🚀 Enterprise AI Gateway Architecture

Multi-Region • Multi-Model • Multi-Subscription • Documented Limits

📋 Key Azure OpenAI Documented Limits

Resources per Region per Subscription: 30 max
Source: Azure OpenAI quotas documentation

Default Enterprise Agreement Limits:
- GPT-4o: 30M TPM, 180K RPM
- GPT-4.1: 5M TPM, 5K RPM
- GPT-4o-mini: 50M TPM, 300K RPM

Default Standard Limits:
- GPT-4o: 450K TPM, 2.7K RPM
- GPT-4.1: 1M TPM, 1K RPM
- GPT-4o-mini: 2M TPM, 12K RPM

APIM Multi-Region:
- Premium tier only
- Separate rate limits per regional gateway
- 15+ minutes for infrastructure changes

🌐 Akamai Global CDN Edge Layer

Intelligent routing, caching, bot protection, and DDoS mitigation

🇺🇸 North America Edge

- Route to: contoso-eastus-01.regional.azure-api.net
- Smart model selection (4o vs 4.1 vs mini)
- Complexity-based routing
- Bot protection & rate limiting

🇪🇺 Europe Edge

- Route to: contoso-westeurope-01.regional.azure-api.net
- GDPR compliance routing
- EU data residency
- Regional optimization

🌏 APAC Edge

- Route to: contoso-southeastasia-01.regional.azure-api.net
- APAC latency optimization
- Regional compliance
- Performance monitoring

⬇️

⚡ Azure API Management - Multi-Region Gateways

Single Premium APIM instance: ~$2,500/month • Regional gateway deployment • Management plane in primary region only

East US Gateway (Primary)

- **Management Plane + Gateway**
- Content Safety Policies (threshold=4)
- Token Rate Limiting (per region)
- Circuit Breakers & Retry Logic
- Policy sync to other regions (<10s)

West Europe Gateway

- **Gateway Only**
- Synchronized policies from primary
- Independent rate limit counters
- Regional backend pools
- Automatic failover support

Southeast Asia Gateway

- **Gateway Only**
- Synchronized policies from primary
- Independent rate limit counters
- Regional backend pools
- Health monitoring

⬇️

🎯 Azure OpenAI Backend Infrastructure

Strategy: Multiple models per subscription for maximum quota utilization (up to 30 resources per region)

🇺🇸 North America (East US)

Subscription A (Enterprise Agreement)

Limit: 30 Azure OpenAI resources per region

Model	Capacity	Type
GPT-4o	30M TPM 180K RPM	PTU
GPT-4o	30M TPM 180K RPM	PAYG
GPT-4.1	5M TPM 5K RPM	PTU
GPT-4o-mini	50M TPM 300K RPM	PTU

Subscription B (Standard Backup)

Limit: 30 Azure OpenAI resources per region

Model	Capacity	Type
GPT-4o	450K TPM 2.7K RPM	PAYG
GPT-4.1	1M TPM 1K RPM	PAYG
GPT-4o-mini	2M TPM 12K RPM	PAYG

🇪🇺 Europe (West Europe)

Subscription C (Enterprise Agreement)

Limit: 30 Azure OpenAI resources per region

Model	Capacity	Type
GPT-4o	30M TPM 180K RPM	PTU
GPT-4.1	5M TPM 5K RPM	PTU
GPT-4o-mini	50M TPM 300K RPM	PTU

Subscription D (Standard Backup)

Limit: 30 Azure OpenAI resources per region

Model	Capacity	Type
GPT-4o	450K TPM 2.7K RPM	PAYG
GPT-4o-mini	2M TPM 12K RPM	PAYG

🌏 APAC (Southeast Asia)

Subscription E (Enterprise Agreement)

Limit: 30 Azure OpenAI resources per region

Model	Capacity	Type
GPT-4o	30M TPM 180K RPM	PTU
GPT-4.1	5M TPM 5K RPM	PTU
GPT-4o-mini	50M TPM 300K RPM	PTU

Subscription F (Standard Backup)

Limit: 30 Azure OpenAI resources per region

Model	Capacity	Type
GPT-4o	450K TPM 2.7K RPM	PAYG
GPT-4o-mini	2M TPM 12K RPM	PAYG

📊 Total Documented System Capacity

255M Total TPM (Enterprise) GPT-4o: 180M + GPT-4.1: 15M + Mini: 150M × 3 regions	**22.05M** Total TPM (Standard Backup) GPT-4o: 2.7M + GPT-4.1: 3M + Mini: 12M × 3 regions	**6** Azure Subscriptions 2 per region × 3 regions	**180** Max Resources 30 per region × 6 subscriptions	**3x** Rate Limit Multiplier Independent counters per APIM gateway

<10s
Policy Sync Time
Primary to regional gateways