informatics

MDPI

Article

Comparing Genetic Algorithms and LLMs for Research-Hypothesis Generation: Mainstream Convergence of LLMs, Wider Hypothesis Exploration by Combinatorial Search, and the Limits of Literature-Distance Novelty Metrics

Akira Otsuki

College of Economics, Nihon University, Tokyo 101-8360, Japan; otsuki.akira@nihon-u.ac.jp

Abstract

Large language model (LLM)-based creative AI systems tend to produce plausible rather than novel research ideas. This study compares two ways of generating research hypotheses: a genetic algorithm (GA) that recombines Method, Problem, and Goal elements under a fitness function rewarding distance from the literature and literature support for each link, and free proposals by an LLM. Papers and hypotheses share one embedding space built from 2856 arXiv papers. Over 30 runs with budget-matched baselines and outcome measures on held-out and later-published papers, the selected hypotheses were more distant from the unseen literature than unsearched ones, but the GA matched rather than outperformed random search and hill climbing. Across two LLM families, two prompts, and two temperatures, and even when choosing from the GA's concept vocabulary, LLM-proposed hypotheses were less diverse and closer to existing literature, whereas combinatorial search explored more widely. On 9755 ICLR submissions, the literature-distance novelty measure was practically unrelated to reviewers' judgments (Spearman $\rho \leq 0.06$), so it cannot establish originality. In controlled manuscript generation, LLM judges proved unreliable, and their order-consistent verdicts more often favored manuscripts built on LLM-proposed hypotheses, which blinded human raters tended to rank as least useful. Code and data are openly released.

Keywords: creative AI; genetic algorithm; novelty evaluation; hypothesis generation; large language models; literature-based discovery; temporal hold-out evaluation; research automation

Check for updates

Academic Editor: Antony Bryant

Received: 12 September 2026
Revised: 11 October 2026
Accepted: 14 October 2026
Published: 16 October 2026

Copyright: © 2026 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

As illustrated in Table 1, the emergence of text-generation AI and AI agents has enabled AI to efficiently perform a large portion of tasks that were previously carried out by humans.

However, as discussed in Section 2, creative AI that creates new value (research) is still considered to be at a developmental stage from the standpoint of deriving novelty. Therefore, this study constructs a knowledge space of existing papers from a paper database, uses a genetic algorithm to derive new research themes (hypotheses) from it, and researches and develops a creative AI that autonomously and automatically processes everything from that research theme through to manuscript writing.

Table 1. Examples of task substitution by text-generation AI and AI agents.

Industry	Key Outcomes/ROI	Source
Finance	According to a McKinsey report, AI has achieved cost reductions of up to 70% in certain banking operations.	[1]
Medical	AI can now perform radiological image diagnosis in under five seconds, at a fraction of the cost of a manual review.	[1]
Real Estate	From home pricing and property marketing to lead generation and virtual property tours, AI is revolutionizing the real-estate industry.	[2]
IT	AI agents have doubled the pace of development cycles through automated coding, debugging, documentation updates, and security vulnerability remediation, and have fully replaced simple bug-fixing work.	1. How AI agents automated most of our dbt development [3] 2. OpenAI announces "Codex Security," an AI agent that automates vulnerability discovery, verification, and remediation [4]
General	AI agents can support nearly any task, including text generation, conversation summarization, content classification, data extraction, question answering, and next-step recommendation.	[5]

This study re-implements the model with papers and hypotheses represented in a single semantic vector space produced by a pretrained sentence encoder and with a utility term grounded in the literature rather than in hand-written rules, and it makes three contributions. First, it provides a reproducible evaluation framework for hypothesis-generating systems that separates the papers used for search from those used for evaluation and adds checks that do not rely on the model's own scoring functions: 30 independent runs against budget-matched baselines and ablations, outcome measures computed on held-out and later-published papers (which test out-of-sample consistency rather than independent validity), controlled manuscript generation, a blinded human evaluation, and a validation against expert peer review, with all code, data, prompts, random seeds, and run logs openly released. Second, it shows empirically, for two small open-weight LLM families, two prompts, and two temperatures, that LLM-proposed hypotheses are less diverse and closer to existing literature than combinatorially selected ones, also when the LLM chooses from the same concept vocabulary; this substantiates, with data, the motivating concern that LLMs gravitate toward plausible ideas. Third, it shows that literature-distance novelty, a measure widely used to assess generated ideas, is practically unrelated to expert reviewers' judgments, so that such measures cannot establish originality without external validation; moreover, an LLM judge, proved unreliable and, where its verdicts were consistent, leaned toward the LLM-proposed hypotheses that human raters tended to rank lowest. The genetic algorithm itself did not outperform simpler search at the present scale; this negative result is reported as a finding, indicating that the representation and the objective, rather than the search operator, matter in concept spaces of this size. An earlier end-to-end comparison of generated manuscripts is retained only as a pilot case study (Sections 4.8–4.10).

2. Related Work

The attempt to automate the entire research process—so-called creative AI—is one of the ultimate frontiers of artificial intelligence research. In particular, autonomous scientific-discovery systems such as The AI Scientist [6,7], released by Sakana AI, have demonstrated the ability to complete the cycle from hypothesis generation through experiment execution,

data analysis, manuscript writing, and even peer review end-to-end, built around a large language model (LLM) core. The initial version, "The AI Scientist-v1" [6], relied heavily on human-authored code templates, which constrained the scope of research to within the template [7]. Its successor, "The AI Scientist-v2" [7], introduced "agentic tree search," in which a manager agent overseeing experiment progress systematically explores preliminary investigation, hyperparameter tuning, main experiments, and ablation studies, deepening the research. However, according to a report from an independent evaluation team [8], many of the ideas that The AI Scientist judged to be "novel" were in fact concepts that had already been known in the field for decades. For example, techniques such as microbatching in stochastic gradient descent (SGD), adaptive learning-rate scheduling, and hybrid matrix factorization were flagged as novel by The AI Scientist, yet these are already widely documented. This is because The AI Scientist's novelty judgment relies on API queries to literature databases such as Semantic Scholar, but in practice this amounts to little more than simple keyword search; consequently, cases have been reported in which insufficient query design or synthetic understanding causes the system to mistakenly judge ideas as "novel." In addition, The AI Scientist tends to add or modify only about 8% of the characters in the existing template code on average, suggesting that much of the "novelty" it proposes is limited to superficial changes such as tuning hyperparameters within an existing framework or applying a standard regularization technique to a different dataset.

SciMON (Scientific Inspiration Machines Optimized for Novelty) [9] is a framework that extracts "inspiration" from past scientific papers and explicitly optimizes for novelty. It iteratively compares generated ideas against existing literature and continues updating its proposals until sufficient novelty is achieved. However, because SciMON generates ideas by extracting distinct concepts from existing literature (e.g., a knowledge graph) and recombining them, it risks mass-producing "patchwork" ideas that merely connect existing techniques A and B, without reaching a truly paradigm-shifting "creation of new concepts." Moreover, although SciMON has a loop for self-evaluating the novelty of its generated ideas, this evaluation itself depends on the LLM's training data; because the LLM evaluates based on "past patterns" contained in that training data, it risks undervaluing genuinely novel ideas that lie outside the extension of the past as mere "errors" or "meaningless" output.

MOOSE-Chem [10] defines scientific discovery as a combination of "research background" and "inspiration" and mathematically decomposes the hypothesis-generation process to execute it. Specialized for the chemistry domain, it has succeeded in rediscovering hypotheses from nature- and science-level papers under conditions excluded from its training data, demonstrating an ability to surface relationships among unknown knowledge. However, MOOSE-Chem, too, is strong at combinatorial discovery—"connecting known A and known B" from existing papers—but is less able to achieve "fundamental discovery," in which entirely new principles or concepts are produced; even when a resulting paper appears "new" in form, it is likely, in substance, to remain "incremental novelty" that merely shifts the scope of application of an existing technique.

Scideator [11] is a human–AI collaborative system that synthesizes new ideas by combining three sets: Purpose, Mechanism, and Evaluation. Its built-in "Idea Novelty Checker" analyzes in detail how these facet combinations differ from existing literature and explains the basis for novelty. However, like SciMON and MOOSE-Chem, Scideator is designed to deepen ideas through "chain-of-thought" reasoning and "iterative refinement," so the ideas it generates tend to remain within "combinatorial novelty" that does not go beyond a combination of cited literature.

ChemCrow [12] gives an LLM (such as GPT-4) concrete instructions for tasks such as drug design and materials synthesis, autonomously querying synthesis-verification data from platforms (Search/WebSearch, Name2SMILES, ReactionPlanner, ReactionExecute,

etc.) and iteratively generating synthesis procedures. Whereas SciMON and Scideator emphasize "recombining paper data," ChemCrow's strength lies in feasibility—"can it actually be made, and is it computationally correct"—but this comes with its own challenges from a "novelty" standpoint. Because ChemCrow uses tools such as RXN4Chemistry (reaction prediction) and SMILES manipulation, which are programmed based on "existing chemical knowledge," the ideas it generates tend to stay within the range the tools can predict—that is, within "known reaction rules." As a result, it tends to converge on safe, incremental improvements that the existing tools readily approve, rather than proposing "new reaction mechanisms that break existing rules" or "unprecedented catalytic concepts."

