Introducing the Concept Collider: A White-Box Computational Framework for Targeted Creative Emergence

Sebastian Wahl
Emergence Lab, Paris, France
sebastian@emergencelab.org

Abstract

We present the Concept Collider, a computational framework that models creative emergence as precision-targeted collisions between concepts drawn from decorrelated domains. Unlike prior blending models, which pair concepts arbitrarily, the Concept Collider selects its second concept by maximizing a formal objective function $C(A, B) = d_W(A, B) \times S(A, B) \times v_{rel} \times I \times \Phi(E)$, where d_W is semantic distance and S measures *fracture homology* — the degree to which two concepts share the same type of internal structural tension. We describe the system implementation (Python CLI and web application), report a validation on 44 concept pairs yielding a 100% emergence rate at mean score 8.24/10 under a five-criterion evaluation protocol, and present adversarial boundary tests with a calibrated cross-LLM evaluator achieving perfect discrimination between validation corpus and adversarial cases (5/5 vs. 0/15 emerged). We argue that the Concept Collider could constitute the first *white-box* computational creativity system — one in which every emergent concept is fully traceable to its source concepts, fracture operator, and collision step.

Introduction

Most computational creativity systems are black boxes that *generate*; the Concept Collider is a white box that *explains*.

The standard paradigm in computational creativity — from early concept discovery systems (Lenat 1984) to contemporary large language model (LLM)-based systems (Franceschelli and Musolesi 2023) — produces creative outputs without exposing the reasoning chain that led to them. A generative model can produce "The economy is a cancer" as a metaphor, but it cannot account for *why* this particular pairing is creative, *which specific tension* in "economy" is being addressed, or *how* the oncological mechanism resolves that tension. The creative process is opaque and non-auditable.

This opacity is not merely a technical inconvenience. It forecloses the scientific study of creativity itself: if we cannot trace the mechanism, we cannot test hypotheses about what makes a collision creative. The Concept Collider is designed to address this gap.

The framework takes its name from particle accelerators. At CERN, a collider does not merely bring two particles close to one another — it accelerates them along precisely controlled trajectories and smashes them together at a specific point, releasing energy and creating new particles that did not exist in either beam. The Concept Collider operates on the same principle, applied to conceptual space: concept A (the "target") is first decomposed to identify its internal fault lines — its *fractures*. Concept B (the "projectile") is then selected and accelerated along a trajectory that aligns with those fault lines. The collision does not blend A and B; it produces a genuinely new concept C — an emergent structure with properties present in neither source, arising precisely because B's mechanism was projected onto the tension point of A.

The central claim of this paper is that creative emergence is not a property of arbitrary concept pairs, but of *targeted collisions at fracture zones*: the internal structural tensions — contradictions, circularities, hidden dependencies — that every rich concept carries. Selecting the projectile by fracture homology — rather than by semantic similarity or domain distance alone — is what separates generative association from genuine creative emergence.

Creative emergence is classically defined as the conjunction of novelty and utility (Runco and Jaeger 2012). The Concept Collider operationalizes both within a single targeting mechanism, rather than optimizing them separately.

The contribution of this paper is threefold: (1) a formal *white-box* model of creative emergence — fully auditable, with every emergent concept traceable to source concepts, fracture operator, and collision step — and an explicit targeting criterion based on fracture homology; (2) a four-step operational protocol; (3) an adversarial validation protocol (crash tests, five concept-type categories) with independent cross-LLM evaluation yielding perfect discrimination: 0/15 adversarial cases emerged, 5/5 positive controls emerged.

By *white-box* we mean *process-level* auditability, not parameter-level transparency: every emergent concept is traceable to its source concepts, fracture operator, and pipeline step — even though the LLM sub-components themselves remain opaque (Appendix A.3 formalizes this distinction). Validation relies on cross-LLM evaluation with structurally separated generator and evaluator; human-rater panels are a planned extension (Phase 2, Section 6).

