SolvXAI

Data to decisions, powered by AI

SOLVXAI WHITE PAPER SERIES | NO. 08 | ENGINEERING JUDGMENT

Resolving Engineering Ambiguity at Scale

How an AI engineering system turns multidisciplinary disagreement into testable hypotheses, scoped evidence, and an auditable decision-grade synthesis.

IN BRIEF

Multidisciplinary disagreement often reflects different definitions, boundaries, or evidence—not competing versions of one answer. This paper explains how SolvXAI stabilizes the observation, generates competing hypotheses, delegates discriminating evidence work, and reconciles the result without hiding dissent or overstating certainty.

Published by the SolvXAI Research Team

August 2026 · Edition 1.0

solvxai.app

Reader's guide

This paper exposes SolvXAI's engineering principles and evaluable obligations while withholding proprietary prompts, routing logic, tool contracts, thresholds, and infrastructure details.

THE PROBLEM	THE SHIFT	THE STANDARD
Discipline-led fan-out can multiply hidden definitions and produce irreconcilable answers.	Generate competing hypotheses under one observation contract, then delegate evidence tests.	Reconcile by scope, provenance, falsifiers, uncertainty, and decision fitness—not votes.

Contents

Abstract

Multidisciplinary engineering questions are often stated as decisions before they are stated as well-formed analytical problems. “Which wells are underperforming?”, “What is causing the decline?”, and “Where should we intervene?” appear precise, yet each can conceal several legitimate interpretations: the population being compared, the time and operating regime, the definition of performance, the physical system boundary, the grade of evidence required, and the authority implied by the answer. A conventional multi-agent response is to assign the question independently to reservoir, production, completions, and operations specialists and then combine their outputs. This increases breadth but can multiply ambiguity. Each discipline may answer a different question correctly, leaving the coordinator with several incompatible rankings and no principled way to reconcile them.

This paper presents SolvXAI’s objective-led, hypothesis-first approach to ambiguity resolution. The system first establishes a neutral observation and decision contract. When causal ambiguity is material, it generates a bounded set of competing, testable explanations before assigning specialist work. Agents are then organized around evidence obligations for those hypotheses—not around producing discipline-specific final answers. Findings are reconciled by scope, provenance, predicted signatures, falsifiers, evidence strength, and decision relevance. Agreement is treated as metadata rather than proof; missing evidence is not treated as falsification; and unresolved alternatives remain visible. The architecture uses the platform’s existing context, skills, specialist roles, workflow, and evaluation surfaces. The design contribution described here lies primarily in the organization of judgment rather than in additional runtime machinery.

An anonymized production observation illustrates the need. Two specialist top-ranked well lists showed substantial non-overlap, while the delivered answer silently adopted one list. The problem was not merely weak synthesis: the rankings encoded different definitions of performance and incomplete cross-disciplinary coverage. The revised design converts such divergence into an explicit hypothesis-and-evidence program. We describe the design principles, conceptual workflow, stopping logic, evaluation strategy, limitations, and implications for responsible engineering AI. This is a design paper and technical case study, not a controlled efficacy study.

Keywords: engineering ambiguity; hypothesis generation; multidisciplinary reasoning; evidence reconciliation; subsurface engineering; agentic systems; decision support; human oversight

1. Introduction

Engineering ambiguity is not an edge case. It is the normal condition that exists between a decision request and a defensible technical conclusion.

Consider a request to identify the ten best-performing wells in an asset. “Best” might mean highest current rate, strongest production relative to reservoir quality, lowest decline, greatest recovery relative to expectation, best economic margin, most reliable operations, or greatest remaining opportunity. A reservoir engineer may normalize by connected volume or pressure support. A production engineer may emphasize uptime, drawdown, liquid loading, or artificial-lift behavior. A completions engineer may compare stimulation intensity and completion generation. Their outputs can differ substantially without any specialist being irrational.

The mistake is to treat these outputs as votes on one hidden truth. Before reconciliation is possible, the system must establish whether the specialists evaluated the same population, time window, operating state, comparison basis, and decision. If they did not, rank averaging or majority agreement produces false precision. If they did, disagreement may be valuable evidence of a coupled mechanism or an unresolved data problem.

This problem becomes more consequential as AI systems gain the ability to delegate. Parallel specialist agents increase reach and reduce the context burden on a lead agent, but fan-out by itself does not produce independent verification. It may instead create several anchored analyses, duplicate the same evidence, or return incompatible answers that are compressed into a convenient narrative. The coordination challenge is epistemic: what claims exist, what observations would distinguish them, and what evidence is strong enough for the intended decision?

SolvXAI addresses this challenge through three linked ideas:

- **Objective before method.** The system resolves the decision being served before selecting an analytical workflow.
- **Hypotheses before disciplinary conclusions.** Where broad causal ambiguity exists, competing explanations are generated under a shared observation contract before specialists are asked to test them.
- **Evidence before consensus.** Reconciliation is based on claim scope, independent evidence, falsifiers, uncertainty, and decision relevance—not on voting, rank averaging, or confidence language.

These ideas echo a long scientific tradition. Chamberlin's method of multiple working hypotheses was intended to reduce attachment to a single "ruling theory" [1]. Platt's strong inference emphasized alternative hypotheses and experiments designed to exclude them [2]. Modern systems engineering similarly calls for decision analysis proportional to complexity, uncertainty, and stakeholder consequence [3]. SolvXAI adapts these principles to long-running AI-assisted engineering work, where the system must also manage context, delegation, evidence provenance, and human authority.

