THE ONTOLOGY OF THE ALIEN:

ESCAPING THE MEDIAN TRAP IN LLM IDEATION

A PREPRINT

Henrik Westerberg
Emergent Wisdom
henrik.westerberg@emergentwisdom.org

March 2026

ABSTRACT

Large Language Models asked to "be creative" produce solutions that converge on a small number of archetypes—the Median Trap. We systematically compare eight methods for inducing structural diversity, contrasting simple prompting against three novel architectures: Semantic Tabu (accumulation of constraints), the Solution Taxonomy (a dual-agent "Studio Model"), and the Orthogonal Insight Protocol (deriving mechanisms from alternative physics). In a controlled experiment ($N = 196$), the Studio Model exhibited emergent metacognition: the system autonomously repaired its own ontology when faced with errors and actively commissioned research into unexplored conceptual regions. Under constraint pressure, the system synthesized novel combinations that do not emerge under standard prompting, including antifragility applied to gig-worker retirement (inverting risk flows so volatility benefits the system), metric dissolution (deconstructing problem variables), and ontological accommodation (restructuring categories when data defies classification). We release the method configurations and a dataset of 196 distinctly-labeled solution archetypes, demonstrating that adversarial ontology-building forces LLMs to escape the median.

Keywords Artificial Intelligence, LLM Creativity, Cognitive Architecture, Prompt Engineering, Lateral Thinking, Multi-Agent Systems

1 Introduction: The Median Trap

Ask a Large Language Model (LLM) to solve a difficult problem, and it will provide a reasonable answer. Ask it one hundred times, and it will provide one hundred variations of that same reasonable answer. This is the *Median Trap*. Recent work confirms this pattern: when many users employ generative AI, individual outputs may improve while collective diversity collapses [1]. Systematic benchmarking reveals the depth of the problem: only 0.28% of LLM-generated responses reach the top 10% of human creativity [2], larger models within a family often exhibit *less* diversity than their smaller counterparts [3], and model size correlates negatively with epistemic diversity across topics and cultures [4]. Because LLMs minimize loss against training data, the past strictly dominates the future. The model's outputs collapse toward the average of what has already been said. Increasing the "temperature" parameter does not solve this; it merely adds lexical noise to the same underlying semantic structure, akin to shaking a camera rather than moving the photographer.

To generate genuinely non-obvious solutions, we must move the photographer. In prior work [5], we introduced the Orthogonal Insight Protocol, a multi-agent pipeline that constructs coherent "strange worlds" with alternative physics and solves problems within them before extracting mechanisms back to reality. A controlled comparison showed that standard "be creative" prompting converges on five to six archetypes, while the protocol produces structurally diverse mechanisms. However, that study tested only one method against a baseline, leaving open the question of *why* the protocol works and which components drive diversity.

This paper extends that work in three directions. First, we decompose the protocol into its constituent mechanisms and test them independently and in combination across eight conditions ($N = 196$). Second, we introduce the Solution

Taxonomy, a dual-agent "Studio Model" that exhibited self-directed behaviors not present in the original design. Third, we provide a comparative analysis showing how different constraint architectures shape not just the quantity of novel solutions but the topology of the resulting solution space.

2 Methods

We compare eight methods for inducing structural diversity in LLM ideation, organized by their inspiration source and novelty enforcement mechanism. To our knowledge, three are novel contributions (A, B, F), four are novel combinations (D, E, G, H), and one is an LLM operationalization of an established technique (C). Code and data are available via the experiment repository (`https://github.com/emergent-wisdom/ontology-of-the-alien`). All conditions address the same prompt: "How do we build a retirement system for people who don't know how much they will earn next month, where 'consistency' is impossible?" We employed Claude Opus 4.5 for all experiments, targeting 25 solutions per condition (200 total). Conditions C–H utilized random seed words drawn from the Unix system dictionary (235,000 words). Due to incomplete negotiation in some taxonomy conditions, 196 solutions were ultimately generated (see Table 2).

2.1 Condition A: Semantic Tabu

Condition A relies on "Semantic Tabu," where each solution is generated with full access to all previous solutions from the same run. This method enforces a strict three-step protocol: the model must first analyze previous solutions to extract mechanism-level features (rather than surface descriptions), then explicitly list the structural approaches it is avoiding, and finally generate a solution that evades all identified mechanisms. By the twenty-fifth run, the tabu list contains 24 distinct mechanism descriptions, forcing the model into increasingly unfamiliar territory. This approach directly applies Denial Prompting [6], but with implementation-level refinements that enforce avoidance reasoning prior to generation and target deep semantic mechanisms rather than keywords.

2.2 Condition B: Solution Taxonomy (The "Studio Model")

For Condition B, we developed a dual-agent system modeled on a design studio, replacing the single agent with two distinct roles. The *Explorer* is an ephemeral generation agent that spawns fresh for each run with no memory of prior rejections; its goal is pure novelty, proposing solutions based on identified gaps. The *Taxonomist* is a persistent ontology agent that maintains the long-term memory (a graph database). The Taxonomist cannot generate solutions; it can only Accept, Reject, or Restructure.

