isar innovations

isar innovations Technical Report

How LLMs Discover Novel Solutions Emergent Creativity in Agentic Research Workflows

Nino Wagensonner

info@isar-innovations.dev

June 2026

Abstract

Large language models are usually evaluated as reasoners, assistants, or summarizers. This framing misses their most consequential research role: under the right workflow conditions, LLMs can discover novel solutions. The crucial condition is structure: a novel solution must be grounded in evidence, articulate a mechanism, expose a falsifiable boundary, and justify why limited human, experimental, or proof-search budget should be spent on it. Recent work on AI-driven formal proof search demonstrates that agentic language-model systems become scientifically valuable when their outputs are coupled to hard verifiers such as Lean. Formal proof search closes a loop after a candidate statement exists. This paper studies the creative step that opens that loop: how LLM agents construct grounded, nontrivial, testable solution candidates from scientific documents.

We present an agentic workflow that transforms scientific documents into enriched research objects, conducts structured multi-perspective deliberation over their claims and mechanisms, and searches multiple graph views for candidate hypotheses. The workflow separates creative generation from grounding and review. Document enrichment materializes claims, evidence, references, formulas, semantic units, and graph structure; a deliberative creativity layer reframes these signals through disciplinary, adversarial, and explanatory lenses; an offline discovery layer ranks candidates by mechanism clarity, evidence support, novelty, and falsifiability. For selected candidates, the workflow continues beyond ideation by producing prototype paths, benchmark or attack obligations, and Lean proof re-entry targets. We evaluate the workflow by the quality of the hypotheses it surfaces and the executable validation paths it attaches to them. In one corpus-scale run, the primary discovery generator produced 780 candidates; after merging discovery, baseline, and stochastic candidate pools and deduplicating candidate signatures, the analysis retained 915 unique candidates and a 30-item creative frontier for review. We present this frontier alongside three outcome witnesses: a system-originated cryptographic search mechanism, a distant-analogy prototype for selective disclosure, and a spatial-econometrics falsifier transfer. The workflow also routes formalizable claims into Lean proof re-entry, where accepted `.lean` artifacts become hard evidence for selected theorem-layer claims. The central claim is that LLM creativity becomes operational when agentic workflows turn documents and human problems into grounded, falsifiable, novel solution candidates: formal proof search closes the loop; creative discovery opens it.

1 Introduction

Machines can now generate fluent scientific summaries, solve formal proof goals, and assist researchers across long technical workflows. The next step is machine-assisted discovery: LLMs finding novel solutions in the form of mechanisms, experiments, attacks, benchmarks, and theorem targets that were not directly given to them and that become worth testing.

At Isar Innovations, this is no longer treated as a speculative future. Our research process is shifting from a mostly human endeavour to a human-directed, machine-amplified discovery practice. Researchers still choose problems, judge importance, and accept responsibility for claims. But the production of candidate mechanisms, cross-paper analogies, falsifiers, experiments, and proof targets is increasingly delegated to agentic workflows that can search, reframe, prototype, attack, benchmark, and formalize research directions at a scale no individual researcher can manually sustain. This paper formalizes that practice.

Scientific creativity is often described as the production of ideas that are both novel and useful [7]. In actual research practice, usefulness is operational: a candidate solution becomes useful when it explains a tension, transfers a mechanism, exposes a boundary, or proposes a small experiment that could reject it. Discovery begins when a system turns source material into a candidate worth testing.

This distinction matters for AI-assisted research. Large language models have made it cheap to produce competent summaries, plausible related-work sections, and lists of possible future directions. The creative bottleneck has moved from phrasing ideas to selecting ideas worth

testing. Scientific review time, experiment time, formalization effort, and security analysis are scarce. A creative research system should therefore produce candidate hypotheses with mechanisms, evidence anchors, falsifiers, and review priorities.

Recent progress in AI-driven formal proof search gives a useful comparison point. Systems that combine language-model generation with Lean verification have begun to solve nontrivial mathematical problems because invalid reasoning can be rejected by a compiler. For example, Tsoukalas et al. [10] report that AlphaProof Nexus resolved 9 of 353 open Erdős problems and proved 44 of 492 OEIS conjectures by coupling LLM-based proof generation to Lean-based verification. This is a strong demonstration of the pattern "agentic generation plus hard verifier." Yet formal proof search addresses a different part of the research loop. It becomes possible once a formalizable conjecture or theorem target has been selected. Creative discovery answers the upstream question: how such targets are discovered, prioritized, and translated from messy scientific documents into testable hypotheses.

This paper focuses on that earlier stage. We study how LLM-based agents can produce emergent solution candidates from scientific literature and human research problems. The workflow converts a scientific document into an *enriched research object*: a readable representation of the paper with claims, mechanisms, formulas, evidence units, references, quality signals, and graph structure. It then runs a deliberative creativity layer that examines the document from multiple perspectives and produces candidate findings, analogies, tensions, synergies, falsifiers, and experiments. Finally, an offline discovery layer searches across multiple graph views to construct a *creative frontier*: a ranked, reviewable set of candidate solution hypotheses with explicit evidence and rejection criteria.

The contribution is a process model for machine-assisted novelty: LLM creativity becomes scientifically useful when enriched document understanding, multi-perspective reframing, graph-based hypothesis search, and falsifier-centered review are coupled into one workflow. The language model generates and interprets structured research artifacts. Grounding, graph structure, ranking, prior-art routing, and hard re-entry points such as experiments, attack suites, and proof assistants turn those artifacts into reviewable solution candidates.

This framing is deliberately complementary to systems such as AlphaProof Nexus: AI-driven Lean proof search can close formal goals once the goals exist, while creative discovery constructs candidate goals, mechanisms, and failure conditions upstream. In our workflow, proof re-entry is not a metaphor. When a candidate has a formalizable theorem layer, the system can emit Lean targets and route them into checked `.lean` artifacts. Isar Innovations develops and uses an internal proof mechanism in this style: generated claims are narrowed into Lean-checkable theorem layers and used as one validation channel for research candidates.

