OPENAI ARCHITECTURE PRIMER

Background Mode

Asynchronous Responses execution

A visual guide to long-running requests, result retrieval, workflow handoffs, and enterprise control.

FIRST-PARTY DOCUMENTATION / INDEPENDENT LUMINITY SYNTHESIS



Background Mode in One View

The application starts long-running work, continues its own flow, then observes and retrieves a terminal response.

THE ASYNCHRONOUS REQUEST PATH

01
Start
background request

02
Continue
application work

03
Observe
webhook or poll

04
Retrieve
terminal response

Completion Window

WINDOW

soonest is default. Eligible 2h and 8h requests target completion at 30% and 40% below Standard model token rates.

Handle Outcomes

STATUS

Handle completed, failed, incomplete, or canceled responses. Poll or subscribe to signed webhooks.

IN THIS PRIMER

Key Points

01 Run long work asynchronously

Save the response ID, release the initiating path, and retrieve the result after a terminal event.

02 Application owns the workflow

Background mode does not choose the plan, split work, satisfy dependencies, or execute application tools.

03 Results require handling

Monitor every terminal state, validate retrieved output, and authorize each downstream consequence separately.

The request path is Luminity synthesis. OpenAI documents the background lifecycle, completion windows, retrieval methods, and terminal states.

© 2026 Luminity Digital, Inc. | Independent synthesis of official OpenAI documentation.

Source: https://developers.openai.com/api/docs/guides/background · verified Nov 18, 2026

OpenAI Select Partner

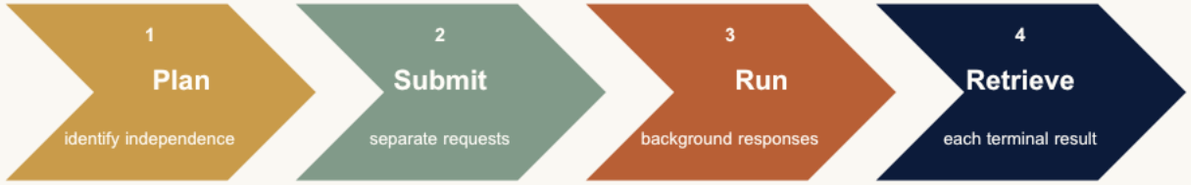
Independent Work and Dependent Handoffs

The application decides which requests can run separately and which steps must wait for prior output.

01

Independent Requests

separate inputs | application-selected | concurrent work



Separate Requests — ASYNC

The application may submit independent requests without waiting between submissions.

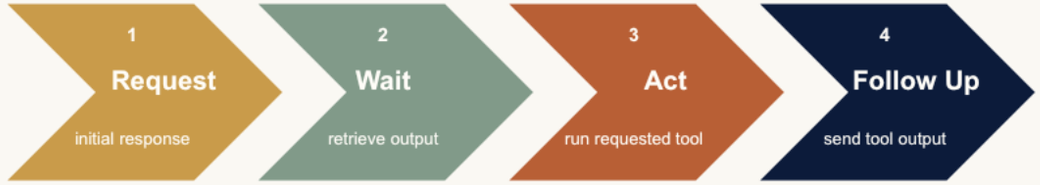
Upfront Set — BATCH

OpenAI directs upfront independent request sets to the Batch API.

02

Dependent Follow-Up

tool call | saved output | continuation



Preserve State — STATE

Save the original input and full output when later continuation may outlive retrieval.

Continue Workflow — HANDOFF

Bind the follow-up to the prior response and the correct function call output.

03

The Coordination Decision

Independence — SPLIT

Split only work the application knows can proceed independently.

Dependency — WAIT

Wait when the next step requires a prior result or tool output.

Batch Boundary — BATCH

Use Batch API when all independent requests can be submitted together up front.

Application Ownership — OWNER

Background mode does not create the plan or coordinate dependencies.

© 2026 Luminity Digital, Inc. | Independent synthesis of official OpenAI documentation.

Source: https://developers.openai.com/api/docs/guides/background · verified Nov 18, 2026

OpenAI Select Partner

Asynchronous Execution and Multi-Agent Work

Background mode controls request timing. Separate capabilities and products may coordinate subagents.

01

Background Mode

background=true	ASYNC RESPONSE
save response ID	RETRIEVAL KEY
webhook or polling	OBSERVE STATUS
terminal state	HANDLE OUTCOME
follow-up request	APPLICATION MANAGED

Lifecycle | Timing | Retrieval

02

Separate Orchestration

Responses Multi-agent | Codex subagents | Agent harness



RESPONSES MULTI-AGENT

A separate beta capability on supported models, enabled per request

CODEX DELEGATION

Triggered by a direct request or applicable project or skill instruction

03

Ownership Boundary

Background Mode — EXECUTION

Runs and exposes a Responses request asynchronously. It does not create the work plan.

Multi-Agent — ORCH

Separately enabled capabilities may delegate parallel, bounded work and synthesize results.

Application — AUTHORITY

Defines dependencies, tool handoffs, monitoring, validation, and downstream authority.

BOUNDARY

- Async ≠ decomposition
- Concurrent ≠ coordinated
- Reasoning ≠ multi-agent
- Tool output ≠ approval

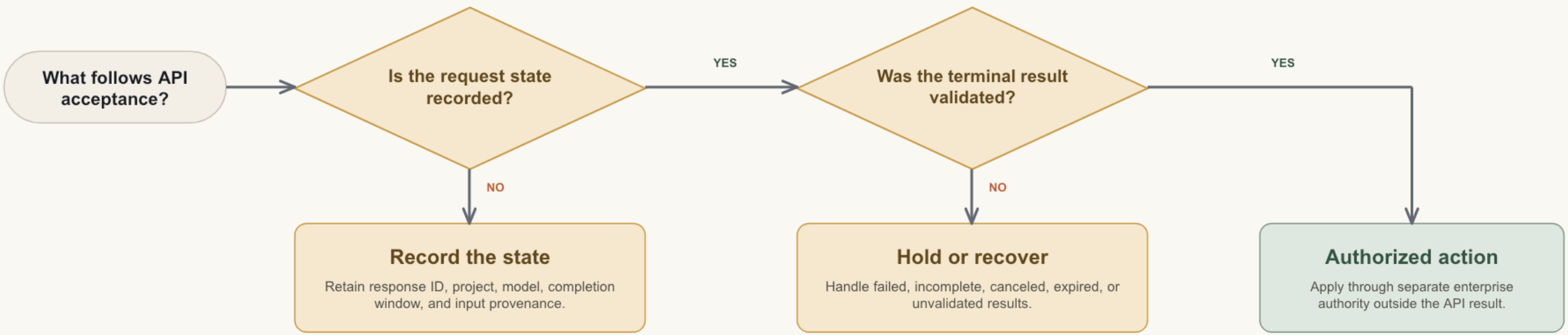
© 2026 Luminity Digital, Inc. | Independent synthesis of official OpenAI documentation.

Source: https://developers.openai.com/api/docs/guides/background · verified Nov 18, 2026

OpenAI Select Partner

Enterprise Control Around Asynchronous Work

Retain request state, handle every terminal outcome, validate results, and authorize consequences separately.



Luminity Digital
INTELLIGENCE, ENGINEERED, DELIVERED.

© 2026 Luminity Digital, Inc. | Independent synthesis of official OpenAI documentation.

Source: https://developers.openai.com/api/docs/guides/background · verified Nov 18, 2026

OpenAI Select Partner