

AI G...

Search... K

Getting Started

Expand menu

Projects and Deployments

Use a Template

Import Existing Project

Add a Domain

Buy a Domain

Transfer an Existing Domain

Collaborate

Next Steps

Supported Frameworks

Expand menu

Full-stack

Expand menu

Next.js

SvelteKit

Nuxt

Remix

Frontends

Expand menu

Astro

Vite

React Router

Gatsby

Create React App

Backends

Expand menu

Nitro

Hono

All Frameworks

Incremental Migration

Production Checklist

Guides

Access

Expand menu

Account Management

Activity Log

Deployment Protection

Expand menu

Bypass Deployment Protection

Expand menu

Exceptions

OPTIONS

Allowlist

Protection

Bypass for Automation

Sharable Links

Protect Deployments

Expand menu

Password Protection

Trusted IPs

Vercel Authentication

Directory Sync

SAML SSO

Two-factor (2FA)

AI

Expand menu

v0

AI SDK

Agents

AI Gatewaybeta

Expand menu

Models & Providers

Observability

Pricing

Provider Options

OpenAI-Compatible API

Authentication

BYOK

Community Frameworks

Expand menu

LangChain

Mastra

Model Variants

MCP

Expand menu

Deploy MCP servers

Vercel MCP serverbeta

Expand menu

Tools

Integrations

Expand menu

Adding a Provider

Adding a Model

xAI

Groq

fal

Deep Infra

ElevenLabs

LMNT

Copy page

▼

AI Gateway

AI Gateway is available in Beta on all plans and your use is subject to Vercel's Public Beta Agreement and AI Product Terms.

The AI Gateway provides a unified API to access 100+ models through a single endpoint. It gives you the ability to set budgets, monitor usage, load-balance requests, and manage fallbacks.

The design allows it to work seamlessly with AI SDK 5, OpenAI SDK, or your preferred framework.

Some of the key features of the AI Gateway include:

- Unified API: helps you switch between providers and models with minimal code changes
- High Reliability: automatically retries requests to other providers if one fails
- Embeddings Support: generate vector embeddings for search, retrieval, and other tasks
- Spend Monitoring: monitor your spending across different providers

Getting started

Create an API key

First, let's create an API key in the AI Gateway tab of the Vercel Dashboard:

- From the Vercel dashboard, click the AI Gateway tab
- Click API keys on the left side bar
- Click Create key and proceed with Create key from the Dialog

Once you have the API key, you'll need to provide it to your application, which we will do in the next section.

①This guide uses API keys. You can also use OIDC tokens to authenticate your requests.

Using AI SDK

The AI SDK is a Typescript library that helps developers build AI-powered applications. In the below quickstart, you'll build a simple AI-chatbot with a streaming user interface.

Prerequisites

To follow this quickstart, you'll need:

- Node.js 18+ and `pnpm` installed on your local development machine.
- An AI Gateway API key (created in the previous step).

Setup your application

Start by creating a new directory using the `mkdir` command. Change into your new directory and then run the `pnpm init` command. This will create a `package.json` in your new directory.

>_terminal

1mkdir demo

2cd demo

3pnpm init

Install Dependencies

Install `ai`, along with other necessary dependencies.

The AI SDK is designed to be a unified interface to interact with any large language model. This means that you can change model and providers with just one line of code! Learn more about available providers and building custom providers in the providers section.

>_terminal

1pnpm add ai zod dotenv

2pnpm add -D @types/node tsx typescript

The `ai` package contains the AI SDK. You will use `zod` to define type-safe schemas that you will pass to the large language model (LLM). You will use `dotenv` to access environment variables (your AI Gateway API key) within your application. There are also three development dependencies, installed with the `-D` flag, that are necessary to run your Typescript code.

Configure AI Gateway API key

Create a `.env` file in your project's root directory and add your AI Gateway API Key. This key is used to authenticate your application with the AI Gateway service.

>_terminal

touch .env

Edit the `.env` file:

env

AI_GATEWAY_API_KEY=your_ai_gateway_api_key

Replace `your_ai_gateway_api_key` with your actual AI Gateway API key from the previous step.

The AI Gateway provider will default to using the `AI_GATEWAY_API_KEY` environment variable.

