Azure AI Cost Cheat Sheet - December 2025

Quick Pricing Reference

GPT Model Pricing (per 1,000 tokens)

Model	Input	Output	Cost vs GPT-3.5
GPT-3.5-Turbo	$0.002	$0.002	1x (baseline)
GPT-4o	$0.005	$0.015	4x
GPT-4 Turbo	$0.01	$0.02	7.5x
GPT-4 (32K)	$0.06	$0.12	45x

Hidden Costs Calculator

Fine-Tuned Model Hosting:

- $2.52-$3.00 per hour
- $1,836-$2,160 per month (regardless of usage)
- Auto-deleted after 15 days of inactivity

Infrastructure Overhead:

- Cognitive Services resource: $0-$12/month
- Key Vault: ~$3/month
- Virtual Network (private endpoints): $7.20/month per endpoint
- Storage Account: $2-5/month
- Azure Monitor: $5-50/month

Real Cost Formula

`Total Monthly Cost = (Token Usage Cost) + (Fine-tuned Model Hosting × Models × 730 hours) + (Infrastructure Overhead) + (Error retry overhead: ~10%)`

Example: Production Chatbot

Scenario:

- 1M interactions/month
- 100 input + 300 output tokens per interaction
- GPT-4 Turbo model
- 1 fine-tuned model

Calculation:

- Input: 1M × 100 / 1,000 × $0.01 = $1,000
- Output: 1M × 300 / 1,000 × $0.02 = $6,000
- Fine-tuning: $1,840/month
- Infrastructure: $35/month
- Retry overhead: $700/month
- **Total: $9,575/month**

Microsoft Calculator Shows: $7,000 Difference: $2,575/month = $30,900/year

When to Use Each Model

GPT-4 Turbo:

- Complex analysis requiring reasoning
- High-stakes content (legal, financial)
- Tasks where mistakes are expensive

GPT-4o:

- Balance of quality and cost
- General-purpose applications
- Mixed workloads

GPT-3.5:

- Simple summarization
- Data transformation
- High-volume, low-complexity tasks

PTU Pricing (Enterprise)

Provisioned Throughput Units:

- Starting at $2,448/month per PTU
- Save up to 70% vs pay-as-you-go
- Requires annual commitment
- Breakeven: ~$5,000/month workload

Cost Optimization Tips

1. **Start with GPT-3.5** - Prove value, then upgrade selectively
2. **Delete unused fine-tuned models** - They cost $1,836/month even when idle
3. **Optimize prompts, not responses** - Reduce input tokens by 60%+
4. **Use PTUs for production** - 50-70% savings with annual reservations
5. **Monitor per application** - Tag deployments, track with KQL

Common Cost Traps

▯ **Don't:**

- Deploy fine-tuned models without active monitoring
- Use GPT-4 for everything "because it's better"
- Trust the pricing calculator alone
- Ignore infrastructure costs

▯ **Do:**

- Run 2-week pilot with real logging
- Measure actual input/output ratios
- Include all dependent Azure services
- Test multiple models for each use case

Download Complete Guide

For the full article with detailed examples and production deployment strategies: https://azure-noob.com/blog/azure-openai-pricing-real-costs

Azure Noob - December 2025 Production-tested on 31,000+ resources across 44 Azure subscriptions