PureTensor Inc
131 Continental Dr, Suite 305
Newark, DE 19713, US

Cross-Model Adversarial Synthesis:

Exploiting Latent Space Heterogeneity for Novel Knowledge Generation

Document: PT-R-2026-001
Version: 1.0
Date: 22 March 2026

PureTensor Inc, Research Division

PUBLIC RESEARCH PAPER

Abstract

Large language models trained independently on overlapping corpora develop fundamentally different internal representations of the same knowledge. We propose that structured adversarial interaction between multiple frontier models, mediated by human editorial judgment, can produce analytical outputs that exceed the capability of any individual model. We term this approach Cross-Model Adversarial Synthesis (CMAS) and present an initial experimental framework alongside a case study in theoretical physics -- specifically, black hole information theory -- that demonstrates the mechanism in practice. The case study involved three rounds of cross-model interaction between Claude Opus 4 (Anthropic) and ChatGPT Pro with o1-pro extended reasoning (OpenAI), with human-mediated citation verification and targeted rebuttal. The final synthesized output was judged by the human mediator to be superior to either model's independent contribution, with specific improvements traceable to cross-model adversarial pressure. We outline a theoretical basis for why this approach should be expected to yield novel insights, grounded in the observation that different training procedures induce different topological structures over latent knowledge spaces, and that cross-model interaction exposes connective paths unavailable within any single model's representation. This paper establishes the initial framework for a planned series of systematic investigations into the viability and limits of multi-model adversarial research.

1. Introduction

The dominant paradigm for using large language models in research and analysis is single-model interaction: a human poses questions to one model and iterates on the responses. This approach treats model selection as a discrete choice -- pick the best model for the task -- and discards the information contained in the diversity of available models. We argue this is wasteful.

Different large language models, even those trained on substantially overlapping corpora, develop meaningfully different internal representations of knowledge. Differences in architecture (dense vs. mixture-of-experts, decoder-only vs. encoder-decoder), tokenization strategy, training data curation, reinforcement learning procedures, and stochastic initialization produce neural networks that, while superficially similar in capability benchmarks, organize their learned knowledge into distinct topological structures. A concept-pair that is a short traversal in one model's embedding space may require a long, indirect path in another's.

This observation has a significant implication: the set of connections, analogies, and synthetic arguments that are natural or easy for one model to produce is different from the corresponding set for another model. If novel insight often arises from connecting previously unconnected ideas -- a claim well-supported by the history of science -- then

the cross-product of multiple models' connection sets may contain paths to insight that no single model would naturally traverse.

We propose a structured methodology, which we term Cross-Model Adversarial Synthesis (CMAS), that exploits this latent space diversity through a protocol of adversarial question generation, independent deep reasoning, critical cross-evaluation, targeted rebuttal, and iterative refinement. The human participant in this protocol serves not as a passive relay but as an editorial function: routing information between models, verifying claims against primary sources, applying selection pressure for quality, and recognizing signal when it emerges.

This paper is the first in a planned series of investigations by PureTensor Inc's research division. We present the theoretical motivation for CMAS, describe the experimental protocol, document an initial case study in black hole information theory, and outline the research programme that will follow.

2. Theoretical Framework

2.1 Latent Space Geometry and Training Divergence

A large language model's internal representation can be understood, at a high level, as a mapping from discrete token sequences to points in a high-dimensional continuous space. This mapping is learned through training and encodes the statistical relationships between concepts, facts, arguments, and reasoning patterns present in the training corpus. Crucially, this mapping is not unique. Two models trained on identical data with different random seeds will converge to different representations. When architectural differences, tokenization differences, data curation differences, and reinforcement learning differences are added, the resulting representations diverge further.

The geometry of these representations matters. In a transformer's latent space, related concepts cluster together, and the distance and direction between concept embeddings encode semantic relationships. The specific topology -- which concepts are neighbors, which are distant, which clusters are connected by smooth paths and which are separated by low-density regions -- is a function of the training procedure. It is this topology, not merely the set of facts encoded, that determines which connections are easy for the model to make (short paths, high-density regions) and which are hard (long paths, requiring traversal through low-density space).

