An Addition Is Not An Improvement.

If a component is there to improve the workflow, define what should change relative to a relevant baseline.

Start with the baseline.

Different Inference Paths. A Shared Output.

A component can use direct, conditioned or iterative inference, but it should always produce the same type of output.

DIRECT

One-shot response.

CONDITIONED

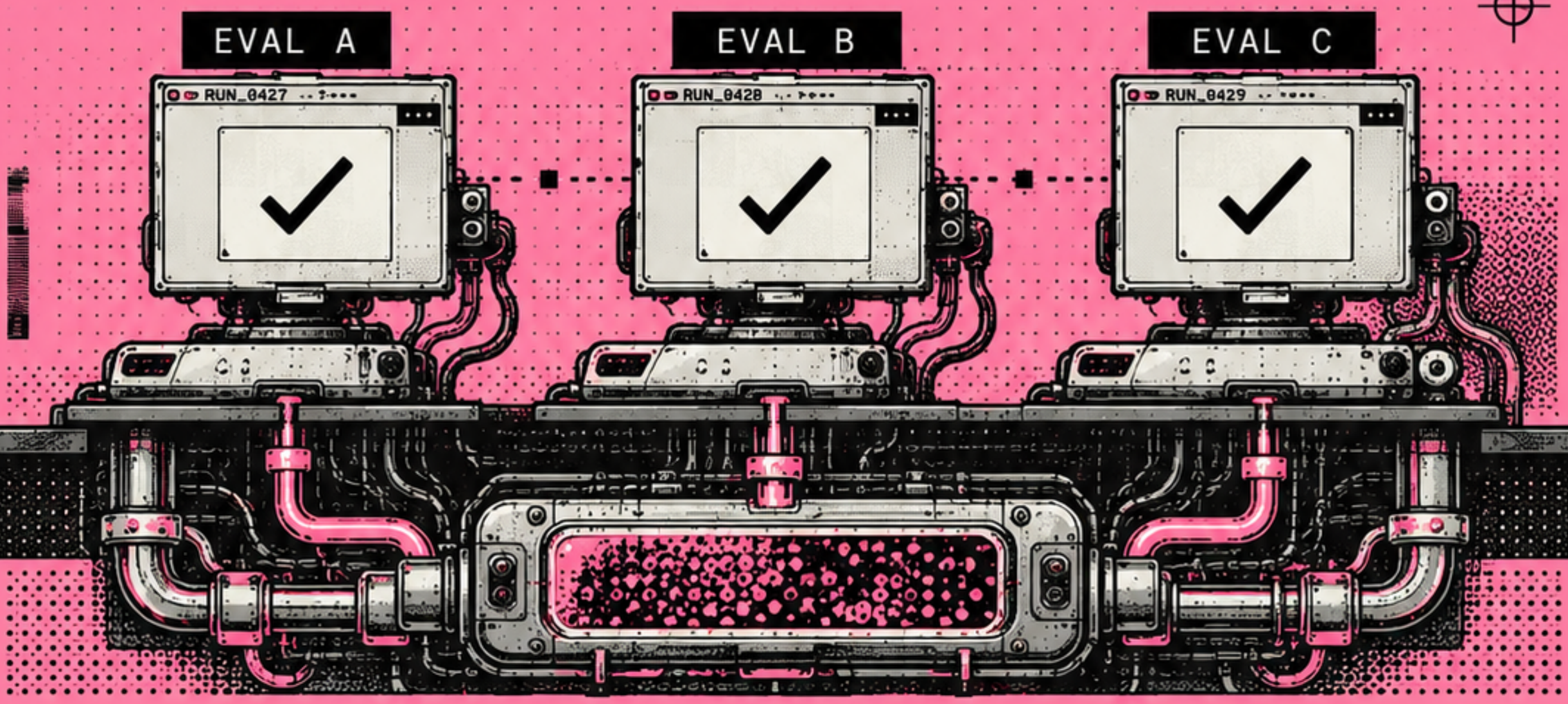
Uses context, instructions or tools.

ITERATIVE

Refines through multiple steps.

Agreement Doesn't Establish Independent Failure Paths.

Multiple evaluators can improve performance. Shared evidence, prompts, model families, assumptions, or dependencies can create correlated failure risk.



SHARED EVIDENCE

SHARED MODEL FAMILY

SHARED PROMPT

SHARED DEPENDENCY

Evaluate the ensemble's behavior, not just the vote count.

Human Review Can Address Different Risks.