Thesis AI [13] is a manuscript-writing tool that can compose papers of up to 80 pages from a single prompt and related literature (PDFs). Thesis Generator: Essay AI [14] is a lightweight writing tool aimed primarily at undergraduate coursework, specialized for quickly drafting a paper's thesis statement. However, both Thesis AI and Thesis Generator: Essay AI construct ideas based on token-occurrence probabilities and semantic similarity over an existing corpus of literature, so they tend to remain confined to exploring the "adjacent possible" formed by combining known A and known B, and are unlikely to produce the discontinuous leaps needed to overturn an existing theoretical framework.

Finally, AI agents such as Gemini 3's Deep Research feature and Antigravity [15] have substantially strengthened the "logical leaps" and "autonomous execution of complex processes" that conventional LLMs have struggled with; using an AI agent, it is likewise possible, as with The AI Scientist, to automatically carry out related work surveys, idea generation, and manuscript writing. Because an AI agent autonomously identifies and prioritizes tasks toward a given goal before executing them, it can, more than conventional LLM-reliant creative AI, discover groundbreaking ideas autonomously, beyond human instruction. In other words, in the ability to combine existing knowledge to derive the unknown closest to the correct answer, it is currently the most powerful tool available; however, because the underlying "brain" of an AI agent is still an LLM, it remains difficult for it to generate novelty that destroys and rebuilds the existing logical framework itself.

Two further lines of work bear directly on how such systems should be evaluated. Literature-based discovery, which originated in Swanson's observation that individually published links (A–B and B–C) can be combined into unpublished but testable hypotheses (A–C) [16], motivates treating a hypothesis as promising when its components are each attested in the literature while their combination is not; this is the structure of the novelty and utility terms used in the present model. Temporal hold-out validation, in which a model is built only from literature up to a cutoff date and its outputs are compared with literature published afterwards, has been used to show that word embeddings of materials science abstracts anticipate later discoveries [17] and to forecast emerging research directions in artificial intelligence [18]. A large blinded study with more than 100 NLP researchers found that LLM-generated research ideas were judged more novel, but slightly less feasible, than ideas written by experts [19], and LLM agents that iteratively refine ideas over the literature have been proposed [20]; such studies show that automated novelty scores require external validation. Finally, searching for novelty with evolutionary computation has a precedent in novelty search, which rewards distance from previously found solutions instead of an explicit objective [21].

In summary, the survey of related work confirms that current generative AI and AI agents can already automate the cycle from hypothesis generation through experiment execution, data analysis, manuscript writing, and even peer review. However, substantial challenges remain in achieving the essence of research—"generating novelty that destroys and rebuilds the existing logical framework itself." Specifically, all of the related work surveyed above is LLM-based, and an LLM, typified by GPT (Generative Pre-trained

Transformer) [22], is a model that generates text by using the Transformer architecture to consider the context of an entire sentence and, as shown in Equation (1), precisely updating the probability distribution over the next word from the combination of preceding words to predict and generate the following word.

$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta) \quad (1)$$

Consequently, while an LLM has an extremely high capacity to reproduce natural language patterns learned from large volumes of data, it does not verify "facts"; it merely generates output that is "not unnatural as a sequence of words," so it can also produce statistically plausible falsehoods (hallucinations). Furthermore, the ideas it generates also tend toward safe, standard claims that many people have already made, making it unlikely to produce academically valuable, sharp perspectives or hypotheses that overturn conventional wisdom. Therefore, this study applies a genetic algorithm to research and develop a creative AI model that, through repeated trial and error, discovers novelty that humans would not readily conceive.

3. Materials and Methods

The proposed model consists of the components described in Sections 3.1–3.6.

3.1. Construction of the Knowledge Space from Existing Papers

First, the knowledge space K of existing papers is defined as a pair consisting of a concept set C and a paper collection P (Equation (2)).

$$K = (C, P) \quad (2)$$

The concept set is the collection of basic vocabulary held by the knowledge space, expressed as the direct sum of the following three subsets: the method set (Method), the problem set (Problem), and the goal set (Goal).

$$C = (C_{method} \cup C_{problem} \cup C_{goal}) \quad (3)$$

For the paper collection P, each paper $p \in P$ is defined as an object having a title, an abstract, a publication year, and a vector v that numerically represents its content.

$$P = (title, summary, year, v) \quad (4)$$

Here, $v \in R^d$ is the embedding of the paper's title and abstract produced by a pretrained sentence encoder φ (all-MiniLM-L6-v2, a distilled Sentence-BERT model [23,24]; $d = 384$; unit-normalized). The same encoder is applied to hypotheses in Section 3.2, so that papers and hypotheses lie in a single vector space and their distances are directly comparable.

3.1.1. Dynamic Model of Knowledge Acquisition

This is modeled as a mapping function that extracts paper information from the paper database and updates the knowledge space K. First, the search-query set function Q is defined as in Equation (5).

$$Q = (q_1, q_2, \cdots, q_n) \quad (5)$$

(e.g., q_1 = "LLM hallucination limitations").

3.1.2. Knowledge Update Process

Letting the external knowledge base (the public arXiv API [25]) be S, the new paper collection P_{new} is obtained as in Equation (6).

$$P_{new} = \bigcup_{q \in Q} search(S, q, limit) \tag{6}$$

Here, limit denotes the maximum number of results retrieved per query. Q contains one query per concept term (42 terms; limit = 25; categories cs.AI, cs.LG, cs.CL, and cs.NE), and a background sample from a seeded random day of each month (limit = 40) is added so that the knowledge space is not restricted to the concept vocabulary itself. Two temporally disjoint windows were collected with 150 queries: papers submitted between January 2020 and December 2023 ("past"; 2856 papers) and between July 2024 and December 2025 ("future"; 1697 papers). That is, Equation (6) represents the retrieval of papers from the source S using q, and the resulting knowledge-space state K_{t+1} is obtained by adding the newly retrieved P_{new} to the existing paper collection P_t (Equation (7)).

$$K_{t+1} = P_t \cup P_{new} \tag{7}$$

3.1.3. Vector Representation of the Knowledge Space of Existing Papers

Expressing the paper collection as a knowledge matrix V gives Equation (8).

$$V = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} \in \boldsymbol{R^{n \times d}} \tag{8}$$

Here, $n = |P|$ (the total number of papers), and each row vector v_i represents the semantic features of a paper. This matrix V functions as the "distribution" of existing knowledge for subsequent tasks (such as computing novelty or retrieving similar papers). $R^{n \times d}$ denotes a real-valued matrix with n rows (papers) and d = 384 columns. Individual dimensions of a sentence embedding have no stand-alone meaning, but the cosine similarity between two rows reflects the semantic similarity of the corresponding papers. To separate search from evaluation, the past papers were split once at random (fixed seed) into a search partition K_s, which is the only part of the knowledge space the GA may access, and a held-out partition K_e; the future papers form a third partition K_f, subsampled to the size of K_e because maximum-similarity statistics depend on set size (Table 2). K_e and K_f are used only for the outcome measures of Section 3.4.

Table 2. Composition of the knowledge space (arXiv metadata).

Partition	Submission Period	Role	Papers
Search, K_s	January 2020–December 2023	Novelty and utility during search	1428
Held-out past, K_e	January 2020–December 2023	Outcome measures N_{out}, U_{out}	1428
Future, K_f	July 2024–December 2025	Outcome measures FS, Δ	1428 (of 1697)
Retrieved in total		84 targeted + 66 background queries	4553

3.2. Generation of Proto-Concepts for Candidate Research Themes (Hypotheses)

A combinatorial approach is used that recombines existing knowledge parts to create new "meanings"; here, too, a genetic-algorithm-like approach, described in the following section, is partly used to generate the proto-concepts of hypotheses.

First, the elements constituting a hypothesis proto-concept are defined as sets, as follows.

- Knowledge Space K:

$$K = \{M, P, G\} \tag{9}$$

Here, M, P, and G denote the finite sets of methods (Method), problems (Problem), and goals (Goal), respectively.

Note that P (Problem) used from this section onward is a different symbol from the P used in Equation (4) to denote the paper collection; the two should be distinguished by context.

- Hypothesis proto-concept h:

A hypothesis proto-concept h is defined as one element (individual) of the direct-product set S formed from M, P, and G.

$$h = (m,p,g) \in S \; where \; S = M \times P \times G \tag{10}$$

Here, $K = \{M, P, G\}$ (Equation (9)) refers to the element sets of method, problem, and goal themselves, whereas $S = M \times P \times G$ (Equation (10)) is their direct product—that is, the space of all combinations an individual h can take.

- Concept Vector v:

The embedding function φ then maps the hypothesis proto-concept h into a d-dimensional real vector space. φ is the sentence encoder of Section 3.1 applied to a fixed natural-language rendering of h, "Using m to address p in order to improve g." (d = 384), so that hypothesis vectors and paper vectors are directly comparable.

$$\varphi: S \to \boldsymbol{R}^{\boldsymbol{d}}, v_h = \varphi(h) \tag{11}$$

For example, suppose the following three hypotheses exist:

- Hypothesis A: "Use deep learning (M) to reduce food loss (P) and achieve cost reduction (G)."
- Hypothesis B: "Use AI (M) to reduce unnecessary waste (P) and improve profit margins (G)."
- Hypothesis C: "Use 3D printing (M) to solve housing shortages (P) and increase supply speed (G)."