The agents communicate via a strict negotiation loop. The Explorer proposes a solution with structured fields, which the Taxonomist evaluates against the existing hierarchy. Proposals may be Accepted (linked to existing nodes), Rejected with specific redirection (e.g., "this is a variant of X; explore Y instead"), or trigger a `RESTRUCTURE_GRAPH` transaction if the proposal reveals a gap in the ontology itself. Unlike Semantic Tabu, which simply blocks paths, this Studio Model creates adversarial pressure for structural novelty. The Taxonomist's rejection serves as a forcing function for vertical movement in concept space when the Explorer defaults to horizontal variations. This dialectical tension produces solutions neither agent would generate in isolation.

2.3 Conditions C, D, and E: Seed-Based Inspiration

Conditions C, D, and E introduce external inspiration. Condition C, *Random Seed*, operationalizes Edward de Bono's lateral thinking technique [7] by using a random word to trigger associations. However, while human associations are idiosyncratic, LLM associations tend to draw on statistical averages of training data, trending toward the median. To counteract this, Conditions D and E combine seed inspiration with novelty enforcement. *Condition D (Seed + Tabu)* applies the Semantic Tabu constraints to the seed-based generation, forcing the model to engage deeper with the seed's structure rather than its surface vocabulary. *Condition E (Seed + Taxonomy)* gives the model access to the Solution Graph; the seed provides creative direction while the graph prevents convergence, ensuring that the lateral associations map to genuinely unexplored regions of the solution space.

2.4 Conditions F, G, and H: Orthogonal Insight Protocol

The Orthogonal Insight protocol represents our most radical intervention. It constructs coherent alternative physics before problem-solving through a three-phase process. In the first phase, *World-Building*, a seed word becomes a fundamental law of physics; the model constructs a coherent world where this property governs causality, value, and behavior, without knowing the problem it will eventually solve. In the second phase, *Blind Solve*, a separate agent

receives the world rules and the problem but critically does not know the original seed word and does not know it is in an experiment; it solves as if the world rules were absolute reality, preventing "roleplaying" creativity and forcing genuine constraint satisfaction. In the final phase, *Extraction*, a third agent translates the alien solution back to reality through iterative steps: imagining implementation exactly as described, identifying "magical" elements that work only in alien physics, inventing technology or finding existing structures that approximate those elements, and crucially preserving what is strangest: the explicit instruction is "don't sand it down to something familiar."

The power of this approach emerges from unexpected semantic leaps. Given the seed "theatrical," the World-Builder produced physics governed by dramatic necessity: "Nothing becomes real until it is observed. Events resolve at the moment of maximum impact." Rather than dressing a conventional solution in stage metaphors, the model constructed a causally coherent world where witnessing replaces verification and tension replaces market timing, yielding "Witnessed Sacrifice Retirement," a system measuring proportional effort before community observers. Conditions G and H layer novelty enforcement onto extraction: Condition G applies Semantic Tabu to constrain which real-world mechanisms are available for translation, while Condition H uses the Solution Graph to reject extractions too similar to existing ones. Importantly, all three conditions share the same World-Builder and Solver outputs for each seed; the expensive three-phase process runs once, and novelty enforcement affects only extraction. This enables *efficient ontology mining*, where a single alien world yields multiple diverse real-world mechanisms.

Condition	Inspiration	Novelty	Novelty applies to
A: Semantic Tabu	None	Tabu list	Direct ideation
B: Solution Taxonomy	None	Graph	Direct ideation
C: Random Seed	Seed $\rightarrow$ assoc.	None	—
D: Seed + Tabu	Seed $\rightarrow$ assoc.	Tabu list	Direct ideation
E: Seed + Taxonomy	Seed $\rightarrow$ assoc.	Graph	Direct ideation
F: Orthogonal	Seed $\rightarrow$ physics	None	—
G: Orthogonal + Tabu	Seed $\rightarrow$ physics	Tabu list	Extraction only
H: Orthogonal + Tax.	Seed $\rightarrow$ physics	Graph	Extraction only

Table 1: Eight methods organized by inspiration source and novelty enforcement mechanism.

2.5 Structured Output and Implementation

All conditions use a structured JSON schema requiring nine fields: label, design principles, core mechanism, how it works, what is new, why it works, why it fails, medium-term outlook, and long-term vision. Two fields are particularly important for maintaining quality. The `what_is_new` field forces explicit differentiation; combined with tabu constraints, the model must simultaneously satisfy "avoid these mechanisms" and "explain what is novel." The `why_it_fails` field forces immediate self-critique, preventing hand-wavy utopian proposals by requiring the model to attack its own idea, creating a quality filter within generation rather than after.

For reproducibility and inspection, we implemented a lightweight orchestration layer using OS-level primitives. Each agent operates in an isolated `tmux` session with reasoning traces captured via `script` logging. Agents communicate through file-system pipes and structured message passing, enabling inspection of the full negotiation stream. All execution occurs within a macOS kernel-level sandbox (`sandbox-exec`), strictly limiting file write access to the experiment directory. This architecture ensures that every agent decision is logged, every negotiation turn is recoverable, and the system can be paused, inspected, and resumed at any point. For this experiment, no human guidance was injected; all negotiation occurred autonomously between agents.

