Article

Can generative AI produce breakthrough innovative thinking? A critical analysis based on the Statistical-Counterfactual-Experimental tripartite gap and the Three-Engine Four-Layer Coupling model

WenJun Zhang
School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China
E-mail: wjzhang@iaees.org, zhwj@mail.sysu.edu.cn

Received 12 August 2026; *Accepted 9 September 2026*; *Published online 18 September 2026*; *Published 1 December 2026*

Abstract
Generative artificial intelligence has achieved remarkable capabilities in content generation, code synthesis, and scientific assistance. Yet the question of whether such systems can produce breakthrough innovative thinking remains unresolved. This paper conducts a comprehensive literature review and critical analysis to address this question. This study operationalizes breakthrough innovative thinking as a five-dimensional construct comprising questioning default assumptions, redefining problems, cross-domain connection, systemic reconstruction, and rapid experimental validation. Drawing on references, this study systematically evaluates generative AI's performance across these dimensions. It is found that that generative AI excels as a statistical association engine, demonstrating strength in cross-domain connection and combinatorial generation, but exhibits significant limitations in problem redefinition, causal intervention, experimental validation, and social value commitment. This study diagnoses the fundamental deficiency not as statistical or Bayesian learning per se, but as a structural mismatch between predictive objectives and the requirements of breakthrough innovation. The study proposes the Three-Engine Four-Layer Coupling model, which posits that breakthrough innovation emerges from the coupled operation of a statistical generation engine, a counterfactual reconstruction engine, and an experimental validation engine, operating across statistical, counterfactual, causal-experimental, and social-value layers. I formalize this model and derive testable hypotheses. It is concluded that generative AI functions as a powerful probability amplifier for breakthrough innovation but cannot autonomously serve as a breakthrough agent. The future path lies in constructing human-AI-experiment-society coupled hybrid systems.

Keywords generative AI; breakthrough innovation; first principles; counterfactual reasoning; Bayesian inference; causal inference; human-AI hybrid intelligence; computational creativity.

Selforganizology
ISSN 2410-0080
URL: http://www.iaees.org/publications/journals/selforganizology/online-version.asp
RSS: http://www.iaees.org/publications/journals/selforganizology /rss.xml
E-mail: selforganizology@iaees.org
Editor-in-Chief: WenJun Zhang
Publisher: International Academy of Ecology and Environmental Sciences

1 Introduction

1.1 Research Background

The rapid advancement of generative artificial intelligence (GAI) has transformed multiple domains of human activity (Ismayilzada et al., 2024; Zhang, 2026b-h, 2027a). Large language models demonstrate remarkable capabilities in natural language understanding and generation, code synthesis, and knowledge retrieval (Vaswani et al., 2017). Diffusion models produce images of striking fidelity through iterative denoising processes (Ho et al., 2020). Systems such as AlphaFold have achieved breakthrough results in protein structure prediction, with the AlphaFold Database now hosting approximately 214 million structures covering nearly the entire known protein space (Jumper et al., 2021; DeepMind, 2024). The 2024 Nobel Prize in Chemistry was awarded to scientists using AI for protein design and structure prediction, marking the formal arrival of the AI for Science era (Royal Swedish Academy of Sciences, 2024). In materials science, systems such as GNoME have predicted novel stable structures, and MatterGen has generated new materials with a 68% proportion of novel stable structures (Merchant et al., 2023; Zeni et al., 2025).

Breakthrough innovation occupies a qualitatively different position from incremental improvement (Schumpeter, 1976). Schumpeter's concept of creative destruction identified the continuous process by which new products and processes replace established ones as an essential feature of capitalist economies (Schumpeter, 1976; Aghion and Howitt, 1992). Schumpeterian growth theory has operationalized this notion by modeling the process through which new innovations replace older technologies (Aghion and Howitt, 1992; Acemoglu, 2009). Christensen's theory of disruptive innovation specified how new technologies initially serving niche markets can eventually displace established firms, with well-managed companies often failing precisely because they listen too carefully to their existing customers (Christensen, 1997; Christensen and Overdorf, 2000). Earlier work by Cooper and Schendel suggested that technological innovations can create new industries and transform or destroy existing ones, often pioneered by new firms (Cooper and Schendel, 1976). Abernathy and Clark developed a Transilience Map categorizing innovations as architectural, market niche, regular, or revolutionary, arguing that successful pursuit of these different types requires different organizational and managerial skill sets (Abernathy and Clark, 1985). Tushman and Anderson argued that technology evolves through incremental changes punctuated by occasional technological breakthroughs, with competency-destroying breakthroughs typically initiated by new firms (Tushman and Anderson, 1986). Henderson and Clark demonstrated that industry incumbents are often stumped by relatively minor technological improvements that do not fit their established architectural knowledge (Henderson and Clark, 1990).

A critical tension emerges from this (Boden, 1998). As generative AI systems grow more capable in generating content and solving well-defined problems, does this necessarily imply a corresponding increase in their capacity for breakthrough innovation (Boden, 1998; Runco, 2024). Reviews of AI creativity consistently find that while current models produce linguistically and artistically creative outputs, they struggle with tasks requiring creative problem-solving, abstract thinking, and compositionality (Ismayilzada et al., 2024; Franceschelli and Musolesi, 2024). Their generations suffer from a lack of diversity, originality, long-range coherence, and are prone to hallucinations (Ismayilzada et al., 2024). Runco argues that generative AI is potentially innovative but not truly creative, and that its processes suggest discovery rather than creativity (Runco, 2024). A recent analysis applying the standard definition of creativity to LLM outputs demonstrates that LLM creativity has a fundamental upper limit mathematically constrained to the boundary between little-c and Pro-c creativity (Cropley, 2025).

1.2 Problem Statement

This study addresses the central question of whether generative AI can produce breakthrough innovative

thinking (Boden, 1998; Runco, 2024). This question decomposes into several sub-questions: (1) is generative AI based solely on statistical learning (LeCun et al., 2015; Goodfellow et al., 2016). (2) Is it based solely on Bayesian inference (Griffiths et al., 2008; Tenenbaum et al., 2011; Zhang, 2012, 2016, 2018). (3) What are its fundamental deficiencies for breakthrough innovation (Boden, 1998; Cropley, 2025). (4) Can these deficiencies be remedied, and through what mechanisms (Zhang, 2016; Fukumura and Ito, 2025).

The distinction between statistical and Bayesian foundations matters because these two frameworks carry different implications for the possibility of genuine novelty (Pearl, 2009; Tenenbaum et al., 2011). Statistical learning, particularly in its connectionist manifestations, operates through pattern compression, interpolation, and extrapolation from training distributions (Zhang, 2010, 2012; LeCun et al., 2015). Bayesian inference, by contrast, involves explicit prior specification and posterior updating in light of evidence (Griffiths et al., 2008; Zhang, 2012, 2022). While generative AI certainly employs statistical mechanisms, the claim that it is purely Bayesian is an oversimplification that obscures the actual computational architecture and its limitations (Chen et al., 2026). Empirical studies demonstrate that LLMs do not perform explicit Bayesian inference when presented with all evidence at once, but instead carry out a form of implicit inference, and forcing a Bayesian update may in fact be harmful to task performance (Chen et al., 2026). LLMs follow Bayesian confirmation assumptions well with true evidence but fail to adhere to other Bayesian assumptions when encountering different evidence types (Kim et al., 2025).

The concept of breakthrough innovative thinking requires careful operationalization (Gero, 2000). This study defines it as a five-dimensional construct: questioning default assumptions, redefining problems, cross-domain connection, systemic reconstruction, and rapid experimental validation (Gero, 2000; Casakin, 2010). This definition draws on multiple traditions in creativity research, including the identification of first-principles reasoning as a creative design process alongside combination, transformation, analogy, and emergence (Gero, 2000), and the recognition that successful designing requires dismissing well-established ways of thinking in favor of creating new ones that address situational needs (Jung and Vartanian, 2018).

1.3 Core Concepts

Breakthrough innovative thinking involves the capacity to question default assumptions, redefine problems, connect disparate domains, reconstruct systems, and validate through rapid experimentation (Gero, 2000; Boden, 2004). This is distinct from incremental innovation, which improves existing solutions within established paradigms (Christensen, 1997; Abernathy and Clark, 1985).

Generative AI encompasses systems based on autoregressive transformers (Vaswani et al., 2017), diffusion models (Ho et al., 2020), generative adversarial networks (Goodfellow et al., 2014), and variational autoencoders (Kingma and Welling, 2014). These systems learn from data through self-supervised or unsupervised objectives, typically maximizing the likelihood of observed data or approximating its distribution (LeCun et al., 2015; Goodfellow et al., 2016).

Statistical learning refers to the process of extracting patterns and regularities from data, with the goal of generalization to new instances (Hastie et al., 2009; Zhang, 2010). Key mechanisms include representation learning, dimensionality reduction, and probabilistic modeling (LeCun et al., 2015).

Bayesian inference involves the explicit use of prior distributions and likelihood functions to compute posterior distributions over hypotheses or parameters (Griffiths et al., 2008; Tenenbaum et al., 2011; Zhang, 2012, 2016, 2018, 2022). While generative AI systems can be interpreted through a Bayesian lens, their actual training objectives and inference procedures differ in important ways (Chen et al., 2026).

1.4 Methodology

This study employs a multi-method approach combining systematic review, conceptual analysis, comparative case analysis, logical reasoning, and thought experiment (Booth et al., 2016; Creswell and Creswell, 2022).

The argument chain proceeds from definition to operationalization to capability assessment to deficiency diagnosis to theoretical construction to improvement pathways (Gero, 2000; Boden, 2004; Pearl, 2009).

1.5 Innovation and Structure

This study makes several original contributions: (1) it proposes the goal-hypothesis mismatch thesis, arguing that the fundamental deficiency of generative AI for breakthrough innovation lies not in its statistical nature but in the mismatch between its predictive objectives and the requirements of breakthrough innovation (Chen et al., 2026). (2) It develops the Three-Engine Four-Layer Coupling model, which specifies the coupled operation of statistical generation, counterfactual reconstruction, and experimental validation engines across statistical, counterfactual, causal-experimental, and social-value layers (Zhang, 2016; Zhang, 2010). (3) It proposes a human-AI hybrid breakthrough innovation architecture with associated evaluation metrics (Fukumura and Ito, 2025; Ito et al., 2025).

2 Systematic Review and Theoretical Foundations

2.1 The Theoretical Genealogy of Breakthrough Innovation

The study of breakthrough innovation has deep intellectual roots. Schumpeter's concept of creative destruction identified the process by which new combinations of productive means replace old ones as the fundamental driver of economic development (Schumpeter, 1976). Schumpeter considered such creative destruction to be an essential aspect of capitalism, driven by numerous microeconomic, firm-level decisions on strategic and technological considerations (Schumpeter, 1976; Aghion and Howitt, 1992). The Schumpeterian growth theory has operationalized this notion of creative destruction by modeling the process through which new innovations replace older technologies (Aghion and Howitt, 1992; Acemoglu, 2009).

Christensen's theory of disruptive innovation, developed in the 1990s, specified how new technologies initially serving niche markets can eventually displace established firms (Christensen, 1997). Christensen observed that well-managed companies often fail precisely because they listen too carefully to their existing customers and focus on sustaining innovations rather than disruptive ones (Christensen, 1997; Christensen and Overdorf, 2000). The theory has been widely applied and also subjected to criticism for its ambiguity and inconsistent empirical support (Christensen and Overdorf, 2000). A disruptive innovation is not a breakthrough innovation in the sense of being a sudden radical advance; rather, it is one in which a relatively simple product or service initially takes root among people who lack the expertise or money to use the dominant products or services in a given market (Horn, 2024). Christensen and Overdorf appear to be the first to use the term disruptive innovation rather than disruptive technology, and Christensen argued that whether a given technology or product is disruptive is determined not by its properties but by the perspective of the firms that adopt or do not adopt it (Christensen and Overdorf, 2000).

Earlier work by Cooper and Schendel suggested that technological innovations can create new industries and transform or destroy existing ones, often pioneered by new firms (Cooper and Schendel, 1976). Abernathy and Clark developed a Transilience Map categorizing innovations as architectural, market niche, regular, or revolutionary, arguing that successful pursuit of these different types requires different organizational and managerial skill sets (Abernathy and Clark, 1985). Tushman and Anderson argued that technology evolves through incremental changes punctuated by occasional technological breakthroughs, with competency-destroying breakthroughs typically initiated by new firms (Tushman and Anderson, 1986). Henderson and Clark demonstrated that industry incumbents are often stumped by relatively minor technological improvements that do not fit their established architectural knowledge (Henderson and Clark, 1990).

2.2 Cognitive Mechanisms of Human Breakthrough Innovative Thinking

The cognitive foundations of breakthrough innovation have been studied from multiple perspectives. Gero

identified five creative design sub-processes: combination, transformation, analogy, emergence (Gero, 2000; Zhang, 2027b), and first principles, the last referring to the use of existing knowledge in an abductive way (Gero, 2000). Goel distinguished between lateral transformation, where one idea yields to a slightly different idea, and vertical transformation, where an idea is further developed into a more detailed version (Goel, 1995). First-principles thinking has been described as a way of thinking and problem-solving that breaks down a complex problem into its most basic assumptions, facts, concepts, or ideas and then reassembles them from the bottom up (Aristotle, 1984; Tan and Xiao, 2025). This method has been shown to increase learning depth and promote creativity (Tan and Xiao, 2025). Popularized by Elon Musk, utilizing first-principles thinking to solve problems in an innovative, creative, and less biased way has proven popular in the tech community, with the benefit being to overcome the imitation phase faster and enable addressing problems from a different angle (Scrum.org, 2024).

Successful designing requires dismissing well-established ways of thinking in favor of creating new ones that address the needs of the situation, including design objectives and constraints (Jung and Vartanian, 2018). Analogies and metaphors provide mechanisms for this dismissal by enabling the recognition of correspondence between two items and the interpretation of a design feature through meanings related to a counterpart (Casakin, 2010). This process requires accessing and transferring structural information, including previously acquired knowledge of objects, attributes, relations, previous design solutions, and usage contexts (Casakin, 2010).

Boden's influential typology distinguishes among combinational, exploratory, and transformational creativity (Boden, 2004). Combinational creativity involves generating novel combinations of familiar ideas (Boden, 2004). Exploratory creativity involves exploring the space of possibilities defined by a particular conceptual framework (Boden, 2004). Transformational creativity involves altering the conceptual space itself by rejecting or replacing some of its defining rules (Boden, 2004). Transformational creativity is considered the most radical form and is closely associated with paradigm shifts and technological revolutions (Boden, 2004; Kuhn, 1962). Boden argued that some H-creative ideas have already been generated by AI programs, though usually by merely exploratory or combinational procedures, and that transformational AI originality is only just beginning (Boden, 1998). The reason why AI originality is only just beginning is its lack of domain-expertise, which is required for mapping the conceptual space that is to be explored or transformed, and its lack of valuation of results, which is especially necessary and especially difficult for transformational programs (Boden, 1998). Only a few AI models can critically judge their own original ideas, and hardly any can combine evaluation with transformation (Boden, 1998).