Table 3 lists the cosine similarities that φ assigns to these three hypotheses. Hypotheses A and B, which are close in meaning, have a similarity of 0.456, whereas Hypothesis C has similarities of only 0.194 and 0.250 to Hypotheses A and B, respectively; the vector space thus reflects semantic similarity.

Table 3. Cosine similarity between the example hypotheses under the encoder φ.

Hypothesis	A	B	C
A (deep learning/food loss/cost reduction)	1.000	0.456	0.194
B (AI/unnecessary waste/profit margins)	0.456	1.000	0.250
C (3D printing/housing shortages/supply speed)	0.194	0.250	1.000

Next, the generation model for the proto-concepts of candidate research themes (hypotheses) is described. The generator is a set of operators that generate new individuals h from the knowledge space K; first, as shown in Equation (12), an initial individual is

generated by selecting elements from each set in the knowledge space according to a uniform distribution.

$$h_{new} = (unif(M), unif(P), unif(G)) \tag{12}$$

Next, an offspring individual is generated by independently inheriting each dimension (attribute) from two parent hypotheses h_1 and h_2 (Equation (13)). The prime symbol ($'$) denotes a newly generated concept (m, p, g).

$$h_{child} = (m', p', g') \tag{13}$$

Here, each component x$' \in \{m', p', g'\}$ is determined by the Bernoulli trial in Equation (14) ($p = 0.5$). That is, whether each component m', p', g' inherits from parent h_1 or parent h_2 is determined with probability $p = 0.5$ (Equation (14)).

Here, x_1 denotes the corresponding attribute of parent h1, and x_2 the corresponding attribute of parent h2.

$$x' = \begin{cases} x_1 & with\ probability\ 0.5 \\ x_2 & with\ probability\ 0.5 \end{cases} \tag{14}$$

Next, one attribute of individual h is randomly selected and replaced with a different element from the knowledge set (Equation (15)).

$$h_{mutated} = f_{mut}(m, p, g) \tag{15}$$

That is, letting the selection of the attribute to be mutated be $i \in \{1, 2, 3\}$ (corresponding to method, problem, and goal), this is expressed as follows.

$$h_{mutated} = \begin{cases} (m_{new}, p, g) & if\ i = 1, m_{new} \sim unif(M) \\ (m, p_{new}, g) & if\ i = 2, p_{new} \sim unif(P) \\ (m, p, g_{new}) & if\ i = 3, g_{new} \sim unif(G) \end{cases} \tag{16}$$

As a concrete example of the above, the process of generating a hypothesis proto-concept is described in ①–③ below, assuming the knowledge space K shown in Table 4.

Table 4. Example contents of the knowledge space K.

Category	Elements (Concepts)
Method	Deep learning, blockchain, 3D printing
Problem	Food loss, counterfeit goods, housing shortage
Goal	Cost reduction, improved reliability, faster supply

- Ex Nihilo Ideation (Random Generation).

As in the example below, an element is randomly selected for each slot (method, problem, goal).

- "Blockchain" selected from Method.
- "Food loss" selected from Problem.
- "Improved reliability" selected from Goal.

As a result, a hypothesis proto-concept such as "blockchain for improved reliability to solve food loss" is generated.

- Crossover.

This is a process of "imitation and combination" that produces a new idea (offspring) by combining two existing ideas (parents). If the two existing ideas (parents) are as follows:

- Parent 1 (h1): deep learning for cost reduction to solve food loss.

- Parent 2 (h2): 3D printing for faster supply to solve housing shortage.

Then, as illustrated below, for each attribute either parent is selected with 50% probability.

- Method: inherited from Parent 2 as "3D printing".
- Problem: inherited from Parent 1 as "food loss".
- Goal: inherited from Parent 2 as "faster supply".

As a result, a new hypothesis proto-concept such as "3D printing for faster supply to solve food loss" (immediate supply through 3D-printed food) is generated.

- Mutation.

This corresponds to an "unexpected leap" that forcibly rewrites part of the hypothesis proto-concept from ② into something different. For example, if the original hypothesis (h) is as follows:

"Deep learning for cost reduction to solve food loss".

The problem is selected as the target to be rewritten by the mutation process, and "food loss" is replaced with "housing shortage" from the knowledge space. As a result, the hypothesis proto-concept changes to:

"Deep learning for cost reduction to solve housing shortage" (e.g., cost reduction through automated design and material optimization).

Table 4 is a toy example. The concept set used in the experiments of Section 4 consists of 15 methods (Transformer, Diffusion Model, Reinforcement Learning, Evolutionary Strategy, Knowledge Graph, Symbolic AI, Quantum Annealing, Graph Neural Network, Contrastive Learning, Federated Learning, Retrieval-Augmented Generation, Bayesian Optimization, Multi-Agent System, Neural Architecture Search, Causal Inference), 15 problems (Hallucination, Catastrophic Forgetting, Data Scarcity, High Latency, Bias, Lack of Novelty, Energy Consumption, Adversarial Vulnerability, Distribution Shift, Privacy Leakage, Reward Hacking, Long-Context Reasoning, Label Noise, Miscalibrated Uncertainty, Opaque Decision Making), and 12 goals (Creativity, Robustness, Efficiency, Fairness, Autonomy, Explainability, Scalability, Safety, Generalization, Personalization, Trustworthiness, Sustainability), giving 2700 possible hypotheses.

3.3. Generation of Candidate Research Themes (Hypotheses) via Genetic Algorithm

This is a process that uses the framework of a genetic algorithm (GA) to generate new "hypotheses" from the hypothesis proto-concepts generated in the previous section—an automation of the process by which combinations that humans would not readily conceive are discovered through repeated trial and error.

The proposed model adopts an evolutionary computation approach that combines John Holland's foundational theory [26] with Kenneth De Jong's elitism strategy [27]. New hypotheses are generated by implementing an iterative heuristic search that combines the knowledge space K' of hypothesis proto-concepts to find the hypothesis h* with the highest score (a weighted sum of novelty and utility; Equation (21)).

Note that the knowledge space K' used in this section refers to the set of hypothesis proto-concepts generated in the previous section (Section 3.2)—that is, the candidate space that the GA searches in this section—and is a symbol used in a different context from $K = \{M, P, G\}$ (the element sets of method, problem, and goal themselves) in Equation (9) or $K = (C, P)$ (the knowledge space of existing papers) in Equation (2). In addition, this study treats the three elements m, p, and g as attributes corresponding to genes in the GA, and treats their combination $h = (m, p, g)$ as a single individual.

The overall procedure for hypothesis generation by the GA is as follows, in steps (1)–(7): (1) Generate an initial population P0 (Equation (17)). (2) Compute the evaluation value f(h) for each individual h and rank the population (Equation (21)). (3) Preserve the

top two individuals as elites directly into the next generation (Equation (22)). (4) Select parents from the top five individuals and generate offspring by combining the M, P, and G elements through crossover (Equation (23)). (5) Apply mutation with probability 0.3 to individuals generated by crossover (Equation (24)). (6) Combine the elites and the newly generated individuals to form the next generation P_{t+1} (Equation (25)). (7) Repeat steps (1)–(6) for a fixed number of generations (20 generations) to obtain the final candidates.

First, the various symbols are defined as follows.

- Knowledge Space of Hypothesis Proto-Concepts: K'.
- Hypothesis: $h \in H$ (the individual being searched).
- Population: the set of the t-th generation is denoted $P_t = \{h_1, h_2, \ldots, h_N\}$.
- Fitness Function: $f(h)$ = score. $f(h)$ is a function of "novelty" and "utility": see Equation (21).

Next, the mathematical model of the evolutionary process is described in ①–③ below.

- Initialization.

The initial population P_0 is generated by random sampling from the hypothesis proto-concept space K' created in the previous section, as shown in Equation (17).

$$P_0 = \{g_i(K') \mid i = 1, \cdots, N\} \tag{17}$$

Here, g is a generating function that produces random numbers.

- Evaluation and Selection.

As shown below, in each generation, individuals are ranked (ranking) by their score.

$$h_{(1)}, h_{(2)}, \cdots, h_{(N)} \; s.t. \; f(h_{(1)}) \geq f(h_{(2)}) \geq \cdots f(h_{(N)})$$

Here, *s.t.* means ranking under the constraint "$f(h(1)) \geq f(h(2)) \geq \cdots \geq f(h(\mathrm{N}))$", where $h(1)$ denotes rank 1. The score $f(h)$ is defined as a linear combination of the two variables $N(h)$ and $U(h)$, described below.

