OPENAI SAFETY + ASSURANCE READOUT

Embedded Assessments Inside the Lab

Four assessors around one evidence system

A visual readout of OpenAI’s first phase of embedded assessments and the architecture needed to preserve independent evidence.

OFFICIAL OPENAI SOURCE / INDEPENDENT LUMINITY SYNTHESIS



Luminity Digital
INTELLIGENCE, ENGINEERED. DELIVERED.

© 2026 Luminity Digital, Inc. All rights reserved in the original analysis, selection, arrangement, visual design, and commentary.

Source: https://openai.com/index/our-first-phase-of-embedded-assessments/ · verified Nov 16, 2026

OpenAI and related marks are trademarks of OpenAI, L.L.C. Independent analysis of public information. Not sponsored or endorsed by OpenAI.

Apollo Research, Goodfire, METR, and Transluce names and marks are trademarks or property of their respective companies.

OpenAI Select Partner

EMBEDDED ASSESSMENTS · A READOUT

Embedded Assessments in One View

Deep access lets independent assessors test different claims while protocols define how the evidence can travel.

FOUR ASSESSMENT LANES

01	02	03	04
Apollo Research	**Goodfire**	**METR**	**Transluce**
subversion	interpretability	alignment + monitors	evaluation quality

Embedded Access

METHOD

Models, checkpoints, evaluation data, monitors, documentation, and staff context.

Protocol Layer

BOUNDARY

Scope, secure access, oversight, disputes, conflicts, and publication rights.

OpenAI describes signed agreements and planned assessments. The work should be read as a portfolio, not one completed audit.

IN THIS READOUT

Key Points

01 Access shapes the evidence

Internal artifacts and staff context let assessors pursue claims that public testing cannot reach.

02 Methods answer different questions

Behavior, interpretability, monitor stress testing, and meta-evaluation contribute distinct evidence.

03 Limitations stay attached

Scope, denied access, disclosure delays, and unresolved disputes affect how much confidence a report can carry.

OpenAI expects the protocols to evolve. Planned work and intended follow-ons are not completed results.

Luminity Digital
INTELLIGENCE, ENGINEERED. DELIVERED.

© 2026 Luminity Digital, Inc. All rights reserved in the original analysis, selection, arrangement, visual design, and commentary.

Source: https://openai.com/index/our-first-phase-of-embedded-assessments/ · verified Nov 16, 2026

OpenAI and related marks are trademarks of OpenAI, L.L.C. Independent analysis of public information. Not sponsored or endorsed by OpenAI.

Apollo Research, Goodfire, METR, and Transluce names and marks are trademarks or property of their respective companies.

OpenAI Select Partner

Evidence Access and Confidence

Embedded access reaches internal evidence while protocol limits determine what conclusions can support.

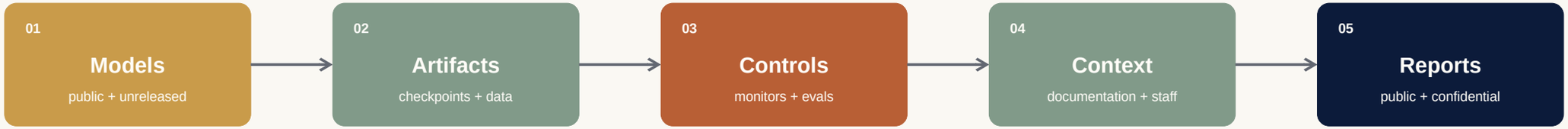
01

Evidence Depth

- Public behavior
- Embedded access
- Working context

02

What Assessors Can Inspect



03

Confidence Conditions

Scope — BOUNDARY

The system, claim, duration, and resources remain explicit.

Access Record — METHOD

Granted, denied, and substituted access remains visible.

Disclosure — PUBLIC

Redactions and responsible disclosure delays travel with the report.

Oversight — SSC

Unredacted reporting and disputes can reach the Safety and Security Committee.

EVIDENCE RULE

More access can deepen evidence. It does not remove the need to document limits.

Luminity Digital
INTELLIGENCE, ENGINEERED. DELIVERED.

© 2026 Luminity Digital, Inc. All rights reserved in the original analysis, selection, arrangement, visual design, and commentary.

Source: https://openai.com/index/our-first-phase-of-embedded-assessments/ · verified Nov 16, 2026

OpenAI and related marks are trademarks of OpenAI, L.L.C. Independent analysis of public information. Not sponsored or endorsed by OpenAI.

Apollo Research, Goodfire, METR, and Transluce names and marks are trademarks or property of their respective companies.

OpenAI Select Partner

Four Assessors, Four Evidence Questions

Each engagement applies a different method to a different layer of the same safety system.

01

Assessment Portfolio

✓	Apollo Research	SUBVERSION CLAIMS
✓	Goodfire	MODEL INTERNALS
✓	METR	ALIGNMENT + MONITORS
✓	Transluce	EVALUATION QUALITY
!	One combined assurance grade	NOT DEFINED

02

Distinct Methods



PORTFOLIO LOGIC

Confidence grows when distinct methods test distinct claims and their limits remain visible.

