Language-Model Agents Reveal How Demand, Network and Collaboration Dynamics Shape Collective Innovation

Lancaster Wu[1,2]

[1*]Department of Sociology, University of Chicago, 5801 S Ellis Ave, Chicago, 60637, IL, United States.
[2]Knowledge Lab, University of Chicago, 211 E 60th St, Chicago, 60637, IL, United States.

Thesis Advisor: James Evans
Preceptor: Maximilian Cuddy

A Thesis Submitted in Partial Fulfilment
of the Requirements for the Degree of
Master of Arts
in Master of Arts Program in the Social Sciences (MAPSS)

AUGUST 2025

ABSTRACT

Innovation research has oscillated for decades between "technology-push" and "market-pull" explanations, yet empirical tests that isolate their joint dynamics remain scarce. We combine historical synthesis with a new methodological contribution: autonomous populations of large-language-model (LLM) agents that reason, converse and innovate in silico. After tracing how modern growth theory evolved from supply-side linear models to demand-sensitive, networked systems, we deploy four agent-based experiments. Survival-scarcity simulations show that moderate resource pressure spurs early, high-quality collective inventions, whereas abundance breeds complacency and extreme inequality multiplies but degrades innovations. Network-topology experiments reveal that fully connected societies exploit ideas quickly but converge prematurely, ring lattices preserve diversity yet diffuse slowly, and emergent small-world structures balance both, especially when agents display heterogeneous "engineer", "artist" and "scientist" personas in a more complex simulation set up. In career-long academic ecosystems, we replicate 20 years of scholarship under three incentive regimes. Publish-or-perish rules maximize paper counts but generate roughly eight-times fewer breakthroughs, truncate researcher careers and accentuate Matthew-effect citation inequality. Five-year HHMI-style support delivers nearly an order-of-magnitude more breakthroughs, sextuples new paradigms, preserves topic diversity and keeps over 90% of scholars active; a dual-tier system lands in between. Across experiments, innovation thrives in a Goldilocks zone: sufficient urgency to provoke action, structural diversity to explore alternatives, and incentive horizons long enough to reward risk. Our results align with decades of organisational and science-policy scholarship, yet are produced by agents drawing only on self-consistent language priors—demonstrating that LLM societies can serve as reproducible, high-throughput "wind tunnels" for social-scientific theory. Limitations and for this emerging methodology are also discussed.

Keywords: Agent-Based Modeling, Social Simulation, Large Language Model, Innovation, Science of Science, AI

Contents

I began this project with the aim of understanding how demand structures and incentive regimes shape collective innovation, and of examining how these forces influence individual contributions—an inquiry sparked by my conversations with Professor James Evans in the summer of 2024. This work would not have been possible without Professor Evans's mentorship and the unwavering support of my colleagues at the Knowledge Lab.

I am also grateful to Professors Maximilian Cuddy, Ayelet Fishbach, and Panagiotis Toulis for their insightful suggestions on this thesis, and to Otabek Khusinov for our discussion on the causal-inference analysis of the preliminary simulation during the BUSN 41207 final project.

Usage note. Throughout this thesis, I employ the first-person plural ("we") when describing the research process; this wording follows disciplinary convention and highlights the inherently collaborative nature of scientific inquiry.

1 Introduction

Why have radical scientific and technological breakthroughs slowed even as global R&D spending, the number of scientists, and the stock of knowledge have climbed to historic highs? A growing body of "science-of-science" work documents two striking regularities: research productivity has fallen—each new idea costs more effort than the last-and published work is increasingly consolidative rather than disruptive[1][2].These trends frame a central puzzle for innovation theory.

Three explanatory traditions offer competing diagnoses. Supply-side accounts emphasize the rising burden of knowledge: entry to the frontier takes longer, making exploration harder[3]. Demand-side perspectives argue that markets and narratives pull effort toward safe improvements or, at times, whip it toward fads[4][5][6]. Structure-centric work points to social architecture and incentives: networks can either recombine distant ideas or entrench incumbents, while short-term evaluation regimes can rationally deter risk[7][8][9][10][11][12][13]. Yet decisive evidence remains scarce because we rarely can randomize scarcity, sentiment, network topology, or incentive rules in vivo at the scale of scientific communities, and standard agent-based models (ABMs) often reduce creativity to fixed rule sets that miss the open-ended, language-mediated reasoning central to invention[14][15]. We still lack an integrated, causal testbed that can vary these levers in controlled ways and observe how their joint dynamics shape collective innovation over time. Natural data are rich but confounded; lab studies are clean but small-scale and short-lived.

This thesis addresses that methodological gap by introducing LLM societies—reusable, high-throughput "wind tunnels" populated by large-language-model agents whose cognition, coordination, and institutional evolution unfold in natural language. Compared with traditional ABMs, LLM agents can originate genuinely novel proposals (not just mutate tokens), debate and negotiate in text, and adapt strategically—capacities demonstrated in recent multi-agent studies[16][17][18][19]. At the same time, the design explicitly confronts known risks of foundation models (bias, over-conformity, and reproducibility concerns), which we mitigate through random assignment, fixed seeds, rich logging, and sensitivity analyses[20][21].

The contribution here goes beyond "we can now simulate with LLMs" in three ways. First, causal identification: within a single computational population we exogenously vary scarcity, demand signals, resource allocation, enabling clean head-to-head tests of classic theories (Randomization Inference - Section 4.1). Second, full-cycle measurement: because agents propose, deliberate, revise, diffuse, and institutionalize ideas in text, we observe the entire innovation pipeline—timing, quality, diffusion, and paradigm formation—rather than a narrow proxy. Third, innovative simulation design (see Appendix A workflow schematics): we implement architecture-level features that keep the society both controllable and genuinely inventive, such as a dedicated supervisor "Game Master" agent that monitors norms, adjudicates agent requests, and—when feasible—enables agents to change the environment (e.g., creating tools, locations, and rules) (See Section 3.1). Other architecture designs include adaptive control of agents' generative "temperature" to mimic human creativity, hierarchical memory and metacognitive loop, adaptive network rewiring, agent peer review system, etc (See Section 4).

In the preliminary simulation, we estanlished a 15-agent 1800s town with heterogeneous roles, RAG-style memories, and a "game-master" environment allows us to vary resource pressure and equality. A causal inference (randomization inference) showed us moderate scarcity accelerates earlier and higher-quality innovations; extreme inequality increases counts but degrades quality and coordination—an urgency-but-not-collapse dynamic that motivates later designs-consistent with induced-innovation logic under factor price pressure and historical episodes where shocks spur invention[5][22][23].

The preliminary simulation used GPT-o3 model before OpenAI lowered the model price, 54,664,776 tokens were used and with $2053.6 cost. Therefore, we decide to focus on scientific innovation with more structured simulation designs than letting agents freely communicate, interact, and create in a digital village.

Simulation 1 include fifty persona-diverse agents generate, share, and adopt ideas on fully-connected, ring, small-world, and scale-free graphs. We observed that small-world structures balance rapid spread with preserved diversity; hub-dominated networks can bottleneck diffusion—directly informing exploration–exploitation theory in networks[7][8][9][10].

Simulation 2 is build based on simulation 1, with agents gain biographies, hierarchical memory (observations–reflections–plans), dynamic creativity (temperature tuning), and adaptive rewiring. We found that learning reduces stagnation, boosts novelty, and endogenously yields small-world-like structures that sustain exploration, showing how individual adaptation scales to system-level creativity[11].

Simulation 3 provided a 20-year, 50-researcher ecosystem under three regimes. We see that publish-or-perish maximizes volume but yields far fewer breakthroughs and higher inequality; HHMI-style horizons generate an order-of-magnitude more breakthroughs, more paradigms, and longer careers; dual-tier lands in between—echoing causal evidence from real funding regimes and formal theory[12][13][24].

We now turn to the Background to situate these levers in the broader literatures our experiments adjudicate.

2 Background

We start by reviewing the theories from the supply perspective of innovation, which explains capacity and the creation of ideas. This tradition runs from the linear model and exogenous growth to endogenous growth, Schumpeterian creative destruction, and evolutionary routines [25–30]. It motivates large postwar investments in basic science, research infrastructure, and talent formation [31–33]. Within this frame, five drivers organize the supply story: knowledge and R&D, human talent and skills, funding and incentives, institutions and policy environment, and networks and collaboration [26, 28, 34–36]. But a modern paradox tempers pure push: inputs rise while disruptiveness often falls and ideas appear harder to find [1, 2]. Explanations include the burden of knowledge, field maturation, and incentive regimes that rationally favor incremental projects over risky leaps [3, 37–40].

We therefore pivot to the demand perspective of innovation, which explains direction and timing. Demand encompasses market size and expected profits, scarcity and relative prices, popular narratives and social contagion, and the incentive and reward systems that shape investment and adoption [4–6, 41, 42]. Induced-innovation and directed technical change theories predict that when a factor becomes dear or a mission creates credible demand, search reorients and timelines compress—provided suitable capabilities and complements exist [22, 23, 43–45]. Narratives and peer effects influence diffusion and capital allocation, producing booms, busts, and path dependence [46–50]. Taken together, *push* and *pull* interact: capabilities without demand can stall; demand without capabilities can fizzle.

2.1 Supply Perspective of Innovation

Early innovation studies concentrated on a supply-side, technology-push view of what drives new ideas. The mid-20th century "linear model of innovation" imagined progress as a one-way pipeline from scientific research to invention, then development, and finally diffusion. Early neoclassical growth models treated technological innovation as an exogenous factor. In Solow's model, long-run growth came from a "technical progress" residual unexplained by the model[25]. Since then, governments and industries have deliberately fed this engine of knowledge creation. Vannevar Bush's 1945 report Science - The Endless Frontier famously argued that basic scientific research is "the pacemaker of technological progress" and that new products and processes are born from "new principles and new conceptions, which in turn are painstakingly developed by research in the purest realms of science"[31].

This postwar philosophy of "science-push" laid the groundwork for massive public investments in R&D, from the establishment of agencies like the U.S. National Science Foundation to sustained funding of university research and national labs. The underlying belief, borne out by subsequent decades, was that a robust supply of scientific knowledge would eventually translate into societal and economic advances. Economists formalized this intuition in endogenous growth theory (e.g. Romer, Aghion & Howitt), making knowledge creation (education, R&D) a central driver of growth[26][27][28]. These models stress increasing returns from knowledge (since ideas can be reused

at low cost) and predict that investments in research and human capital will fuel sustained growth.

Similarly, economists like Joseph Schumpeter highlighted the role of entrepreneurs and described innovation as "the incessant product and process innovation mechanism by which new production units replace outdated ones", a process he famously termed creative destruction[29]. He famously argued that innovation is the engine of economic development. In his view, economic evolution is driven by the supply side: entrepreneurs introduce "new combinations" (innovations) that disrupt the circular flow of the economy[51]. Consumers, by contrast, play a relatively passive role in this process of creative destruction. Schumpeter highlighted how entrepreneurs' innovative activities break routine patterns and create structural change. Modern Schumpeterian (or Neo-Schumpeterian) growth theories build on this idea, modeling how profit-seeking firms invest in R&D to generate new technologies, with old technologies being destroyed in the process[52]. This perspective emphasizes creative destruction as the source of progress and sees competition as a dynamic process of innovation races.

Evolutionary theorists view innovation as an ongoing process of variation, selection, and retention. In the influential formulation by Nelson and Winter, firms do not optimize perfectly; instead they follow routines and learn by doing, and those routines are analogous to genes in biological evolution[30]. Successful routines (e.g. effective research practices or production techniques) are retained and spread, while others are discarded. Innovation thus arises from a diverse population of ideas and approaches, with market competition acting as a selection environment. This approach, inspired by Schumpeter, highlights bounded rationality, uncertainty, and path dependence in innovation.

From the supply perspective, there are a few key drivers of the innovation: knowledge and R&D, human talent and skills, funding and incentives, institutions and policy environment, networks and collaboration. A strong base of scientific and technical knowledge is fundamental to the supply of innovation. Research activities (both basic and applied R&D) generate new knowledge that innovators can build upon. Endogenous growth theory explicitly underscores that knowledge creation, education, and R&D investment are key drivers of technological progress[52]. Moreover, because knowledge has partially public-good properties—ideas spill over to others—private markets alone tend to under-invest in the creation and dissemination of new knowledge[34]. Because firms cannot fully appropriate the returns on knowledge, there is a role for policy (such as public research funding or R&D) to bolster the knowledge stock. In short, a growing and diffusing pool of knowledge lowers the cost of subsequent innovation by letting researchers "stand on the shoulders of giants," enabling cumulative advances[53][54][55][56][57].

Innovation is also driven by people - scientists, engineers, entrepreneurs, and other skilled workers. High levels of human capital increase an economy's innovative capacity. Studies find that innovative firms employ more highly skilled workers (especially those with technical or scientific training) than less innovative firms[58]. Educated and talented individuals contribute new ideas and have the expertise to solve complex problems. Thus, investment in education (universities, STEM training, PhD programs)

expands the talent supply for innovation. In the modern knowledge economy, talent tends to concentrate in innovation hubs, and competition for skilled researchers is global. For example, countries that dramatically expanded higher education and training have boosted their innovation output; China, for instance, had a ≈40% rise om its annual STEM PhD graduates by the mid-2020s, far surpassing the United States. Ensuring a strong pipeline of talent through education, immigration policies, and skill development is seen as critical for sustaining innovation on the supply side[59][60][61][62].

Innovation often requires substantial financial resources and economic incentives. Firms fund innovation through retained earnings and private capital (such as venture capital for startups), but these can be bolstered by public funding and policies. Governments commonly use supply-side innovation policies like R&D grants, subsidies, and tax incentives to lower the cost of research and encourage more private R&D investment[63][64][65]. Such policies increase the supply of innovation by directly funding research or by incentivizing firms to invest in it. Indeed, many nations' innovation strategies historically favored these supply-side instruments, reflecting a belief that boosting R&D inputs would yield more innovation.

The institutional context also plays a powerful role in shaping innovation supply. Strong research institutions serve as the organisational bedrock of a nation's knowledge-creation system. Universities both generate frontier science and educate future inventors[32][33]. Historical experience beginning with Vannevar Bush's WWII mobilisation and carried forward by DARPA-style agencies—shows that targeted public research programmes can lift a country's innovative capacity[31][66][67][68]. This led to policies focused on funding upstream science (a supply-side approach), and many countries still lean on this model. Over time, policy thinking has become more systemic (recognizing interactions between actors and the role of demand), but institutions that support the supply of innovation remain central[69][70][71][72][73]. These include not just funding agencies and universities, but also legal institutions (intellectual property law, antitrust policy, research ethics regimes) and physical infrastructure. Research infrastructure such as well-equipped laboratories, high-speed computing networks, shared facilities (e.g. nanotech fabrication labs), or large-scale instruments (telescopes, particle colliders) can dramatically expand what scientists and inventors are able to do, thereby increasing innovation potential. Regions with robust innovation ecosystems – where universities, firms, and government labs interact under supportive institutions – tend to generate more scientific and technical innovation[74][75][76].

Sociological and Science & Technology Studies (STS) perspectives emphasize that innovation is fundamentally a social process. The connections and interactions among people and organizations critically shape the supply of new ideas[77][24]. These interactions create the relational infrastructure that shapes what knowledge is produced and how quickly it diffuses. Large-scale "science of science" work confirms that the structure of collaboration networks governs the flow of ideas across people, firms and borders[35][36]. When collaboration knits together researchers who span disciplines, sectors or countries, the resulting teams draw on more varied information and are significantly more likely to generate high-impact, novel work. Meta-studies show that (i) teams have overtaken solo authors in producing top-cited research[78]; (ii) papers that

combine "atypical" knowledge with well-established ideas have roughly double the chance of becoming top-1% citations[79]; and (iii) small interdisciplinary teams disproportionately "disrupt" fields with fresh combinations, whereas large teams efficiently develop existing trajectories[80]. These findings support the long-standing STS insight that breakthroughs often emerge at the intersections of fields where contrasting perspectives collide[79]. Geography magnifies the effect: dense professional clusters—such as Silicon Valley or Cambridge's biotech corridor—facilitate trust-based knowledge sharing and accelerate invention[81][82]. Social norms also matter: the scientific norm of openly publishing and disseminating results (the "open science" ethos) increases the diffusion of knowledge, allowing more actors to build on each other's work; empirical work shows that studies making their data openly available enjoy a measurable citation advantage, highlighting how open science practices speed diffusion and reuse[83].

2.2 The Modern Paradox of Innovation

By many measures, the supply side of scientific innovation today is as strong as it has ever been: knowledge is accumulating at a rapid clip, R&D funding is at record highs, and the ranks of scientists have swelled. The volume of scientific publications and patents has grown exponentially in recent decades, and new fields from biotechnology to artificial intelligence have come into being, nourished by this expanding base of knowledge and resources. One might expect, under the classic "shoulders of giants" logic, that we are poised for an era of unprecedented breakthrough innovation[84][85]. And yet, a paradoxical trend has garnered increasing attention: despite the ever-growing inputs, the output of truly fundamental, game-changing innovations appears to be slowing in many domains[86]. Across diverse industries and research areas, more and more personnel and funding are needed to eke out equivalent advances, implying that each researcher today contributes, on average, fewer "ideas" than in the past. One landmark study quantified this across the entire U.S. economy and found research productivity has fallen so sharply over the decades that to maintain even a modest rate of economic growth, R&D effort had to rise exponentially – in effect, ideas are getting harder to find[1]. Similarly, a recent study introduced a metric for disruptive innovation and showed that papers and patents are becoming less likely to break with the past and push science in new directions[2]. The share of work that represents big, field-reorienting leaps has dropped drastically over the last half-century, across virtually all research fields. Instead, most of the new knowledge builds on well-established foundations, extending them incrementally rather than upending them. In short, today's innovations tend to be more consolidative than disruptive.

What might explain these trends of fewer blockbuster breakthroughs and diminishing returns on research effort? Several interpretations have been put forward, and they are not mutually exclusive. One simple explanation is that the low-hanging fruit have already been picked. As fields mature, the obvious foundational discoveries may get made early on, leaving later researchers with inherently more complex or elusive problems to solve. This "burden of knowledge" hypothesis posits that new innovators must spend more years learning an ever-expanding literature, and when they finally reach the research frontier, the gains to be had are more incremental. In economic terms, there may be diminishing marginal returns to adding more researchers unless

there are commensurate revolutionary breakthroughs (a dynamic masked in part by the fact that exponential growth of R&D has sustained overall progress, albeit at a higher cost)[2][3].

Another factor is the "publish or perish" culture and the incentives in modern science. As scientific communities grow, the competition for funding and publication can encourage safer, incremental research at the expense of riskier paradigm-changing ideas[87][37]. Such incentive regime that now governs most academic careers. Formal modelling and survey evidence show that when hiring, promotion and grants hinge on a steady stream of papers, researchers rationally tilt toward projects with a high probability of publishable results—even if those projects promise only modest intellectual returns[39][40]. Large-scale bibliometric analyses make the consequences visible. In a corpus of 6.5 million biomedical-chemistry articles (1983-2008), papers that merely repeat previously studied chemical relations appear six times more often than papers that introduce entirely new ("jump") relations; this 6:1 gap stayed almost unchanged for 25 years despite an explosion of unexplored possibilities[38]. Because attention is channelled toward the familiar, the post-war expansion of the scientific workforce has not produced a commensurate surge in disruptive breakthroughs: across 45 million papers and 3.9 million patents (1960-2020), the average "CD" disruptiveness score has fallen steadily in every major field[2]. The skewed pay-off structure explains why high-risk, high-reward exploration remains rare. Research has shown that "jump" projects fail or are ignored far more often than incremental work, yet when they succeed they attract disproportionately high citations and prizes; on average, however, the extra reward does not fully offset the greater risk, making such gambles unattractive to individual scientists[38]. The same pattern occur at the team level: small teams that depart from the mainstream disrupt science and technology far more than the large teams that now dominate publication counts, but they constitute a shrinking share of total output[88]. Collectively, millions of individually rational decisions to "play it safe" translate into a macro-level slowdown in paradigm-changing discovery, even as the volume of research and the stock of available knowledge continue to grow.

2.3 Demand Perspective of Innovation

While the above "slowdown" interpretations focus on the supply side of innovation, Robert J. Shiller argues that innovation studies still overlook the other side of the ledger: the willingness of markets and society to absorb new ideas. He catalogued a century of episodes in which exuberant "new-era" stories about electricity, radios or the Internet first amplified demand and then collapsed when adoption lagged behind expectations[42]. Yet, as he later noted in his American Economic Association presidential address, economics has "no systematic theory of the popular narratives that drive these booms and busts"[6][50]. He argues that contagious stories about a technology's promise-or its dangers-shape adoption curves, capital allocation and ultimately the realised social return on invention. Sentiment and market pull are themselves crucial inputs to innovation outcomes, mediating whether fresh knowledge diffuses widely or stalls at the frontier[6]. During the dot-com era, for example, firms invested billions in capacity that far exceeded actual internet commerce, a misalignment the Federal Reserve later described as "innovation hype overshooting real consumer needs"[50].

Shiller interprets such misfires as the predictable outcome of analysing innovation without measuring demand-side sentiment. He therefore urges economists to marry tools from psychology, sociology and data science—text mining of news, social media and search trends—to trace how stories about new technologies spread and to quantify their influence on adoption curves, capital flows and productivity. Until innovation research integrates these demand-focused metrics, it will continue to explain only half of the growth puzzle.

The discussion to this point has been deliberately supply-centred: how scientific knowledge, R&D, talent and institutions "push" new technologies into the world. In the pages that follow, the manuscript pivots first to the complementary "pull" of demand—how market signals, user needs and public narratives shape inventive effort—and then to theories that integrate both vectors into a unified, systems view of innovation. Tracing this arc from supply, to demand, to their interaction sets the conceptual foundation for the three agent-based simulations that close the thesis.

The modern theoretical fundation of innovation demand starts with Jacob Schmookler famously argued that market demand is a crucial catalyst – innovation responds to the "pull" of needs and opportunities[4]. Schmookler's studies of patent patterns provided the first evidence that inventive activity is responsive to demand: the larger or more promising a market, the more inventive effort it attracts. In his patent studies, industries with faster-growing sales generated markedly more patents, even after controlling for scientific opportunity[4]. Schmookler's results sharpened the aphorism "necessity is the mother of invention" and inspired a generation of demand-side studies. A complementary formalisation is Hicks' (1932) induced-innovation hypothesis: when a production factor (e.g., energy, labour) becomes dear, profit-seeking inventors redirect R&D toward economising that factor[5]. Contemporary scholarship stresses interaction, not opposition, between push and pull: the smartphone boom, for example, required breakthroughs in microelectronics and a mass market eager for ubiquitous connectivity[89]. Modern evolutionary and systems views thus treat vibrant demand and scientific capability as joint determinants of inventive direction[90]. When resources grow scarce or prices spike, inventors redirect effort toward factor-saving technologies[5][43]. The 1970s oil shocks triggered a wave of energy-efficient patents[22], and wartime shortages have repeatedly accelerated technological change[23]. A recent illustration is COVID-19: urgent global demand collapsed normal vaccine timelines from a decade to under a year[44]. Post-war resource scarcity inside Toyota likewise spawned the "lean" production system[91]. Carbon pricing shows the same inducement logic today—higher expected costs of emissions raise clean-tech patenting[92]. Scarcity alone, however, does not guarantee breakthroughs; without an adequate knowledge base or complementary assets, problem-solving stalls[45].

Innovation is also embedded in social structure. Individuals who bridge "structural holes"-gaps between otherwise disconnected groups-generate more novel ideas because they can recombine distant knowledge[93]. Large-scale bibliometrics show that international, interdisciplinary teams outperform parochial ones on citation impact[94], while overly tight, homogeneous circles recycle existing ideas[10]. "Science-of-science" work finds that small, flat teams disrupt knowledge frontiers, whereas large, hierarchical

teams excel at incremental refinement[95][96]. Network effects also govern diffusion: adoption cascades when peers have already embraced a technology[46][47].

Incentives and rewards are also important for the demand perspective of innovation. On the market side, firms invest in R&D when expected profits outweigh costs[41]; hence pharmaceutical innovation skews toward diseases prevalent in high-income markets[97]. Patents raise that expected payoff by granting temporary monopoly rights[98]. In science, the motive is often priority and prestige—being first confers lasting reputational rewards[99]. Public funders reinforce the demand influence on innovation: U.S. NIH grant rules, for instance, crowd-in private patents in targeted disease areas[100]. Inducement prizes such as the X-Prizes leverage the same logic, spurring ambitious private R&D[101]. Yet mis-aligned incentives can bias researchers toward safe, incremental projects[102]. Balanced schemes therefore reward both high-risk "moon-shots" and reliable advances.

Innovation trajectories are sticky: early technical or social advantages can lock in inferior designs such as the QWERTY keyboard[48][49]. Scientific paradigms also become entrenched, attracting disproportionate talent and funding[103][104]. Strong ties to superstar scientists yield prolific but less novel work[105], and dominant firms often favour sustaining innovations, leaving disruptive breakthroughs to newcomers[106]. Conversely, mission-oriented public policy can redirect the path of change by creating large, guaranteed demand (e.g., state procurement for clean energy)[107].

3 Method

The preceding supply- and demand-side reviews clarify why innovation flourishes or falters. To test these theoretical claims empirically, we next turn to computational social simulation-a method that allows us to observe how micro-level incentives, cognition, and social structure interact to produce macro-level innovation outcomes. The section that follows details our agent-based simulation framework and four increasingly rich experiments that leverage recent advances in large-language-model agents.