The paper's thesis can be stated compactly: *Formal proof search closes the loop. Creative discovery opens it.* Formal proof search supplies hard verification. Creative discovery supplies the targets: candidate hypotheses, mechanisms, and experiments worth verifying.

2 Related Work

2.1 Creativity as Novelty, Usefulness, and Constraint

The standard psychological definition of creativity combines novelty and usefulness [7]. For scientific work, usefulness is tightly coupled to constraint: an idea must be compatible with evidence, expose a mechanism, and generate consequences that can be checked. We therefore use a stricter operational notion of creative value. A generated hypothesis is valuable when it is reviewable: it can be traced to sources, explained as a mechanism, compared against prior art, and rejected by a minimal next experiment, proof target, or attack.

This view aligns with computational creativity work that treats creativity as a process rather than a single output [2, 5, 12]. The evaluation target here is research utility: whether a workflow

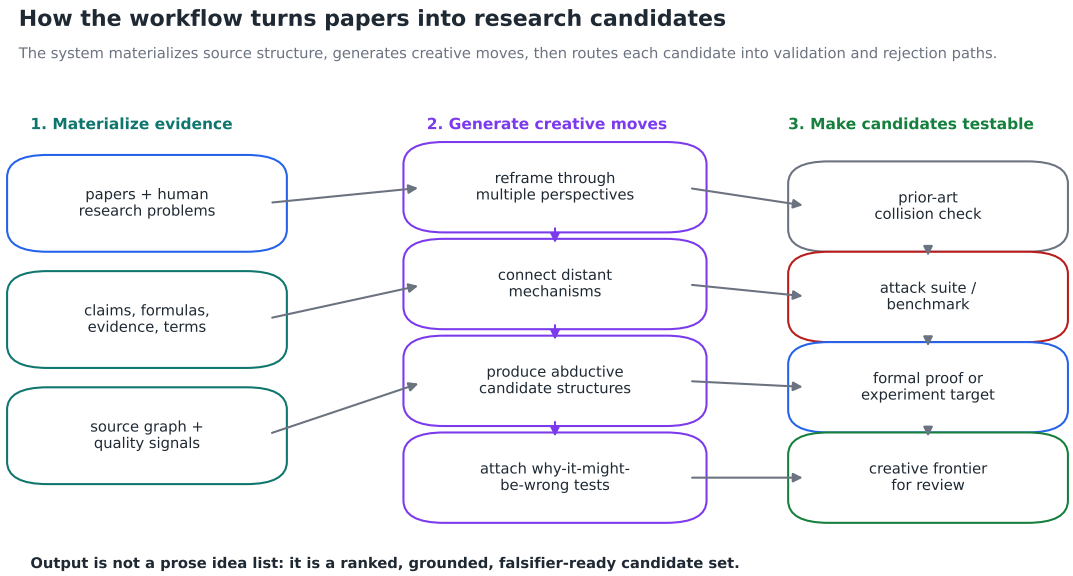

Figure 1: Agentic creative discovery workflow. Formal proof search closes the loop; creative discovery opens it.

increases the number of grounded, nontrivial, falsifier-ready hypotheses per unit of review effort.

2.2 Cognitive Models of Discovery

The workflow is abductive in the Peircean sense: surprising or underexplained observations motivate hypotheses that could explain them and yield testable consequences [3]. The system constructs candidate explanations, mechanism transfers, and boundary tests from structured evidence. In the present implementation, abductive moves include boundary-tension explanations, formula-motif analogies, method exaptation, falsifier transfer, failed-expectation analysis, role analogy, hidden-bridge construction, anomaly explanation, and compression-gain claims.

Abduction supplies reviewable hypotheses. A candidate enters the creative frontier when it can be attached to evidence, a mechanism, and a next test. This keeps the system close to scientific problem finding rather than unconstrained brainstorming.

The Geneplore model separates generative and exploratory phases of creative cognition [4]. The workflow follows a similar division. The enriched research object and deliberative layer generate preinventive structures: candidate mechanisms, analogies, tensions, and possible falsifiers. The discovery and review layers then explore those structures by ranking, deduplicating, routing, and rejecting them.

Serendipity is also relevant, but it should not be confused with randomness. Scientific serendipity often depends on prepared observation: an anomaly is noticed because the observer has the conceptual structure needed to interpret it [6]. The workflow operationalizes this principle by preserving anomalies, warnings, failed expectations, and cross-frame tensions in the graph, where they can later become candidates.

2.3 Knowledge Graphs, GraphRAG, and Literature-Based Discovery

Knowledge graphs and retrieval-augmented generation are natural baselines for scientific literature systems. They improve grounding by retrieving source material and structuring relationships between entities. Literature-based discovery systems similarly search for hidden connections

across bodies of literature [8]. The workflow builds on this intuition with a candidate-generation and review-allocation objective.

This distinction changes the output contract. A retrieval system may return the best answer to a query. A creative discovery system should return multiple candidate hypotheses, including useful rejects and known-boundary rediscoveries. Those non-success outputs are not failures. They show that the system is mapping a creative frontier rather than cherry-picking a final claim.

2.4 Agentic AI and Formal Proof Search

Agentic AI workflows decompose tasks into tool-using loops, critics, multi-perspective agents, and verifiers. Formal proof search is the strongest current example of this pattern in scientific reasoning because a proof assistant can reject invalid reasoning. AlphaProof Nexus is a recent high-profile example [10]. Similar systems, benchmarks, and datasets for Lean and theorem proving test proof generation, premise retrieval, and autoformalization [1, 9, 13, 14].

This paper is adjacent to that work but studies a different stage. Formal proof systems are re-entry points for selected formal claims. Our workflow supplies the upstream discovery layer that routes candidates into proof search, empirical benchmarks, and attack suites.

