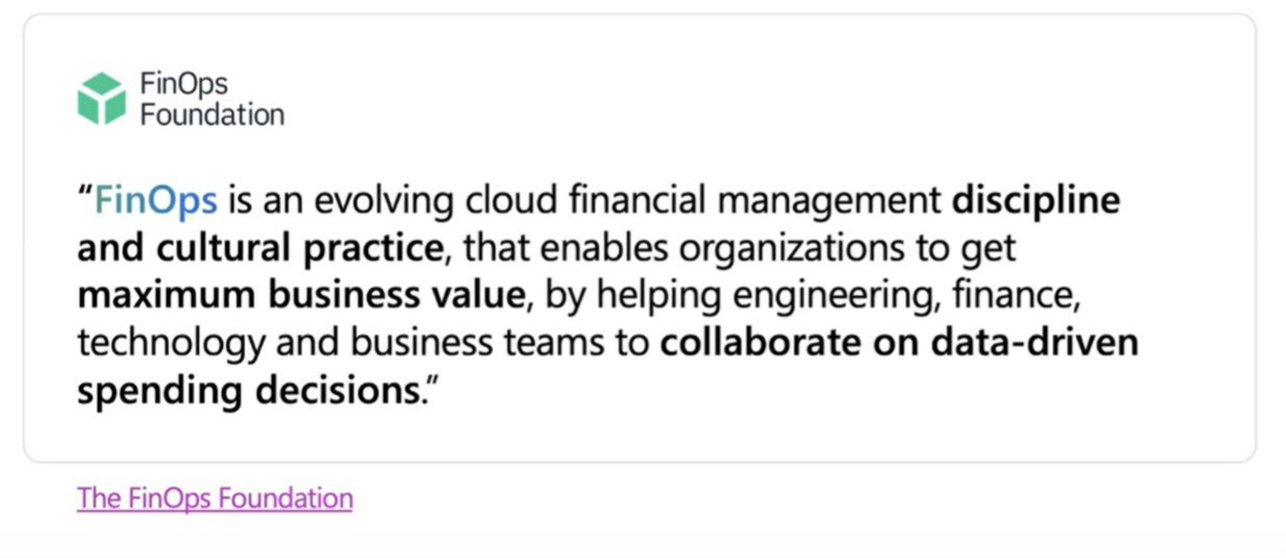
Other infrastructure cost considerations

- **AKS** – application autoscaling, cluster autoscaling, node scaling, etc.
- **Serverless costs** – Azure Functions etc.
- **Availability zones**
- **Storage**
- **Load balancing**
- **Networking**
- **Security products**
- **Monitoring** (e.g. sending data to Azure Monitor Logs)
- **Backups**