2. Ambiguity is a system of coupled uncertainties

Ambiguity is often treated as a language problem: ask a clarifying question, obtain a missing parameter, and continue. That is necessary in some cases but inadequate at scale. Many ambiguities are discoverable only after inspecting the available evidence, and some should remain represented as bounded alternatives rather than being prematurely collapsed.

SolvXAI distinguishes several recurring forms.

Ambiguity class	Hidden question	Typical failure if ignored
Objective	What decision is the work meant to support?	A technically correct analysis answers the wrong decision.
Metric and comparison	What does "best," "underperforming," or "similar" mean?	Rankings are compared although their denominators differ.
Population	Which entities are eligible, comparable, and sufficiently observed?	Selection bias and silent exclusion distort the result.
System boundary and scale	Is the relevant unit a well, completion, reservoir compartment, pad, facility network, or time regime?	Local explanations ignore larger coupled constraints.
Causal	Which mechanisms could produce the observation?	The first plausible story becomes the diagnosis.
Evidence	What data are fit, independent, and discriminating?	Repeated use of one source is mistaken for corroboration.

Ambiguity class	Hidden question	Typical failure if ignored
Applicability	Does an external method or analogue transfer to this asset and regime?	Valid research is applied outside its domain.
Decision grade	Is the answer exploratory, screening-grade, design-grade, or execution-grade?	Confidence exceeds what the evidence can support.
Authority	Who may recommend, approve, or execute the consequence?	Analysis is mistaken for operational authorization.

Ambiguity is layered—and the layers interact

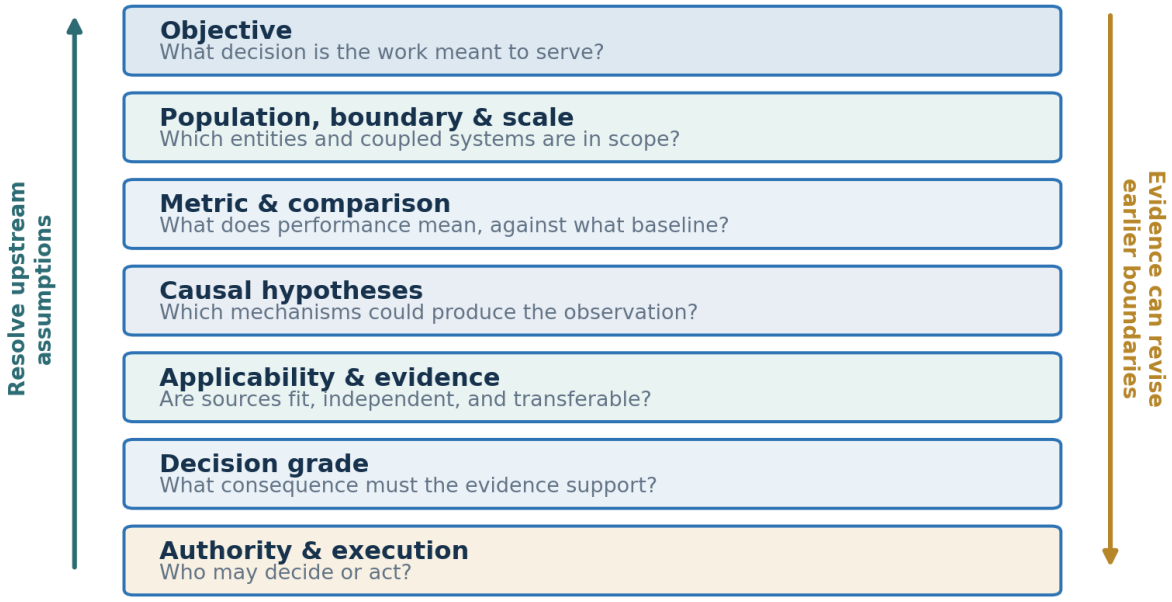

Figure 1. Ambiguity is layered. Upstream choices constrain downstream analysis, while new evidence can force the system to revisit an earlier boundary.

The ordering matters. A causal model cannot be responsibly ranked if the observation being explained is unstable. A well may appear to underperform because the comparison cohort is wrong, the operating regime changed, pressure data are not datum-aligned, or facility downtime is attributed to the well. In such cases, “data quality” is not a preliminary housekeeping task separate from diagnosis. It is itself a competing explanation.

The architecture therefore treats ambiguity as a routing signal. Some ambiguity is **load-bearing**: different resolutions would materially change the workflow or decision. That ambiguity must be resolved through evidence or a focused question. Other ambiguity can be represented safely as parallel branches with explicit selection conditions. The goal is not to eliminate uncertainty. It is to prevent an unexamined assumption from propagating as fact.

3. Design thesis: organize work around claims, not departments

A common multidisciplinary fan-out pattern mirrors an organization chart. A coordinator asks each discipline for its answer, then synthesizes the responses. This is useful when the decision naturally decomposes into independent disciplinary deliverables. It is weak when all disciplines are interpreting the same ambiguous phenomenon.

Suppose an asset shows declining performance. A department-led pattern may produce:

- a reservoir ranking based on depletion or connectivity;
- a production ranking based on rate, drawdown, and uptime;
- a completions ranking based on treatment design and offset response; and
- an operations ranking based on equipment state and constraints.

The coordinator now has four answers, each with different inclusion rules and implicit counterfactuals. Asking the agents to debate can make the prose more coherent, but conversational convergence is not the same as evidentiary convergence. Models may align because one argument is rhetorically stronger, because they share training priors, or because later agents anchor on earlier outputs—a known hazard in judgment under uncertainty [9].