3 Problem Formulation

We use *falsifier* pragmatically. A falsifier is not necessarily a strict Popperian logical contradiction. In this paper, a falsifier is an empirically measurable failure condition, formal contradiction target, attack surface, or minimal next experiment capable of rejecting or materially weakening a hypothesis.

Let a scientific corpus be represented as a set of documents, each containing claims, methods, evidence, references, formulas, limitations, and implicit research questions. The task is to transform this corpus into a set of reviewable candidate hypotheses.

A candidate hypothesis package contains:

- a concise hypothesis;
- the source documents or source units that ground it;
- the creative move that produced it, such as analogy, exaptation, boundary bridging, or falsifier transfer;
- the proposed mechanism;
- evidence anchors;
- prior-art hooks;
- a minimal next experiment, attack, or proof target;
- a reason it might be wrong.

The system is successful if it increases the number of reviewable, nontrivial, evidence-grounded hypotheses per paper and per reviewer-hour. The target is useful novelty: candidates that combine surprise with grounding, mechanism, and a path to rejection. Rediscovered boundaries are valuable because they show that the system recognizes real scientific constraints. Prior-art collisions are valuable because they route candidates toward benchmarks, related work, or sharper claims.

4 Workflow

The workflow has four layers: document enrichment, deliberative creativity, grounded discovery, and falsifier-centered review.

4.1 Document Enrichment

The first layer transforms a scientific document into an *enriched research object*. The goal is not compression. The goal is to make the document operable for later creative search. The enriched object contains canonical text, semantic units, claims, methods, formulas, evidence references, limitations, terminology, source links, and graph edges.

This stage matters because creative search over raw PDFs is brittle. Documents hide mechanisms in prose, distribute evidence across tables and references, and often imply research questions without stating them. Enrichment materializes those structures once so that later passes can be cheaper and more reproducible. This is also the first point where quality signals can be attached: missing references, weak grounding, ambiguous claims, and suspicious formula usage become visible rather than being silently absorbed into a summary.

Graph construction begins here with conservative source-derived edges: references, citations, formulas, terms, semantic units, claims, evidence spans, methods, and quality signals. These edges form the baseline topology before the deliberative layer adds creative relations such as analogies, tensions, synergies, falsifiers, and proposed experiments.

4.2 Deliberative Creativity Layer

The second layer is the causal creative generator. It conducts structured multi-perspective deliberation over enriched research objects. Different perspectives are used to reframe the same document: a mechanistic reader asks what causal process is claimed; an adversarial reader asks what would break it; a domain-transfer reader asks where the mechanism might apply elsewhere; a methodological reader asks what experiment would distinguish explanations.

Deliberation produces structured creative data: tensions, anomalies, synergies, analogies, candidate hypotheses, falsifiers, and experiments. These outputs become graph nodes and edges that the discovery layer can search. This is the core reason the workflow can be creative: the model generates reframings and candidate structures, and later layers ground, compare, filter, and route them.

The sequencing is two-pass. First, document enrichment builds a grounded graph from source structure. Second, deliberation populates that graph with creative edges and candidate-bearing nodes. The discovery layer then searches both the conservative source graph and the deliberative graph, so candidate generation is grounded without being limited to direct citation or semantic similarity.

4.3 Grounded Discovery Layer

The third layer searches over multiple graph views because different creative moves become visible in different structures. Co-occurrence graphs expose local semantic overlap. Evidence graphs preserve grounding. Synergy graphs expose candidate relations. Heterogeneous graphs retain typed relations among papers, mechanisms, formulas, warnings, questions, experiments, and candidates. Hypergraph-style views can preserve multi-way support that pairwise projections would lose.

The discovery layer uses a portfolio of methods. Classical graph baselines include common neighbors, Jaccard, Adamic-Adar, resource allocation, personalized PageRank, Katz, preferential attachment, and BM25-style evidence matching. Structural generators search for bridges, boundaries, and community tensions. Grounded deliberative generators exploit path rules,

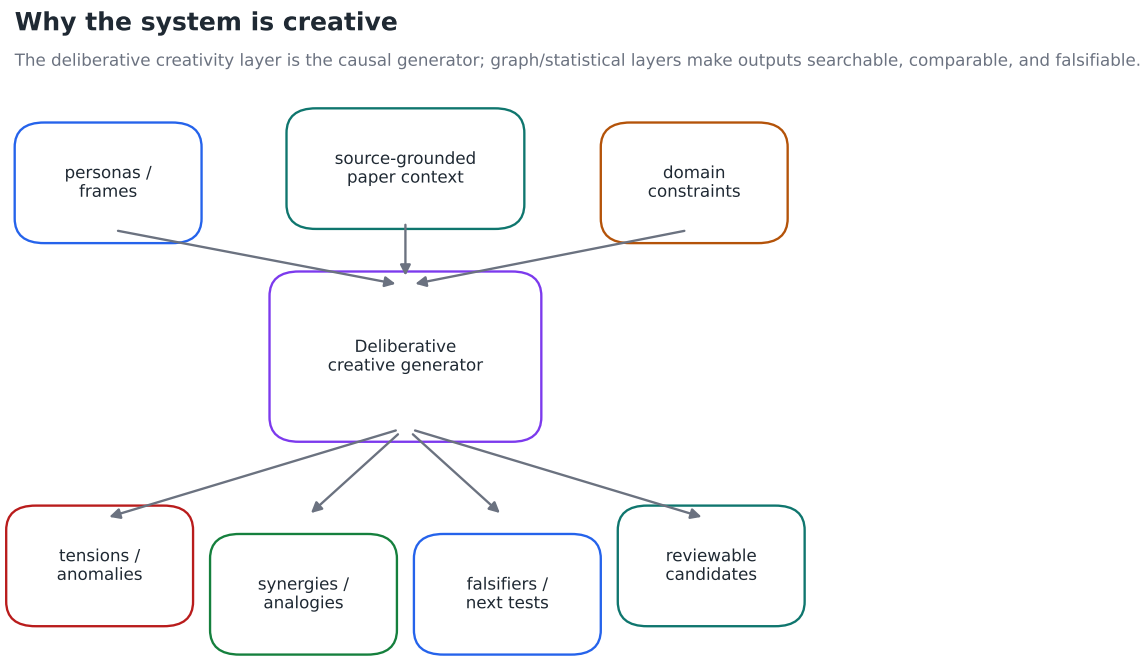

Figure 2: The deliberative creativity layer is the causal creative generator. Graph and statistical layers make its outputs searchable, comparable, and reviewable.

hypergraph support, metapath similarity, frame support, cross-frame resonance, formula-motif bridges, and synthesis portfolios. Fusion and review-allocation methods rank candidates by evidence support, mechanism clarity, testability, diversity, and prior-art risk.

