Asynchronous Batch API

- Demo application
- Supported Models

import { Render, PackageManagers, WranglerConfig, CURL } from "~/components";

Asynchronous batch processing lets you send a collection (batch) of inference requests in a single call. Instead of expecting immediate responses for every request, the system queues them for processing and returns the results later.

Batch processing is useful for large workloads such as summarization or embeddings when there is no human interaction. Using the batch API will guarantee that your requests are fulfilled eventually, rather than erroring out if Cloudflare does not have enough capacity at a given time.

When you send a batch request, the API immediately acknowledges receipt with a status like `queued` and provides a unique `request_id`. This ID is later used to poll for the final responses once the processing is complete.

You can use the Batch API by either creating and deploying a Cloudflare Worker that leverages the Batch API with the AI binding, using the REST API directly or by starting from a template.

:::note[Note]

Ensure that the total payload is under 10 MB.

:::

Demo application

If you want to get started quickly, click the button below:

Deploy to Workers

This will create a repository in your GitHub account and deploy a ready-to-use Worker that demonstrates how to use Cloudflare's Asynchronous Batch API. The template includes

preconfigured AI bindings, and examples for sending and retrieving batch requests with and without external references. Once deployed, you can visit the live Worker and start experimenting with the Batch API immediately.

Supported Models

Refer to our model catalog for supported models.

□□□ #2
□ □□□ □□□ 2026-11-01 05:27:33 UTC
□ □□□ □□□ 2026-11-01 05:27:35 UTC