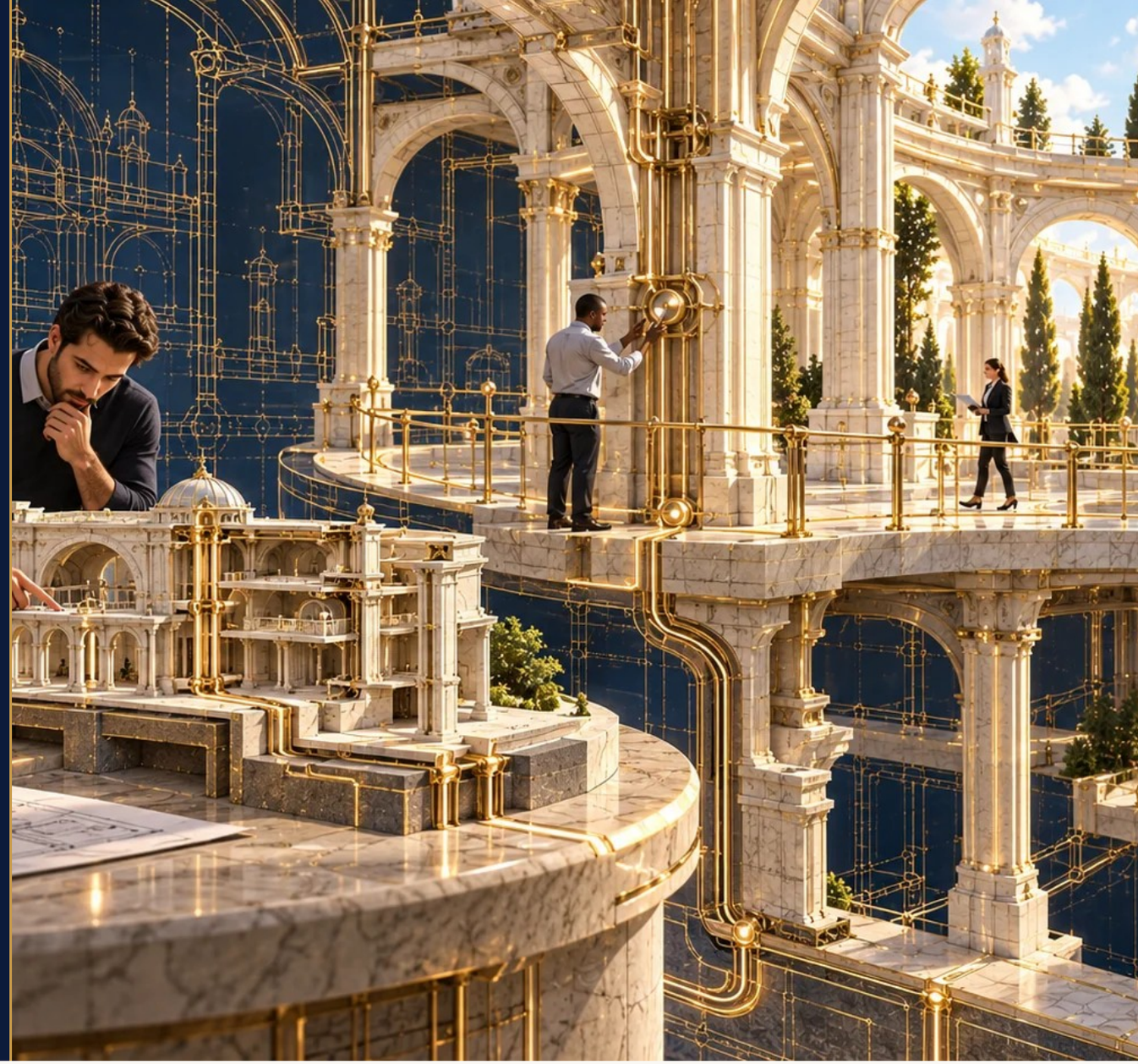
OPENAI ARCHITECTURE PRIMER

Safety Checks and Enforcement

Classifiers, project alerts, and organization notices

A visual guide to OpenAI checks, request effects, alert scopes, and the records application teams need to investigate.