Learn more

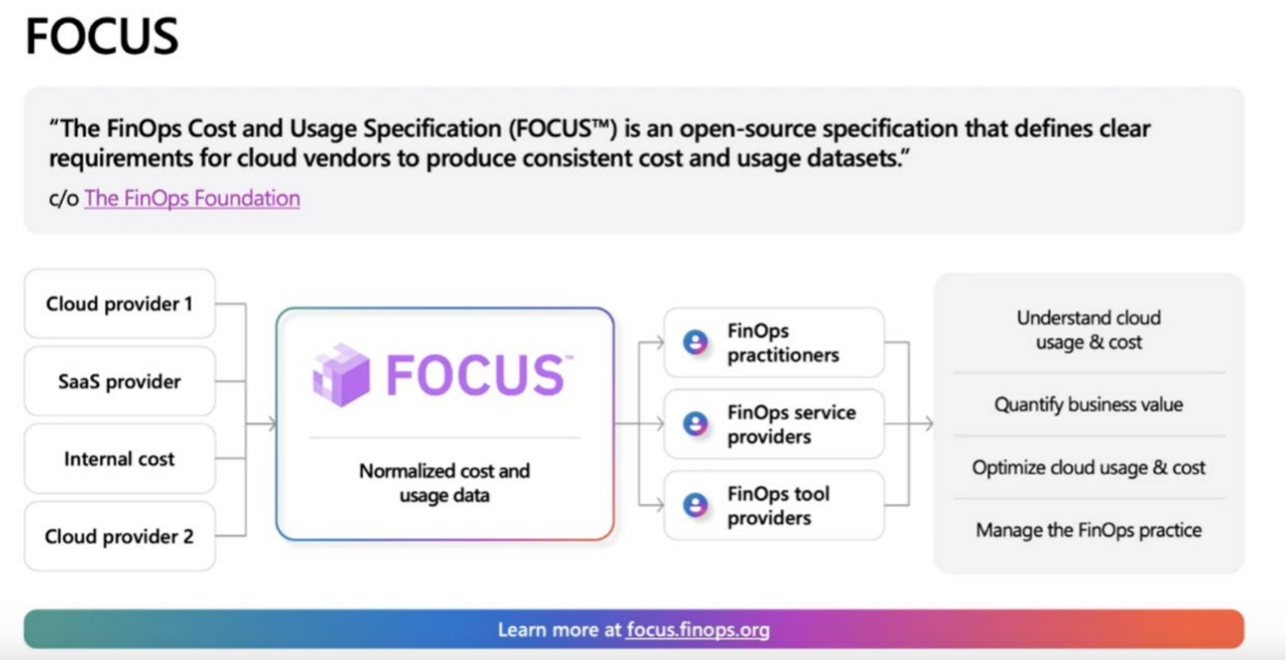
- **Azure Pricing Calculator**
  https://azure.microsoft.com/pricing/calculator
- **Azure OpenAI pricing**
  https://azure.microsoft.com/pricing/details/cognitive-services/openai-service
- **Plan to manage costs for Azure OpenAI Service**
  https://learn.microsoft.com/azure/ai-services/openai/how-to/manage-costs
- **Azure AI Search pricing**
  https://azure.microsoft.com/pricing/details/search/
- **Plan and manage costs of an Azure AI Search service**
  https://learn.microsoft.com/azure/search/search-sku-manage-costs



FinOps Foundation

"FinOps is an evolving cloud financial management **discipline and cultural practice**, that enables organizations to get **maximum business value**, by helping engineering, finance, technology and business teams to **collaborate on data-driven spending decisions**."