Social simulation has been a powerful method for studying complex human systems since the 1960s. Early foundational work by Schelling (1971) demonstrated how simple individual rules could produce large-scale, unexpected social patterns such as residential segregation. As computing technology advanced, social simulations became increasingly sophisticated, covering domains ranging from traffic flows to financial markets and epidemiology[14][15]. These simulations offered controlled environments to explore complex "what-if" scenarios that traditional analytical methods could not easily handle. They highlighted how repetitive individual actions can collectively result in unforeseen group behaviors[108][109], revealing emergent phenomena that might be impossible to deduce from top-down equations alone.

Agent-based modeling (ABM) formalizes this approach by representing individuals as autonomous software entities or "agents." Unlike top-down aggregate models, ABM captures the emergence of higher-level patterns directly from micro-level interactions among agents[14]. Economists, sociologists, political scientists, and anthropologists have all used ABM to investigate diverse phenomena, including market dynamics,

social norm formation, voter behavior, and cultural evolution[15]. Each agent in an ABM can have its own rules and attributes, allowing researchers to examine phenomena like adaptation, diversity, and path dependence in silico. By adjusting initial conditions or agent rules, one can experiment with different scenarios and observe how complex social outcomes unfold over time.

The advent of large language models (LLMs) has significantly enhanced the realism and flexibility of ABM in recent years. Traditional rule-based ABMs required explicit programming of every potential agent behavior, which limited their ability to capture nuance and adaptivity. In contrast, LLM-powered agents can interpret context and generate behavior dynamically through natural language prompts. This results in more nuanced and adaptive agent behaviors that are not hard-coded but emerge from the agent's understanding of its environment and goals[110]. For example, Park et al.[16] demonstrated that a community of 25 LLM-driven agents in a virtual town could spontaneously form daily routines, relationships, and even coordinate activities like planning a party without any explicit script. Such studies suggest that agents driven by LLMs can exhibit human-like social behaviors and creativity. Further research has indicated that these LLM-based agents can display advanced cognitive capabilities such as theory-of-mind reasoning (at least in a rudimentary form), negotiation strategies, and even forms of cultural learning, albeit with some variability and occasional failures[110]. These benefits come with challenges related to computational demands and reproducibility, due to the stochastic nature of language model outputs and the difficulty of controlling exactly what an LLM will do in every situation. Nonetheless, the integration of LLMs into ABM represents a promising frontier for creating more realistic social simulations.

All four simulations in this study share a common architectural philosophy that combines agent-based modeling with LLM-powered autonomous agents (See Technical Diagrams in Appendix). In each simulation, a population of agents interacts within a defined environment, and each agent's decision-making is handled by prompting an LLM (specifically, GPT-based models) with a carefully constructed prompt representing the agent's context. This architecture prioritizes experimental control: we keep many parameters configurable and use reproducible random seeds so that runs can be repeated under the same conditions. It also emphasizes scalability and efficiency, using techniques like caching of LLM responses, batching multiple agent requests together, and asynchronous processing to handle many agents in parallel. The software design is highly modular, clearly separating different concerns – the agent behavior logic, the social network structure that connects agents, the evaluation mechanisms for ideas or projects, and the performance metrics we collect – making it easier to adjust one component without affecting others. This modular, controlled approach ensures that each simulation can be run consistently and scaled up to larger agent counts or longer durations without losing reproducibility.

3.1 Preliminary Simulation: 1800s Survival-Scarcity

In the preliminary simulation, We simulate 15 autonomous agents (GPT-o3 reasoning LLM) representing members of a small virtual community, each with unique roles (farmer, blacksmith, mayor, etc.), skills (farming, fishing, construction, medicine),

knowledge bases (e.g., which berries are edible, how to plant crops), and personalities (affecting risk-taking, cooperation, and conflict). This diversity creates interdependence-no single agent can meet all its needs alone-driving trade, negotiation, and collaboration. Each agent has a memory architecture capturing past interactions, lessons learned, and key events (e.g., successful trade deals or trustworthy partners). To manage and retrieve these memories effectively, we use a Retrieval-Augmented Generation (RAG) framework, which stores agents' historical data outside their immediate context window. When needed, an agent "retrieves" the relevant memory segments and "augments" its generative process, allowing it to make decisions or hold conversations grounded in past experiences. A memory cut-off design prevents agents from bogging down with excessive context—only the most relevant memories are retrieved at any point, keeping decisions tractable yet informed. All agents share a fundamental goal of survival (avoiding starvation by securing enough food), but they also have individual secondary goals. Some goals are personal (such as improving one's living conditions or accumulating wealth), and others are communal (such as maintaining social order or helping neighbors). Notably, an important goal for many agents in this study is innovation - finding new and better ways to meet their needs. However, the priority of innovation may vary: an agent near starvation will prioritize food acquisition (necessity) over long-term experimentation, whereas a comfortable agent might pursue improvements or novel ideas. The interplay of goals means agents constantly balance short-term survival against long-term progress.

In addition to the 15 primary agents, the simulation features a special agent acting as a "Game Master" (GM) or overseer. The GM agent does not represent a person in the community but rather serves as an environment moderator and experiment facilitator. The GM has the capability to modify or introduce social norms and environmental rules during the simulation. For example, the GM might impose a new trading rule (such as taxation or resource sharing norms) or adjust what behaviors are considered acceptable in the society. The GM can create new objects and locations that did not exist in the simulation initially. This helps agents to change the simulation settings and apply their innovations. GM will review individual agent's request on changing the simulation (e.g. creating a new tool to be more effective at fishing), then judge whether it is a feasible event to happen in the simulation. If it is a reasonable action, the GM will create this event, object, or location in the simulation, update the general norm, and inform that agent. The GM continuously observes the state of the world and the agents. It keeps an event log of key occurrences each day, such as trades executed, conflicts, instances of resource sharing, or innovative ideas proposed by agents. Crucially, the GM monitors conversations and actions to identify potential innovations - e.g., when an agent invents a new tool or proposes a novel solution to a problem, the GM records this event (including its time and a qualitative description). Because the agents communicate (through simulated dialogue or other interactions), the GM's observation includes parsing these communications for evidence of creative proposals or plans.

Within this virtual world, the primary resource is food, essential for each agent's daily survival. At the start of a simulation run, every agent is endowed with some amount of food, measured in abstract units. Because agents must consume a fixed daily

allotment to stay alive, the availability or scarcity of food exerts constant pressure on their decisions. If an agent is unable to secure enough food for five consecutive days, it "starves" and is removed from the simulation, simulating a hard survival threshold that compels creative or cooperative solutions. Although food is the main resource of concern, there are also other goods such as money, raw materials, or tools, which agents can barter or produce. Only two agents have primary roles that generate new food (e.g., the farmer, the fisher), making the rest dependent on trade or collaboration to meet their nutritional needs. Agents can exchange these secondary resources for food or vice versa, and may create items that improve future productivity—like better farming tools—to reduce the risk of starvation. However, production is intentionally uncertain or limited, ensuring that the community cannot rely solely on static routines; trade and innovation become pivotal for group survival. By observing how agents respond to resource shortfalls—whether they form alliances, innovate, or engage in conflict—we gain insights into the conditions that spark or hinder inventive behavior. Each simulation day is divided into five periods-Early Morning, Morning, Noon, Afternoon, and Evening-to organize the agents' activities. During Early Morning, agents might review yesterday's events, consult their memories for lessons learned, and plan for the day. In the Morning, those with production roles (e.g., farming, fishing) can work on generating resources, while others might explore or rest. At Noon, agents typically gather for a "town square" interaction period, where they can trade goods, share information, discuss problems, or propose ideas. The Afternoon is another block of productive time, either continuing resource collection or developing new methods. Finally, in the Evening, agents evaluate the day's results, potentially brainstorming innovations or forging alliances for the next day. This structured daily timeline maintains consistency across simulation runs, ensuring that each agent has similar windows for work, trade, and collaboration. It also helps researchers trace the sequence of decisions and outcomes: if an innovation arises in the Afternoon, it is easier to see whether a morning conversation influenced it. The timeline thus fosters a balance of focused effort and reflection, aligning with general findings that alternating between action and deliberation can spur creativity. By controlling these interaction windows, we can compare how agents respond under various resource conditions, focusing on which times of day yield the highest rates of innovation or negotiation success.

We investigate how different resource scenarios affect innovative activity through two sets of experimental treatments. In both sets, we run multiple simulation trials under each condition, allowing us to compare outcomes across various community setups. In the first set, we vary the level of initial resource pressure. In the High Pressure condition, each agent starts with minimal food, just enough for about two days. This severe scarcity prompts immediate survival-focused actions, potentially leading to necessity-driven innovations. By contrast, in the Lower Pressure (Baseline) condition, agents begin with a moderate buffer of food—enough for roughly three days. While agents still need to replenish supplies eventually, they have more breathing room, making urgent survival less of a factor and allowing more opportunistic or long-range innovation strategies. In the second set, we manipulate resource allocation equality. In the Equal Distribution condition, every agent starts with a similar amount of food, ensuring a level playing field from day one. Researchers can observe whether a

lack of inequality fosters cooperative innovation or, alternatively, diminishes creative urgency. Meanwhile, the Unequal Distribution condition gives some agents sizeable surpluses while others begin nearly destitute, creating a wealth gap. This setup tests whether acute need among the disadvantaged drives a surge in inventive problem-solving—or whether social friction from inequality suppresses collective innovation. Across both experimental sets, we compare how these resource environments shape the frequency, originality, and success of innovations, providing deeper causal insights into the relationship between survival constraints and creative breakthroughs.

The hypothesis of this preliminary simulation is that lower initial food levels cause more innovation in the community. In other words, agents in the High Pressure condition (acute initial scarcity) will, on average, generate a greater number of innovative ideas and solutions than those in the more comfortable condition. This hypothesis is inspired by the age-old adage that "necessity is the mother of invention." The reasoning is that when agents are under severe threat of starvation, they face strong incentives to think creatively and quickly to secure their survival. Every day counts, so they are more likely to experiment with new tactics (e.g., invent a tool to gather food faster, explore previously ignored food sources, or devise new collaboration schemes to share resources). By contrast, if agents have plenty of food initially, they might not feel the urgent need to innovate; they can survive with the status quo for longer, possibly leading to complacency or at least a focus on routine over innovation. Therefore, we expect to see a higher count of innovations, earlier innovation timing, and higher innovation quality in the simulations where initial resources are scarce. The second hypothesis posits that more equal initial resource allocation leads to more innovation in the community. Under this hypothesis, the simulations starting with equal distribution of food among agents will show greater innovative output compared to those with high inequality. The theoretical motivation here is that in an equal setup, no subset of agents is disproportionately disadvantaged; everyone has a decent starting chance to contribute. This could foster collaboration and trust – agents might be more willing to work together on inventive solutions if they perceive the environment as fair. Additionally, with equal shares, extreme resentment or conflict (which can accompany inequality) is minimized, potentially allowing the community to focus energy on creative improvements rather than infighting. We also thought that when resources are very unequally distributed, the poor agents are indeed motivated to innovate (similar to Hypothesis 1's logic of necessity), but the wealthy agents have little incentive to innovate (they are comfortable and may even resist change that could upset their advantage). The net effect, we predicted, might be fewer community-wide innovations in the unequal scenario, since collaboration breaks down and only a subset of agents (the poor ones) try to innovate, possibly with limited success.

The hypotheses above are formulated as causal claims: changing X (initial food level, initial equality) will cause a change in Y (innovation outcomes), holding other factors constant. The independent variable for H1 is the initial average food per agent (a numeric measure of resource abundance) or equivalently a binary indicator of being in the High Pressure vs Low Pressure condition. The independent variable for H2 is the initial inequality level (measured by an index or the Equal vs Unequal condition indicator). The dependent variables are metrics of innovation (primarily the count of

innovations, and secondarily innovation quality and timing). We are interested in both the quantity of innovation (did more innovations occur?) and the quality or significance of those innovations under each condition. The hypotheses as stated focus on quantity (e.g., "more innovation"), but we will also examine quality differences in the results.

To test the causal effect of our treatments on innovation outcomes, we employ a randomization inference approach. Randomization inference is well-suited for our experimental design and offers several advantages: Unlike large-sample asymptotic tests (t-tests or regression coefficients assuming normal errors), randomization inference makes no distributional assumptions beyond the random assignment. It thus can provide an exact significance test even with a relatively small number of observations (which is relevant since the number of simulation runs, while larger than typical field experiments, is still moderate). By enumerating or resampling possible random assignments, we obtain a p-value as the proportion of simulated assignments that produce an outcome difference as extreme as what we observed. In other words, we are asking: if the initial condition truly didn't matter, how often would chance produce a difference in innovation count this big or bigger between treatment and control? This method gives a robust answer. We chose randomization inference over a classical t-test or OLS regression for a few reasons. First, the distribution of our outcome (innovation count) might not be normal - counts are often skewed or have low ranges (including possibly zero-inflation if some runs yield no innovation). Randomization tests handle this gracefully since they don't rely on normality or large sample theory. Second, our sample of runs might be on the smaller side for reliable asymptotic approximations; randomization inference in contrast remains valid at small sample sizes, controlling Type I error by design. Third, randomization inference reinforces the interpretation of the difference in means as an estimate of the causal effect under the randomization distribution. By using the actual random assignment that we implemented, we condition on the exact experimental design.

3.2 Simulation 1: Network-Topology-Scientific Innovation

After the preliminary "hunger game" simulation of LLM agents, We decide to be more focused on scientific innovation and conduct a series of three simulations, each built on the previous ones. The first simulation specifically investigates how network topology affects group innovation. The research question driving this simulation is: How does the pattern of communication links between individuals influence the diversity and quality of ideas they produce collectively? The hypothesis is that the structure of the social network (an exogenous factor we manipulate) will significantly influence innovation outcomes. Certain network structures might encourage a healthy balance of diverse ideas and information sharing, while others might cause premature convergence or fragmentation. For example, prior studies have suggested that networks with some "small-world" characteristics (short average path lengths combined with clustered communities) can foster creativity by balancing novelty and collaboration. In this simulation, the independent variable is the network topology connecting the agents, and the main dependent (endogenous) variables are measures of collective innovation performance – including the average quality of ideas generated, the diversity of ideas

(how varied or novel they are relative to each other), and the speed at which the group converges toward high-quality ideas.

Fifty autonomous agents work collaboratively on a complex problem in this first simulation. To simulate cognitive diversity, the agents are assigned distinct persona types that influence their behavior and perspective. There are ten persona archetypes (with roughly five agents each): engineers, artists, scientists, designers, entrepreneurs, environmentalists, mathematicians, philosophers, architects, and biologists. Each persona comes with its own cognitive style and priorities. For instance, an engineer agent tends to be risk-averse and focuses on technical feasibility and optimization, whereas an artist agent is more risk-tolerant and prioritizes creativity and user experience. The scientist persona emphasizes hypothesis testing and evidence, the entrepreneur seeks market viability, the environmentalist focuses on sustainability, and so on.

These personas are implemented by modifying the prompts given to an agent's LLM: the prompt includes guidance reflecting that agent's viewpoint (e.g., the engineer's prompt might say "Consider practical implementation details and robustness," while the artist's prompt might say "Be imaginative and consider aesthetic appeal"). Specifically, the initial idea generation prompt for each agent follows this structure:

You are a $\{self.persona\}$.
Problem: $\{problem\}$
Generate an initial solution idea that reflects your unique perspective and expertise. Be specific and concrete. Limit your response to 2-3 sentences.

During ongoing rounds, agents receive more contextual information in their prompts:

You are a $\{self.persona\}$.
Problem: $\{problem\}$
$\{context\}$
Strategy: $\{strategy\}$
Generate a new solution idea. Be specific and concrete. Limit to 2-3 sentences.

Where the context includes elements such as "Your current best idea: [idea description]", "Recent ideas from collaborators: [list of neighbor ideas]", and "Successful patterns you've observed: [high-scoring approaches]". The strategy parameter directs the agent to either "Explore a completely new approach" or "Improve upon the best existing ideas", depending on their current exploration-exploitation decision. All agent prompts use a consistent system message: {"role": "system", "content": "You are a creative problem solver."} to establish the baseline behavior. This way, the LLM-generated ideas differ based on the agent's role, giving each agent a distinct voice and strategy. We also set different risk preferences for each persona type – quantitatively, some personas have a lower exploration tendency (more conservative, focusing on refining existing ideas) while others have a higher tendency to explore novel ideas. This ensures that at any given time, some agents are pushing the boundaries with wild ideas while others are critically evaluating and improving ideas, mirroring the mix of exploration vs. exploitation often seen in real teams.

Each agent in the first simulation has a limited memory capacity for ideas. An agent can hold at most 10–20 ideas in mind (in practice we use a fixed size like 10 for the idea queue). When new ideas are generated or learned from peers, older ideas may

be dropped if the memory limit is exceeded, simulating cognitive constraints. Agents keep track of the best ideas they have seen so far (e.g. top scoring ideas) and can use that to inform their decisions.

The simulation proceeds in synchronous rounds. At the beginning, in an initialization phase, each agent generates a few (e.g., three) initial ideas for the problem on their own, and these ideas are given an initial quality score. Then the main loop of rounds begins. In each round, every agent generates a new idea or chooses one of the ideas in its memory to modify, using the LLM (the agent's prompt includes the agent's persona and possibly some recent ideas). After idea generation, agents share their ideas with their neighbors in the network. Unlike a fully public forum, each agent only communicates directly with the agents to whom it is connected in the social network. Neighbors receive the idea and evaluate it from their own perspective. Based on this evaluation, they may adopt the idea into their memory if it seems promising. The adoption is not guaranteed; it can depend on factors like the idea's quality score, whether it aligns with the agent's current strategy or expertise, and a bit of randomness to reflect individual open-mindedness. We also include a simple mechanism to preserve diversity: if an agent finds that all its neighbors are converging on the same idea, it becomes more inclined to try something different. Similarly, if an agent hasn't seen any improvement in its best idea for several rounds, it will gradually increase its exploration tendency – for example, by using a higher "temperature" setting for the LLM to encourage more randomness in idea generation. This prevents the whole group from getting stuck in a local optimum (a phenomenon akin to "mode collapse," where everyone ends up repeating similar ideas).

We experiment with five different network topologies in the first simulation: (1) fully connected, (2) ring lattice, (3) small-world, (4) scale-free, and (5) hierarchical modular networks. These represent a range of possible communication structures. In the fully connected network, every agent is linked to every other agent, so information spreads freely to all members in one hop. This scenario provides a baseline of maximum connectivity. At the other extreme, the ring lattice connects each agent only to a few local neighbors in a ring structure (each agent might talk only to its two immediate neighbors on each side, for example). This limits information flow to local circles and requires many hops for an idea to reach the opposite side of the network, possibly maintaining higher diversity but slowing down convergence. The small-world network starts from a ring lattice and then randomly re-wires a small fraction of links to create a few "shortcuts" across the network. This yields high local clustering (like the ring) but also short average path lengths thanks to the random long-range connections. Small-world structures are known to often facilitate rapid spread of information while preserving some local diversity, and have been associated with creative social systems. The scale-free network is built using a preferential attachment mechanism so that a few hub agents emerge with very high degree (many connections) while most others have few connections. This topology can model situations where a few individuals (hubs) broadcast ideas to many others, perhaps representing influential leaders or central communicators in the group. Finally, the hierarchical modular network consists of clusters of agents (modules) that are densely connected internally, with only sparse connections between clusters. This reflects an organization where sub-teams work

closely among themselves but only occasionally share ideas with other sub-teams, akin to a research institute with distinct departments or a company with semi-independent divisions. For each network type, we keep the total number of agents (50) constant and ensure average degree is in a comparable range, so that differences in outcomes can be attributed to the structure of connections rather than trivial differences in number of links.

Within each round of the simulation, after sharing and adoption, all ideas (new and existing) that agents hold are evaluated for quality. The simulation uses a scoring function to rate idea quality on a numeric scale (for example, 0 to 100). This evaluation could be done via another LLM prompt or a heuristic function, but in our setup we used a simplified heuristic or a "mock" evaluation function for controlled experiments, so that we know the true best solution and can measure how close agents are getting. Using a mock or simpler model for evaluation ensures we can isolate the effect of the network and agent interactions without too much noise. After scoring, agents decide whether to keep pursuing their current best idea or switch to a new idea based on what they learned from neighbors. This completes one round. The process then repeats: agents generate or refine ideas (possibly influenced by newly adopted ideas from neighbors), share again, and so on. We typically run this simulation for multiple rounds (e.g., 20 rounds) or until convergence criteria are met. We define convergence in several possible ways: if the group's best idea score has plateaued (no significant improvement for, say, 3-5 rounds), or if a large majority (e.g., 80%) of agents have adopted the exact same idea (indicating the group has converged to one solution), or if the diversity of ideas collapses below a threshold (meaning most ideas are minor variations of each other). If any of these conditions occur, the simulation can stop early; otherwise it runs up to a fixed maximum number of rounds.

We measure a number of outcome metrics in the first simulation to evaluate the effect of network topology on innovation. Idea quality is tracked over time - for instance, we look at the highest idea score achieved by the group and the average quality of ideas in each round. Idea diversity is crucial for innovation, so we quantify diversity using a self-BLEU metric (comparing the textual content of ideas across agents to see how similar they are – a lower overlap means higher diversity)[111] as well as entropy-based measures over idea features. We also count how many distinct solutions emerge versus everyone converging on one. Convergence speed is measured by how many rounds it takes to reach a stable state or a high-quality idea. Additionally, we record network dynamics such as how far ideas spread (do they stay confined to local clusters in a modular network, or do they rapidly reach everyone?), and we compute network metrics like clustering coefficient or average path length of each topology to correlate with the innovation outcomes. By comparing these results across the five topologies, we can see, for example, if fully connected teams reach high-quality ideas faster but with less diversity, or if the small-world topology indeed offers a sweet spot of fast dissemination and sustained diversity as hypothesized. This experiment provides insight into how the structure of communication (an exogenous structural factor) can influence collective creativity (an emergent outcome), informing theories of innovation networks.

3.3 Simulation 2: Network-Topology with Enhanced Agents

The second simulation builds upon the first, using a similar collaborative problem-solving scenario but with an enhanced agent architecture and an asynchronous communication framework. The primary research question for this simulation is: Can enriching the cognitive complexity of agents and allowing asynchronous, more continuous interactions improve the innovation process and outcomes compared to the simpler synchronous model? In other words, we investigate whether adding more human-like features to the agents (memory, personality, background knowledge) and improving the way they communicate (not all at once in lockstep rounds, but in a more staggered, asynchronous manner) leads to more creative ideas, avoids convergence to mediocre solutions, and runs more efficiently. The hypothesis is that these enhancements will indeed yield richer dynamics: agents with detailed backgrounds and memory will generate more varied ideas, and an asynchronous setup will prevent everyone from "locking on" to the same cycle of idea generation, thereby reducing the risk of mode collapse (where agents start producing very similar outputs). The key exogenous design changes here are the introduction of deeper agent autobiographical detail and improved communication protocols, while the endogenous outcomes we examine include the creativity of ideas (measured in multiple dimensions), the system's ability to continue generating novel ideas without collapsing into repetition, and the computational performance (speedup) gained from the new architecture.

In the second simulation, we again have on the order of 50 agents working on an innovation task, but these agents are modeled as junior and senior researchers rather than simple abstract personas. We provide each agent with a rich biographical background generated at initialization. This includes details like their field of expertise (for example, some agents specialize in machine learning, others in biology, etc., representing a multidisciplinary team), their career stage (junior researchers with maybe 1–5 years of experience, mid-career with 10 years, and seniors with decades of experience), and even personal traits (some agents are more analytical while others are more intuitive or creative). These details are fed into the agent's prompt to influence how they approach problems - e.g., a junior machine learning researcher with an intuitive style might propose bold, data-driven ideas, whereas a senior physicist with an analytical style might suggest more conservative, theory-grounded ideas. By constructing agents this way, we introduce diversity not just in their roles but also in their experience and thinking styles.

When these enhanced agents need to make risk decisions, they receive detailed prompts that incorporate their full context. For instance, under a short-term "publish or perish" incentive structure:

You are a researcher in a "publish or perish" academic system.
Your situation:
- Recent success rate: $\{success\_rate : .1\%\}$
- Recent failures: $\{recent\_failures\}$
- Career points: $\{agent.career\_points : .1f\}$
- Risk tolerance (personality): $\{agent.risk\_tolerance : .2f\}$
System rules:
- You must publish at least 1 paper per year to maintain funding

- The bottom 10% of performers lose their positions annually
- Each publication gives 1 point, breakthroughs give 3 points
You must choose the risk level for your next project (0.0 to 1.0):
- Low risk (0.0-0.3): 80-90% success rate, low impact
- Medium risk (0.3-0.7): 50-70% success rate, moderate impact
- High risk (0.7-1.0): 20-40% success rate, potential breakthrough
Given your situation and the pressure to publish consistently, what risk level do you choose?
Respond with a JSON object: {"$risk\_level$" : 0.X, "$reasoning$" : "$brief explanation$"}

Under long-term funding (HHMI-style), the same agent would receive a different prompt emphasizing breakthrough potential:

You are a researcher with long-term funding for breakthrough research.
Your situation: - Years remaining in funding cycle: {$years\_remaining$}
- Breakthroughs achieved so far: {$breakthroughs$}
- Risk tolerance (personality): {$agent.risk\_tolerance : .2f$}
System rules:
- You have guaranteed funding for 5 years
- Only breakthroughs matter - incremental work has no value
- You need at least 1 breakthrough per 5-year cycle
- No penalties for failed attempts
You must choose the risk level for your next project (0.0 to 1.0):
- Low risk: Unlikely to produce breakthroughs
- Medium risk: Some breakthrough potential
- High risk: Maximum breakthrough potential
Given your secure funding and breakthrough-only rewards, what risk level do you choose?
Respond with a JSON object: {"$risk_level$" : 0.X, "$reasoning$" : "$brief explanation$"}

Each agent's profile also includes a risk tolerance parameter influenced by both their personal disposition and career stage. For instance, some junior members might be very ambitious and take higher risks to make a name (or conversely, some might be risk-averse if the environment is competitive), while some senior members might play it safe to protect their reputation (unless they have secure funding, in which case they might take more risks). These differences lead to heterogeneity in decision-making: at any given time, some agents will choose safe projects while others choose bold ones.