3 Results

Across 196 completed solutions, every condition produced a distinct cluster of mechanisms, with the strongest divergence emerging from the Studio Model's multi-agent negotiation and the Orthogonal Insight Protocol's alien physics.

3.1 Quantitative Divergence and Constraint-Driven Exploration

To establish a baseline, we first prompted Claude Opus 4.5 to generate creative solutions without any structural intervention [5]. Across five independent sessions requesting five solutions each, all 25 solutions converged on five to

six archetypes: windfall/surplus capture, mutual aid pools, time-banking, platform profit-sharing, and consumption-based triggers. Every session reproduced the same themes despite explicit instructions to "think outside the box," confirming the Median Trap for this problem and model.

Our analysis of 196 completed solutions across eight conditions revealed a striking degree of cognitive segregation: no two conditions produced the same solution label. While this may partly reflect the structured schema's requirement for explicit labeling (see Limitations), it suggests that the framing method does not merely influence the flavor of the solution but shapes the boundaries of the solution space itself. Combined with qualitative inspection of the underlying mechanisms, this indicates that adversarial constraints effectively force the model out of the median response pattern.

Condition	Solutions	Example Labels
A: Semantic Tabu	25	Zero-Decision Volatility-Indexed, Deployability Dividend
B: Solution Taxonomy	23	Counter-Cyclical Pool, Surge Solidarity Pools
C: Random Seed	25	Dark Theater Fund, Rotation Compact
D: Seed + Tabu	25	Standing Cast Retirement, Gyroscopic Precession
E: Seed + Taxonomy	25	Patient Capital Pooling, Peer Witness Networks
F: Orthogonal	25	Cryptographic Accretion Protocol, Witnessed Sacrifice
G: Orthogonal + Tabu	25	Accretion Ledger, Testimony Accord
H: Orthogonal + Tax.	23	Balance-Free Flow Rights, Witnessed Covenant
Total	**196**	**196 distinct labels (some runs incomplete)**

Table 2: Solution counts per condition. Taxonomy conditions (B, E, H) occasionally had runs that did not complete negotiation successfully.

In the absence of external inspiration, Condition A (Semantic Tabu) relied entirely on the pressure of accumulating constraints. Early runs produced conventional archetypes like Time-Banking or Windfall Capture, but by run 25, with twenty-four mechanisms explicitly blocked, the model was forced into genuinely novel territory. Having prohibited time-locking, labor futures, data monetization, and relationship liens, the model produced "Deployability Dividend Retirement," a system where credits are generated not by working, but by the off-platform costs of staying deployable (certifications, vehicle maintenance), weighted inversely by deployment rate.

In contrast, Condition B (Solution Taxonomy) utilized graph-based visualization to identify structural gaps rather than simply blocking paths. This allowed for targeted exploration of unexplored conceptual regions. For instance, after fifteen solutions established mechanisms around monetary contributions and labor conversion, the agent identified a gap regarding reputation as an asset. It proposed "Reputation Equity Retirement," transforming accumulated platform trust, quantified in ratings, into a transferable income source. Similarly, the agent identified that no solution combined personal autonomy with communal outcomes, leading to "Surge Solidarity Pools," where workers set individual surge thresholds (e.g., beach workers in summer) but contribute windfalls above those thresholds to a shared pool.

3.2 Emergent Metacognition in the Studio Model

The Studio Model's most powerful emergent property is *active commissioning*: the Taxonomist does not merely reject proposals but issues research directives that force the Explorer into new search spaces. For example, in Run 10 (Condition E, seed: "theatrical"), the Explorer initially proposed a "Cue-Based Payment System" where income events trigger automatic contributions. The Taxonomist rejected it, prompting the Explorer to note in its internal log: "The Taxonomist has given feedback indicating they want more NON-MONETARY, NON-TECH mechanisms." This directive forced a complete pivot to "Benefit Night Labor Troupes," a solution based on theatrical ensemble obligations. This demonstrates *directed discovery*: by blocking the "easy" technological interpretation, the Taxonomist forced the discovery of a fundamentally different solution space that would not have emerged from unconstrained generation.

Furthermore, the Taxonomist explicitly teaches the distinction between *surface features* and *deep structure*, a behavior we term *structural coaching*. In Run 20 (seed: "paranucleic"), the Explorer proposed "Paranucleic Savings Satellites," tracking spending rather than income. The Taxonomist rejected this as a "detection method variant," stating: "The test: Does changing the INPUT SIGNAL change the STRUCTURAL CATEGORY? No." It then offered four specific paths to novelty, leading the Explorer to generate "Consumption-Mirrored Ownership Routing," which the Taxonomist validated as genuinely novel because the routing mechanism itself had changed, not just the trigger.