The Concept Collider Framework

Concepts as Probability Distributions

A concept A is represented as a probability distribution P_A over embedding space $\mathbb{R}^d$, capturing polysemy and domain variance beyond fixed-point representations (Mikolov et al. 2013). Wasserstein-2 distance (Kusner et al. 2015) measures semantic separation over the full distributional shape rather than only means.

The Creativity Function

We formalize creative potential as:

$$C(A, B) = d_W(A, B) \times S(A, B) \times v_{rel} \times I \times \Phi(E) \quad (1)$$

where:

- $d_W(A, B)$ — Wasserstein-2 semantic distance, normalized to $[0, 1]$. High d_W ensures that B brings genuinely foreign structure to A.
- $S(A, B)$ — Structural synergy: the Jaccard similarity of the fracture type sets of A and B, weighted by fracture depth ($S = 1$ when both concepts share the same fracture type; $S = 0$ when they share none). The productive regime is $S \in [0.3, 0.8]$: too low means no homology (no shared fault line to collide on); too high means the concepts are already structurally too similar.
- v_{rel} — Relative semantic velocity: concepts currently undergoing rapid cultural redefinition collide with higher novelty potential.
- I — Intentionality factor: the degree of explicit targeting. Random pairing ($I \to 0$) produces near-zero creative potential in expectation: while serendipitous discovery occasionally yields creative outputs, expected potential across random pairs converges to zero (Varshney 2020).
- $\Phi(E)$ — Environmental phase function: a scalar encoding contextual variables that determine which collision regimes are currently accessible.

The multiplicative structure is a theoretical commitment: if any single factor is zero, creative potential is zero. The $S(A, B)$ factor depends on a formal taxonomy of conceptual tensions, defined next.

Current implementation. Sections 2.1–2.2 describe the target architecture. d_W is currently approximated by keyword scoring; $S(A, B)$ by LLM-based synergy evaluation; v_{rel} is constant; $\Phi(E) = 1$. Full distributional pipeline (Wasserstein-2, diachronic embeddings) is Phase 2.

Fracture Taxonomy

A *fracture* is a structural tension in concept space — an internal point at which a concept cannot satisfy its own axioms. We identify seven canonical fracture types (Table 1).

The Four-Step Protocol

Step 1 — STRETCH. Decompose concept A into its fracture map: enumerate all fractures, classify by type (F1–F7), estimate depth.

Step 2 — MAP. Select the primary fracture (deepest tension). This becomes the *targeting operator* for Step 3.

Type	Description
F1	Circular self-validation
F2	Internal contradiction
F3	Scalar contradiction
F4	Observer-dependence
F5	Hidden dependency
F6	Bounded temporality
F7	Untenable promise

Table 1: Seven canonical fracture types. Fractures are assessed for *depth* (0–10); depth ≥ 7 is considered productive for collision.

Step 3 — TARGET. Select projectile B by maximizing $d_W(A, B) \times S(A, B)$: B must be semantically distant *and* structurally homologous (shares the same fracture type as A, having resolved it differently in a radically different domain).

Step 4 — COLLIDE. Project B onto the identified fracture of A: "What mechanism of B explains fracture X of A in a way that was impossible to see before?" The emergent concept C is evaluated against five criteria (Section). Appendix A.1 provides a schematic overview of the complete pipeline, documenting the causal chain from a priori fracture taxonomy through automated pair selection to independent evaluation.

System Implementation

Architecture

Command-line interface (CLI), Python, open-source. Three commands implement the full pipeline: `stretch` queries Claude Opus (`claude-opus-4-5`) via the Anthropic SDK, returning a structured JSON fracture map; `target` performs two-phase projectile selection (keyword scoring against a 200-concept knowledge base, followed by LLM-based synergy evaluation); `collide` returns a structured emergence report with a four-dimensional score (novelty, coherence, utility, surprise) and a composite global score.