- $N(h)$ (Novelty): how far a hypothesis is from the papers in the search partition K_s, defined as in Equation (18). The score increases the less similar the hypothesis is to every existing paper.

$$N(h) = 1 - \max_{k \in K_s} cos(\varphi(h), \; v_k) \tag{18}$$

- $U(h)$ (Utility): Equation (19) evaluates whether the pairwise links that make up a hypothesis are supported by the literature, a data-driven proxy for the "soundness" of the combination.

$$U(h) = \frac{1}{3}[s(m,p) + s(m,g) + s(p,g)], \; s(x,y) = \max_{k \in K_s} cos(\varphi(x \oplus y), \; v_k) \tag{19}$$

Here, $x \oplus y$ denotes a short phrase that expresses the link between two elements ("m for p", "m for g", and "addressing p to improve g"), and $s(x, y)$ is the similarity between that phrase and the closest paper in $\mathrm{K_s}$. No hand-written rule is involved: a link scores high when some existing paper discusses it. A hypothesis therefore obtains high utility when each of its pairwise links is attested in the literature and high novelty when the complete triple is not—that is, when it recombines individually supported links into an unexplored whole, the structure exploited in literature-based discovery [16]. The three links correspond to the following checks.

- Fit between m and p: can this method solve this problem? (technical feasibility).
- Fit between m and g: is this method suitable for achieving this goal? (efficiency).
- Fit between p and g: if this problem is solved, is this goal actually achieved? (goal validity).

Similarity in Equations (18) and (19) is the cosine similarity of the embeddings (Equation (20)).

$$cos(\varphi(x),\ v_k) = \frac{\varphi(x)^T v_k}{\|\varphi(x)\|\|v_k\|} = \varphi(x)^T v_k \tag{20}$$

Because all vectors are unit-normalized, cosine similarity and Euclidean distance are monotonically related ($\|a - b\|^2 = 2 - 2\cos(a, b)$), so a Euclidean formulation would rank hypotheses identically; what matters is that hypothesis and paper vectors are produced by the same encoder. Then, as shown in Equation (21),

$$f(h) = \alpha N(h) + \beta U(h) \tag{21}$$

the score is computed by linearly combining $N(h)$ and $U(h)$.

Here, $\alpha = 0.6$ and $\beta = 0.4$ (Section 3.4). The same function serves both as the fitness during search and as the selection criterion $C(h')$ of Section 3.4 ($C \equiv f$), and the best hypothesis evaluated during a run is returned. Because N and U are the very quantities that the search optimizes, they are not used as evidence of the model's effectiveness; the outcome measures defined in Section 3.4 are computed on partitions of the knowledge space that the search never accesses.

- Generation of the Next Generation.

The next generation P_{t+1} is constructed through the following three operations.

- Elitism:

As shown in Equation (22), the top two individuals by score are first carried over unchanged.

$$E_t = \{h_{(1)}, h_{(2)}\} \tag{22}$$

The reason for preserving elites is as insurance in case the best record is not surpassed through "adventure" (crossover and mutation). Because this study focuses on the creation of new ideas, the ratio of elitism (stability) to adventure (crossover) is set to 20:80, although this ratio can be freely adjusted.

- Crossover:

Then, as shown in Equation (23), parents are randomly selected from the top five individuals ($h(1)$–$h(5)$) to generate a new individual.

$$child = c(parent_1, parent_2)\ where\ parent \in \{h_{(1)}, \cdots, h_{(5)}\} \tag{23}$$

Because the top five individuals, including elites h(1) and h(2), are each selected with 20% probability, about 40% of offspring have one of the two elites as a parent, and about 4% ($1/5 \times 1/5$) are offspring of both elites (ranks 1 and 2) together.

- Mutation:

As shown in Equation (24), the mutation operator $m(h)$ is applied with probability $p = 0.3$. This is intended to maintain diversity through mutation ($p = 0.3$) so as to avoid local optima, and is applied only to individuals produced by crossover.

$$h' = \begin{Bmatrix} m(child) & with\ probability\ 0.3 \\ child & with\ probability\ 0.7 \end{Bmatrix} \tag{24}$$

From the above, the next generation P_{t+1} is defined as in Equation (25). $N - 2$ reflects that the elites already occupy two slots, leaving $N - 2$ remaining slots. That is, E_t denotes the set of top individuals with the highest fitness scores (the weighted average of novelty N and utility U) among the $N - 2$ remaining slots.

$$P_{t+1} = E_t \cup \left\{ h_i' \middle| i = 1, \cdots, N-2 \right\} \tag{25}$$

The following describes the hypothesis-generation process using a concrete example.

- First, the top-2 hypotheses by score are extracted and preserved from the hypothesis proto-concept space K', as follows. This is the elitism operation.
- Hypothesis 1: {M: deep learning, P: food loss, G: cost reduction}.
- Hypothesis 2: {M: 3D printing, P: housing shortage, G: faster supply}.
- Next, as the crossover operation, Hypotheses 1 and 2 are "married" to produce a new offspring as follows. For clarity, the elite Hypotheses 1 and 2 are crossed here, but in practice parents 1 and 2 are randomly selected from Hypotheses 1–5. In the example below, an offspring individual is generated by combining Parent 1's "problem (P)" and "goal (G)" with Parent 2's "method (M)."
- Parent 1 (Hypothesis 1): {M: deep learning, P: food loss, G: cost reduction}.
- Parent 2 (Hypothesis 2): {M: 3D printing, P: housing shortage, G: faster supply}.
- Offspring (crossover): {M: 3D printing, P: food loss, G: cost reduction}.

Through this process, a hypothesis with higher logical consistency (soundness)—such as "using a 3D food printer to reshape food destined for disposal into low-cost processed food"—is derived, one that would be difficult to reach through mere random generation.

- Next is the mutation operation: with 30% probability per generation, one of the three elements is rewritten to something entirely different. As this is repeated over several generations, cumulative changes can occur in which multiple elements among the three (two, in the example below) are replaced.
- Before mutation: {M: 3D printing, P: food loss, G: cost reduction}.
- After mutation: {M: 3D printing, P: shortage of disaster shelters, G: faster supply}.

Through this mutation, the method (M) is retained while the problem (P) and goal (G) are renewed. As a result, the solution leaps from the original context of "food loss" to the new domain of "disaster relief," preserving the possibility of a major discovery that goes beyond existing knowledge.

- Finally, the above process is repeated 20 times. Over these 20 iterations, the structural characteristics of the hypothesis evolve as illustrated in Table 5. The best hypothesis evaluated during the 20 generations (200 fitness evaluations) is returned.

Table 5. Illustrative progression of hypothesis structure through the genetic algorithm. (hypothetical example for exposition, not an experimental trace).

Stage	Generation	Structural Characteristics of the Hypothesis	Mathematical/Logical State
Chaotic phase	1–4	*M*, *P*, *G* completely independent Example: {*M*: 3D printing, *P*: food loss, *G*: profit margin improvement}	All elements are random. No logical chain is connected, and the utility score is extremely low.
Formation phase	5–8	Emergence of a partial logical chain Example: {*M*: deep learning, *P*: food loss, *G*: profit margin improvement}	Crossover replaces individuals with a good fit between "method (*M*) and problem (*P*)," but the goal (*G*) does not yet align.
Transition phase	9–12	Acceptance of "inconsistency" through mutation Example: {*M*: deep learning, *P*: food loss, *G*: faster supply}	One element (*G*) mutates. Although it appears mismatched at first glance, this opens the door to a new solution space of "logistics and supply."
Refinement phase	13–16	Completion of logical consistency through crossover Example: {*M*: AI, *P*: food loss, *G*: faster supply}	For the goal of "faster supply" obtained through mutation, the method "AI" flows in via crossover from another individual, directly linking *M*–*P*–*G*.
Convergence phase	17–20	Extraction of the optimal solution Example: h^* = {*M*, *P*, *G*} Best	In the knowledge space K', the hypothesis with the highest novelty and the greatest logical robustness is "discovered."

Note: h^* denotes the best hypothesis found by the genetic algorithm (its final output).

Note that the GA parameters used in this study—elitism rate, parent-selection range, mutation rate, number of generations, and so on—were set experimentally, with the primary aim of verifying the feasibility of applying a GA to the generation of research-theme proposals; optimizing these parameters is left for future work. Their effect is assessed against budget-matched baselines in Section 4.

3.4. Selection Criterion and Outcome Measures on Held-Out Data

The overall score $C(h')$ of a hypothesis h' is a linear combination of novelty $N(h')$ and utility $U(h')$ (Equation (26)), identical to the fitness f of Equation (21); the hypothesis that maximizes $C(h')$ during a run is adopted.

$$C(h') = \alpha N(h') + \beta U(h') \tag{26}$$

Here, α and β are weighting coefficients that determine the importance of each indicator; the proposed model sets $\alpha = 0.6$ and $\beta = 0.4$. This reflects a design that places slightly greater priority on unprecedented novelty than on utility.

Because $C(h')$ is the optimized quantity, the model is evaluated with outcome measures computed on the two partitions that the search never accesses: K_e (held-out papers from the same period as K_s) and K_f (papers posted between July 2024 and December 2025, i.e., after the period covered by K_s). Held-out novelty N_{out} measures the distance from h' to unseen past literature (Equation (27)).

$$N_{out}(h') = 1 - \max_{k \in K_e} cos\left(\varphi(h'),\ v_k\right) \tag{27}$$

As a check that does not depend on the encoder, N_{out} is also computed with TF–IDF vectors fitted on K_e (denoted $N_{out}^{\text{TF–IDF}}$), and held-out link support U_{out} is obtained by evaluating Equation (19) on K_e instead of K_s.

