

Introduction

Vivgrid is an AI Agent and LLM infrastructure platform that helps developers build, deploy, and operate AI-powered applications with speed, reliability, and cost efficiency.

Snapshot source: <https://docs.vivgrid.com/introduction>

Generated: January 23, 2026

Build AI Agents

AI Agents are rapidly gaining adoption as large language models become more capable of executing complex tasks through reliable tool calling and reasoning. High-quality AI Agents require continuous optimization across three core dimensions:

- Accuracy
- Latency
- Cost efficiency

Sustained improvement across these dimensions is essential for production-grade agent systems. Vivgrid follows a decoupled architecture: agent logic is separated from client applications. This significantly reduces iteration friction, allowing teams to evolve agent capabilities independently while maintaining stable integrations. Vivgrid provides the following capabilities to help developers build, deploy, and operate AI Agents at scale:

Capability	What it does
Serverless Tools	Build strongly typed LLM tools and run them as managed serverless functions with minimal operational overhead.
Models	Access state-of-the-art models through a unified, production-ready endpoint.
Agent Evaluation	Run recursive and regression evaluations to validate agent behavior before release.
Observability & Insights	Analyze agent performance with detailed request logs, latency breakdowns, and cost metrics.

Model API

Vivgrid also provides a unified LLM Model API, enabling direct access to commonly used models for coding, text, speech, and multimodal workloads. The API can be seamlessly integrated into third-party tools such as Claude Code, Cursor, OpenCode, n8n, and Dify.

Vibe Coding Models: Supported coding models include:

- gpt-5.1-codex
- gpt-5.1-codex-max
- gpt-5.2-codex

These models are optimized for code generation, completion, refactoring, and interactive development workflows.

Additional Models: Vivgrid supports a broad range of model capabilities, including:

- Text-to-Speech (TTS)
- Speech-to-Text (Whisper)
- OCR
- Reranking

For the complete and up-to-date model specification, refer to the Models documentation and check out the API reference for detailed usage instructions.