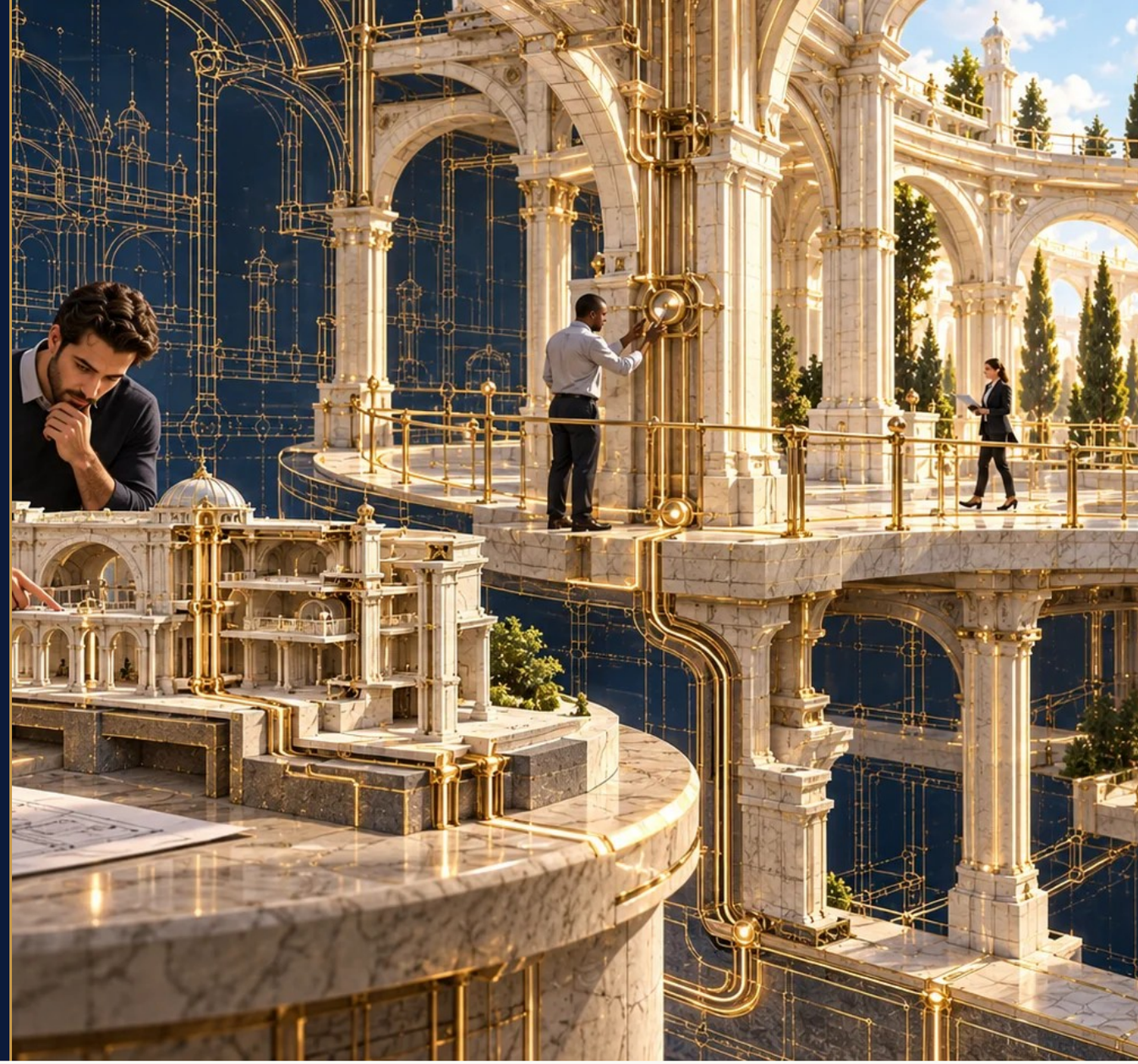
OPENAI ARCHITECTURE PRIMER

# Deployment Safety

Use case limits, adversarial tests, and release gates

A visual guide to the safety work application teams should complete before release and repeat as real usage changes.

FIRST-PARTY DOCUMENTATION / INDEPENDENT LUMINITY SYNTHESIS



Luminity Digital
INTELLIGENCE, ENGINEERED. DELIVERED.

**OpenAI Deployment Safety**

Source verified Nov 5, 2026

OpenAI Select Partner

# Deployment Safety Is a Maintained Release Decision

Application safeguards, evaluation evidence, and adversarial testing answer different parts of the release question.

AT A GLANCE

**3**
evidence layers:
safeguards + evals + red team

**2**
test lenses:
intended + adversarial behavior

**1**
stable safety identifier
per end user or session

**4**
events that reopen the
release decision

### Application Boundary
01

Define the users, inputs, outputs, oversight, reporting, and traceability controls around the model call.

### Release Evidence
02

Use representative evals and authorized adversarial testing, then preserve failures as regression cases.

IN THIS SET

## Three chapters

**01 Application boundary**

How moderation, constrained input and output, human review, reporting, and safety identifiers work together.

**02 Evaluation + red teaming**

How representative cases and adversarial probes produce different evidence before release.

**03 Release decision**

How fit, safeguards, evidence, operations, and change triggers become one maintained gate.

Models, moderation surfaces, red-teaming offerings, and safety guidance can change. Recheck the live OpenAI documentation before deployment.



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

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

OpenAI Select Partner

# Application Safety Boundary

Identity, constrained interaction, moderation, human review, and reporting reduce misuse opportunities around the model call.