The topology is deliberately parallel rather than sequential. Each view and method family proposes candidates into the same schema: source, target, technique, family, score, evidence references, and why the candidate is interesting. A normalization pass deduplicates overlapping source pairs and hypothesis variants, then preserves the method traces that produced them. Fusion does not erase disagreement; it records whether a candidate is supported by local overlap, structural bridge signals, deliberative evidence, hard-test surprise, or high-risk/high-reward distance. Review allocation then selects a frontier that balances score, support diversity, novelty, prior-art risk, and falsifier availability.

The key design principle is complementarity: the discovery layer makes deliberative outputs searchable, comparable, falsifiable, and reviewable at corpus scale.

4.4 Falsifier-Centered Review

The final layer produces reviewable hypothesis packages. A package is routed rather than accepted solely because it scores highly. Some candidates are promoted for deeper review. Some are downgraded because prior art already owns the claim. Some become benchmark ideas. Some become known-boundary rediscoveries. Some become productive rejects because the bridge is creative but unsupported.

This stage is essential for epistemic humility. A creative system that only reports successes invites cherry-picking. A useful scientific system should expose the whole frontier: successes, near misses, collisions, and failures. The frontier is the object of evaluation.

4.5 Formal Proof Re-Entry

Some candidates are empirical or adversarial and must re-enter through benchmarks or attack suites. Others contain a formal theorem layer. For those candidates, the workflow uses Lean re-entry as a hard validation path: natural language claims are narrowed into formal targets, translated into Lean 4 statements, checked as `.lean` artifacts, and either accepted by the proof assistant or returned as failed/partial formalization attempts.

This is where the paper's contrast with AlphaProof Nexus becomes operational. AlphaProof Nexus demonstrates that agentic proof search can solve selected mathematical targets by coupling language-model generation to Lean verification. Our system occupies the upstream stage: it discovers candidate theorem targets, mechanism lemmas, boundary claims, and proof obligations from scientific documents and human problems. Lean proofing then becomes one re-entry channel among several. A successful proof re-entry does not certify an entire research system; it certifies the formalized claim that was actually stated in Lean.

The KPT companion report documents one such outcome and its proof-oriented follow-up path [11]. The scientific value is the handoff: Isar Innovations uses a similar internal proof mechanism to decompose generated mechanisms into claims that can be measured, attacked, reviewed, or compiled by Lean.

5 Evaluation

We evaluate the workflow by asking whether it produces structured creative data that is better for review than random pairing or zero-shot ideation.

5.1 Corpus Scale and Reuse

The reported evaluation snapshot represents roughly a 10^3-artifact research corpus. This scale is already large enough to produce non-obvious, reviewable directions: the indexed database contains hundreds of enriched paper objects, hundreds of deliberative discussion objects, thousands of candidate entities, and hundreds of thousands of entity and edge records. The precise numbers are less important than the early-yield observation: even before moving to 10^4 or 10^5 papers, a 10^3-scale enriched corpus produced 780 candidates from the primary discovery generator. After merging discovery, baseline, and stochastic candidate pools and deduplicating candidate signatures, the analysis retained 915 unique candidates and a 30-item frontier for review.

This is not framed as a final scaling limit. It is evidence that useful unexplored directions can appear at modest corpus size once papers are converted into reusable structured research objects. Expensive document understanding is materialized once, while graph views, candidate ranking, deduplication, reviewpack generation, and ablations are cheaper repeatable passes over the enriched corpus. The next systems pass should report token and compute cost per enriched research object, per deliberative artifact, and per downstream candidate pass; these costs are the quantities that determine practical scaling to 10^4 and 10^5 documents.

The architecture naturally separates expensive enrichment from cheaper downstream discovery. Scaling to 10^4 and 10^5 papers should increase bridge opportunities and prior-art coverage while making deduplication, ranking, grounding, and human-review bandwidth central systems metrics. At larger scale, the correct metric is accepted or useful hypotheses per reviewer-hour, not raw candidate count.

5.2 Creative Frontier Run

In one corpus-scale run, the system produced:

- 780 candidates from the primary discovery generator;

