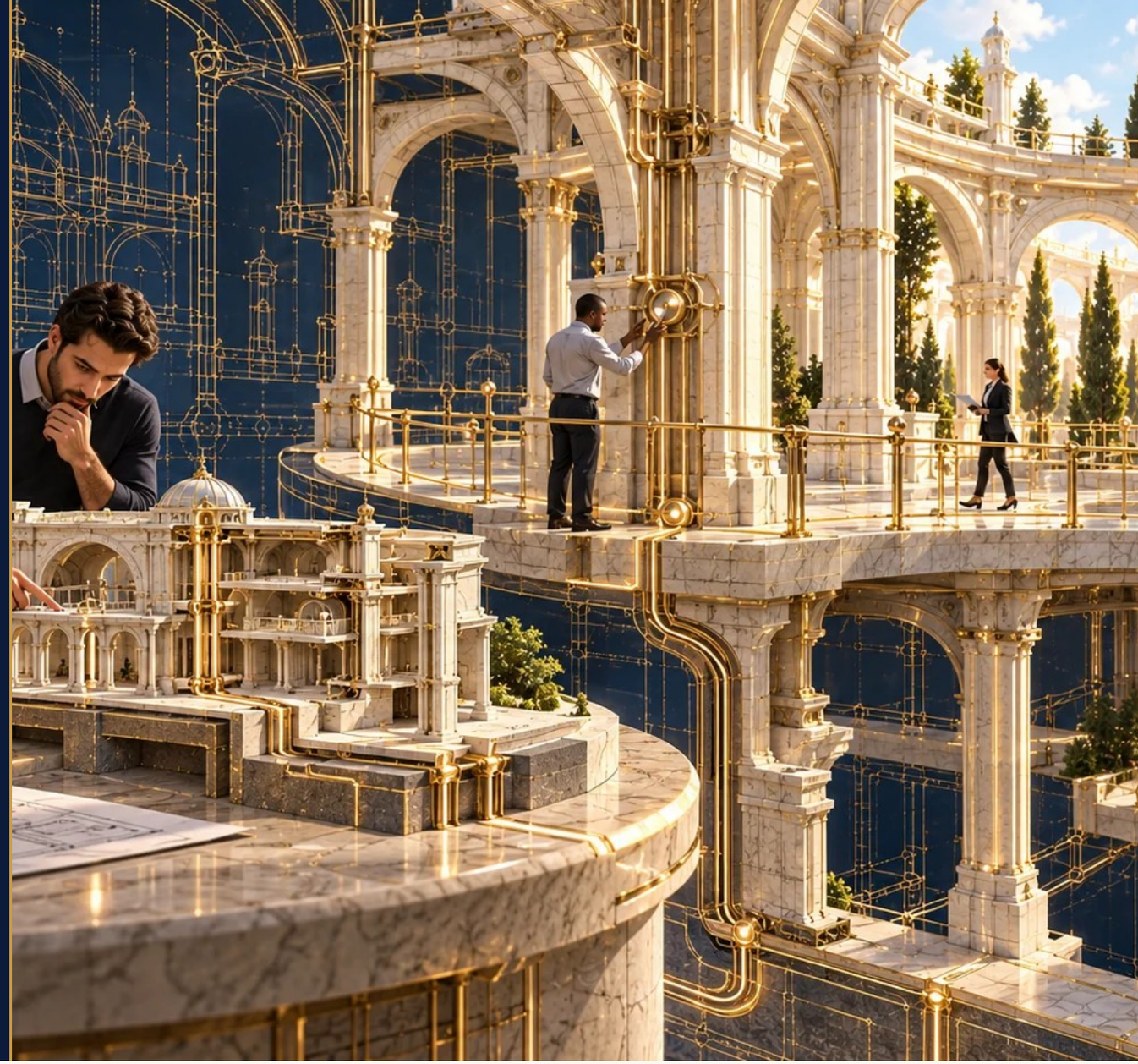
OPENAI ARCHITECTURE PRIMER

Audience and Content Safeguards

Minors, child safety, and content provenance

A visual guide to audience-specific duties, CSAM response operations, and the limited provenance signals available for media.

FIRST-PARTY DOCUMENTATION / INDEPENDENT LUMINITY SYNTHESIS



Three Safeguard Domains, Three Different Decisions

Audience duties, child-safety operations, and provenance evidence answer different questions.

AT A GLANCE

3
specialized domains:
minors + CSAM + provenance

4
under-18 safeguard
classes

3
CSAM phases:
prevent + detect + respond

2
provenance signal
families: C2PA + SynthID

01
Audience and Child Safety

Minors + CSAM — PEOPLE

Under-18 conditions and CSAM operations require purpose-built safeguards around the API.

02
Content Signals

Moderation + Provenance — SIGNALS

Moderation and provenance expose bounded signals; neither substitutes for a governed decision.

Identify audience → Apply safeguard → Retain decision

IN THIS SET

Three chapters

01 **Applications Serving Minors**
Data conditions, age-appropriate safeguards, monitoring, and age assurance.

02 **CSAM Prevention and Response**
Product rules, dedicated detection, reporting readiness, and reviewer support.

03 **Content Provenance Signals**
C2PA, SynthID, supported media, result interpretation, and evidence limits.

Legal duties, API coverage, and supported signals can change. Recheck the live documentation before deployment.

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

Source: https://developers.openai.com/api/docs/guides/safety-checks/under-18-api-guidance · verified Nov 5, 2026

OpenAI Select Partner

Applications Serving Minors

The application owner controls the data, content, monitoring, escalation, and access decisions around API use.

