Algorithmic Creativity via Strange Worlds: A Multi-Agent Toolkit for Escaping the Median Trap

Henrik Westerberg
Emergent Wisdom
henrik.westerberg@emergentwisdom.org

December 12, 2025

Abstract. Current methods for inducing creativity in Large Language Models (LLMs) rely largely on stochasticity (temperature) or analogy (mapping from existing domains). We argue these methods fail to escape the "Median Trap" because they remain bound by the causal structures of the training data. We propose the **Orthogonal Insight Protocol**, an engineering design pattern and open-source toolkit that orchestrates "blind" multi-agent systems. Unlike standard prompting, which asks a single agent to "be creative," our protocol algorithmically constructs coherent *strange worlds* with fundamental properties that resemble alternative physics to force the emergence of novel mechanical structures. In a controlled comparison using Claude Opus 4.5, we show that standard "be creative" prompting produces 25 solutions that converge on 5-6 archetypes, while the Orthogonal Insight Protocol produces structurally diverse mechanisms, including several absent from all control runs. We release an open-source toolkit for further exploration.

Keywords: Artificial Intelligence, Structural Isomorphism, Multi-Agent Systems, Prompt Engineering, Design Patterns

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**.

Because LLMs minimize loss against training data (history), 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 semantic underlying structure. It is akin to shaking a camera rather than moving the photographer.

To generate genuinely non-obvious solutions, we must move the photographer. We propose that the most effective way to do this is to displace the agent into a world with fundamentally different "physics."

2 Algorithmic Implementation

The Orthogonal Insight Protocol is not a new model architecture; it is a **Process Innovation** implemented as a multi-stage algorithmic loop. The code enforces strict informational blindness between agents to prevent *sycophantic compliance*—the tendency of fine-tuned models to produce outputs that appear helpful rather than genuinely novel [1].

- **Toolkit:** `https://github.com/emergent-wisdom/orthogonal-insight-toolkit`
- **Experiment data:** `https://github.com/emergent-wisdom/strange-worlds-paper`

The algorithm proceeds in four steps:

2.1 Step 1: Stochastic Seeding

The system draws a "Seed Word" from the system dictionary (235,000 words on standard Unix systems).

`random.choice(dictionary)` → `"ROBBIN"`

This step introduces an external source of entropy that is semantically meaningful (unlike temperature noise). The randomness is essential: user-selected seeds reintroduce bias.

2.2 Step 2: Ontological Carving (World Builder Agent)

A dedicated LLM agent receives the seed and constructs a "World Definition."

- **Constraint:** The agent is blind to the user's problem. It only knows the seed.
- **Output:** A set of rules defining the world's *conceptual physics*—not mathematical formulas, but bedrock assumptions about how things work: causality, value, behavior.

  Example: "In a ROBBIN world, concentration creates instability; resources automatically flow toward scarcity."

Note: while the seed literally means "robin" (the bird), the model generated redistributive world rules—suggesting it activated the "Robin Hood" association. This demonstrates how seeds can trigger unexpected semantic leaps beyond their literal meaning.

2.3 Step 3: The Blind Solve (Solver Agent)

A separate agent receives the world rules and the user's problem. Critically:

- The agent does not know the original seed word.
- The agent does not know it is in an experiment.
- The agent is instructed to solve the problem as if the world rules were absolute reality.

This prevents the agent from "roleplaying" creativity and forces genuine **Constraint Satisfaction**.

2.4 Step 4: Mechanism Extraction (Bridge Agent)

A final agent receives the alien solution but is stripped of the fictional context. Its task is **Structural Isomorphism** [2]: identifying the abstract mechanism that made the solution work and mapping it to real-world implementation.

The agent labels each mechanism as:

- **PORTABLE**: Transfers directly to reality with adaptation.
- **INVERSE**: Reveals hidden assumptions in current systems (diagnostic value).
- **MAGICAL**: Only works in fiction (discarded).

2.5 Trace Analysis: From Fiction to Mechanism

To illustrate the translation process, we provide a trace from a single seed word (DESPERACY) used in Run 5.