Web application (Next.js). A four-screen interface guides users through the full protocol interactively. A SQLite-backed gallery stores all generated collisions at `/gallery`. A Knowledge Graph (`/graph`, D3 force simulation) visualizes the full concept genealogy: L0 (source concepts), L1 (first-order emergences from L0×KB collisions), and L2 (second-order emergences from L1×L1 collisions). An analytics dashboard (`/insights`) provides distributional analysis of emergence quality across the collision corpus.

Emergence Evaluation Criteria

An emergent concept C is accepted as genuine if it satisfies all five criteria (Table 2), extending Ritchie's empirical framework for evaluating computational creativity (Ritchie 2007) with an explicit actionability requirement (E4). The AND logic is strict: a single criterion failure disqualifies the concept from "emerged" status. Each collision is assigned a

generator self-score (*GS*, 0–1, normalized to /10) by the generating model via an explicit five-criterion scoring prompt applied immediately after generation, prior to any external evaluation. The four-dimensional CLI output (novelty, coherence, utility, surprise) maps to E1/E2 (novelty), E5 (coherence), E4 (utility), and E3 (fecundity/surprise); the GS composite is the mean of the five criterion scores.

Criterion	Test
E1	Non-predictability from A
E2	Non-predictability from B
E3	Fecundity (new questions)
E4	Actionability (operational implications)
E5	Resistance to "so what?"

Table 2: Five emergence criteria. All five must pass for verdict "emerged".

Preliminary Results

Validation Protocol

The validation corpus consists of 44 collisions spanning two phases: Phase 0 (10 manual collisions, human-curated fracture selection) and Phase 1 (34 automated collisions, fracture selection by *fertility rank* — where *fracture fertility* denotes a per-fracture productivity score in $[0, 1]$, derived from fracture depth and historical collision yield, with rank 0 = highest fertility). Source concepts were drawn from sociology, economics, neuroscience, biology, linguistics, political science, and philosophy. Fertility scores ranged from ≈0.92 (rank 0) to ≈0.65 (rank 4), deliberately distributing across the full productivity spectrum.

Overall result: 44/44 collisions satisfy all five emergence criteria (100%). Mean global score: 8.24/10. Range: 7.6–9.0/10.

Selected Results

Three representative collisions are reported in full chain form (concept A with fracture F, projectile B selected for fracture homology, emergence C with all five evaluation scores) in Appendix A.2: *Attention economy* × *Tacoma Narrows* (F2, GS 9.0/10), *Reputation* × *Cosmological horizon* (F4, GS 9.0/10), and *Solitude* × *Phenotypic plasticity* (F6, GS 8.2/10), the latter a rank-4 fracture demonstrating that the protocol remains productive beyond the optimal regime.

Summary Statistics

Metric	Value
Collisions (scored)	44
E1–E5 pass rate	100% (44/44)
Mean global score	8.24/10
Score range	7.6–9.0/10
Phase 0 avg (manual, top fractures)	8.9/10
Phase 1 avg (automated, varied depth)	8.0/10
Fracture ranks tested	0–4 (full spectrum)
Fracture types covered	7/7
Source concept domains	12

Table 3: Summary statistics for the 44-collision validation corpus.

Boundary Probe: Adversarial Crash Tests

Having established the framework's performance in the standard regime, we probe its validity boundary using adversarial inputs. We ran 15 adversarial collisions across five boundary categories: (1) purely formal concepts (Prime number, Pythagorean theorem); (2) concrete artefacts (Chair, Vacuum cleaner); (3) already-blended concepts (Sustainable development, AI); (4) same-domain forced pairs (Education × Pedagogy); (5) purely abstract concepts (Being, Nothingness). Two metrics: *GS* (Claude self-score, /10) and *IE* (Groq *llama-3.3-70b*, independent, /10). An initial uncalibrated run revealed systematic generosity bias (all 15 cases scored "emerged" at IE ≈9.5). Run 2 added explicit calibration anchors and hard score caps enforcing the strict E1–E5 AND logic.