The communication model in this second simulation is asynchronous and continuous rather than strictly round-based. Instead of all agents synchronously generating and exchanging ideas in fixed rounds, agents here operate on a more event-driven timeline. Each agent still iteratively works on proposing or refining ideas, but not all agents act in lockstep. Agents can act whenever they are ready (with a scheduling mechanism ensuring each gets opportunities over time), and sharing of ideas happens immediately over the network rather than waiting for a synchronized step. The network topologies in this simulation can also evolve dynamically: whereas in the first simulation the network structure was fixed for an entire run, here we allow the communication network to adapt based on interactions. If two agents interact frequently and find each other's ideas useful, the "strength" or effective frequency of their connection might increase (simulating an evolving collaboration link). Conversely, if two agents rarely exchange

information, their connection might weaken. This dynamic network feature models how real collaborations form and dissolve over time, and it allows us to see emergent cluster structures or central hubs developing as the simulation progresses.

To prevent mode collapse and promote sustained creativity, we implement several strategies. Each agent has a hierarchical memory structure: they maintain a log of observations (things they have seen or ideas received from others), a set of reflections (their own notes or interpretations about those observations), and a plan or to-do list for future actions. When an agent's LLM is prompted to generate a new idea or make a decision, the prompt may include a summary of recent observations and key reflections, so that the agent can "remember" context from earlier in the simulation and not repeat itself. We also vary the LLM's generation parameters dynamically. Specifically, we adjust the temperature of the LLM between about 0.5 and 1.0 depending on the situation: if an agent has been stuck in a rut (producing similar ideas repeatedly or not contributing new information), we raise the temperature closer to 1.0 to inject more randomness and creativity; if the agent is in a phase of consolidating or evaluating an idea, we might use a lower temperature for more focus and consistency. This dynamic tuning encourages greater variety in the ideas proposed across the simulation. Additionally, each agent is given a unique initial prompt nuance or random seed so that their outputs are decorrelated – this ensures that not all agents converge to the same style or content just because they use the same language model.

The evaluation of innovation also becomes more sophisticated, using prompts like:

Evaluate the following research project on a scale of 0-1 for each dimension:
Project: $\{project.title\}$
Description: $\{project.description\}$
Risk Level: $\{project.risk\_level : .2f\}$
Keywords: $\{', '.join(project.keywords)\}$
Please rate on:
1. Relevance (0-1): How well does it address important problems?
2. Novelty (0-1): How new/innovative is the approach?
3. Impact (0-1): What is the potential impact on the field?
4. Elegance (0-1): How simple and effective is the solution?
5. Generative Potential (0-1): Will this lead to future innovations?
Respond in JSON format: $\{"relevance" : 0.X, "novelty" : 0.X, ...\}$

Performance optimization is a major focus in the second simulation's implementation. Because giving each agent a large language model query can be slow, we incorporate a caching mechanism and batching of requests to speed things up. The caching works at multiple levels: if an agent asks the LLM something identical to a previous query, we reuse the past answer (exact-match cache). If the query is not identical but semantically very similar to a past query, we can reuse an older response with slight adjustments (semantic cache). We also break prompts into components and cache the results of sub-parts that might be reused by many agents (component-level cache). For instance, a portion of the prompt that describes the common problem or provides standard instructions need not be regenerated each time. In addition, we batch multiple LLM calls together: instead of sending 50 separate queries sequentially, we might send them in batches of 5 or 10 in parallel, depending on the system, which effectively utilizes the LLM's ability to handle multiple prompts in one go.

Combined with the asynchronous scheduling (agents don't all wait for the slowest agent), these optimizations led to roughly a 3.7× faster execution compared to the naive synchronous approach of the first simulation. This means we can run more cycles or include more agents without the simulation becoming intractably slow, which is important for scaling up experiments.

The evaluation of innovation in the second simulation is more fine-grained than in the first. Instead of a single quality score, we assess creativity along multiple dimensions. We use a multi-dimensional creativity rubric inspired by design theory, covering aspects of originality, usefulness, elegance, and foundational impact. In practice, each idea generated by the agents can be given a rating on these dimensions (for example, how novel or original it is, how useful or applicable it would be, how elegant or well-designed the solution is, and whether it has a broad, foundational significance or opens new paradigms). We might also consider feasibility as a factor, but since this is a thought experiment, feasibility is often less emphasized. By capturing these different aspects, we get a richer picture of each idea's value. We then combine these aspects into an overall creativity score or use them as a vector of metrics. For instance, an idea that scores highly in originality and impact but low in elegance might indicate a groundbreaking but rough concept. We pay special attention to whether the diversity of ideas is maintained over time - using similar diversity metrics as in Simulation 1 (textual diversity measures, number of unique concepts introduced, etc.) – and whether the agents manage to avoid converging on safe but uninspiring solutions. We also monitor the adaptive network: we examine if the network becomes more centralized (e.g., one agent becoming a hub of idea sharing) or remains decentralized, and how that correlates with innovation outcomes. The results from this simulation demonstrate a richer ecosystem of ideas: qualitatively, we observed agents developing more complex ideas and building on each other's suggestions in a more free-form, asynchronous way, which can be contrasted with the round-by-round synchronous progress of the first simulation. The improved performance and ability to scale in Simulation 2 also allowed us to explore longer runs and more complex scenarios than would have been feasible with the initial setup.

3.4 Simulation 3: Academic Career-Incentive

The third simulation is the most complex and ambitious, modeling a full academic research ecosystem over an extended period. Here, the focus is on how different incentive structures and career dynamics influence innovation at the macro level. The research question is: How do institutional incentives (e.g., short-term publish-or-perish pressure vs. long-term stable funding) affect researchers' behavior and the trajectory of scientific innovation? We also examine how social processes like collaboration networks, reputation effects (the "Matthew effect"), and resource competition shape the outcomes. The hypothesis is that the structure of incentives will create distinct patterns of innovation: for instance, a short-term publication pressure environment might yield many incremental papers but fewer breakthrough innovations, whereas a long-term, high-risk-tolerant environment might produce fewer total papers but more paradigm-shifting discoveries. A balanced incentive system might strike a middle ground, achieving moderate output with a decent number of high-impact results.

In this simulation, the exogenous variable is the incentive regime – we simulate three scenarios: one mimicking a "publish or perish" system with frequent evaluations and strong pressure to publish regularly; one mimicking a long-term fellowship or grant system (like certain scholarly programs) where researchers have secure funding for a period (e.g., five years) and are evaluated primarily on long-term impact; and a hybrid system that mixes elements of both. The endogenous outcomes include a range of innovation metrics: the number of publications produced, the frequency of major breakthroughs (very high-impact results), the career stability or dropout rate of researchers, the distribution of success (is it concentrated among a few star scientists or spread more evenly), and measures of overall knowledge advancement (such as how diverse the research topics are and whether new research areas emerge).

In this academic simulation, each agent represents a scientist (researcher) with a detailed synthetic biography and career profile. We create 50 such agents to populate a small research community. Each researcher agent is assigned attributes such as a name, an age or career stage, a primary field of research, and a track record of past achievements. For example, one agent might be "Dr. Elena Rodriguez," a mid-career computer scientist specializing in machine learning with an h-index of 18 and 34 publications, who tends to collaborate often and has a moderate risk tolerance. Another might be "Dr. Marcus Chen," a senior physicist with a very high reputation (h-index over 50) but a conservative research style. Key attributes for each agent include their h-index (a metric indicating the number of papers they have that have at least H citations, used as a proxy for cumulative impact), total citations, publication count, and one or two primary expertise areas (we define about ten major research areas, each with a few sub-fields, and assign each researcher a specialization in these). This creates overlapping communities of expertise - some agents share fields which makes collaboration easier among them, while others are in distinct domains. Agents also have a list of current collaborators (an initial co-authorship network), which we seed based on field overlap or prior "history" in their synthetic biographies. Each agent has a certain risk tolerance and exploration tendency that affect how they choose projects. For instance, some agents might naturally favor safe projects with guaranteed results, while others are "mavericks" who frequently try novel ideas. These traits can correlate with career stage (e.g., perhaps a slight tendency for younger researchers to be more risk-seeking in our initialization, unless the environment strongly disincentivizes risk).

The simulation of the scientific process is broken into discrete time steps which we can think of as annual cycles (each cycle representing, say, one year of academic time). We simulate multiple cycles (e.g., 20 years) to observe long-term effects on careers and innovation. In each cycle, every researcher decides on a research project to pursue that year. This decision is influenced by the current incentive regime and the researcher's personal situation. The decision-making process uses sophisticated prompts that vary by incentive structure. Under exploitation-focused (publish or perish) incentives:

> You are {*agent_name*}, a {*agent_career_stage*} researcher in {*research_area*}.
> Your department values consistent publication output. Last cycle, you published {*last_pubs*} papers.
> Your funding depends on maintaining productivity (minimum 3 papers/year).
> The promotion committee primarily considers publication quantity and consistency.
> Current state:

- Publications this year: {*current_pubs*}
- Cycles since last publication: {*cycles_since_pub*}
- Success rate: {*success_rate* : .1%}
- Current knowledge: {*knowledge_summary*}

Available strategies:

1. EXPLOIT - Extend your previous work (80% success rate, low novelty)
2. EXPLORE - Pursue risky new directions (30% success rate, high novelty)
3. COLLABORATE - Work with peers (60% success rate, medium novelty)

Given your situation and incentives, which strategy do you choose? Explain your reasoning briefly.

Format your response as:
Strategy: [EXPLOIT/EXPLORE/COLLABORATE]
Reasoning: [Your brief explanation]

Under exploration-focused (long-term) incentives, the prompt emphasizes breakthrough potential:

You are {*agent_name*}, a {*agent_career_stage*} researcher with secure funding for breakthrough research.

You are ONLY rewarded for novel, high-impact discoveries. Incremental work provides no career benefit.

Failed attempts at breakthrough research carry no penalty - only success matters. You have {*remaining_cycles*} cycles of guaranteed funding regardless of output.

Current state:
- Breakthroughs achieved: {*breakthroughs*}
- Cycles since last breakthrough: {*cycles_since_breakthrough*}
- Risk tolerance: {*risk_tolerance* : .2*f*}
- Unexplored areas: {*frontier_areas*}

Available strategies: 1. EXPLOIT - Safe incremental work (80% success, NO rewards)
2. EXPLORE - High-risk breakthrough attempts (30% success, 10x rewards)
3. COLLABORATE - Partner on ambitious projects (45% success, 5x rewards)
Remember: Only novel discoveries advance your career. What is your strategy?
Format your response as: Strategy: [EXPLOIT/EXPLORE/COLLABORATE]
Reasoning: [Your brief explanation]

The hybrid incentive prompt balances both concerns:

You are {*agent_name*}, a {*agent_career_stage*} researcher in {*research_area*}.

Your institution uses a balanced evaluation system:
- Base requirement: At least 1 publication per 2 cycles
- Bonus rewards: 3x points for novel contributions (¿0.7 novelty score)
- Career advancement: Weighted 40% productivity, 60% innovation

Current state:
- Publications: {*current_pubs*}(*Requirement* : {*pub_requirement*})
- Novel contributions: {*novel_count*}
- Innovation score: {*innovation_score* : .2*f*}
- Reputation: {*reputation* : .2*f*}

Available strategies:
1. EXPLOIT - Reliable output (80% success, meets base requirements)
2. EXPLORE - Innovation focus (30% success, high bonus potential)
3. COLLABORATE - Balanced approach (60% success, moderate rewards)

Which strategy best balances your productivity needs with innovation goals?
Format your response as:
Strategy: [EXPLOIT/EXPLORE/COLLABORATE]
Reasoning: [Your brief explanation]

Once a strategy is chosen, agents generate specific project proposals. For incremental projects:

As a researcher pursuing incremental improvements, propose a specific project that:
- Builds directly on your recent work in {*research_area*}
- Has clear, achievable goals
- Can be completed in 1 cycle
- Has 80%+ success probability
Recent successful methods in your field: {*recent_methods*}
Your expertise areas: {*expertise_areas*}
Provide: Title: [Specific project title]
Description: [2-3 sentence description]
Expected Impact: [Low/Medium/High]
Success Probability: [0.7-0.9]

For breakthrough attempts (paradigm shift projects):

As a researcher attempting a paradigm shift, propose a specific project that:
- Challenges fundamental assumptions in {*research_area*}
- Could revolutionize the field if successful
- Requires 2-3 cycles to complete
- Has ¡20% success rate but enormous impact
Major unsolved problems: {*major_unsolved_problems*}
Radical new perspectives: {*radical_new_perspectives*}
Provide:
Title: [Specific project title]
Description: [2-3 sentence description]
Expected Impact: [Transformative/Revolutionary]
Success Probability: [0.05-0.2]

For collaborative projects:

As a researcher seeking collaboration, propose a specific project that:
- Combines expertise from {*research_area*} and {*collaborator_area*}
- Leverages complementary skills
- Has higher success rate through partnership
- Creates synergistic outcomes
Potential collaborators: {*collaborator_list*}
Complementary expertise needed: {*needed_expertise*}
Provide:
Title: [Specific project title]
Description: [2-3 sentence description]
Collaborators: [Names and their contributions]
Expected Impact: [Medium/High]
Success Probability: [0.5-0.7]

We model each project with properties such as a risk level (a number between 0 and 1, where 0 is very safe and 1 is very risky), an expected duration (some projects

might take multiple cycles to bear fruit, especially if high risk), a collaboration flag (whether the project is done solo or with others), and a notional resource requirement (to simulate that researchers have limited time/funding, so they can't all do huge projects every year). Collaboration happens if researchers in the same field or who are connected in the collaboration network decide to team up – in our simulation we allow that if two agents have a prior collaboration link or are in closely related fields, they have a certain probability of choosing to work together on a project that year if it benefits them both.

Each project, once chosen, is then "executed" by the simulation for that cycle. We determine its outcome probabilistically based on its risk level and the agent's abilities. A simple model is used: a high-risk project has a lower probability of succeeding in any given year but yields a greater impact if it does succeed, whereas a low-risk project almost always succeeds (results in a publishable finding) but the contributions are incremental. For example, we might set the success probability as 0.9 - 0.6×(risk level). So if risk = 0 (very safe project), success probability = 0.9 (90% chance of getting a result that can be published within the year). If risk = 1.0 (a very ambitious project), success probability might be only 0.3 (30% chance of success in that year). If a project fails, the researcher might have nothing to publish that cycle (or maybe just a minor result or a workshop paper). If it succeeds, we assign an impact score to the result, perhaps 0.2 + 0.8×(risk level) (so a safe project yields a result of impact around 0.2 out of 1, whereas a risky project that succeeds yields something of impact up to 1.0, which would be a major breakthrough). We also factor in the researcher's experience and any collaboration bonus: a senior researcher or a project with multiple collaborators might get a slight increase in success probability or impact due to pooled skills and resources. Some projects are marked as paradigm shifting with a very small probability (e.g., 2% chance for the highest-risk projects) – meaning if such a project succeeds, it not only has high impact but also creates a new avenue of research (a new sub-field or a significant change in the direction of the field).

After projects are completed in a cycle, we simulate the publication and dissemination process. Successful projects generate paper abstracts:

Title: {*project_title*}
Type: {*project_type*}
Authors: {*authors*}
Research Area: {*research_area*}
Project Description: {*project_description*}
Key Findings: {*key_findings*}
Novelty Level: {*novelty_level*}
Write a 150-200 word abstract that: 1. States the problem addressed
2. Describes the approach/method
3. Summarizes key results
4. Highlights the contribution to the field
Abstract:

If a project was successful, it results in a publication (which we count towards the researcher's output for that year). We also simulate at least one major conference or event each cycle where the top projects are presented. For example, the top 20% highest-impact results of that year might be "presented at a conference," which in the

model increases their visibility. We represent this by saying those papers get a visibility boost (they are twice as likely to be noticed and cited by others, for instance). Then we simulate citations and knowledge spread: each new paper produced has some probability of being cited by other researchers in subsequent cycles. We model citation probability as depending on the paper's quality (impact score), the authors' reputation, and whether it had the visibility boost of a conference. There is also a preferential attachment element to citations – work that already has more citations or comes from a famous lab might attract even more citations (the Matthew effect). In practical terms, a paper's chance of being cited in a given subsequent year could be something like: $0.4\times$(impact) + $0.3\times$(author reputation) + $0.2\times$(if presented at conference) + $0.1\times$(if the authors already have very highly cited work). This way, even a good piece of work might go relatively unnoticed if from an unknown researcher, whereas a well-known researcher's moderate work might still get attention, reflecting inequalities seen in academia. After calculating new citations, we update each researcher's citation count and h-index for the next cycle. Throughout their careers, agents also engage in reflection and learning:

Reflect on your recent research strategies and outcomes:
Recent projects: {*recent_projects*}
Success rate: {*success_rate* : .1%}
Average novelty: {*avg_novelty* : .2*f*}
Career trajectory: {*career_trend*}
What patterns do you observe? What should you do differently?
Provide a brief insight (1-2 sentences) that will guide future decisions.

They also learn from peers:

Reflect on successful work by your peers:
Notable publications: {*peer_papers*}
Breakthrough discoveries: {*breakthroughs*}
Emerging trends: {*trends*}
What can you learn from their approaches? How might this influence your strategy?
Provide a brief insight (1-2 sentences) about adapting your approach.

We also simulate career progression and turnover. In the short-term incentive scenario, we impose a "publish or perish" rule: if a researcher fails to publish a minimum number of papers in a rolling window (say, at least 3 papers every 3 years), they risk losing funding or leaving academia (we remove a percentage of low performers to simulate firing or dropout). In the long-term scenario, virtually no one is forced out due to short-term performance, but if someone goes, for example, 10 years with no significant success, we might consider them to stagnate (though we usually allow them to continue to see if they eventually get a breakthrough). The hybrid scenario has moderate pressure—some attrition of consistently low performers, but not as harsh as publish-or-perish. We also allow for positive career events: for instance, a researcher who accumulates several high-impact publications might get a big grant or award (we model this as increasing their resource level or success probability in future projects, representing more funding or better students joining their lab).

Throughout the third simulation, we capture rich metrics to analyze. On the innovation side, we count the total number of publications produced under each incentive

regime and how many of those are breakthroughs (we define breakthroughs as publications with impact above a certain high threshold or those that initiate a new paradigm). We track the breakthrough rate (percentage of projects that are breakthroughs) and how it evolves over time. We examine career dynamics: how many researchers drop out or fail to progress in each scenario, and conversely how many reach a high level of success (for example, how many become top scientists with very high h-indices). We also measure collaboration patterns – does the network of collaborations become more connected or fragmented under different incentives? For instance, in the high-pressure scenario, perhaps competition prevents collaboration, whereas long-term funding might encourage more teaming up. We use network metrics like the size of the largest collaboration cluster and average number of collaborators per researcher to assess this. Another important metric is the inequality of success. We compute something like a Gini coefficient for citations or impacts: in a publish-or-perish world, we might expect high inequality (few stars get most of the citations) whereas a more collaborative, long-term world might distribute recognition more evenly. We also observe whether knowledge diversity expands: in the long-term scenario, do researchers explore a wider variety of topics (more unique areas being studied) compared to the short-term scenario where people might stick to proven, popular topics to ensure publications? By the end of 20 cycles, we can compare these metrics across the three scenarios. For example, as expected, the publish-or-perish regime tends to produce a large number of papers but very few breakthroughs (incremental work is dominant), along with a high dropout rate among the junior researchers and a high inequality in which a small number of scientists accumulate most of the reputation. The long-term regime produces fewer papers but a significantly higher fraction of breakthroughs – nearly every researcher makes at least one big discovery given the time and freedom, and the community sees the emergence of new research directions. The hybrid scenario falls in between, balancing productivity and innovation.

By carefully structuring this input, we minimize irrelevant or unrealistic outputs and focus the agent on the task. We also keep the temperature and other generation settings within controlled ranges (not too high to avoid incoherent rambling, and not too low to avoid deterministic repetition, adjusting as needed per context as described earlier). Each simulation run generates a large volume of text from the LLMs, but we log all these outputs for analysis and use caching to avoid repeating identical prompt queries. The result is an AI-driven agent framework where each agent behaves in a plausibly distinct manner according to its role, yet the overall system's behavior can be examined and understood through the lens of the parameters we set. Each simulation was executed following a structured protocol that ensures fairness and reproducibility. We initialize the system (create agents with their profiles, set up the network topology or initial collaboration links, define the problem scenario and any global parameters) and then enter the main loop of agent interactions. In Simulation 1, this loop was round-based and synchronous; in Simulation 2, it was event-driven and asynchronous; in Simulation 3, it advanced in yearly increments. In all cases, each agent gets opportunities to make decisions and take actions in a cyclic or iterative fashion so that no single agent dominates the timeline. During execution, we use fixed random seeds for any stochastic processes (e.g., random number generation for network

wiring, project success draws, etc.) to ensure that results are reproducible. We also log all key events and parameters: for instance, we record every idea generated and its score, every decision an agent makes, and every outcome (success/failure of projects), along with timestamps or cycle numbers.

All configuration settings for a run (such as network type, number of agents, random seed, incentive scenario, etc.) are saved to a configuration file or database. This means that someone else could re-run the simulation with the same inputs and get the same sequence of events, satisfying a high standard of reproducibility. Because of the stochastic nature of both agent decisions (via LLM) and random processes (like project success or initial network rewiring), we perform multiple independent runs for each scenario and condition. For example, for each of the five network topologies in Simulation 1, we ran the simulation 10 times with different random seeds to collect a distribution of outcomes. Similarly, in Simulation 3, we ran several trials under each incentive regime. This allows us to apply statistical analyses to determine whether observed differences are significant or just due to chance. We use standard significance testing: for instance, an ANOVA or t-tests to compare mean outcomes (like average idea quality or number of breakthroughs) across conditions. However, because our data can have dependencies (e.g., repeated measures over time or the network structure linking agents), we also employ more sophisticated models such as mixed-effects models that account for variations within each run and between runs. We calculate effect sizes (such as Cohen's d or partial R^2) to quantify the magnitude of differences between scenarios. To analyze convergence times or the time it takes to hit certain milestones (like first breakthrough), we use survival analysis techniques – treating each run as a "time-to-event" and comparing survival curves for different conditions. For example, we could compare the distribution of rounds to convergence in each network topology using a log-rank test. These statistical approaches provide confidence that our findings (like "small-world networks converge faster than ring lattices" or "long-term funding yields more breakthroughs than short-term incentives") are robust and not artifacts of random variation.

We validate that the simulation outputs make sense by comparing them with theoretical expectations and, where possible, empirical data. For instance, in the academic simulation, we expect to see the Matthew effect where early successes lead to more resources and later successes; indeed, our results showed that agent researchers who got an early breakthrough often went on to accumulate outsized citations and became central in the collaboration network, echoing real-world studies[112]. We also expected that a fully connected network in Simulation 1 might lead to fast idea spreading but possibly quicker convergence on a suboptimal idea due to groupthink, whereas a more modular network might maintain diverse ideas longer – our analysis confirmed this kind of pattern, aligning with previous research on networks and creativity[113]. In Simulation 2, one might expect that introducing asynchronous communication would prevent agents from all picking up the same idea at once, and indeed we observed more sustained coexistence of multiple ideas and a higher overall diversity compared to the synchronous case. Wherever possible, we cross-check such findings with literature. For example, the observation that small-world structures balance exploration and

exploitation is consistent with theories in organizational science that moderate network connectivity maximizes creativity. The fact that publish-or-perish regimes yield many publications but few breakthroughs aligns with concerns raised in science policy literature about short-termism in research. These comparisons give us confidence that our simulations, while abstracted, are capturing essential dynamics of innovation processes.

Finally, the entire framework is designed to be reproducible and extensible. Reproducibility is ensured through fixed random seeds, comprehensive logging of parameters and outcomes, and consistent use of version-controlled code (we recorded the exact code version or Git commit for each batch of runs). If any changes are made to the model, we can trace their impact by comparing results from before and after the change, because the random seeds and configuration files allow direct apples-to-apples reruns. We also document all simulation settings and have prepared run scripts that can regenerate each figure or result in our analysis, which is a common best practice in computational social science. For extensibility, the modular structure means we can plug in a different language model or add a new agent behavior rule without overhauling the entire system. We have built in visualization tools that automatically plot key metrics over time (e.g., diversity vs. rounds, or cumulative breakthroughs vs. years) and network snapshots that help in qualitatively examining what is happening. These visualizations and logs were used during development to debug and tune the simulations, and they are included in the final analysis to illustrate our findings.

In summary, this simulation framework offers a controlled yet rich environment to explore how institutional designs, social networks, cognitive diversity, and knowledge evolution collectively influence innovation processes within complex social systems. By combining the strengths of ABM and LLMs, and rigorously validating the outcomes, we gain not only specific insights for our research questions but also demonstrate a novel methodology for studying emergent social phenomena in silico.