Use human review where it meaningfully reduces risk, either for performance-related issues or for decision / governance reasons.



PERFORMANCE-RELATED

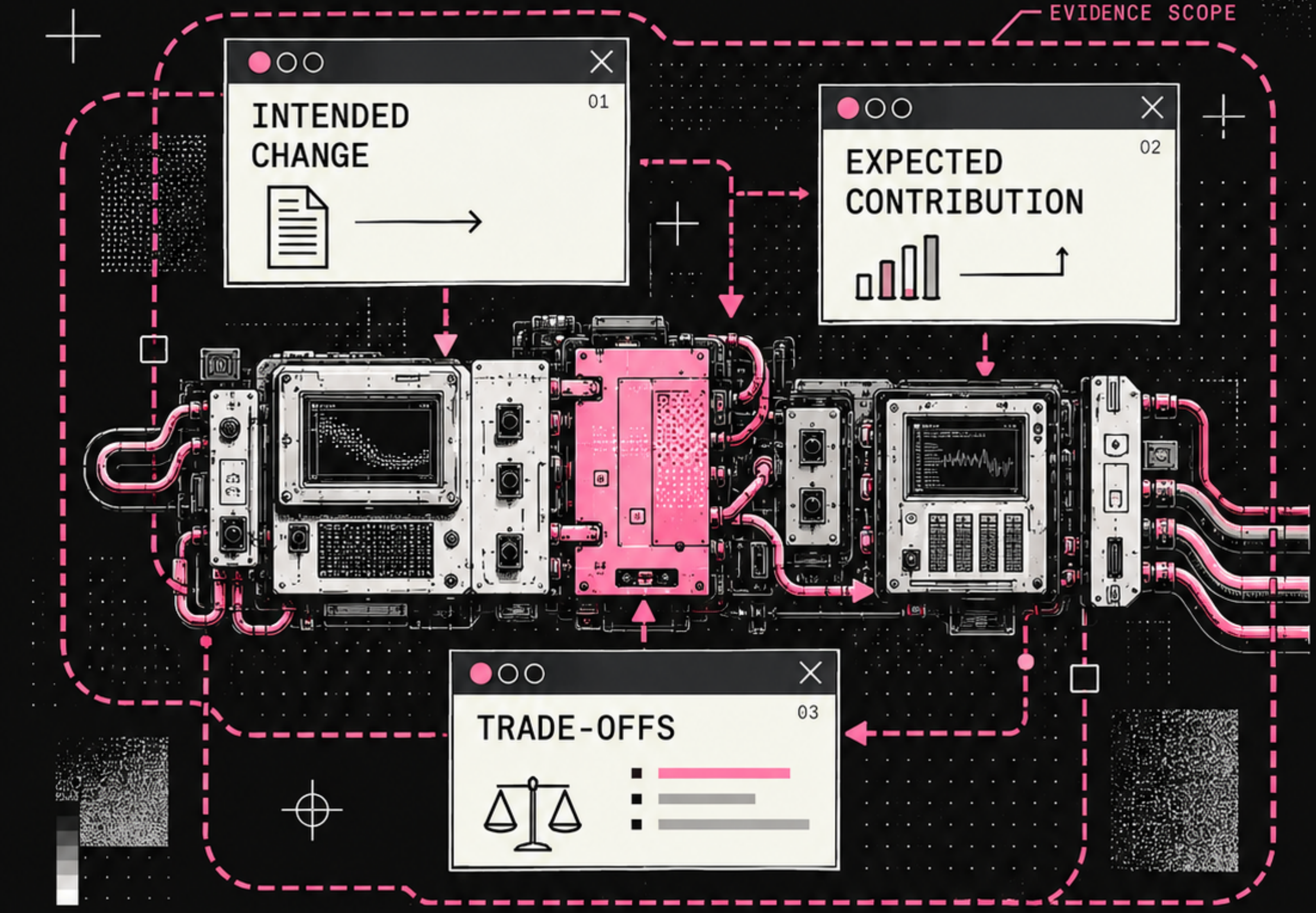
- Subjective judgment
- Nuanced interpretation
- Factual uncertainty
- Domain expertise

DECISION / GOVERNANCE

- High-stakes decisions
- Policy or compliance
- Legal or regulatory impact
- Ethical considerations

Evaluate Within The Right Scope.

Assess the component in the context of the modified workflow, and align the evidence with the claim.



Evidence Should Match The Scope Of The Claim.