On this page

Getting started

Create an API key

Using AI SDK

Prerequisites

Setup your application

Install Dependencies

Configure AI Gateway API key

Create your application

Run your application

Next steps

Using OpenAI SDK

Using other community frameworks

- [OpenAI](#)
- [Perplexity](#)
- [Pinecone](#)
- [Replicate](#)
- [Together AI](#)
- API
 - ▼ Expand menu
 - [REST API](#)
 - [Vercel SDK](#)
 - [Build Output API](#)
 - ▼ Expand menu
 - [Build Output Configuration](#)
 - [Features](#)
 - [Vercel Primitives](#)
- Build & Deploy
 - ▼ Expand menu
 - [Builds](#)
 - ▼ Expand menu
 - [Build Features](#)
 - [Build Image](#)
 - ▼ Expand menu
 - [Build Image Installed Packages](#)
 - [Build Queues](#)
 - [Configuring a Build](#)
 - [Managing Builds](#)
 - [Deploy Hooks](#)
 - [Deployment Retention](#)
 - [Deployments](#)
 - ▼ Expand menu
 - [Environments](#)
 - [Generated URLs](#)
 - [Managing Deployments](#)
 - [Promoting Deployments](#)
 - [Troubleshoot Build Errors](#)
 - [Accessing Build Logs](#)
 - [Claim Deployments](#)
 - [Inspect OG Metadata](#)
 - [Preview Deployment Suffix](#)
 - [Sharing a Preview Deployment](#)
 - [Environment Variables](#)
 - ▼ Expand menu
 - [Framework Environment Variables](#)
 - [Managing Environment Variables](#)
 - [Reserved Environment Variables](#)
 - [Sensitive Environment Variables](#)
 - [Shared Environment Variables](#)
 - [System Environment Variables](#)
 - [Git Integrations](#)
 - ▼ Expand menu
 - [GitHub](#)
 - [Azure DevOps](#)
 - [Bitbucket](#)
 - [GitLab](#)
 - [Instant Rollback](#)
 - [Microfrontendsbeta](#)
 - ▼ Expand menu
 - [Getting Started](#)
 - [Local Development](#)
 - [Path Routing](#)
 - [Managing Microfrontends](#)
 - ▼ Expand menu
 - [Security](#)
 - [Using Vercel Toolbar](#)
 - [Testing & Troubleshooting](#)
 - [Monorepos](#)
 - ▼ Expand menu
 - [Turborepo](#)
 - [Remote Caching](#)
 - [Nx](#)
 - [Monorepos FAQ](#)
 - [Package Managers](#)
 - [Protected Git Scopes](#)
 - [Rolling Releases](#)
 - [Skew Protection](#)
 - [Webhooks](#)
 - ▼ Expand menu
 - [Webhooks API Reference](#)
 - CDN
 - ▼ Expand menu
 - [Domains](#)
 - ▼ Expand menu
 - [Working with](#)

Create your application

Create an `index.ts` file in the root of your project and add the following code:

Anthropic OpenAI Google Grok

```
index.ts

import { ModelMessage, streamText } from 'ai';
import 'dotenv/config';
import * as readline from 'node:readline/promises';

4

5const terminal = readline.createInterface({
6input: process.stdin,
7output: process.stdout,
8});
9
10const messages: ModelMessage[] = [];
11
12async function main() {
13while (true) {
14const userInput = await terminal.question('You: ');
15
16  messages.push({ role: 'user', content: userInput });
17
18  const result = streamText({
19    model: 'anthropic/claude-opus-4-20250514',
20    messages,
21  });
22
23  let fullResponse = '';
24  process.stdout.write('\nAssistant: ');
25  for await (const delta of result.textStream) {
26    fullResponse += delta;
27    process.stdout.write(delta);
28  }
29  process.stdout.write('\n\n');
30
31  messages.push({ role: 'assistant', content: fullResponse });
32
33}
34}
35
36main().catch(console.error);
37
```

Run your application

Now you can run your application:

```
>_terminal

pnpm tsx index.ts
```

You should see a prompt where you can start chatting with the AI. The responses will be streamed in real-time!

Next steps

Continue with the [AI SDK documentation](#) to learn advanced configuration, set up provider routing and fallbacks, and explore more integration examples.

Using OpenAI SDK

The AI Gateway provides OpenAI-compatible API endpoints that allow you to use existing OpenAI client libraries and tools with the AI Gateway.

The OpenAI-compatible API includes:

- Model Management: List and retrieve the available models
- Chat Completions: Create chat completions that support streaming, images, and file attachments
- Tool Calls: Call functions with automatic or explicit tool selection
- Existing Tool Integration: Use your existing OpenAI client libraries and tools without needing modifications

Learn more about using the OpenAI SDK with the AI Gateway in the [OpenAI-Compatible API page](#).