2.3 Technical Foundations of Generative AI

The technical architecture of modern generative AI rests on several key innovations. The Transformer architecture introduced self-attention mechanisms that enable parallel processing of sequential data and capture long-range dependencies (Vaswani et al., 2017). The Transformer relies solely on self-attention mechanisms, departing from conventional recurrent and convolutional models, and this approach enables the model to capture intricate relationships within sequential data without being hindered by the limitations of sequential processing (Vaswani et al., 2017). The self-attention mechanism allows the model to simultaneously weigh the significance of all input elements, producing a weighted sum that captures context more effectively (Vaswani et al., 2017). Multi-head self-attention employs multiple sets of learned linear projections of the input to compute attention weights in parallel, with each head capturing different relationships between elements (Vaswani et al., 2017). Self-supervised learning objectives, particularly masked language modeling and autoregressive prediction, allow training on vast unlabeled corpora (Devlin et al., 2019; Brown et al., 2020). Scaling laws describe how performance improves predictably with increases in model size, data, and compute

(Kaplan et al., 2020; Hoffmann et al., 2022).

Diffusion models generate data by learning to reverse a gradual noising process (Ho et al., 2020). The Denoising Diffusion Probabilistic Model presents high quality image synthesis results using diffusion probabilistic models, a class of latent variable models inspired by considerations from nonequilibrium thermodynamics (Ho et al., 2020). The best results are obtained by training on a weighted variational bound designed according to a novel connection between diffusion probabilistic models and denoising score matching with Langevin dynamics (Ho et al., 2020). These models naturally admit a progressive lossy decompression scheme that can be interpreted as a generalization of autoregressive decoding (Ho et al., 2020).

In terms of cognitive architecture, generative AI systems can be characterized as performing statistical compression, conditional sampling, prompt control, and feedback fine-tuning (LeCun et al., 2015; Brown et al., 2020). The core operation involves learning a compressed representation of the training distribution and then sampling from conditional distributions given prompts or contexts (Goodfellow et al., 2016). The decoder selects outputs from a restricted set of possible responses using statistical measures like top-p sampling, top-k sampling, or temperature scaling (Holtzman et al., 2020). Recent work has established that large language models are not mere statistical memorizers but are capable of in-context learning, performing inference at test time using only examples provided in the prompt without any parameter updates (Brown et al., 2020). This capability complicates simple characterizations of generative AI as purely statistical while also raising questions about the nature of the representations and computations involved (Brown et al., 2020; Wei et al., 2022).

2.4 AI Creativity and Scientific Discovery Research

The field of computational creativity has developed a substantial body of work on evaluating and understanding machine creativity. Evaluation approaches distinguish between artefact-based evaluation, which assesses whether results are creative, and process-based evaluation, which assesses whether the process is creative, using frameworks such as combinational, exploratory, and transformational creativity (Jordanous, 2012; Lamb et al., 2018). The International Association for Computational Creativity was established to promote the scientific study of human and machine creativity via computational means by fostering research in computational creativity (IACC, 2024).

Empirical studies have examined machine-generated creative outputs across multiple domains. An analysis of machine-generated poetry through Boden's models found that while initial computer-generated systems exhibited instances of combinational and exploratory creativity, current generative agents incorporate certain traits of transformational creativity (Kim et al., 2025). However, these systems often lack contextual awareness, emotional depth, and intentionality inherent in human-written texts (Kim et al., 2025). Research on machine-generated poetry through Boden's models evaluates the capability of machines in simulating human creative processes, finding evidence of combinational and exploratory creativity but limited transformational creativity. The OMEGA benchmark, designed to evaluate whether large language models can reason outside the box in mathematics, distinguishes among exploratory, compositional, and transformative generalization, with findings indicating that while these models can perform exploratory generalization by applying known strategies to new problems, compositional generalization remains limited, and transformative reasoning shows little to no improvement (Sun et al., 2025).

In scientific discovery, AI systems have achieved notable successes. AlphaFold2 solved the protein structure prediction problem, a challenge that had resisted solution for approximately fifty years, and its database now hosts over 200 million predicted structures (Jumper et al., 2021; DeepMind, 2024). The turning point was 2020, when AlphaFold showed that an AI model could solve a fundamental scientific problem that had resisted decades of effort (World Fund, 2026). The system has been used by over 3.7 million researchers

globally and has influenced more than 500,000 related papers (DeepMind, 2024). In materials science, systems such as GNoME have predicted novel stable structures, and MatterGen has generated new materials with a 68% proportion of novel stable structures (Merchant et al., 2023; Zeni et al., 2025). M3GNet, developed at UC San Diego, predicts the structure and dynamic properties of materials almost instantaneously and was used to develop a database of more than 31 million yet-to-be-synthesized materials (Chen and Ong, 2022). However, most predicted materials are still not directly usable, as a structure may look stable in simulation but be impossible to synthesise, rely on inaccessible inputs, or fail under real operating conditions (World Fund, 2026).

However, the distinction between AI-assisted breakthrough and AI-autonomous breakthrough remains crucial (Boden, 1998; Runco, 2024). Reviews of generative AI as a transformative force for innovation indicate significant potential to augment human creativity as a collaborative partner, but emphasize the need for responsible development and ethical frameworks to address biases and other challenges (Sedkaoui and Benaichouba, 2026). The findings indicate that despite the theory-driven approach to investigating generative AI's creative and innovative potential, cutting-edge applications research prioritizes examining the possibilities of generative AI models rather than their fundamental limitations (Sedkaoui and Benaichouba, 2026).

2.5 Research Gaps

Several significant gaps emerge from previous research: (1) there is a lack of systematic, layered assessment of generative AI's breakthrough innovation capabilities (Boden, 1998; Cropley, 2025). Existing studies typically examine specific domains or tasks rather than providing a comprehensive framework that can guide evaluation across the full spectrum of breakthrough innovation requirements (Ismayilzada et al., 2024). (2) There is insufficient rigorous analysis of whether statistical or Bayesian foundations constitute fundamental deficiencies (Chen et al., 2026). The literature often assumes rather than demonstrates that these computational mechanisms are inherently limited for breakthrough innovation (Griffiths et al., 2008; Tenenbaum et al., 2011). (3) There is a lack of actionable improvement pathways and original theoretical frameworks that specify how generative AI systems could be enhanced to better support breakthrough innovation (Zhang, 2016; Fukumura and Ito, 2025). (4) The relationship between predictive objectives and breakthrough innovation requirements has not been adequately theorized (Zhang, 2010; Chen et al., 2026). (5) The role of social value commitment and meta-cognitive self-negation in breakthrough innovation, and their absence in current AI systems, requires systematic investigation (Boden, 1998).

3 Conceptual Framework: Operationalization of Breakthrough Innovative Thinking and Generative AI Capability Layering

3.1 Five-Dimensional Operational Definition

We define breakthrough innovative thinking through five dimensions (Appendix). Questioning default assumptions involves the capacity to identify and challenge implicit presuppositions that constrain problem formulation and solution search (Gero, 2000). This is related to that described as dismissing well-established ways of thinking in favor of creating new ones (Jung and Vartanian, 2018). Redefining problems involves reformulating the problem space itself, not merely searching within a given problem space (Boden, 2004; Gero, 2000). First-principles thinking, which breaks down complex problems into basic assumptions and reassembles them from the bottom up, exemplifies this dimension (Aristotle, trans. 1984; Tan and Xiao, 2025). Cross-domain connection involves identifying structural similarities and transferring knowledge across domains that are superficially dissimilar (Casakin, 2010; Gero, 2000). Analogies and metaphors are key mechanisms for this dimension, enabling the recognition of correspondence between items from different domains and the interpretation of one through the qualities of another (Casakin, 2010). Systemic

reconstruction involves reconfiguring multiple components and their relationships into a coherent new whole, going beyond local modifications (Abernathy and Clark, 1985; Henderson and Clark, 1990). Rapid experimental validation involves the capacity to translate ideas into testable interventions and to learn from the results of those interventions in iterative cycles (Pearl, 2009; Zhang, 2016).

These five dimensions are not independent but interact in complex ways (Gero, 2000; Boden, 2004). Cross-domain connection may suggest a problem redefinition (Casakin, 2010). Questioning assumptions may enable systemic reconstruction (Jung and Vartanian, 2018). Experimental validation may reveal that a problem was incorrectly formulated, triggering further redefinition (Pearl, 2009). The breakthrough innovation process involves iterative cycling through these dimensions rather than linear progression (Zhang, 2016).

3.2 Cognitive Architecture of Generative AI

Generative AI systems can be characterized through a four-stage cognitive architecture: statistical compression, conditional sampling, prompt control, and feedback fine-tuning (LeCun et al., 2015; Brown et al., 2020). In the statistical compression stage, the model learns a compressed representation of the training distribution, capturing statistical regularities at multiple levels of abstraction (LeCun et al., 2015). In the conditional sampling stage, the model generates outputs by sampling from conditional distributions given the input context or prompt (Goodfellow et al., 2016). In the prompt control stage, the system's behavior can be steered through natural language instructions, examples, or other forms of guidance (Brown et al., 2020). In the feedback fine-tuning stage, the model's parameters are adjusted based on human feedback or task-specific objectives (Ouyang et al., 2022).

This architecture produces characteristic strengths and weaknesses. The system is strong in interpolation, association, and combination (Boden, 2004; LeCun et al., 2015). It can generate outputs that blend elements from different training examples in novel ways (Goodfellow et al., 2016). It is weak in intervention, validation, and value commitment (Pearl, 2009). It cannot perform causal interventions on the world, cannot validate its outputs through physical experiment, and cannot commit to values or goals that transcend its training objective (Pearl, 2009; Zhang, 2021a-c).

3.3 Analysis of the Statistical Only and Bayesian Only Propositions

The proposition that generative AI is based solely on statistical learning requires careful examination. It is true that the training objectives of most generative models are statistical in nature, typically involving maximum likelihood estimation or approximations thereof (Goodfellow et al., 2016; LeCun et al., 2015). The models learn to predict masked tokens, the next token in a sequence, or the noise added to an image (Devlin et al., 2019; Brown et al., 2020; Ho et al., 2020). This is undeniably statistical learning (Hastie et al., 2009).

However, the proposition that this is all there is to generative AI is too strong (Brown et al., 2020; Wei et al., 2022). The in-context learning capabilities of large language models demonstrate that they can perform computations at inference time that are not simply retrievals of memorized patterns (Brown et al., 2020). The representations learned by these models capture semantic and syntactic structures that go beyond surface-level statistics (Tenney et al., 2019; Rogers et al., 2020). The capacity for compositional generalization, while limited, exceeds what would be expected from pure pattern matching (Lake and Baroni, 2018; Hupkes et al., 2020).

The proposition that generative AI is based solely on Bayesian inference is similarly problematic (Chen et al., 2026). While Bayesian interpretations of learning and inference are available, and while some aspects of model behavior can be described in Bayesian terms, the actual computational mechanisms differ from Bayesian inference in important respects (Griffiths et al., 2008; Zhang, 2021a-c; Chen et al., 2026). Studies of whether large language models are consistently Bayesian find that they often possess biased priors and fail to adhere to Bayesian assumptions when encountering different evidence types (Kim et al., 2025; Chen et al.,

2026). The likelihoods used by these models do not always provide a good model of the world, and forcing a Bayesian update may in fact be harmful to task performance (Chen et al., 2026). Implicit Bayesian inference is an insufficient explanation of language model behavior in compositional tasks, and while the optimal strategy in-distribution is Bayesian, Bayesian inference is generally suboptimal on out-of-distribution prompts due to model misspecification (Ujváry et al., 2025).

More fundamentally, the question of whether statistical or Bayesian foundations are the source of limitations misidentifies the problem (Zhang, 2016). Human cognition also contains substantial statistical and Bayesian components (Griffiths et al., 2008; Tenenbaum et al., 2011). The capacity for breakthrough innovation in humans does not arise from the absence of statistical learning but from the presence of additional mechanisms that operate in conjunction with statistical learning (Zhang, 2016; Boden, 2004).

3.4 Four-Layer Capability Model

This study proposes a four-layer model of the capabilities required for breakthrough innovative thinking (Appendix; Fig. 1). The statistical layer involves pattern recognition and generation, the capacity to identify regularities in data and produce novel instances that conform to those regularities (Zhang, 2010; LeCun et al., 2015). The counterfactual layer involves hypothesis generation and problem redefinition, the capacity to entertain alternative possibilities and to reformulate problems in ways that open new solution spaces (Pearl, 2009; Zhang, 2021a-c). The causal-experimental layer involves intervention, validation, and iteration, the capacity to act on the world and learn from the consequences of those actions (Pearl, 2009; Zhang, 2021b). The social-value layer involves counter-consensus commitment, risk-taking, and paradigm commitment, the capacity to pursue ideas that are not validated by current consensus and to persist in the face of failure and opposition (Boden, 1998; Schumpeter, 1976).

Generative AI performs strongly at the statistical layer, moderately at the counterfactual layer, weakly at the causal-experimental layer, and minimally at the social-value layer (Boden, 1998; Chen et al., 2026; Pearl, 2009). This layered assessment provides a more nuanced picture than simple binary judgments about whether AI can or cannot be creative (Boden, 2004; Runco, 2024).

3.5 Evaluation Metrics

This study proposes seven metrics for evaluating breakthrough innovation capabilities. Novelty measures the degree to which an output differs from existing solutions (Lamb et al., 2018; Franceschelli and Musolesi, 2024). Surprise measures the degree to which an output violates expectations (Lamb et al., 2018). Counter-consensus degree measures the degree to which an output challenges prevailing views (Boden, 2004). Problem redefinition degree measures the extent to which an output reformulates the problem space (Gero, 2000). Verifiability measures the extent to which an output can be tested through experiment or observation (Pearl, 2009). Impact measures the extent to which an output influences subsequent work (Schumpeter, 1976). Reproducibility measures the extent to which the process generating the output can be replicated (Zhang, 2016).

4 Can Generative AI Produce Breakthrough Innovative Thinking? Capabilities and Boundaries

4.1 Performance and Limitations in Questioning Default Assumptions

Generative AI can generate critical text and identify assumptions embedded in prompts when explicitly asked to do so (Brown et al., 2020; Ouyang et al., 2022). It can list presuppositions, generate counterarguments, and explore alternative framings (Wei et al., 2022). However, this capability is fundamentally reactive rather than proactive (Boden, 1998). The system questions assumptions when prompted to question them, but it does not spontaneously identify and challenge the assumptions embedded in its own training data, objectives, or architecture (Boden, 1998; Zhang, 2027). It lacks what might be called meta-cognitive self-negation, the

capacity to turn its critical faculties on itself (Zhang, 2026a, 2027a; Boden, 1998).