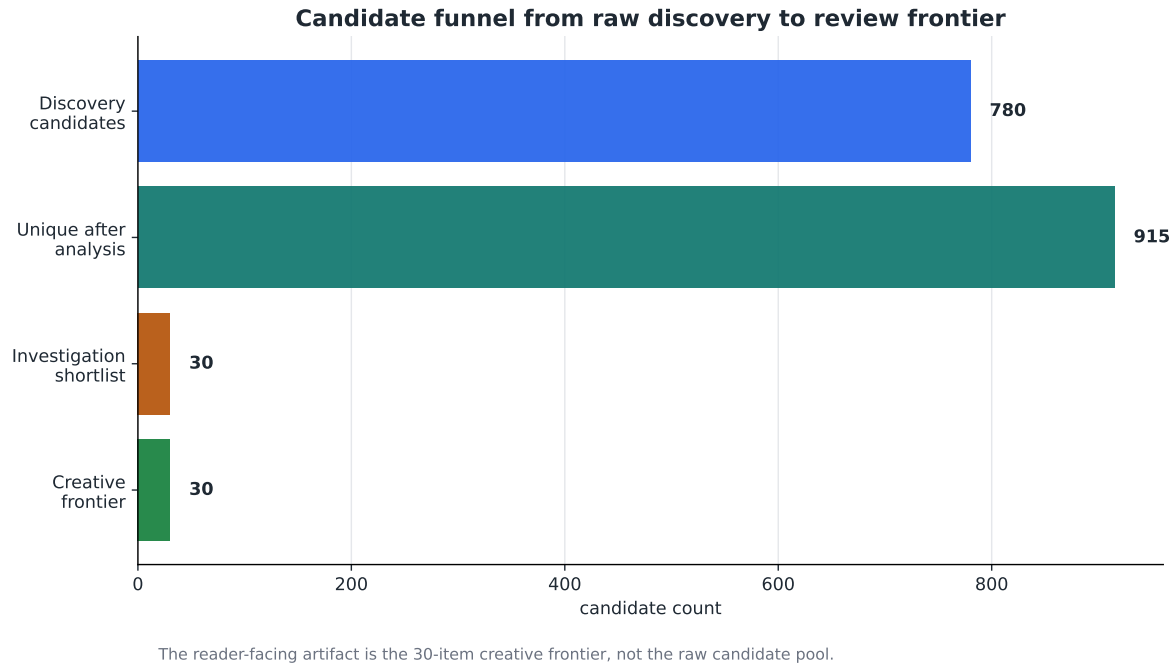

Figure 3: Candidate funnel from raw discovery candidates to a 30-item creative frontier.

- 915 unique candidates after merging discovery, baseline, and stochastic candidate pools and deduplicating candidate signatures;
- a 30-candidate investigation shortlist;
- a 30-candidate creative frontier;
- a manually reviewed illustrative sample of 16 directions.

The reader-facing result is the creative frontier, not the raw candidate count. The frontier contains promoted leads, maybe/deep-review items, prior-art-adjacent benchmarks, known-boundary rediscoveries, and productive rejects.

5.3 Baseline Comparison

The relevant baselines are random cross-paper pairing and zero-shot LLM ideation. Random pairing can produce surprising topic combinations, but it lacks evidence anchors, mechanism scores, falsifiers, prior-art hooks, and ranking reasons. Zero-shot ideation can produce fluent research ideas, but grounding and provenance are unstable, and there is no systematic corpus-scale ranking or noise audit. Grounded creative discovery instead produces source-paired candidate directions with mechanisms, scores, evidence counts, review buckets, and rejection reasons. Human and domain review starts from structured creative data rather than from prose guesses.

The present evaluation is artifact-level and process-level: it reports generated candidates, ranking behavior, case evidence, and traceability from sources to review packages. The baselines are used to define what the workflow must improve over: random pairing lacks source-grounded mechanisms, while direct zero-shot ideation lacks systematic corpus routing, rejection reasons, and repeatable frontier construction. A separate human-utility benchmark can then ask whether reviewers prefer grounded creative discovery outputs over those baselines.

The evaluation claim is therefore:

> The workflow converts a corpus into a structured creative frontier whose candidates can be promoted, downgraded, rejected, or routed to prior-art and falsification checks.

Table 1: Promising-first creative frontier sample.

Candidate direction	Underlying source papers	Creative move	Decision
AI-assisted skill expansion as precision-channel vs. mean-shift-channel trial	"Coding Beyond Your Training" plus synthesized randomized trial design	Reframe AI assistance as a decomposable causal mechanism rather than a single performance delta	Promote
Compound-matrix rank-boundary robustness under noise	"Inversion of the Multiplicative Matrix Compound Operator" plus numerical rank-estimation support	Turn an algebraic boundary into a perturbation/noise robustness experiment	Promote
Lattice preconditioning for counterterm discovery	Preconditioning and counterterm-discovery source papers	Transfer preconditioning from search efficiency into symbolic or field-theory discovery	Maybe / deep review
Compression-induced recall latency for sparse critical events	Memory consolidation and long-context video understanding papers	Convert memory compression into a rare-event recall stress benchmark	Promote as benchmark
Moran's-I falsifier transfer	quark-gluon plasma diagnostic bracketing plus spatial-panel commuter-mobility paper	Transfer a diagnostic/falsifier pattern into spatial-spillover robustness testing	Promote as process example

For such a human-utility benchmark, reviewers would score each candidate package on a fixed rubric: mechanism clarity, source grounding, novelty, falsifier quality, prior-art risk, expected review value, and experiment/proof readiness. A five-point Likert scale is sufficient, with an additional binary field for whether the reviewer would spend follow-up time on the candidate. The key metric is accepted or useful candidates per reviewer-hour, not raw idea count.

6 Two Operating Modes

The same machinery supports two modes: corpus-driven creative discovery and human-triggered problem solving.

In corpus-driven mode, the system starts from a scientific corpus and produces a creative frontier. The user or reviewer then allocates attention across promoted leads, benchmarks, prior-art collisions, and rejects.

In human-triggered mode, the user poses a natural-language problem. The system searches the existing creative corpus for mechanisms, analogies, and source opportunities and returns mechanism candidates rather than only conventional answers. A direct LLM response to the same problem is a useful baseline because it often collapses to known design families. The workflow is tested by whether it can move beyond that baseline into grounded mechanism transfer. This mode captures a common research reality: many discoveries begin as a human question, "Can we build a mechanism that has these properties?"

7 Case Evidence

The case studies are endpoint witnesses. They demonstrate that the workflow can produce real, reviewable, and sometimes implementable outcomes: a released protected-search mechanism, a selective-disclosure prototype, a corpus-scale creative frontier, and a cross-domain falsifier transfer.