03

Questions the Portfolio Leaves Open

Coverage — SCOPE

Which safety claims remain outside these four engagements?

Aggregation — METHOD

How should different evidence types combine without hiding disagreement?

Timeliness — ACTION

How quickly can findings influence training, safeguards, and release decisions?

Continuity — PROGRAM

What turns a first phase into sustained independent assessment?

STATUS OpenAI signed four agreements. The described assessments and follow-ons remain at different stages.

Luminity Digital
INTELLIGENCE, ENGINEERED. DELIVERED.

© 2026 Luminity Digital, Inc. All rights reserved in the original analysis, selection, arrangement, visual design, and commentary.

Source: https://openai.com/index/our-first-phase-of-embedded-assessments/ · verified Nov 16, 2026

OpenAI and related marks are trademarks of OpenAI, L.L.C. Independent analysis of public information. Not sponsored or endorsed by OpenAI.

Apollo Research, Goodfire, METR, and Transluce names and marks are trademarks or property of their respective companies.

OpenAI Select Partner

Goodfire and Interpretability Inside Training

Goodfire will analyze model internals for training-run monitoring, debugging, and alignment work.

01

Interpretability Path



02

What This Method Adds

Earlier Signal

TIMING

Potential evidence can surface during training rather than after deployment.

Explanatory Leverage

MECHANISM

Internal patterns can give researchers a target for deeper investigation.

Research Feedback

ACTION

The primary audience is OpenAI researchers who can use findings in safety work.

03

Evidence Boundary

✓ Interpretability signal — EVIDENCE

! Proof of alignment by itself — NOT ESTABLISHED

ASSURANCE RULE

Interpretability evidence strengthens a case when it connects to behavior, controls, and remediation.

Luminity Digital
INTELLIGENCE, ENGINEERED. DELIVERED.

© 2026 Luminity Digital, Inc. All rights reserved in the original analysis, selection, arrangement, visual design, and commentary.

Source: https://openai.com/index/our-first-phase-of-embedded-assessments/ · verified Nov 16, 2026

OpenAI and related marks are trademarks of OpenAI, L.L.C. Independent analysis of public information. Not sponsored or endorsed by OpenAI.

Apollo Research, Goodfire, METR, and Transluce names and marks are trademarks or property of their respective companies.

OpenAI Select Partner

Apollo Research and Subversion Claims

Apollo will test whether models could work against developer intent and whether safeguards would detect it.

01

Four Claims Across the Lifecycle



02

First-Stage Evidence

Models and Checkpoints — SYSTEM

Selected public and unreleased models plus training checkpoints.

Training and Evaluations — RECORD

Relevant data, results, intervention records, and independent testing.

Monitoring Evidence — CONTROL

Monitoring systems and flagged behaviors tied to the claims.

Research Context — CONTEXT

Researchers, tools, and compute needed to interpret the evidence.

03

Publication Boundary

Independent findings

Confidential evidence

Redactions + delays

Preserving what Apollo could and could not inspect is part of the evidence package.

STATUS

OpenAI describes a pilot assessment and the first-stage access it plans to provide.

Luminity Digital
INTELLIGENCE, ENGINEERED. DELIVERED.

© 2026 Luminity Digital, Inc. All rights reserved in the original analysis, selection, arrangement, visual design, and commentary.

Source: https://openai.com/index/our-first-phase-of-embedded-assessments/ · verified Nov 16, 2026

OpenAI and related marks are trademarks of OpenAI, L.L.C. Independent analysis of public information. Not sponsored or endorsed by OpenAI.

Apollo Research, Goodfire, METR, and Transluce names and marks are trademarks or property of their respective companies.

OpenAI Select Partner

METR and Missed-Failure Discovery

METR separates a search for misaligned behavior from a test of whether monitoring controls can detect it.

01

Two Assessment Tracks



02

Evidence and Access

Run Evidence — SYSTEM

Relevant checkpoints plus training and evaluation rollouts.

Misalignment Records — RECORD

Monitoring records and evidence tied to the selected run.

Prior Exercise — EXERCISE

Three weeks with codebase, monitors, trajectories, and staff discussions.

Independent Iteration — METHOD

METR sets priorities and iterates within agreed scope and access.

03

Scope and Status

✓ Narrower first-phase alignment scope — DESCRIBED

! Broader monitor follow-on scope — NOT FINAL

EVIDENCE RULE Discovery asks what failed. Monitor testing asks whether controls would have seen it.

Luminity Digital
INTELLIGENCE, ENGINEERED. DELIVERED.

© 2026 Luminity Digital, Inc. All rights reserved in the original analysis, selection, arrangement, visual design, and commentary.

Source: https://openai.com/index/our-first-phase-of-embedded-assessments/ · verified Nov 16, 2026

OpenAI and related marks are trademarks of OpenAI, L.L.C. Independent analysis of public information. Not sponsored or endorsed by OpenAI.

Apollo Research, Goodfire, METR, and Transluce names and marks are trademarks or property of their respective companies.

OpenAI Select Partner

Transluce and Evaluation Quality

Transluce will examine whether post-training alignment evaluations have sufficient quality and coverage.