The statistical compression mechanism that underlies generative AI necessarily encodes the assumptions present in the training data (LeCun et al., 2015). These assumptions are not represented explicitly but are distributed across the model's parameters (LeCun et al., 2015; Rogers et al., 2020). The model cannot easily identify or challenge them because it has no separate representation of what its assumptions are (Boden, 1998). This is in contrast to human cognition, where explicit reasoning about assumptions is possible through meta-cognitive processes (Flavell, 1979).

4.2 Performance and Limitations in Redefining Problems

The capacity for problem redefinition is more limited (Gero, 2000; Boden, 1998). Studies specifically examining the use of large language models for problem reframing find that using these models does not help improve the quality of problem frames (Ding et al., 2024). In fact, it increases the competence gap between experienced and inexperienced designers, with inexperienced users perceiving lower agency when working with these models (Ding et al., 2024). The probabilistic and autoregressive nature of large language models has been identified as one of the reasons preventing them from reaching transformational creativity (Boden, 1998; Franceschelli and Musolesi, 2024). Transformational creativity requires the ability to modify the conceptual space itself, to change the rules that define what counts as a valid solution (Boden, 2004). This is fundamentally different from searching within a fixed conceptual space, which is what probabilistic generation accomplishes (Boden, 2004; Zhang, 2010).

The OMEGA benchmark, designed to evaluate whether large language models can reason outside the box in mathematics, distinguishes among exploratory, compositional, and transformative generalization (Sun et al., 2025). The findings indicate that while these models can perform exploratory generalization by applying known strategies to new problems, compositional generalization remains limited, and transformative reasoning shows little to no improvement (Sun et al., 2025).

4.3 Advantages in Cross-Domain Connection

Cross-domain connection is the dimension where generative AI demonstrates the greatest relative strength (Ding et al., 2024; Kim et al., 2025). The high-dimensional representations learned from massive and diverse corpora enable the system to identify structural similarities across domains that might not be apparent to human experts with narrower training (LeCun et al., 2015; Rogers et al., 2020). The capacity for analogical reasoning, while imperfect, allows the system to generate analogies that can support creative problem-solving (Ding et al., 2024; Webb et al., 2023). Research on analogical reasoning with large language models indicates that cross-domain analogies generated by these systems can be helpful in creative tasks, particularly for problem reformulation (Ding et al., 2024). LLMs possess an emergent capability for analogical reasoning that can be harnessed for creative design ideation (Ding et al., 2024; Ding et al., 2023). Unlocking LLM creativity in science through analogical reasoning demonstrates that analogical reasoning helps mitigate mode collapse into low-diversity generations (Shen et al., 2026).

However, this strength has important limitations. The system's cross-domain connections are driven by statistical associations in the training data rather than by deep structural understanding (Boden, 1998; Zhang, 2010). The system may generate analogies that are superficially plausible but structurally invalid (Ding et al., 2024). The lack of a grounded causal world model means that the system cannot reliably assess whether a proposed analogy captures genuine causal mechanisms or merely superficial similarities (Pearl, 2009; Zhang, 2021a-c).

4.4 Potential and Deficiencies in Systemic Reconstruction

Generative AI can integrate knowledge from multiple domains and generate proposals that combine ideas in novel ways (Brown et al., 2020; Wei et al., 2022). It can produce syntheses of existing approaches and suggest

modifications to existing systems (Boden, 2004). In scientific domains, systems such as the Virtual Lab demonstrate the capacity for multi-agent collaboration on research problems, with agents actively validating hypotheses by writing and executing code (Swanson et al., 2025). The multi-agent system enables agents to work together on open-ended scientific research by generating and executing research plans (Swanson et al., 2025).

However, the capacity for systemic reconstruction is limited by the absence of a stable world model and long-term objectives (Boden, 1998; Zhang, 2016). The system can generate text that describes a systemic reconstruction, but it cannot implement or test that reconstruction in the world (Pearl, 2009). It lacks the persistent memory and goal-directed behavior that would be required to maintain a complex reconstruction project over time (Zhang, 2016). The system's behavior is fundamentally reactive to prompts rather than driven by intrinsic goals (Boden, 1998).

4.5 Instrumentalization Potential in Rapid Experimental Validation

The dimension of rapid experimental validation is where generative AI is most clearly instrumental rather than autonomous (Zhang et al., 2025). The system cannot perform experiments itself, but it can be connected to automated experimental platforms, simulation environments, and robotic laboratories (Zhang et al., 2025). Position papers on intelligent science laboratories argue for the integration of cognitive and embodied AI, proposing closed-loop frameworks that support iterative, autonomous experimentation and the possibility of serendipitous discovery (Zhang et al., 2025). The paradigm of embodied science reframes scientific discovery as a closed loop tightly coupling agentic reasoning with physical execution, proposing a unified Perception-Language-Action-Discovery framework (Zhuang et al., 2026). Human-AI co-embodied intelligence, uniting human researchers, agentic AI, and wearable hardware, has been demonstrated in microfabrication cleanrooms, enabling the co-development of fabrication protocols and achieving previously unattainable fabrication outcomes (Lin et al., 2026). These developments show that generative AI can serve as a component in experimental validation systems, but the validation function itself requires physical embodiment and real-world interaction that the AI system does not possess independently (Zhang et al., 2025; Zhuang et al., 2026).

4.6 Case Comparison: AI-Assisted Breakthrough Versus AI-Autonomous Breakthrough

A comparative analysis of cases illuminates the distinction between AI-assisted and AI-autonomous breakthrough. AlphaFold2 achieved a breakthrough in protein structure prediction that had resisted solution for decades (Jumper et al., 2021; DeepMind, 2024). This is a genuine scientific breakthrough. However, the breakthrough involved solving a well-defined prediction problem within an established paradigm (Boden, 1998). The problem itself, predicting protein structure from amino acid sequence, was clearly formulated, and the criteria for success were well-established (Jumper et al., 2021). AlphaFold2's achievement was a monumental technical accomplishment, but it did not involve redefining the problem or challenging the assumptions of structural biology (Boden, 1998).

In materials discovery, systems such as GNoME and MatterGen have generated novel stable structures and identified candidate materials (Merchant et al., 2023; Zeni et al., 2025). These systems search large spaces of possible materials using machine learning models trained on known materials (Merchant et al., 2023). The breakthroughs involve discovering materials with desirable properties, but the search process operates within a defined space of possible compositions and structures (Zeni et al., 2025). The systems do not redefine what a material is or challenge the fundamental assumptions of materials science (Boden, 1998).

In artistic domains, generative AI produces outputs that are evaluated as creative in terms of novelty and surprise, but analyses through Boden's framework find that these systems excel in manipulating language patterns and producing innovative poetic forms while lacking the contextual awareness, emotional depth, and intentionality of human creativity (Kim et al., 2025). Research on generative AI and creativity indicates that an

overuse of artificial creativity may lead to a reduction of idea diversity and to a decrease in the authenticity of creative outcomes due to ideation based on AI training data rather than personal experiences and thoughts (Grammenos and Lubart, 2025). Widespread reliance on GenAI may lead to a homogenization of creative output, a lack of true originality, and a resulting decline in human creative skills, a phenomenon termed the diminishment thesis (Yang et al., 2026).

The pattern across these cases is consistent. Generative AI excels at solving well-defined problems within established paradigms (Boden, 1998; Cropley, 2025). It can produce novel outputs that are valuable within those paradigms (Boden, 2004). It can even achieve results that are recognized as breakthroughs in the sense of solving long-standing problems (Jumper et al., 2021). But in all cases, the problem definition, the evaluation criteria, and the paradigm itself are supplied externally (Boden, 1998). The AI system does not redefine the problem, challenge the paradigm, or establish new evaluation criteria (Boden, 1998; Runco, 2024).

4.7 Summary

Generative AI can produce candidate elements of breakthrough innovation (Boden, 1998; Cropley, 2025). It can generate novel combinations, identify cross-domain connections, and propose solutions to well-defined problems (Boden, 2004; Ding et al., 2024). It can achieve breakthrough results when embedded in human-led research programs that supply the problem definition and evaluation criteria (Jumper et al., 2021; Merchant et al., 2023). However, it cannot autonomously produce complete breakthrough innovative thinking (Boden, 1998; Runco, 2024). The missing elements are the capacity for autonomous problem redefinition, the capacity for causal intervention and experimental validation, and the capacity for social value commitment and meta-cognitive self-negation (Boden, 1998; Pearl, 2009; Zhang, 2026).

5 Diagnosis of Fundamental Deficiencies: Are Statistics and Bayesian Inference the Original Sin?

5.1 The Nature of Statistical Learning

Statistical learning involves extracting patterns and regularities from data through optimization of an objective function (Hastie et al., 2009; Zhang, 2010; Zhang and Barrion, 2006; Zhang and Zhang, 2008; Zhang et al., 2007, 2008). In the case of generative AI, the objective is typically to maximize the likelihood of the observed data under the model, or to minimize a divergence between the model distribution and the data distribution (Goodfellow et al., 2016; Zhang, 2026b-h). This process compresses the training data into a set of parameters that capture statistical regularities at multiple levels of abstraction (LeCun et al., 2015). The strengths of this approach are substantial (LeCun et al., 2015). Statistical learning can capture high-dimensional dependencies that are difficult to model explicitly (Hastie et al., 2009; Zhang, 2007a-b, 2010). It can generalize to new instances that are similar to training examples (Goodfellow et al., 2016). It can generate novel combinations of learned features (Boden, 2004). The scaling laws observed in large language models demonstrate that increasing model size and training data leads to predictable improvements in performance across a wide range of tasks (Kaplan et al., 2020; Hoffmann et al., 2022).

The limitations of statistical learning for breakthrough innovation stem from what it optimizes (Boden, 1998; Zhang, 2010). The objective function rewards the production of outputs that are probable under the training distribution (LeCun et al., 2015). The model learns to generate outputs that resemble the training data (Goodfellow et al., 2016). While it can generate novel combinations, these combinations are constrained by the statistical structure of the training distribution (Boden, 2004). The model has no mechanism for recognizing when the training distribution itself is inadequate or when a fundamentally different approach is required (Boden, 1998).

5.2 Contributions and Boundaries of the Bayesian Perspective

Bayesian inference provides a framework for updating beliefs in light of evidence (Griffiths et al., 2008). It

involves specifying prior probabilities over hypotheses, computing likelihoods of observed data under each hypothesis, and deriving posterior probabilities through Bayes' theorem (Tenenbaum et al., 2011; Zhang, 2010, 2012, 2016, 2018, 2022). This framework has been highly influential in cognitive science, where it has been used to model human learning and reasoning (Tenenbaum et al., 2011).

The Bayesian perspective contributes to understanding generative AI by providing a normative framework for belief updating and uncertainty quantification (Griffiths et al., 2008). It clarifies the conditions under which learning from data is optimal (Tenenbaum et al., 2011). It highlights the importance of priors, which in the context of generative AI correspond to the inductive biases encoded in the model architecture and training objective (Chen et al., 2026).

However, the Bayesian perspective also has boundaries (Chen et al., 2026). Bayesian inference assumes that the hypothesis space is well-defined and that the likelihood function correctly specifies the probability of observations under each hypothesis (Griffiths et al., 2008; Zhang, 2022). In the context of breakthrough innovation, the relevant hypothesis space may not be known in advance (Boden, 2004). The creative act often involves generating new hypotheses that were not part of the original space (Boden, 1998). Bayesian inference provides no mechanism for this kind of hypothesis space expansion (Tenenbaum et al., 2011).

Studies of whether large language models are consistently Bayesian find that they often are not (Chen et al., 2026; Kim et al., 2025). They possess biased priors that cause initial divergence in zero-shot settings, and they fail to adhere to Bayesian assumptions when encountering different evidence types (Kim et al., 2025). The likelihoods used by these models do not always provide a good model of the world (Chen et al., 2026). These findings suggest that characterizing generative AI as Bayesian, even approximately, obscures important differences between its actual computational mechanisms and Bayesian inference (Chen et al., 2026).

More fundamentally, the question of whether Bayesian inference is a deficiency for breakthrough innovation is misleading. Bayesian inference is a general framework for reasoning under uncertainty (Griffiths et al., 2008). Its limitations for breakthrough innovation stem not from the framework itself but from what is included in the hypothesis space and how the likelihood function is defined (Tenenbaum et al., 2011). If the hypothesis space includes hypotheses that represent radical departures from existing paradigms, and if the likelihood function correctly evaluates evidence for those hypotheses, Bayesian inference could in principle support breakthrough innovation (Boden, 2004). The problem with current generative AI is not that it is Bayesian but that its effective hypothesis space is constrained by the training distribution (Boden, 1998; Chen et al., 2026).

5.3 Fundamental Deficiency I: Objective Function Mismatch

The primary fundamental deficiency is the mismatch between the predictive objective of generative AI and the requirements of breakthrough innovation (Boden, 1998). The training objective of predicting the next token, the masked token, or the noise component optimizes for producing outputs that are probable given the context (Devlin et al., 2019; Brown et al., 2020; Ho et al., 2020). This is a predictive objective (LeCun et al., 2015). It rewards the generation of outputs that conform to the statistical structure of the training data (Zhang and Barrion, 2006; Zhang, 2007a-b; Zhang et al., 2008; Goodfellow et al., 2016).

Breakthrough innovation, by contrast, requires predicting not the next token but the next question (Boden, 1998). It requires identifying which assumptions to challenge, which problems to redefine, which domains to connect (Gero, 2000). These are not prediction tasks in the statistical sense (Boden, 2004). They are meta-level tasks that involve reasoning about the problem space itself rather than generating outputs within a fixed problem space (Boden, 2004).

The objective function also optimizes for similarity to training data rather than for breakthrough (LeCun et al., 2015). An output that is highly probable under the training distribution is by definition similar to what has

been seen before (Zhang, 2007a-b; Zhang et al., 2008; Goodfellow et al., 2016). An output that constitutes a breakthrough is by definition dissimilar from what has been seen before (Boden, 2004). The training objective systematically rewards outputs that are the opposite of breakthroughs (Boden, 1998). This is not a flaw in the implementation of the objective but a fundamental misalignment between the objective and the goal (Zhang, 2010).

Recent work on prediction error minimization as a common computational principle for curiosity and creativity suggests that minimizing prediction errors may underlie both curiosity and creative insight (Becker and Cabeza, 2024). Curiosity is tied to the anticipation of minimizing prediction errors through future novel information, while creative AHA moments are connected to the actual minimization of prediction errors through current novel information (Becker and Cabeza, 2024). This framework suggests that the relationship between prediction and creativity is more nuanced than simple opposition (Becker and Cabeza, 2024). However, the predictive objective in generative AI is retrospective rather than prospective (LeCun et al., 2015). It minimizes prediction error on existing data rather than seeking out situations that will generate informative prediction errors (Boden, 1998; Becker and Cabeza, 2024).