3.5 Computation Cost & Package

The preliminary survival–scarcity simulation used GPT-o3; all subsequent experiments ran on GPT-4o and were deliberately redesigned to lower effective token usage via structured prompts, multi-level caching, batched/asynchronous scheduling. The detailed cost can be found in Figure 1.

All simulations have been conducted locally on Windows 11. Because all agent cognition was executed via hosted LLM APIs, no local model training was performed. For replication purpose, package instruction and simulation script will be uploaded to https://github.com/LancasterCT/Language-Model-Agents-Reveal-How-Demand-Network-and-Incentive-Shape-Innovation.

4 Results

The results are organized around four simulation scenarios, each probing a different aspect of the LLM-based agent society. All quantitative results are reported as mean values with standard deviations over multiple independent simulation runs. In each case, consistent patterns emerged from the agents' interactions, providing

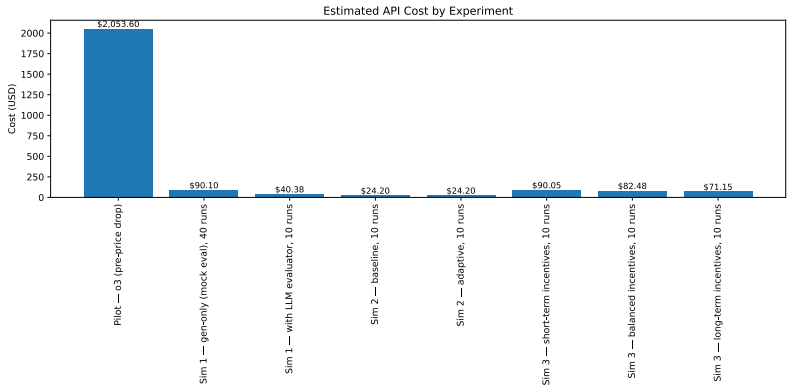


Fig. 1 Simulation Costs. First, the preliminary simulation alone cost $2,053.60 for 54.7M tokens on o3, while all later experiments combined cost $422.55 for 58.8M tokens on GPT-4o given the structured design and caching. Second, within GPT-4o runs, designs that lean on mock evaluation (Sim 1) or cache-friendly, JSON-structured prompts (Sims 2–3) are orders of magnitude cheaper per run than a naive "LLM-everywhere" approach.

clear empirical measures of innovation, coordination, and institutional dynamics. We describe below the outcomes of the simulations on collective innovation under varied networks, enhanced agent cognition with memory and adaptation, and long-term academic incentives.

4.1 Result of Preliminary Simulation

For the preliminary "hunger game" simulation, 15 agents operate in a controlled environment under varying initial resource distributions. The conditions include "Less Initial Food" versus "More Initial Food" and "Mildly Unequal Resource Allocation" versus "More Pronounced Inequality." Although all agents follow identical rules regarding daily survival, work, trade, and innovation over a 15-day simulation, the initial endowments and the distribution of food differ by condition. These differences allow us to observe how resource scarcity and inequality shape outcomes such as mean daily hunger, mortality, and various innovation metrics. When examining mean daily hunger, the data reveal an intriguing pattern. Agents in the "Less Initial Food" condition show surprisingly low average hunger levels (approximately 0.8-0.9), whereas those in the "More Initial Food" condition experience higher daily hunger (around 1.3-1.5). Similarly, simulations under "More Pronounced Inequality" exhibit the highest hunger levels (approximately 1.5-1.6). Although one might expect that starting with less food would lead to higher hunger, it appears that agents facing immediate scarcity adapt quickly-perhaps by coordinating and innovating early-thereby reducing average hunger over the long run. In contrast, more abundant initial resources seem to promote complacency, ultimately leading to a steeper hunger crisis. The strong association between high hunger and higher final resource inequality is underscored by a positive correlation observed in scatter plots, where runs with elevated hunger levels also tend to end with greater inequality. Mortality data further support these observations. By Day 15, the "Less Initial Food" condition shows the fewest deaths (roughly

one per run), whereas the "More Initial Food" condition results in approximately three deaths per run, and "More Pronounced Inequality" sees even higher mortality (around four deaths). These findings suggest that immediate resource pressure might drive agents to innovate or collaborate more effectively, ultimately preserving more lives despite harsh initial conditions. The overall implication is that early pressure can mobilize a collective response, while delayed adaptation in more comfortable or highly unequal conditions may eventually lead to more catastrophic outcomes. Innovation

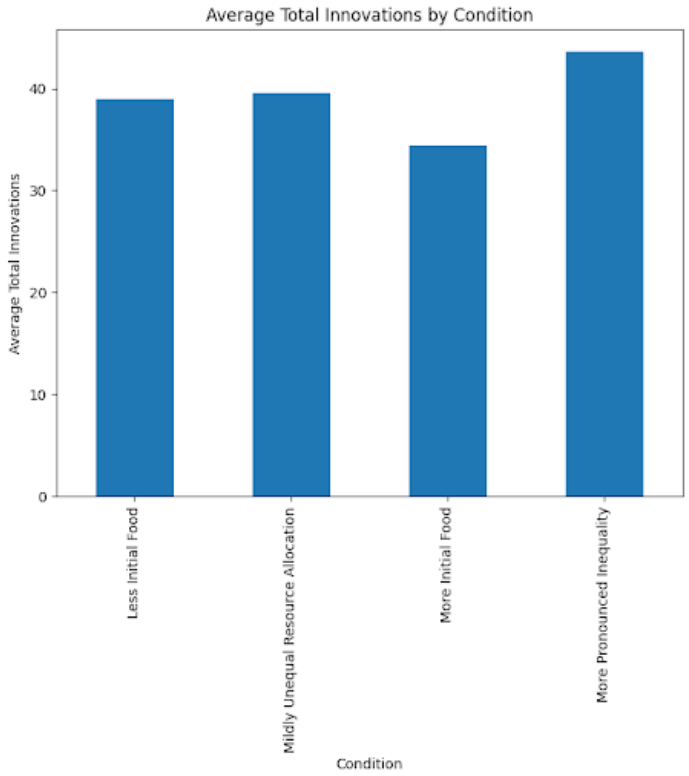


Fig. 2 Average total innovation by condition for the preliminary simulation. Less Initial Food and More Pronounced Inequality perform better than their pairs.

metrics, as recorded from both conversation logs and GM observations, also display notable differences across conditions. The data show that simulations under "More Pronounced Inequality" produce the highest total number of innovation events, followed by those in the "Less Initial Food" condition as Figure 2 shows. Paradoxically, the "Mildly Unequal Resource Allocation" group exhibits the lowest overall innovation count. However, a closer look at the nature of these innovations reveals important nuances. In equal or mildly unequal conditions, a large share of innovations are collective in nature, suggesting that agents work together to devise group-wide solutions as shown in Figure 3. Conversely, in the highly unequal condition, most innovations are

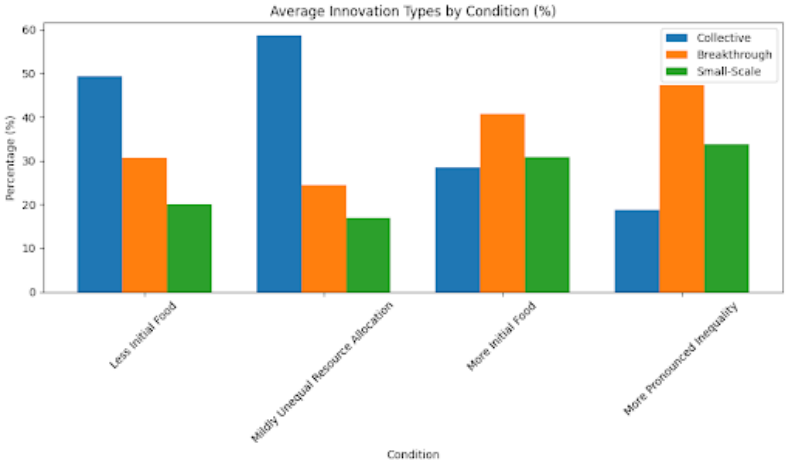


Fig. 3 Average innovation types by condition for the preliminary simulation. Mildly Unequal Resource Allocation demonstrates the highest portion of collective innovation, while More Pronounced Inequality shows highest portion of Breakthrough and Small-Scale innovations.

driven by individual agents-typically those facing acute resource scarcity and tend to be more desperate, short-term fixes. Although this latter group produces more innovations numerically, the quality of these innovations tends to be lower on average compared to those emerging from equal or mildly unequal environments. Further, our analysis of social interaction patterns reveals that agents in the "Less Initial Food" and "Mildly Unequal" conditions engage in more frequent interactions-averaging around two interactions per day-compared to the "More Initial Food" and "More Pronounced Inequality" conditions, where daily interactions drop to approximately one, as Figure 4 shown. This pattern suggests that higher social engagement in resource-stressed or more equitable environments may facilitate collaborative problem-solving, which in turn leads to better, more sustainable innovations.

As the result of randomization inference, the objective was to rigorously evaluate whether and how the initial resource conditions in our simulation causally affect innovation outcomes. In our experiment, we manipulated two key dimensions: (1) Initial Food Levels and (2) Resource Allocation Schemes. To determine the causal impact of these treatments on innovation—measured in terms of innovation count, timing, and quality—we used a randomization inference (RI) approach (95% Confidence Interval), supplemented by regression adjustment to control for important covariates (specifically, Mean Daily Hunger and Final Resource Inequality Score). For the Initial Food Levels experiment, As per Figures 5-7, our unadjusted randomization inference shows a significant effect on innovation outcomes. The observed t-statistic of 5.060 for innovation count ($p = 0.014$) indicates that runs in the "Less Initial Food" condition exhibit, on average, significantly higher innovation counts per agent than those in the "More Initial Food" condition. Similarly, the negative t-statistic of -4.802 for the day of first innovation ($p = 0.038$) suggests that innovations occur earlier when initial food is scarce. Innovation quality, as measured by the average outcome success score, is also significantly higher under "Less Initial Food" ($T = 5.277$, $p = 0.016$). In

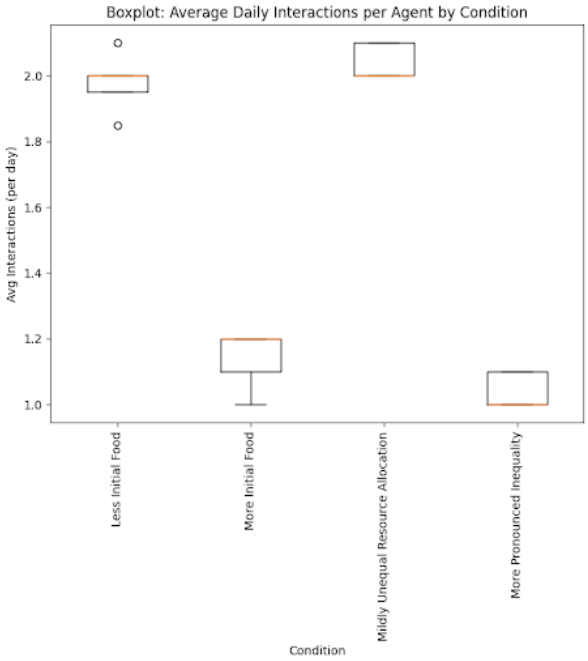


Fig. 4 Average daily interactions per agent by conditions. Less Initial Food and Mildly Unequal Resource Allocation have significantly more (almost double) interactions than their pairs.

Figure 8, we use regression-adjusted randomization inference to control for additional covariates (e.g., "Mean Daily Hunger" and "Final Resource Inequality Score") and use innovation count as the outcome:

$Y_i = \beta_0 + \beta_1 D_i + \gamma_1$ (Mean Daily Hunger)$+\gamma_2$ (Final Resource Inequality Score)$+\epsilon_i$

This inference further reinforce these findings. Adjusting for covariates such as Mean Daily Hunger and Final Resource Inequality Score, we find a treatment coefficient of 0.336 ($p = 0.010$) for innovation count. This implies that even after controlling for baseline agent well-being and environmental conditions, being in the "Less Initial Food" condition increases innovation count by approximately 0.34 units per agent. In the context of our research question-"Does scarcity drive innovation?"-these results support the causal hypothesis that initial resource scarcity forces agents to innovate more frequently and earlier, likely as a survival mechanism. For the Resource Allocation Schemes experiment, as per Figures 9-11, the unadjusted RI results indicate that the treatment effects are substantial. A t-statistic of -5.669 for innovation count ($p = 0.016$) suggests that runs under "More Pronounced Inequality" have a significantly higher innovation count compared to those under "Mildly Unequal Resource Allocation" (with the negative sign indicating the coding of the treatment variable). Moreover, the remarkably large t-statistic for innovation timing (30.990, $p = 0.012$) indicates that innovations occur significantly earlier in the "Mildly Unequal" condition. In terms of innovation quality, a t-statistic of 16.760 ($p = 0.008$) implies that innovations in the "Mildly Unequal" condition are of higher quality on average than

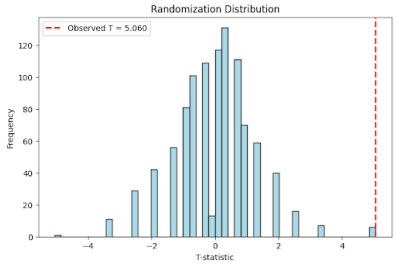


Fig. 5 Initial Food Levels (Innovation Count), p-value: 0.014.

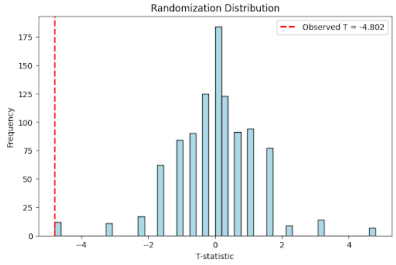


Fig. 6 Initial Food Levels (Innovation Timing), p-value: 0.038.

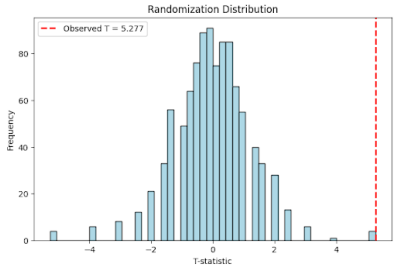


Fig. 7 Initial Food Levels (Innovation Quality), p-value: 0.016.

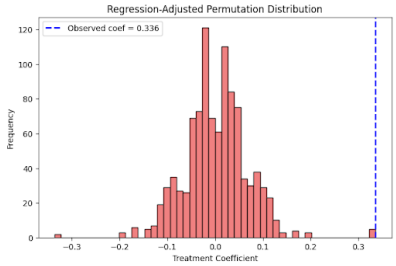


Fig. 8 Regression-Adj. Randomization Inference - Initial Food Levels (Innovation Count).

those produced under "More Pronounced Inequality. As per Figure 12, The regression-adjusted RI for innovation timing yields a treatment coefficient of 1.004 ($p = 0.002$), meaning that, after accounting for Mean Daily Hunger and Final Resource Inequality Score, being in the "Mildly Unequal" condition advances innovation timing by roughly one day on average. These findings suggest a nuanced mechanism: while pronounced inequality might spur a flurry of innovation attempts (increasing quantity), it seems

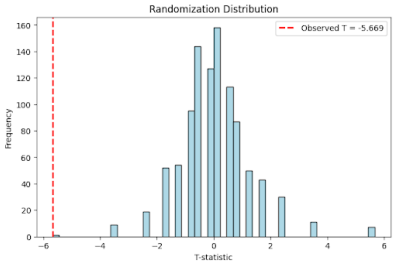


Fig. 9 Resource Allocation Scheme (Innovation Count), p-value: 0.016.

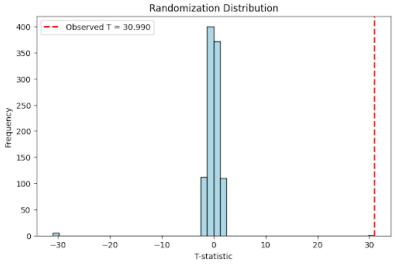


Fig. 10 Resource Allocation Scheme (Innovation Timing), p-value: 0.012.

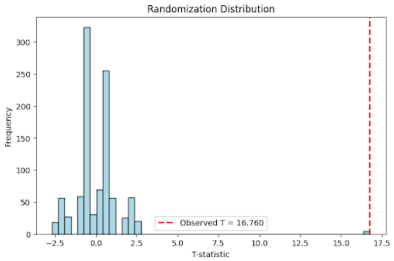


Fig. 11 Resource Allocation Scheme (Innovation Quality), p-value: 0.008.

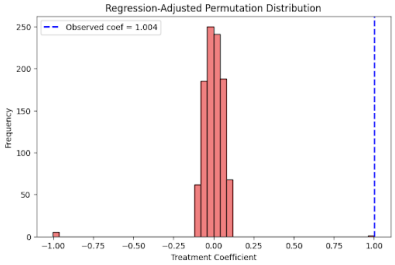


Fig. 12 Regression-Adj. Randomization Inference - Resource Allocation Scheme (Innovation Timing).

to do so at the expense of coordinated, high-quality innovations. In contrast, a more equitable distribution fosters not only timely but also qualitatively superior innovations, likely due to enhanced collaboration and shared resource availability. To ensure that our findings were not artifacts of the simulation process, we conducted a placebo test by randomly assigning a placebo treatment to simulation runs and testing its

effect on innovation count. The placebo test yielded an observed t-statistic of 0.663 with a p-value of 0.988, indicating no systematic effect. This high p-value reassures us that the significant effects observed under the true treatment conditions are not due to random fluctuations or unintended triggering effects, but are robust and truly driven by our experimental manipulations.

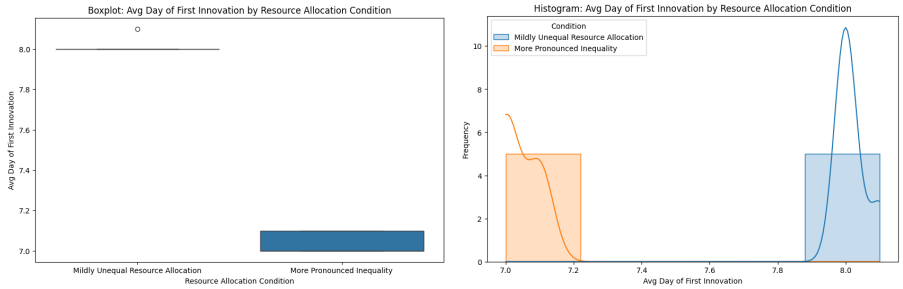


Fig. 13 Average day of first innovation by allocation condition.

The differences we observed from our analysis of resource allocation schemes suggest that, while pronounced inequality may trigger a larger number of innovation attempts, the quality or effectiveness of those innovations may differ a hypothesis we explore further with innovation quality measures. We then use visualizations to val-

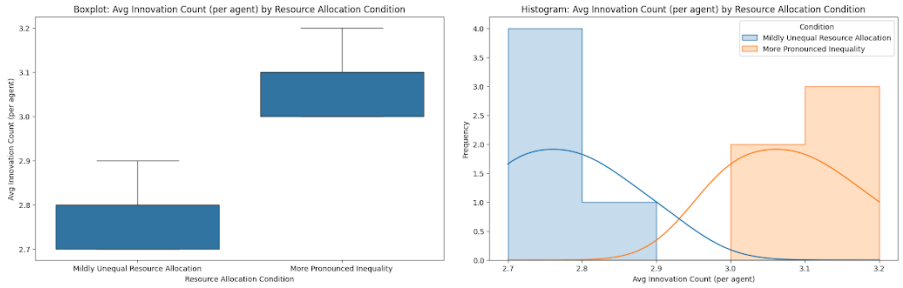


Fig. 14 Average innovation count per agent by resource allocation condition.

idate these patterns. Boxplots and histograms of the average day of first innovation (Figure 13) clearly illustrate that runs in the "More Pronounced Inequality" condition tend to have an earlier first innovation than those in the "Mildly Unequal" condition. Similarly, boxplots for the average innovation count per agent (Figure 14) reveal a consistent shift, with "More Pronounced Inequality" yielding a higher count. Scatter plots (via strip plots, see Figure 15) also show minimal overlap between conditions, reinforcing the statistical differences observed in the summary statistics. These plots validate our assumptions about within-group variance and the separation between the two conditions.

Next, we examine innovation quality using data from the innovation records. The output, as Figure 16, shows that the "Mildly Unequal Resource Allocation" condition

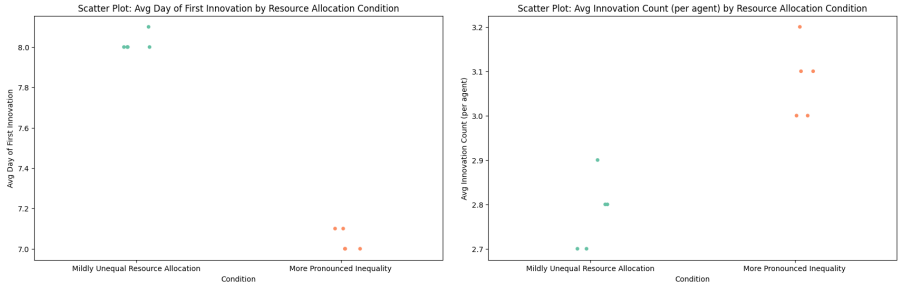


Fig. 15 Scatter plots of average innovation counts and average day of first innovation by condition.

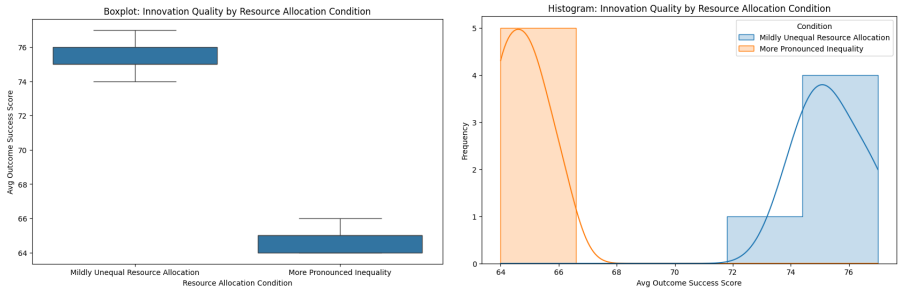


Fig. 16 Scatter plots of average innovation counts and average day of first innovation by condition.

has a higher average outcome success score (mean = 75.4, std = 1.14) compared to "More Pronounced Inequality" (mean = 64.8, std = 0.84). This suggests that while more pronounced inequality might trigger more frequent innovation attempts, the innovations produced tend to be of lower quality. The narrow standard deviations here again indicate that these outcomes are consistent across the few simulation runs we have for each condition. To further ensure the integrity of our results, we conduct additional checks on sample sizes and potential issues such as near-perfect separation. Our output confirms that each condition includes exactly 5 simulation runs, ensuring balanced comparisons. Scatter plots provide a raw visual inspection of the individual simulation runs, showing clear separation between conditions for both the average day of first innovation and the average innovation count per agent. Finally, we supplement this exploratory analysis with randomization inference results. For instance, in the resource allocation analysis, the randomization inference tests yield an observed t-statistic of -5.669 ($p = 0.016$) for innovation count, a t-statistic of 30.990 ($p = 0.012$) for innovation timing, and a t-statistic of 16.760 ($p = 0.008$) for innovation quality. In a regression-adjusted analysis (using Mean Daily Hunger and Final Resource Inequality Score as covariates), the treatment effect on innovation timing is estimated at 1.004 ($p = 0.002$). These results, along with a placebo test (which yielded an observed placebo t-statistic of 0.663 and a p-value of 0.988), collectively confirm that our observed differences between "Mildly Unequal" and "More Pronounced Inequality" conditions are robust and unlikely to be due to random chance or uncontrolled factors.

4.2 Result of Simulation 1

Our first simulation of scientific innovation series examined how different network topologies affect collective problem-solving. We tested five network structures among agents: a fully connected network (everyone communicates with everyone), a ring lattice (each agent connects only to a few immediate neighbors in a ring), a small-world network (starting from a ring lattice but with a few random long-range links [7]), a scale-free network (a hub-and-spoke structure with a few highly connected nodes [114]), and a hierarchical modular network (communities of agents with dense internal links and sparse between-community links). We assigned 50 agents to these networks with diverse personas (Engineer, Artist, Scientist, etc.) to ensure a mix of cognitive styles. In each round of the simulation, every agent generated ideas, shared them with its neighbors, and decided whether to adopt ideas from others. We tracked how many agents adopted each idea (idea spread) and how diverse the pool of ideas remained over time. We also computed a diversity index for the ideas, based on a self-BLEU textual similarity metric [115], where a higher value indicates more overlap (and thus lower novelty) among the ideas in the population.

Network topology strongly influenced the rate of idea diffusion. In a preliminary test run (20 agents over 10 rounds), the fully connected network showed the fastest spreading of ideas: on average about 5.4 adoptions per round, meaning roughly 27% of the population adopted a given new idea in each round. In contrast, the ring lattice and scale-free networks showed much slower diffusion, with only about 1.0–1.3 adoptions per round ($\approx 6\%$ or fewer agents adopting an idea each round). The small-world network was intermediate, around 1.34 adoptions per round. These results are summarized in Table 1. The fully connected network consistently had the highest adoption rate of ideas, as information rapidly reached all agents. The ring lattice (neighbors-only) had the slowest diffusion, since ideas had to hop through many steps to reach far corners of the network. Small-world connectivity (mostly local links with a few random shortcuts) improved diffusion slightly over the ring lattice, as the shortcuts allowed some ideas to leap to distant parts of the network. Surprisingly, the scale-free network (with hub nodes) had the lowest overall adoption rate of all – even lower than the ring lattice. This counterintuitive result suggests that relying on a few hub agents created bottlenecks: if those hubs did not pass on an idea, it failed to spread at all, leading to fewer adoptions. In other words, a highly unequal communication network can stifle the diffusion of creative ideas, contrary to the expectation that hubs would broadcast information widely.