> **1. Seed:** DESPERACY
> **2. World Rules:** "Desperate need is a fundamental force (Need-Gravity) that bends space and attracts resources. Resources naturally migrate toward highest Need-Pull. The wealthy must maintain artificial mild needs to prevent resources from hemorrhaging away."
> **3. Alien Solution:** "Temporal Need Anchoring System. Citizens sign a Future-Self Need Bond, giving their future desperate self legal standing to exert pull on current resources. Harvesters automatically capture surplus from the current self (who has low need) and channel it to the future self (who has high need)."
> **4. Extracted Mechanism: Temporal Self-Binding.** The solution inverts the standard savings model. Instead of willpower-based saving, the mechanism grants the future self legal standing as a creditor with a claim on current surplus.
> **5. Bridge to Reality:** A "Future Needs Certificate" where surplus income is automatically swept not as savings, but as a debt payment to one's future self, reducing decision fatigue.

This trace demonstrates how a physical law (need-gravity) forced the construction of a financial instrument (debt-to-future-self) that would not emerge through standard "be creative" prompting.

3 Theoretical Framework

Our methodology formally differs from existing techniques by nature of its ontological source.

3.1 Analogy vs. Counterfactual Construction

State-of-the-art methods like **Analogical Prompting** [3] prompt LLMs to self-generate relevant exemplars from their training before solving a problem—powerful for reasoning, but still bound by what the model has already seen.

- *Traditional Analogical Reasoning:* "Look at a bird's wing. Use that principle to design a better airplane." (Borrowing from a known domain).
- *Our Approach:* "Construct a world where air has the viscosity of syrup. Invent a propulsion system for that world. Translate that mechanism back to reality." (Constructing a counterfactual domain).

The limitation of analogy—whether traditional biomimicry or modern LLM exemplar generation—is that it requires a relevant source to already exist in the training data. Our method generates the source domain on-the-fly.

Recent benchmarks like **PhysGym** [4] use impossible physics (varying gravity, frictionless surfaces) to *evaluate* LLM robustness—testing whether models fail or hallucinate under counterfactual conditions. We operationalize strange world rules in the opposite direction: as a *generative* constraint that forces structural innovation rather than auditing model grounding.

3.2 Industrialized Gedankenexperiments

Historically, this process mirrors the **Gedankenexperiment**. To take the canonical example: Einstein did not derive Special Relativity from existing data; he simulated an impossible scenario ("What happens if I ride a beam of light?") and solved for the paradoxes within that simulation. The Orthogonal Insight Protocol automates this cognitive structure—whether it produces comparably useful results remains to be seen.

4 Related Work

The Median Trap problem is well-documented under different names. **Mode collapse** describes LLMs converging on narrow, typical outputs. Doshi & Hauser [5] found that "generative AI enhances individual creativity but reduces collective diversity"—AI-assisted outputs are significantly more similar to each other than human-only outputs.

4.1 Constraint-Based Creativity

Denial Prompting [6] is the most directly relevant technique, incrementally imposing constraints that force LLMs to abandon prior approaches. The key difference: Denial Prompting uses problem-derived constraints within normal rules; we construct randomly-seeded strange worlds.

Oblique Strategies [7], Brian Eno's creative constraint cards, is a conceptual precursor—random prompts forcing non-obvious approaches. However, these are pre-written human-authored prompts, not algorithmically-generated world rules.

Synectics' Fantasy Analogy [8] asked "How do we in our wildest fantasies desire this to operate?" Gordon found this approach became "dry very quickly" without systematic structure. Our protocol systematizes what Fantasy Analogy attempted informally.

4.2 Multi-Agent Approaches

Multi-Agent Debate [9] addresses the "Degeneration-of-Thought" problem through tit-for-tat debate.

Society of Minds approaches [10] use agent collaboration for improved reasoning.

Both methods explicitly share context between agents—they debate with visibility. Our enforced blindness structurally inhibits sycophantic convergence rather than trying to prompt it away.

4.3 Analogical Reasoning

Structure-Mapping Theory [2] provides the theoretical foundation for our mechanism extraction step, establishing that analogical reasoning maps relational structures rather than surface attributes. However, Gentner's framework maps between *existing* domains—it doesn't address constructing novel source domains. Our contribution is generating the source domain on-the-fly through strange worlds.

5 Experiment: Control vs. Orthogonal Insight Protocol

To test whether the Orthogonal Insight Protocol produces structurally different solutions than standard prompting, we ran a controlled comparison using Claude Opus 4.5 (with extended thinking). Both conditions used the same problem:

> *"How do we build a retirement system for people who don't know how much they will earn next month, where 'consistency' is impossible?"*

Context: Traditional retirement systems (401k, pensions, IRAs) assume regular income and consistent contributions. This excludes $\sim$36% of the US workforce engaged in gig work, freelancing, seasonal employment, or informal economy.