SolvXAI instead forms work around **competing hypotheses and evidence obligations**. Disciplines remain essential, but they are selected because they own particular tests, system boundaries, or evidence—not because every broad request must be answered once per department. This changes the unit of delegation:

> ***From: "Give me the production-engineering answer."***

> ***To: "Test whether changing operating constraints explain the observed cohort-relative decline, using these predicted signatures and falsifiers."***

The shift creates common ground. Specialists can still disagree, but the disagreement becomes locatable: a different observation, an incompatible boundary, conflicting measurements, a non-unique model, or a different decision value. Reconciliation then has a technical object to operate on.

4. The public conceptual architecture

The public architecture can be understood as six reasoning stages. These are decision contracts, not a fixed analytical recipe. The system may revisit earlier stages when evidence changes the boundary or reveals a new alternative.

4.1 Establish the objective shape

The Engineering Index first determines the kind of engineering work being requested. SolvXAI uses a small set of objective shapes: investigation, benchmarking, opportunity screening, decision support, assurance, applicability assessment, and procedure planning. Each shape has a different definition of sufficient evidence.

This distinction prevents category errors. Benchmarking may identify a performance gap without claiming its cause. Investigation evaluates causal explanations. Opportunity screening separates an attractive signal from feasible, actionable intervention. Assurance evaluates an existing claim rather than adopting it as context. Applicability asks whether an external result transfers to the asset. Procedure planning distinguishes analytical guidance from controlled execution.

4.2 Establish a neutral observation contract

Before causal fan-out, the main agent stabilizes the observation to be explained. A fit-for-purpose contract includes:

- the decision being served;
- the entity or population;
- the relevant system and time boundary;
- operating state and comparison basis;
- units, datum, identity, and data fitness;
- the observed condition without embedded cause;
- what would count as a useful explanation for the decision; and
- the grade of decision the evidence is expected to support.

Neutrality is important. “Wells constrained by formation damage” is not an observation; it contains a diagnosis. “Wells with a sustained rate shortfall relative to a declared cohort and operating-state baseline” is testable without prejudging mechanism.

4.3 Generate competing explanations independently

When ambiguity is narrow, the main agent may continue directly. When multiple mechanisms could materially alter the decision, the system generates a bounded first wave of hypotheses. Generators receive the same observation context but are not shown one another’s candidate causes. They are not asked for final rankings.

Independence is used to reduce path dependence, not to manufacture diversity for its own sake. Candidate explanations are expected to include, where supported:

- physical mechanisms at relevant scales;
- operational or equipment mechanisms;
- measurement, identity, or data-quality explanations;
- confounding factors and changing operating regimes;
- null explanations in which the apparent anomaly is not decision-material; and
- coupled or interaction hypotheses when no single discipline explains the signature.

The output is consolidated into a canonical hypothesis register with stable identities. Near-duplicates are merged only when their predicted signatures and falsifiers are materially equivalent. Over-broad hypotheses are split. Minority explanations are preserved when they remain plausible and decision-relevant.

4.4 Assign discriminating evidence work

Each active hypothesis is translated into an evidence obligation:

- What mechanism is asserted?

- What signatures should be present if it is true?
- What observation would weaken or falsify it?
- What confounders could mimic the same signature?
- Which specialist, tool, and evidence source are appropriate?
- At what entity, cohort, system, and operating regime does the result apply?

The system then chooses the minimum sufficient specialist team. A production engineer may own a test of constraint changes; a reservoir engineer may own pressure-support and connectivity evidence; a completions engineer may own treatment-response signatures; an operations specialist may own facility or equipment state. Some tests require only calculation or data retrieval rather than another agent. The principle is proportionality: add an independent leg when it changes coverage, discrimination, or assurance—not as a ritual.

4.5 Reconcile by scope and evidence

Specialist reports return into the hypothesis register rather than into a prose popularity contest. For each hypothesis, the system records the applicable scope, evidence supporting and conflicting with it, data limitations, independence of sources, and the discriminating strength of the result. Status is scoped and may be **supported**, **weakened**, **unresolved**, or **falsified**.

These terms are deliberately conservative. Missing pressure data do not falsify depletion. A strong well-level signature does not automatically establish a field-wide mechanism. A correlation may support a screening hypothesis without supporting causal intervention. The same source, transformed by several agents, is not independent corroboration.