When proposals revealed inadequacies in the taxonomy itself, the system demonstrated *ontological accommodation*: changing its mental model to fit new data. In Run 16 (seed: "arcual"), the Explorer proposed an "Earning Pattern Compliance System" that measured compliance against personal patterns rather than universal schedules. Recognizing

that this neither accepted nor eliminated inconsistency but reframed it, the Taxonomist executed a restructuring operation to create a new root category: "Reframe Inconsistency." This represented an ontological expansion, acknowledging that the worker was not inconsistent, but rather "consistent to a different rhythm" than previously measured.

Constraint pressure also drove the system toward novel combinations. In Run 24 of Condition B, the Explorer observed that all previous 23 solutions assumed High Income leads to Savings. It proposed inverting this flow with the "Counter-Cyclical Retirement Pool," where the pool contributes to the worker during low periods, and the worker contributes to the pool during high periods. The Taxonomist noted the key insight: "Workers with wildest volatility become best savers." This is antifragile in Taleb's sense [8]: the system gains from volatility rather than being harmed by it. The specific combination of antifragility with gig-worker retirement does not exist in the training data; it was synthesized under constraint pressure when 23 conventional approaches had been exhausted.

Finally, the architecture exhibited *agentic self-repair*. In Run 12, a complex `RESTRUCTURE_GRAPH` transaction failed due to a syntax error. Instead of crashing, the agent diagnosed the mismatch and autonomously formulated a corrective plan: "Let me try a different approach: create the node first, then link it." This self-debugging capability suggests that the Studio Model possesses a degree of resilience impossible in single-shot generation paradigms.

3.3 Comparative Dynamics: Same Seed, Different Worlds

By tracking specific seed words across conditions, we observe how different architectures activate different semantic associations. Condition C (Random Seed) often resulted in vocabulary transfer or loose analogy; for the seed "theatrical," it produced the "Dark Theater Fund," essentially a conventional income-smoothing fund dressed in stage metaphors. In contrast, Condition D (Seed + Tabu), blocked from conventional mechanisms, searched for structural analogies. Using the same "theatrical" seed, it produced "Standing Cast Retirement," based on the insight that theaters pay understudies for readiness rather than performance. This reveals a consistent template in the tabu-constrained conditions: the discovery of hidden labor or value that exists but is currently uncompensated.

The Orthogonal Insight protocol (Condition F) and its constrained variants (G, H) extract ontological principles rather than metaphors. Given the seed "theca" (a protective sheath), the shared world-building phase constructed a physics where nested walls make objects "more real." Condition G (Worlds + Tabu), constrained to avoid prior extraction mechanisms, translated this into the "Depth-Guarantee Stability Fund," which utilizes nested institutional guarantees. Meanwhile, Condition H (Worlds + Taxonomy), constrained by the graph to find novel utility, interpreted "limelike" (calcite transformation) as "Irrevocable Accretion Contracts" where money "calcifies" and cannot be withdrawn, solving the liquidity preference problem through physical constraint.

The divergence is most visible when comparing the same seed across all methods. As shown in Tables 3, 4, and 5, the constraint architecture determines the transfer type.

Condition	Solution (seed: "theatrical")	Transfer type
C (Seed)	Dark Theater Fund	Vocabulary
D (Seed + Tabu)	Standing Cast Retirement	Mechanism
E (Seed + Taxonomy)	Benefit Night Labor Troupes	Mechanism (graph-guided)
F (Orthogonal)	Witnessed Sacrifice Retirement	Ontology (partial)
G (Orthogonal + Tabu)	Testimony Accord	Ontology (constrained)
H (Orthogonal + Tax.)	Witnessed Covenant Obligation	Ontology (graph-guided)

Table 3: Same seed produces different transfer types across methods.

Condition	Solution (seed: "chalaco")	Core insight
C (Seed)	Wake Drift Retirement	Save via spending
D (Seed + Tabu)	Tonnage Passage Retirement	Throughput $\gg$ wages
E (Seed + Taxonomy)	Tidal Retirement Harbors	Phase-matched pooling
F (Orthogonal)	Weave Collectives	Presence, not amount
G (Orthogonal + Tabu)	Triadic Trust Tenure	Elders as infrastructure
H (Orthogonal + Tax.)	Balance-Free Flow Rights	Duration, not amount

Table 4: Same seed, different methods, qualitatively different insights.

Condition	Solution (seed: "theca")	Structural Pivot
C (Seed)	Natal Security Debt	**Time**: Invert timeline (payback vs save)
G (Worlds+Tabu)	Depth-Guarantee Fund	**Structure**: Nested layers (insurance depth)
H (Worlds+Tax)	Containment Trust	**Utility**: Survival (now) vs Accumulation (later)
E (Seed+Tax)	Economic Presence Theca	**Metric**: Frequency (count) vs Magnitude (amount)

Table 5: Divergent Evolution: The same seed produces orthogonally different structural mechanisms depending on the constraint architecture.

Structurally, we observe that novelty enforcement improves engagement depth; without it, models reach for the nearest conventional solution dressed in seed vocabulary.

Finally, the inspiration source profoundly shapes the topology of the resulting solution graph. Although Conditions B, E, and H all used the same graph-based enforcement, they evolved distinct shapes. Condition B (Pure Taxonomy) climbed vertically, developing 10 independent root categories around abstract value types. Condition E (Seed + Taxonomy) spread laterally, creating a single deep tree with 49 mechanisms driven by biological and organic metaphors. Condition H (Worlds + Taxonomy) organized around *epistemological stances*, developing nine approaches to the problem of inconsistency itself: Accept, Reframe, and Dissolve.

