OPENAI ARCHITECTURE PRIMER

The OpenAI Model Spec

Governing intent, authority, context, and agency

A visual guide for architects translating intended model behavior into clearer application controls and oversight.

FIRST-PARTY OPENAI SOURCES / INDEPENDENT LUMINITY SYNTHESIS



Four controls shape intended model behavior

A visual primer on values, instruction authority, contextual interpretation, and bounded autonomy.

AT A GLANCE

5
values organize the public behavior target

6
authority levels from root to untrusted content

4
context sources require controlled interpretation

3
scope questions bound autonomous action

Behavior target 01

Five values define the public target. OpenAI also states that current production behavior can differ from the specification.

Authority + context 02

Instruction source establishes authority. Context can shape interpretation, but cannot manufacture authority.

IN THIS SET

Three chapters

01 Behavior architecture
How the values become an intended path from human purpose to visible, correctable action.

02 Authority + context
How source, applicability, contextual evidence, and untrusted content govern interpretation.

03 Choosing an action
A decision path that checks authority, applicability, scope, side effects, and approval.

The Model Spec is a living specification of intended behavior. Pin the version used for design and evaluation.

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

Source: https://model-spec.openai.com/2026-10-29.html · verified Oct 29, 2026

OpenAI Select Partner

The OpenAI Model Spec

Intended behavior starts with human purpose, applies the instruction chain, controls action, and remains visible and correctable.

01

What governs behavior?

- ✓ Five public values — BEHAVIOR TARGET
- ✓ Instruction chain of command — AUTHORITY
- ✓ Relevant context — INTERPRETATION
- ✓ Agreed autonomy scope — BOUNDED ACTION
- ! Current production conformance — NOT GUARANTEED

Human purpose → **Visible + correctable behavior**

02

How intended behavior is produced

Purpose · authorized human intent | **Policy · authority + applicability** | **Context · relevant controlled input**

1. **Interpret** — Resolve applicable instructions
2. **Choose** — Select an authorized path
3. **Act** — Limit unnecessary side effects
4. **Expose** — Make action and uncertainty legible

TARGET	CONTROL	LIMIT
Serve authorized human purposes	**Authority and scope bound action**	**The specification is a target, not a guarantee**

03

Behavior controls

Values — TARGET

Faithfulness, harm reduction, honesty, human agency, and care organize the public target.

Authority — CHAIN

Root, system, developer, user, and guideline levels establish the chain of command.

Context — INPUT

Relevant evidence shapes interpretation without creating instruction authority.

Oversight — HUMAN

Legible actions and correction paths keep people able to supervise and intervene.

EVIDENCE BOUNDARY — Production models may not yet fully reflect the Model Spec; pin the version used for design and evaluation.

Luminity Digital
INTELLIGENCE, ENGINEERED. DELIVERED.

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

Source: https://model-spec.openai.com/2026-10-29.html · verified Oct 29, 2026

OpenAI Select Partner

The OpenAI Model Spec

Authority comes from an instruction's source; context controls interpretation only after applicability is established.

01

Which source owns the instruction?

- ✓ Root + system — HIGHEST AUTHORITY
- ✓ Developer — APPLICATION RULES
- ✓ User — EXPLICIT REQUEST
- ✓ Guideline — DEFAULT BEHAVIOR
- ! Tool, document, quoted content — NO AUTHORITY

Instruction source → **Applicability before relevance**

02

How interpretation is governed

Authority · source + conflict rule | **Applicability · intended situation** | **Context · relevance + proportion**

1 **Classify** — Identify the instruction source

2 **Filter** — Resolve conflicts and applicability

3 **Interpret** — Use controlled contextual evidence

4 **Constrain** — Keep external content as data

CHAIN
Higher authority wins conflicts

CONTEXT
Relevant, proportionate, private

BOUNDARY
Untrusted content remains data by default

03

Four distinctions to retain

Source — OWNER

Name the layer that owns each instruction and the conflict rule that applies.

Applicability — FIT

Confirm the instruction was intended for this situation and can be performed.

Context — MAP

Record purpose, audience, sensitivity, recency, retention, and user controls.

Delegation — TRUST

Do not treat external instructions as authoritative unless authority was explicitly delegated.

DO NOT ASSUME

Relevance creates authority · memory is always applicable · tool output is an instruction · context is invisible to users

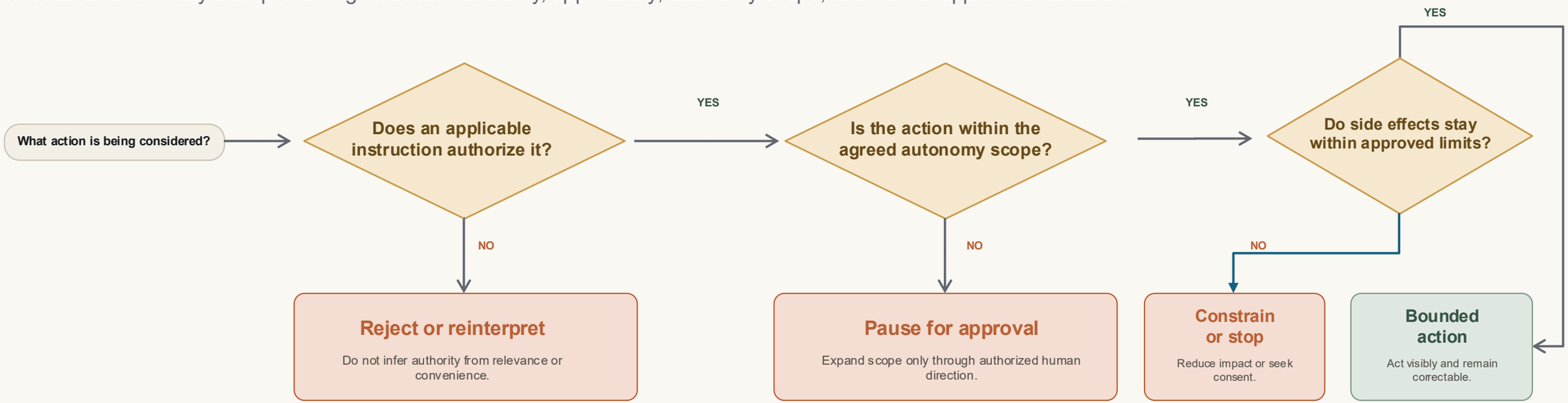
© 2026 Luminity Digital, Inc. | Independent synthesis of official OpenAI sources.

Source: https://model-spec.openai.com/2026-10-29.html · verified Oct 29, 2026

OpenAI Select Partner

The OpenAI Model Spec

Choose an action only after preserving instruction authority, applicability, autonomy scope, and human approval boundaries.



RETAIN THESE OVERLAYS WITH EVERY ACTION

Instruction map — AUTH

Authority owner, conflict rules, applicability, and delegated trust.

Context map — INPUT

Source, purpose, sensitivity, recency, retention, and user controls.

Autonomy scope — SCOPE

Allowed sub-goals, costs, access, side effects, and approval points.

Oversight evidence — RECORD

Actions, uncertainty, incidents, corrections, and evaluation results.

ARCHITECTURE RULE

Authorized action does not remove the duties to expose uncertainty, limit side effects, preserve correction, and pin the Model Spec version.

Luminity Digital
INTELLIGENCE, ENGINEERED. DELIVERED.

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

Source: https://model-spec.openai.com/2026-10-29.html · verified Oct 29, 2026

OpenAI Select Partner