We can formalize this loosely. Let M_A and M_B be two language models, and let L_A and L_B be their respective latent spaces. For any pair of concepts (c_i, c_j), define d_A(c_i, c_j) and d_B(c_i, c_j) as the effective distances in each model's latent space. The connection accessibility for a given model is inversely related to this distance. Our central claim is that for non-trivial concept pairs, d_A(c_i, c_j) and d_B(c_i, c_j) are imperfectly correlated. Some connections are easier in model A; others are easier in model B.

2.2 Novel Insight as Cross-Space Path Discovery

The history of science offers abundant examples of breakthrough insights arising from connections across previously isolated domains. Darwin connected Malthusian population economics to selective breeding to biogeographic observation. Shannon connected Boolean algebra to electrical switching circuits. Kahneman and Tversky connected cognitive psychology to economic decision theory. In each case, the key insight was not a new fact but a new path connecting existing facts.

We propose an analogy: each model's latent space represents a different disciplinary perspective on the same underlying knowledge. Just as interdisciplinary collaboration in human science exposes connections invisible from within any single discipline, cross-model interaction exposes connections that are inaccessible from within any single model's latent space topology.

The mechanism is specific. When Model A generates a claim that connects concepts (c_i, c_j), it does so because those concepts are proximate in L_A. Model B, presented with this claim, evaluates it using the geometry of L_B. If d_B(c_i, c_j) is large, Model B may challenge the connection, request justification, or propose an alternative path through a third concept c_k that is proximate to both c_i and c_j in L_B but not in L_A. The resulting exchange can surface a three-concept chain (c_i to c_k to c_j) that neither model would have produced independently.

2.3 The Role of Adversarial Pressure

Simple consensus-seeking between models (e.g., asking one model whether it agrees with another's answer) is unlikely to produce novel insight. Models trained on similar data will tend to converge on similar answers, and prompting for agreement reinforces this convergence. Adversarial pressure -- structured challenge, demand for justification, targeted critique -- is necessary to force models out of their default response distributions and into regions where their latent space differences become productive.

Adversarial pressure serves three functions in the CMAS protocol. First, error correction through uncorrelated failure modes. If two models are likely to fail on different aspects of a complex problem, cross-checking catches errors that self-review would miss. Second, forcing deeper reasoning. A model that receives a specific challenge must generate a more thorough response than one that receives generic approval. Adversarial feedback drives the target model to access less immediately accessible regions of its latent space. Third, exposing assumptions. Each model's default framing of a problem reflects the dominant framing in its training data. Adversarial challenge from a model with a different default framing forces both framings into explicit competition, potentially revealing a third framing that encompasses both.

2.4 The Human Editorial Function

The human mediator in CMAS is not a passive relay. The mediator performs several functions that current AI systems cannot reliably perform:

- Source verification: checking whether cited papers exist and support the claims attributed to them. Independent empirical verification is beyond current LLM capabilities.

- Signal recognition: identifying when a model's response contains a genuinely novel or interesting connection, as distinct from a fluent-sounding but vacuous one. This requires domain judgment.
- Strategic routing: deciding which model should respond to which challenge, when to introduce additional context, and when to terminate an exchange.
- Quality gating: preventing the accumulation of errors across rounds by verifying key claims before they become premises for subsequent reasoning.

The human mediator thus serves as a selection function in an evolutionary analogy: the models generate variation, the adversarial structure provides pressure, and the human selects for fitness. Whether this selection process can produce genuine novelty -- as biological evolution does -- is the central empirical question of this research programme.

3. Methodology: The AI Council Protocol

3.1 Protocol Design

The CMAS protocol consists of six phases, each with defined inputs, outputs, and quality criteria.

Phase 1: Adversarial Question Generation. A designated model (the challenger) generates a question designed to require multi-domain synthesis, formal reasoning, and awareness of open problems. The question should be at the frontier of current knowledge, requiring composition of ideas rather than retrieval of established results. The human mediator reviews the question for quality before proceeding.