Using other community frameworks

The AI Gateway is designed to work with any framework that supports the OpenAI API or AI SDK 5.

Read more about using the AI Gateway with other community frameworks in the [AI Gateway with community frameworks](#) section.

Last updated on August 6, 2025

Domains

▼ Expand menu

- [Adding a Domain](#)
- [Adding a Domain to an Environment](#)
- [Assigning a Domain to a Git Branch](#)
- [Deploying & Redirecting Domains](#)
- [Removing a Domain](#)
- [Renewing a Domain](#)
- [Transferring Domains](#)
- [Viewing & Searching Domains](#)

- [Working with DNS](#)
- [Managing DNS Records](#)
- [Working with Nameservers](#)
- [Managing Nameservers](#)
- [Working with SSL](#)
- [Custom SSL Certificates](#)
- [Pre-Generate SSL Certificates](#)
- [Supported Domains](#)
- [Troubleshooting Domains](#)

◦ [Edge Network](#)

▼ Expand menu

- [Regions](#)
- [Compression](#)
- [Manage Usage](#)

◦ [Encryption](#)

◦ [Headers](#)

▼ Expand menu

- [Security Headers](#)
- [Cache-Control Headers](#)
- [Request Headers](#)
- [Response Headers](#)

◦ [Image Optimization](#)

▼ Expand menu

- [Getting Started](#)
- [Limits and Pricing](#)
- [Managing Usage & Costs](#)
- [Legacy Pricing](#)

◦ [Incremental Static Regeneration](#)

▼ Expand menu

- [Getting Started](#)
- [Usage & Pricing](#)

◦ [Redirects](#)

◦ [Rewrites](#)

◦ [Vercel Cache](#)

• Collaboration

▼ Expand menu

◦ [Comments](#)

▼ Expand menu

- [Enabling Comments](#)
- [Using Comments](#)
- [Managing Comments](#)
- [Integrations](#)

◦ [Draft Mode](#)

◦ [Edit Mode](#)

◦ [Feature Flags](#)

▼ Expand menu

- [Flags Explorer](#)

▼ Expand menu

- [Getting Started](#)
- [Reference](#)
- [Pricing](#)
- [Flags SDK](#)
- [With Runtime Logs](#)
- [With Vercel Platform](#)
- [With Web Analytics](#)

◦ [Toolbar](#)

▼ Expand menu

- [Add to Environments](#)

▼ Expand menu

- [Add to Localhost](#)
- [Add to Production](#)
- [Managing Toolbar](#)
- [Browser Extensions](#)
- [Accessibility Audit Tool](#)
- [Interaction Timing Tool](#)
- [Layout Shift Tool](#)

• Compute

▼ Expand menu

◦ [Fluid Compute](#)

▼ Expand menu

- [Pricing](#)

- [Functions](#)
 - ▼ Expand menu
 - [Getting Started](#)
 - [Streaming](#)
 - [Runtimes](#)
 - ▼ Expand menu
 - [Node.js](#)
 - ▼ Expand menu
 - [Advanced Node.js Usage](#)
 - [Supported Node.js versions](#)
 - [Python](#)
 - [Go RuntimeGo](#)
 - [Ruby](#)
 - [Wasm](#)
 - [Edge Runtime](#)
 - [Configuring Functions](#)
 - ▼ Expand menu
 - [Duration](#)
 - [Memory](#)
 - [Runtime](#)
 - [Region](#)
 - [Advanced Configuration](#)
 - [API Reference](#)
 - ▼ Expand menu
 - [@vercel/functions](#)
 - [Logs](#)
 - [Limits](#)
 - [Concurrency Scaling](#)
 - [Data Cache](#)
 - [Routing Middleware](#)
 - ▼ Expand menu
 - [Getting Started](#)
 - [API](#)
 - [Cron Jobs](#)
 - ▼ Expand menu
 - [Getting Started](#)
 - [Managing Cron Jobs](#)
 - [Usage & Pricing](#)
 - [OG Image Generation](#)
 - ▼ Expand menu
 - [@vercel/og](#)
 - [Examples](#)
 - [Sandbox](#)
 - ▼ Expand menu
 - [Examples](#)
 - [Pricing and Limits](#)
 - [Multi-tenant](#)
 - ▼ Expand menu
 - [Domain Management](#)
 - [Limits](#)
 - [Observability](#)
 - ▼ Expand menu
 - [Overview](#)
 - ▼ Expand menu
 - [Insights](#)
 - [Observability Plus](#)
 - [Logs](#)
 - ▼ Expand menu
 - [Runtime](#)
 - [OpenTelemetry](#)
 - [Session Tracing](#)
 - [Query](#)
 - ▼ Expand menu
 - [Query Reference](#)
 - [Monitoring](#)
 - ▼ Expand menu
 - [Getting Started](#)
 - [Monitoring Reference](#)
 - [Limits and Pricing](#)
 - [Notebooks](#)
 - [Speed Insights](#)
 - ▼ Expand menu
 - [Getting Started](#)
 - [Using Speed Insights](#)
 - [Metrics](#)
 - [Privacy](#)
 - [@vercel/speed-insights](#)
 - [Limits and Pricing](#)
 - [Troubleshooting](#)
 - [Migrating from Legacy](#)
 - [Log Drains](#)
 - ▼ Expand menu
 - [Configure Log Drains](#)
 - [Correlate Logs and Traces](#)
 - [Log Drains Reference](#)
 - [Web Analytics](#)
 - ▼ Expand menu