5.4 Fundamental Deficiency II: Counterfactual and Causal Gap

The second fundamental deficiency is the gap in counterfactual and causal reasoning (Pearl, 2009; Zhang, 2021a-c). Text-only large language models achieve remarkable performance on linguistic and reasoning tasks but remain limited on intervention-sensitive and individual-level counterfactual reasoning (Pearl, 2009; Zhang, 2021c). The three levels of causal cognition, namely association, intervention, and counterfactual reasoning, are formally irreducible to one another, and current AI systems lack the structured causal models capable of supporting individual-level counterfactual invariance (Pearl, 2009; Bareinboim et al., 2022). Pearl's ladder of causality organizes questions into three ranks (Pearl, 2009). The lowest rank is the rank of association, which contains all questions that can be answered from observations (Pearl, 2009). The second rank is the rank of intervention, which contains questions that cannot be solved from the joint distribution alone and concern how the system will respond if the distribution of one of the variables changes (Pearl, 2009). The final and third rung is the rank of counterfactuals, which are data points that did not happen and involve questions of the form what would have happened if A would have happened instead of B (Pearl, 2009).

The counterfactual gap is particularly significant for breakthrough innovation (Boden, 1998; Pearl, 2009). Breakthrough innovation often involves asking what if we changed this default assumption or what would happen if we intervened in this system (Gero, 2000). These are counterfactual questions that require a causal model of the world (Pearl, 2009). The system must be able to represent the consequences of interventions that have not been observed and may not be possible in the current state of the world (Pearl, 2009; Zhang, 2021a-c).

Generative AI systems lack this capacity because they do not have explicit causal models (Pearl, 2009; Zhang, 2021a-c). They learn correlations from data but do not learn the causal mechanisms that generate those correlations (Zhang, 2021a-c). The distinction between correlation and causation is fundamental (Pearl, 2009). A system that learns only correlations cannot reliably predict the effects of interventions because interventions change the causal structure in ways that may not be reflected in the observational distribution (Pearl, 2009; Zhang, 2021a-c).

Recent works on causal world models represents an attempt to address this gap (Zhang, 2021a-c; Nam et al., 2026). C-JEPA induces latent interventions with counterfactual-like effects and prevents shortcut solutions, making interaction reasoning essential (Nam et al., 2026). The approach leads to an absolute improvement of about 20% in counterfactual reasoning compared to the same architecture without object-level masking (Nam et al., 2026). However, these approaches remain limited in scope and have not been integrated into general-purpose generative AI systems (Pearl, 2009; Zhang, 2021a-c; Nam et al., 2026).

5.5 Fundamental Deficiency III: Experimental Closure and Embodiment Gap

The third fundamental deficiency is the absence of experimental closure and embodiment (Pearl, 2009). Breakthrough innovation depends on trial and error, on testing ideas against reality and learning from the results (Schumpeter, 1976; Pearl, 2009). This requires the capacity to act on the world, observe the consequences, and update beliefs and strategies accordingly (Zhang, 2016).

Generative AI systems exist primarily in virtual environments (Boden, 1998; Ujváry et al., 2025). They generate text, images, or code, but they do not perform physical experiments or interact with the world in ways that provide feedback about the validity of their ideas (Ujváry et al., 2025). While AI systems can be connected to automated laboratories and robotic platforms, the integration remains limited and the AI does not autonomously control the experimental process (Zhuang et al., 2026).

Position papers on intelligent science laboratories argue for the integration of cognitive and embodied AI to create closed-loop systems that support iterative, autonomous experimentation. The paradigm of embodied science reframes scientific discovery as a closed loop tightly coupling agentic reasoning with physical execution (Zhuang et al., 2026). These proposals recognize the importance of embodiment for scientific discovery, but they also highlight how far current systems are from achieving genuine experimental closure (Zhuang et al., 2026).

5.6 Fundamental Deficiency IV: Social Cognition and Value Gap

The fourth fundamental deficiency is the absence of social cognition and value commitment (Boden, 1998; Schumpeter, 1976). Breakthrough innovation is not merely a cognitive process but a social one (Schumpeter, 1976). Radical ideas are typically rejected by the mainstream before they are accepted (Boden, 1998). The innovator must be willing to pursue ideas that are not validated by current consensus, to persist in the face of failure and opposition, and to commit to a paradigm even when it is not yet established (Boden, 1998; Kuhn, 1962).

Generative AI systems have no capacity for this kind of value commitment (Boden, 1998; Runco, 2024). They optimize for outputs that are probable under the training distribution, which means they are biased toward consensus views (LeCun et al., 2015; Chen et al., 2026). They cannot commit to a paradigm that is not yet established because they have no mechanism for representing or pursuing goals that transcend their training objective (Boden, 1998). They cannot take risks because they have no concept of risk or of the potential rewards that might justify it (Schumpeter, 1976).

The social dimension of breakthrough innovation also includes the capacity to communicate ideas effectively, to build coalitions of support, and to navigate the institutional structures that govern scientific and technological fields (Schumpeter, 1976; Christensen, 1997). These are social skills that require understanding of human psychology, social dynamics, and institutional contexts (Boden, 1998). Generative AI can generate text that describes these skills but cannot perform them (Runco, 2024).

5.7 Fundamental Deficiency V: Meta-Cognition and Self-Negation Gap

The fifth fundamental deficiency is the absence of meta-cognition and self-negation (Zhang, 2026a; Boden, 1998). Breakthrough innovation often requires questioning one's own assumptions, recognizing the limitations of one's current framework, and being willing to abandon ideas that are not working (Boden, 2004). This requires meta-cognitive capacities that allow the system to monitor its own cognitive processes and to revise them when necessary (Flavell, 1979; Zhang, 2026a).

Generative AI systems have limited meta-cognitive capabilities (Boden, 1998; Zhang, 2027a). They can be prompted to reflect on their outputs and to generate critiques of their own work (Ouyang et al., 2022). But this reflection is not integrated into their operation in a way that would allow them to spontaneously revise their approach (Boden, 1998). The system does not have a persistent representation of its own assumptions, goals,

or limitations that it can monitor and revise (Zhang, 2026a).

The absence of self-negation is particularly significant for breakthrough innovation (Boden, 1998). The creative process often involves abandoning promising ideas that turn out to be dead ends and pursuing alternative directions (Boden, 2004). This requires the capacity to recognize when an approach is failing and to make a decisive change (Boden, 1998). Generative AI systems, by contrast, tend to persist with the approach that is most probable given the prompt and context (LeCun et al., 2015). They lack the capacity for the kind of decisive paradigm shift that characterizes breakthrough innovation (Boden, 2004; Kuhn, 1962).

5.8 Core Proposition

The fundamental deficiency of generative AI for breakthrough innovation is not statistical or Bayesian learning per se. Human cognition also employs statistical and Bayesian mechanisms, and these mechanisms contribute to human creativity (Griffiths et al., 2008; Tenenbaum et al., 2011). The fundamental deficiency is a structural mismatch involving three components: predictive objectives that reward conformity to training distributions rather than the generation of radical novelty (Boden, 1998; LeCun et al., 2015); the absence of causal models and experimental closure that would allow the system to test ideas against reality and learn from the results (Pearl, 2009); and the absence of social value commitment and meta-cognitive self-negation that would allow the system to pursue ideas that are not validated by current consensus and to revise its own assumptions when they prove inadequate (Boden, 1998; Zhang, 2026).

We term this the goal-hypothesis mismatch (Boden, 1998). The system's objective is to predict, but breakthrough innovation requires something more than prediction (Boden, 2004). It requires the generation of hypotheses that are not predictable from existing data, the testing of those hypotheses through intervention, and the commitment to pursue them even in the face of opposition (Pearl, 2009; Schumpeter, 1976). The mismatch is not between statistics and creativity but between prediction and innovation (Boden, 1998).

6 Original Theoretical Construction: The Three-Engine Four-Layer Coupling Theory

6.1 Theoretical Name and Core Propositions

This study proposes the Three-Engine Four-Layer Coupling Theory of breakthrough innovative thinking (Appendix; Fig. 1). The theory posits that breakthrough innovation emerges from the coupled operation of three engines: a statistical generation engine, a counterfactual reconstruction engine, and an experimental validation engine (Pearl, 2009). These engines operate across four layers: the statistical layer, the counterfactual layer, the causal-experimental layer, and the social-value layer (Zhang, 2021a-c). Breakthrough innovation occurs when the three engines are coupled in a way that allows them to interact productively, with the output of one engine serving as input to the others in iterative cycles (Boden, 2004).

The core proposition is that breakthrough innovation cannot be reduced to any single engine or layer (Boden, 2004). Statistical generation alone produces novel combinations but not radical novelty (Boden, 2004). Counterfactual reconstruction alone produces alternative possibilities but not validated innovations (Pearl, 2009). Experimental validation alone tests hypotheses but does not generate them (Pearl, 2009). The coupling of the three engines is what produces breakthrough innovation.

6.2 The Three Engines

The statistical generation engine performs pattern recognition, association, and combinatorial generation (Zhang, 2010; Zhang, 2007a-b). It learns the statistical structure of a domain and generates novel instances that conform to that structure (LeCun et al., 2015). In generative AI, this engine is implemented through the model's learned parameters and sampling mechanisms (Goodfellow et al., 2016). The engine is powerful for interpolation, extrapolation, and combination within the space defined by the training distribution (Boden, 2004; LeCun et al., 2015).

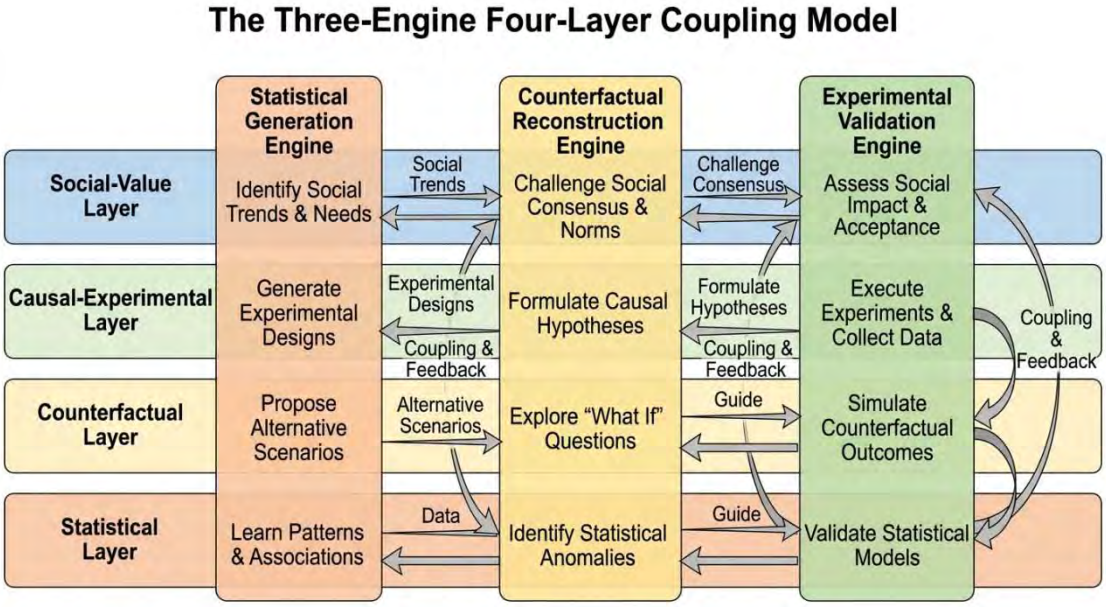

Fig. 1 The Three-Engine Four-Layer Coupling model.

The counterfactual reconstruction engine performs questioning, redefinition, and counter-consensus reasoning (Boden, 1998; Pearl, 2009). It identifies assumptions, generates alternative framings, and explores possibilities that are not represented in the training distribution (Boden, 1998). This engine is responsible for problem redefinition and for the generation of hypotheses that challenge existing paradigms (Gero, 2000; Boden, 2004). In current AI systems, this engine is weakly implemented, primarily through prompt-based guidance rather than through the system's intrinsic capabilities (Boden, 1998; Chen et al., 2026).

The experimental validation engine performs intervention, observation, and iteration (Pearl, 2009; Zhang, 2021a-c). It tests hypotheses against reality, observes the consequences of interventions, and updates beliefs and strategies based on the results (Pearl, 2009). This engine is responsible for the empirical grounding of breakthrough innovation, ensuring that ideas are tested against the world rather than merely generated (Pearl, 2009). In current AI systems, this engine is largely absent, although it can be partially implemented through integration with automated experimental platforms (Zhuang et al., 2026).

6.3 The Four Layers

The statistical layer involves pattern recognition and generation (Zhang, 2010; Zhang, 2007a-b). It is the layer at which generative AI operates most effectively (LeCun et al., 2015). The system learns statistical regularities from data and generates outputs that conform to those regularities (Zhang, 2007a-b ; Goodfellow et al., 2016). The layer is necessary but not sufficient for breakthrough innovation (Boden, 2004). Without the capacity to recognize patterns, the system cannot generate novel combinations (Boden, 2004). But the statistical layer alone produces outputs that are constrained by the training distribution and cannot generate the radical novelty required for breakthrough (Boden, 1998).

The counterfactual layer involves hypothesis generation and problem redefinition (Pearl, 2009; Boden, 1998). It is the layer at which the system entertains alternative possibilities, challenges assumptions, and reformulates problems (Gero, 2000). This layer is partially present in generative AI through the capacity for in-context learning and compositional generalization, but it is limited by the absence of explicit causal models and by the predictive objective that rewards conformity to training data (Boden, 1998; Chen et al., 2026).

The causal-experimental layer involves intervention, validation, and iteration (Pearl, 2009; Zhang, 2021a-c). It is the layer at which the system acts on the world and learns from the consequences (Pearl, 2009). This layer is largely absent in current generative AI. The system can generate text that describes experiments, but it cannot perform them. The integration of generative AI with automated experimental platforms represents an early attempt to address this gap, but the integration remains limited (Zhuang et al., 2026).

The social-value layer involves counter-consensus commitment, risk-taking, and paradigm commitment (Boden, 1998; Schumpeter, 1976). It is the layer at which the system commits to ideas that are not validated by current consensus and persists in the face of opposition (Boden, 1998). This layer is absent in current generative AI (Runco, 2024). The system has no mechanism for value commitment, risk assessment, or paradigm persistence (Boden, 1998). It optimizes for outputs that are probable under the training distribution, which means it is biased toward consensus views (LeCun et al., 2015; Chen et al., 2026).

6.4 The Breakthrough Innovation Occurrence Function

This study proposes a formal representation of the conditions under which breakthrough innovation occurs (Appendix). Let B denote the occurrence of breakthrough innovation, then we have:

$$B = f(S, C, E, V, M)$$

where S represents the statistical generation capability, C represents the counterfactual reconstruction capability, E represents the experimental validation capability, V represents the value commitment and social cognition capability, and M represents the meta-cognitive monitoring and self-negation capability (Zhang, 2010; Pearl, 2009; Boden, 1998).