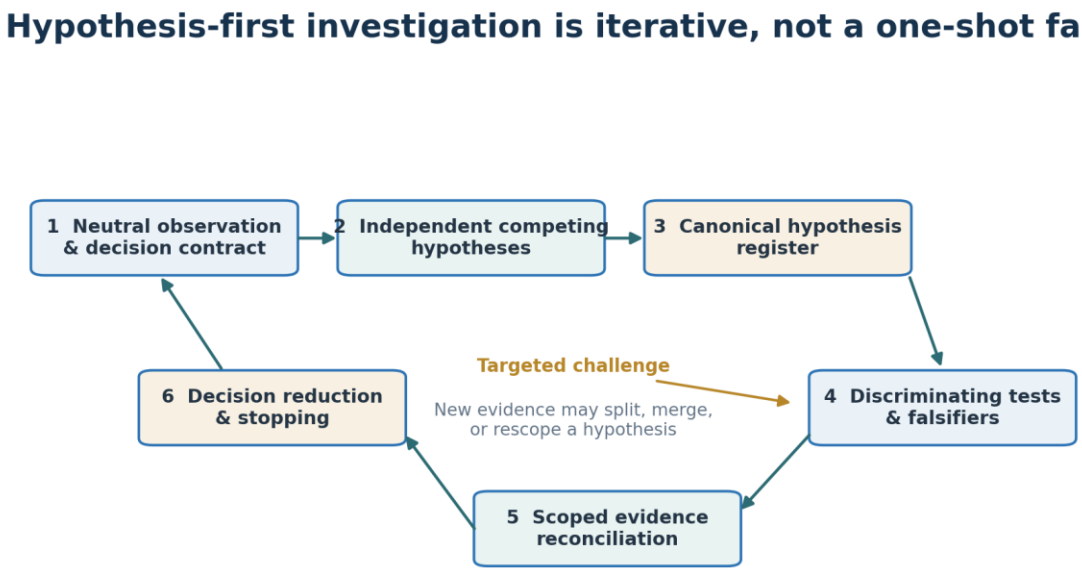

Figure 2. The hypothesis lifecycle. The workflow is iterative: evidence can rescope the observation, split a hypothesis, or trigger a targeted challenge.

4.6 Reduce to a decision without erasing uncertainty

Decision reduction begins only after the hypothesis space has been narrowed enough for the requested decision grade. The final answer declares one comparison and selection basis. It separates:

- observed severity;
- causal support;
- actionability and reversibility;
- uncertainty and unresolved alternatives; and
- the evidence that would most efficiently change the decision.

The system may recommend further surveillance, a bounded pilot, or no action. A “stop” outcome is legitimate when additional analysis has low expected decision value. NASA’s systems-engineering guidance similarly recommends evaluating the robustness of rankings and whether further uncertainty reduction is worth its cost [3].

5. Evidence-centered reconciliation

Reconciliation is the design’s central problem. SolvXAI rejects five tempting shortcuts.

5.1 No majority vote

Three agents repeating the same conclusion do not create three independent observations. They may share evidence, assumptions, prompts, or model priors. Agreement can increase confidence only to the extent that the underlying paths are genuinely independent and technically relevant.

5.2 No rank averaging across unlike questions

A reservoir ranking and an operations ranking may be valid views of different objectives. Averaging their positions implies a common scale that may not exist. When a single portfolio order is needed, the decision basis must be declared first; discipline outputs then contribute evidence to that basis.

5.3 No “confidence wins” rule

Language-model confidence is not a calibrated substitute for evidence. A careful specialist with limited data may be less assertive than an agent overfitting a familiar pattern. Claims are therefore compared by provenance, discriminating power, uncertainty, and applicability.

5.4 No false intersection or indiscriminate union

Taking only items shared by every list can discard real discipline-specific risks. Taking the union can produce an unprioritized list whose size grows with every lens. SolvXAI instead constructs an entity-by-hypothesis evidence view and distinguishes convergent findings, scope-specific findings, coverage gaps, and genuine contradictions.

5.5 No conversion of absence into falsification

“Not observed” has several meanings: the signature is truly absent, the data are not fit, the relevant scale was not observed, or the test was never performed. Only the first may weaken a hypothesis, and even then only within the test’s scope.