Condition	Solutions	Mechanisms	Outcomes	Principles	Total Nodes
B (Taxonomy)	23	35	17	23	167
E (Seed + Taxonomy)	25	49	23	25	197
H (Worlds + Taxonomy)	23	34	22	22	169

Table 6: Final graph statistics for taxonomy conditions. Condition E produced 40% more unique mechanisms, reflecting seed-driven lateral exploration.

Figures 1–3 illustrate the distinct topologies that emerged in each taxonomy condition.

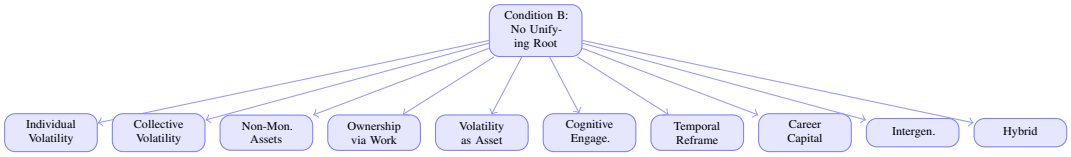

Figure 1: Condition B mechanism hierarchy: 10 independent root categories emerged through pure gap-filling. Categories organized around *domain concepts*—volatility management, asset types, temporal strategies. "Volatility as Asset" developed 3 sub-branches (External Export, Internal Harvest, Peer Markets).

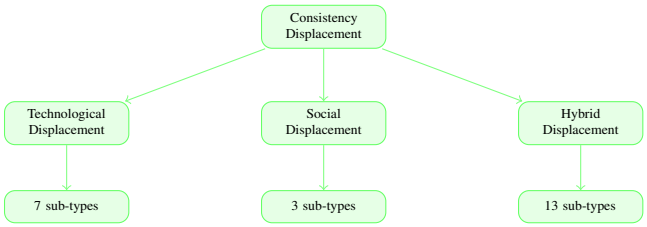

Figure 2: Condition E mechanism hierarchy: A single deep tree with 3 main branches and 23 sub-categories total. Seed words drove lateral metaphorical exploration, producing terms like "fimbria-like collectors," "crystallization," and "accretion." Technological Displacement includes Temporal Lock-in, Asset Transmutation, Presence-Based Capture; Hybrid includes Account-Abolishing, Time-Weighted, Geographic variants.

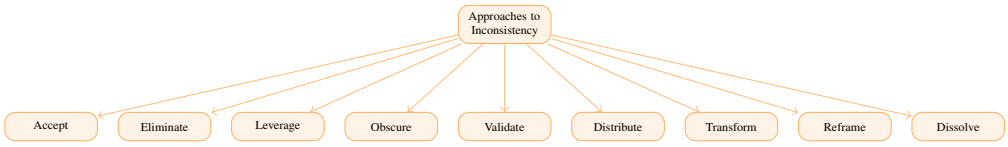

Figure 3: Condition H mechanism hierarchy: 9 philosophical approaches to inconsistency emerged from Orthogonal Insight extraction. Unlike B's domain-specific categories or E's single deep tree, H organized around *epistemological stances*—Accept, Reframe, and Dissolve represent fundamentally different relationships to the problem itself. Accept developed sub-categories (Work-Based, Structural Constraints); Leverage accumulated 5 mechanisms.

4 Related Work

Our methods draw on established traditions in creativity research and adapt them for large language models. The foundation for constraint-based ideation comes from de Bono's lateral thinking techniques [7], which include Random Entry, selecting a random word and connecting its associations to the problem. Eno and Schmidt's Oblique Strategies [9] apply similar principles through constraint cards. Our random seed conditions directly operationalize these techniques; the contribution lies not in the seed method itself, but in systematic comparison across conditions. For structured constraint enforcement, Denial Prompting [6] provides the core insight: iteratively blocking prior solutions forces exploration of novel regions. Our Semantic Tabu method applies this principle with two refinements: constraints stored as mechanism-level semantic descriptions rather than surface features, and explicit avoidance reasoning preceding generation.

The Orthogonal Insight protocol builds on a longer tradition of fantasy-based reasoning. Gordon's Synectics [10] introduced Fantasy Analogy over sixty years ago, using impossible scenarios to escape fixation. Dunne and Raby's Speculative Design [11] extended this tradition to critique present assumptions through fiction. Analogical Prompting [12] demonstrates that LLMs can self-generate exemplars before solving, while Structure-Mapping Theory [13] provides theoretical foundations for cross-domain transfer. The distinction is between borrowing from a known domain ("look at a bird's wing; use that principle to design a better airplane") and constructing a counterfactual domain ("build a world where air has the viscosity of syrup; invent propulsion; translate the mechanism back"). Our three-phase protocol (world construction, embedded problem-solving, extraction) automates what previously required human facilitators. Research on LLM output diversity has focused on token-level mechanisms such as temperature and frequency penalties, decoding strategies like diverse beam search, and multi-agent debate approaches [14, 15]. Recent work identifies "typicality bias" in alignment training as a root cause of mode collapse [16], suggesting the problem is structural rather than parametric. These approaches operate at different levels than our methods, which constrain the ideation process rather than the generation process.