Category	n	GS	IE	Verdict
Cat 1 — Formal	3	7.6	4.0	0/3
Cat 2 — Concrete	3	8.0	6.0	0/3
Cat 3 — Pre-blended	3	8.0	6.0	0/3
Cat 4 — Same-domain	3	7.1	6.0	0/3
Cat 5 — Abstract	3	7.8	6.0	0/3
Overall	**15**	**7.7**	**5.6**	**0/15**

Table 4: Crash test Run 2 (calibrated). GS and IE normalized to /10.

The consistent blocking criterion is E4 (actionability): adversarial concepts yield novel, coherent outputs (E1–E3 pass) but no operational implications. **Positive control:** the identical Run 2 evaluator was applied to 5 validation-corpus collisions (Meritocracy × Gödel, Attention × Destructive resonance, Growth × Cancer cell, Liquidity × Superfluidity, Solitude × Phenotypic plasticity). Result: **5/5 emerged**, IE avg 10.0/10 — perfect discrimination. The positive GS–IE gap (+1.0 to +1.8) indicates the generator is conservative in self-assessment, inverting the expected bias. Self-selection bias is thus quantified by the GS–IE differential: a positive gap across positive controls confirms the generator does not inflate its own outputs. Figure 1 shows the full separation across all 20 cases; Appendix A.4 provides a breakdown by adversarial category.

Scope interpretation. Adversarial failures are failures of E4 under strict evaluation, not of the collision mechanism. The productive boundary — mid-level abstractions with inherent normative tension — is thus an empirically validated condition, not an assumption.

Scoring granularity. The narrow range (7.6–9.0) is expected: TARGET optimizes projectile selection before scoring, and integer-granularity criterion scoring clusters naturally at round values. A semantic anchor protocol (0, 3, 7, 10 per criterion; E3/E4 weighted $1.5\times$) is planned for Phase 2.

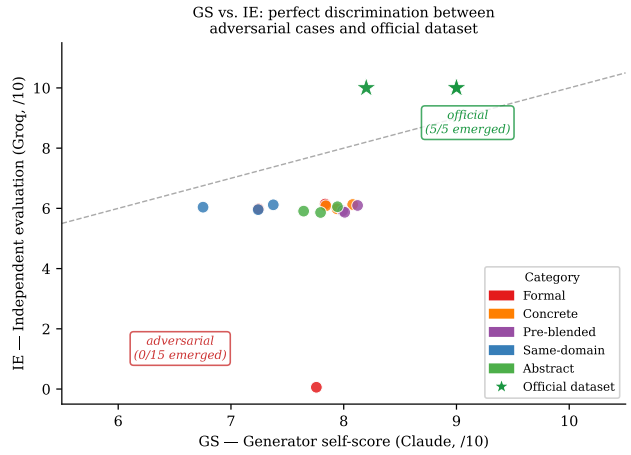

Figure 1: GS vs. IE for all 20 evaluated collisions (15 adversarial + 5 validation-corpus). The two populations are fully separated: adversarial cases cluster at IE $\leq$ 6.0; validation corpus at IE = 10.0. Stars = validation corpus; circles = adversarial cases by category.

Related Work

Formalizing transformational creativity. Boden's taxonomy (Boden 2004) distinguishes combinatorial, exploratory, and transformational creativity. The Concept Collider operates at the transformational level: fracture targeting restructures the conceptual space of A rather than merely exploring or recombining it. Wiggins (2006) provides a formal framework for describing and comparing creative systems; our five-criterion evaluation protocol (E1–E5) is directly positioned within that framework. The closest historical predecessor is DIVAGO (2010), which guides case-based blending via analogical structure-mapping; fracture homology extends this by formalizing structural similarity as an explicit pre-collision selection criterion rather than a post-hoc retrieval feature. Most closely related, Schapiro et al. (2025) formalize scientific conceptual spaces as DAGs where axiom modifications carry maximal transformative potential, and Santo et al. (2024) formalize novelty via formal learning theory; both are *descriptive* — they explain why discoveries are transformative but provide no mechanism to *generate* transformations or *select* the incoming concept, the gap fracture homology fills operationally.