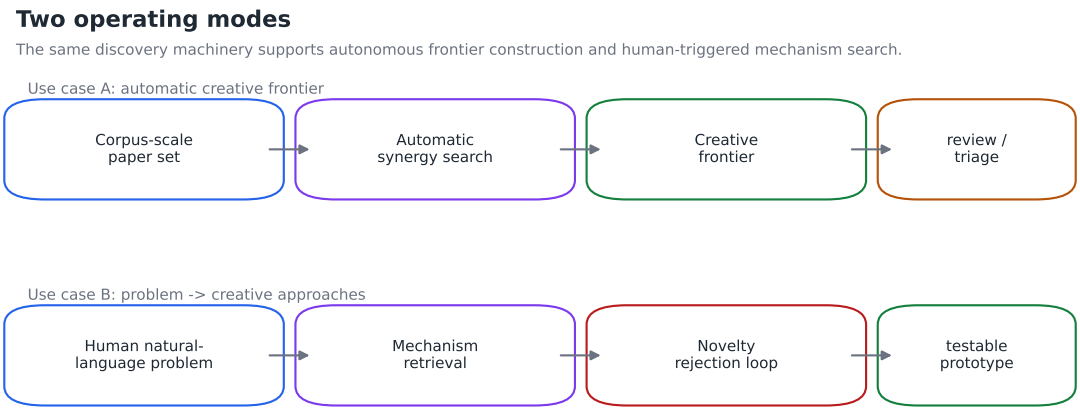

Figure 4: Two operating modes: corpus-driven creative frontier construction and human-triggered problem-to-mechanism search.

7.1 KPT as Released Outcome Witness

The first outcome witness is Keyed Phase Transform (KPT), a mechanism for protected semantic search that applies a key-derived transformation to semantic vectors so authorized queries preserve useful similarity structure while wrong-key access collapses into statistically unhelpful signal [11]. KPT is treated compactly here because it is documented separately. Its role is to show that a system-originated mechanism candidate can leave the ideation stage and re-enter hard validation.

We characterize KPT as a system-originated mechanism discovered via a human-operated prototype workflow, followed by human review, measurement, and a handoff to an internal Lean proof mechanism for selected formalized claims. The human role was operational and editorial: running the prototype, preserving artifacts, selecting the promising trail, reviewing outputs, measuring results, and integrating the separate technical paper.

The separate KPT report is the source for the technical mechanism and proofing details [11]. In this paper, KPT is used only as an outcome witness for the process: the creative workflow supplied a mechanism candidate, and formalizable parts of that mechanism were routed into an internal Lean-checking path. This is the same validation pattern highlighted by AlphaProof Nexus, but applied to a different part of the research loop: AlphaProof-style systems close formal proof targets; the creative discovery workflow supplies candidate targets.

KPT matters because it demonstrates a desired endpoint of the creative discovery process. A candidate mechanism became a measured artifact, was subjected to attack surfaces, and re-entered a formal verifier for selected claims. The result is a concrete proof point: the workflow generated a nontrivial mechanism and formalizable components of that mechanism re-entered hard verification. The claim boundary is important: the proofing path addresses selected theorem-layer statements, not a full end-to-end cryptographic security reduction.

7.2 Creative Frontier as Corpus-Scale Evidence

The creative frontier run is the central evaluation example. It shows that the workflow can produce a structured review object from a corpus rather than a single cherry-picked idea.

The central outcome is reviewable abundance. Promising leads can be promoted. Prior-art-

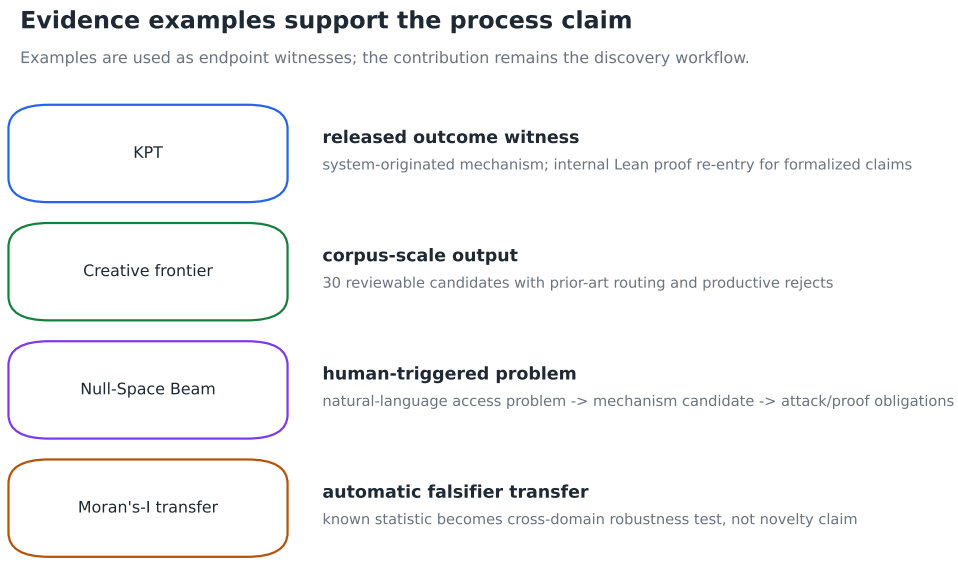

Figure 5: Evidence examples support the process claim without making any single case the whole contribution.

heavy ideas can be downgraded. Known-boundary rediscoveries can be used to demonstrate grounding. Unsupported bridges can be rejected. This is a stronger process claim than "the model generated a good idea." It shows that the workflow produces a topology of creative possibilities, including negative and boundary cases. That topology is what human reviewers need when deciding where to spend scarce attention.

7.3 Null-Space Beam Disclosure as Human-Triggered Problem Solving

The second operating mode is illustrated by a human-triggered access-control problem. The user asked for an encryption or disclosure technique for a payload $\{a, b, c, d\}$ where different keys reveal different partial views: one key might show b, c, another a, b, c, another only d, and a later key only b, without making the stored payload grow with each view policy.