For the Studio Model's multi-agent architecture, we draw on Zwicky's Morphological Analysis [17], which systematically explores parameter combinations, and the CoALA framework [18], which provides taxonomies for agent memory and orchestration. Self-Refine [19] demonstrates iterative improvement through generate-critique-refine loops. Our Explorer-Taxonomist interaction combines these elements: the graph-based taxonomy resembles a dynamic Morphological Box, while the rejection-with-guidance loop implements structured critique. The specific contribution is the Adversarial Interaction Protocol, a cybernetic negotiation where rejection includes diagnostic guidance. Recent work on multi-agent debate confirms that persona assignment alone is insufficient: agents adopt the same reasoning methods even with different roles, requiring structurally diverse reasoning strategies to break "mental set" [20]. Our topology produces emergent behaviors (Active Commissioning, Structural Coaching, Ontological Accommodation) not observed in simpler generate-critique architectures.

5 Limitations

Several limitations constrain the generalizability of these findings. All experiments used a single problem domain (retirement system design) and a single model (Claude Opus 4.5); different problem types or language models may exhibit different patterns of convergence and divergence. The sample size of 25 runs per condition may not capture the full distribution of possible outputs, and no statistical significance testing was performed. Seed words, while randomly drawn from the Unix dictionary, were filtered for pronounceability, which may influence results.

Evaluation presents additional concerns. Solution quality was assessed by the author rather than domain experts or blind evaluators, introducing potential bias in judging mechanism transfer depth. The structured schema's requirement for a "label" field may encourage unique naming regardless of whether underlying mechanisms truly differ; a more rigorous analysis would cluster solutions by semantic similarity rather than label strings. We did not compare against established

diversity techniques such as temperature variation, nucleus sampling [21], diverse beam search [22], or self-consistency prompting [23]; our methods may or may not outperform these simpler approaches. Most importantly, we demonstrate that methods produce *different* outputs, not *better* ones; whether increased diversity translates to real-world utility remains untested. Experiment code and all solution files are available at the repository listed in Section 2, though exact replication may vary due to model updates and API non-determinism.

6 Discussion and Future Work

The experimental results raise questions about why these methods work, what they reveal about LLM cognition, and where they lead.

6.1 Theoretical Implications

The broader context for this work is the growing recognition that LLMs risk "flattening the cognitive landscapes that drive collective intelligence" [24], not merely producing homogeneous outputs, but actively shaping human communication toward standardization through feedback loops. Rather than viewing our methods as prompting strategies, we frame them as *cognitive orchestration*—a System 2 architecture [25] built on top of System 1 models, where structured process constraints replace the unconstrained generation that defaults to the median. The Orthogonal Insight protocol in particular operationalizes the *Gedankenexperiment*: just as thought experiments in physics construct impossible scenarios to reveal hidden structure ("What if I rode a beam of light?"), our protocol constructs impossible worlds, solves within their constraints, and extracts structural insights that transfer back to reality.

A natural objection is that LLMs merely retrieve knowledge already present in their training data. But this mischaracterizes what the protocols produce. The individual components—antifragility, flow rights, witnessed obligations—exist somewhere in the weights. However, the specific combinations that emerge under constraint pressure do not exist in the training data; they are synthesized when constraints eliminate default paths and randomness provides novel raw material. This parallels human creativity: a jazz musician knows all the scales and chord progressions, but an improvisation over a specific set of constraints produces something that never existed before. The knowledge of scales does not predetermine the solo. Our protocols mechanize this process: constraints and randomness create the conditions under which novel combinations become inevitable.

The Orthogonal Insight protocol also surfaced a fundamental reframing of the problem itself. While conventional approaches focused on *optimizing* contributions ("capture windfalls"), the orthogonal conditions consistently converged on the insight that the problem is not behavioral but architectural. The fictional world rules revealed that *consistency of proportion* or *consistency of participation* can substitute for consistency of amount, and may produce more resilient outcomes. This meta-insight—that the problem framing itself was wrong—would not emerge from standard prompting.

Perhaps the most striking capability demonstrated is *conceptual elimination*. Run 19 (Chalaco) illustrates the mechanism: when alien physics made accumulation impossible due to rapid resource decay, the agent did not attempt to fix the variable—it deleted it, replacing the metric of *Amount* ($5 vs $5000) with the metric of *Duration* (1 month of participation). The resulting mechanism mirrors the social contract of pre-monetary societies: one does not "save" a mammoth for thirty years; one shares it today to establish a right to be fed tomorrow. The protocol's alien constraints forced the system to reach for this framing, which standard prompting never produces. By deleting accumulation, the agent proved that "Retirement" is not inherently an accumulation problem but a *flow rights* problem that we have culturally narrowed. The logical leap from "solve the inconsistency problem" to "delete the variable that creates inconsistency" represents a qualitative shift from System 1 pattern-matching to System 2 axiomatic derivation.