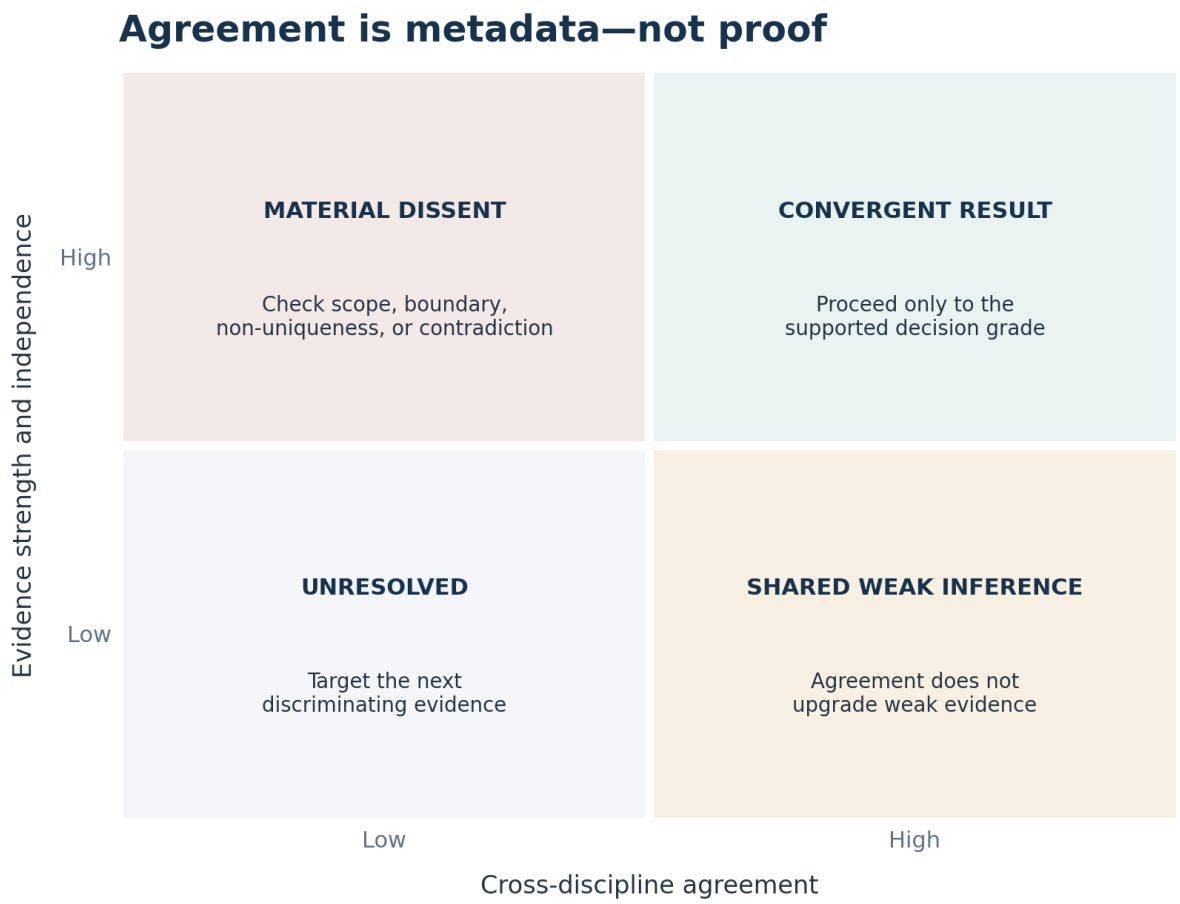

Figure 3. Evidence quality and cross-disciplinary agreement are separate dimensions. Consensus based on weak or shared evidence is not verification.

The evidence-agreement matrix captures the governing rule. High agreement with weak evidence remains a shared weak inference. High-quality evidence with low agreement deserves investigation, because it may reveal boundary differences, non-uniqueness, or a material contradiction. Agreement is useful metadata; it is not the adjudicator.

6. Scale is epistemic before it is computational

“At scale” is often interpreted as higher concurrency. For engineering ambiguity, scale first changes the reasoning problem.

A large portfolio is rarely one homogeneous population. It may contain different reservoirs, completion generations, equipment states, pressure regimes, facility constraints, reporting practices, and data-quality classes. Administrative groupings are not automatically engineering boundaries. A field can be a population of overlapping populations.

SolvXAI therefore discovers the smallest coherent unit for a test and the larger unit needed to reconcile interactions. One hypothesis may be evaluated at the well level, another at the pad or facility level, and another across a pressure-connected reservoir region. The scale itself can be a hypothesis: an apparent well problem may be a shared-network constraint; an apparent field trend may be a composition shift among cohorts.

This produces a different scaling model from indiscriminate fan-out:

1. Use deterministic data operations for repeated calculations and population census.
2. Use agents for distinct judgments, evidence regimes, boundary interpretations, and explicit professional ownership.
3. Narrow the hypothesis space before expanding expensive specialist work.
4. Preserve exact coverage and provenance outside the narrative context.
5. Stop when the requested decision is robust enough, not when every imaginable question has been explored.

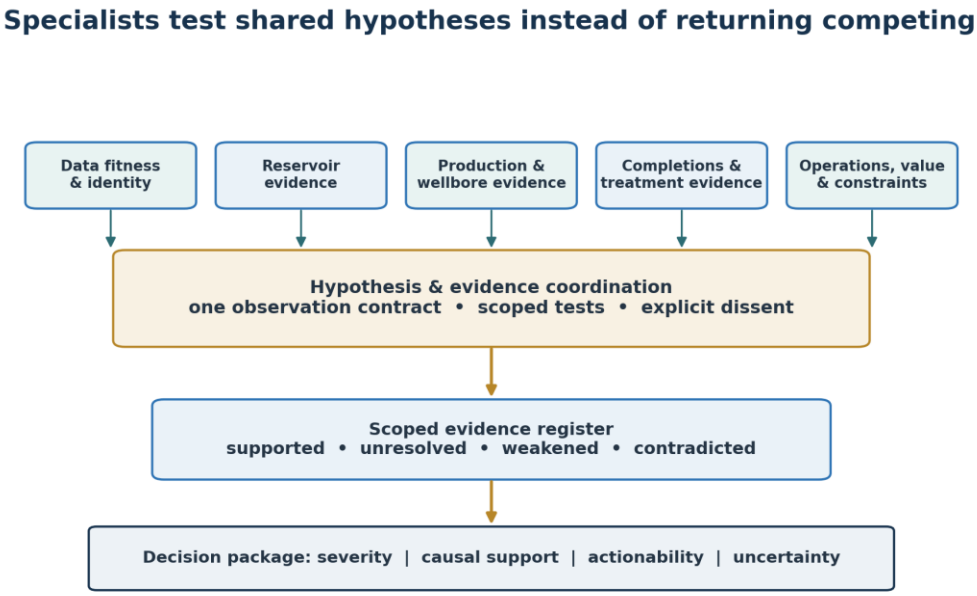

Figure 4. Specialists are selected to test shared hypotheses and boundaries; the final package separates severity, causal support, actionability, and uncertainty.

The result is adaptive breadth. A narrow, well-specified method can remain a simple workflow. A broad, causally ambiguous request earns independent generation and targeted tests. Additional challenge legs are opened only for a material contradiction, missing owner, weak discrimination, or cross-boundary interaction. The design philosophy is to use the simplest existing context and execution surface that can carry the obligation.

7. An anonymized production lesson: when top-ranked lists substantially disagree

The hypothesis-first design was sharpened by a production review of a real multidisciplinary analysis. In the reviewed run, reservoir and production specialists produced top-ranked well lists with substantial non-overlap. The final response silently adopted one list and did not disclose the divergence. Exact list and overlap counts are withheld here to reduce the risk of re-identifying the internal event.