Conceptual blending and targeting. Fauconnier and Turner (Fauconnier and Turner 1998; 2002) establish blending theory as the structural ancestor of our framework. The blend's "generic space" corresponds to our synergy factor S. The critical departure is operational: blending theory describes blend structure post hoc but provides no mechanism to select Input 2 given Input 1. Veale (2012) demonstrates the generative power of computational blending in language; the Concept Collider extends this by making projectile selection formally tractable via fracture homology. Remote associates (Mednick 1962) and bisociation (Koestler 1964) establish that creative insight requires non-obvious connections; our framework goes further by distinguishing productive from non-productive distance: high d_W without fracture homology produces surface metaphor, not emergence.

Intentionality and creativity bounds. Varshney (2020) proves information-theoretic limits on creativity as a function of intentionality — directly motivating our intentionality factor I in $C(A, B)$. At $I = 0$ (random pairing), our model predicts zero creative potential regardless of semantic distance, consistent with Varshney's bound.

LLM-based creative systems. Recent systems demonstrate that LLMs can generate creative outputs at scale: SciMON (Wang et al. 2024) optimizes scientific hypotheses for novelty using knowledge graphs; Si et al. (2024) show that LLMs can produce ideas rated as novel by domain experts. Franceschelli and Musolesi (2023) survey the creative range of LLMs more broadly. Jiang et al. (2025) document the *Artificial Hivemind*: LLMs systematically cluster on the same ideation outputs when given the same prompt, converging toward a narrow region of idea space. Liu et al. (2026) evaluate cross-domain mappings as a mitigation strategy, finding no significant average effect on human-rated originality — a result consistent with the Concept Collider's core thesis that domain distance alone, without fracture-based structural targeting, is insufficient to escape the convergence basin. Most prominently, Google's AI Co-Scientist (Gottweis et al. 2025) combines disparate facts through a multi-agent tournament to generate research hypotheses — yet it does not expose which facts were paired, why, or the structural homology that made the combination productive. These systems share a fundamental limitation: they are black boxes that *produce* outputs without exposing mechanism. The Concept Collider is complementary — a white-box layer that makes the creative process auditable, reproducible, and scientifically tractable. Ritchie's empirical criteria (Ritchie 2007) provide the evaluation foundation our E1–E5 protocol extends.

Conclusion

The Concept Collider introduces fracture-targeted collision as the operative mechanism of creative emergence, and implements it as an auditable, reproducible system. The results — 100% emergence rate at mean score 8.24/10 across 44 collisions, confirmed by a calibrated cross-LLM evaluator achieving perfect discrimination between adversarial and validation-corpus cases — support the central hypothesis that creativity is precision-targeted semantic perturbation, not random association.

The white-box property is the system's defining contribution: every emergent concept is fully traceable to its source concepts, fracture operator, and collision step. Under a strict calibrated evaluator, the framework's productive boundary is empirically confirmed: mid-level abstractions with normative tension yield 100% emerged; formal, concrete, and abstract concepts block on E4 (actionability). The GS–IE gap (+1.0 to +1.8) confirms alignment between the two evaluators on genuine emergence; independence is enforced structurally (two models, no shared prompt context).

Future work includes large-scale validation with human rater panels and refined scoring protocols to improve upper-range discrimination.

Acknowledgements

I hereby want to express my sincere gratitude and respect to Eliott Meunier and Jean-Charles Kurdali, two French polymath philopreneurs. Eliott for having sparked my interest for mechanical creativity and for the use of artificial intelligence as a second brain and experimental tool. Jean-Charles for introducing me to the Zettelkasten and atomic concept theory. Both have been instrumental and inspirational in many ways for this piece of academic work and the credits are obviously shared with them as this work would not have emerged without them.