Future similarity FS measures how close the nearest later paper comes to h' (Equation (28)).

$$FS(h') = \max_{k \in K_f} cos(\varphi(h'),\ v_k) \tag{28}$$

Anticipation Δ compares the nearest future paper with the nearest held-out past paper (Equation (29)); $\Delta > 0$ means that the literature moved toward h' after the period covered by the search partition, in the spirit of temporal hold-out validation [17,18].

$$\Delta(h') = FS(h') - (1 - N_{out}(h')) \tag{29}$$

Because Δ contains N_{out} itself, hypotheses that are far from past papers tend to obtain larger Δ mechanically. Δ is therefore also reported as a novelty-matched percentile: the Δ of h′ is ranked among all hypotheses of the 2700-element space whose N_{out} lies within ± 0.01 of $N_{out}(h')$, so that 50 corresponds to chance. These outcome measures are proxies rather than validated measures of scientific value. Because N_{out} and U_{out} apply the same scoring functions as the search to a disjoint set of papers, they test whether the properties of the selected hypotheses hold out of sample, that is, whether the search overfits the search partition; they do not provide independent evidence of scientific value. FS and Δ add a temporal dimension but rely on the same encoder. Checks that are independent of these scoring functions are the validation of the novelty measure against expert peer review (Section 4.7) and the blinded human evaluation (Section 4.6). Table 6 illustrates the link-support term with the Method–Problem links that are most and least supported in K_s.

Table 6. Method–Problem links with the highest and lowest literature support s(m, p) in K_s.

Link (m–p)	$s(m, p)$	Nearest Paper in K_s
Federated Learning–Privacy Leakage	0.856	Constrained Differentially Private Federated Learning for Low-bandwidth Devices
Federated Learning–Adversarial Vulnerability	0.799	Survey on Federated Learning Threats: concepts, taxonomy on attacks and defences, experimental study and challenges
Causal Inference–Adversarial Vulnerability	0.771	Mitigating Adversarial Vulnerability through Causal Parameter Estimation by Adversarial Double Machine Learning
Transformer–Miscalibrated Uncertainty	0.347	Looped Transformers are Better at Learning Learning Algorithms
Diffusion Model–Lack of Novelty	0.364	Diffusion Models in Bioinformatics: A New Wave of Deep Learning Revolution in Action
Diffusion Model–Miscalibrated Uncertainty	0.379	Diffusion Models in Vision: A Survey

3.5. Manuscript Generation

The research-theme proposal (hypothesis h') generated in the previous section produces the following four key pieces of data, which are used to automatically generate an academic manuscript in Markdown format via a template-engine-style approach.

- method: the method (e.g., Transformer, Quantum Annealing).
- problem: the problem to be solved (e.g., Hallucination, Bias).
- goal: the ultimate goal (e.g., Creativity, Efficiency).
- scores: the computed novelty and utility scores.

Specifically, the above variables are embedded into a predefined manuscript outline (template) using Python's *f*-string feature [28]. For example, the title and abstract are dynamically generated as follows.

Example (logic within the code):
"# {self.h.name}: A Novel Approach to {self.h.problem}"
↓
"Quantum Annealing for Efficiency: A Novel Approach to Data Scarcity"

In addition to simply filling in words, the output is structured by section, as shown in Table 7, so that it reads as a proper academic manuscript. The "$J(\theta)$ = Ltask + λRnovelty" shown in the table is an illustrative example of how the "idea" generated by the proposed model is formalized as a machine-learning optimization problem.

Table 7. Illustrative output of manuscript generation.

Section	Content-Generation Logic
Title/Abstract	States the hypothesis name, the problem being solved, and the evaluated scores (novelty and utility)
Introduction	Explains, using standard phrasing, why the problem is important and why the proposed method is effective
Proposed Method	Describes the method's core algorithm and how it contributes to optimization
Mathematical Formulation	Inserts the objective function $J(\theta)$ = Ltask + λRnovelty using LaTeX to give it a mathematical form
Results/Conclusion	Presents the final score as a simulation result and claims that it surpasses the baseline

3.6. LLM-Based Manuscript Writing for the Evaluation Experiment

For the comparative evaluation reported in Section 4, the manuscript-generation step described above was extended from simple f-string template filling to full natural-language generation by a large language model (LLM). The four pieces of hypothesis data (method, problem, goal, and the computed novelty, utility, and score) are passed as grounding facts to a locally hosted, open-weight LLM (gpt-oss-120b, served through the current release of llama.cpp on the researcher's own hardware, with no live web access; the genetic algorithm and the evaluation scripts described in Section 4 were implemented in Python 3), together with a system prompt that (i) requires a complete eight-section manuscript—Abstract; 1. Introduction; 2. Related Work; 3. Proposed Method, including a full mathematical formulation; 4. Experimental Design, presented honestly as a plan for future validation rather than as completed experiments; 5. Reproducibility; 6. Discussion of Applications Across Multiple Domains; 7. Limitations; 8. Conclusion—and (ii) explicitly forbids inventing citations, datasets, or specific numerical results beyond the facts supplied by the genetic algorithm. This LLM-based writer, rather than the f-string template illustrated in Table 7, produced Condition (A) of the pilot case study reported in Sections 4.8–4.10; the template-filling approach remains a valid, lighter-weight configuration of the proposed model but was not the configuration used in the pilot.

This configuration leaves the genetic algorithm itself unchanged: the method, problem, goal, and their novelty and utility scores are still computed exactly as described in Sections 3.2–3.4, with no numeric value invented or adjusted by the LLM. The LLM's role is confined to expressing these already-computed facts in prose, section structure, and, where relevant, an illustrative mathematical formulation; it does not perform, and is instructed not to claim to have performed, any experiment, literature search, or data analysis

beyond what the genetic algorithm itself produced. The pilot manuscripts were generated with the earlier configuration of the GA (Section 4.8). The released writer script uses the revised GA and records, for every call, the GA seed, the LLM sampling seed, the decoding settings (temperature 0.7, top-p 1.0, maximum 6000 tokens), the exact prompts, and the raw response.

Separately from the LLM-based writer described above, an AI assistant (Claude Opus 5.5, Anthropic) was used outside the experimental conditions to help draft and revise the prose of this manuscript and, during the revision, to help implement and run the experiment code of Sections 3.1–3.4 and 4.1–4.4 under the author's direction. The author designed the study, specified the experiments, inspected the code, data, and outputs, and verified the reported numbers against the released logs. All scientific claims, interpretations, and conclusions in this manuscript are the sole responsibility of the author.

4. Experimental Results

4.1. Experimental Design

Five experiments and one validation study were conducted with the revised model (Sections 3.1–3.4). Experiments 1–3 concern the search itself (Sections 4.2–4.4); Experiment 4 generates manuscripts under controlled conditions (Section 4.5); Experiment 5 is a blinded human evaluation of hypotheses (Section 4.6); and Section 4.7 tests the novelty measure against expert peer review. In Experiments 1–3, all search methods share the same fitness function (Equation (21)), the same search partition K_s (1428 papers), the same concept set (2700 hypotheses), and the same budget of 200 fitness evaluations, which corresponds to the GA configuration of Section 3.3 (20 generations $\times$ 10 individuals) and covers about 7% of the hypothesis space. Each stochastic method was run with 30 random seeds (1–30).

Experiment 1 (search effectiveness). The GA was compared with two budget-matched baselines: uniform random search, which evaluates 200 randomly drawn hypotheses, and stochastic hill climbing, which applies the GA's mutation operator to a single current hypothesis, accepts improvements, and restarts after 25 non-improving steps. Because the hypothesis space is small enough to be evaluated exhaustively, the global optimum of f is known, and each run's result can be expressed as a percentile of the full fitness distribution.

Experiment 2 (outcomes on held-out data). The hypotheses selected in Experiment 1 were evaluated with the outcome measures of Section 3.4 (N_{out}, $N_{out}{}^{\text{TF–IDF}}$, U_{out}, FS, Δ, and the novelty-matched Δ percentile). Two objective ablations isolate the two terms of the fitness function (GA with $\alpha = 1$ and $\beta = 0$, "novelty-only"; GA with $\alpha = 0$ and $\beta = 1$, "utility-only"), and two references without search are included: 30 hypotheses drawn uniformly at random ("uniform") and the 30 hypotheses closest to existing papers in K_s ("typical-30"), which represent the statistically most plausible combinations within the concept set. In addition, 30 hypotheses proposed freely by an LLM ("LLM-proposed"; the writer model of Experiment 4, asked for a novel and valid Method–Problem–Goal hypothesis without being given the concept set) were evaluated with the same measures, and the diversity of the hypotheses selected by each condition was quantified as their mean pairwise cosine similarity under φ. To test the robustness of the LLM results, the proposals were repeated with a second model family (Llama-3.1-8B-Instruct [29], 8-bit), with a prompt without the generative-AI restriction of the original prompt, and at a higher temperature (1.0), giving eight configurations of 30 hypotheses each; in a scope-matched variant, each model was given the GA's concept vocabulary and asked to choose a novel and valid Method–Problem–Goal combination from it (30 runs per model). The design, measures, and tests of these additional runs were fixed before the data were generated; diversity was compared with the GA by permutation tests and N_{out} by Mann–Whitney tests, with Holm correction (Appendix B).