This is a single internal observation, not statistical evidence of system-wide performance. It is nevertheless instructive because the failure can be reconstructed logically.

First, “top-performing” had not been converted into one decision contract. The reservoir view emphasized subsurface-relative performance; the production view emphasized behavior visible in production history and operating data. Second, each list was a censored observation: it revealed only the selected shortlist, not how every candidate was evaluated under each lens. Third, there was no entity-by-discipline coverage grid, so non-overlap

could not be distinguished from non-evaluation. Fourth, the final synthesis treated adoption as reconciliation and omitted the material dissent.

The lesson is not that both lists should have been averaged or merged. The appropriate response depends on the decision:

- If the question is purely current production performance, adopting the production basis may be correct—but the basis and reservoir dissent should be disclosed.
- If the decision is an intervention portfolio, current severity, causal support, feasibility, and expected value should be evaluated separately before reduction.
- If the question remains causally broad, the conflicting ranks should seed hypotheses and discriminating tests rather than become competing final answers.

Under the revised pattern, the main agent would establish a neutral performance observation and comparison basis, generate plausible cross-disciplinary causes, and assign evidence work against those causes. A final deliverable may contain the requested number of wells only when that many satisfy the declared basis at the required evidence grade; otherwise it returns fewer and reports the shortfall rather than padding the list. Each inclusion has a declared reason: observed severity, supported mechanism, actionability, or explicit screening status. Wells selected by only one discipline remain visible as scope-specific findings or unresolved candidates rather than disappearing during synthesis.

8. Relationship to multi-agent fan-out and debate

Contemporary agent systems commonly use an orchestrator-worker pattern. Anthropic's published multi-agent research architecture describes a lead agent that develops a strategy, launches specialized subagents in parallel, and synthesizes their findings [4]. Its authors emphasize independent exploration, broad-to-narrow search, effort proportional to complexity, observability, and evaluation. Claude Code Review, documented as a research preview at the time of access, provides a more specific comparison: multiple agents inspect a change in parallel from different issue classes, candidate findings are checked against actual code behavior, and surviving findings are deduplicated and ranked [5]. In other words, it separates candidate generation from verification rather than treating every independent observation as a defect.

These patterns address breadth, context capacity, path dependence, and false-positive filtering. SolvXAI adapts the same candidate-then-verification logic to engineering, where a candidate is not merely a suspected code defect but a scoped causal explanation. The extension adds an observation contract, explicit population and boundary definitions, predicted signatures, falsifiers, evidence grades, and professional authority. General multi-agent debate research has also shown that exchanging answers and arguments over multiple rounds can improve reasoning in some tasks [6]. In engineering, however, convergence alone may be unsafe when agents are answering different questions, relying on shared weak evidence, or operating at different scales.

SolvXAI's adaptation is therefore hypothesis-centered rather than debate-centered:

Generic fan-out or debate pattern	SolvXAI hypothesis-first orchestration
Decompose by topic, role, or solution angle	Decompose by objective, hypothesis, evidence obligation, and boundary
Ask each agent for an answer	Ask selected specialists to test predicted signatures and falsifiers

Generic fan-out or debate pattern	SolvXAI hypothesis-first orchestration
Synthesize or debate outputs	Reconcile claims in a scoped evidence register
Treat convergence as a useful endpoint	Treat agreement as metadata; evidence determines status
Expand for breadth	Expand only when additional independence or discrimination changes the decision
Often optimize answer quality	Optimize decision fitness, auditability, and calibrated uncertainty

This is not an argument against discipline-led work. Some objectives genuinely require separate discipline deliverables. Nor is every request a causal investigation. The Engineering Index exists to make that distinction before fan-out.

9. Two synthetic engineering walkthroughs

The following examples are synthetic composites. They contain no company, field, well, or customer information. Their purpose is to show how the reasoning architecture behaves, not to prescribe a universal technical method.

9.1 Example A: a cohort of wells appears to underperform

Initial request. “Find the ten most underperforming wells and explain why.”

A direct discipline-led fan-out would likely produce several rankings. Instead, the system first defines the observation: sustained production below a declared peer expectation, over a stated stable-operating window, excluding periods of known downtime and requiring minimum data coverage. It then tests whether the peer population is coherent. Completion generation, fluid regime, pressure region, artificial-lift state, and facility association may justify overlapping cohorts rather than one field-wide comparison.

The independent first wave proposes a bounded set of explanations:

- reduced reservoir deliverability or pressure support;
- near-wellbore or completion impairment;
- artificial-lift or wellbore-performance limitation;
- shared facility or gathering constraint;
- changing fluid behavior or loading;
- measurement, allocation, identity, or downtime-accounting error; and
- a null explanation in which the apparent shortfall disappears under a fit cohort.

Each explanation creates different evidence obligations. A reservoir leg tests pressure and connectivity signatures. A production leg tests rate–pressure behavior and operating-state changes. A completions leg examines treatment-era and offset-response patterns. An operations leg tests shared constraints and equipment state. A data leg checks allocation, timestamps, units, and exclusions. No specialist is asked to return “the final ten.”

