

[Back to Models](#)**OpenAI: GPT-6 Astra** Active[openai/gpt-6-astra](#)[Compare](#)[API Request](#)[Chat](#)

GPT-6 Astra is OpenAI's most capable model for complex reasoning, coding, computer use, research, and document creation, with strong performance on multistep workflows across code, browsers, and professional software. It supports reasoning effort levels of low, medium, high, xhigh, and max, and achieves strong results while often using fewer output tokens for lower estimated cost per task than earlier models. It is also OpenAI's most aligned model yet, designed to follow intent carefully, respect task boundaries, adapt to changing requirements, and communicate transparently.

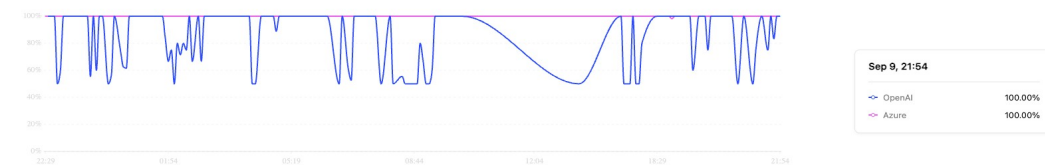
By : OpenAI Input Type : [Text](#) [Image](#) [Audio](#) Output Type : [Text](#) Publish time : 2026-09-04**Providers**

Route requests across multiple providers. Copy a provider slug to set your preference.

Provider	Cache Hit Rate	Input	Output	Cache read	Context	Latency	Throughput
OpenAI	86.8%	\$10-20 / M tokens	\$50-75 / M tokens	Read: 1-2 / M tokens Write: 12.5-25 / M tokens	1.05M	1.2s	38.4tps
Azure	88.2%	\$10-20 / M tokens	\$50-75 / M tokens	Read: 1-2 / M tokens Write: 12.5-25 / M tokens	1.05M	5.2s	11.7tps

Uptime 24hours

Direct request success rate on AI Gateway and per-provider.



ZenMux Launches GPT-6 Astra API Access Through Unified Platform

September 17, 2026

Singapore - September 17, 2026 -

ZenMux has added OpenAI's GPT-6 Astra to its unified model routing platform. The listing went live following OpenAI's early September release of the model, giving developers a single environment to evaluate GPT-6 Astra alongside other systems already on ZenMux.

The addition targets technical teams that need to test a newly released model without building a separate integration path for each provider. Rather than maintaining individual API connections for OpenAI, Azure, and other endpoints, teams accessing the ZenMux GPT-6 Astra API can review provider availability, context limits, supported interfaces, and current operating indicators from a single model page before deciding how to route requests in an application or evaluation workflow.

OpenAI describes GPT-6 Astra as a model built for complex reasoning, coding, computer use, research, and professional document creation. According to OpenAI's announcement, the model is available through its API under the identifier gpt-6-astra, as well as through Microsoft Azure and Amazon Bedrock. OpenAI lists

computer and browser use, software engineering, cybersecurity, scientific work, and multistep professional tasks among its primary use cases.

On ZenMux, the model page currently lists OpenAI and Azure as available providers, alongside a 1.05-million-token context window and support for text and image input. Provider-level indicators covering latency, throughput, cache activity, and uptime are displayed alongside these specifications, giving engineering teams a way to compare routing options side by side. ZenMux notes that these indicators are meant to support evaluation rather than replace it ? organizations should still test cost, output quality, and policy requirements against their own workloads before moving anything into production.

The model can be reached through response, chat completion, and messages-based API pathways, depending on the provider selected and the parameters supported. Teams already using a compatible integration pattern can evaluate GPT-6 Astra within the same broader environment used for other models, without maintaining a separate discovery and routing process just for this one release.

The launch coincides with an account-level activity currently running on ZenMux. It includes a service-fee adjustment for qualifying credit top-ups and separate referral benefits, with full conditions displayed in the account interface. ZenMux describes this as an account-level program, not a change to the model's published token rates, and advises users to review applicable terms before making routing or account decisions.

Provider availability is only one part of a model evaluation. Request format, context requirements, latency, throughput, caching behavior, and internal governance requirements can all determine whether a given model fits a specific workload. ZenMux organizes these details around each model listing, including the ZenMux GPT-6 Astra API page, so teams can document tests and compare results while working from a common API layer rather than juggling separate provider dashboards.

The page also sits alongside a broader catalog spanning OpenAI, Anthropic, Google, DeepSeek, Z.ai, and other providers. ZenMux says this structure is meant for organizations that expect their model choices to shift as applications move from prototype to production, and as different tasks call for different combinations of reasoning depth, speed, context capacity, and cost control.

ZenMux continues expanding its catalog as providers release new systems and update existing endpoints. The GPT-6 Astra listing is now available for developers and organizations reviewing model options for coding, research, computer-use, and other multistep workflows.

For more information, visit <https://zenmux.ai/openai/gpt-6-astra>.

About ZenMux:

ZenMux is an enterprise-grade large model aggregation platform with an insurance payout mechanism. The platform provides one-stop access to the latest models across providers. When issues such as poor output quality or excessive latency occur during use, our intelligent insurance detection and payout mechanism automatically compensates, addressing enterprise concerns around AI hallucinations and unstable quality.

###

For more information about ZenMux, contact the company here: ZenMuxEmberember@zenmux.ai Singapore

ZenMux

ZenMux is an enterprise-grade AI model aggregation platform with a built-in insurance payout mechanism, offering one-stop access across providers and automatic compensation for quality or latency issues.

Website: <https://zenmux.ai/>

Email: ember@zenmux.ai