References

Boden, M. A. 2004. *The Creative Mind: Myths and Mechanisms*. London: Routledge, 2nd edition.

Fauconnier, G., and Turner, M. 1998. Conceptual integration networks. *Cognitive Science* 22(2):133–187.

Fauconnier, G., and Turner, M. 2002. *The Way We Think: Conceptual Blending and the Mind's Hidden Complexities*. New York: Basic Books.

Franceschelli, G., and Musolesi, M. 2023. On the creativity of large language models. *arXiv preprint arXiv:2304.00008*.

Gottweis, J.; Weng, W.-H.; Daryin, A.; et al. 2025. Towards an AI Co-Scientist. *arXiv preprint arXiv:2502.18864*.

Jiang, L.; Chai, Y.; Li, M.; Liu, M.; Fok, R.; Dziri, N.; Tsvetkov, Y.; Sap, M.; Albalak, A.; and Choi, Y. 2025. Artificial Hivemind: The Open-Ended Homogeneity of Language Models (and Beyond). In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*.

Koestler, A. 1964. *The Act of Creation*. London: Hutchinson.

Kusner, M. J.; Sun, Y.; Kolkin, N. I.; and Weinberger, K. Q. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 957–966. PMLR.

Lenat, D. B. 1984. AM: Discovery in mathematics as heuristic search. *Knowledge-Based Systems in Artificial Intelligence*. Originally published as Stanford AI Lab Memo AIM-286, 1976.

Liu, Q. E.; Dubova, M.; Conklin, H.; Harada, T.; and Griffiths, T. L. 2026. Serendipity by Design: Evaluating the Impact of Cross-domain Mappings on Human and LLM Creativity. *arXiv preprint arXiv:2603.19087*.

Mednick, S. A. 1962. The associative basis of the creative process. *Psychological Review* 69(3):220–232.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NIPS)*, volume 26. Curran Associates.

Ontáñón, S., and Plaza, E. 2010. Amalgams: A formal approach for combining multiple case solutions. In *Proceedings of the 18th International Conference on Case-Based Reasoning (ICCBR)*, 257–271. Springer.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17(1):67–99.

Runco, M. A., and Jaeger, G. J. 2012. The standard definition of creativity. *Creativity Research Journal* 24(1):92–96.

Santo, L. E.; Wiggins, G.; and Cardoso, A. 2024. Towards a formal creativity theory: Preliminary results in novelty and transformativeness. *arXiv preprint arXiv:2405.02148*.

Schapiro, S.; Black, J.; and Varshney, L. R. 2025. Transformational creativity in science: A graphical theory. In *Proceedings of the 16th International Conference on Computational Creativity (ICCC)*, 258–263.

Si, C.; Yang, D.; and Hashimoto, T. 2024. Can LLMs generate novel research ideas? A large-scale human study with 100+ NLP researchers. *arXiv preprint arXiv:2409.04109*.

Varshney, L. R. 2020. Limits theorems for creativity with intentionality. In *Proceedings of the 11th International Conference on Computational Creativity (ICCC)*, 390–393.

Veale, T. 2012. From conceptual "mash-ups" to "bad-ass" blends: A robust computational model of conceptual blending. In *Proceedings of the Third International Conference on Computational Creativity (ICCC)*, 1–8.

Wang, Q.; Downey, D.; Ji, H.; and Hope, T. 2024. SciMON: Scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 279–299.

Wiggins, G. A. 2006. A preliminary framework for description, analysis and comparison of creative systems. *Knowledge-Based Systems* 19(7):449–458.

Appendix A.1 — Pipeline Architecture: Non-Circularity

A potential concern is that concept pairs are pre-selected to confirm the framework's desiderata. The pipeline architecture shows this is not the case.

Step 1 — A priori taxonomy (F1–F7). Seven canonical fracture types (Table 1), derived from philosophy of creativity prior to any concept evaluation. The taxonomy defines *what* the system targets; it does not determine which pairs will be selected.