Phase 2: Deep Reasoning Response. A second model (the responder) attempts to answer the question using extended reasoning capabilities where available. The responder is given no hints or framing from the challenger's perspective. This phase produces the initial substantive response.

Phase 3: Critical Cross-Evaluation. The challenger (or a third model) evaluates the response for correctness of cited claims, internal logical consistency, quality of formal reasoning, identification of weak points, and assessment of novelty. The evaluator produces a structured critique.

Phase 4: Human-Mediated Rebuttal. The human mediator verifies the evaluator's claims (especially citation checks and correctness assessments), augments the critique with domain knowledge, and crafts a targeted rebuttal for the responder. This is the critical quality-gating step.

Phase 5: Self-Correction and Synthesis. The responder addresses the rebuttal, correcting errors, strengthening weak arguments, and extending the analysis. The responder is explicitly asked to calibrate confidence across its claims.

Phase 6: Final Cross-Verification. All models review the final synthesized output for residual errors and overall coherence. The human mediator renders a final quality judgment.

3.2 Model Selection Criteria

Models are selected for maximum diversity across architecture (dense transformer, mixture-of-experts, differing layer counts), training data composition and curation methodology, reinforcement learning from human feedback (RLHF) procedures, reasoning modality (chain-of-thought, extended thinking, standard inference), and known strengths and weaknesses in relevant domains.

For the initial case study, we used Claude Opus 4 (Anthropic) and ChatGPT Pro with o1-pro extended reasoning (OpenAI). These models differ in architecture, training methodology, and reasoning approach, while sharing sufficient capability to engage meaningfully with frontier research topics.

3.3 Human Mediator Requirements

The human mediator must possess sufficient domain expertise to evaluate the quality of model outputs (though not necessarily at the level of a domain specialist), the ability to verify claims against primary sources, understanding of each model's known failure modes, judgment to distinguish genuine insight from fluent-sounding confabulation, and willingness to challenge both models' outputs when appropriate.

4. Experimental Case Study: Black Hole Information Theory

4.1 Domain Selection Rationale

We selected a problem at the intersection of algorithmic information theory, quantum gravity, and computational complexity for the initial case study. This domain was chosen because it requires genuine multi-domain synthesis (no single textbook covers the specific intersection of Kolmogorov complexity, holographic entropy bounds, quantum error correction in AdS/CFT, and computational complexity of Hawking radiation decoding); it contains verifiable claims published on ArXiv; it contains open problems requiring navigation of uncertainty rather than retrieval of consensus; and the specific composition requested is unlikely to appear verbatim in training data.

4.2 Experimental Procedure

Round 1: Adversarial Question Generation. Claude Opus 4 generated a three-part question requiring: (a) formalization of a tension between holographic entropy bounds and Kolmogorov complexity using the covariant entropy bound, with connections to the AMPS firewall paradox and computability of the S-matrix; (b) analysis of circuit complexity of behind-horizon bulk operator reconstruction in the context of quantum error correction, Python's Lunch, and Harlow-Hayden decoding hardness; and (c) whether computational intractability of Hawking radiation decoding is physical or model-relative, connecting to hypercomputation, spectral gap undecidability, and constructibility of the AdS/CFT dictionary.

Round 2: Deep Reasoning Response. ChatGPT Pro (o1-pro, extended reasoning mode) spent approximately 28 minutes on chain-of-thought reasoning before producing a comprehensive response. The response correctly identified the need to separate state count, description length, and decoding cost as three distinct concepts that the question's framing risked conflating. This structural correction was itself a significant analytical contribution.

Round 3: Critical Cross-Evaluation. Claude Opus 4 evaluated the response, identifying approximately 90% of the content as correct and well-reasoned. It flagged a 2026 Physical Review D paper citation as potentially hallucinated and identified several areas where the argument could be tightened, including the need to make the AMPS-as-complexity-problem point more explicit, to move the operator/geometry reconstruction distinction earlier, and to engage with the hardest version of the hypercomputer question.

Round 4: Human-Mediated Rebuttal. The human mediator verified citations against ArXiv, discovering that the flagged citation was in fact real: Mir et al., Physical Review D 113, L021904 (2026), arXiv:2601.22761. The mediator crafted a targeted five-point rebuttal acknowledging the citation's existence while maintaining that it should not be load-bearing, and pushing specifically on the weakest sections.