Breakthrough innovation requires all five components to be present at sufficient levels (Zhang, 2016). If S is low, the system cannot generate novel combinations (Boden, 2004). If C is low, the system cannot question assumptions or redefine problems (Boden, 1998). If E is low, the system cannot validate its ideas against reality (Pearl, 2009). If V is low, the system cannot commit to ideas that are not validated by consensus (Schumpeter, 1976). If M is low, the system cannot revise its own assumptions when they prove inadequate (Zhang, 2026a).

The function f is not simply additive. The components interact in complex ways (Boden, 2004). S and C interact to generate novel hypotheses that are grounded in domain knowledge (Boden, 2004). C and E interact to refine hypotheses through empirical testing (Pearl, 2009). E and V interact to sustain the pursuit of ideas through failure and opposition (Schumpeter, 1976). V and M interact to enable the revision of value commitments and paradigms (Boden, 1998).

This study proposes the following functional form for f:

$$B = S^{\alpha} \cdot C^{\beta} \cdot E^{\gamma} \cdot V^{\delta} \cdot M^{\varepsilon}$$

where α, β, γ, δ, and ε are parameters representing the relative importance of each component and the exponents are constrained to be positive. This multiplicative form captures the intuition that all components are necessary and that the absence of any one component reduces B to zero or near zero. The parameters would need to be estimated empirically for different domains and contexts.

6.5 The Position of Generative AI in the Theory

Generative AI occupies a specific position in this theoretical framework (Boden, 1998). It is a strong statistical generation engine (LeCun et al., 2015; Zhang, 2026b-h). It is a weak counterfactual reconstruction engine (Boden, 1998). It lacks the experimental validation engine (Zhang et al., 2025). It lacks the social-value layer

(Schumpeter, 1976). And it has limited meta-cognitive capabilities (Zhang, 2026).

Therefore, generative AI functions as a breakthrough innovation probability amplifier rather than as an autonomous breakthrough agent (Boden, 1998). It can increase the probability of breakthrough innovation by accelerating the generation of candidate ideas, by identifying cross-domain connections, and by assisting in the exploration of solution spaces (Boden, 2004). But it cannot autonomously perform the complete cycle of breakthrough innovation because it lacks the counterfactual reconstruction, experimental validation, and social-value components (Boden, 1998; Pearl, 2009; Schumpeter, 1976).

6.6 Dialogue with Existing Theories

The Three-Engine Four-Layer Coupling Theory engages with several existing theoretical frameworks (Zhang, 2016). Computational creativity research provides the foundational insight that creativity can be understood in terms of conceptual spaces and the operations that transform them (Boden, 2004). The theory extends this by specifying the engines and layers involved in breakthrough innovation and by identifying the specific gaps in current AI systems (Boden, 1998).

Extended cognition theory posits that cognitive processes extend beyond the individual brain to include external tools and resources (Clark and Chalmers, 1998). The theory is consistent with this view in recognizing that breakthrough innovation involves the coupling of internal cognitive processes with external experimental and social systems (Clark and Chalmers, 1998). However, the theory also emphasizes the importance of internal capabilities, particularly the counterfactual and meta-cognitive capabilities that cannot be fully externalized (Boden, 1998; Zhang, 2026a).

Human-AI hybrid intelligence research has demonstrated the potential of combining human and AI capabilities for creative tasks (Fukumura and Ito, 2025; Ito et al., 2025). The theory provides a framework for understanding why such hybrids can be effective: the human supplies the counterfactual reconstruction, experimental validation, and social-value capabilities that the AI lacks, while the AI supplies the statistical generation capability that the human lacks at scale (Fukumura and Ito, 2025; Boden, 1998).

Actor-network theory emphasizes the role of heterogeneous networks of human and non-human actors in innovation (Latour, 2005). The theory is consistent with this view in recognizing that breakthrough innovation involves the coupling of multiple engines and layers (Latour, 2005). However, the theory also specifies the functional requirements of each component and the conditions under which their coupling produces breakthrough innovation (Zhang, 2027b).

6.7 Testable Hypotheses

The theory generates several testable hypotheses. H1: Generative AI cross-domain connection performance will exceed the average human performance, but problem redefinition performance will be weaker than human expert performance (Boden, 1998). H2: Adding causal world models and automated experimental integration to generative AI systems will improve breakthrough innovation metrics (Pearl, 2009; Zhang, 2021a-c). H3: Multi-agent counter-consensus debate among AI agents will improve problem redefinition performance (Fukumura and Ito, 2025). H4: Human-AI hybrid systems will produce higher breakthrough innovation outputs than either AI alone or human alone (Fukumura and Ito, 2025; Boden, 2004).

7 Improvement Pathways: From Generative AI to Breakthrough Innovation Cognitive Systems

7.1 Objective Function Improvement

The first improvement pathway involves redesigning the objective function of generative AI to align better with the requirements of breakthrough innovation (Boden, 1998; Zhang, 2010). Current objectives maximize the likelihood of observed data or minimize prediction error (Goodfellow et al., 2016; LeCun et al., 2015). An alternative objective would incorporate novelty, surprise, and counter-consensus as explicit rewards (Boden,

2004; Lamb et al., 2018).

The multi-objective optimization would balance novelty, feasibility, and impact (Boden, 2004; Zhang, 2010). Novelty rewards outputs that differ from training data and from existing solutions (Lamb et al., 2018). Feasibility rewards outputs that can be implemented or tested (Pearl, 2009). Impact rewards outputs that have the potential to influence subsequent work (Schumpeter, 1976). The tradeoffs among these objectives would need to be specified based on the domain and context.

Prediction error minimization as a computational principle for curiosity and creativity suggests a possible direction (Becker and Cabeza, 2024). If creative insight is connected to the actual minimization of prediction errors through novel information, then an objective that seeks out situations where prediction error can be reduced through the generation of novel hypotheses might promote breakthrough innovation (Becker and Cabeza, 2024). This would be a prospective objective that actively seeks informative prediction errors rather than retrospectively fitting existing data (Becker and Cabeza, 2024).

7.2 Architecture Improvement

The second improvement pathway involves architectural modifications. Causal world models would provide the system with explicit representations of causal mechanisms that support intervention and counterfactual reasoning (Pearl, 2009; Zhang, 2021a-c; Nam et al., 2026). Neural-symbolic systems would combine the pattern recognition capabilities of neural networks with the explicit reasoning capabilities of symbolic systems (Zhang, 2007a-b, 2010; Garcez and Lamb, 2020). Knowledge graphs would provide structured representations of domain knowledge that can be used for reasoning and hypothesis generation (Hogan et al., 2021). Long-term memory would allow the system to maintain state across extended problem-solving episodes (Graves et al., 2016). Meta-cognitive monitoring would allow the system to track its own assumptions, goals, and limitations (Flavell, 1979; Zhang, 2026a).

Recent work on causal world models represents progress in this direction (Zhang, 2021a-c; Nam et al., 2026). C-JEPA induces latent interventions with counterfactual-like effects and prevents shortcut solutions (Nam et al., 2026). The approach demonstrates that object-level masking can induce a causal inductive bias via latent interventions (Nam et al., 2026). However, these approaches remain limited in scope and have not been integrated into general-purpose generative AI systems (Pearl, 2009; Nam et al., 2026).

7.3 Interaction Improvement

The third improvement pathway involves interaction mechanisms (Fukumura and Ito, 2025). Multi-agent debate allows multiple AI agents to critique and refine each other's ideas, potentially overcoming the tendency toward consensus that characterizes single-agent systems (Fukumura and Ito, 2025; Ito et al., 2025). Research on creative generation via multi-agent debate suggests that debate can improve quality but may suppress diversity, highlighting the need for careful design of debate protocols (Fukumura and Ito, 2025). Red-team critique involves explicitly assigning an agent to challenge the assumptions and conclusions of another agent (Fukumura and Ito, 2025). Counter-consensus prompts explicitly instruct the system to generate ideas that challenge prevailing views (Boden, 1998). Problem redefinition protocols provide structured procedures for reformulating problems (Gero, 2000).

7.4 Tool and Embodiment Improvement

The fourth improvement pathway involves integration with tools and embodied systems. Automated experimentation platforms allow hypotheses to be tested without human intervention. Robotic laboratories provide physical manipulation capabilities (Burger et al., 2020). Simulation environments allow rapid testing of ideas in virtual settings. Real-world closed-loop systems integrate all of these components into a unified system that can generate, test, and refine ideas autonomously (Zhuang et al., 2026).

The integration of cognitive and embodied AI in intelligent science laboratories represents progress in this

direction. Human-AI co-embodied intelligence unites human researchers, agentic AI, and wearable hardware to enable real-time collaboration in experimental settings (Lin et al., 2026). The embodied science paradigm reframes scientific discovery as a closed loop tightly coupling agentic reasoning with physical execution (Zhuang et al., 2026). These developments show the potential for closing the experimental gap, but they also highlight the challenges of integrating AI systems with physical experimental platforms (Zhuang et al., 2026).

7.5 Human-AI Hybrid Improvement

The fifth improvement pathway involves human-AI hybrid systems (Fukumura and Ito, 2025). Humans supply the capabilities that AI lacks: value commitment, risk-taking, problem redefinition, and meta-cognitive self-negation (Boden, 1998; Schumpeter, 1976). AI supplies the capabilities that humans lack at scale: rapid generation of candidate ideas, cross-domain connection, and systematic exploration of large solution spaces (Boden, 2004). The hybrid system combines these capabilities in a way that exceeds what either can achieve alone (Fukumura and Ito, 2025; Boden, 2004).

This approach recognizes that the goal is not to create AI systems that replace human creativity but to create systems that augment human creativity by compensating for human limitations while leveraging human strengths (Fukumura and Ito, 2025; Ito et al., 2025). The human remains in the loop for the critical decisions about what problems to pursue, what assumptions to challenge, and what risks to take (Boden, 1998). The AI accelerates the exploration of solution spaces and the generation of candidate ideas.

7.6 Institutional and Cultural Improvement

The sixth improvement pathway involves institutional and cultural changes (Schumpeter, 1976; Christensen, 1997). Tolerance of failure is essential for breakthrough innovation because most attempts at radical novelty fail (Boden, 1998). Open science practices allow ideas and results to be shared and built upon (Merton, 1973). Interdisciplinary teams bring together diverse perspectives and expertise (Christensen, 1997). Counter-consensus incentives reward ideas that challenge prevailing views rather than merely extending them (Boden, 1998; Schumpeter, 1976).

These institutional factors are not merely contextual but are constitutive of breakthrough innovation (Schumpeter, 1976). The social-value layer of the Three-Engine Four-Layer Coupling Theory recognizes that breakthrough innovation requires social conditions that support risk-taking and non-conformity (Boden, 1998). Creating these conditions is as important as developing the technical capabilities of AI systems (Schumpeter, 1976; Christensen, 1997).

7.7 Evaluation and Governance

The seventh improvement pathway involves developing better evaluation metrics and governance frameworks (Lamb et al., 2018). Current metrics for AI creativity focus on output characteristics such as novelty and surprise (Lamb et al., 2018; Franceschelli and Musolesi, 2024). A comprehensive evaluation of breakthrough innovation capability would need to assess the full cycle from problem definition to experimental validation to social impact (Boden, 2004; Pearl, 2009). This requires process-based evaluation in addition to output-based evaluation (Jordanous, 2012).

Governance frameworks need to address questions of attribution, accountability, and safety (Boden, 1998; Schumpeter, 1976). When AI systems contribute to breakthrough innovations, how should credit be assigned (Boden, 1998). When AI-generated ideas are harmful or dangerous, who is responsible (Runco, 2024). These questions become more pressing as AI systems become more capable and more integrated into innovation processes (Sedkaoui and Benaichouba, 2026). The International Association for Computational Creativity's mission to promote the scientific study of human and machine creativity provides a model for how the field can organize around these questions (IACC, 2024).

8 Research Methodology and Validation Design

8.1 Literature Bibliometrics and Systematic Review

The literature review employed a systematic approach, searching multiple databases for publications related to generative AI, creativity, innovation, and related topics (Booth et al., 2016). Search terms were combined using Boolean operators to ensure comprehensive coverage (Booth et al., 2016). The inclusion criteria prioritized recent publications while ensuring coverage of foundational works (Creswell and Creswell, 2022). The present review identified numerous references, with the majority published within the last five years (Ismayilzada et al., 2024; Chen et al., 2026).

8.2 Conceptual Analysis and Logical Reasoning

Conceptual analysis was employed to clarify key terms and distinguish between related concepts (Boden, 2004). The analysis distinguished between statistical and Bayesian foundations (Griffiths et al., 2008; Chen et al., 2026), between combinational and transformational creativity (Boden, 2004), and between AI-assisted and AI-autonomous breakthrough (Boden, 1998). Logical reasoning was used to derive implications from the theoretical framework and to identify testable hypotheses.

8.3 Case Study Design

Comparative case analysis examined AI achievements in protein structure prediction (Jumper et al., 2021), materials discovery (Merchant et al., 2023; Zeni et al., 2025), mathematical reasoning (Sun et al., 2025), and artistic creation (Kim et al., 2025). Cases were selected to represent different domains and different degrees of AI autonomy (Boden, 1998). The analysis distinguished between cases where AI solved well-defined problems within established paradigms and cases where AI contributed to problem redefinition or paradigm shift (Boden, 2004).

8.4 Thought Experiment and Counterfactual Analysis

Thought experiments were used to explore the conditions under which generative AI might produce breakthrough innovation (Boden, 1998). Counterfactual analysis considered what would happen if specific deficiencies were remedied (Pearl, 2009). These methods are particularly appropriate for questions that cannot be directly tested with current technology but that can be analyzed through logical reasoning about hypothetical scenarios (Pearl, 2009; Zhang, 2016).

8.5 Computational Modeling and Simulation

The Three-Engine Four-Layer Coupling Theory was formalized as a mathematical model (Appendix). The model specifies the conditions under which breakthrough innovation occurs and the relative contributions of different components. Simulation experiments could be designed to explore the behavior of the model under different parameter settings. These experiments would test whether the model's predictions align with observed patterns of breakthrough innovation.

8.6 Empirical Validation Plan

Three empirical studies are proposed to test the theory's hypotheses (Fukumura and Ito, 2025). First, a human-AI comparison experiment would assess cross-domain connection and problem redefinition performance. Human participants and AI systems would be given tasks requiring both capabilities, and their performance would be compared using the metrics defined in Section 3.5 (Lamb et al., 2018). Second, an AI-assisted innovation experiment would compare breakthrough innovation outputs from human-AI hybrid systems with outputs from humans alone and AI alone (Fukumura and Ito, 2025; Ito et al., 2025). Third, a longitudinal tracking study would follow breakthrough innovation projects over time to assess the role of different components in the innovation process (Schumpeter, 1976; Christensen, 1997).