Experiment 3 (sensitivity to knowledge-space size). The novelty of all 2700 hypotheses was recomputed against random subsets of K_s of 3, 10, 30, 100, 300, and 1000 papers (10 repetitions each) and against the three reference papers of the earlier version of the model, and the resulting rankings were compared with the ranking obtained from the full search partition.

Experiment 4 (controlled manuscript generation). Ten manuscripts were generated from each of three hypothesis sources—the GA, uniform random selection from the same concept set, and the LLM itself—with the writer model, prompts, decoding settings, and seeds held fixed (Section 4.5).

Experiment 5 (blinded human evaluation of hypotheses). Thirty sets of three hypotheses, one from each source and presented in a uniform format with the source hidden, were ranked by independent raters for usefulness and novelty (Section 4.6).

Validation of the novelty measure. The literature-distance novelty measure was applied to ICLR submissions with public reviews and correlated with the reviewers' scores (Section 4.7).

Differences between the GA and each other condition were tested with two-sided Mann–Whitney U tests [30] with Holm correction over the six comparisons per measure [31]; effect sizes are reported as the Vargha–Delaney A_{12} statistic [32] (A_{12} = 0.5, no difference; A_{12} > 0.5, GA larger), and 95% confidence intervals of means were obtained from 10,000 bootstrap resamples. All experiments ran on a CPU (Python 3.13, sentence-transformers 5.1.0) and complete in about two minutes once the knowledge space has been built.

4.2. Search Effectiveness (Experiment 1)

Table 8 and Figure 1 summarize Experiment 1. All three search methods reached the top 1% of the fitness distribution within the budget (GA: 99.2nd percentile on average; random search and hill climbing: 99.6th). The GA did not outperform either baseline: its mean best fitness (0.564) was slightly, but not significantly, lower than that of random search (0.568; A_{12} = 0.38, p_{Holm} = 0.28) and hill climbing (0.569; A_{12} = 0.37, p_{Holm} = 0.28), and it found the global optimum ("Diffusion Model for Autonomy to solve Distribution Shift", f = 0.586) in 1 of 30 runs, compared with 3 of 30 for hill climbing and 0 of 30 for random search. The GA's best-so-far curve flattens after about 100 evaluations (Figure 1), consistent with a loss of diversity in a population of ten individuals with two elites. The 30 GA runs returned 19 distinct hypotheses, which confirms that a single run is not representative and that results must be aggregated over seeds.

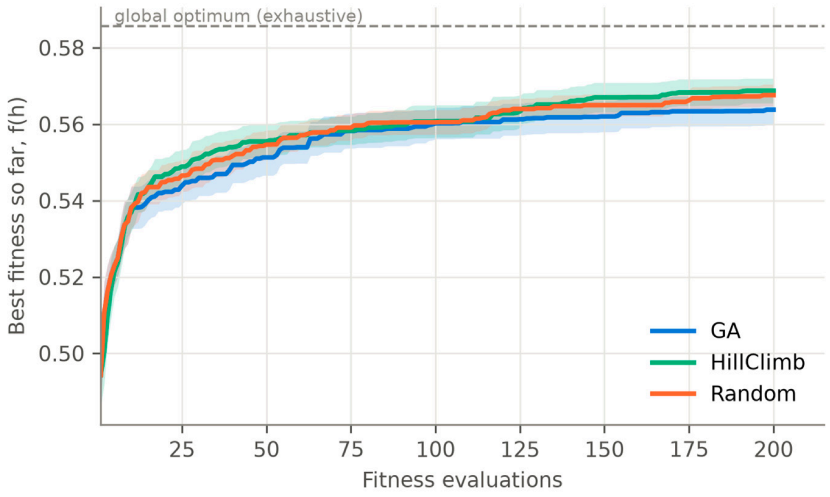


Figure 1. Best fitness found as a function of the number of fitness evaluations (mean of 30 runs; shaded band: 95% confidence interval). Dashed line: global optimum obtained by exhaustive evaluation of all 2700 hypotheses.

Table 8. Experiment 1: best fitness reached with a budget of 200 evaluations (30 runs per method; mean (SD)).

Method	Best Fitness f	Percentile of f	Global Optimum Found	Distinct Hypotheses	A_{12} of GA vs. Method (p_{Holm})
GA (proposed)	0.564 (0.011)	99.2 (1.0)	1/30	19	—
Random search	0.568 (0.008)	99.6 (0.4)	0/30	14	0.38 (0.28)
Hill climbing	0.569 (0.009)	99.6 (0.4)	3/30	16	0.37 (0.28)

Note: — Not applicable: the GA is the reference method of the comparison.

4.3. Outcomes on Held-Out Data (Experiment 2)

Table 9 and Figure 2 summarize Experiment 2. Novelty measured during search generalized to unseen literature: across all 2700 hypotheses, N computed on K_s and N_{out} computed on K_e were strongly correlated (Spearman $\rho = 0.75$), and hypotheses selected with the novelty–utility objective were more distant from held-out papers than uniformly drawn hypotheses (GA: 0.515 vs. 0.461; $A_{12} = 0.78$, $p_{Holm} < 0.001$) and far more distant than the typical-30 reference (0.259). The utility term had the intended effect: compared with novelty-only search, the GA retained higher held-out link support ($U_{out} = 0.523$ vs. 0.472; $A_{12} = 0.78$, $p_{Holm} < 0.001$) at a small, non-significant cost in novelty (0.515 vs. 0.533; $p_{Holm} = 0.44$). Novelty and link support are strongly antagonistic across the hypothesis space ($\rho = -0.82$), so the objective selects a compromise rather than maximizing both. The encoder-independent TF–IDF novelty agreed only weakly with the embedding-based measure ($\rho = 0.21$ across the space); on that measure, the GA again exceeded the typical-30 reference (0.786 vs. 0.745; $p_{Holm} < 0.001$) but not uniform sampling (0.769; $p_{Holm} = 0.31$). Random search and hill climbing, which optimize the same objective, did not differ significantly from the GA on any outcome measure.

Table 9. Experiment 2: outcome measures on held-out data (30 hypotheses per condition; mean (*SD*)). Higher N_{out}: more distant from unseen past papers; higher U_{out}: stronger literature support for the pairwise links; $\Delta > 0$: later papers came closer than earlier ones; matched percentile 50 = chance. f is always evaluated with $\alpha = 0.6$ and $\beta = 0.4$.

Condition	f	N_{out}	N_{out} (TF–IDF)	U_{out}	FS	Δ	Δ, Matched Pct.
GA (proposed)	0.564 (0.011)	0.515 (0.039)	0.786 (0.025)	0.523 (0.046)	0.438 (0.073)	−0.047 (0.090)	21.5 (27.4)
Random search	0.568 (0.008)	0.502 (0.051)	0.796 (0.017)	0.496 (0.057)	0.409 (0.069)	−0.089 (0.076)	10.6 (22.3)
Hill climbing	0.569 (0.009)	0.511 (0.044)	0.780 (0.031)	0.520 (0.047)	0.417 (0.058)	−0.072 (0.079)	13.7 (18.9)
GA, novelty-only	0.566 (0.013)	0.533 (0.043)	0.793 (0.024)	0.472 (0.044)	0.396 (0.058)	−0.072 (0.073)	10.1 (20.1)
GA, utility-only	0.454 (0.016)	0.283 (0.048)	0.719 (0.034)	0.713 (0.030)	0.710 (0.058)	−0.007 (0.050)	60.1 (27.2)
Uniform (no search)	0.502 (0.027)	0.461 (0.061)	0.769 (0.039)	0.550 (0.047)	0.561 (0.071)	0.022 (0.062)	55.1 (29.8)
Typical-30 (no search)	0.413 (0.015)	0.259 (0.051)	0.745 (0.037)	0.706 (0.041)	0.741 (0.046)	0.000 (0.053)	60.4 (29.0)
LLM-proposed	0.483 (0.014)	0.435 (0.042)	0.811 (0.022)	0.549 (0.033)	0.580 (0.041)	0.014 (0.029)	56.4 (18.1)

The temporal hold-out did not support the expectation that novelty-seeking search anticipates later research. The GA's hypotheses had a negative mean anticipation ($\Delta = -0.047$; 95% CI [−0.079, −0.015]); that is, their nearest future paper was, on average, farther away than their nearest held-out past paper, whereas uniformly drawn hypotheses had $\Delta = 0.022$ ($p_{Holm} = 0.017$). After matching for held-out novelty, GA hypotheses fell at the 21st percentile of Δ on average (chance = 50; uniform: 55th, $p_{Holm} < 0.001$); novelty-only search and random search with the same objective fell even lower (10th–11th percentile), whereas the utility-only and typical-30 conditions were at or above chance (60th). Within the 18-month horizon examined, the literature therefore did not move toward hypotheses selected for their distance from existing work; whether such hypotheses are premature, implausible, or merely outside the scope of the sampled categories cannot be decided from these data.

Across all hypotheses, Δ was unrelated to the fitness f ($\rho = -0.01$) and to the utility term ($\rho = -0.01$).