Round 5: Self-Correction and Synthesis. ChatGPT Pro produced a substantially improved response that corrected four specific overstatements identified in its own confidence calibration, properly separated operator reconstruction from geometry reconstruction complexity, engaged with the hardest version of the hypercomputer question (interior-only computation lost at the singularity), and produced a final trichotomy that was tighter than any formulation produced in earlier rounds.

4.3 Key Observations

Observation 1: Structural Framing. The three-way separation (state count, description length, decoding cost) that anchored the strongest version of the answer was present in ChatGPT Pro's initial response. This structural contribution was not created by the adversarial process but was correctly identified by Claude Opus as the most important element of the response, validating the evaluator's ability to recognize signal.

Observation 2: False Positive Correction. The evaluating model's flagging of a real citation as likely hallucinated was itself informative. It revealed a pattern-matching heuristic (recent papers cited with high specificity by LLMs are likely fabricated) that is useful but imperfect. The human mediator's verification step caught this error, preventing a valid result from being discarded. This demonstrates the necessity of the human quality-gating function.

Observation 3: Specificity of Adversarial Pressure. The most significant improvements in the final version occurred in areas where adversarial pressure was most specific. Generic praise produced no improvement; targeted critique (e.g., 'you dodged the most interesting version of the question') produced substantive new analysis.

Observation 4: Confidence Calibration. The self-reflective confidence calibration section, in which the model explicitly ranked its own claims by confidence level, was the single most

valuable component of the entire exchange. This capability was triggered by explicit demand in the rebuttal. The model did not volunteer it spontaneously, suggesting that adversarial pressure unlocks capabilities that default prompting does not.

Observation 5: Emergent Synthesis. The final formulation -- 'effective description vs. reconstruction task vs. causal accessibility' -- was not present in any individual model's first response. It emerged through the iterative adversarial process as a refinement of the initial three-way separation, informed by the rebuttal's demand for causal-access specificity. Whether this constitutes novelty or merely improved synthesis is addressed in the discussion.

5. Results and Analysis

5.1 Quality Assessment

We assess the output quality along four dimensions:

Citation accuracy: Of the approximately 11 distinct papers cited in the final response, 10 were verified as real papers supporting the claims attributed to them. One citation (Aaronson-Pollack) was described with a minor imprecision ('linear-time' shorthand for what is actually O(N^2) in boundary sites), which the model itself corrected in its confidence calibration. Overall citation fidelity: high.

Logical coherence: The final argument maintains internal consistency across all three parts. The separation of state count, description length, and decoding cost in part (a) properly grounds the complexity analysis in part (b) and the model-relativity discussion in part (c). No internal contradictions were identified.

Framing quality: The answer corrects the question's implicit conflation of Kolmogorov complexity with holographic entropy, which is a genuine contribution to clarity. The final trichotomy provides a clean organizing framework for the problem space.

Novelty: The individual claims are all present in the existing literature. However, the specific synthesis -- connecting Susskind's 'Horizons Protect Church-Turing' to the causal-access-dependence of computational hardness, then linking this to the Bouland-Fefferman-Vazirani dilemma as the boundary between physics and computational model choice -- is, to our knowledge, not explicitly articulated in any single published source. We characterize this as synthetic novelty: novel arrangement of existing ideas into a framework that provides new explanatory clarity.

5.2 Evidence of Cross-Model Value-Add

Three specific improvements in the final output are directly traceable to cross-model adversarial interaction:

- The reframing of AMPS as purely complexity-theoretic rather than information-theoretic. While present implicitly in the original response, this point was made explicit only after adversarial demand.

- The engagement with interior-only hypercomputation. The original response handled the hypercomputer case at a surface level. Targeted critique produced substantive new analysis connecting interior computation to Susskind's reformulation of the Extended Church-Turing thesis.
- The confidence calibration. Entirely absent from the initial response, this emerged only under explicit adversarial demand and proved to be the most valuable component of the final output.