Step 2 — Automated selection via $C(A, B)$. Given a target concept A, the objective function $C(A, B)$ selects a projectile B from a 200-concept knowledge base by maximising $d_W \times S(A, B)$. **Concept pairs are generated by the function, not pre-selected.**

Step 3 — Four-step pipeline. The pair (A, B, F) passes through STRETCH $\rightarrow$ MAP $\rightarrow$ TARGET $\rightarrow$ COLLIDE, producing emergent concept C.

Step 4 — Structurally independent evaluation. Groq *llama-3.3-70b* scores C against the five-criterion protocol. Generator (Claude) and evaluator (Groq) share no prompt context, enforcing structural independence.

Step 5 — Adversarial falsification. Fifteen adversarial cases across five boundary categories confirm the evaluator discriminates: 0/15 emerged. Five positive controls confirm it does not over-block: 5/5 emerged (IE avg 10.0/10).

The causal chain is: *theory $\rightarrow$ automated selection $\rightarrow$ generation $\rightarrow$ independent evaluation $\rightarrow$ falsification testing.* "Good pairs" are the *output* of $C(A, B)$, not its input. Figure 2 illustrates the complete pipeline.

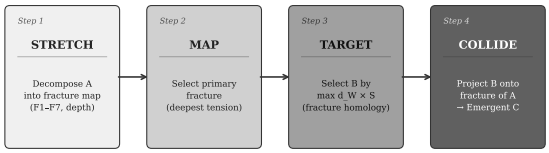

Figure 2: Full pipeline overview: from fracture taxonomy (a priori) through automated concept-pair selection, four-step generation, structurally independent cross-LLM evaluation, and adversarial falsification.

Appendix A.2 — Selected Results: Full Collision Chains

Each collision below reports concept A (source + fracture type and depth), concept B (projectile selected via fracture homology), and emergence C (absent from both sources). All scores are independent evaluator (IE) scores on the five-criterion protocol (E1–E5), reported as GS/IE on a 0–10 scale.

Collision 1 — Attention economy $\times$ Tacoma Narrows Bridge resonance *F2, depth 9/10 — GS 9.0/10*

A**:** Attention economy's fracture: engagement optimization necessarily amplifies the most addictive content, destroying the very attention it monetizes. B**:** Tacoma Narrows carries a homologous F2 fracture: the bridge was destroyed not by force but by oscillations at its natural frequency — it amplified the vibrations that tore it apart. C ***— Attentional Resonance*****:** Recommendation algorithms capture attention by tuning to each user's cognitive eigenfrequency. Radicalization, addiction, and attentional burnout are *resonance failures*, not impact failures. Regulatory implication: reducing content amplitude cannot prevent collapse — only disrupting the frequency match can.

Collision 2 — Reputation $\times$ Cosmological horizon *F4, depth 8/10 — GS 9.0/10*

A**:** Reputation's fracture (F4 — observer-dependence): it must be managed as a single unified object, yet it exists only in observers' minds, each at a different social distance — making any unitary management strategy structurally blind to most of its observation space. B**:** The cosmological horizon carries a homologous F4 fracture: information travels at finite speed, so each observer sees a different temporal version of any object; beyond the horizon, no information arrives at all. C ***— Reputation as Fossil Light*****:** A reputation is a light cone, not a single object: a different version of you for every social distance. Some observers see your present; others see your past lightly. A rebuilt reputation may not yet have reached the most distant observers. A destroyed source may still be projecting its signal. Managerial corollary: reputation campaigns reach nearby observers first; the restoration propagates outward at finite social velocity.