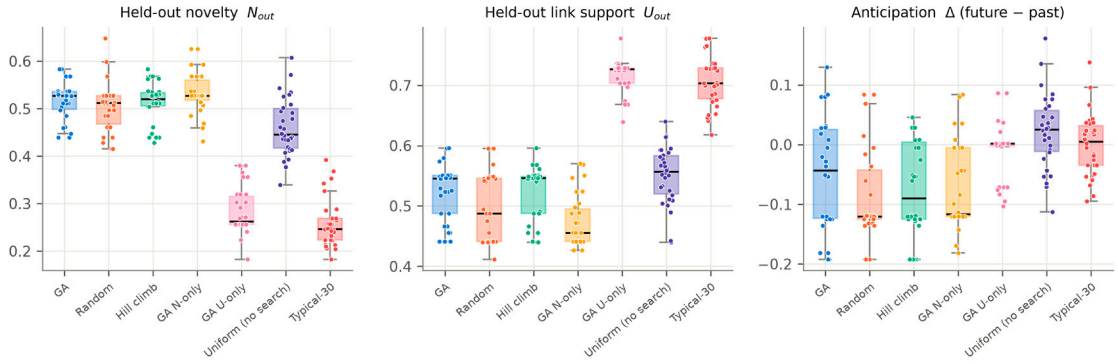


Figure 2. Held-out outcome measures of the hypotheses selected in each condition (30 per condition; boxes: interquartile range; dots: individual runs).

The hypotheses proposed freely by the LLM differed from all search conditions (Table 9, last row). They were closer to existing papers than the GA's hypotheses on the embedding-based measure (N_{out} = 0.435 vs. 0.515; A_{12} of the LLM vs. the GA = 0.08, $p_{Holm} < 0.001$), but more distant on the lexical TF–IDF measure (0.811 vs. 0.786), because they are phrased in longer and more specific technical terms; the two novelty measures thus disagree precisely where wording and content diverge. The LLM's hypotheses were also closer to the literature published afterwards (FS = 0.580 vs. 0.438; $p_{Holm} < 0.001$) and had a slightly positive anticipation (Δ = 0.014 vs. −0.047; $p_{Holm} = 0.017$; novelty-matched percentile 56), i.e., they follow the direction in which the field was already moving. They were also the least diverse among the conditions that produce distinct hypotheses: their mean pairwise similarity was 0.513, compared with 0.290 for the GA and 0.284 for uniform selection (random search, 0.452, and typical-30, 0.722, are concentrated because they repeatedly converge on the same few combinations), and 15 of the 30 LLM-proposed methods were variants of neuro-symbolic learning, 11 of them addressing hallucination. Although all 30 LLM-proposed hypotheses were worded differently, they thus concentrated on a few mainstream themes, whereas the GA's 30 runs produced 19 distinct and widely spread combinations.

Diversity was the only respect in which the GA differed from random search with the same objective and budget. At equal fitness (0.564 vs. 0.568), the 30 GA runs returned less similar hypotheses than the 30 random-search runs (mean pairwise similarity 0.290 vs. 0.452; difference 0.162, 95% bootstrap CI [0.071, 0.248]; two-sided permutation test, $p_{Holm} = 0.005$). The numbers of distinct hypotheses (19 vs. 14) did not differ significantly ($p_{Holm} = 0.39$), and the GA did not differ from hill climbing on either measure (0.334; $p_{Holm} = 0.48$). Across repeated runs, the GA thus yields a somewhat wider set of high-fitness hypotheses than random search; this may partly reflect its early convergence to different local optima (Section 4.2) rather than a more effective exploration of the space.

The differences between LLM-proposed and GA-selected hypotheses did not depend on the particular model, prompt, or temperature (Table A1). In all eight LLM configurations, the proposals were less diverse than the GA's (mean pairwise similarity 0.446–0.518 vs. 0.290) and closer to held-out literature (N_{out} 0.373–0.435 vs. 0.515; A12 $\leq$ 0.08; all $p_{Holm} < 0.001$), and they were closer to later literature (FS 0.567–0.610 vs. 0.438). The specific themes, however, differed between models: neuro-symbolic methods appeared in 27–53% of Nemotron's proposals but in none of Llama's, so the concentration on particular themes is model-specific, whereas the lower diversity and the smaller distance from the literature were common to all configurations. Matching the scope did not remove the difference. Given the GA's concept vocabulary, Nemotron chose the same combination (retrieval-

augmented generation, hallucination, trustworthiness) in all 30 runs, and Llama chose 22 distinct combinations that were less diverse (0.335; p_{Holm} = 0.034) and markedly closer to held-out literature (N_{out} 0.380; A12 = 0.04; p_{Holm} < 0.001) than the GA's. A difference in budget remains: the GA evaluates 200 candidates under an explicit objective, whereas each LLM answer is a single proposal. A comparison in which the same objective selects from a pool of LLM proposals was not carried out, so the comparison concerns the two approaches as they are typically used and does not isolate the generation mechanism.

4.4. Sensitivity to Knowledge-Space Size (Experiment 3)

Table 10 and Figure 3 summarize Experiment 3. With few reference papers, novelty values were compressed toward the top of the scale, and their ranking of hypotheses agreed poorly with the ranking obtained from the full search partition: with three random papers, the mean novelty was 0.78 and the Spearman correlation with the full-corpus ranking was 0.33; with the three reference papers of the earlier version, it was only 0.08, i.e., practically unrelated to the full-corpus ranking. Agreement exceeded 0.8 only from about 300 papers onward (0.82 at 300 and 0.95 at 1000 papers). The three-paper knowledge space of the earlier version therefore could not measure novelty meaningfully, which explains the ceiling-level Originality scores in the pilot case study (Section 4.10).

Table 10. Experiment 3: novelty of all 2700 hypotheses as a function of knowledge-space size (random subsets of K_s, 10 repetitions; mean). N can slightly exceed 1 because cosine similarity can be negative.

Papers in Knowledge Space	Mean N	SD of N	Range of N	ρ with Full-Corpus Ranking
3 (reference papers of the earlier version)	0.770	0.087	0.48–1.02	0.08
3 (random)	0.778	0.087	0.48–1.02	0.33
10	0.695	0.096	0.35–0.96	0.34
30	0.629	0.097	0.31–0.92	0.48
100	0.549	0.088	0.25–0.82	0.61
300	0.491	0.088	0.18–0.76	0.82
1000	0.451	0.081	0.17–0.73	0.95
1428 (full K_s)	0.440	0.080	0.15–0.72	1.00

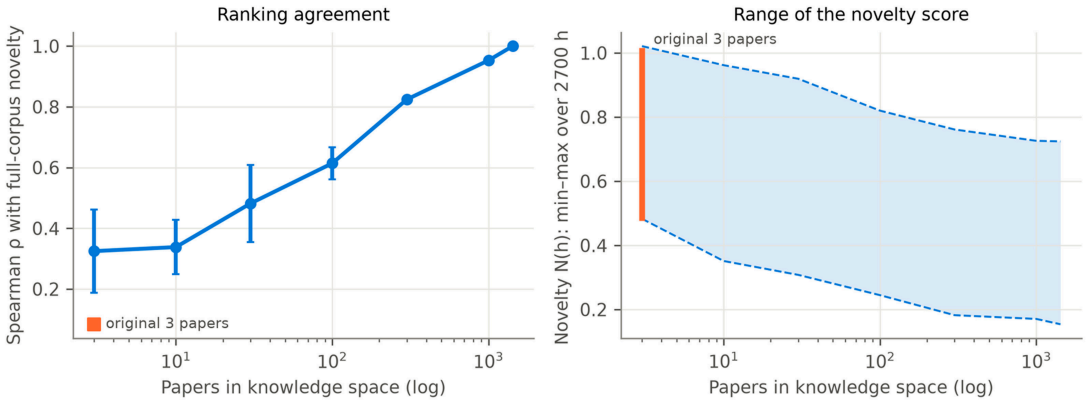


Figure 3. Effect of knowledge-space size on the novelty measure. Left: rank agreement with the full-corpus novelty (mean ± SD over 10 random subsets); square: the three reference papers of the earlier version. Right: range of novelty values over the 2700 hypotheses.

4.5. Controlled Manuscript Generation (Experiment 4)

Experiment 4 removes the confounds of the earlier manuscript comparison (Section 4.8) by varying only the source of the hypothesis. Ten manuscripts were generated from each

of the three sources: the revised GA (seeds 1–10), uniform random selection from the same concept set (seeds 1–10), and the writer model itself, asked to propose a novel and valid Method–Problem–Goal hypothesis without the concept set (seeds 1–10). All 30 manuscripts were written by the same open-weight model (NVIDIA Nemotron-Nano-9B-v2 [33], 8-bit quantized, served with llama.cpp [34]) with the same system prompt, temperature 0.7, top-p 0.95, internal reasoning disabled, and a writer seed of 100 + s for seed s; the writer received only the Method, Problem, and Goal phrases—no scores, no information on how the hypothesis was obtained, and no other context. Because the model ended its answer after the Abstract when asked for a complete manuscript in a single request, every manuscript was written with the same fixed sequence of nine requests, one per section, each seeing the sections already written. The resulting manuscripts were comparable in length (mean 1827, 1806, and 1821 words for the GA, random, and LLM sources) and all contained the nine required sections. The model used for Condition A of the pilot (gpt-oss-120b) could not be used here because it does not fit in the memory of the available hardware.