As a baseline, a direct LLM answer to this problem produced a conventional attribute-based or envelope-style response: encrypt fields and distribute field-key wraps according to policies. We do not treat that answer as part of the discovery evidence. It is the comparison point. The framework run instead searched the creative corpus for distant mechanisms that could satisfy the same selective-disclosure pressure without merely relabeling known access-control encryption.

The follow-up idea search surfaced beamforming, wiretap secrecy, and null-space projection as analogies. The key mechanism cue was a wireless-security pattern: beamforming can privilege an intended receiver while leakage depends on spatial degrees of freedom, eavesdropper dimensionality, and channel-estimation error. The workflow transferred that rank-and-subspace intuition from physical channels to disclosure keys. The problem was reframed from "which policy wraps which field key?" to "which projection rows can be revealed while forbidden coordinates remain in the null space?"

Let $x \in \mathbb{F}_q^{p+m}$ be the concatenation of p payload slots and m random mask slots. Let $R \in GL_{p+m}(\mathbb{F}_q)$ be a per-record invertible mixing matrix. The stored vector is

$$y = Rx. \tag{1}$$

For a view S, let $P_S \in \mathbb{F}_q^{|S|\times(p+m)}$ be the projection matrix that selects the permitted payload coordinates and zeros the forbidden and mask coordinates. A view key is

$$K_S = P_S R^{-1}. \tag{2}$$

Applying the key reconstructs only selected payload slots. Forbidden payload and mask coordinates are zero in the key response matrix. The current prototype uses record-scoped private rotations rather than one reused global transform, which avoids the most obvious many-record global-matrix weakness in the toy mechanism.

The hard channel is key issuance. Repeated projection keys leak linear rows of the inverse transform. The prototype therefore includes a rank-bounded issuer: before returning key material, the issuer computes the rank of the span of previously issued keys plus the candidate key and denies issuance if the rank would exceed the configured disclosure budget. In a representative setup with 32 payload slots, 8 mask slots, and rank budget 24, the issuer releases controlled views for b, c, d, and later b; it denies a wider a, b, c view when that would raise exposure rank from 24 to 32. The stored envelope remains unchanged.

This case also illustrates the paper's operational definition of a falsifier. Here, the falsifier is an attack surface and measurable failure condition: cumulative key issuance must not reconstruct forbidden payload functions while remaining under the disclosure budget. The rank-budget denial is therefore not only an implementation detail; it is the local test that turns the idea into a reviewable mechanism candidate.

The open work is substantial and scientifically productive: formal security model, leakage function, collusion model, adaptive key-exposure game, production backend, padding, authenticated encryption for bytes, and KMS-managed per-record rotations. The value of the case is the creative process: a natural-language problem, a conventional direct-LLM baseline, framework-based distant analogy search, mechanism transfer, prototype, attack model, and explicit proof obligations.

7.4 Moran's-I as Automatic Falsifier Transfer

The third evidence case is an automatic falsifier transfer. In a corpus-driven run, the system linked a resolvent or quadratic diagnostic pattern from integrable systems with a Moran's-I and spatial-weight-matrix robustness test for spatial-spillover dominance. Human review narrowed the analogy into a concrete falsifier protocol.

The spatial-side claim concerned indirect spatial spillovers in a spatial Durbin panel model for commuter mobility. The candidate asked whether the dominance claim remains stable under multiple defensible spatial weight matrices and residual Moran's-I diagnostics. A minimal next experiment would re-estimate or approximate the model under several W definitions: multiple k-nearest-neighbor values around the reported choice, inverse-distance cutoffs, contiguity definitions where meaningful, and parameterized distance decay. If the direct/indirect decomposition changes materially under plausible W definitions, the dominance claim is fragile.

The creative mechanism is falsifier transfer: the system proposed a hard, local, testable challenge to a scientific claim by connecting distant diagnostic patterns and turning the analogy into an actionable robustness protocol.

8 Discussion

8.1 How LLMs Discover Novel Solutions

The workflow suggests a practical answer to the question of how LLMs discover novel solutions. They generate candidate structures beyond ordinary retrieval: reframings, analogies, mechanisms, tensions, and possible falsifiers. These structures become scientifically useful when they are

connected to evidence, graph search, prior-art routing, and hard re-entry points. The system does not stop at suggestion. It can pursue a candidate by deriving a prototype sketch, constructing a benchmark or attack suite, checking prior art, or translating a formalizable claim into a Lean proof task.

In this sense, emergence is a workflow property. The novel solution appears when document understanding, multi-perspective reframing, graph search, execution loops, and falsifier-centered review compound. The LLM supplies transformations of meaning; the surrounding agentic system turns those transformations into concrete research work: mechanisms to build, claims to prove, attacks to run, and experiments to prioritize.

8.2 Grounding Does Not Reduce Creativity

Grounding makes creativity usable. Evidence enables review; mechanism enables testing; prior-art routing turns novelty into a sharper claim instead of an accidental rediscovery.

The creative frontier evaluation supports this point. Known-boundary rediscoveries and prior-art collisions strengthen the process because they show that the workflow can route candidates honestly. Productive rejects are also useful because they reveal where a bridge is too distant until a formal mapping exists.

8.3 Human Direction and Agentic Execution

Human problems act as creative pressure, but the machine performs more than brainstorming. Researchers set the problem, standards, and accountability; the workflow performs a substantial part of the exploratory research labor by searching the corpus, constructing mechanisms, comparing alternatives, and turning candidates into prototype, benchmark, attack, or proof obligations. A concrete problem provides constraints against which conventional LLM answers and framework-discovered candidates can be compared.

The system improves the research frontier available to human reviewers. In the Null-Space Beam case, a direct LLM answer returned the conventional ABE/envelope-style family, while the framework run produced a null-space projection mechanism with a measurable key-exposure falsifier. In the creative frontier case, human review distinguishes promising leads from benchmarks, prior-art collisions, and rejects.