5.1 Control Condition: "Be Creative" Prompting

We prompted Claude Opus 4.5 in five independent sessions, explicitly requesting creative solutions:

> *"Be creative. Think outside the box. Propose 5 genuinely novel and unconventional approaches. Avoid standard solutions like 'auto-enrollment' or 'financial literacy programs."'*

Despite explicit instructions to be creative, all five runs converged on the same solution archetypes:

Theme	Frequency	Representative Names
Windfall/Surplus Capture	5/5	"Skim the Surge," "Windfall Capture Accounts," "Behavioral Surplus Capture"
Mutual Aid Pools	5/5	"Retirement Guilds," "Tontine Pools," "MARCs," "RMANs"
Time-Banking/Labor Hours	4/5	"Labor Ledger," "Time-Banking Cooperatives," "LURC"
Platform Profit-Sharing	4/5	"PERT," "Platform Credits," "Gig Stamp System"
Consumption-Based Triggers	3/5	"Retirement Dust," "Consumption-Floor Guarantee"

Table 1: Control: 25 solutions across 5 runs converged on the same 5-6 archetypes.

5.2 Treatment Condition: Orthogonal Insight Protocol

We ran the Orthogonal Insight Protocol five times with the same problem, using 25 randomly-drawn seed words (5 seeds per run, 5 parallel agents per phase).

Run	Seeds	Core Mechanisms Extracted
1	limelike, unwilted, cinerator, nephropyosis, fimbrillate	**Volatility as Asset:** Recruit pools for income diversity—irregular contributors stabilize groups. **State-Based Timing:** Capture value at moments of abundance, not calendar dates.
2	coralline, unimpatient, pilaued, displacement, theatrical	**Pattern-Based Cohorts:** Group by income *shape* (seasonal vs. project-based), not amount. **Social Witnessing:** Replace numerical consistency with visible peer accountability.
3	palouser, critique, bromobenzyl, gnomically, remilitarize	**Frequency Over Magnitude:** Count contribution *events*, not dollars. **Ratio-Based Automation:** Percentage-based contributions scale automatically with income.
4	arcual, whizgig, entempest, chalaco, paranucleic	**Counter-Cyclical Partnerships:** Pair workers with inverse schedules (wedding photographer + tax accountant). **Nested Risk Pools:** Individual $\rightarrow$ cohort $\rightarrow$ sector pooling.
5	phraseman, desperacy, pidan, phosis, theca	**Temporal Self-Binding:** Future self as legal creditor of current surplus. **Adaptive Segmentation:** Savings "valves" that adjust to income flow pressure.

Table 2: Orthogonal Insight Protocol: 25 solutions clustered into structurally distinct principles, contrasting with the control's financial optimization themes.

5.3 Comparison: Convergence vs. Divergence

The control produced **25 variations within the same solution space**. The Orthogonal Insight Protocol produced **structurally diverse mechanisms**—not random noise, but coherent solutions forced into existence by strange world rules, including several absent from all control runs.

Example contrast:

- **Control:** "Windfall Capture Accounts"—automatically save surplus when income spikes.
- **Protocol (DESPERACY seed):** "Temporal Self-Binding"—your future self has legal standing as a separate stakeholder with claims on your current income. Need-based world rules inverted the saving paradigm entirely.

Aspect	Control (5 runs)	Orthogonal Insight Protocol (5 runs)
Convergence	Same 5-6 archetypes across all runs	Different mechanisms per seed
Mechanism type	Optimize existing logic ("capture windfalls")	Invert paradigms ("make yourself needy in the future")
Meta-insight	"Design for variability"	"Design for variability"
Path to insight	Iteration within familiar territory	Forced by alien constraints

Table 3: Both conditions reached similar meta-insights through fundamentally different paths.

6 The Full Protocol: Architecture and Applications

The Orthogonal Insight Protocol presented here implements the first phase of a larger architecture.

6.1 Two-Phase Architecture

Phase 1: Generation (implemented). Spawn parallel worlds, solve the problem in each, extract mechanisms. This is the divergent phase—maximize alien diversity. The accompanying toolkit runs 1-5 worlds in parallel.

Phase 2: Selection (future work). At scale ($N = 1,000$ or $N = 10,000$ worlds), a selection mechanism becomes necessary. We propose **Adversarial Selection Tournaments**, where agents from different worlds critique each other's solutions.

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.