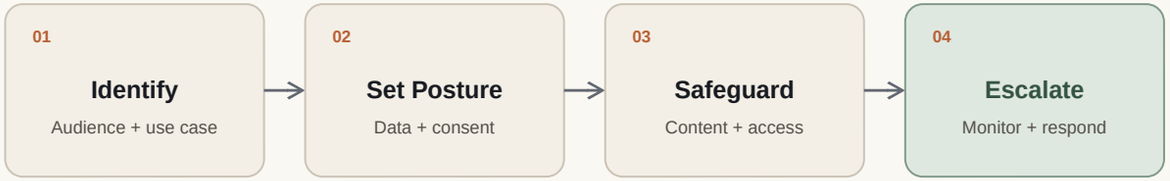
01

Audience Conditions

✓ Application serves minors	HEIGHTENED CARE
! Under 13 or applicable digital-consent age	ZDR BEFORE PERSONAL DATA
✓ Age-appropriate disclosures	EXPLAIN AI USE
✓ Age-appropriate content filters	ADDRESS SENSITIVE CONTENT
✓ Monitoring, reporting, and escalation	ACT ON HIGH-RISK USE

02

Safeguard Path



Use age assurance where required or otherwise appropriate for the use case. The organization remains responsible for applicable law.

Disclosure — INFORM

Explain the AI system and responsible use in language appropriate for minors.

Content Filtering — FILTER

Apply age-appropriate filters and test likely failure modes.

Monitoring + Reporting — RESPOND

Maintain monitored queues and escalation paths staff can act on.

Age Assurance — ACCESS

Control product access when law or the use case requires it.

BOUNDARY

- Use current flagship models
- ZDR condition is audience-specific
- Application owns compliance
- Architectural guidance, not legal advice

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

Source: https://developers.openai.com/api/docs/guides/safety-checks/under-18-api-guidance · verified Nov 5, 2026

OpenAI Select Partner

CSAM Prevention and Response

Products that accept uploads or enable interaction need a dedicated child-safety operating model.

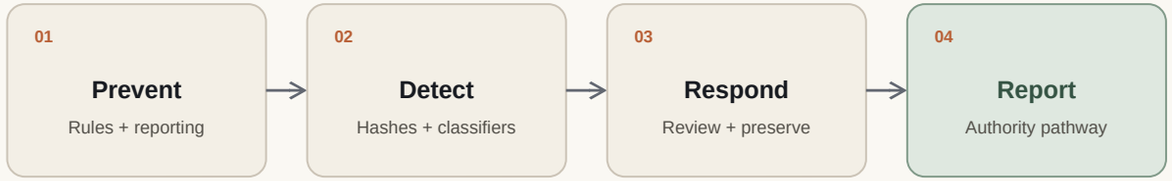
01

Operating Requirements

- ✓ Clear rules + visible user reporting — PREVENT
- ✓ Every upload linked to a user — ATTRIBUTE
- ✓ Hash matching + novel-content classifiers — DETECT
- ✓ Repeat-actor detection + enforcement — DISRUPT
- ✓ Authority reporting + playbook + reviewer support — RESPOND

02

Child-Safety Path



Preserve useful records and identifiers, define who can decide and escalate, and support the humans reviewing harmful material.

CSAM Material — DO NOT ROUTE

Do not send known or suspected CSAM to the Moderation API; that endpoint is not designed for handling it.

sexual/minors — TEXT ONLY

The moderation category covers sexual content involving people under 18 and is text-only.

Moderation Result — LIMITED

A signal can support blocking, review, account action, or friction; it is not dedicated detection.

Human Review — PEOPLE

Trained reviewers, escalation routes, and wellness support remain necessary.

OPERATING RULE

- Keep usable identifiers
- Reporting duties vary by jurisdiction
- Moderation is not a substitute
- Support reviewer well-being

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/under-18-api-guidance · verified Nov 5, 2026

OpenAI Select Partner

A Bounded Provenance Decision

The API checks supported OpenAI signals in images and audio; each result must be interpreted independently.

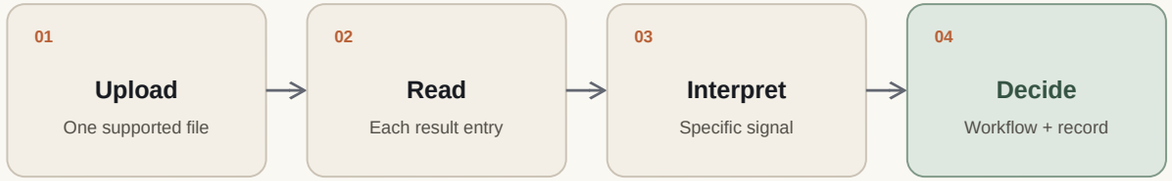
01

Select the Check by Content Type

✓ Image upload	C2PA + SYNTHID
✓ Audio upload	SYNTHID
! Text verification	APPROVED ACCESS ONLY
! Unsupported or blocked file	400
! Organization without access	404

02

Evidence Path



There is no top-level outcome. Read each applicable entry, its issuer or validation state, and any available model or generation time.

C2PA Result — IMAGE

A detected result requires a trusted or valid OpenAI-issued manifest with an AI-generation action.

SynthID Result — MEDIA

Detected means the verifier found a supported watermark in an image or audio file.

not_detected — LIMIT

It means no supported signal was found; it does not prove human origin or rule out AI modification.

Restricted Text Access — ELIGIBILITY

Text verification is reviewed case by case for approved organizations, including research and academia.

EVIDENCE RULE

- Read entries independently
- Check C2PA issuer + validation
- Signals can be removed or degraded
- Retain an accountable decision

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/under-18-api-guidance · verified Nov 5, 2026

OpenAI Select Partner