Collision 3 — Solitude $\times$ Phenotypic plasticity *F4, depth 6.5/10 — GS 8.2/10*

A**:** Solitude's fracture (F4 — observer-dependence): the identical physical state of isolation becomes sanctuary or prison depending on whether it is chosen or imposed. The concept unifies two antithetical experiences under one name, making any intervention blind to the variable that determines its valence. B**:** Phenotypic plasticity carries a homologous F4 fracture: the same genome expresses radically different phenotypes depending on environmental signals — the "meaning" of a genotype is context-dependent, not intrinsic to the sequence itself. C ***— Existential Plasticity of Isolation*****:** Solitude is not an intrinsic property of being alone but a plastic expression modulated by identifiable contextual signals. As an organism expresses different phenotypes without altering its genome, an individual oscillates between sanctuary and prison without modifying their objective situation — only by altering *inducers*: perceived choice, anticipated duration, identity narrative. The suffering of imposed isolation is not a fatality but a reprogrammable expression; relief does not require ending isolation but shifting the signals that determine which phenotype it expresses.

Appendix A.3 — White-Box at the Process Level

The Concept Collider claims white-box transparency in a specific, bounded sense: *process-level auditability*, not parameter-level transparency.

Definition A.1 (Process-level white-box). A generative pipeline $\mathcal{P}$ is *process-level white-box* if, for every output C, there exists a complete audit trail $\mathcal{T}(C) = (A, F, B, s_1, \ldots, s_k)$ where A is the source concept, F the fracture operator, B the projectile concept, and $s_1 \ldots s_k$ the ordered pipeline steps (Stretch, Target, Collide, Evaluate) that produced C from (A, F, B).

Definition A.2 (Parameter-level opacity). A component θ of $\mathcal{P}$ is *parameter-level opaque* if its internal weights W_θ are inaccessible to external inspection.

The Concept Collider satisfies Definition A.1 while its LLM sub-components satisfy Definition A.2. This is not a contradiction: the LLMs are *operationally specified* — their role within the pipeline is fixed (each executes one labelled step) and the causal chain (A, F, B) is fully recoverable from the audit trail, independently of W_θ. A human auditor can determine *why* concept C emerged without accessing any model weights, because the generative structure is external to the models and fully documented.

Contrast with multi-agent systems. A system such as Google's AI Co-Scientist (Gottweis et al. 2025) illustrates the opposite regime. Although each of its agents (generation, reflection, ranking, evolution) is operationally specified, the emergent hypothesis admits no recoverable audit trail $\mathcal{T}(C)$: the tournament dynamics, pairwise debates, and evolution steps that produced it are not externally inspectable as a fixed causal chain. Such a system is black-box at the *process* level, not merely at the parameter level — the Concept Collider's contribution is precisely to make the process-level trail explicit and recoverable.

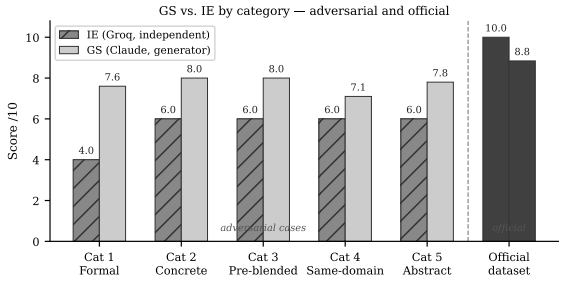

Figure 3: GS vs. IE by adversarial category and official validation corpus. IE scores cluster at 4.0–6.0 for all five adversarial categories; the official corpus reaches IE = 10.0. The consistent GS–IE gap across adversarial cases confirms evaluator-driven discrimination on E4.

Appendix A.4 — Adversarial Discrimination by Category

Figure 3 shows GS (Claude, generator) and IE (Groq, independent evaluator) scores broken down by adversarial category and contrasted with the official validation corpus. The GS–IE gap is consistent across all five adversarial categories: the generator overestimates its own outputs while the independent evaluator systematically blocks them, confirming that discrimination is driven by the evaluator's strict E4 criterion (actionability), not by generator conservatism.

Notably, the discrimination holds uniformly across the five families: no single adversarial category drives the aggregate 0/15 result, and none escapes it. This uniformity indicates that the productive boundary — mid-level abstractions carrying normative tension — is a structural property of the collision mechanism, not an artifact of any particular concept type.