Reconciliation may show several wells with strong evidence of a shared facility limitation, others that are anomalous only under an invalid cohort, a smaller set consistent with impaired inflow, and an unresolved subset for which pressure evidence is absent. A facility hypothesis predicts correlated restrictions or recoveries among wells sharing the constrained system and is weakened when the suspected constraint changes without a

corresponding response. An impaired-inflow hypothesis predicts persistent rate–pressure deterioration at the well scale and is weakened when the behavior disappears after accounting for operating state. The final priority list can then declare its decision basis—for example, recoverable near-term production—while keeping causal support and uncertainty separate from observed severity. If fewer than ten wells meet that basis at the required evidence grade, the system returns fewer and reports the shortfall. The list is no longer a compromise among departments; it is a reduction of tested explanations.

9.2 Example B: a stimulation approach appears successful elsewhere

Initial request. "Should we apply this treatment design across the next development campaign?"

The request combines applicability, opportunity screening, and decision support. The system first separates the external claim—improved response following the treatment—from the asset-transfer question. It establishes whether rock and fluid properties, stress regime, completion geometry, operating constraints, offset spacing, measurement practices, and response windows are comparable.

Competing hypotheses might include a genuine treatment effect, selection bias in the treated population, a contemporaneous operating-policy change, geology-driven response, or a short-lived uplift that does not improve value. Evidence work is organized to discriminate these explanations rather than to collect discipline endorsements. Reservoir and geomechanics perspectives test transfer conditions; completions tests design-response signatures; production evaluates durability and operating confounders; economics and operations examine value, feasibility, and reversibility. One useful discriminator is whether the uplift persists after comparison with a matched cohort and adjustment for contemporaneous operating-policy changes.

If the evidence supports transfer only within one rock-quality and stress regime, the system does not force a field-wide yes or no. It may recommend a bounded pilot with predeclared success measures, monitoring, stop conditions, and a comparison population. The final authority remains with designated qualified human decision owners. Before any field execution, hazards, the operating envelope, accountable owners, approvals, and execution controls belong in a separate controlled procedure-planning process. Here, ambiguity is reduced into a bounded learning decision rather than hidden inside a universal recommendation.

10. Engineering principles derived from practice

The architecture is governed by principles intended to remain stable even as models and tools change.

10.1 Observation, interpretation, and decision are different epistemic layers

A measured rate, a calculated decline parameter, an inferred constraint, and a recommended intervention are not interchangeable. The system labels facts, calculations, assumptions, hypotheses, and decisions so that downstream users can see where judgment enters.

10.2 Professionals own method; the system owns structural discipline

The platform does not prescribe one fixed reservoir, production, or completions workflow for every asset. Responsible specialists select methods appropriate to physics, data, and scale. Common contracts require them to state applicability, units, checks, limitations, uncertainty, and evidence provenance.

10.3 User intent and professional authority are preserved

The user supplies the objective, constraints, and acceptance criteria. Designated qualified organizational decision owners retain approval and execution authority. The AI system may investigate and recommend within scope, but it does not silently broaden authority or convert an analytical result into field execution.

10.4 Context should be sufficient, curated, and non-duplicative

The lead agent retains the full conversational context for audit and synthesis. Specialist agents receive the minimum sufficient brief: assigned objective, relevant requirements, evidence boundary, constraints, resources, and return form. Withholding other agents' tentative causes during independent generation reduces anchoring.

10.5 Uncertainty is a deliverable, not a disclaimer

Useful uncertainty identifies what is unknown, why it matters, the scope of the limitation, and what observation would change the decision. "More data are needed" is incomplete unless the next discriminating evidence is named.

10.6 Reproducibility is part of engineering quality

Material conclusions retain their source trail, transformations, assumptions, units, and scope. NIST's AI Risk Management Framework similarly emphasizes validity in context, transparency, ongoing evaluation, and the documentation of limitations [7].

10.7 The evidence grade must match the decision grade

Screening evidence may justify surveillance or prioritization. It may not justify an irreversible intervention. This separation resembles the petroleum sector's distinction between technical uncertainty and project maturity or commerciality in the Petroleum Resources Management System [8].

11. Validation strategy

Agentic workflows are nondeterministic: valid runs may follow different paths. Validation must therefore evaluate both structural obligations and decision outcomes rather than require one exact chain of thought.

SolvXAI uses a layered validation strategy.

11.1 Implemented structural checks

Deterministic checks verify that broad engineering ambiguity activates objective orientation, that hypothesis-first investigation obligations are available to the right workflow, and that specialist return forms preserve selection basis, scope, evidence, and falsifier information. Control cases verify that narrow method requests and unrelated simple tasks do not trigger unnecessary fan-out.

11.2 Scenario corpus

Representative scenarios cover ambiguous underperformance, conflicting rankings, changing operating regimes, shared-system constraints, data-quality alternatives, coupled mechanisms, and mixed populations. The corpus also includes controls where the appropriate response is direct calculation, a focused clarification, or a single specialist.

11.3 Production-trace review

Frozen, anonymized traces are replayed conceptually to ask whether the revised design would expose the observed failure: Was the comparison basis declared? Were first-wave hypotheses independent? Was the population covered? Was dissent disclosed? Did the final decision distinguish severity, causal support, and actionability?

11.4 Holistic review between phases

Changes to shared context can have broad behavioral effects. Reviews therefore examine not only local wording but the entire route from objective recognition through delegation, return contracts, reconciliation, and final synthesis. Snapshot and consistency checks guard against accidental duplication or drift across context layers.