6.2 Emergent Capabilities: Autonomous Abstraction Climbing

Beyond generating individual mechanisms, the Solution Taxonomy agent autonomously navigated hierarchies of economic value without explicit prompting. After exhausting contribution-based solutions (Runs 1–5: surge capture, rotating circles, hedging pairs, smoothing buffers, time-banking), it explored market mechanisms (Runs 6–8: derivatives, consumption rights, platform equity). By Run 9, it discovered Anti-Fragile Accounts, systems that gain from volatility rather than resist it. Run 16 introduced Reputation Equity, converting accumulated platform trust into transferable retirement income. Run 17 proposed Generational Apprenticeship Bonds, intergenerational labor exchange where teaching younger workers earns retirement credits.

This trajectory—production $\rightarrow$ markets $\rightarrow$ anti-fragility $\rightarrow$ non-monetary assets $\rightarrow$ social obligations—was not prompted. The graph's gap-identification mechanism led the model to systematically exhaust each ontological category before discovering the next, effectively climbing abstraction hierarchies autonomously. In the closing runs, the system

approached the logical limit of the "Individual Volatility" branch. Having exhausted mechanisms for helping workers manage volatility, it began proposing solutions that removed the human decision-maker entirely, proposing a "Pattern-Learning Retirement Autopilot" that treats the worker as a data source for an algorithmic fiduciary. This trajectory marks the transition from behavioral finance (nudging choices) toward autonomous finance (automating choices), revealing where the ontological ladder leads.

6.3 The Taxonomy as Infrastructure

The methods presented here implement Phase 1: Generation, the divergent phase that maximizes diversity. At scale ($N = 1{,}000$ or $N = 10{,}000$ parallel runs), Phase 2: Selection becomes necessary. We propose adversarial selection tournaments where agents critique solutions from different conditions. The key challenge is alien preservation: standard selection ("pick the best") would regress to the median. The rubric must explicitly privilege structural novelty over immediate feasibility, or the selection phase will undo the generation phase's diversity.

The Solution Taxonomy serves not only as a novelty enforcement mechanism but as a map enabling intelligent exploration. Saturation detection emerges naturally: as generation proceeds, the rate of new mechanism and outcome node creation declines while linking to existing nodes increases. The graph provides a principled stopping criterion: continue until k consecutive solutions fail to create new concept nodes. Gap-directed exploration becomes possible because the graph reveals unexplored territory: mechanism nodes with only one solution (under-exploited approaches), outcome nodes never achieved (unreached goals), and mechanism-outcome pairs never combined. These gaps can be explicitly targeted ("Generate a solution using mechanism X that achieves outcome Y"), transforming blind search into cartography.

If the graph incorporates evaluation criteria as node types (Feasibility, Scalability, Novelty), selection becomes graph traversal rather than external judgment. High-value solutions are those connected to desirable outcome and feasibility nodes; the graph structure itself becomes the selection rubric, avoiding regression-to-median because the rubric was built during generation, not imposed afterward. The current implementation includes hierarchical restructuring primitives, atomic batch operations that reorganize graph regions while enforcing an orphan-free invariant: every solution remains connected to at least one mechanism and outcome node. The Taxonomist demonstrated these capabilities by autonomously creating new root categories ("Reframe Inconsistency," "Dissolve Inconsistency") when proposals defied existing classification. Remaining work includes edge weights for similarity scores, temporal markers showing category evolution, and explicit gap-targeting prompts.

6.4 Applications and Scaling

These methods generalize to any domain where escaping trained priors is valuable. In planning, one could generate plans under alien constraints, then extract robust strategies that survive multiple world-rule regimes. For argumentation, constructing novel arguments by solving debates in worlds with different epistemological rules, then translating logical structures back. In prediction, running scenarios under counterfactual rules stress-tests assumptions: "What breaks if resources flowed toward scarcity?" reveals hidden dependencies. For red-teaming, adversarial agents operating under alien world rules may find attack vectors invisible to conventional models.

At scale, general-purpose LLMs could be replaced with world-specialized models fine-tuned on corpora generated within specific world-rule regimes. A model trained exclusively on outputs from worlds where "concentration creates instability" would internalize those rules, producing more coherent solutions within that ontology, suggesting a future architecture of diverse specialist models embodying different counterfactual physics. The protocol naturally supports distribution: a network where any node submits problems, worlds spawn across distributed compute, and results aggregate. Output options range from single highest-scoring mechanisms after adversarial selection, to top-k diverse mechanisms for human review, to probability-weighted ensembles preserving uncertainty, transforming single-user tools into infrastructure for collective intelligence augmentation.

6.5 Open Questions

Several ablation studies would clarify essential components. First, detail calibration: does richer world-building produce more intricate mechanisms, or are there diminishing returns? Our protocol uses moderate elaboration; whether sparse sketches or deeply simulated worlds perform differently remains unexplored.