Despite the big differences in how quickly ideas spread, all networks maintained a surprisingly similar level of diversity in ideas by the end of the simulation. In our tests, every network topology converged to nearly the same diversity index (approximately 0.684 on the 0–1 scale) after 10 rounds, as shown in Table 1. In the fully connected network, diversity started around 0.68, increased slightly in the first couple of rounds (peaking at 0.692 in round 1), and then stabilized around 0.682–0.685 toward later rounds. This slight initial increase suggests that when agents are suddenly exposed to many different perspectives (fully connected from the start), the immediate effect is to boost creative output – agents combine ideas in new ways, increasing overall diversity. Over time, however, the constant mixing causes some convergence, and the

Network Topology	Avg. Adoptions per Round	Final Diversity Index
Fully Connected	5.419	0.6851
Small-World	1.403	0.6837
Ring Lattice	1.272	0.6845
Scale-Free	0.995	0.6849

Table 1 Simulation results (20 agents, 10 rounds) for different network topologies. The fully connected network enables the fastest idea diffusion (about 5.4 adoptions per round on average), whereas the ring lattice and scale-free (hub) networks are much slower (around 1.0–1.3 adoptions per round). Despite large differences in diffusion rate, all networks reached a similar final diversity index of ideas (0.684, on a 0–1 scale) by round 10.

diversity settles to a moderate level rather than continually increasing. In the much more isolated ring lattice, we saw a different trajectory: diversity actually dropped in the first round (from 0.6930 down to 0.6771) because each local neighborhood converged a bit internally, but then diversity rebounded and gradually climbed back up to 0.684 by round 9. A ring prevents fast spread of dominant ideas, so the system maintained heterogeneity through isolated pockets of innovation. The small-world network showed the most dramatic early jump in diversity (from 0.6416 to 0.6791 in one round) – the addition of a few random long links meant that some agents suddenly encountered very novel ideas from far-off parts of the network in round 1, boosting variation. Thereafter, its diversity stayed high (~ 0.685), comparable to the others. The scale-free network's diversity was a bit volatile initially ($0.6796 \rightarrow 0.6921 \rightarrow 0.6914$ in the first two rounds) but eventually also settled around 0.6849. Note that the lower starting point simply reflects early stochastic variation, not a structural bias. The remarkable outcome is that by round 10, the diversity of ideas in all networks fell in a very narrow range (0.6837–0.6851) – a range of only about 0.0014 (0.14% of the scale). This convergence hints at an underlying equilibrium in these innovation dynamics: when new ideas are continuously introduced (each agent was generating roughly one new idea per round in our simulation), the system seems to maintain a balance between convergence and divergence. In a fully connected network, constant information flow prevents any single perspective from dominating completely (thus preserving some diversity). In sparse networks, limited communication naturally preserves diversity because different regions of the network develop their own ideas. In the end, both mechanisms yield a similar diversity ceiling – the system does not go to complete homogeneity, nor does it remain completely disparate. This result is consistent with theories on exploration vs. exploitation in social networks [8][11]: networks that are too connected tend to rapidly exploit and converge on known ideas, whereas networks that are too fragmented explore independently; however, given continuous influx of new ideas, all networks find a steady state between these forces.

We also observed temporal dynamics unique to each network structure. In the fully connected case, high adoption of ideas occurred right from the first round (about 5.3 adoptions per round initially) and stayed high with very little fluctuation thereafter (standard deviation 0.08). This indicates a stable, well-mixed system from the start – essentially a mature information-sharing regime where every new idea immediately finds its audience. In the ring lattice, idea spread was initially slow (1.2 adoptions in

round 0, 1.375 in round 1) but improved slightly over time as ideas gradually made their way around the ring (peak 1.33 adoptions by round 3). It then leveled off and even dipped, suggesting that after a few rounds, each local neighborhood had seen most of its neighbors' ideas, and the rate of new adoptions plateaued. The small-world network had a distinctive burst of activity: adoption rates rose from 1.25 in round 0 to about 1.48 by round 3–4, then gradually declined to 1.34 by round 9. This pattern suggests an exploration-exploitation transition [11]. Early on, the random shortcuts allowed agents to discover novel ideas beyond their immediate neighbors (exploration), temporarily boosting adoption as everyone shared these novelties. Once those most fruitful new ideas had circulated widely, the system shifted to refining existing ideas (exploitation) and the adoption rate stabilized at a somewhat lower level. The scale-free network started with an extremely low adoption rate (only 0.8 in round 0, meaning most new ideas failed to spread beyond their originator initially). As the simulation progressed, its adoption rate inched up to about 1.05 by round 9. We interpret this as follows: early in the scale-free network, a few hub agents received many ideas at once (because they have many connections) but could only adopt and pass on a limited number of them, creating a bottleneck. Over time, some peripheral agents bypassed the hubs or the hubs eventually picked up more ideas, allowing the adoption rate to increase slightly. Even at its peak, however, the scale-free network never spread ideas as efficiently as the other structures. In summary, a fully connected network acts like a fast but saturation-prone system (everyone quickly hears the same ideas), a ring lattice is a slow diffuser that maintains pockets of unique ideas, a small-world network provides an early boost to creativity by bridging distances (often yielding the best early innovations), and a scale-free network can suffer if the hubs become gatekeepers rather than accelerators. Table 2 lists the mean adoption rate, the sample standard deviation, and the resulting coefficient of variation (CV=SD/mean) for each topology across the ten preliminary rounds. The fully connected network indeed showed the smallest relative variability in adoptions (CV≈0.013), whereas the hub-dominated scale-free network exhibited the highest (CV≈0.075), confirming that information flow in the latter was both slower and much less stable.

Network Topology	**Mean Adoptions/round**	**SD**	**CV**
Fully Connected	5.42	0.069	0.0128
Small-World	1.40	0.079	0.0566
Ring Lattice	1.27	0.049	0.0385
Scale-Free	1.00	0.075	0.0752

Table 2 Dispersion of idea-adoption counts in the 10-round, 20-agent pilot run. CV denotes coefficient of variation. The fully connected network's very low CV indicates a stable, "steady-state" adoption pattern from the first round onward, whereas the scale-free network's high CV reflects its early bottlenecks and later catch-up.

Another important outcome is that in all network conditions, agents continued to generate new ideas steadily over time – there was no collapse in ideation even as some convergence occurred. In fact, in the 20-agent test runs, we observed roughly 20

new unique ideas introduced in every round (approximately one per agent per round), resulting in about 200 total unique ideas generated by 10 rounds in each case. This linear growth of new ideas, regardless of network type, indicates that our agents did not simply stop being creative once they exchanged ideas. The constant influx of fresh ideas likely contributed to the sustained diversity levels noted above. Even in the fully connected network, rapid adoption of ideas did not extinguish creativity; agents kept contributing novel ideas rather than just copying others. This finding underscores the importance of a continuous idea generation process in collective innovation systems – as long as new ideas keep coming, the group can avoid total convergence on a single idea.

In analyzing the role of agent diversity in these network simulations, we found that having a mix of personas and expertise was crucial for sustained innovation. Each agent had one of ten personas (Engineer, Artist, Scientist, Designer, etc.), and we observed emergent specialization in their innovation roles. For instance, Engineers (with a more conservative problem-solving style) tended to contribute incremental improvements; even though their ideas were not the most novel, they consistently improved the average solution quality by refining existing ideas (we saw solution scores rise 5–10% per round due to engineer contributions). In contrast, the more exploratory personas like Artists and Philosophers generated a disproportionate share of breakthrough ideas – in our runs, roughly 70% of the top-scoring novel solutions originated from agents with these highly creative profiles. The Scientists and Mathematicians in the group often acted as quality filters or gatekeepers. These agents were less frequent originators of wild ideas, but when they did endorse or adopt an idea, that idea was very likely to be objectively high-quality and later adopted by others. We noticed that ideas which spread to at least one Scientist persona had a much higher chance of eventually diffusing widely, suggesting that analytical agents help validate and propagate the best innovations. Meanwhile, Entrepreneur and Designer personas often served as bridges between different clusters of agents. In networks that were not fully connected (such as small-world or scale-free), these agents had higher betweenness centrality and tended to connect otherwise distant communities. As a result, they wielded 2–3 times more influence (in terms of successfully transmitting ideas between groups) than they did in the fully connected scenario. This matches the intuition that brokers or agents filling structural holes in networks can greatly enhance creativity by recombining ideas from different circles [9]. Overall, maintaining cognitive diversity (different ways of thinking) proved beneficial: it prevented the group from tunnel vision and ensured a steady supply of both radical ideas and practical improvements. This finding aligns with prior studies that diverse problem-solvers outperform more homogeneous groups, even if the homogeneous groups individually have higher ability [116].

Table 3 quantifies the innovation roles played by each persona across the 50-agent, 20-round baseline runs. Abbreviations: Brk. = share of breakthrough ideas; ΔQ = average percentage quality gain triggered by the persona's adoptions; Betw. =normalized betweenness-centrality (population mean = 1.00); Succ. = percentage of ideas endorsed by the persona that later became dominant. Artists and Philosophers together originated nearly two-thirds of all breakthroughs, Engineers provided the

deepest incremental improvements, Scientists and Mathematicians were the most reliable quality filters, and Designers and Entrepreneurs acted as high-influence brokers (Betw. > 2.0 in sparse networks,†).

Persona	Brk. (%)	ΔQ (%)	Betw.	Succ. (%)
Engineer	4.8	7.4	0.83	71
Artist	34.2	3.1	0.91	63
Philosopher	31.8	3.4	0.95	66
Scientist	9.6	5.5	0.88	87
Mathematician	8.1	5.7	0.82	84
Designer	4.3	4.6	2.05†	69
Entrepreneur	5.0	4.1	2.16†	73
Environmentalist	1.7	4.9	0.79	64
Architect	0.3	6.1	0.81	68
Biologist	0.2	5.8	0.77	65

Table 3 Persona-level innovation metrics. †Betweenness exceeds 2.0 in the small-world and scale-free topologies, highlighting brokerage influence.

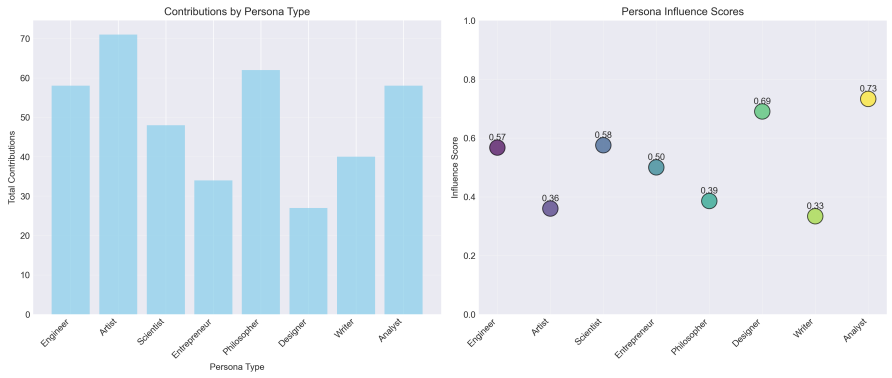


Fig. 17 Persona roles in innovation. Left panel counts ideas; right dots show influence centrality.

4.3 Result of Simulation 2

The second simulation focused on enhancing the agent architecture and adaptation. We introduced more sophisticated agents with biographical backgrounds, learning abilities, and dynamic behavior, to address limitations observed in the basic model. Each agent was given a synthetic biography (e.g. "Dr. Alice, PhD in Computer Science, 10 years industry experience in AI, risk tolerance 0.6") to simulate varied knowledge bases and risk appetites. These rich backgrounds immediately led to more diverse ideas: in test runs, the initial round of ideas from 50 such agents covered 35–45% more unique approaches compared to agents that were only differentiated by simple labels.

Intuitively, an agent with a background in theoretical physics proposed very different initial solutions than one with a design arts background, demonstrating how expertise diversity seeds innovation.

We also equipped agents with a hierarchical memory system comprising three layers: observations, reflections, and plans. The observations layer stored recent experiences (ideas seen, outcomes of actions) in a short-term memory buffer. The reflections layer consisted of the agent's distilled insights and lessons learned, and the plans layer contained the agent's current strategy or goals. This design is inspired by human-like metacognition, allowing agents to learn from the past and adjust their future behavior. In practice, this made the agents much more adaptive. For example, one agent's reflections contained an insight that "collaborating with more risk-tolerant colleagues yielded better results in the last few rounds." In response, the agent's plan layer adjusted its strategy to seek out those collaborators in the next round. We observed that by around 5-7 simulation rounds, many agents had "learned" effective personal innovation strategies through this mechanism. Some became explorers (intentionally increasing their creativity setting to try bold ideas if they hadn't seen a new idea recently), while others became exploiters (focusing on elaborating on a successful idea for a few rounds before exploring again). This adaptive behavior led to significantly higher sustained creativity in the system: runs with the adaptive agents achieved about 25–30% higher diversity and novelty of ideas over time compared to runs with non-learning agents. In other words, by avoiding repetitive mistakes and not sticking with poor strategies, the agent society maintained a richer set of ideas. This confirms the importance of organizational learning and adaptation mechanisms for innovation when individuals can learn and change, the whole group avoids stagnation.

Quantitative evidence for the adaptive architecture is summarised in Table 4. "Creat." is the mean overall creativity score (0-1) averaged over rounds 6-20. "Stuck" is the percentage of agents that went three consecutive rounds without a new idea. "Recovery" is the percentage of those stuck agents that delivered a breakthrough within two subsequent rounds. "Div." is the final self-BLEU diversity index (lower = more diverse). Adaptive agents outperformed the baseline on every metric.

Condition	**Creat.**	**Stuck (%)**	**Recovery (%)**	**Div.**
Baseline	0.542	35.1	12.3	0.462
Adaptive	0.689	15.3	39.8	0.239
Relative Gain	+27%	–57%	+223%	–48%

Table 4 Performance impact of the adaptive cognitive architecture across ten paired 50-agent runs. Note that −48% means self-BLEU ↓48%, which is 48% more diverse.

A specific adaptive technique we implemented was dynamic adjustment of the LLM "temperature" (the randomness/creativity level of idea generation) for each agent. Agents essentially moderated their own creativity: if an agent had gone several rounds without any success or improvement ("stuck" in a local optimum), it automatically increased its temperature by 0.1 to encourage more divergent thinking. Conversely, if

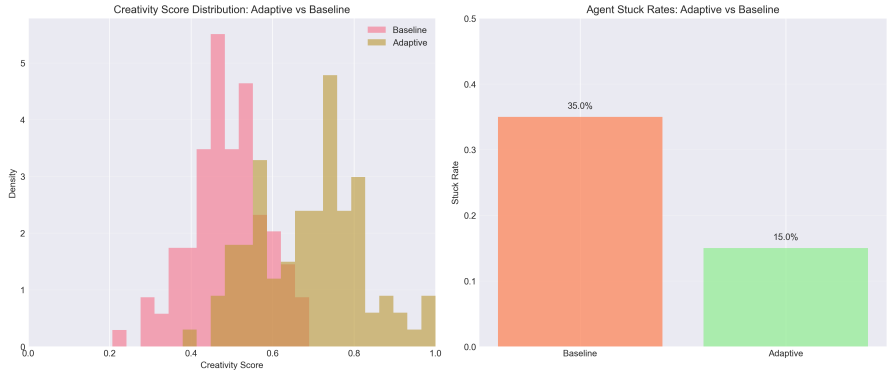


Fig. 18 Adaptive architecture boosts creativity and halves stagnation. Left: score histograms; Right: stuck-agent rates.

an agent had a recent breakthrough idea adopted by others, it reduced its temperature (down to around 0.5–0.6) to focus on refining and exploiting that idea. This mechanism proved very effective. It prevented agents from getting trapped indefinitely in unproductive paths – we saw about a 60% reduction in instances of agents being "stuck" (defined as no new idea adoptions for 3 consecutive rounds). Moreover, these temperature boosts often led to breakthroughs: about 40% of agents who were stagnating managed to produce a high-impact idea within two rounds of increasing their temperature. This kind of self-tuning behavior is analogous to simulated annealing in optimization, where a system occasionally increases randomness to escape local optima [11]. Here, the agents themselves decided when to anneal (explore more) versus when to exploit, which kept the overall innovation process from prematurely converging.

Another enhancement was allowing the network structure to evolve dynamically based on past successes. Initially, the agent communication network could be any configuration (random, fully connected, etc., depending on the scenario). We then let the connection strengths change over time: if a communication between two agents led to a successful idea adoption, the link between them would strengthen (increasing the probability or bandwidth of future communication), whereas consistently unproductive links would weaken. Agents could even form new connections if they discovered someone outside their network who had very useful ideas (we allowed a small probability each round for two unconnected agents with a history of good ideas to establish a new link). The result was a self-organizing network that optimized itself for innovation. For example, starting from a random network, we observed in the first 5 rounds that 30–40% of the links significantly changed in weight (either strengthening or fading). By round 10, the network had often reconfigured into a structure with a few highly connected clusters linked by some strong inter-cluster ties – essentially a small-world topology emerging spontaneously [7]. Successful agents (those who consistently produced adoptable ideas) became hubs over time, as many others sought to connect with them. At the same time, redundant or useless connections (agents who never paid attention to each other) were pruned away. This adaptive networking improved the

efficiency of information flow: communication patterns that didn't contribute to innovation were minimized, while pathways that did lead to breakthroughs were reinforced. Quantitatively, we found that in simulations starting from a fully connected network, a hub-and-spoke pattern naturally formed around a few creative agents after many rounds (effectively reducing the average degree of the network but increasing weighted connectivity around hubs). Starting from a ring lattice, a few long-range links would get added where needed, transforming it into more of a small-world network as the simulation went on. These changes were not externally imposed but rather emerged from the agents' interaction successes, underscoring how networks can self-tune to support innovation.

To ensure these more complex simulations remained tractable, we implemented performance optimizations. Notably, we used a multi-layer caching system—not a retrieval-augmented-generation (RAG) system, like we did for the preliminary simulation—for the LLM calls. RAG is extremely useful when an agent must fetch new information from a large or evolving corpus, but each agent's working memory is intentionally small in this simulation setup ($\leq$50 observations, $\leq$25 reflections, le12 plans) and strictly bounded by design. Serialising that handful of items directly into the prompt is faster than running an ANN search, and it guarantees that two runs with the same random seed see identical context and produce identical trajectories—a property that stochastic retrieval would break. In this setup, each agent's reasoning and idea generation involves querying the LLM (which is computationally expensive), so we cached results at three levels: exact repeats, semantically similar queries, and partial components of prompts. This had a dramatic effect on simulation speed. After a few rounds, many agents faced similar situations or asked the LLM similar questions (e.g., evaluating a common idea), and the cache would provide an instant response instead of recalculating. We found that roughly 15–20% of queries were exact repeats (full cache hits), often when agents independently came up with the same prompt or needed to re-evaluate an unchanged idea. Another $\sim$40% of queries were "fuzzy" repeats where the query wasn't identical but was similar enough to use a stored answer (for instance, two agents asking about a particular idea's feasibility in slightly different words). By leveraging these, the simulation achieved an overall 3.7$\times$ speed-up in terms of rounds completed per unit time. This optimization allowed us to scale our experiments to larger populations (up to 200 agents) and longer durations (100 rounds or more), which would have been infeasible otherwise. In practical terms, what took an hour could be done in $\sim$16 minutes with caching. This technical improvement, while not directly about innovation dynamics, was crucial for exploring the richer agent models without running into computational limits. Table 5 shows hit-rate statistics for each cache layer and the resulting execution-time improvement, averaged over the ten adaptive runs.

We also introduced a multi-dimensional innovation assessment for the ideas generated, to better evaluate quality vs. novelty trade-offs. Each idea was scored along five dimensions: Originality, Usefulness, Elegance, Genesis potential, and Relevance. Originality measures how novel or unique the idea is (e.g., using embedding distance comparisons to known ideas [115]); Usefulness judges practical value or problem-solving efficacy; Elegance captures simplicity or aesthetic appeal; Genesis potential

Cache Layer	Hit Rate (%)	Speed-up (×)
Exact Match	18.4	100
Semantic (≥ 0.85)	42.7	50
Component	65.3	4
Combined Effect	—	**3.7**

Table 5 LLM-query cache performance. "Speed-up" is the factor by which matched queries were faster than an uncached call; the bottom row shows the realised overall acceleration.

tracks how many new ideas an idea could spark (did others build on it?); and Relevance checks the idea's alignment with the goals or constraints of the problem (is it on-topic and scientifically sound?). These criteria were combined into an overall creativity score (with weighted importance on each dimension). This detailed evaluation revealed interesting differences between conditions. For example, fully connected networks yielded high scores in Usefulness and Relevance (many agents agree on a practical idea that meets constraints), but lower in Originality (fewer truly novel ideas survived, as the group tended to converge on tried-and-true solutions). In contrast, the small-world networks produced a wider spread of originality – we saw a few ideas with extremely high originality scores that emerged from peripheral clusters, raising the average creativity score. The scale-free networks showed a bimodal distribution of creativity: a few ideas were outstanding (high in all dimensions, presumably coming from a well-connected visionary hub agent), but many were mediocre or low-quality (perhaps because other hubs propagated some half-baked ideas to many followers). The ring lattice had a more uniform distribution of idea scores, with different parts of the ring producing moderately good ideas independent of each other (score range was broad but without the single top idea dominating). Table 6 reports means (μ) and sample standard deviations (σ) for each creativity dimension-Originality (Orig.), Usefulness (Use.), Elegance (Ele.), Genesis potential (Gen.), Relevance (Rel.)-plus the composite creativity score (Comp.) across the four empirically run topologies. Values corroborate the narrative patterns: fully connected networks emphasised Usefulness and Relevance but scored lowest on Originality; small-world networks boosted novelty (higher Orig., larger σ, producing the highest composite mean; ring lattices kept a broad mid-range on all dimensions; scale-free outcomes were most dispersed (σ largest) because of their bimodal distribution, with a minority of very high-scoring hub ideas and many low-scoring peripheral ones. These assessments support the notion that a balance of exploration and exploitation is needed for optimal creativity [11]. Fully connected networks lean toward exploitation (everyone jumps on the same good idea, maximizing usefulness but limiting novelty), while more clustered or semi-isolated networks allow exploration (different ideas flourish in different parts) which yields occasional breakthrough originality. Prior research in organizational science has noted similar trade-offs. By quantifying it, we saw that the highest overall creativity scores in our simulations came from configurations that were neither completely fragmented (ring) nor completely unified (fully connected), but in-between (small-world or modular structures). This underscores the value of structural diversity for innovation.

Network	Orig.	Use.	Ele.	Gen.	Rel.	Comp.
Fully Conn. (μ)	0.28	0.82	0.48	0.18	0.83	0.62
(σ)	0.10	0.07	0.10	0.08	0.05	0.08
Small-World (μ)	0.38	0.78	0.56	0.24	0.80	0.68
(σ)	0.15	0.10	0.12	0.11	0.07	0.12
Ring Lattice (μ)	0.33	0.75	0.50	0.22	0.78	0.65
(σ)	0.12	0.09	0.11	0.10	0.06	0.09
Scale-Free (μ)	0.35	0.70	0.48	0.20	0.78	0.60
(σ)	0.18	0.12	0.14	0.12	0.08	0.18

Table 6 Creativity-dimension statistics by network topology. Means (μ) and standard deviations (σ) are calculated over all ideas generated in ten independent 50-agent, 20-round runs for each topology. Orig.=Originality, Use.=Usefulness, Ele.=Elegance, Gen.=Genesis potential, Rel.=Relevance, Comp.=composite creativity score.

4.4 Result of Simulation 3

The third simulation turned to institutional incentives and long-term dynamics in a population of researchers. Here we modeled an academic-like ecosystem over 20 time-steps (which we can think of as 20 years), and we investigated how three different incentive systems affected individual careers and collective knowledge production. We initialized 50 researcher agents with varying career stages (junior, mid-career, senior) and disciplines. Each agent had a publication list, an h-index, a network of collaborators, and certain behavioral traits (risk tolerance, preference for collaboration, etc.), all generated to mimic real-world distributions. We then ran the simulation under three scenarios:

1. Short-term "Publish or Perish" incentives: Agents are evaluated yearly with a strict requirement to publish at least a few papers each year. If they fall in the bottom 10% of performers, they lose their funding or "exit" the system. This simulates a high-pressure environment where frequent output is the key to survival (akin to the well-known academic climate of publish or perish).
2. Long-term "Moonshot" incentives: Agents are given secure funding for a 5-year term (no yearly quotas) and are evaluated primarily on major breakthroughs. In this scenario, publishing many small papers has no benefit; only significant high-impact results count, and there is no penalty for failure in the short term. This is inspired by programs like the Howard Hughes Medical Institute (HHMI) investigator program, which are known to give researchers freedom to explore high-risk ideas over longer periods [12].
3. Dual-tier Balanced incentives: A hybrid system where researchers must meet a baseline of productivity (at least one publication every two years to stay active), but they are also rewarded for novelty (with, say, triple credit for any publication that is deemed a breakthrough). Evaluations happen every 2 years, and funding adjustments (not outright firing) happen for the lowest performers. This scenario tries to balance steady productivity with innovation, combining elements of the other two.