FIRST-PARTY DOCUMENTATION / INDEPENDENT LUMINITY SYNTHESIS



Safety Signals Operate at Different Scopes

Request effects, project alerts, and organization cases create different integration duties.

AT A GLANCE

4
documented safety paths across the API

3
operating scopes: request + project + organization

2
webhook record types: alert + case

1
stable safety identifier per user and API surface

01
Request and Access Effects

Classifiers + Cyber

REQUEST

Request checks can delay, stop, or restrict access; high-capability cyber models add separate safeguards.

02
Alerts and Cases

Provider Records

RECORDS

Project misalignment alerts and organization enforcement cases require distinct retrieval and review paths.



IN THIS SET

Three chapters

01 **Request and Access Effects**
Classifiers, safety identifiers, cybersecurity safeguards, and access boundaries.

02 **Project Misalignment Monitoring**
Coverage, stopping behavior, alert delivery, and evidence preservation.

03 **Organization Enforcement**
Warning and deactivation cases, identifier mapping, and investigation routing.

Coverage and enforcement behavior can change. Recheck the live documentation before integration.

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

Source: https://developers.openai.com/api/docs/guides/safety-checks · verified Nov 5, 2026

OpenAI Select Partner

Request and Access Effects

Safety classifiers and cybersecurity safeguards can affect a response, a user identifier, or broader model access.

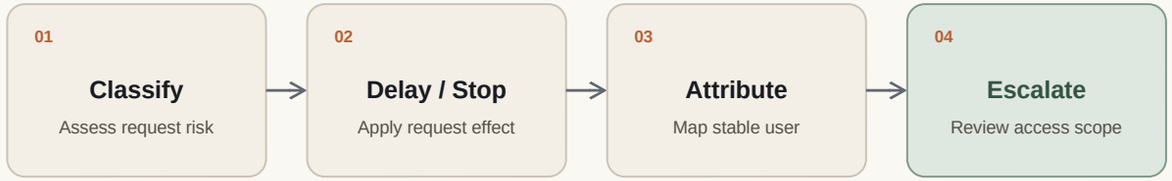
01

What the Application Receives

✓ GPT-5 safety classifiers	REQUEST THRESHOLDS
✓ Safety identifier	STABLE USER ATTRIBUTION
✓ GPT-5.6 Cyber + GPT-6 Cyber	ADDED SAFEGUARDS
! Cyber model ID alone	NO ACCESS GRANT
! Trusted Access for Cyber	DOES NOT IMPLY ZDR

02

How the Signal Moves



A request effect is not the same as an access decision. Access approval is bounded to the approved person or service, organization, project, model, and product surface.

General Classifier — REQUEST

Thresholds can delay or stop output and contribute to later access action.

Cyber Safeguards — MODEL

High and Critical cyber-capability models add documented API safeguards.

Safety Identifier — USER

Use a stable, non-personal value; hash any internal ID before sending.

Access Boundary — SCOPE

Daybreak approval and ZDR posture remain separate determinations.

CHECK

- Errors may arrive midstream
- IDs do not cross APIs or sessions
- Prevent replacement-account bypass
- Tool-action review remains separate

Luminity Digital
INTELLIGENCE, ENGINEERED. DELIVERED.

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

Source: https://developers.openai.com/api/docs/guides/safety-checks · verified Nov 5, 2026

OpenAI Select Partner

Project Misalignment Monitoring

Coverage depends on the API and continuity method. A flag requires review; it does not prove a policy violation.