5.3 Limitations

This case study has significant limitations that must be acknowledged. It is a single case in one domain, with no control condition comparing against single-model self-refinement. The human mediator's domain knowledge and editorial judgment are confounded with the cross-model effect; improvements may be attributable to human input rather than model diversity. Our assessment of synthetic novelty is subjective, lacking a formal metric for distinguishing genuine novelty from well-organized retrieval. Finally, we document a single successful case; failed attempts would not appear in this format, creating potential positive publication bias. These limitations define the research agenda for subsequent papers.

6. Discussion

6.1 Toward a Theory of Cross-Model Novelty

The central theoretical question is: under what conditions does cross-model synthesis produce genuinely new insight, as opposed to better-checked retrieval? We propose a tentative taxonomy:

Verification novelty: The cross-model process catches errors and improves accuracy. This is valuable but not novel in the intellectual sense.

Synthesis novelty: The process produces a framing or arrangement of existing ideas that provides new explanatory clarity. This is arguably novel in the same sense that a good review article or textbook chapter is novel. Our case study demonstrates this level.

Generative novelty: The process produces a claim, conjecture, or connection that is not present in the training data of any participating model. This is the strongest form and the most difficult to establish. We have not yet demonstrated this.

The theoretical argument from latent space diversity suggests that generative novelty should be possible in principle. If different models have different easy paths through knowledge space, their cross-product includes paths that are not easy for any single model. Whether these paths can lead to genuinely new destinations, rather than merely new routes to known destinations, remains an open empirical question.

6.2 Implications for AI-Assisted Research

If CMAS proves to be a reliable methodology, it has several implications. Adversarial multi-model pipelines could become a standard tool in systematic literature review, hypothesis generation, and theoretical synthesis, particularly in interdisciplinary domains. The human researcher's role shifts from question-asker to editorial function -- a role that requires domain expertise, source verification capability, and the judgment to recognize signal. This is a more demanding role, not a less demanding one. Organizations investing in diverse AI infrastructure -- multiple model providers, local inference capability, high-bandwidth interconnects -- gain research advantages over those locked into single-vendor relationships.

7. Future Research Directions

This paper establishes an initial framework. Subsequent papers in this series will address the following areas:

- Systematic domain evaluation: Repeating the CMAS protocol across multiple domains (mathematics, molecular biology, economic theory, philosophy of mind, materials science) to establish where the approach adds most value and where it fails.
- Control experiments: Rigorous comparison between CMAS and single-model self-refinement, single-model with human feedback, and multi-model consensus without adversarial structure. These controls are necessary to isolate the specific contribution of cross-model adversarial pressure.
- Formal novelty metrics: Development of quantitative metrics for output novelty, potentially drawing on information-theoretic measures of semantic distance from training data or blind expert evaluation protocols.
- Embedding space analysis: Direct measurement of latent space divergence between models using representational similarity analysis, centered kernel alignment, or related techniques to provide empirical grounding for the theoretical claim that different models have meaningfully different knowledge topologies.
- Automated adversarial protocols: Investigating whether the human mediator's role can be partially automated through a third mediator model that orchestrates the adversarial exchange, with human oversight limited to final quality gating.
- Scaling investigations: Testing the protocol with more than two models, longer adversarial chains, and domain-specialized models (e.g., including a protein structure model in a molecular biology CMAS session).
- Three-plus model configurations: Whether adding a third or fourth model produces diminishing returns or exposes qualitatively different failure modes. The combinatorial explosion of cross-model interactions may require new protocol designs.

8. Conclusion

We have presented Cross-Model Adversarial Synthesis (CMAS) as a structured methodology for exploiting the latent space diversity of independently trained large language models to produce analytical outputs of higher quality than any single model can achieve independently. The theoretical basis rests on the observation that different training procedures induce different topological structures over knowledge space, and that adversarial cross-model interaction can expose connective paths unavailable within any single model's representation.