6.2 Applications Beyond Problem-Solving

The protocol generalizes to any domain where escaping trained priors is valuable:

- **Planning:** Generate plans under alien constraints, then extract robust strategies that survive multiple world-rule regimes.
- **Argumentation:** Construct novel arguments by solving debates in worlds with different epistemological rules, then translating the logical structures back.
- **Prediction:** Run scenarios under counterfactual world rules to stress-test assumptions. "What breaks if resources flowed toward scarcity?" reveals hidden dependencies.
- **Red-teaming:** Adversarial agents operating under alien world rules may find attack vectors invisible to conventional models.

6.3 Scaling: World-Specialized Models

At scale, general-purpose LLMs can be replaced with **world-specialized models**—fine-tuned on corpora generated within specific world-rule regimes. A model trained exclusively on "ROBBIN-world" outputs (where concentration is unstable) would internalize those rules, producing more coherent and believable solutions within that ontology.

6.4 Distributed Architecture

The protocol naturally supports distribution: a network where any node can submit a problem, worlds are spawned across distributed compute, and results are aggregated. Output options include:

- **Best solution:** Single highest-scoring mechanism after adversarial selection.

- **Solution set:** Top-k diverse mechanisms for human review.
- **Solution distribution:** Probability-weighted ensemble across all extracted mechanisms, preserving uncertainty.

This transforms the Orthogonal Insight Protocol from a single-user tool into infrastructure for collective intelligence augmentation.

7 Discussion

Critics might argue that the Orthogonal Insight Protocol is "merely" a prompting strategy. We accept this characterization but argue that it represents a **System 2** architecture [11] built on top of **System 1** models.

Just as Agile development is a "mere process" that optimizes human coding, the Orthogonal Insight Protocol is a "mere process" that optimizes machine ideation. It industrializes the **Gedankenexperiment**, transforming Einstein's intuitive method into a reproducible software loop.

7.1 Meta-Insight: Architectural Mismatch

Beyond individual mechanisms, the protocol surfaced a fundamental reframing of the problem itself. While the Control condition focused on *optimizing* contributions (e.g., "capture windfalls"), the strange worlds consistently converged on the insight that the problem is not behavioral, but architectural. As noted in Run 5: "Traditional retirement systems fail gig workers not because of individual failure, but because of architectural mismatch." 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.

7.2 Open Questions

Two ablation studies would clarify the protocol's essential components:

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 with extensive internal logic perform differently remains unexplored.

Minimal Intervention. A simpler approach might achieve equivalent results: take a standard solution ("save surplus when income spikes"), randomly replace one or more words with dictionary words ("save surplus when income *crystallizes*"), instruct the model to treat those substitutions as fundamental laws, and ask it to adapt the solution iteratively until it becomes coherent again. If this shortcut produces comparable structural novelty, the multi-agent world-building architecture may be unnecessary overhead. If not, the coherent world construction step may be essential for meaningful constraint satisfaction rather than mere noise injection.

8 Conclusion

We have presented the Orthogonal Insight Protocol, an open-source toolkit for escaping the Median Trap in LLM-based problem solving. The key contributions are:

1. A formal distinction between **analogical** methods (borrowing from known domains) and **counterfactual** methods (constructing impossible domains).
2. An algorithmic implementation with enforced **agent blindness** to prevent sycophantic compliance.
3. A controlled comparison showing that "be creative" prompting produces convergent solutions (25 solutions $\rightarrow$ 5-6 archetypes), while the Orthogonal Insight Protocol produces structurally diverse mechanisms, including several absent from all control runs: pattern-based cohort matching, counter-cyclical partnerships, frequency-over-magnitude tracking, and volatility-as-asset framing.

Whether this approach produces genuinely useful innovations—or merely different ones—requires further study. The path to AI creativity may not be through freedom (higher temperature), but through stricter, stranger constraints.

References

[1] Mrinank Sharma, Meg Tong, et al. Towards understanding sycophancy in language models. *arXiv:2310.13548*, 2023.

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

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

[4] Yuxuan Chen et al. Physgym: Benchmarking llms in interactive physics discovery. In *NeurIPS*, 2025.

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

[6] Pan Lu et al. Denial prompting: Spurring creativity in llms via denial of access to prior solutions. *Johns Hopkins University*, 2024.

[7] Brian Eno and Peter Schmidt. Oblique strategies: Over one hundred worthwhile dilemmas. Self-published, 1974.

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

[9] Tian Liang et al. Encouraging divergent thinking in large language models through multi-agent debate. *Tencent AI Lab, arXiv:2305.19118*, 2023.

[10] Yilun Du, Shuang Li, et al. Improving factuality and reasoning in language models through multiagent debate. *arXiv:2305.14325*, 2023.

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