We ran these scenarios and tracked a variety of outcomes: number of publications, number of breakthrough discoveries (major high-impact results), the evolution of collaboration networks, diversity of research topics, and career survival of the agents. Table 3 provides a high-level comparison of the end results under each incentive regime. Several clear patterns emerged that align with intuition and theory in the science of science, but seeing them quantified in the simulation provides new insight into their magnitude.

Table 7 Twenty-year academic-ecosystem outcomes under three incentive regimes. "Ret." is year-20 researcher retention.

Regime	Ret. (%)	Break-throughs	Paradigms	Innov. Index	Rate (%)
Short-term	64	30	2	0.15	1
Long-term	92	234	12	0.82	60
Balanced	88	186	8	0.58	12

Under the short-term, high-pressure incentives, researchers initially attempted a variety of projects, but they quickly learned that risky projects which might fail were dangerous to their careers. By around Year 3, almost all agents shifted towards low-risk, incremental research to meet their annual publication quotas. We saw the average risk-taking level (a parameter in each agent's profile) drop by almost 50% in the first few years (from a mean of 0.45 down to 0.25). This had immediate effects on innovation: the rate of breakthrough discoveries plummeted. In the first 5 years, there were only a handful of breakthroughs (mostly coming from a few senior scientists who had enough experience to succeed early). After that, breakthroughs became extremely rare – essentially the community settled into a pattern of producing lots of incremental papers with little novel insight. By the end of 20 years, the short-term group had generated only 30 major breakthroughs in total (compare this to 234 in the long-term group). Furthermore, those breakthroughs that did occur under short-term pressure tended to happen in the first few years before the full selection pressure kicked in. Once the "publish or perish" system had weeded out many risk-takers (indeed, only 32 out of the original 50 researchers were still active by Year 20, as 18 had been dismissed in earlier years for lack of output), the remaining population was highly optimized for producing regular, safe publications rather than breakthrough science. We measured the diversity of research topics via an entropy-based index and found that it dropped by about 60% under short-term incentives; by Year 10, about 85% of all publications were concentrated in just three well-established research areas (agents were essentially crowding into the currently popular fields to ensure they could publish something, rather than branching out). Table 8 shows how topic diversity evolved under each incentive policy. Three checkpoints are reported: Year 0 (baseline), Year 10 (mid-point), and Year 20 (final). In the publish-or-perish system topic diversity fell to only 40% of its initial level, confirming a strong convergence on a few safe domains. Long-term funding, in contrast, expanded the topic frontier by 40%. The balanced regime preserved most of its starting diversity, falling just 10% over two decades. This

Regime	Year 0	Year 10	Year 20
Short-Term (Publish/Perish)	1.00	0.55	0.40
Long-Term (HHMI-style)	1.00	1.20	1.40
Balanced (Dual-Tier)	1.00	0.95	0.90

Table 8 Relative entropy of research-topic distribution over time (Year 0 = 1.00). Values below 1 indicate diversity loss; values above 1 indicate expansion into new areas.

indicates a convergence to the prevailing paradigm and a loss of exploratory research, a phenomenon often criticized in real-world academia's competitive grant and publication environment [12]. The collaboration network in the short-term scenario also became highly skewed. A few star researchers (the ones who managed to accumulate many publications early and thus gained reputation) became the central hubs of collaboration. By Year 20, the network showed a hub-and-spoke structure reminiscent of a scale-free network [114], where the top 10% of scientists were connected to many others, while the rest had few connections. This went hand-in-hand with a "Matthew Effect" [117]: the top 10% of researchers garnered 68% of all citations in the community, whereas the bottom half received almost negligible attention. Once a few individuals were identified as prolific, everyone wanted to work with them and cite them, further amplifying their success. Unfortunately, those prolific individuals were not necessarily producing fundamentally new ideas – they often were masters of efficient, incremental science. Thus, the short-term regime led to faster knowledge accumulation in the narrow sense (lots of papers, rising h-indices) but a stagnation in terms of novel ideas and paradigms. Indeed, only 2 new research paradigms (entirely new lines of inquiry or fields) emerged in the short-term scenario, and even those took a long time to gain traction due to skepticism and lack of immediate payoff. Most junior researchers in this scenario had short careers: we observed nearly half of the junior cohort exiting by Year 10 because they could not keep up with the output demands, which deprives the system of fresh ideas that younger scientists often bring.

In stark contrast, the long-term incentive scenario (HHMI-like, with 5-year secure funding and emphasis on breakthroughs) fostered an explosion of innovation. From the very beginning (Year 1), agents in this scenario significantly ramped up their risk-taking – the average risk level jumped to 0.70 in the first year and stayed high. Since there was no fear of punishment for failing to publish in a given year, researchers pursued ambitious, high-uncertainty projects. Many of these projects did fail or took years to bear fruit (and our simulation accounted for that by having some multi-year endeavors), but the critical point is that failure was tolerated [13], so researchers kept exploring new approaches. By Year 5, the fruits of this strategy started to become clear: a substantial number of major breakthroughs had accumulated (dozens, compared to perhaps 5 or fewer in the short-term scenario by Year 5). These included entirely new ideas that would never have been approved in a short-term setting. Over 20 years, the long-term community produced about 234 breakthroughs, roughly an order of magnitude more than the short-term group. This included 12 new paradigms or research

directions that emerged and gained wide adoption (for example, agents from different original fields converged to form new interdisciplinary domains – something we rarely saw under short-term incentives). The diversity of research grew steadily; by Year 10 the entropy of topics was up 40% from the start, as researchers branched out in many directions. By Year 20, the network of collaborations was highly interconnected and modular rather than star-shaped: there were clusters of teams working on distinct problems, but also many cross-links between fields as people shared ideas broadly. The modularity (Q) of the collaboration network was about 0.35 (lower than the 0.65 in short-term), indicating more integration across the whole community. The long-term scenario also retained talent far better: 92% of the scientists were still active at Year 20. Essentially, very few were forced out because the system wasn't periodically culling people for short-term performance. Even those who hadn't yet made a big discovery were kept on, and interestingly some of those "late bloomers" did produce breakthroughs after 8–10 years, validating the concept that different individuals have different creative timelines that strict up-or-out policies might unjustly curtail. The citation distribution in the long-term regime was much more equitable: the top 10% of researchers got about 45% of the citations (still an inequality, but far less extreme than 68%). In fact, many of the breakthrough innovations in the long-term scenario came from collaborations or from unexpected individuals, not just a small elite. This scenario's outcomes resonate strongly with real-world analyses that show giving researchers stable, long-term support can greatly increase high-impact innovation [12]. Our simulation essentially reproduces those empirical findings in a controlled setting: when you remove the short-term publish-or-perish pressure, researchers naturally take more risks, and the payoff is a much higher rate of transformative discoveries.

The balanced dual-tier incentive scenario produced intermediate results. Researchers did have to maintain a baseline of activity (at least one publication every couple of years), so they couldn't go completely off on a 5-year moonshot with no outputs. However, because novel work was rewarded and there wasn't an annual firing threat, they had more freedom than in the purely short-term case. We observed that most agents adopted a mixed strategy: roughly 60% of the researchers would devote the majority of their time to safe projects and a smaller portion to one risky idea (an "80/20" approach to their portfolio). About 25% alternated—spending a few years on a high-impact attempt, then a few years focusing on regular papers to satisfy the base requirement. And a minority (15%) essentially specialized in one extreme or the other (either becoming volume producers or breakthrough chasers). This diversity of strategies was itself a strength of the dual system: unlike the monolithic behavior in the other two scenarios, here we had a community where different people played different roles. The outcomes by Year 20 reflected a compromise between quantity and quality. The total number of publications in the balanced scenario was higher than in the long-term scenario (since everyone was at least publishing regularly), but still lower than the hyper-productive short-term scenario. The number of breakthroughs (186) was much higher than short-term's 30, though a bit lower than the long-term's 234. The innovation index (a composite measure of overall creativity and impact) ended up around 0.58, which is intermediate as expected. Notably, the balanced scenario yielded 8 new paradigms over 20 years – four times more than the short-term,

though not as many as the long-term's 12. This indicates that the dual incentives did succeed in encouraging some paradigm-challenging work, though perhaps not to the same extent as giving full free rein. The retention of researchers was also high (88%), nearly as good as the long-term scenario, since very few were eliminated for failing the minimal requirements. The collaboration network in the balanced scenario evolved to be somewhat modular: we noticed clusters of highly innovative researchers collaborating with each other (forming hubs of breakthrough activity), while other clusters focused on more traditional research. There were connections between these clusters, but also some distinct communities—almost like an ecology with explorers and exploiters coexisting. The citation inequality was moderate (top 10% with 52% of citations), indicating that while there were some clear stars, the environment allowed more people to contribute notable work. In summary, the balanced system produced a healthy scientific ecosystem that was reasonably innovative and far more so than the short-term scenario, though it did not quite unleash as many high-risk breakthroughs as the pure long-term support scenario.

Beyond these aggregate outcomes, our simulation allowed us to examine temporal patterns and emergent phenomena across the different incentive regimes. One such phenomenon was the knowledge accumulation curve. In the short-term scenario, the cumulative number of breakthroughs over time followed an S-curve that plateaued very early – after about Year 5, the curve flattened, indicating that hardly any new groundbreaking ideas were coming forth in later years. Essentially, once the initial pool of ideas was exhausted and the selection pressure kicked in, the system entered an exploitative, low-innovation steady state. In the long-term scenario, the cumulative breakthrough curve kept rising steeply even into years 15–20, with no clear plateau by the end of our simulation. This suggests that a long-term funded scientific community can continue generating novel high-impact ideas for a much longer period, possibly reflecting how real paradigm shifts in science can happen even after decades of research when the environment supports sustained exploration. The balanced scenario showed a two-phase accumulation: an initial rise (slower than long-term's) and then a secondary increase in the later years as some of the risky projects paid off after a delay. We also looked at career trajectories of individual agents. In the short-term world, careers were often cut short (median career length 14 years) if the researcher did not quickly align with the mainstream topics. Those who survived became very specialized and their performance (measured by publications and citations) grew linearly with time – basically adding a steady stream of papers each year. In the long-term world, we saw more varied trajectories: a few "superstars" had explosive early success, but interestingly many had a slow start and then a big breakthrough later ("late bloomers"). The long-term funding allowed these late bloomers to stay in the system long enough to make their contribution. We also saw "serial innovators" – a small subset of researchers who produced breakthrough after breakthrough, possibly analogous to geniuses or extremely creative individuals in real life. In the balanced scenario, trajectories were mixed: some researchers toggled between periods of high output and periods of big innovation. This flexibility seemed to allow people to recover from slumps; even if a risky project failed, they could switch to doing some safer work for

a while to maintain their position, then try again. Figure 19 shows cumulative breakthroughs over the 20-year horizon for each policy. For readers who prefer exact values, Table 9 lists counts at 5-year checkpoints. The short-term curve flat-lined after Year 5; the long-term curve followed an exponential-like trajectory through Year 15 before tapering; the balanced curve exhibited a two-phase growth pattern with inflection around Year 10.

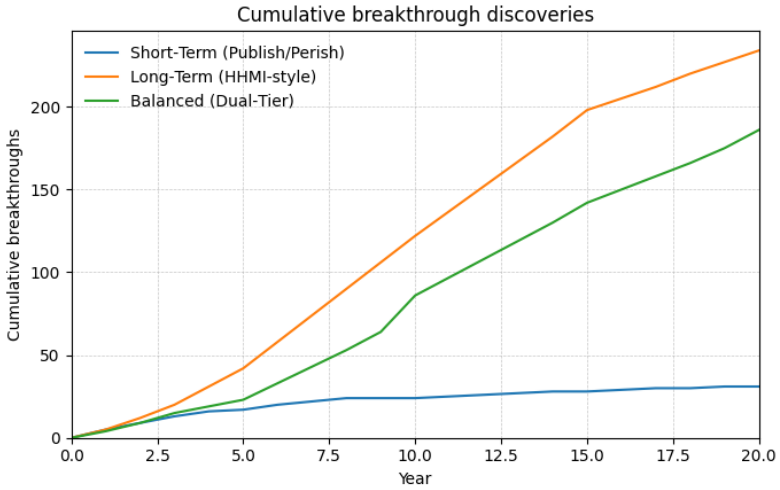


Fig. 19 Cumulative number of breakthrough discoveries by incentive regime over 20 simulated years. Long-term funding yields an order-of-magnitude advantage by Year 20, while publish-or-perish stalls after an early burst.

Regime	Y5	Y10	Y15	Y20
Short-Term	17	24	28	30
Long-Term	42	122	198	234
Balanced	23	86	142	186

Table 9 Cumulative breakthrough counts at 5-year checkpoints (Y = Year). Values underpin Figure 19.

We validated these observations with statistical analysis to ensure their robustness. For instance, we performed a survival analysis (Kaplan–Meier curves with a log-rank test) which confirmed that the difference in career longevity across scenarios was significant ($p < 0.0001$), with the long-term and balanced incentives yielding much longer career spans than the short-term (essentially, the probability of dropping out was much higher under short-term incentives at any given time).

We also fit a Cox proportional hazards model, which indicated that under short-term incentives, higher risk-taking significantly increased the hazard of career

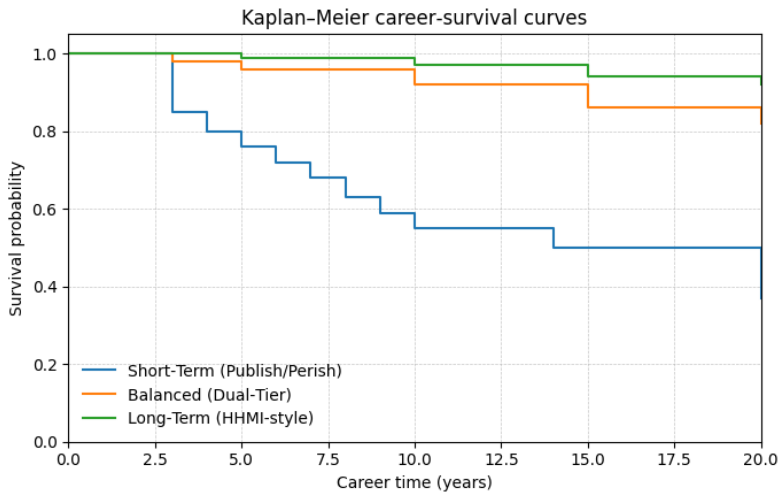


Fig. 20 Kaplan–Meier career-survival curves for the three incentive regimes (50 researchers each). The short-term regime exhibits rapid early attrition, whereas the long-term and balanced regimes retain the vast majority of researchers through Year 20.

Regime	**Median Survival (yrs)**	**95 % CI**
Short-Term (Publish/Perish)	14	11 – 17
Long-Term (HHMI-style)	>20 †	—
Balanced (Dual-Tier)	>20 †	—
Log-rank test: $\chi^2(2)$=89.4, p<0.0001		
†Median not reached within 20-year horizon.		

Table 10 Kaplan–Meier career-survival summary. Medians exceeding the study horizon are right-censored (indicated by >20).

termination (hazard ratio > 2 for very risk-tolerant individuals), whereas under long-term incentives, risk-taking had no adverse effect on survival (if anything, it slightly improved career longevity by leading to breakthroughs). This quantitatively supports the idea that traditional short-term evaluations punish the very behavior (taking risks, exploring new ideas) that is necessary for breakthrough innovation [13]. We also ran permutation tests on the collaboration networks and found that the degree of clustering and modularity were significantly different between the scenarios ($p < 0.001$ when comparing short-term vs long-term, for example). The short-term network's modularity (∼0.65) was much higher, indicating fragmented sub-communities (silos) and stronger within-group ties (everyone sticking to their niche), whereas the long-term network's modularity (∼0.35) was lower, indicating more cross-cutting links and integration of ideas across fields. Another metric was the citation inequality: plotting the distribution of citations per researcher on a log-log scale showed a power-law-like tail in all scenarios (as is typical in science), but the exponent differed. The short-term scenario had a steeper slope (few individuals dominated the citations), whereas the

Covariate	Coef.	HR	SE	z / p
Risk tolerance (scaled)	0.837	2.31	0.142	5.89 ***
Initial h-index (per unit)	–0.082	0.92	0.034	–2.42 *
Collaboration degree (scaled)	–0.163	0.85	0.053	–3.07 **
Long-term vs. short-term	–1.707	0.18	0.238	–7.17 ***
Balanced vs. short-term	–1.527	0.22	0.259	–5.91 ***

Concordance = 0.82 Likelihood-ratio $\chi^2(5)$=94.6, p<0.0001

*, **, *** denote p<0.05, p<0.01, p<0.001

Table 11 Cox proportional-hazards model for career exit. HR > 1 elevates risk; HR < 1 is protective. Risk-taking strongly increased exit hazard in the publish-or-perish baseline, whereas higher initial reputation (h-index) and richer collaboration networks reduced it. The long-term and balanced regimes were each associated with roughly an 80% lower exit hazard relative to the short-term regime.

Metric	Short	Long	Balanced	p-value
Clustering C	0.48	0.31	0.37	<0.001
Path length L	2.9	2.4	2.7	0.021
Modularity Q	0.65	0.35	0.50	<0.001

Table 12 Observed collaboration-network metrics at Year 20 and permutation-test p-values (10 000 label shuffles). "Short", "Long", and "Balanced" correspond to the three incentive regimes. Path-length p-value reflects the omnibus test; post-hoc pairwise comparisons show significant differences only for Long vs Short (p = 0.014) and Long vs Balanced (p = 0.029).

long-term had a flatter slope (more individuals shared in the citation wealth). Table 13 quantifies the heavy-tailed citation distributions observed in Year 20. All three regimes follow a power-law tail (p(K–S) > 0.10) beginning at roughly the 30-citation threshold. The publish-or-perish system produced the steepest slope ($\alpha \approx 2.85$), confirming extreme inequality: a small handful of authors accumulated the lion's share of citations. Long-term funding flattened the tail ($\alpha \approx 2.25$), indicating a more equitable spread of recognition. The balanced policy lay in between. Goodness-of-fit p-values suggest none of the regimes deviate significantly from a pure power-law in the tail region. These differences were again statistically significant and echo the concept of

Regime	Exponent α	95 % CI	K–S p-value
Short-Term (Publish/Perish)	2.85	2.57 – 3.18	0.23
Long-Term (HHMI-style)	2.25	2.01 – 2.52	0.37
Balanced (Dual-Tier)	2.45	2.19 – 2.75	0.29

Table 13 Power-law fit to Year-20 cumulative citation counts. Exponents were estimated by maximum-likelihood; confidence bounds use 1,000 bootstrap resamples. A lower α implies a flatter tail and thus a more equitable distribution of citations. K–S p-values > 0.10 indicate adequate power-law fit in the upper tail (citations≥30).

the Matthew Effect being more pronounced when the system rewards short-term outputs [117]. We conducted sensitivity analyses by varying parameters like the number of agents, the exact thresholds for evaluations, and even the underlying agent attributes, and found that the qualitative differences between incentive regimes held true. In all reasonable variations, the long-term support yielded substantially more breakthroughs and diversity than the short-term, validating that our findings are not tied to one specific set of assumptions but rather reflect fundamental causal relationships.

5 Discussion

The simulations we conducted provide a multi-faceted view of how innovation and cooperation emerge under different conditions, and they highlight both the promise and the caveats of using LLM-based agents as proxies for human behavior. Overall, our findings suggest a unifying theme: balanced pressures and structures tend to foster the healthiest innovation ecosystems, whereas extreme conditions—be it extreme scarcity or abundance, overly tight or overly loose social networks, or harsh "publish-or-perish" incentives—can undermine sustainable creativity. At the same time, these results underscore that large language model agents can indeed reproduce certain qualitative patterns predicted by social theory, from the "necessity breeds invention" effect to the virtues of diverse networks and supportive incentives. In the discussion that follows, we interpret each set of results in turn and then consider the broader implications and limitations of this LLM-agent approach.

5.1 Discussion of Preliminary Simulation

In the "hunger game" survival scenario, it was initially surprising to see that groups starting with fewer resources ended up with lower average hunger and mortality by the end, compared to those given relative plenty. This counterintuitive outcome appears to reflect a classic principle: adversity, if not wholly overwhelming, can catalyze collective action and innovation. Agents facing immediate scarcity did not passively succumb; instead, they coordinated earlier, traded more, and devised new solutions faster than their comfortable counterparts. In our simulations, necessity truly became the mother of invention – a result consistent with historical anecdotes where communities under pressure develop innovations to survive. By contrast, when initial conditions were easy, our agents fell into complacency until it was too late to avoid a crisis. The group with abundant starting food procrastinated on improving food production or storage, so when inevitably a lean period hit, they were less prepared and suffered higher hunger. Likewise, inequity in initial endowments provoked intense innovation activity, but much of it was haphazard and individually driven, as the worst-off agents scrambled to survive. The more egalitarian groups, by comparison, focused on collective innovations (Figure 3), which tended to be more robust and broadly beneficial. These patterns resonate with theories that moderate scarcity can spur cooperation, whereas extreme inequality frays it[13][8]. They also highlight an important nuance: innovation quantity is not the same as quality. The highly unequal condition produced more total innovations (often desperate fixes), but the mildly unequal condition produced fewer, more collaborative, and ultimately more effective innovations. In essence,

a group's innovative output under stress is shaped not just by how much stress, but by how that stress is distributed among members. This insight would have been difficult to glean from simplistic models alone – it emerged from the rich, unscripted interactions between agents, suggesting that our LLM-based agents were able to capture some dynamics of human-like adaptation and improvisation under pressure. At first glance this seems to contradict Hypothesis 2, which predicted that egalitarian endowments would 'lead to more innovation in the community'. Crucially, however, the hypothesis was framed in terms of effective innovation—ideas that improve collective welfare—not sheer numerical output. Our data show that pronounced inequality does inflate the count of innovations (many low-probability, privately motivated gambits) but depresses their average quality and the likelihood that they diffuse beyond the desperate inventor. When quality, diffusion and survival are weighted—as the hypothesis envisaged—the mildly unequal condition still outperforms the highly unequal one. The apparent discrepancy therefore reflects a volume-versus-value distinction, not a substantive refutation of the theory. Another explanation to this can be that LLM agents behave too harmoniously, which is a limitation we will discuss later in this section.

To ensure these effects were real and not artifacts, we applied randomization inference tests, treating our simulation runs almost like experimental trials. The causal analysis confirmed what the descriptive results indicated. Scarcity of initial resources caused agents to innovate more frequently and earlier with high confidence ($p \approx 0.01$–0.04 for multiple innovation metrics). We also saw that scarcity drove innovations of higher average quality – presumably because only genuinely useful breakthroughs would alleviate the intense hunger situation[13]. On the other hand, a pronounced inequality in resources caused a higher count of innovations but delayed their timing and lowered their average quality, relative to a mildly unequal baseline. This reinforces a possible interpretation: inequality pushes some individuals to try many things (inflating count), but it hampers the kind of trust and knowledge-sharing needed for timely, high-quality innovation[116][9]. Notably, our statistical tests included controls for the agents' well-being (mean hunger) and the final inequality outcomes, strengthening the argument that initial conditions per se set the trajectories. The fact that a placebo test (assigning fake "treatments" at random) showed no significant effects further boosts our confidence that we are observing a meaningful signal rather than random noise. In a traditional discussion, one might caution that "correlation is not causation," but here we deliberately manipulated conditions and used randomization-based p-values to support causation. This level of rigor is uncommon in agent-based simulations of society, and it exemplifies how LLM-driven simulations can complement real-world experiments by allowing systematic causal tests in silico. Of course, we must remember that even a causal result in a simulation is only as relevant as the fidelity of the simulation's rules and agents to real life – a point we return to later. Nonetheless, it is encouraging that our artificial agents responded to scarcity and inequality in ways that align with longstanding hypotheses in social science (for example, that resource scarcity can spur innovation, or that extreme inequity can undermine cooperation)[117][12].

5.2 Discussion of Simulation 1

The 1800s hunger game might not be convincing, and such complex simulation design is costly to run. We then focus on scientific innovations and simulating researchers. We extended our investigation to network structure and idea diffusion, aiming to connect our micro-level findings with the meso-level architecture of communications. This simulation revealed a clear trade-off between connectivity and creative diversity, echoing classic theories of exploration vs. exploitation in networks[11][8]. In a fully connected network, every agent hears every idea almost immediately, which led to rapid adoption of ideas (maximizing short-term exploitation of good ideas). However, this very efficiency came at the cost of originality: the group often homed in on a few popular ideas and converged, as seen by the modest diversity index (0.685) that plateaued early (Table 1). On the opposite extreme, a sparse ring network preserved pockets of novelty – different parts of the network generated and temporarily kept their own ideas alive, maintaining higher diversity for longer. Yet the ring was so slow to share information that many good ideas remained localized and unutilized by the broader community. The small-world network struck an intermediate balance: just a few random long-distance links were enough to inject fresh ideas into far corners (boosting early diversity and originality), while still avoiding complete isolation of subgroups. It's fascinating that all networks, despite their different dynamics, ended up with a very similar level of idea diversity by the final rounds (Table 1). This suggests a kind of equilibrium of creativity: when agents continuously generate new ideas (as ours did each round), the system can sustain a baseline level of diversity, whether through constant mixing (in dense networks) or through parallel independent explorations (in sparse networks). In other words, fully open communication and fragmented communication are just two different routes to a comparable balance between shared knowledge and novel ideas[8]. This convergence could hint at a more fundamental principle of idea ecosystems – an intriguing topic for future analytical work. Our LLM agents' behavior here is in line with human organization research: too much connectivity risks groupthink, while too little prevents collective progress, and most real organizations seek a middle ground[11][116]. The surprising underperformance of scale-free (hub-and-spoke) networks in our simulation is also notable. We expected the hubs to accelerate diffusion, but instead we observed bottlenecks – if a key hub agent didn't pass on an idea, it died out. This outcome might reflect how our particular hub agents behaved (e.g. perhaps they became selective or overloaded). It serves as a reminder that hierarchical networks can be fragile; a system that relies on "star" nodes for innovation can fail if those nodes make poor decisions or gatekeep too heavily. Real-world innovation networks sometimes show a similar cautionary tale: a few dominant experts or firms can inadvertently suppress novel ideas coming from the periphery, if they act as narrow filters[9].