- [Getting Started](#)
- [Using Web Analytics](#)
- [Filtering](#)
- [Custom Events](#)
- [Redacting Sensitive Data](#)
- [Privacy](#)
- [@vercel/analytic](#)
[s](#)
- [Pricing](#)
- [Troubleshooting](#)
- [Manage & Optimize](#)

- Platform

- ▼ Expand menu

- [Dashboard](#)

- ▼ Expand menu

- [Navigating the Dashboard](#)
 - [Support Center](#)
 - [Using the Command Menu](#)

- [Notifications](#)

- [Projects](#)

- ▼ Expand menu

- [Managing projects](#)
 - [Project Dashboard](#)
 - [Transferring a project](#)

- [Project Configuration](#)

- ▼ Expand menu

- [General Settings](#)
 - [Project Settings](#)
 - [Git Configuration](#)
 - [Git Settings](#)
 - [Global Configuration](#)
 - [Security settings](#)

- [Product Changes](#)

- [Checks](#)

- ▼ Expand menu

- [Checks API](#)
 - [Checks Reference](#)

- [CLI](#)

- ▼ Expand menu

- [Deploying from CLI](#)
 - [Project Linking](#)
 - [Telemetry](#)
 - [Global Options](#)
 - [vercel alias](#)
 - [vercel bisect](#)
 - [vercel blob](#)
 - [vercel build](#)
 - [vercel cache](#)
 - [vercel certs](#)
 - [vercel deploy](#)
 - [vercel dev](#)
 - [vercel dns](#)
 - [vercel domains](#)
 - [vercel env](#)
 - [vercel git](#)
 - [vercel help](#)
 - [vercel init](#)
 - [vercel inspect](#)
 - [vercel install](#)
 - [vercel integration](#)
 - [vercel integration-resource](#)
 - [vercel link](#)
 - [vercel list](#)
 - [vercel login](#)
 - [vercel logout](#)
 - [vercel logs](#)
 - [vercel project](#)
 - [vercel promote](#)
 - [vercel pull](#)
 - [vercel redeploy](#)
 - [vercel remove](#)
 - [vercel rollback](#)
 - [vercel rolling-release](#)
 - [vercel switch](#)
 - [vercel teams](#)
 - [vercel telemetry](#)
 - [vercel whoami](#)

- [Glossary](#)

- [Integrations](#)

- ▼ Expand menu

- [Extend Vercel](#)

- ▼ Expand menu

- [Add a Connectable Account](#)
 - [Add a Native Integration](#)
 - [Permissions and Access](#)

- [Integrate with Vercel](#)

- ▼ Expand menu

- [Native integration concepts](#)
 - [Create a Native Integration](#)
 - [Deployment integration actions](#)
 - [Native Integration](#)