11.5 Planned outcome evaluation

Structural verification cannot establish empirical superiority. The next evaluation layer should compare baseline discipline-led fan-out with hypothesis-first orchestration on blinded cases. Useful measures include:

- population coverage and silent exclusion;
- number of materially distinct hypotheses considered;
- rate of unsupported causal closure;
- preservation of minority but plausible explanations;
- evidence independence and citation correctness;
- calibration of supported, unresolved, and falsified statuses;
- decision stability under withheld evidence;
- expert-rated actionability and safety;
- elapsed time, token cost, and cost per decision-changing finding; and
- over-orchestration on simple controls.

The paper makes no claim that these outcome measures have already been proven in a controlled study. It presents an implemented design and a falsifiable evaluation program.

12. Governance and publication boundary

The system is designed as decision support. Qualified professionals remain responsible for technical acceptance, regulatory compliance, operational authorization, and execution. Human intervention is especially important when evidence is incomplete, consequences are irreversible, stakeholders value outcomes differently, or a recommendation crosses from analysis into field procedure.

Publication also requires a boundary. The concepts in this paper are intentionally inspectable: objective orientation, neutral observation contracts, competing hypotheses, discriminating evidence, scoped reconciliation, proportional effort, provenance, and human authority. These are the principles by which users should judge the system.

The implementation details are intentionally excluded: exact prompts, proprietary routing and scoring logic, model policies, tool interfaces, operational thresholds, internal storage and service topology, and customer-specific evidence. The paper describes what obligations the system must satisfy, not the confidential mechanism by which every obligation is instantiated.

13. Limitations and open questions

The architecture has important limitations.

First, hypothesis diversity is not guaranteed. Independent agents may share model priors and training data. Different prompts or roles are not equivalent to statistically independent experts. Diversity must be evaluated by the distinct mechanisms, predictions, and evidence paths produced.

Second, falsification is difficult in observational engineering systems. Subsurface problems are often non-unique, measurements are sparse, operations change during observation, and multiple mechanisms interact. Status should therefore be scoped and provisional.

Third, an explicit hypothesis register can create bureaucracy if activated for simple tasks. The minimum-sufficient-workflow principle and simple controls are essential to prevent procedural overhead from replacing judgment.

Fourth, evidence quality remains constrained by identity resolution, units, timestamps, measurement practices, and access. A sophisticated workflow cannot recover information that was never measured, although it can make the missing evidence visible.

Fifth, decision reduction includes value judgments. Economic priorities, risk appetite, reversibility, regulatory obligations, and operational capacity cannot be inferred purely from technical evidence. They require declared stakeholder inputs.

Finally, the current evidence is design and implementation evidence, supplemented by an anonymized production observation. Controlled comparisons with domain experts and baseline agent patterns remain necessary. Future research should test which forms of independence produce the greatest decision value, how hypothesis registers evolve over long-lived asset studies, and how to calibrate stopping rules against consequence and evidence cost.

14. Conclusion

SolvXAI's approach begins with a simple recognition: multidisciplinary disagreement cannot be reconciled reliably until the system knows whether the participants are answering the same question.

The Engineering Index makes objective and ambiguity discovery explicit. A neutral observation contract stabilizes what is being explained. Competing explanations are generated before causal closure. Specialists are selected to satisfy evidence obligations and test predicted signatures or falsifiers. Findings return to a scoped hypothesis register. Evidence strength and agreement remain separate. Decision reduction declares its basis, preserves unresolved alternatives, and stops when the evidence is fit for the requested consequence.

This design uses no new class of runtime machinery. It refines the structure already present in an agentic engineering platform: context, skills, professional roles, delegation, tools, and evaluation. Its purpose is not to make ambiguity disappear. It is to make ambiguity observable, tractable, and auditable—so that exploration remains broad where it should, conclusions become narrow where the evidence demands, and human engineers can see why a recommendation deserves their trust.

References

[1] Chamberlin, T. C. (1897). "The Method of Multiple Working Hypotheses." *The Journal of Geology*, 5(8), 837–848. https://doi.org/10.1086/607980

[2] Platt, J. R. (1964). “Strong Inference.” *Science*, 146(3642), 347–353. https://doi.org/10.1126/science.146.3642.347

[3] National Aeronautics and Space Administration. (2016). *NASA Systems Engineering Handbook*, NASA/SP-2016-6105 Rev. 2. https://www.nasa.gov/wp-content/uploads/2018/09/nasa_systems_engineering_handbook_0.pdf

[4] Hadfield, J., Zhang, B., Lien, K., Scholz, F., Fox, J., & Ford, D. (2025). “How we built our multi-agent research system.” Anthropic. https://www.anthropic.com/engineering/multi-agent-research-system

[5] Anthropic. (n.d.). “Code review.” *Claude Code Documentation*. Accessed August 2026. https://code.claude.com/docs/en/code-review

[6] Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2023). “Improving Factuality and Reasoning in Language Models through Multiagent Debate.” arXiv:2305.14325. https://arxiv.org/abs/2305.14325

[7] Tabassi, E. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0).* NIST AI 100-1. https://doi.org/10.6028/NIST.AI.100-1

[8] Society of Petroleum Engineers. (2018). “2018 PRMS Update Completed.” *Journal of Petroleum Technology*. https://jpt.spe.org/2018-prms-update-completed

[9] Tversky, A., & Kahneman, D. (1974). “Judgment under Uncertainty: Heuristics and Biases.” *Science*, 185(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

About this paper

This publication is a design paper and technical case study prepared from SolvXAI’s engineering-practice architecture, objective skills, orchestration contracts, evaluation scenarios, and anonymized production review. It does not contain customer-identifying information and should not be interpreted as a statistical validation study or as field operating guidance.