

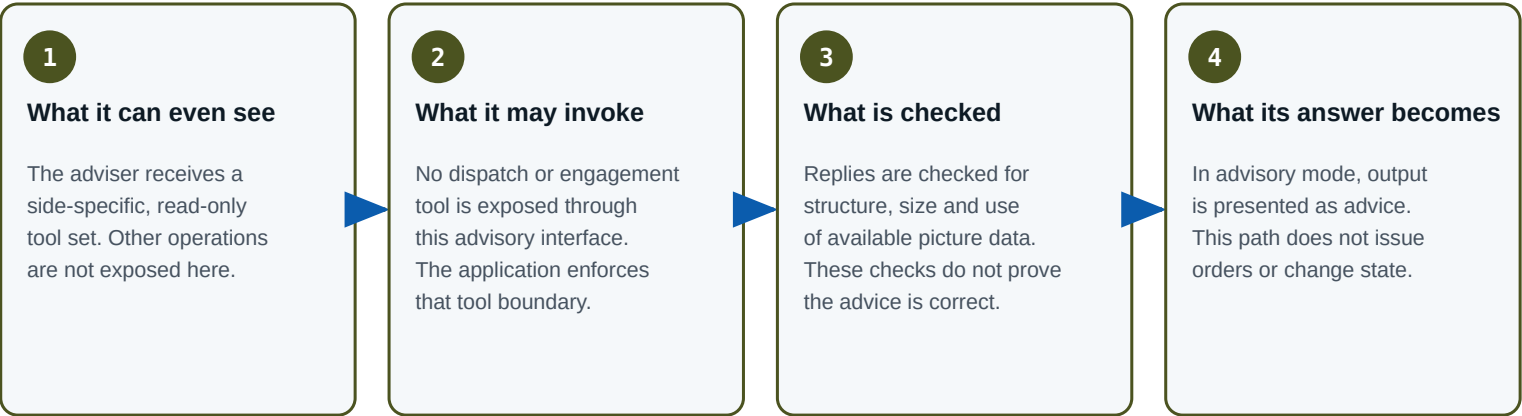
Limiting what an AI adviser is permitted to do

Application controls define available tools and validate responses. Software assurance can help establish which controls are running.



BOUNDED AUTHORITY, IN FOUR GATES

The advisory interface has defined limits



Operations excluded from the published advisory tool interfaces

× publish a proposal

× replicate to a peer

× authorise anything

× touch the audit record

× task or dispatch

× see a credential

× engage

× see attestation evidence

× write a civilian alert

× control media

ASSURANCE AROUND THE APPLICATION

Check the software enforcing the limits

Confidential computing can protect a defined application boundary from its host. It supplements organisational and application controls; it does not remove the need to trust the hardware, verifier policy, approved code or deliberate data recipients.

- The hardware demonstration places application logic and tool execution inside an enclave. The external model and its provider remain outside that boundary; data sent to them is not hidden from those recipients.
- An independent client has appraised hardware evidence and completed a protected call. The in-page verifier is served from unsealed assets, so its panel alone is not an independently trusted browser bootstrap.
- Software evidence can identify the measured controls; it does not certify their correctness or the safety of every model response.

SCOPE OF THE DEMONSTRATION

READ-ONLY ADVISORY INTERFACES

Two distinct modes

The advisory tools let a model inspect a bounded picture. They do not give it direct command or effect operations. The external model is not inside the attested application boundary.

A separate synthetic wargame controller can use validated model outputs to take actions after an operator authorises that invocation. This is distinct from the advisory-only path.

Authorise the mode and constrain its actions

An authorised exercise is not per-action approval. The scope must be clear to the operator.

What the evidence establishes

The recorded application and tool boundary is distinct from the external, unattested model. Two advisers completed structured tool-call cycles in the recorded demonstration. This is not evidence that every model response is safe.

Read-only tools limit direct action through that interface. The separate synthetic controller has a different, explicitly authorised scope.

Neither mode is a guarantee against rogue AI.

The assurance argument

Restrict the operations available to the adviser, validate its output and check the software enforcing the boundary. State separately who authorises any controller that turns advice into actions. Hardware checking does not make advice correct.