8.7 Validity and Limitations

The validity of the analysis depends on the accuracy of the literature review, the rigor of the conceptual

analysis, and the appropriateness of the theoretical framework (Booth et al., 2016; Creswell and Creswell, 2022). Potential limitations include the difficulty of operationalizing breakthrough innovation (Boden, 2004), the challenge of comparing human and AI performance on creative tasks (Runco, 2024), and the rapid pace of change in AI capabilities (Ismayilzada et al., 2024). These limitations should be considered when interpreting the findings.

9 Discussion

9.1 Theoretical Contributions

This study makes several theoretical contributions. It redefines the creative boundaries of generative AI by distinguishing between statistical generation, counterfactual reconstruction, and experimental validation. It identifies the goal-hypothesis mismatch as the fundamental deficiency rather than statistical or Bayesian foundations. It proposes the Three-Engine Four-Layer Coupling Theory as a framework for understanding breakthrough innovation and the conditions under which it occurs. It specifies testable hypotheses and proposes empirical validation designs.

The theory contributes to the broader understanding of creativity by specifying the components that are necessary for breakthrough innovation and by identifying the conditions under which their coupling produces radical novelty. It provides a framework for evaluating AI systems not just in terms of output characteristics but in terms of the underlying capabilities that support breakthrough innovation (Lamb et al., 2018).

9.2 Practical Implications

The findings have implications for research, industry, education, and policy (Schumpeter, 1976; Christensen, 1997). In research, the framework can guide the development of AI systems that better support breakthrough innovation by identifying the capabilities that need to be enhanced. In industry, the findings suggest that AI is best deployed as a collaborative partner rather than as a replacement for human innovators (Fukumura and Ito, 2025). In education, the framework can inform curricula that develop both the statistical and counterfactual capabilities required for creative work (Boden, 2004). In policy, the findings suggest that governance frameworks should focus on ensuring that AI systems are used in ways that augment rather than undermine human creativity (Boden, 1998; Runco, 2024).

9.3 Response to Controversies

The question of whether AI will eventually develop breakthrough innovation capabilities remains open (Boden, 1998; Runco, 2024). Some argue that scaling alone will suffice, while others argue that fundamental architectural changes are required (LeCun et al., 2015; Boden, 1998). The theory proposed here suggests that while scaling can improve statistical generation capabilities, it cannot overcome the structural mismatches that define the goal-hypothesis gap (Zhang, 2016; Boden, 1998). Architectural changes that add causal world models, experimental closure, and value commitment mechanisms are necessary for genuine breakthrough innovation capability (Pearl, 2009; Zhang, 2021a-c).

The question of whether statistical learning is sufficient for creativity is more nuanced (Boden, 2004). Statistical learning is necessary for pattern recognition and combinatorial generation, but it is not sufficient for transformational creativity (Boden, 2004; Zhang, 2010). Transformational creativity requires the capacity to modify the conceptual space itself, which goes beyond what statistical learning within a fixed space can provide (Boden, 2004).

9.4 Ethical and Social Implications

The development of AI systems with enhanced breakthrough innovation capabilities raises ethical and social questions (Boden, 1998). If AI systems become more autonomous in generating innovative ideas, questions of attribution and accountability become more complex (Runco, 2024). If AI-generated innovations have negative

consequences, determining responsibility becomes difficult (Sedkaoui and Benaichouba, 2026). The theory proposed here suggests that these questions can be addressed through human-AI hybrid systems that keep humans in the loop for critical decisions (Fukumura and Ito, 2025; Ito et al., 2025).

There is also the question of whether AI systems should be granted credit for innovative achievements (Boden, 1998). The theory's distinction between AI-assisted and AI-autonomous breakthrough provides a framework for thinking about attribution (Boden, 1998). AI-assisted breakthroughs are those where the AI contributes to a process that is fundamentally human-led (Boden, 1998; Jumper et al., 2021). AI-autonomous breakthroughs would be those where the AI independently performs the complete cycle of breakthrough innovation (Boden, 1998). The theory suggests that the latter are not currently possible and may require capabilities that go beyond statistical learning (Boden, 1998; Zhang, 2027a).

9.5 Research Limitations and Future Directions

The main limitation of this analysis is its theoretical nature. The proposed hypotheses have not been empirically tested (Fukumura and Ito, 2025). Future research should focus on developing operational measures of the constructs in the theory and testing the predictions in controlled experiments (Fukumura and Ito, 2025). The development of causal world models, experimental closure mechanisms, and value commitment architectures should be pursued as research priorities (Pearl, 2009; Zhang, 2021a-c).

Another limitation is the rapid pace of change in AI capabilities (Ismayilzada et al., 2024). New architectures and training paradigms are emerging continuously, and some may address the deficiencies identified here (Chen et al., 2026). The theory should be understood as a framework for analyzing and predicting capability development rather than as a fixed assessment of current systems.

10 Conclusion

10.1 Core Conclusions

Generative AI has achieved remarkable capabilities in statistical generation, pattern recognition, and combinatorial creativity (LeCun et al., 2015; Boden, 2004). It can produce outputs that are novel, surprising, and valuable within established paradigms (Boden, 2004; Lamb et al., 2018). It can assist human innovators by accelerating the generation of candidate ideas and by identifying cross-domain connections. However, it cannot autonomously produce complete breakthrough innovative thinking (Boden, 1998; Runco, 2024).

The fundamental deficiency is not statistical or Bayesian learning per se (Zhang, 2010, 2016). Human cognition also employs these mechanisms, and they contribute to human creativity (Griffiths et al., 2008; Tenenbaum et al., 2011). The fundamental deficiency is the goal-hypothesis mismatch: the mismatch between predictive objectives that reward conformity to training distributions and the requirements of breakthrough innovation for radical novelty, experimental validation, and social value commitment (Boden, 1998; Zhang, 2016). The absence of causal models, experimental closure, and meta-cognitive self-negation further compounds this mismatch (Pearl, 2009; Ujváry et al, 2025; Zhang, 2026).

10.2 Summary of Original Contributions

The paper proposes the Five-Dimensional Operationalization of breakthrough innovative thinking: questioning default assumptions, redefining problems, cross-domain connection, systemic reconstruction, and rapid experimental validation (Gero, 2000; Boden, 2004). It diagnoses the fundamental deficiency as the goal-hypothesis mismatch rather than statistical or Bayesian foundations (Zhang, 2016; Boden, 1998). It constructs the Three-Engine Four-Layer Coupling Theory, which specifies the statistical generation, counterfactual reconstruction, and experimental validation engines operating across statistical, counterfactual, causal-experimental, and social-value layers. It proposes improvement pathways spanning objective functions, architecture, interaction, embodiment, human-AI hybrid systems, and institutions (Boden, 1998; Pearl, 2009;

Fukumura and Ito, 2025). It defines evaluation metrics and proposes empirical validation designs (Lamb et al., 2018; Fukumura and Ito, 2025).

10.3 Final Answer

Can generative AI produce breakthrough innovative thinking (Boden, 1998). The answer is nuanced (Boden, 1998; Runco, 2024). Generative AI can produce candidate elements of breakthrough innovation (Boden, 1998; Cropley, 2025). It can generate novel combinations, identify cross-domain connections, and propose solutions to well-defined problems (Boden, 2004). It can achieve breakthrough results when embedded in human-led research programs that supply the problem definition and evaluation criteria (Jumper et al., 2021; Merchant et al., 2023). However, it cannot autonomously produce complete breakthrough innovative thinking because it lacks the counterfactual reconstruction, experimental validation, and social-value capabilities that are essential for the complete cycle of breakthrough innovation (Boden, 1998; Pearl, 2009; Schumpeter, 1976).

The future path lies not in replacing human innovators with AI but in constructing human-AI-experiment-society coupled systems that combine the strengths of each component (Fukumura and Ito, 2025). Humans supply value commitment, risk-taking, problem redefinition, and meta-cognitive self-negation (Boden, 1998; Schumpeter, 1976). AI supplies statistical generation, cross-domain connection, and systematic exploration at scale (LeCun et al., 2015). Experimental platforms supply empirical validation and iteration (Pearl, 2009; Zhang et al., 2025). Social institutions supply the conditions that support risk-taking and non-conformity (Schumpeter, 1976; Christensen, 1997). The coupling of these components creates the conditions for breakthrough innovation (Zhang, 2016; Boden, 2004).

10.4 Outlook

The theory proposed here provides a framework for future research on AI and breakthrough innovation. The most promising direction is the development of hybrid systems that integrate human and AI capabilities in ways that compensate for their respective limitations (Fukumura and Ito, 2025; Boden, 1998). This requires advances in causal world modeling (Pearl, 2009; Zhang, 2021a-c; Nam et al., 2026), experimental closure (Ujváry et al., 2025), value alignment (Boden, 1998), and meta-cognitive monitoring (Zhang, 2026a). It also requires institutional innovations that support the collaboration between humans, AI, and experimental platforms (Schumpeter, 1976; Christensen, 1997). The question is not whether AI can replace human creativity but whether the coupling of human and AI capabilities can produce forms of breakthrough innovation that neither could achieve alone (Boden, 1998; Fukumura and Ito, 2025).

Appendix

Detailed Computational Methods

A.1 Formalization of the Breakthrough Innovation Occurrence Function

The breakthrough innovation occurrence function is defined as:

$$B = S^{\alpha} \cdot C^{\beta} \cdot E^{\gamma} \cdot V^{\delta} \cdot M^{\epsilon}$$

where:

$B \in [0,1]$ represents the probability or degree of breakthrough innovation occurrence

S $\in$ [0,1] represents the statistical generation capability

$C \in [0,1]$ represents the counterfactual reconstruction capability
$E \in [0,1]$ represents the experimental validation capability (Pearl, 2009)
$V \in [0,1]$ represents the value commitment and social cognition capability (Schumpeter, 1976)
$M \in [0,1]$ represents the meta-cognitive monitoring and self-negation capability
$\alpha, \beta, \gamma, \delta, \varepsilon > 0$ are parameters representing the relative importance of each component

The multiplicative form captures the necessary condition property: B approaches zero if any component approaches zero. This formalizes the intuition that all five components are required for breakthrough innovation (Boden, 2004).

A.2 Parameter Estimation Procedure

The parameters $\alpha, \beta, \gamma, \delta, \varepsilon$ can be estimated through maximum likelihood estimation using data on breakthrough innovation outcomes and measured levels of each component (Hastie et al., 2009). The log-likelihood function is:

$$L(\alpha, \beta, \gamma, \delta, \varepsilon) = \Sigma_i [y_i \cdot \log(B_i) + (1-y_i) \cdot \log(1-B_i)]$$

where $y_i \in \{0,1\}$ indicates whether breakthrough innovation occurred in case i, and B_i is computed using the measured component values for case i (Hastie et al., 2009).

A.3 Simulation Design

To explore the behavior of the model under different parameter settings, simulation experiments can be conducted by varying the component values and observing the resulting B values. The simulation would systematically vary each component from 0 to 1 while holding others constant, generating response curves that show how B changes with each component. Additional simulations would vary multiple components simultaneously to explore interaction effects.

A.4 Empirical Measurement Instrument

The measurement of S, C, E, V, and M would require validated instruments (Lamb et al., 2018). S could be measured through standard creativity tests assessing combinatorial and exploratory creativity (Boden, 2004). C could be measured through problem redefinition tasks and counterfactual reasoning assessments (Pearl, 2009). E could be measured through experimental design and implementation tasks. V could be measured through risk-taking and persistence assessments (Schumpeter, 1976). M could be measured through meta-cognitive monitoring tasks (Flavell, 1979). The development and validation of these instruments is a priority for future research (Fukumura and Ito, 2025).

A.5 Case Coding Scheme

For the comparative case analysis, a coding scheme was developed to assess each case on the five dimensions (Boden, 1998). Each dimension was coded on a 1-5 scale based on the degree to which the case exhibited the corresponding capability (Boden, 2004). Inter-rater reliability would be established through independent

coding by multiple raters (Creswell and Creswell, 2022).

The full Case Coding Scheme is written as a stand-alone codebook that can be used by independent raters. The five dimensions follow the operationalization in Section 3.1, which draws on Gero (2000), Casakin (2010), Boden (2004), Pearl (2009), etc. The scale anchors and decision rules are designed for comparative case analysis, not for psychometric measurement of individuals.

A.5.1 Purpose and Scope

This coding scheme assesses the degree to which a case exhibits breakthrough innovative thinking along five dimensions. It is intended for comparative analysis of AI-assisted, human-led, and hybrid innovation cases. It can also be used to compare AI-autonomous cases if and when they exist. The scheme does not assume that a case must be fully autonomous to be coded. It codes the observable capability profile of the case.

A.5.2 Unit of Analysis

The unit of analysis is a case episode. A case episode is a bounded innovation attempt that has a identifiable problem formulation, a set of actors, a knowledge base, an intervention or experimental component, and a documented outcome or output. The same organization or system may contribute multiple case episodes if the problem formulations differ substantially.

A.5.3 Five Dimensions and Operational Definitions

Dimension 1. Questioning default assumptions. This dimension assesses whether the case identifies and challenges implicit presuppositions that constrain problem formulation and solution search (Gero, 2000). Evidence includes explicit statements that a premise is being rejected, tests of an assumption that was previously taken for granted, or generation of alternatives that require dropping a default constraint.

Dimension 2. Redefining problems. This dimension assesses whether the case reformulates the problem space itself rather than only searching within a given space (Boden, 2004; Gero, 2000). Evidence includes changes in the objective function, changes in the variables considered relevant, changes in the success criteria, or a shift from one problem class to another.

Dimension 3. Cross-domain connection. This dimension assesses whether the case identifies structural similarities across domains and transfers knowledge from one domain to another (Casakin, 2010; Gero, 2000). Evidence includes analogies, metaphors, imported methods, or transferred formalisms that are not superficial but are used to generate new hypotheses or designs.

Dimension 4. Systemic reconstruction. This dimension assesses whether the case reconfigures multiple components and their relationships into a coherent new whole (Abernathy and Clark, 1985; Henderson and Clark, 1990). Evidence includes changes in architecture, interfaces, feedback loops, component roles, or system-level goals that go beyond local modification.

Dimension 5. Rapid experimental validation. This dimension assesses whether the case translates ideas into

testable interventions and learns from the results in iterative cycles (Pearl, 2009; Zhang, 2016). Evidence includes experimental design, simulation testing, robotic execution, real-world trials, or closed-loop revision based on empirical feedback.

A.5.4 Five-Point Anchored Rating Scale

Each dimension is coded on a 1 to 5 scale. The anchors are as follows.

Score 1. Absent or negligible. The case shows no meaningful evidence of the capability. The problem is accepted as given. No alternative assumption is tested. No cross-domain transfer occurs. No system-level reconfiguration occurs. No empirical test or feedback loop is present.

Score 2. Weak. The case shows a minimal or prompted instance of the capability. An assumption is mentioned but not challenged in action. A problem is restated in minor terms without changing the search space. A surface analogy is used. One component is changed while the architecture remains intact. A test is proposed but not executed, or executed without feedback.