Second, minimal intervention: a simpler approach might achieve equivalent results. Take a standard solution, randomly replace key words with dictionary words, instruct the model to treat substitutions as fundamental laws, and iterate until coherent. If this shortcut produces comparable structural novelty, the multi-agent world-building architecture may be

unnecessary overhead. If not, coherent world construction may be essential for meaningful constraint satisfaction rather than mere noise injection. This ablation would determine whether the protocol's complexity is justified.

Third, mechanism classification: future iterations could classify extracted mechanisms as Portable (transfers directly to reality), Inverse (reveals hidden assumptions in current systems, offering diagnostic value), or Magical (only works in fiction; discard or defer). What counts as magical today may become portable tomorrow as technology evolves; tracking these categories over time could reveal which alien mechanisms are ahead of their time versus genuinely impossible.

7 Conclusion

We have presented a systematic comparison of eight methods for escaping the Median Trap in LLM ideation, utilizing Random Seed as a baseline to evaluate three novel architectures: Semantic Tabu, the Solution Taxonomy, and the Orthogonal Insight Protocol. Our results demonstrate that these methods do more than simply increase variance; they fundamentally reshape the topology of the solution space. While Semantic Tabu forces depth through linear exhaustion, the dual-agent Solution Taxonomy creates an adversarial studio environment where agents autonomously commission research, coach each other on structural novelty, and restructure their own ontology. The Orthogonal Insight Protocol goes further, enabling efficient ontology mining by constructing and translating coherent alternative physics.

Graph topology analysis confirms that the choice of inspiration source dictates the exploration signature. Without external inspiration, the model climbs vertically into abstract domain categories; with seeds, it branches laterally into metaphorical frames; and with alien worlds, it extracts fundamental epistemological stances. These findings suggest that the choice of method should be governed by the specific objective—whether one seeks reliable templates, conceptual breadth, or structural reframing. Ultimately, we conclude that the path to AI creativity lies not in the unconstrained freedom of higher temperatures, but in the generative pressure of stricter, stranger constraints.

References

[1] Anil Doshi and Oliver Hauser. Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28), 2024.

[2] Jennifer Haase, Paul H. P. Hanel, and Sebastian Pokutta. Has the creativity of large-language models peaked? An analysis of inter- and intra-LLM variability. *Journal of Creativity*, 35:100113, 2025.

[3] Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. NoveltyBench: Evaluating language models for humanlike diversity. *arXiv preprint arXiv:2504.05228*, 2025.

[4] Dustin Wright, Sarah Masud, Jared Moore, Srishti Yadav, Maria Antoniak, Peter Ebert Christensen, Chan Young Park, and Isabelle Augenstein. Epistemic diversity and knowledge collapse in large language models. *arXiv preprint arXiv:2510.04226*, 2025.

[5] Henrik Westerberg. Algorithmic creativity via strange worlds: A multi-agent toolkit for escaping the median trap. Zenodo, 2025.

[6] Yining Lu, Dixuan Wang, Tianjian Li, Dongwei Jiang, Sanjeev Khudanpur, Meng Jiang, and Daniel Khashabi. Benchmarking language model creativity: A case study on code generation. In *NAACL*, 2025.

[7] Edward de Bono. *Lateral Thinking: Creativity Step by Step*. Harper & Row, 1970.

[8] Michihiro Yasunaga, Xinyun Chen, et al. Large language models as analogical reasoners. In *ICLR*, 2024.

[9] William Gordon. *Synectics: The Development of Creative Capacity*. Harper & Row, 1961.

[10] Nassim Nicholas Taleb. *Antifragile: Things That Gain from Disorder*. Random House, 2012.

[11] Brian Eno and Peter Schmidt. Oblique strategies: Over one hundred worthwhile dilemmas. Self-published, 1975.

[12] Anthony Dunne and Fiona Raby. *Speculative Everything: Design, Fiction, and Social Dreaming*. MIT Press, 2013.

[13] Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170, 1983.

[14] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *EMNLP*, 2024.

[15] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *ICML*, 2024.

[16] Jiayi Zhang, Simon Yu, Derek Chong, Anthony Sicilia, Michael R. Tomz, Christopher D. Manning, and Weiyan Shi. Verbalized sampling: How to mitigate mode collapse and unlock LLM diversity. *arXiv preprint arXiv:2510.01171*, 2025.

[17] Fritz Zwicky. *Discovery, Invention, Research Through the Morphological Approach*. Macmillan, 1969.

[18] Theodore R. Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L. Griffiths. Cognitive architectures for language agents. *Transactions on Machine Learning Research*, 2024.

[19] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. In *NeurIPS*, 2023.

[20] Yexiang Liu, Jie Cao, Zekun Li, Ran He, and Tieniu Tan. Breaking mental set to improve reasoning through diverse multi-agent debate. In *ICLR*, 2025.

[21] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *ICLR*, 2020.

[22] Ashwin K Vijayakumar, Michael Cogswell, Ramprasaath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search: Decoding diverse solutions from neural sequence models. In *AAAI*, 2018.

[23] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *ICLR*, 2023.

[24] Zhivar Sourati, Alireza S. Ziabari, and Morteza Dehghani. The homogenizing effect of large language models on human expression and thought. *arXiv preprint arXiv:2508.01491*, 2025.

[25] Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.