- Flows
 - Native
 - [Integrations REST API](#)
 - Integration
 - [Approval Checklist](#)
 - Integration Image
 - [Guidelines](#)
 - Requirements
 - [for listing an Integration](#)
 - Upgrade an Integration
 - CMS Integrations
 - Expand menu
 - [Agility CMS](#)
 - [ButterCMS](#)
 - [Contentful](#)
 - [DatoCMS](#)
 - [Formsfree](#)
 - [Makeswift](#)
 - [Sanity](#)
 - [Sitecore](#)
 - Ecommerce Integrations
 - Expand menu
 - [BigCommerce](#)
 - [Shopify](#)
 - [Sign in with Vercel](#)
 - [Building Integrations with Vercel REST API](#)
 - External Platforms
 - Expand menu
 - [Kubernetes](#)
 - Limits
 - Expand menu
 - [Fair use Guidelines](#)
- Pricing
 - Expand menu
 - Plans
 - Expand menu
 - [Hobby Plan](#)
 - [Pro Plan](#)
 - Expand menu
 - [Pro Plan Trial](#)
 - [Billing FAQ](#)
 - [Enterprise Plan](#)
 - Expand menu
 - [Billing FAQ](#)
 - Pricing
 - Expand menu
 - [Regional Pricing](#)
 - Expand menu
 - [Cape Town, South Africa](#)
 - [Cleveland, USA](#)
 - [Dubai, UAE](#)
 - [Dublin, Ireland](#)
 - [Frankfurt, Germany](#)
 - [Hong Kong](#)
 - [London, UK](#)
 - [Mumbai, India](#)
 - [Osaka, Japan](#)
 - [Paris, France](#)
 - [Portland, USA](#)
 - [San Francisco, USA](#)
 - [São Paulo, Brazil](#)
 - [Seoul, South Korea](#)
 - [Singapore](#)
 - [Stockholm, Sweden](#)
 - [Sydney, Australia](#)
 - [Tokyo, Japan](#)
 - [Washington, D.C., USA](#)
 - [Manage and Optimize Usage](#)
 - [Calculating Usage of Resources](#)
 - [Billing & Invoices](#)
 - [Legacy Metrics](#)
 - Spend Management
 - Security
 - Expand menu
 - Overview
 - Expand menu
 - [Security & Compliance Measures](#)
 - [Shared Responsibility Model](#)
 - [PCI DSS iframe Integration](#)
 - [Reverse Proxy Servers and Vercel](#)
 - [Access Control](#)

- [Audit Logs](#)
- [Firewall](#)
 - ▼ Expand menu
 - [Firewall Concepts](#)
 - [DDoS Mitigation](#)
 - [Attack Challenge Mode](#)
 - [Web Application Firewall](#)
 - ▼ Expand menu
 - [Custom Rules](#)
 - [Rate Limiting](#)
 - [Rule Configuration](#)
 - [System Bypass Rules](#)
 - [Rate Limiting SDK](#)
 - [IP Blocking](#)
 - [Managed Rulesets](#)
 - [Examples](#)
 - [Usage & Pricing](#)
 - [Firewall API](#)
 - [Firewall Observability](#)
- [Bot Management](#)
- [BotID](#)
 - ▼ Expand menu
 - [Get Started with BotID](#)
 - [Handling Verified Bots](#)
 - [Advanced BotID Configuration](#)
 - [Form Submissions](#)
 - [Local Development Behavior](#)
- [OIDC](#)
 - ▼ Expand menu
 - [AWS](#)
 - [Azure](#)
 - [Connect your API](#)
 - [Google Cloud Platform](#)
 - [OIDC Reference](#)
- [RBAC](#)
 - ▼ Expand menu
 - [Access Roles](#)
 - ▼ Expand menu
 - [Project Level Roles](#)
 - [Team Level Roles](#)
 - [Access Groups](#)
 - [Managing Team Members](#)
- [Secure Compute](#)
- [Two-factor Enforcement](#)
- [Storage](#)
 - ▼ Expand menu
 - [Blob](#)
 - ▼ Expand menu
 - [Server Uploads](#)
 - [Client Uploads](#)
 - [Using the SDK](#)
 - [Pricing](#)
 - [Security](#)
 - [Examples](#)
 - [Edge Config](#)
 - ▼ Expand menu
 - [Getting Started](#)
 - [Using Edge Config](#)
 - [Edge Configs & REST API](#)
 - [Edge Configs & Dashboard](#)
 - [Edge Config SDK](#)
 - [Limits & Pricing](#)
 - [Integrations](#)
 - ▼ Expand menu
 - [DevCycle](#)
 - [Hypertune](#)
 - [LaunchDarkly](#)
 - [Split](#)
 - [Statsig](#)

Products

AI

Enterprise

Fluid Compute

Next.js

Observability

Previews

Rendering

Security

Turbo

v0 

Resources

Community 

Docs

Guides

Help

Integrations

Pricing

Resources

Solution Partners

Startups

Templates

Company

About

Blog

Careers

Changelog

Events

Contact Us

Customers

Partners

Shipped

Privacy Policy

Legal 

Social

 GitHub

 LinkedIn

 Twitter

 YouTube



All systems normal

Select a display theme:  