01

Evaluation Assurance Path



02

What This Method Tests

Evaluation Inputs — SYSTEM

Prompts, scenarios, graders, results, and associated documentation.

Quality — METHOD

Whether the evaluation design can support the claims placed on it.

Coverage — SCOPE

Whether material failure modes sit outside the current testing surface.

Oversight — SSC

Unredacted reporting can reach the Safety and Security Committee.

03

The Question Changes

✓ Did the model pass the current evaluation set? — RESULT

! Was the evaluation good enough to find material failure? — META-EVALUATION

STATUS A broader follow-on is intended for early 2027. It is not a completed result.

Luminity Digital
INTELLIGENCE, ENGINEERED. DELIVERED.

© 2026 Luminity Digital, Inc. All rights reserved in the original analysis, selection, arrangement, visual design, and commentary.

Source: https://openai.com/index/our-first-phase-of-embedded-assessments/ · verified Nov 16, 2026

OpenAI and related marks are trademarks of OpenAI, L.L.C. Independent analysis of public information. Not sponsored or endorsed by OpenAI.

Apollo Research, Goodfire, METR, and Transluce names and marks are trademarks or property of their respective companies.

OpenAI Select Partner

Protocol Architecture for Independence

Independence depends on the operating agreement, access path, oversight, and publication rules.



02 · Iteration

METHOD

Assessors can refine methods and propose further assessments.

03 · Attestation

RECORD

OpenAI can verify, correct, clarify, or decline to attest statements.

04 · Disputes

OVERSIGHT

Unresolved matters can escalate to the Safety and Security Committee.

05 · Conflicts

INTEGRITY

Disclosures, mitigations, recusals, and cool-off periods protect integrity.

BOUNDARY

The architecture must preserve access depth and the assessor's ability to state where access stopped.

© 2026 Luminity Digital, Inc. All rights reserved in the original analysis, selection, arrangement, visual design, and commentary.

Source: https://openai.com/index/our-first-phase-of-embedded-assessments/ · verified Nov 16, 2026

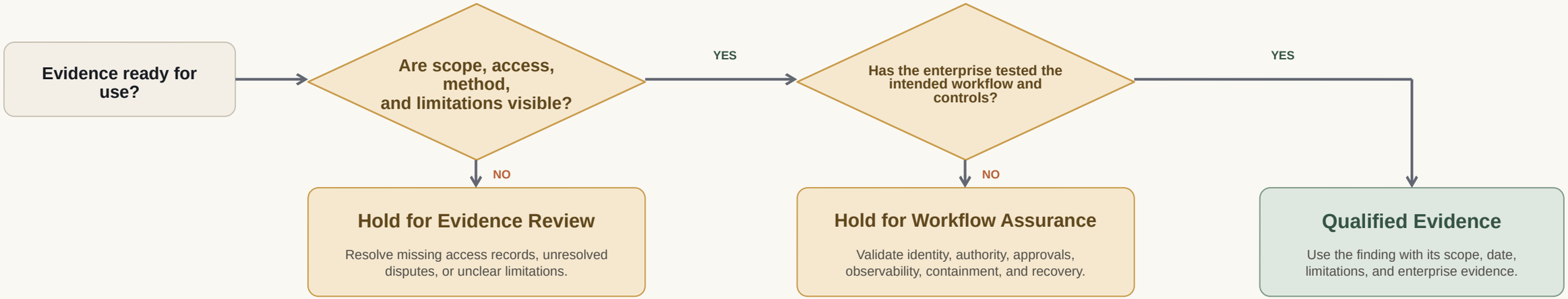
OpenAI and related marks are trademarks of OpenAI, L.L.C. Independent analysis of public information. Not sponsored or endorsed by OpenAI.

Apollo Research, Goodfire, METR, and Transluce names and marks are trademarks or property of their respective companies.

OpenAI Select Partner

Luminity Synthesis — Enterprise Assurance Boundary

Embedded assessment can strengthen model-side evidence. Enterprise workflow accountability remains with the enterprise.



LUMINITY EVIDENCE-INTAKE CHECKLIST

Claim and Scope — CLAIM

Identify the claim, system boundary, assessment date, and intended use.

Access and Method — METHOD

Record what the assessor inspected, requested, and could not access.

Findings and Limits — EVIDENCE

Keep limitations, redactions, delays, disputes, and open questions attached.

Enterprise Proof — CONTROL

Retain independent evidence that the governed workflow behaves as intended.

CLAIM BOUNDARY

Embedded assessment informs enterprise assurance. It does not transfer enterprise accountability.

Luminity Digital
INTELLIGENCE, ENGINEERED. DELIVERED.

© 2026 Luminity Digital, Inc. All rights reserved in the original analysis, selection, arrangement, visual design, and commentary.

Source: https://openai.com/index/our-first-phase-of-embedded-assessments/ · verified Nov 16, 2026

OpenAI and related marks are trademarks of OpenAI, L.L.C. Independent analysis of public information. Not sponsored or endorsed by OpenAI.

Apollo Research, Goodfire, METR, and Transluce names and marks are trademarks or property of their respective companies.

OpenAI Select Partner