



# CDW Documentation

**MAIT-510 - Learn Azure OpenAI: GPT**

---

# MAIT-510 - Learn Azure OpenAI: GPT

## Model Overview and Comparison

Model	Tokens/Minute (TPM)	Requests/Minute (RPM)	Latency	Throughput	Error Handling
<b>gpt-4o</b>	Up to 450K per region	Varies by deployment	Low (real-time)	High (streaming)	Handles large prompts; monitor for 429/500; implement retries/backoff
<b>gpt-4</b>	Varies by deployment	Varies by deployment	Moderate (~1.3s avg)	Moderate	Monitor 429; limit prompt size; retries
<b>gpt-4.1</b>	Up to 30K TPM (enforced)	Varies by deployment	Moderate	Moderate	Known 500s in regions; monitor 429/500
<b>gpt-4.1-mini</b>	Not publicly documented	Not publicly documented	Likely low	Likely high	General best practices apply
<b>gpt-4-32k</b>	Varies by deployment	Varies by deployment	Higher (context size)	Lower	Monitor 429; 32K max prompt
<b>gpt-35-turbo-16k</b>	Varies by deployment	Varies by deployment	Low (~900ms avg)	High	Monitor 429; 16K max prompt
<b>gpt-35-turbo</b>	Varies by deployment	Varies by deployment	Low (~900ms avg)	High	Monitor 429; 4K max prompt
<b>gpt-35-turbo-instruct</b>	Varies by deployment	Varies by deployment	Low	High	Monitor 429; 4K max prompt
<b>gpt-4.5-preview</b>	Not publicly documented	Not publicly documented	Experimental	Experimental	Pre-release; expect bugs; robust error handling
<b>gpt-4.1-nano</b>	Not publicly documented	Not publicly documented	Likely very low	Likely very high	General best practices apply
<b>gpt-image-1</b>	Not publicly documented	Not publicly documented	Moderate	Moderate	Monitor image-specific errors
<b>gpt-4o-mini / tts / audio</b>	Not publicly documented	Not publicly documented	Very low (real-time)	High	Monitor audio errors; use proper input format

## GPT-4o vs GPT-4.1 Turbo Comparison

Category	GPT-4o	GPT-4.1 (Turbo)	Winner
Reasoning	Equal or slightly better	Strong performance	Tie
Coding	Better real-time	Better in benchmarks	GPT-4.1
Math	Better interpretive	Better symbolic precision	Tie / GPT-4.1
Instruction Following	More expressive	More formal	GPT-4o
Multilingual	Better tokenization	Less efficient	GPT-4o
Image Understanding	Native support	Not supported	GPT-4o
Speech/TTS	Built-in STT/TTS	Not supported	GPT-4o
Expressiveness	Dynamic & expressive	Flat tone	GPT-4o
Factual Accuracy	Similar cutoff	Similar cutoff	Tie
Steerability	Strong tone/style ctrl	Text only	GPT-4o
Token Efficiency	Better compression	Slightly worse	GPT-4o

### Summary:

- **GPT-4.1:** Best for symbolic reasoning, coding, structured QA.
- **GPT-4o:** Best for multimodal, expressiveness, efficiency, speech/image.

## Latency Comparison

Model	Avg Latency (Time to First Token)	Notes
GPT-4o	~5 seconds (Azure)	Optimized for low latency + multimodal tasks
GPT-4.1	~45 seconds for 1000-1500 tokens	Higher latency, especially for long completions

## Throughput Comparison

Model	TPM	RPM	Notes
GPT-4o	150,000	900	Higher quotas available via enterprise
GPT-4.1	3,000/PTU	Varies	Dependent on Provisioned Throughput

## Use Cases

### 1. Automated IT Support & Triage

- **Use:** GPT-4o or GPT-4.1
- **Tasks:** Triage tickets, Tier-1 fixes, generate CLI, summarize alerts
- **Benefits:** Faster, reduces L1 work, integrates with ServiceNow or DevOps

### 2. Infrastructure-as-Code Review

- **Use:** GPT-4.1
- **Tasks:** Review Bicep/ARM/Pulumi, validate configs
- **Benefits:** Promotes standardization, catches misconfigs

### 3. Security & Policy Review

- **Use:** GPT-4o / GPT-4.1
- **Tasks:** Analyze IAM, firewalls, audit logs; policy translation
- **Benefits:** Faster audits, stronger compliance, cross-team alignment

## Manual Testing: Thermochemistry Prompt

**Prompt:** Calculate  $\Delta H$  (kJ/mol  $\text{NaNO}_3$ ) using a calorimeter (451 J/°C), 0.0300 mol NaOH, 1000 mL of 0.0300 M  $\text{HNO}_3$ ,  $T \uparrow$  from 23.000°C → 23.639°C. Assume 4.18 J/g°C, 1.00 g/mL.

#### GPT-4.1 Output:

- Heat (solution): **2673.3 J** (should be 2671.02)
- Calorimeter: 288.4 J
- Total q: 2961.7 J
- $\Delta H = -98.7$  kJ/mol

#### GPT-4o Output:

- Heat (solution): **2672.82 J**
- Calorimeter: 288.69 J
- Total q: 2961.51 J
- $\Delta H = -98.7$  kJ/mol

#### Correct Calculation:

- $1000 \times 4.18 \times 0.639 = 2671.02$  J
- Total heat = 2671.02 + 288.69 = **2959.71 J**
- $\Delta H = -2959.71 / 0.0300 = -98.7$  kJ/mol

## Conclusion

- GPT-4.1 = better **explanations**, but made arithmetic errors.
- GPT-4o = better **numerical skill**, but also rounded incorrectly.
- Both models accepted feedback—but repeated the **same mistake**.
- ChatGPT (web version) corrected its error and gave the **correct final answer**.
- Playground versions seem more prone to **repeating numeric errors**.
- **GPT-4.1** = best for detailed QA/debug work.
- **GPT-4o** = best for expressive, real-time, multimodal tasks.