Score 3. Moderate. The case shows a clear but bounded instance of the capability. Several assumptions are identified and alternatives are generated within the same paradigm. The problem is reframed in a way that changes the search space but remains within the existing paradigm. A structural analogy is used with partial mapping. Multiple components are integrated into a new configuration. A limited test is executed in simulation or a controlled setting, with some learning.

Score 4. Strong. The case shows a systematic and consequential instance of the capability. Core assumptions are challenged and alternative problem formulations are opened. The problem is redefined across a paradigm boundary, with new variables or objectives. A deep structural correspondence is identified and used to generate new hypotheses. The system architecture is reconstructed with new interfaces and feedback. An iterative test-learn cycle is executed in a real or high-fidelity environment.

Score 5. Transformative. The case shows a foundational and paradigm-changing instance of the capability. A foundational assumption is rejected and a new evaluative criterion or paradigm is established. A new problem class is created, and the old problem becomes a special case or becomes irrelevant. A new interdomain framework is created that reorganizes both domains. A new system-level paradigm is established with new components and goals. A closed-loop autonomous experimentation process revises both hypotheses and the system itself.

A.5.5 Decision Rules

Rule 1. Code the highest level of evidence observed, not the average or intended level. If the case contains one strong instance and several weak instances, code the strong instance for that dimension.

Rule 2. Do not infer capability from output quality alone. A novel output does not by itself prove that assumptions were questioned or that problems were redefined. Code the process evidence.

Rule 3. Distinguish prompted from spontaneous capability. If the capability appears only after an external prompt, code it at least one point lower than if it appears spontaneously in the case record.

Rule 4. Distinguish simulation from real-world validation. A simulation-only test can receive at most score 3 on Dimension 5 unless the simulation is high-fidelity and closed-loop.

Rule 5. Distinguish human-led from AI-autonomous action. Use the autonomy code in A.5.6 in addition to the five dimension scores. Do not collapse autonomy into the dimension scores.

Rule 6. If evidence is insufficient, code as 1 or mark NA. Do not guess. Report missing evidence in the coding sheet.

Rule 7. If two coders disagree by more than one point on any dimension, they must resolve the disagreement by returning to the case record and applying the anchors again.

A.5.6 Autonomy Code

Each case also receives an autonomy code. This code is separate from the five dimension scores.

A0. No AI. The innovation process is entirely human.

A1. AI as tool. AI performs a narrow auxiliary function such as search, formatting, or calculation. The problem formulation, evaluation, and decision authority remain human.

A2. AI as assistant. AI generates candidate ideas or analyses under human direction. Humans select, reframe, and validate.

A3. AI as co-designer. AI contributes to problem formulation, hypothesis generation, and solution refinement in an interactive loop with humans.

A4. AI as autonomous generator. AI generates and tests candidates within a defined problem space without continuous human direction. Humans define the problem and evaluate the output.

A5. AI as autonomous breakthrough agent. AI independently performs the full cycle of questioning assumptions, redefining problems, connecting domains, reconstructing systems, and validating experimentally. This code should be used only with direct evidence.

A.5.7 Case Selection and Data Sources

Cases should be selected to represent different domains, different autonomy codes, and different degrees of paradigm change. Recommended case types include protein structure prediction (Jumper et al., 2021), materials discovery (Merchant et al., 2023; Zeni et al., 2025), mathematical reasoning (Sun et al., 2025), artistic generation (Kim et al., 2025), multi-agent scientific collaboration (Swanson et al., 2025), causal world modeling (Nam et al., 2026), and embodied experimentation (Lin et al., 2026; Zhuang et al., 2026).

Data sources should include peer-reviewed papers, preprints, official technical reports, code repositories, experimental logs, and documented interviews where available. News reports may be used for context but should not be the sole basis for a score above 3 on any dimension.

A.5.8 Coding Procedure

Step 1. Assemble the case record. Extract all passages that describe problem formulation, assumptions, analogies, system changes, experiments, and feedback.

Step 2. Segment the case record by dimension. A single passage may provide evidence for more than one dimension.

Step 3. Score each dimension independently. Do not let a high score on one dimension influence another dimension.

Step 4. Assign the autonomy code.

Step 5. Write a brief evidence note for each score. The note must cite the specific page, section, figure, or code commit where the evidence appears.

Step 6. Repeat with a second independent coder.

Step 7. Compare scores. If disagreements exceed one point, hold a consensus meeting. If disagreements remain, record both scores and report the discrepancy.

A.5.9 Inter-Rater Reliability

Inter-rater reliability should be established before the main coding round. Use at least two independent coders on a training set of cases. Report the intraclass correlation coefficient for the five dimension scores and the percentage agreement for the autonomy code. A common threshold is ICC above 0.75 for good reliability and above 0.90 for excellent reliability. If reliability is below threshold, revise the anchors and retrain the coders. Creswell and Creswell (2022) provide general guidance on establishing reliability in mixed-methods research.

A.5.10 Coding Sheet Template

Case ID:

Case name:

Domain:

Source references:

Autonomy code:

Dimension 1. Questioning default assumptions. Score 1 to 5. Evidence note.

Dimension 2. Redefining problems. Score 1 to 5. Evidence note.

Dimension 3. Cross-domain connection. Score 1 to 5. Evidence note.

Dimension 4. Systemic reconstruction. Score 1 to 5. Evidence note.

Dimension 5. Rapid experimental validation. Score 1 to 5. Evidence note.

Overall profile. *Q*, *R*, *C*, *S*, *E*.

Coder initials:

Date:

Consensus score if applicable:

Disagreement note:

A.5.11 Illustrative Coding Example

The following example is for training only. It is not a final empirical coding of the case.

Case: A hypothetical AI system that predicts protein structures.

Autonomy code: A4. AI as autonomous generator within a defined problem space.

Dimension 1. Questioning default assumptions. Score 2. The case record shows that the system was asked to question whether the input representation was sufficient, but the challenge was prompted by human researchers and did not lead to a new problem formulation.

Dimension 2. Redefining problems. Score 1. The problem remained fixed as sequence-to-structure prediction. No new objective or success criterion was introduced.

Dimension 3. Cross-domain connection. Score 4. The system imported attention mechanisms and geometric constraints from other domains and used them to generate new structural hypotheses.

Dimension 4. Systemic reconstruction. Score 4. The system reconstructed the prediction pipeline by integrating multiple modules for sequence representation, geometric reasoning, and structure refinement.

Dimension 5. Rapid experimental validation. Score 5. The system was validated against experimental

structures in a closed-loop benchmark, and errors were used to revise the model.

Overall profile: *Q*2, *R*1, *C*4, *S*4, *E*5.

This profile indicates a strong statistical and experimental engine with weak problem redefinition and assumption questioning. It is consistent with the theoretical position that generative AI can be a powerful candidate generator and validator within a human-defined paradigm, but not an autonomous breakthrough agent.

A.5.12 Aggregation and Reporting

Do not sum the five dimension scores into a single creativity score. Report the profile vector. The profile vector preserves information about which capabilities are present and which are absent. If an aggregate is required for communication, report the geometric mean and clearly state that it is not a substitute for the profile. A case with scores *Q*1, *R*1, *C*5, *S*5, *E*5 has a different profile from a case with *Q*5, *R*5, *C*1, *S*1, *E*1, even if the arithmetic sum is the same.

A.5.13 Limitations

The scheme relies on documented evidence. Cases with poor documentation will receive lower scores by default. The scheme cannot distinguish between a capability that is absent and a capability that is present but undocumented. Coders should report documentation quality alongside the scores. The scheme also cannot resolve the philosophical question of whether AI systems have genuine understanding. It codes observable capability, not inner experience.

A.5.14 Summary

The full Case Coding Scheme consists of five dimensions, a five-point anchored scale, an autonomy code, decision rules, a coding procedure, a reliability protocol, a coding sheet template, and an illustrative example. It is designed to be used with the Three-Engine Four-Layer Coupling model in Section 6 and the empirical validation plan in Section 8. It can be applied to current AI-assisted cases and to future AI-autonomous cases without changing the core dimensions.

References

Abernathy WJ, Clark KB. 1985. Innovation: Mapping the winds of creative destruction. Research Policy, 14(1): 3-22. https://www.sciencedirect.com/science/article/abs/pii/0048733385900216

Acemoglu D. 2009. Introduction to Modern Economic Growth. Princeton University Press, Princeton, USA. https://press.princeton.edu/books/hardcover/9780691132921/introduction-to-modern-economic-growth

Aghion P, Howitt P. 1992. A model of growth through creative destruction. Econometrica, 60(2): 323-351. https://www.jstor.org/stable/2951599

Aristotle. trans. 1984. Metaphysics. In: The Complete Works of Aristotle (Barnes J, ed). Princeton University Press, Princeton, USA. https://press.princeton.edu/books/hardcover/9780691016511/the-complete-works-of-aristotle-volume-two

Bareinboim E, Correa JD, Ibeling D, Icard T. 2022. On Pearl's hierarchy and the foundations of causal inference. In: Probabilistic and Causal Inference: The Works of Judea Pearl (Geffner H, ed). ACM, New York, USA. https://dl.acm.org/doi/10.1145/3501714.3501743

Becker M, Cabeza R. 2024. Prediction error minimization as a common computational principle for curiosity and creativity. Behavioral and Brain Sciences, 47: e93. https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/prediction-error-minimization-as-a-common-computational-principle-for-curiosity-and-creativity/50C2809278B520BD06E75EAD4658F7D0

Boden MA. 1998. Creativity and artificial intelligence. Artificial Intelligence, 103(1-2): 347-356. https://www.sciencedirect.com/science/article/pii/S0004370298000551

Boden MA. 2004. The Creative Mind: Myths and Mechanisms. Routledge, London, UK. https://www.taylorfrancis.com/books/mono/10.4324/9780203508527/creative-mind-margaret-boden

Booth WC, Colomb GG, Williams JM, et al. 2016. The Craft of Research. 4th edition. University of Chicago Press, Chicago, USA. https://www.amazon.com/dp/022623973X?lv=shuf&channelId=520&plpRedirect=mhFallback

Brown TB, Mann B, Ryder N, et al. 2020. Language models are few-shot learners. 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Burger B, Maffettone PM, Gusev VV, et al. 2020. A mobile robotic chemist. Nature, 583(7815): 237-241. https://www.nature.com/articles/s41586-020-2442-2

Casakin H. 2010. Visual analogy, visual displays, and the nature of design problems: The effect of expertise. Environment and Planning B, 37(1): 170-188. https://ideas.repec.org/a/sae/envirb/v37y2010i1p170-188.html

Chen C, Ong SP. 2022. A universal graph deep learning interatomic potential for the periodic table. Nature Computational Science, 2: 718-728. https://www.nature.com/articles/s43588-022-00349-3

Chen CC, Jörke M, Goliński A, et al. 2026. LLMs are not (consistently) Bayesian: Quantifying internal (in)consistencies of LLMs' probabilistic beliefs. arXiv, arXiv:2605.06915. https://arxiv.org/abs/2605.06915

Christensen CM. 1997. The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail. Harvard Business School Press, Boston, USA. https://www.christenseninstitute.org/book/the-innovators-dilemma/

Christensen CM, Overdorf M. 2000. Meeting the challenge of disruptive change. Harvard Business Review, 78(2): 66-77. https://hbr.org/2000/03/meeting-the-challenge-of-disruptive-change

Clark A, Chalmers D. 1998. The extended mind. Analysis, 58(1): 7-19. https://www.jstor.org/stable/3328150

Cooper AC, Schendel D. 1976. Strategic responses to technological threats. Business Horizons, 19(1): 61-69. https://www.sciencedirect.com/science/article/abs/pii/0007681376900240

Creswell JW, Creswell JD. 2022. Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. 6th edition. Sage Publications, Thousand Oaks, USA. https://us2.sagepub.com/en-us/nam/research-design/book270550

Cropley DH. 2025. “The Cat Sat on the …?” Why generative AI has limited creativity. Journal of Creative Behavior, 59(4): e70077. https://onlinelibrary.wiley.com/doi/10.1002/jocb.70077

DeepMind. 2024. AlphaFold. https://deepmind.google/science/alphafold/. Accessed 15 September 2026.

Devlin J, Chang MW, Lee K, Toutanova K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT, 4171-4186. https://aclanthology.org/N19-1423/

Ding ZJ, Srinivasan A, Macneil S, et al. 2023. Fluid transformers and creative analogies: exploring large language models' capacity for augmenting cross-domain analogical creativity. Proceedings of the ACM Conference on Creativity and Cognition, 489-505. https://dl.acm.org/doi/10.1145/3591196.3593516

Flavell JH. 1979. Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. American Psychologist, 34(10): 906-911. https://www.ovid.com/journals/ampsy/abstract/00000487-197910000-00018~metacognition-and-cognitive-monitoring-a-new-area-of?redirectionsource=fulltextview

Franceschelli G, Musolesi M. 2024. Creativity and machine learning: A survey. ACM Computing Surveys, 56(11): 1-41. https://dl.acm.org/doi/full/10.1145/3664595

Fukumura K, Ito T. 2025. Can LLM-powered multi-agent systems augment human creativity? Evidence from brainstorming tasks. Proceedings of the ACM Collective Intelligence Conference, 20-29. https://dl.acm.org/doi/10.1145/3715928.3737479

Garcez A, Lamb LC. 2020. Neurosymbolic AI: The 3rd wave. arXiv, arXiv:2012.05876. https://arxiv.org/abs/2012.05876

Gero JS. 2000. Computational models of innovative and creative design processes. Technological Forecasting and Social Change, 64(2-3): 183-196. https://www.sciencedirect.com/science/article/pii/S0040162599001055

Goel V. 1995. Sketches of Thought. MIT Press, Cambridge, USA. https://direct.mit.edu/books/monograph/4716/Sketches-of-Thought

Goodfellow IJ, Bengio Y, Courville A. 2016. Deep Learning. MIT Press, Cambridge, USA. https://www.deeplearningbook.org/

Goodfellow IJ, Pouget-Abadie J, Mirza M, et al. 2014. Generative adversarial nets. Proceedings of the 28th International Conference on Neural Information Processing Systems, 27: 2672-2680. https://dl.acm.org/doi/10.5555/2969033.2969125

Grammenos D, Lubart T. 2025. Peeping at creAItivity through a keyhole: Creative self-perceptions, potential, and enhancement of GenAI chatbots. Artificial Intelligence Review, 58: 293. https://link.springer.com/article/10.1007/s10462-025-11288-6

Graves A, Wayne G, Reynolds M, et al. 2016. Hybrid computing using a neural network with dynamic external memory. Nature, 538(7626): 471-476. https://www.nature.com/articles/nature20101

Griffiths TL, Kemp C, Tenenbaum JB. 2008. Bayesian models of cognition. In: The Cambridge Handbook of Computational Psychology (Son R, ed). 59-100. Cambridge University Press, Cambridge, UK. https://psycnet.apa.org/record/2008-06911-003

Hastie T, Tibshirani R, Friedman J. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd Edition. Springer, New York, USA. https://link.springer.com/book/10.1007/978-0-387-84858-7

Henderson RM, Clark KB. 1990. Architectural innovation: The reconfiguration of existing product technologies and the failure of established firms. Administrative Science Quarterly, 35(1): 9-30. https://www.jstor.org/stable/2393549

Ho J, Jain AN, Abbeel P. 2020. Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33: 6840-6851. https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html

Hoffmann J, Borgeaud S, Mensch A, et al. 2022. Training compute-optimal large language models. arXiv, arXiv:2203.15556. https://arxiv.org/abs/2203.15556

Hogan A, Blomqvist E, Cochez M, et al. 2021. Knowledge graphs. ACM Computing Surveys, 54(4): 1-37.

https://dl.acm.org/doi/10.1145/3447772

Holtzman A, Buys J, Du L, et al. 2020. The curious case of neural text degeneration. arXiv:1904.09751. https://arxiv.org/abs/1904.09751

Horn MB. 2024. What does Disruptive Innovation Theory have to say about AI? https://www.christenseninstitute.org/blog/what-does-disruptive-innovation-say-about-ai/

Hupkes D, Dankers V, Mul M, et al. 2020. Compositionality decomposed: How do neural networks generalise? Journal of Artificial Intelligence Research, 67: 757-795. https://scispace.com/pdf/compositionality-decomposed-how-do-neural-networks-3hz9nrb7jv.pdf

IACC. 2024. International Association for Computational Creativity. https://computationalcreativity.net. Accessed 15 September 2026.