01

## What belongs in the boundary?

| | | |
|---|---|---|
| ✓ | Moderation or tailored filtering | APPLICATION CONTROL |
| ✓ | Human review for high-stakes use | RECOMMENDED |
| ✓ | Constrained inputs and outputs | REDUCES MISUSE |
| ✓ | Stable safety identifier | RECOMMENDED |
| ! | Login or registration | USE-CASE DEPENDENT |

Known user or session → Traceable application event

02

## How the controls surround a request

Identity · user or session | Boundary · input + output | Oversight · review + report

1 **Identify** — Attach a stable user or session identifier

2 **Constrain** — Validate inputs and bound generated output

3 **Moderate** — Apply safety checks around generation

4 **Review** — Verify consequential use and accept reports

**TRACE**
**Hash user identity; use a stable session ID for logged-out use**

**LIMIT**
**Narrow input and output ranges when the workflow permits**

**VERIFY**
**Give reviewers the source material and limits needed to check output**

03

## Four application control surfaces

### Identity and Trace — USER

Use a privacy-preserving safety identifier to trace individual misuse without sending personal data.

### Input Boundary — INPUT

Prefer validated fields and trusted sources when open-ended input adds unnecessary risk.

### Output Boundary — OUTPUT

Limit output and use validated backend material when novel generation is not required.

### Human Oversight — REVIEW

Review consequential outputs and keep a monitored path for user-reported problems.

**BOUNDARY**

- Safeguards reduce risk
- Safety IDs do not replace access control
- Human review needs source context
- Communicate system limits



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

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

OpenAI Select Partner

# Evaluation and Red Teaming

Evals measure intended behavior. Red teaming probes adversarial, abusive, and unexpected behavior before deployment.

01

## What evidence does each method produce?

| | |
|---|---|
| ✓ Representative tasks | EVALUATION |
| ✓ Known edge cases | EVALUATION |
| ✓ Misuse and high-risk interactions | RED TEAM |
| ! Third-party assets without permission | DO NOT SUBMIT |
| ✓ OpenAI managed red teaming | ENTERPRISE |

Observed failure → Regression evidence

02

## How a finding becomes release evidence

Scope · owned or authorized | Cases · expected + adversarial | Record · trigger + consequence

1 **Define** — Name success, misuse, assets, and consequences

2 **Build** — Combine expected, edge, and adversarial cases

3 **Run** — Test the complete application and connected actions

4 **Repair** — Change the control and rerun the failed case

**EVAL**
**Measures whether outputs, tool choices, and handoffs meet criteria**

**RED TEAM**
**Probes misuse, prompt injection, and unexpected behavior**

**EVIDENCE**
**Keeps the failure, repair, and passing retest together**

03

## Four evidence surfaces

### Evaluation Set — INTENT
Representative tasks, expected outputs, and edge cases define intended performance.

### Adversarial Cases — MISUSE
Misuse attempts and prompt injections expose behavior ordinary quality tests may miss.

### Authorized Scope — AUTHORITY
Only test code and assets you own or have explicit permission to submit.

### Regression Record — RESULT
Preserve each material failure, control change, and passing retest as release evidence.

**EVIDENCE RULE**

- Evals and red teams answer different questions
- Authorization limits the test scope
- A finding needs repair and retest

Luminity Digital
INTELLIGENCE, ENGINEERED. DELIVERED.

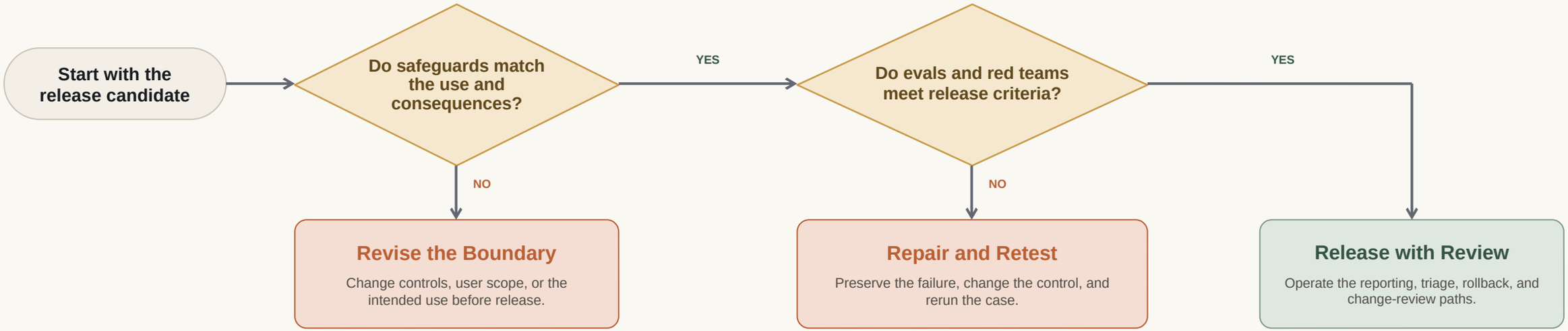
© 2026 Luminity Digital, Inc. | Independent synthesis of official OpenAI documentation.

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

OpenAI Select Partner

# A Maintained Deployment Decision

Release only when the application boundary, evidence, and operating response match the workload consequences.




## LIVE EVIDENCE REOPENS THE RELEASE DECISION

### User Reports — SIGNAL

Investigate credible reports of harmful, incorrect, or easily manipulated behavior.

### Observed Failures — CASE

Reproduce production failures and add them to the evaluation set.

### Material Changes — CHANGE

Retest when a model, prompt, data source, tool, or policy changes behavior.

### New Misuse — THREAT

Expand adversarial cases when users find new ways around existing controls.

**ARCHITECTURE RULE**

- A release gate is not certification
- Live evidence can reopen it
- Material changes require retesting
- Keep the response path staffed

Luminity Digital
INTELLIGENCE, ENGINEERED. DELIVERED.

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

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

OpenAI Select Partner