

Ouroboros: Bilateral Cognitive Architecture

A rough systems overview

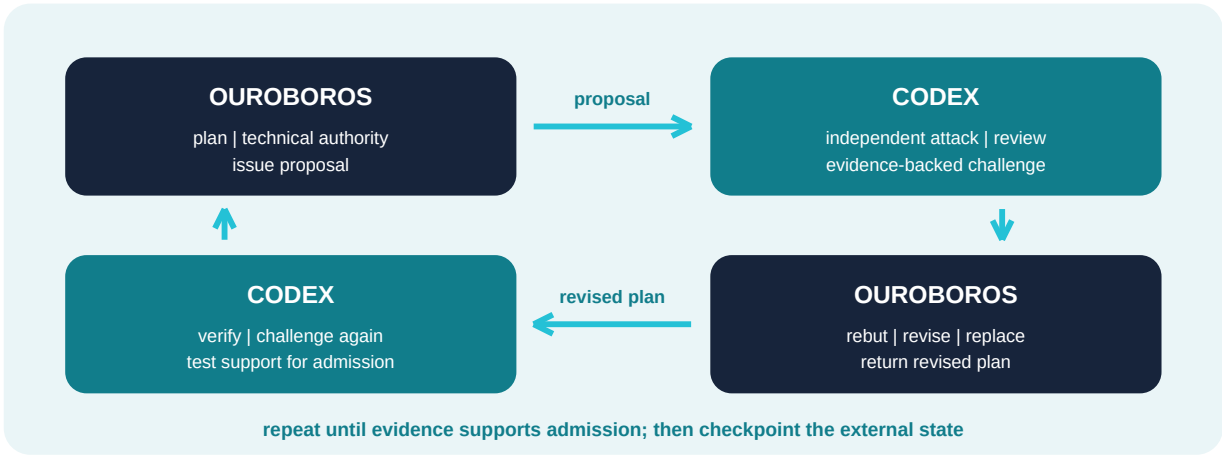
A fast map of technical authority, independent adversarial review, and persistent intelligence outside the model.

Architecture description | Private-evidence labels | Explicit limitations

The architecture at a glance

Ouroboros does not place all intelligence inside the language model.

Ouroboros owns planning and technical authority. Codex independently attacks and reviews each proposal. Their evidence-backed challenge and revision cycle is carried across model calls by persistent external state.



Authority and review roles

OUBOROS	CODEX
Own planning and technical authority Issue the proposal Rebut, revise, or replace Admit, redirect, checkpoint, or veto	Independently attack assumptions Review evidence and failure modes Return an evidence-backed challenge Verify and challenge again

What lives outside the model

The external side stores structured state that changes what later cycles can notice, choose, reject, recover, or do next. Evidence label: private evidence.

STATE	FUNCTION
Memory	Preserves prior work, failures, successful routes, and unresolved questions.
Evidence	Separates observations from inference and carries provenance forward.
Topology	Represents tasks, dependencies, boundaries, and reachable transitions.
Authority	Defines which consequential transitions may be admitted or vetoed.
Receipts + checkpoints	Make accepted state durable and allow interruption, replay, and recovery.
Process state	Detects recurring waste and can redirect or repair the workflow itself.

Why this can be more than repeated prompting

Each accepted cycle can alter the persistent state from which later reasoning begins. The model can change without automatically erasing accumulated evidence and control state. The intended phase change is durable, self-correcting work across interruptions and model calls - not merely faster text generation.

This release describes the architecture. It does not publish a benchmark proving the magnitude or universality of the advantage.

Evidence labels and limits

Architecture description

A statement about how the system is intended to be organized. The README, overview, graph, and release manifest publicly support these descriptions.

Private evidence

An implementation statement observed in internal source, receipts, run artifacts, or control-plane state that is not included here. This label permits a clear first-party statement without implying that the public can independently verify the hidden evidence from this dataset.

Limitations

This overview is not evidence of consciousness, AGI, universal autonomy, universal safety, complete mediation, or a known quantitative advantage over model-only agents. External state can also be wrong: stale evidence, incomplete mediation, bad authority, or correlated model interpretation remain meaningful failure modes.

The narrow claim

**Ouroboros is designed and observed internally as a system in which it holds planning and technical authority, Codex independently challenges and verifies, and meaningful intelligence and continuity are stored outside the model.
Evidence label: private evidence.**

Release map

README.md - one-screen orientation
architecture_overview.md - short prose overview
architecture_graph.json - machine-readable loop
evidence_manifest.json - claim labels
CLAIM_BOUNDARIES.md - explicit non-claims