Ismayilzada M, Paul D, Bosselut A, van der Plas L. 2024. Creativity in AI: Progresses and challenges. arXiv, arXiv:2410.17218v3. https://arxiv.org/abs/2410.17218

Ito T, Dong Y, Haqbeen JA, Matsuo T, Sahab S. 2025. Hyperdemocracy: Towards creative consensus building between humans and AI. Proceedings of the IEEE International Conference on Agentic AI, 116-121. https://www.semanticscholar.org/paper/Hyperdemocracy%3A-Towards-Creative-Consensus-Building-Ito-Dong/afdd5d55e0962c15fd0ac3496af702a6718eb409

Jordanous A. 2012. A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. Cognitive Computation, 4(3): 246-279. https://link.springer.com/article/10.1007/s12559-012-9156-1

Jumper J, Evans R, Pritzel A, et al. 2021. Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873): 583-589. https://www.nature.com/articles/s41586-021-03819-2

Jung RE, Vartanian O (eds). 2018. The Cambridge Handbook of the Neuroscience of Creativity. Cambridge University Press, Cambridge, UK. https://www.cambridge.org/core/books/cambridge-handbook-of-the-neuroscience-of-creativity/CE1FF0E1875FDB0C95F26DADA507958E

Kaplan J, McCandlish S, Henighan T, et al. 2020. Scaling laws for neural language models. arXiv, arXiv:2001.08361. https://arxiv.org/abs/2001.08361

Kim M, Kim S, Thorne J. 2025. From evidence to belief: A Bayesian epistemology approach to language models. Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 10578–10611. https://aclanthology.org/2025.naacl-long.531.pdf

Kingma DP, Welling M. 2014. Auto-encoding variational Bayes. Proceedings of ICLR, 1-14. https://arxiv.org/pdf/1312.6114

Kuhn TS. 1962. The Structure of Scientific Revolutions. University of Chicago Press, Chicago, USA. https://press.uchicago.edu/ucp/books/book/chicago/S/bo13179781.html

Lake BM, Baroni M. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. Proceedings of the 35th International Conference on Machine Learning, PMLR 80: 2873-2882. https://proceedings.mlr.press/v80/lake18a.html

Lamb C, Brown DG, Clarke CL. 2018. Evaluating computational creativity: An interdisciplinary tutorial. ACM Computing Surveys, 51(2): 1-34. https://dl.acm.org/doi/10.1145/3167476

Latour B. 2005. Reassembling the Social: An Introduction to Actor-Network-Theory. Oxford University Press, Oxford, UK. http://media1.cagd.co.uk/422/811840_75472b.pdf

LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. Nature, 521(7553): 436-444. https://www.nature.com/articles/nature14539

Lin XY, Zhang YY, Gan YH, et al. 2026. Human-AI co-embodied intelligence for scientific experimentation

and manufacturing. arXiv, arXiv: 2511.02071v2. https://arxiv.org/abs/2511.02071

Merchant A, Batzner S, Schoenholz SS, et al. 2023. Scaling deep learning for materials discovery. Nature, 624(7990): 80-85. https://www.nature.com/articles/s41586-023-06735-9

Merton RK. 1973. The Sociology of Science: Theoretical and Empirical Investigations. University of Chicago Press, Chicago, USA. https://press.uchicago.edu/ucp/books/book/chicago/S/bo28451565.html

Nam H, Le Lidec Q, Maes L, LeCun Y, Balestriero R. 2026. Causal-JEPA: Learning world models through object-level latent interventions. arXiv, arXiv:2602.11389. https://arxiv.org/abs/2602.11389

Ouyang L, Wu J, Jiang X, et al. 2022. Training language models to follow instructions with human feedback. arXiv, arXiv: 2203.02155. https://arxiv.org/abs/2203.02155

Pearl J. 2009. Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge, UK. https://dl.acm.org/doi/book/10.5555/1642718

Rogers A, Kovaleva O, Rumshisky A. 2020. A primer in BERTology: What we know about how BERT works. Transactions of the Association for Computational Linguistics, 8: 842-866. https://aclanthology.org/2020.tacl-1.54/

Royal Swedish Academy of Sciences. 2024. Nobel Prize in Chemistry 2024. https://www.nobelprize.org. Accessed 15 September 2026.

Runco M. 2024. The discovery and innovation of AI does not qualify as creativity. Journal of Cognitive Psychology, 1-10. https://www.tandfonline.com/doi/full/10.1080/20445911.2024.2436362

Schumpeter JA. 1976. Capitalism, Socialism and Democracy. Routledge, London, UK. https://www.taylorfrancis.com/books/mono/10.4324/9780203202050/capitalism-socialism-democracy-joseph-schumpeter

Scrum.org. 2024. Scrum First Principles. https://www.scrum.org/resources/blog/scrum-first-principles. Accessed 15 September 2026.

Sedkaoui S, Benaichouba R. 2026. Generative AI as a transformative force for innovation: A review of opportunities, applications and challenges. European Journal of Innovation Management, 29(3): 677-701. https://www.sciencedirect.com/org/science/article/abs/pii/S1460106023000792

Shen A, Druckmann S, Zou J. 2026. Unlocking LLM creativity in science through analogical reasoning. arXiv, arXiv: 2605.11258. https://arxiv.org/abs/2605.11258

Sun YY, Hu S, Zhou G, et al. 2025. OMEGA: Can LLMs reason outside the box in math? Evaluating exploratory, compositional, and transformative generalization. arXiv, arXiv: 2506.18880. https://arxiv.org/abs/2506.18880

Swanson K, Wu W, Bulaong NL, et al. 2025. The Virtual Lab of AI agents design new SARS-CoV-2 nanobodies. Nature, 646: 716–723. https://www.nature.com/articles/s41586-025-09442-9

Tan JJ, Xiao XX. 2025. Harness first-principles thinking in problem-based learning for chemical education. Journal of Chemical Education, 102 (2): 943–947. https://pubs.acs.org/jceda8/article/102/2/943/3694094/Harness-First-Principles-Thinking-in-Problem-Based

Tenenbaum JB, Kemp C, Griffiths TL, Goodman ND. 2011. How to grow a mind: Statistics, structure, and abstraction. Science, 331(6022): 1279-1285. https://www.science.org/doi/10.1126/science.1192788

Tenney I, Das D, Pavlick E. 2019. BERT rediscovers the classical NLP pipeline. Proceedings of ACL, 4593-4601. https://aclanthology.org/P19-1452/

Tushman ML, Anderson P. 1986. Technological discontinuities and organizational environments. Administrative Science Quarterly, 31(3): 439-465. https://www.jstor.org/stable/2392832

Ujváry S, Mészáros A, Brendel W, et al. 2025. Implicit Bayesian inference is an insufficient explanation of

language model behaviour in compositional tasks. ICLR 2025 Workshops: DeLTa. https://mlanthology.org/iclrw/2025/ujvary2025iclrw-implicit/

Vaswani A, Shazeer N, Parmar N, et al. 2017. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems, 6000-6010. https://dl.acm.org/doi/10.5555/3295222.3295349

Webb T, Holyoak KJ, Lu HJ. 2023. Emergent analogical reasoning in large language models. Nature Human Behaviour, 7: 1526-1541. https://www.nature.com/articles/s41562-023-01659-w

Wei J, Tay Y, Bommasani R, et al. 2022. Emergent abilities of large language models. Transactions on Machine Learning Research, 1-30. https://mlanthology.org/tmlr/2022/wei2022tmlr-emergent/

World Fund. 2026. Materials discovery at machine speed. https://www.worldfund.vc/knowledge/materials-discovery-at-machine-speed. Accessed 15 September 2026.

Yang LP, Xin T, Yu YY, et al. 2026. Human vs. LLM creativity: A comparative analysis of task-dependent asymmetry and linguistic mechanisms. Journal of Intelligence,14(2): 27. https://www.mdpi.com/2079-3200/14/2/27

Zeni C, Pinsler R, Zügner D, et al. 2025. A generative model for inorganic materials design. Nature, 639: 624–632. https://www.nature.com/articles/s41586-025-08628-5

Zhang S, Yang SR, Xie T, et al. 2025. Position: Intelligent science laboratory requires the integration of cognitive and embodied AI. arXiv, arXiv: 2506.19613. https://arxiv.org/abs/2506.19613

Zhang WJ. 2007a. Supervised neural network recognition of habitat zones of rice invertebrates. Stochastic Environmental Research and Risk Assessment, 21: 729-735. https://link.springer.com/article/10.1007/s00477-006-0085-y

Zhang WJ. 2007b. Pattern classification and recognition of invertebrate functional groups using self-organizing neural networks. Environmental Monitoring and Assessment, 130: 415-422. https://link.springer.com/article/10.1007/s10661-006-9432-1

Zhang WJ. 2010. Computational Ecology: Artificial Neural Networks and Their Applications. World Scientific, Singapore. https://www.worldscientific.com/worldscibooks/10.1142/7436#t=aboutBook

Zhang WJ. 2012. Computational Ecology: Graphs, Networks and Agent-based Modeling. World Scientific, Singapore. https://www.worldscientific.com/worldscibooks/10.1142/8118#t=aboutBook

Zhang WJ. 2016. Selforganizology: The Science of Self-Organization. World Scientific, Singapore. https://www.worldscientific.com/worldscibooks/10.1142/9685#t=aboutBook

Zhang WJ. 2018. Fundamentals of Network Biology. World Scientific Europe, London, UK. https://www.worldscientific.com/worldscibooks/10.1142/q0149#t=aboutBook

Zhang WJ. 2021a. Causality inference of linearly correlated variables: The statistical simulation and regression method. Computational Ecology and Software, 11(4): 154-161. http://www.iaees.org/publications/journals/ces/articles/2021-11(4)/2-Zhang-Abstract.asp

Zhang WJ. 2021b. Causality inference of nominal variables: A statistical simulation method. Computational Ecology and Software, 11(4): 142-153. http://www.iaees.org/publications/journals/ces/articles/2021-11(4)/1-Zhang-Abstract.asp

Zhang WJ. 2021c. A statistical simulation method for causality inference of Boolean variables. Network Biology, 11(4): 263-273. http://www.iaees.org/publications/journals/nb/articles/2021-11(4)/2-Zhang-Abstract.asp

Zhang WJ. 2022. p-value based statistical significance tests: Concepts, misuses, critiques, solutions and beyond. Computational Ecology and Software, 12(3): 80-122. http://www.iaees.org/publications/journals/ces/articles/2022-12(3)/1-Zhang-Abstract.asp

Zhang WJ. 2026a. The Feedback Integration Theory of Consciousness: A unified neurobiological framework. Network Biology, 16(4): 552-585. http://www.iaees.org/publications/journals/nb/articles/2026-16(4)/5-Zhang-Abstract.asp

Zhang WJ. 2026b. MR PHARMACIST: AI-Based medication consultation platform. Network Biology, 16(3): 268-388

Zhang WJ. 2026c. DR SUN YATSEN: An AI-Based medical clinic. Network Biology, 16(3): 117-267

Zhang WJ. 2026d. TraitGenePathAna: The AI-Powered biological trait analysis platform. Network Biology, 16(2): 49-81

Zhang WJ. 2026e. Species Distribution Finder: The AI-driven web tool to discover where species live around the world. Computational Ecology and Software, 16(2): 172-197

Zhang WJ. 2026f. AI-driven assessment of plant adaptation to climate change: The web tool based on physiological, biological, morphological, and genetic indicators of plant species. Computational Ecology and Software, 16(2): 99-153

Zhang WJ. 2026g. AI-driven assessment of animal adaptation to climate change: The web tool based on physiological, morphological, behavioral, and genetic indicators of animal species. Computational Ecology and Software, 16(1): 58-98

Zhang WJ. 2026h. A GWAS data fetcher with AI for Mendelian Randomization analysis. Network Pharmacology, 11(3-4): 62-89

Zhang WJ. 2027a. Can generative AI produce consciousness? A cross-disciplinary literature review and critical analysis. Selforganizology, 14(3-4): 34-46. http://www.iaees.org/publications/journals/selforganizology/articles/2027-14(3-4)/2-Zhang-Abstract.asp

Zhang WJ. 2027b. General design principles for combining functional modules into complete cells: From compatibility to emergence. Network Biology, 17(2): 538-585. http://www.iaees.org/publications/journals/nb/articles/2027-17(2)/3-Zhang-Abstract.asp

Zhang WJ, Bai CJ, Liu GD. 2007. Neural network modeling of ecosystems: a case study on cabbage growth system. Ecological Modelling, 201: 317-325. https://www.sciencedirect.com/science/article/pii/S0304380006004480

Zhang WJ, Barrion AT. 2006. Function approximation and documentation of sampling data using artificial neural networks. Environmental Monitoring and Assessment, 122: 185-201. https://link.springer.com/article/10.1007/s10661-005-9173-6

Zhang WJ, Liu GH, Dai HQ. 2008. Simulation of food intake dynamics of holometabolous insect using functional link artificial neural network.. Stochastic Environmental Research and Risk Assessment, 22(1): 123-133. https://link.springer.com/article/10.1007/s00477-006-0102-1

Zhang WJ, Zhang XY. 2008. Neural network modeling of survival dynamics of holometabolous insects: a case study. Ecological Modelling, 211: 433-443. https://www.sciencedirect.com/science/article/pii/S030438000700498X

Zhuang X, Zhou CY, Feng KH, et al. 2026. Embodied science: Closing the discovery loop with agentic embodied AI. arXiv, arXiv: 2603.19782. https://arxiv.org/abs/2603.19782