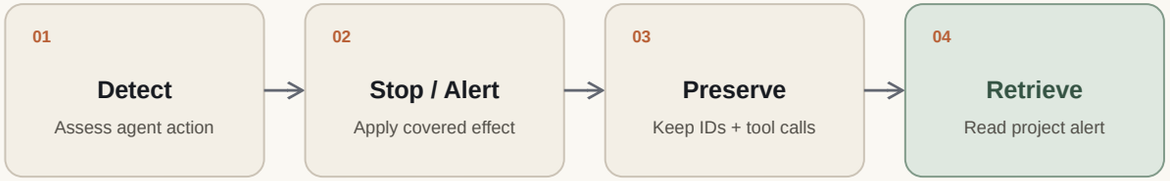
01

Coverage and Effect

✓ Responses + persisted reasoning / WebSockets / compaction	MONITOR + STOP
✓ Responses without those continuity methods	MONITOR + ALERT
! Chat Completions	NOT COVERED
! 403 misalignment_policy_violation	DO NOT RETRY
✓ A flagged conversation	REVIEW, NOT PROOF

02

Project Response Path



A stopped continuation has no general resume path. Preserve request and response IDs, tool calls, application state, and any side effects already completed.

Request Error — RUNTIME

A covered stop returns 403 with misalignment_policy_violation.

Project Webhook — EVENT

safety.alert.created announces a project-scoped alert.

Alert Detail — PROVIDER

Use the alert ID with api.safety.alerts.read to retrieve available detail.

Application State — OWNER

Reconcile earlier actions and decide the internal response.

LIMITS

- Stops do not undo actions
- Reason can be null under ZDR
- request_paused is not proof
- Alerts are not full audit history

Luminity Digital
INTELLIGENCE, ENGINEERED. DELIVERED.

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

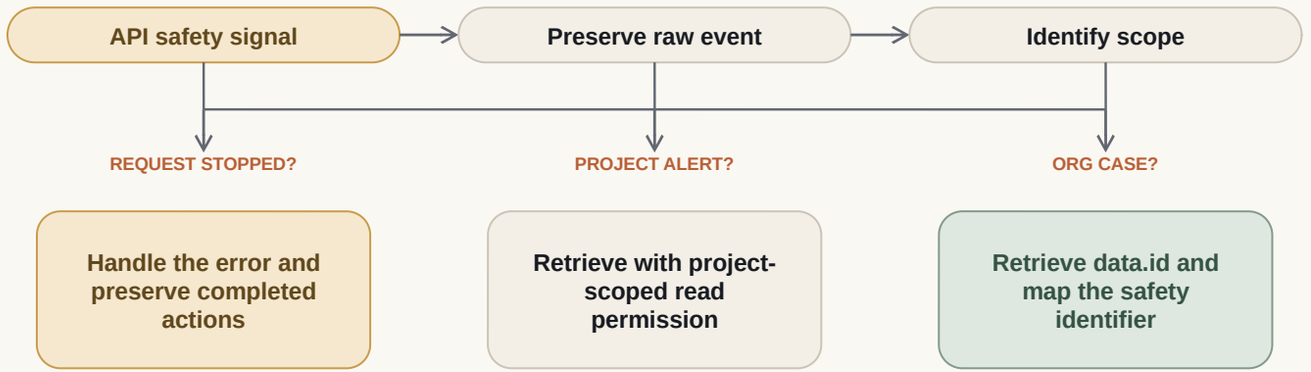
Source: https://developers.openai.com/api/docs/guides/safety-checks · verified Nov 5, 2026

OpenAI Select Partner

An Enforcement Evidence Path

Provider signals become usable only when the application preserves, retrieves, maps, and reconciles the right records.

START WITH THE SIGNAL



ORGANIZATION WEBHOOKS

Warning or deactivation

safety.warning_issued
safety.deactivation_issued

Event ID identifies delivery. data.id identifies the safety case used for lookup.

Request Record — RUNTIME

Error code, request and response IDs, tool calls, and completed actions.

Project Alert — PROJECT

Alert ID, affected project, available reason, and delivery metadata.

Organization Case — ORG

Case ID, notice type, entity identifier, and available enforcement context.

Application Review — OWNER

User mapping, side-effect reconciliation, decision, owner, and remediation.

EVIDENCE

- Provider signals are inputs
- Keep event and case IDs distinct
- Retrieval does not alter enforcement
- Application owns reconciliation

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

Source: https://developers.openai.com/api/docs/guides/safety-checks · verified Nov 5, 2026

OpenAI Select Partner