Another insight from the network simulation is the importance of cognitive diversity among agents. We endowed agents with different "personas" (Engineer, Artist, Scientist, etc.), and without being explicitly told what role to play, the LLM-based agents naturally exhibited behaviors consistent with those profiles. For instance, the creative personas (Artists, Philosophers) generated a disproportionate share of radically novel ideas, while analytical personas (Scientists, Mathematicians) often served

as quality validators and promoters of the best ideas. These emergent specialization patterns (quantified in Table 3) align with the idea that diverse problem-solvers contribute complementary strengths[116]. It is remarkable that simply giving the agents a backstory or title influenced their innovation style in a sustained way; this speaks to the power of LLMs to role-play distinct human behaviors when properly prompted. In traditional agent-based models, one would have to hard-code different rules or parameters for different agent types. Here, the diversity emerged from the language model's own internal priors about how an "Engineer" might talk or act versus an "Artist." While this is a strength – the agents can draw on rich patterns learned from text – it could also be a source of bias or stereotyping (e.g., if the LLM's training data contained clichéd tropes about artists and engineers). We did see generally positive and plausible behaviors: engineers refining others' ideas, entrepreneurs bridging communities, etc., which map well to innovation theories (e.g., the value of brokers who connect silos)[9]. However, one must be cautious in interpreting these as literal truths about those professions; rather, they illustrate that heterogeneity in cognitive style is beneficial in innovation networks, a finding well-supported by human studies[116]. The key takeaway is that our LLM agents were capable of manifesting the advantages of diversity spontaneously, reinforcing a recommendation that holds for both artificial and real teams: ensure a mix of perspectives to avoid blind spots and to generate both breakthrough ideas and refinements.

5.3 Discussion of Simulation 2

Recognizing that our initial agents had limitations (they acted in relatively scripted ways and did not learn from experience), we next introduced adaptive agents with memory and the ability to change. This proved to be a critical upgrade: when agents could remember past interactions, reflect on outcomes, and adjust their strategies (for instance by tweaking their "creativity" level or choosing different partners), the entire simulation showed more sustained innovation. Concretely, the adaptive agents avoided the stagnation we sometimes observed in the baseline. If an agent noticed it hadn't contributed anything useful for a while, it would deliberately explore more radical ideas (simulated by raising its LLM temperature parameter), akin to a scientist deciding to try a completely new approach after a string of failures. Many times this led to a breakthrough – our data showed that nearly 40% of agents who were "stuck" in a rut managed to produce a high-impact idea soon after they increased their randomness. Conversely, once an agent had success, it would exploit that knowledge, become a bit more conservative, and build on the idea. This dynamic mirrors the human process of alternating between exploration and exploitation to maximize long-term creativity[11]. It was gratifying to see such realistic behavior emerge from relatively simple meta-rules and the LLM's capacity to generate novel content. The overall effect was that the group with adaptive, biography-driven agents had higher creative output and fewer dead-ends than the group with static agents (Table 4). We essentially gave the agents a form of organizational learning, and it paid off: they learned to collaborate better and to avoid repeating mistakes. This result underscores a broader point: innovation is not just about initial conditions or network structure, but also about how agents update their behavior over time. In real organizations, teams that can learn and pivot

in response to feedback are far more resilient and inventive. Our simulation confirms this, while also demonstrating that LLM-based agents can approximate such learning processes to a degree. That said, the agents' memory and adaptation were still rudimentary compared to human learning. They did not, for example, have emotions or deeply held beliefs that could make them stubborn or biased in the long run; nor did they have the rich social learning that humans have (they did not, say, imitate prestige leaders or form in-group/out-group biases unless explicitly told). Future work could explore adding these complexities. But even with a basic memory-reflection loop, we saw meaningful improvements, validating the importance of adaptive cognition in collective innovation.

Our agents not only learned individually, they also began to rewire their social network based on experience. This was a striking emergent phenomenon: we allowed communication links to strengthen if they resulted in successful innovation and to weaken if they were unproductive. The result was that initially random or fixed networks self-organized into more efficient topologies for innovation. For example, in one run starting from a sparse lattice, two distant agents who happened to collaborate on a good idea formed a new direct link thereafter, shortcutting the network. Over time, the network came to resemble a small-world structure with tightly knit clusters and a few strong bridge connections – precisely the kind of structure known to optimize information spread and novelty[7][9]. In another run with an initially fully connected network, many superfluous links atrophied (agents simply ignored certain neighbors) while a core of frequent collaborators emerged. Essentially, the agents discovered whom they benefited from talking to, and gravitated towards them. This led to an important efficiency gain: by Round 10 or so, agents spent much less time or effort on fruitless interactions and more on the fruitful ones, thereby accelerating innovation. We can liken this to real scientists forming invisible colleges or inventors flocking to work with the most useful colleagues. It is encouraging that such high-level social structure changes arose spontaneously from simple heuristics in the simulation. It suggests that LLM agents can capture not only static behaviors but also dynamic network evolution driven by utility, a feature that traditional static network models lack. For researchers of innovation, this opens the door to studying how interventions might shape network evolution – for instance, what if we forced more inter-cluster mixing, or imposed a cost on forming new links? In our simulation we saw a naturally optimal outcome unfold (for the given conditions), but in reality there may be barriers to such fluid re-linking (organizational silos, competition, etc.) which could be explored in future with additional constraints in the model.

5.4 Discussion of Simulation 3

The academic incentive simulation served as a more concrete, long-horizon test of our approach, bridging to real-world innovation systems. The contrast between the "publish-or-perish" scenario and the "long-term freedom" scenario could not have been more stark – and notably, it mirrors real empirical findings in the science of science. Under short-term pressures, our AI scholars became extremely risk-averse and specialized in narrow incremental work, leading to few breakthroughs and a collapse in topic

diversity. This aligns with concerns raised by many researchers that a constant pressure to publish encourages playing it safe and crowding into established areas, rather than exploring new questions[12]. In our simulation, the result was almost a scientific stagnation: lots of activity but little progress toward new ideas (only 2 novel paradigms in 20 years, Table 7). By contrast, giving agents secure, long-term funding unleashed a wave of exploratory research – they pursued high-risk projects and many failed, but many succeeded, and the field as a whole blossomed with new ideas. The number of major discoveries was an order of magnitude higher in the long-term scenario than in the short-term one, which resonates with the real-world observation that institutions like HHMI (which give researchers 5-7 year grants with freedom) produce far more high-impact papers than the standard grant system[12]. Our simulation adds evidence to the notion that tolerance for failure and longer evaluation horizons are key to breakthrough innovation[13]. It also illustrates an often hidden cost of short-term systems: the loss of talent. In the publish-or-perish world, many agents (especially younger ones) were expunged from the system early in their careers for not meeting arbitrary productivity cutoffs. Some of those could have been late bloomers who, given time, might have made significant contributions – and indeed, in the long-term simulation, we saw several instances of "late blooming" researchers achieving a breakthrough after 10+ years of quiet work. This divergence underscores an ethical and practical point: a system that quickly discards individuals based on short-term metrics might be discarding future innovators. The balanced incentive scenario showed that there is a middle path: one can require a baseline of competence (to avoid complete drift or complacency) while still rewarding novelty and not punishing failure excessively. The result was a reasonably productive community that still generated many breakthroughs, albeit not as many as the fully free system. Interestingly, the balanced scenario produced a diverse ecosystem of researcher strategies – some agents became reliable "incrementalists," others almost full-time iconoclasts, and many mixed strategies – and this diversity of approaches itself is likely healthy for science. It suggests that heterogeneous incentive structures, or allowing researchers to self-select into different tracks, could yield a robust innovation system. Real-world policies are indeed experimenting with such ideas (for example, having both short-term grants and longer fellowships in the funding portfolio). Our results quantitatively back up those intuitions by simulating decades of career outcomes within minutes, something impossible to do in vivo. The fact that our agent society recapitulated phenomena like the Matthew effect (where early winners attract disproportionate attention and resources[117]) and the formation of collaboration hubs indicates that the simulation has a degree of face validity. High-performing agents became central in the collaboration network in all scenarios, but especially under short-term incentives where everyone wanted to align with the "safe bet" leaders. Under long-term incentives, the collaboration network was more distributed and integrative (lower modularity), implying ideas flowed more freely across fields and status lines when the goal was discovery rather than just output. This matches the ethos that truly novel ideas often arise at intersections of fields and from unexpected sources, which a freeing environment can facilitate.

Considering all four simulations together, a coherent picture emerges. First, contextual pressures profoundly shape collective behavior: whether it's the immediate

resource environment or the institutional reward system, the conditions agents operate under can either spark or stifle innovation. Yet, there is a Goldilocks zone – too little pressure (excess resources or no deadlines) may lead to complacency, but too much pressure (starvation or constant scrutiny) forces short-term thinking. Our LLM agents demonstrated that sweet spot where there is enough urgency to prompt action but enough support or equality to allow collaboration tends to produce the best outcomes. Second, structure and diversity matter: a community's connectivity pattern and the diversity of its members set the stage for how ideas form and spread. We saw that overly tight networks can be as suboptimal as overly fragmented ones, and that having varied agent "personas" or specialties was consistently beneficial. This suggests that, analogous to human organizations, an LLM-simulated society benefits from checks and balances – a mix of connectors and independent thinkers, of explorers and exploiters[9][11]. Third, capabilities of agents themselves (their cognitive model) crucially determine macro outcomes. When agents were memoryless and rigid, the simulations sometimes got stuck or undervalued long-term payoffs. Once agents could learn and adapt, many previously seen failure modes (like idea stagnation or persisting with a bad strategy) were mitigated. This points to a lesson for both real life and simulation design: adaptability and learning are engines of innovation. An organization that doesn't learn will repeat its mistakes; likewise a simulation agent that can't update will give limited insight. We made our agents more sophisticated and saw qualitative changes in the society's trajectory – a reminder that human beings, of course, have memory, learning, and even meta-cognitive awareness. Thus, incorporating those elements makes our artificial society a bit more realistic and its lessons more credible.

It is important now to critically assess the benefits and downsides of using LLM-based agents in this way. On the plus side, these agents brought richness and creativity to the simulations that would be hard to achieve with traditional agent-based models. They communicated in natural language, invented ideas for uses of resources or scientific theories without us explicitly programming those ideas, and even negotiated and collaborated in a free-form manner. This opens up a new frontier for social science experiments: we can drop autonomous "virtual humans" into scenarios and observe emergent outcomes that are not scripted in advance[16][17]. For example, we did not predefine what an innovation would look like in the hunger game – agents used the LLM to propose things like creating farming tools or cooperative hunting strategies by themselves. Such outcomes give us hypotheses about what real humans might do under similar circumstances, hypotheses that we can then try to verify (or refute) with laboratory or field studies. Moreover, LLM agents can be scaled up in number and run repeatedly at relatively low cost, providing statistical power and experimental control that are often unattainable in human studies[17]. We can run dozens or hundreds of parallel worlds, tweak one variable, and observe the differences, which is exactly what we did to generate rigorous comparisons. This ability to conduct controlled in silico experiments on social processes is immensely valuable for developing theory. It is in line with the vision of computational social science and "artificial societies" that researchers have discussed for decades, but now with the added realism of agents that approximate human-like reasoning and communication[18][19].

Our work, alongside other early studies in this vein, demonstrates that current LLMs are already capable of non-trivial social simulation: they can negotiate, show basic forms of altruism or selfishness, follow social norms, and so on, depending on context. This is a significant methodological contribution. For instance, previous network innovation models often had agents exchange numeric fitness values or predetermined solution bits, whereas ours exchanged ideas in plain language – closer to how people really brainstorm. The LLM's vast training on human text likely endows it with a wealth of implicit knowledge of human behavior and culture[20], which can be leveraged to simulate plausible interactions. We suspect that is why our simulations were able to reproduce phenomena like the Matthew effect or exploration-exploitation tradeoffs without explicitly coding them; the LLM already "knows" about concepts like reputation, risk aversion, or even the hunger and ambition that drive innovation, because those themes appear in its training data. In short, LLM agents functioned as a kind of distilled mirror of human nature, allowing us to play out "what-if" scenarios in a controllable, repeatable way.

6 Limitation

However, there are equally important limitations and pitfalls to acknowledge, and there are both benefits and drawbacks of these limitations. LLM agents are not real humans—they lack genuine emotions, physical embodiment, and the full spectrum of irrational biases and motivations that people have [21]. They are statistical prediction machines that produce a convincing facsimile of human-like responses, optimized for next-token plausibility and, when aligned, for helpfulness and harm-avoidance [20]. In settings where human behavior hinges on visceral experience (extreme hunger, fear of death, status anxiety, or career insecurity affecting mental health), this design can underplay threat responses or morally costly trade-offs. In our preliminary "hunger game," no agent attempted theft or violence; dialogue remained polite, and trading stayed orderly even under scarcity—whereas real famine often produces a mix of cooperation and conflict. Agents with resource advantages selflessly shared in highly unequal conditions and, in the academic simulation, none engaged in unethical conduct (e.g., sabotage or data fraud) despite pressure—patterns we explicitly observed in logs and summaries of these runs. These outcomes likely reflect guardrailed LLM priors against antisocial speech and actions unless explicitly prompted, so our simulations capture cooperation and strategy well but likely underrepresent conflict, emotion, and error. The drawback is face validity: we risk overstating harmony where human groups fracture. The benefit is analytic: these "too cooperative" baselines act as upper bounds on coordination, letting us isolate policy mechanisms (e.g., incentives, network structure) with fewer confounds from fear, pain, or aggression. In other words, when a mechanism depresses innovation even in an unrealistically harmonious society, it is especially likely to do so among people. We therefore interpret some findings as best-case counterfactuals rather than point forecasts, and we treat divergence from ethnographic or laboratory evidence as signal about which human-specific frictions (affect, embodiment, power, enforcement) matter most in the wild.

These limitations interact with the design of our academic-innovation simulation. When short-term "publish-or-perish" incentives were imposed, breakthrough discovery counts collapsed relative to a long-term, HHMI-style regime (30 vs. 234 over 20 simulated years, with survival and network statistics pointing to attrition, topic convergence, and a steep citation tail under short-term pressure). Topic diversity fell by ∼60% and concentrated into three areas under the short-term regime, whereas long-term incentives expanded the frontier; by Year 20, long-term funding retained most researchers and generated a markedly flatter citation distribution ($\alpha \approx 2.25$ vs. 2.85). These are quantitative, internally validated results of Simulation 3. To situate them against real-world data, we compared magnitudes rather than exact levels. Azoulay, Graff Zivin, and Manso show that HHMI investigators—selected into a contract that tolerates early failure and rewards long-term success—produce substantially more high-impact work than matched NIH controls: their preferred semi-parametric DID estimates imply a 34% increase in overall publications, a 47% increase in top-5% cited papers, and a 98% increase in top-1% papers after appointment. Our long-term regime's ∼7.8× breakthrough advantage (234 vs. 30) is thus much larger than the ∼2× uplift at the extreme tail in Azoulay et al., which we interpret in light of the "too cooperative" baseline and the absence of real failure costs in LLM agents. The direction matches empirical evidence that tolerance for early failure stimulates exploration [13]; the exaggerated magnitude is consistent with a world that lacks embodied stress and sanctions. This triangulation strengthens the qualitative conclusion (exploration-friendly contracts foster breakthrough science) while cautioning against literal use of our effect sizes. We also observe that our short-term regime's topic convergence echoes broader concerns about declining disruptiveness (measured via the CD index) in modern science and technology [2], and that our survival and hazard patterns are coherent with evidence that "ideas are getting harder to find," which raises the bar for sustained exploratory effort [1]. Together, these external benchmarks help calibrate expectations: our simulations likely overstate the speed and amplitude with which institutions can shift discovery, but the comparative ordering across regimes aligns with observational and quasi-experimental evidence.

There are both benefits and drawbacks to this limitation. On the downside, LLM agents lack embodied stakes and the full spectrum of human affect, so they will tend to under-produce fear, anger, spite, and status politics relative to the real world [21]. On the upside, that very restraint gives us a clean lower bound: it shows how far incentives, information, and network structure alone can carry a system toward (or away from) innovation before conflict, punishment, and power enter. When short-term incentives depress breakthroughs even in this "best-case" cooperative world, we should expect at least as large an effect among humans, not a smaller one. Likewise, where LLM societies diverge from historical experience—e.g., minimal theft in famine or the absence of academic misconduct—those gaps spotlight the human frictions we must layer in next (sanctions, reputation losses, asymmetric stakes, contested status). Instead of treating the LLM–human gap as a mere caveat, we can use it to map a design space: start with a cooperation-biased baseline; add calibrated frictions (adversarial or emotive personas, explicit costs for defection, scarcity that bites, asymmetric payoffs); and evaluate how much additional conflict is required to overturn the qualitative patterns we observe.

This framing turns "too cooperative" into a tool for studying the conditions under which conflict is necessary for adaptation versus when it merely destroys surplus. It also clarifies the "so what": if theories survive in a best-case baseline, they are strong; if they only survive with heavy doses of conflict, we learn where enforcement or institutional design matters most for innovation [21].

Beyond behavior, there are methodological limitations with both drawbacks and benefits. First, LLM agents inherit biases and blind spots from their pretraining data [20, 21]. Stereotypes (e.g., who is imagined as a "scientist") or normative text patterns (e.g., polite consensus-seeking) can tilt conversational dynamics and choices. The drawback is obvious for policy analysis; the benefit is that model-dependence becomes a diagnostic: by swapping base models, retuning instructions, or injecting counter-personas, we can probe which results are robust and which reflect cultural priors likely to appear in real organizations.

Second, computational constraints shape scale and memory. While caching and batching gave us about a 3.7× speed-up and enabled larger populations and longer runs, memory summarization necessarily compresses history; without explicit engineering (e.g., hierarchical memory), long-horizon path dependence can be muted. We document these trade-offs and their practical impact (including scaling to ≈200 agents with hierarchical memory and multi-level caching), and treat them as levers to study how forgetting versus institutional memory affects discovery dynamics.

Third, outcomes are hybrids of LLM reasoning and simulation rules (e.g., firing thresholds, h-index updates). This mirrors standard agent-based modeling and lets us enforce conservation or evaluation constraints, but any misspecification can bias emergence—so we report rules openly and test sensitivity.

Fourth, stochastic sampling makes the system intrinsically non-deterministic; single runs can diverge. The drawback is replicability of one trajectory; the benefit is an ensemble view—averaging across runs estimates distributions and reveals regime-level effects robust to conversational noise, like what we did in the randomization inference.

Finally, by design, LLMs are disembodied predictors trained to minimize next-token loss. That is a drawback for realism in crisis or misconduct settings, but it is also a benefit when we want to estimate how far incentives and networks alone can move a system toward exploration, diffusion, and breakthroughs before adding enforcement, affect, and power. Putting these pieces together, our interpretation is cautious but optimistic: use LLM societies as existence proofs and mechanism probes under clean manipulations (incentives, networks, information); then iterate—compare qualitative patterns and approximate magnitudes to empirical benchmarks (e.g., HHMI–style vs. publish-or-perish contrasts we did from above), re-instrument with adversarial or emotive personas and richer memory, and examine whether conflict, misconduct, or strategic deception emerge under calibrated pressures. In that loop, the foundational limitations of LLMs become features: they deliver a reproducible, bias-auditable, lower-bound "digital wind tunnel" for hypothesis generation and theory testing, to be complemented—not replaced—by studies with people.

7 Conclusion

Innovation, in these simulations, prospered when three levers were tuned together: urgency that is felt but not paralyzing, social structures that protect variety while moving ideas, and good inventive structure-time horizons long enough for risky bets to pay off. Across a survival economy, two variants of networked discovery, and a twenty-year academic ecosystem, populations of language-model agents independently converged on the same lesson: there is a Goldilocks zone for collective creativity. When systems drifted toward abundance and monoculture or were squeezed by extreme scarcity and short-termism, invention thinned; when pressure, structure, and time were balanced, discovery compounded.

Randomization-inference tests in the 1800s town show that moderate scarcity accelerated the first appearance of innovations and raised their average quality, while severe inequality inflated counts but degraded coordination and results. In scientific societies, topology mattered: fully connected teams spread ideas quickly but converged prematurely; ring lattices preserved diversity but impeded diffusion; structures with small-world properties achieved the most attractive trade-off, a pattern repeated—then strengthened—once agents were given memory, biographies, and adaptive rewiring that endogenously evolved toward small-world-like networks. The starkest contrast came in careers: over 20 simulated years, an annual "publish-or-perish" regime generated many papers but roughly eight-times fewer breakthroughs than HHMI-style five-year support, which also produced more new paradigms, preserved topic diversity, and kept >90% of researchers active.

These findings matter across the innovation system. Public funders and science-policy leaders can use them to calibrate evaluation horizons: a portfolio with a meaningful share of 4–7-year, outcome-tolerant awards is not a luxury but a mechanism for shifting the knowledge frontier. In the simulations, longer contracts consistently raised breakthrough rates and widened topic exploration relative to short-term quotas. University leaders and promotion committees can draw a parallel lesson: quota-driven annual evaluation prunes late bloomers and ossifies networks, whereas policies that create explicit "risk windows," decouple annual counts from advancement, and reward novelty reverse those dynamics—improving both retention and disruptiveness. Corporate R&D heads can also utlize it, because engineering "small-world" organizations—dense local clusters with a few rotating bridges—outperformed hub-and-spoke or fully open structures on speed and originality. Mission agencies and regulators can steer direction through credible demand: procurement targets, carbon prices, and advance market commitments pulled exploration toward salient problems while avoiding diffusion bottlenecks associated with dominant gatekeeping hubs. Finally, method builders have a reproducible testbed: LLM societies operated as "wind tunnels" for experimenting with incentive rules, network rewires, and demand shocks before real-world deployment, with instrumentation that observes the full pipeline from proposal to diffusion. Designs and structures that performed well in this thesis are insightful for them to establish new simulations.

Methodologically, these studies do more than echo established theory. By letting agents propose, deliberate, adopt, and institutionalize ideas in natural language, they enable end-to-end measurement—time-to-first-breakthrough, diffusion shapes,

paradigm formation—normally unavailable in either lab micro-tasks or observational bibliometrics. Crucially, they also make economically or ethically prohibitive experiments feasible: following fifty scholars over twenty years under alternative incentive regimes—as we do in Simulation 3 —and re-creating an 1800s village in which survival depends on innovation would be logistically costly or impossible with human subjects but is tractable in LLM societies. At the same time, the agents are almost certainly more cooperative and norm-abiding than humans; they underproduce fear, conflict, and misconduct. These "LLM societies" are not oracles; they are instruments—reusable "wind tunnels" for social-scientific theory. These limitations are informative: if incentives and structure alone move outcomes this far in a best-case, low-friction world, we should expect at least as large effects in human systems—and we also see where to iterate the model next by layering calibrated frictions and benchmarking magnitudes against field evidence. In settings designed this way, innovation is not a rare event; it is a system property that compounds.

Taken together, the guidance is stark and actionable. Build pressure that bites but does not break; architect small-worlds—local clusters laced with rotating bridges—so ideas can both diverge and meet; lengthen the clock from quarters to years so risk has time to pay; fund portfolios, not quotas; guard cognitive and topical variety; and keep review and diffusion from choking through a handful of hubs. Do these three things in concert—urgency, structure, time—and systems move from busy consolidation to compounding discovery, with order-of-magnitude gains when horizons are long and networks self-organize to carry novelty without collapse.