The FinOps Foundation

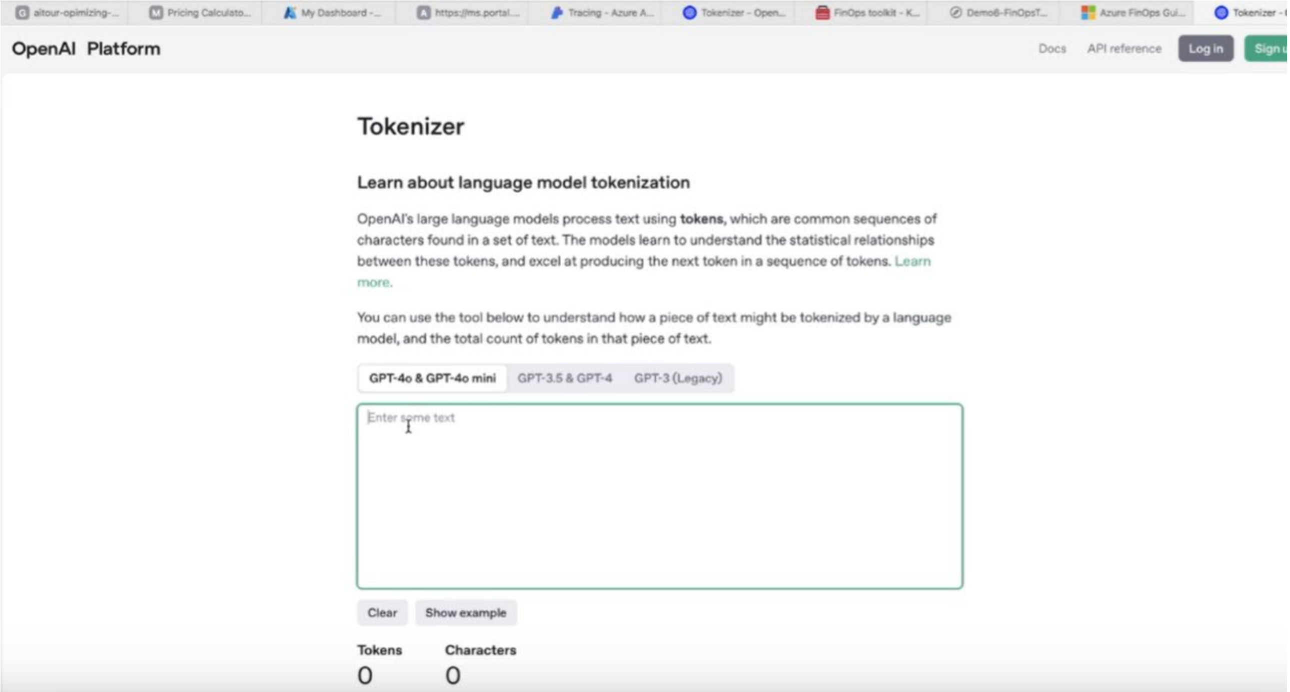


FOCUS

"The FinOps Cost and Usage Specification (FOCUS™) is an open-source specification that defines clear requirements for cloud vendors to produce consistent cost and usage datasets."

c/o The FinOps Foundation

Learn more at focus.finops.org



Tokenizer

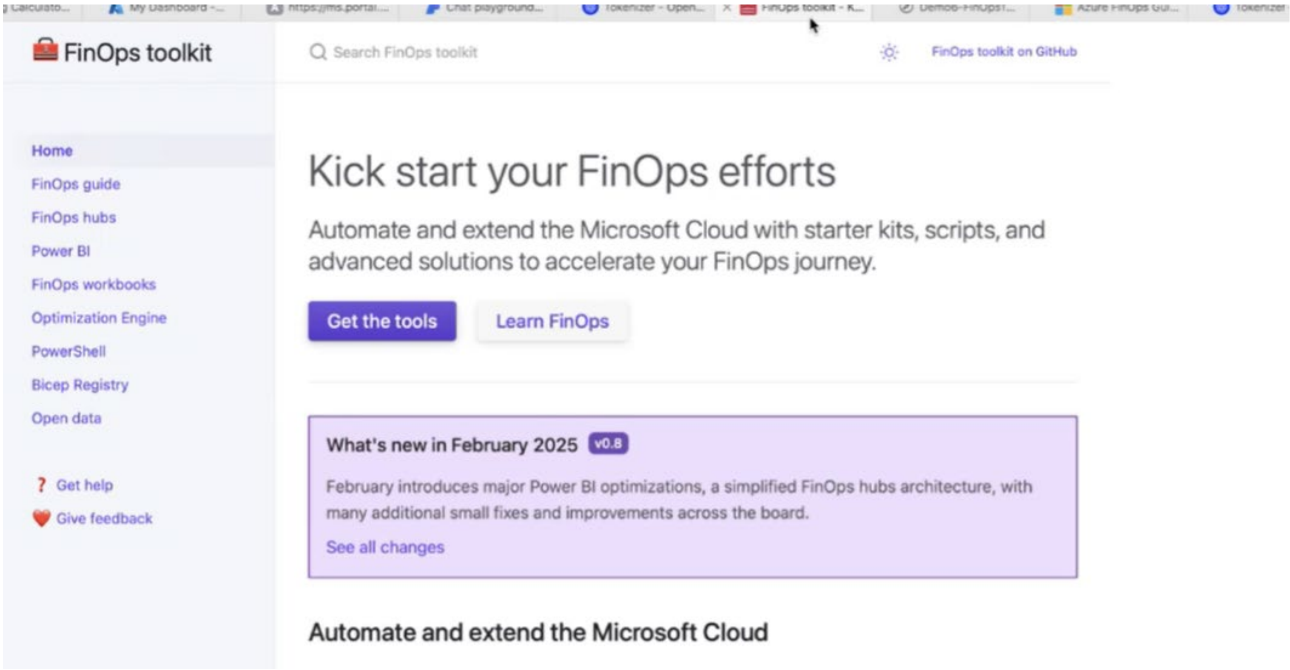
Learn about language model tokenization

OpenAI's large language models process text using **tokens**, which are common sequences of characters found in a set of text. The models learn to understand the statistical relationships between these tokens, and excel at producing the next token in a sequence of tokens. Learn more.

You can use the tool below to understand how a piece of text might be tokenized by a language model, and the total count of tokens in that piece of text.

Tokens 0

Characters 0



Kick start your FinOps efforts

Automate and extend the Microsoft Cloud with starter kits, scripts, and advanced solutions to accelerate your FinOps journey.

What's new in February 2025 v0.8

February introduces major Power BI optimizations, a simplified FinOps hubs architecture, with many additional small fixes and improvements across the board.

See all changes

Automate and extend the Microsoft Cloud