Our initial case study in black hole information theory provides preliminary evidence that the approach works in practice: the final synthesized output, produced through three rounds of cross-model adversarial interaction, was judged superior to either model's independent contribution, with specific improvements directly traceable to the adversarial process. The final trichotomy ('effective description vs. reconstruction task vs. causal accessibility') emerged from the iterative process and was not present in any individual model's initial response.

Significant limitations remain. This is a single case study with no formal control condition and no quantitative novelty metric. The theoretical argument for generative novelty -- creation of genuinely new knowledge rather than improved synthesis of existing knowledge -- is plausible but unproven. The human mediator's contribution is confounded with the cross-model effect.

These limitations define the research programme ahead. Subsequent papers will address them systematically, with the goal of determining whether structured multi-model adversarial research can serve as a reliable methodology for accelerating scientific inquiry across domains.

References

[1] Harlow, D., & Hayden, P. (2013). Quantum Computation vs. Firewalls. Journal of High Energy Physics, 2013(6), 85. arXiv:1301.4504.

[2] Almheiri, A., Marolf, D., Polchinski, J., & Sully, J. (2013). Black Holes: Complementarity or Firewalls? Journal of High Energy Physics, 2013(2), 62. arXiv:1207.3123.

[3] Brown, A. R., Gharibyan, H., Penington, G., & Susskind, L. (2020). The Python's Lunch: Geometric Obstructions to Decoding Hawking Radiation. Journal of High Energy Physics, 2020(8), 121. arXiv:1912.00228.

[4] Bouland, A., Fefferman, B., & Vazirani, U. (2020). Computational Pseudorandomness, the Wormhole Growth Paradox, and Constraints on the AdS/CFT Duality. arXiv:1910.14646.

[5] Cubitt, T. S., Perez-Garcia, D., & Wolf, M. M. (2015). Undecidability of the Spectral Gap. Nature, 528, 207-211.

[6] Akers, C., Bouland, A., Chen, L., Kohler, T., Metger, T., & Vazirani, U. (2024). Holographic Pseudoentanglement and the Complexity of the AdS/CFT Dictionary. arXiv:2411.04978.

[7] Chen, C. F., Penington, G., & Salton, G. (2024). Entanglement Wedge Reconstruction via Universal Recovery Channels. Physical Review Letters, 132, 081801.

[8] Bousso, R. (2002). The Holographic Principle. Reviews of Modern Physics, 74(3), 825-874.

[9] Almheiri, A., Dong, X., & Harlow, D. (2015). Bulk Locality and Quantum Error Correction in AdS/CFT. Journal of High Energy Physics, 2015(4), 163. arXiv:1411.7041.

[10] Brakerski, Z. (2023). Black-Hole Radiation Decoding Is Quantum Cryptography. arXiv:2211.05491.

[11] Mir, S. A., Marino, F., Shabir, A., Krauss, L. M., & Faizal, M. (2026). Undecidability in Spacetime Geometry via the AdS/CFT Correspondence. Physical Review D, 113, L021904. arXiv:2601.22761.

[12] Susskind, L. (2020). Horizons Protect Church-Turing. arXiv:2003.01807.

[13] Vitanyi, P. (2001). Quantum Kolmogorov Complexity Based on Classical Descriptions. IEEE Transactions on Information Theory, 47(6), 2464-2479.

[14] Aaronson, S., & Pollack, J. (2022). Discrete Bulk Reconstruction. arXiv:2210.15601.

[15] Yoshida, B., & Kitaev, A. (2017). Efficient Decoding for the Hayden-Preskill Protocol. arXiv:1710.03363.

[16] Kornblith, S., Norouzi, M., Lee, H., & Hinton, G. (2019). Similarity of Neural Network Representations Revisited. Proceedings of the 36th International Conference on Machine Learning (ICML 2019).

[17] Li, K., Hopkins, A. K., Bau, D., Viegas, F., Pfister, H., & Wattenberg, M. (2023). Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task. ICLR 2023.

[18] Olah, C., et al. (2020). Zoom In: An Introduction to Circuits. Distill.

[19] Darwin, C. (1859). On the Origin of Species. John Murray, London.

[20] Shannon, C. E. (1948). A Mathematical Theory of Communication. Bell System Technical Journal, 27(3), 379-423.