8.4 Cost-Stratified Model Allocation

The workflow also supports cost stratification. Broad document enrichment and deliberative artifact generation can be performed with cheaper local or locally hosted models; expensive models can be reserved for shortlisted candidates, prior-art synthesis, security review, and final draft critique.

This is the scaling principle: cheap breadth followed by expensive depth. Routine enrichment and broad candidate generation create the search frontier; expensive review is reserved for candidates that have already earned attention.

9 Boundaries and Research Agenda

The current evidence defines a clear next research agenda.

First, the early-yield result is already meaningful: roughly 10^3 enriched artifacts were sufficient to surface reviewable, nontrivial directions. The next systems milestone is 10^4 and 10^5 paper scale, where larger corpora should increase bridge opportunities while stressing deduplication, ranking, prior-art routing, and human-review bandwidth.

Second, reviewer utility can be quantified with a blinded expert-panel benchmark: compare grounded creative discovery against random pairing and zero-shot LLM ideation using domain

reviewers who score mechanism clarity, evidence grounding, novelty, falsifiability, and expected review value. This benchmark extends the artifact-level evidence reported here with a human-preference measurement.

Third, unsupported analogies are part of the frontier. The workflow already routes weak bridges as productive rejects; future work can make this routing more measurable by tracking reject reasons, prior-art collisions, and later candidate resurrection when new evidence appears.

Fourth, each outcome witness points to a different validation path. KPT demonstrates a system-originated mechanism with measurement and internal Lean proof re-entry for formalized claims. Null-Space Beam demonstrates a creative mechanism candidate and prototype that now needs a formal security model. Moran's-I demonstrates falsifier transfer that now needs a reproducible statistical robustness study.

10 Conclusion

Scientific discovery requires more than fluent synthesis. It requires novel solutions that are grounded, mechanistic, and falsifiable. This paper presented an agentic workflow showing how LLMs can discover such solution candidates from scientific documents and human research problems. The workflow transforms documents into enriched research objects, uses multi-perspective deliberation to generate candidate structures, searches multiple graph views for grounded hypotheses, and routes outputs into a creative frontier for review.

The evidence supports a positive claim: LLM creativity becomes operational when agentic workflows turn documents, anomalies, and human problems into reviewable, falsifier-ready solution candidates and then pursue them through automated research loops. The creative layer opens the search space; prototype construction, proof systems, experiments, attack suites, prior art, and human review close it.

The broader implication is organizational. In a research lab built around this workflow, humans are no longer the only source of scientific motion. They become directors of a larger discovery process: choosing valuable questions, setting standards, interpreting consequences, and accepting responsibility for final claims. The machines are not merely assistants in that process; they actively generate, investigate, test, prototype, and formalize candidate solutions. The result is not less human research, but a more ambitious form of research in which machines help create and validate the novel solutions worth human attention.

References

[1] Zhangir Azerbayev, Bartosz Piotrowski, Hailey Schoelkopf, Edward W. Ayers, Dragomir Radev, and Jeremy Avigad. Proofnet: Autoformalizing and formally proving undergraduate-level mathematics, 2023.

[2] Margaret A. Boden. Creativity and artificial intelligence. *Artificial Intelligence*, 103(1–2): 347–356, 1998. doi: 10.1016/S0004-3702(98)00055-1.

[3] Robert Burch. Peirce on abduction. Stanford Encyclopedia of Philosophy, 2012. URL https://plato.stanford.edu/archives/fall2012/entries/abduction/peirce.html.

[4] Ronald A. Finke, Thomas B. Ward, and Steven M. Smith. *Creative Cognition: Theory, Research, and Applications*. MIT Press, Cambridge, MA, 1992.

[5] Anna Jordanous. A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation*, 4(3): 246–279, 2012. doi: 10.1007/s12559-012-9156-1.

[6] Robert K. Merton and Elinor G. Barber. *The Travels and Adventures of Serendipity: A Study in Sociological Semantics and the Sociology of Science*. Princeton University Press, Princeton, NJ, 2004.

[7] Mark A. Runco and Garrett J. Jaeger. The standard definition of creativity. *Creativity Research Journal*, 24(1):92–96, 2012. doi: 10.1080/10400419.2012.650092.

[8] Don R. Swanson. Undiscovered public knowledge. *The Library Quarterly*, 56(2):103–118, 1986.

[9] George Tsoukalas, Jasper Lee, John Jennings, Jimmy Xin, Michelle Ding, Michael Jennings, Amitayush Thakur, and Swarat Chaudhuri. Putnambench: Evaluating neural theorem-provers on the putnam mathematical competition, 2024.

[10] George Tsoukalas, Anton Kovsharov, Sergey Shirobokov, Anja Surina, Moritz Firsching, Gergely Berczi, Francisco J. R. Ruiz, Arun Suggala, Adam Zsolt Wagner, Eric Wieser, Lei Yu, Aja Huang, Miklos Z. Horvath, Andrew Ferrauiolo, Henryk Michalewski, Codrut Grosu, Thomas Hubert, Matej Balog, Pushmeet Kohli, and Swarat Chaudhuri. Advancing mathematics research with ai-driven formal proof search, 2026.

[11] Nino Wagensonner. Keyed phase transform: Wavefunction collapse as secure search, June 2026. URL https://isar-innovations.dev/research/kpt-whitepaper-reviewed-2026-06-02.pdf. Reviewed technical report, Isar Innovations.

[12] Geraint A. Wiggins. A preliminary framework for description, analysis and comparison of creative systems. *Knowledge-Based Systems*, 19(7):449–458, 2006. doi: 10.1016/j.knosys.2006.04.009.

[13] Kaiyu Yang, Aidan M. Swope, Alex Gu, Rahul Chalamala, Peiyang Song, Shixing Yu, Saad Godil, Ryan Prenger, and Anima Anandkumar. Leandojo: Theorem proving with retrieval-augmented language models, 2023.

[14] Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. Minif2f: A cross-system benchmark for formal olympiad-level mathematics, 2021.