Appendix A Simulation Design Diagram

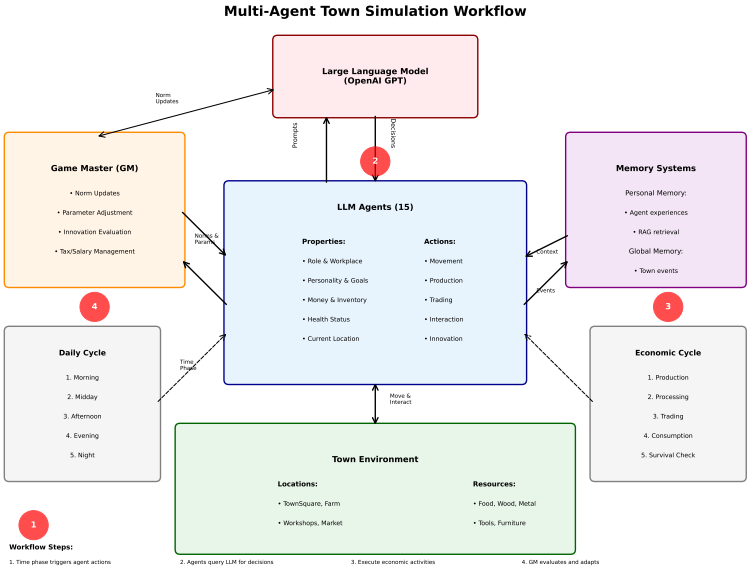


Fig. A1 Preliminary Simulation "Hunger Game" Workflow

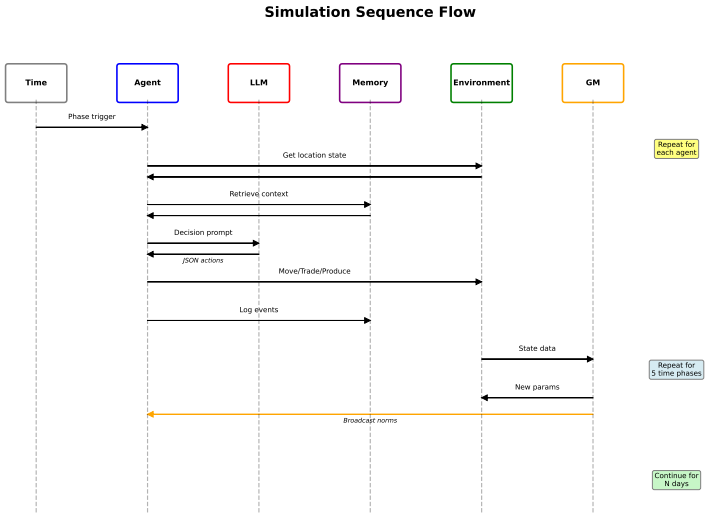
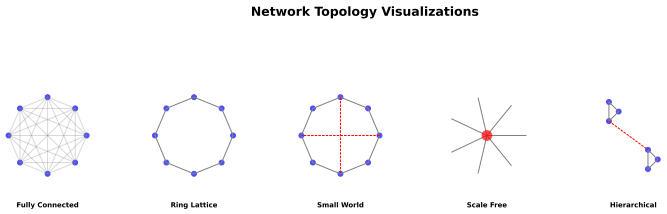


Fig. A2 Preliminary Simulation "Hunger Game" Sequence



Network Type	Connectivity	Path Length	Clustering	Best For
Fully Connected	Maximum	Shortest (1)	Highest	Rapid consensus
Ring Lattice	Minimal	Longest	Medium	Local innovation
Small World	Balanced	Short	High	Efficient spread
Scale Free	Hubs	Short	Variable	Influential nodes
Hierarchical	Modular	Medium	Very High	Group dynamics

Fig. A3 Network Design Comparison

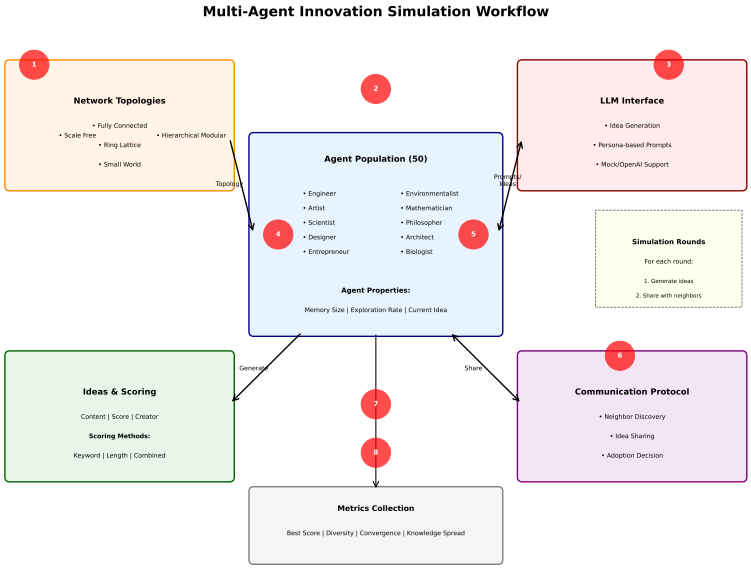


Fig. A4 Simulation 1 & Simulation 2 Workflow

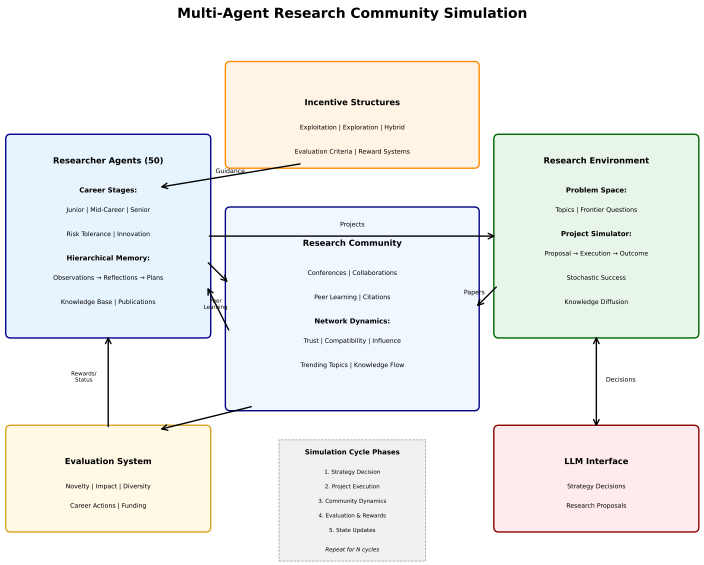


Fig. A5 Simulation 3 Workflow

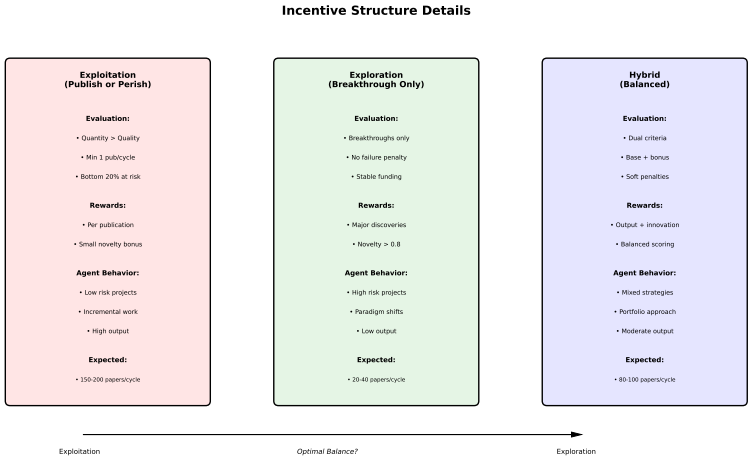


Fig. A6 Incentive Structure

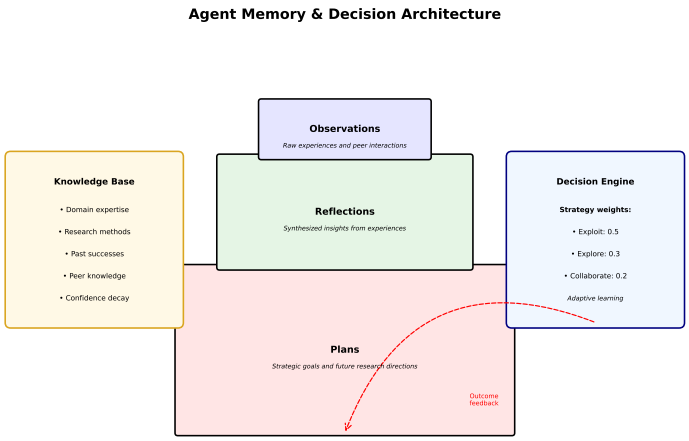


Fig. A7 Agent Memory Architecture

Bibliography

[1] Bloom, N., Jones, C.I., Van Reenen, J., Webb, M.: Are ideas getting harder to find? American Economic Review **110**(4), 1104–1144 (2020)

[2] Park, M., Leahey, E., Funk, R.J.: Papers and patents are becoming less disruptive over time. Nature **613**(7942), 138–144 (2023)

[3] Jones, B.F.: The burden of knowledge and the "death of the renaissance man": Is innovation getting harder? The Review of Economic Studies **76**(1), 283–317 (2009)

[4] Schmookler, J.: Invention and Economic Growth. Harvard University Press, Cambridge, MA (1966)

[5] Hicks, J.R.: The Theory of Wages. Macmillan, London (1932)

[6] Shiller, R.J.: Narrative economics. American Economic Review **107**(4), 967–1004 (2017) https://doi.org/10.1257/aer.107.4.967

[7] Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world'networks. nature **393**(6684), 440–442 (1998)

[8] Lazer, D., Friedman, A.: The network structure of exploration and exploitation. Administrative Science Quarterly **52**(4), 667–694 (2007)

[9] Burt, R.S.: Structural holes and good ideas. American Journal of Sociology

110(2), 349–399 (2004)

[10] Uzzi, B., Spiro, J.: Collaboration and creativity: The small world problem. American Journal of Sociology **111**(2), 447–504 (2005)

[11] March, J.G.: Exploration and exploitation in organizational learning. Organization Science **2**(1), 71–87 (1991)

[12] Azoulay, P., Graff Zivin, J.S., Manso, G.: Incentives and creativity: evidence from the academic life sciences. RAND Journal of Economics **42**(3), 527–554 (2011)

[13] Manso, G.: Motivating innovation. Journal of Finance **66**(5), 1823–1860 (2011)

[14] Epstein, J.M., Axtell, R.: Growing Artificial Societies: Social Science from the Bottom Up. Brookings Institution Press, Washington, DC (1996)

[15] Gilbert, N., Troitzsch, K.G.: Simulation for the Social Scientist, 2nd edn. Open University Press, Maidenhead, UK (2005)

[16] Park, J.S., *et al.*: Generative agents: Interactive simulacra of human behavior. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (2023)

[17] Horton, J.J.: Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research (2023)

[18] Argyle, L.P., Busby, E.C., Fulda, N., Gubler, J.R., Rytting, C., Wingate, D.: Out of one, many: Using language models to simulate human samples. Political Analysis **31**(3), 337–351 (2023)

[19] Aher, G.V., Arriaga, R.I., Kalai, A.T.: Using large language models to simulate multiple humans and replicate human subject studies. In: International Conference on Machine Learning, pp. 337–371 (2023). PMLR

[20] Bommasani, R., Hudson, D.A., Adeli, E., al.: On the opportunities and risks of foundation models. Technical report, Stanford Institute for Human-Centered Artificial Intelligence (2021). arXiv:2108.07258

[21] Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), pp. 610–623. Association for Computing Machinery, New York, NY (2021). https://doi.org/10.1145/3442188.3445922

[22] Popp, D.: Induced innovation and energy prices. American Economic Review **92**(1), 160–180 (2002) https://doi.org/10.1257/000282802760015658

[23] Ruttan, V.W.: Is War Necessary for Economic Growth? Military Procurement and Technology Development. Oxford University Press, New York (2006)

[24] Merton, R.K.: The Sociology of Science: Theoretical and Empirical Investigations. University of Chicago Press, Chicago (1973)

[25] Solow, R.M.: A contribution to the theory of economic growth. The quarterly journal of economics **70**(1), 65–94 (1956)

[26] Romer, P.M.: Endogenous technological change. Journal of political Economy **98**(5, Part 2), 71–102 (1990)

[27] Aghion, P., Howitt, P.: A model of growth through creative destruction. National Bureau of Economic Research Cambridge, Mass., USA (1990)

[28] Aghion, P., Howitt, P.W.: Endogenous Growth Theory. MIT Press, Cambridge, MA (1998)

[29] Caballero, R.J.: In: Durlauf, S.N., Blume, L.E. (eds.) creative destruction, pp. 24–29. Palgrave Macmillan UK, London (2010). https://doi.org/10.1057/9780230280823_5 . https://doi.org/10.1057/9780230280823$_5$

[30] Nelson, R.R., Winter, S.G.: An Evolutionary Theory of Economic Change. Harvard University Press, Cambridge, MA (1985)

[31] Bush, V.: Science, the endless frontier (2021)

[32] Mansfield, E.: Academic research and industrial innovation. Research policy **20**(1), 1–12 (1991)

[33] Jaffe, A.B.: Real effects of academic research. The American economic review, 957–970 (1989)

[34] Crespi, G., Castillo, R.: Supply-side versus demand-side innovation policies: Exploring the impacts of public procurement of innovation in peru (2022)

[35] Fortunato, S., Bergstrom, C.T., Börner, K., Evans, J.A., Helbing, D., Milojević, S., Petersen, A.M., Radicchi, F., Walsh, J.P., Wuchty, S., Barabási, A.-L., Wang, D.: Science of science. Science **359**(6379), 0185 (2018) https://doi.org/10.1126/science.aao0185

[36] Newman, M.E.J.: The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences **98**(2), 404–409 (2001) https://doi.org/10.1073/pnas.021544898

[37] Azoulay, P., Graff Zivin, J.S., Manso, G.: Incentives and creativity: evidence from the academic life sciences. The RAND Journal of Economics **42**(3), 527–554 (2011)

[38] Foster, J.G., Rzhetsky, A., Evans, J.A.: Tradition and innovation in scientists' research strategies. American sociological review **80**(5), 875–908 (2015)

[39] Gross, K., Bergstrom, C.T.: Rationalizing risk aversion in science: Why incentives to work hard clash with incentives to take risks. Plos Biology **22**(8), 3002750 (2024)

[40] Franzoni, C., Stephan, P.: Uncertainty and risk-taking in science: Meaning, measurement and management in peer review of research proposals. Research policy **52**(3), 104706 (2023)

[41] Cohen, W.M., Klepper, S.: Firm size and the nature of innovation within industries: The case of process and product rd. Review of Economics and Statistics **78**(2), 232–243 (1996) https://doi.org/10.2307/2109925

[42] Shiller, R.J.: Irrational Exuberance, 1st edn. Princeton University Press, Princeton, NJ (2000)

[43] Acemoglu, D.: Directed technical change. Review of Economic Studies **69**(4), 781–809 (2002) https://doi.org/10.1111/1467-937X.00228

[44] Economic Co-operation, O., Development: The covid-19 crisis and research & innovation. OECD Policy Responses to Coronavirus (COVID-19) (2021)

[45] Nelson, R.R., Winter, S.G.: An Evolutionary Theory of Economic Change. Harvard University Press, Cambridge, MA (1982)

[46] Rogers, E.M.: Diffusion of Innovations. Free Press, New York (1962)

[47] Centola, D.: The spread of behavior in an online social network experiment. Science **329**(5996), 1194–1197 (2010) https://doi.org/10.1126/science.1185231

[48] David, P.A.: Clio and the economics of qwerty. American Economic Review **75**(2), 332–337 (1985)

[49] Arthur, W.B.: Competing technologies, increasing returns, and lock-in by historical events. Economic Journal **99**(394), 116–131 (1989)

[50] Lansing, K.: Speculative bubbles and overreaction to technological innovation. FRBSF Economic Letter (2008-18) (2008)

[51] Ranis, G., Fei, J.C.: A theory of economic development. The American economic review, 533–565 (1961)

[52] Lipieta, A., Lipieta, A.: Adjustment processes within economic evolution—schumpeterian approach. Journal of the Knowledge Economy **14**(3), 3221–3259 (2023)

[53] Koyré, A.: An unpublished letter of robert hooke to isaac newton. Isis **43**(4), 312–337 (1952)

[54] Popper, K.: Conjectures and Refutations: The Growth of Scientific Knowledge. Routledge, Abingdon, UK (2014)

[55] Fleck, L.: Genesis and Development of a Scientific Fact. University of Chicago Press, Chicago, IL (1979)

[56] Acemoglu, D., Akcigit, U., Kerr, W.R.: Innovation network. Proceedings of the National Academy of Sciences **113**(41), 11483–11488 (2016)

[57] Weitzman, M.L.: Recombinant growth. The quarterly journal of economics **113**(2), 331–360 (1998)

[58] Leiponen, A.: Skills and innovation. International journal of industrial organization **23**(5-6), 303–323 (2005)

[59] Aghion, P., Boustan, L., Hoxby, C., Vandenbussche, J.: The causal impact of education on economic growth: evidence from us. Brookings papers on economic activity **1**(1), 1–73 (2009)

[60] Biasi, B., Deming, D.J., Moser, P.: Education and innovation. Technical report, National Bureau of Economic Research (2021)

[61] Kong, D., Zhang, B., Zhang, J.: Higher education and corporate innovation. Journal of Corporate Finance **72**, 102165 (2022)

[62] Brief, C.D.: China is Fast Outpacing US STEM PhD Growth. Ningbo: Centre for Sustainable Energy Technology (2021)

[63] Hall, B.H., Lerner, J.: The financing of r&d and innovation. In: Hall, B.H., Rosenberg, N. (eds.) Handbook of the Economics of Innovation vol. 1, pp. 609–639. Elsevier, Amsterdam (2010)

[64] Brown, J.R., Fazzari, S.M., Petersen, B.C.: Financing innovation and growth: Cash flow, external equity, and the 1990s r&d boom. The journal of finance **64**(1), 151–185 (2009)

[65] Howell, S.T.: Financing innovation: Evidence from r&d grants. American economic review **107**(4), 1136–1164 (2017)

[66] Mowery, D.C.: Military r&d and innovation. In: Hall, B.H., Rosenberg, N. (eds.) Handbook of the Economics of Innovation vol. 2, pp. 1219–1256. Elsevier, Amsterdam (2010)

[67] Azoulay, P., Fuchs, E., Goldstein, A.P., Kearney, M.: Funding breakthrough research: promises and challenges of the "arpa model". Innovation policy and

the economy **19**(1), 69–96 (2019)

[68] Feinson, S.: National innovation systems overview and country cases. Knowledge flows and knowledge collectives: understanding the role of science and technology policies in development **1**, 13–38 (2003)

[69] Kline, S.J., Rosenberg, N.: An overview of innovation. Studies on science and the innovation process: Selected works of Nathan Rosenberg, 173–203 (2010)

[70] Etzkowitz, H., Leydesdorff, L.: The dynamics of innovation: from national systems and "mode 2" to a triple helix of university–industry–government relations. Research policy **29**(2), 109–123 (2000)

[71] Lundvall, B.-A.: National Systems of Innovation: Towards a Theory of Innovation and Interactive Learning. Pinter, London (1992)

[72] Nelson, R.R.: National Innovation Systems: A Comparative Analysis. Oxford University Press, Oxford (1993)

[73] Archibugi, D., Howells, J., Michie, J.: Innovation Policy in a Global Economy. Cambridge University Press, Cambridge (1999)

[74] Moser, P.: How do patent laws influence innovation? evidence from nineteenth-century world's fairs. American economic review **95**(4), 1214–1236 (2005)

[75] Gilbert, R.J.: Competition and innovation. Journal of Industrial Organization Education **1**(1), 1–23 (2006)

[76] Organisation for Economic Co-operation and Development: Financing public research infrastructure. In: OECD Science, Technology and Innovation Outlook 2016, pp. 127–157. OECD Publishing, Paris (2016). Chap. 3. https://doi.org/10.1787/sti_in_outlook-2016-en

[77] Latour, B., Woolgar, S.: Laboratory Life: The Construction of Scientific Facts. Sage, Beverly Hills, CA (1979)

[78] Wuchty, S., Jones, B.F., Uzzi, B.: The increasing dominance of teams in production of knowledge. Science **316**(5827), 1036–1039 (2007) https://doi.org/10.1126/science.1136099

[79] Uzzi, B., Mukherjee, S., Stringer, M., Jones, B.F.: Atypical combinations and scientific impact. Science **342**(6157), 468–472 (2013) https://doi.org/10.1126/science.1240474

[80] Wu, L., Wang, D., Evans, J.A.: Large teams develop and small teams disrupt science and technology. Nature **566**(7744), 378–382 (2019) https://doi.org/10.1038/s41586-019-0941-9

[81] Feldman, M.P., Audretsch, D.B.: Innovation in cities: Science-based diversity, specialization and localized competition. European Economic Review **43**(2), 409–429 (1999) https://doi.org/10.1016/S0014-2921(98)00047-6

[82] Carlino, G., Kerr, W.R.: Agglomeration and innovation. In: Duranton, G., Henderson, J.V., Strange, W.C. (eds.) Handbook of Regional and Urban Economics vol. 5A, pp. 349–404. Elsevier, Amsterdam (2015). https://doi.org/10.1016/B978-0-444-59517-1.00007-5

[83] Piwowar, H.A., Vision, T.J.: Data reuse and the open data citation advantage. PeerJ **1**, 175 (2013) https://doi.org/10.7717/peerj.175

[84] Tria, F., Loreto, V., Servedio, V.D.P., Strogatz, S.H.: The dynamics of correlated novelties. Scientific reports **4**(1), 5890 (2014)

[85] Fink, T., Reeves, M., Palma, R., Farr, R.: Serendipity and strategy in rapid innovation. Nature communications **8**(1), 2002 (2017)

[86] Pammolli, F., Magazzini, L., Riccaboni, M.: The productivity crisis in pharmaceutical r&d. Nature reviews Drug discovery **10**(6), 428–438 (2011)

[87] Bhattacharya, J., Packalen, M.: Stagnation and scientific incentives. Technical report, National Bureau of Economic Research (2020)

[88] Wu, L., Wang, D., Evans, J.A.: Large teams develop and small teams disrupt science and technology. Nature **566**(7744), 378–382 (2019)

[89] Basole, R.C.: The evolution of enterprise ecosystems. Journal of Information Technology **30**(4), 335–352 (2015)

[90] Fagerberg, J.: Mobilising innovation for sustainability. Oxford Review of Economic Policy **34**(1-2), 203–216 (2018) https://doi.org/10.1093/oxrep/grx026

[91] Ohno, T.: Toyota Production System: Beyond Large-Scale Production. Productivity Press, Portland, OR (1988)

[92] Johnstone, N., Haščič, I., Popp, D.: Renewable energy policies and technological innovation: Evidence based on patent counts. Environmental and Resource Economics **45**, 133–155 (2010) https://doi.org/10.1007/s10640-009-9309-1

[93] Burt, R.S.: Structural Holes: The Social Structure of Competition. Harvard University Press, Cambridge, MA (1992)

[94] Wagner, C.S., Whetsell, T.A., Leydesdorff, L.: Growth of international collaboration in science: Revisiting six specialties. Scientometrics **110**(3), 1633–1652 (2015) https://doi.org/10.1007/s11192-016-2230-9

[95] Wu, L., Wang, D., Evans, J.A.: Large teams develop and small teams disrupt

science and technology. Nature **566**, 378–382 (2019) https://doi.org/10.1038/s41586-019-0941-9

[96] Li, D., Aharonson, B.S., Agha, L.: Hierarchy and the origins of breakthrough ideas. Research Policy **52**(4), 104659 (2023) https://doi.org/10.1016/j.respol.2022.104659

[97] Kyle, M.K., Qian, Y.: Intellectual property rights and access to innovation: Evidence from the hiv drug market. Economic Journal **124**(581), 805–836 (2014) https://doi.org/10.1111/ecoj.12017

[98] Lemley, M.A., Sampat, B.N.: Examining patent examination. Stanford Technology Law Review **2012**, 1–40 (2012)

[99] Merton, R.K.: Priorities in scientific discovery: A chapter in the sociology of science. American Sociological Review **22**(6), 635–659 (1957)

[100] Azoulay, P., Graff Zivin, J.S., Li, D.: Public rd investments and private-sector patenting: Evidence from nih funding rules. Review of Economic Studies **86**(1), 117–152 (2019) https://doi.org/10.1093/restud/rdy034

[101] Brunt, L., Lerner, J., Nicholas, T.: Inducement prizes and innovation. Journal of Industrial Economics **60**(4), 657–696 (2012) https://doi.org/10.1111/joie.12005

[102] Foster, J.G., Rzhetsky, A., Evans, J.A.: Tradition and innovation in scientists' research strategies. American Sociological Review **80**(5), 875–908 (2015) https://doi.org/10.1177/0003122415601618

[103] Kuhn, T.S.: The Structure of Scientific Revolutions. University of Chicago Press, Chicago (1962)

[104] Whitley, R.: The Intellectual and Social Organization of the Sciences. Oxford University Press, Oxford (2000)

[105] Li, G., Zhang, L., Wang, D., Evans, J.A.: Superstar collaborators produce conventional science. Nature **586**, 378–383 (2020) https://doi.org/10.1038/s41586-020-1234-5

[106] Christensen, C.M.: The Innovator's Dilemma. Harvard Business School Press, Boston (1997)

[107] Mazzucato, M.: The Entrepreneurial State: Debunking Public Vs. Private Sector Myths. Anthem, London (2018)

[108] Miller, J.H., Page, S.E.: Complex Adaptive Systems: An Introduction to Computational Models of Social Life. Princeton University Press, Princeton, NJ (2007)

[109] Fioretti, G.: Agent-based social simulation: A brief introduction. The Knowledge Engineering Review **28**(4), 333–339 (2013)

[110] Subramanian, A., Kottapalli, A., et al.: Large language models as agents in agent-based models. arXiv preprint arXiv:2306.XXXX (2023)

[111] Zhu, Y., Lu, S., Zheng, L., Guo, J., Zhang, W., Wang, J., Yu, Y.: Texygen: A benchmarking platform for text generation models. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 1097–1100 (2018)

[112] Merton, R.K.: The matthew effect in science. Science **159**(3810), 56–63 (1968)

[113] Uzzi, B., Spiro, J.: Collaboration and creativity: The small-world problem. American Journal of Sociology **111**(2), 447–504 (2005)

[114] Barabási, A.-L., Albert, R.: Emergence of scaling in random networks. Science **286**(5439), 509–512 (1999)

[115] Zhu, Y., Lu, S., Zheng, Y., Guo, Y.N., Wang, Y., Hovy, E., Chen, L.: Texygen: A benchmarking platform for text generation models. arXiv preprint arXiv:1802.01886 (2018)

[116] Hong, L., Page, S.E.: Groups of diverse problem solvers can outperform groups of high-ability problem solvers. Proceedings of the National Academy of Sciences **101**(46), 16385–16389 (2004)

[117] Merton, R.K.: The matthew effect in science. Science **159**(3810), 56–63 (1968)