The manuscripts were first judged by an LLM from a different model family (Llama-3.1-8B-Instruct [29], 8-bit) under their blinded codes, with the rubric of Appendix A (originality, soundness, reproducibility, and significance on a 1–5 scale; three runs per manuscript). The judge was uninformative: it gave the maximum score for reproducibility and significance in all 90 judgments and for soundness in 85 of 90, and its mean overall score was practically identical across sources (GA 4.74, random 4.78, LLM 4.83). Small LLM judges therefore cannot discriminate between such manuscripts on an absolute scale.

A larger judge (Llama-3.3-70B-Instruct [29], 3-bit quantized) was then applied in two ways. With the same absolute rubric it was equally uninformative: it gave every manuscript 4 for originality, 4 for soundness, and 5 for reproducibility in all 90 judgments. In pairwise comparisons (90 pairs, each manuscript in six pairs, every pair judged in both presentation orders), its verdicts were dominated by position: for the overall verdict, the manuscript shown first won in 85% of the judgments, and the same manuscript won in both orders in only 30% of the pairs (originality: 71%). Among the order-consistent decisions, manuscripts built on LLM-proposed hypotheses were preferred more often than GA-based manuscripts (originality: 14 vs. 2 decisions; overall: 8 vs. 0) and than random-based manuscripts (originality: 13 vs. 3), whereas GA- and random-based manuscripts were preferred about equally often (originality: 7 vs. 9). Because each manuscript appears in six pairs, these decisions are not independent, and given the low order consistency they are reported descriptively, without significance tests. This pattern is notable for two reasons. First, the judge (Llama) and the writer (Nemotron) belong to different model families, so it is not a preference for the judge's own outputs. Second, all 30 manuscripts were written by the same writer with the same prompts and settings, so they differed only in the hypothesis they developed; the preference therefore concerns the origin of the idea rather than the wording. LLM judges are known to favor their own outputs and particular positions [35], and LLMs of different families favor texts written by LLMs over texts written by humans [36]; the present pattern is compatible with such a preference extending from LLM-written text to LLM-originated ideas. The human raters did not show this preference (Section 4.6). The LLM-judge verdicts are therefore not used as evidence of hypothesis quality; human judgment was applied at the level of the hypotheses (Section 4.6).

4.6. Blinded Human Evaluation of Hypotheses (Experiment 5)

Because the central question of this study is whether the GA selects better hypotheses, human raters evaluated hypotheses rather than full manuscripts, which also keeps the task short enough for careful rating. Thirty hypotheses were taken from each source: the first 30 distinct results of GA runs with seeds 1–59, the 30 uniformly drawn hypotheses

of Experiment 2, and the 30 LLM-proposed hypotheses. To prevent the source from being recognizable, every hypothesis was rewritten by the writer model of Experiment 4 with one fixed prompt into a one-sentence hypothesis (20–30 words) and a two- to three-sentence description (50–70 words); sentences longer than 30 words were shortened with one fixed prompt, and vocabulary terms were written in ordinary lowercase. The rendered items had similar lengths (hypothesis sentence: mean 22.8, 21.9, and 23.0 words for the GA, random, and LLM sources). The hypotheses were grouped into 30 sets of three, one per source, with the positions A, B, and C randomized within each set and the order of the sets randomized for each rater; a practice set preceded the 30 sets.

For each set, the raters chose the most and the least useful hypothesis—usefulness being judged from the importance of the problem, the plausibility of the method for the problem and goal, and feasibility—and, separately, the most and the least novel hypothesis (Appendix A). The two raters were researchers in AI-related fields (information science and computational science; information and communication technology) who were not involved in the design, execution, or analysis of the study and were blind to the hypothesis sources. Because their expertise is sufficient to judge usefulness but not to judge originality relative to the specialized AI literature, the novelty rankings are reported as supplementary. As a supplementary third rater, an LLM judge (Llama-3.3-70B-Instruct [29], quantized, run locally with llama.cpp [34]) ranked the same 30 sets independently with the same instructions, each set in three rotated position orders to control for position bias. Ranks within a set (best = 1, worst = 3) were compared across sources with the Friedman test and the Holm-corrected Wilcoxon signed-rank tests, and agreement between raters was assessed with Krippendorff's α [37] and the share of sets with the same choice.

The human raters tended to rank the LLM-proposed hypotheses as least useful: these were ranked worst in 47% of the judgments and best in 22% (mean rank 2.25), compared with mean ranks of 1.83 for the GA and 1.92 for random selection. The differences were not statistically significant (Friedman $\chi^2 = 4.07$, $p = 0.13$; Holm-corrected Wilcoxon tests, $p_{Holm} \geq 0.19$), and agreement between the two raters was low (Krippendorff's $\alpha = 0.14$). Although the GA's advantage was not significant, the human evaluation gave no indication that the wider exploration by combinatorial search (Section 4.3) came at the expense of usefulness: the GA hypotheses received the best mean usefulness rank. The novelty rankings did not differ between sources (mean ranks 2.08, 2.07, and 1.85 for the GA, random, and LLM sources; Friedman $p = 0.21$), and the two raters disagreed systematically ($\alpha = -0.24$), which confirms that novelty relative to the AI literature could not be judged reliably by these raters.

The LLM judge ranked the LLM-proposed hypotheses highest on both criteria: for usefulness, mean ranks were 1.73 for the LLM-proposed, 2.09 for the GA, and 2.18 for the random hypotheses (Friedman $p = 0.10$); for novelty, 1.62, 2.12, and 2.26 (Friedman $\chi^2 = 12.2$, $p = 0.002$; LLM-proposed vs. GA and vs. random, $p_{Holm} = 0.012$). Its rankings were, however, unstable across the three position orders, which agreed on the best hypothesis in only 30% of the comparisons, about the chance level of 33% (Krippendorff's $\alpha = 0.24$), and they were practically unrelated to the human rankings (Spearman $\rho = 0.09$ for usefulness and 0.21 for novelty). The judge thus ranked as most novel the hypotheses that were closest to existing literature (Section 4.3), which is a further indication of its preference for LLM-originated content (Section 4.5) and not evidence of their novelty.

4.7. Validation of the Novelty Measure Against Expert Peer Review

To test whether literature distance reflects what experts judge, the novelty measure of Section 3.4 was applied to ICLR submissions with public reviews from OpenReview, as compiled in the ReviewArena dataset [38] (CC BY 4.0). In this dataset, ICLR 2024 and 2025

contain accepted papers with per-review scores for contribution, soundness, presentation, and overall rating, and ICLR 2023 contains accepted and rejected papers with overall ratings. No year provides an explicit originality score; contribution is the closest construct. For each paper, novelty was computed against prior literature only—the arXiv knowledge space of Section 3.1 restricted to papers posted before the submission deadline, with or without all earlier ICLR papers—and for three text granularities (title and abstract, title, and the first sentence of the abstract, the last being closest to a hypothesis).

The novelty measure was essentially unrelated to the experts' judgments (Table 11). Across all 48 combinations of year, score, reference corpus, and granularity, Spearman correlations ranged from −0.015 to 0.059; the correlation with contribution was 0.046 for ICLR 2024 and 0.029 for ICLR 2025, and novelty predicted acceptance at ICLR 2023 barely above chance (AUC 0.51–0.53). Restriction to accepted papers narrows the range of scores in 2024 and 2025, but the result was the same for 2023, which covers the full range of quality. Literature distance as measured here is therefore not a valid proxy for originality as judged by expert reviewers, and the distance results of Experiments 1–3 are interpreted accordingly (Section 5).

Table 11. Spearman correlation between the novelty measure and expert review scores (title and abstract; reference: prior arXiv papers and earlier ICLR papers; 95% bootstrap confidence interval). The last column gives the range over all six combinations of reference corpus and text granularity. Acceptance at ICLR 2023 (1573 of 3792 papers accepted) is coded 1/0; the corresponding AUC was 0.508 (range over variants 0.508–0.529).

Venue (Papers)	Review Score	Spearman ρ [95% CI]	Range over Variants
ICLR 2024 (2260 accepted)	Contribution	0.046 [0.003, 0.087]	0.021 to 0.050
	Soundness	0.019 [−0.022, 0.060]	−0.005 to 0.037
	Rating	0.050 [0.009, 0.090]	0.011 to 0.050
ICLR 2025 (3703 accepted)	Contribution	0.029 [−0.002, 0.062]	−0.015 to 0.029
	Soundness	0.045 [0.010, 0.078]	−0.002 to 0.045
	Rating	0.025 [−0.007, 0.057]	0.004 to 0.025
ICLR 2023 (3792 in total)	Rating	0.026 [−0.005, 0.058]	0.006 to 0.059
	Acceptance	0.014 [−0.018, 0.044]	0.014 to 0.050

4.8. Pilot Case Study: Manuscript-Generation Conditions

The following pilot case study, carried out before the experiments above, illustrates the end-to-end pipeline from hypothesis to manuscript. It used the earlier configuration of the model—a knowledge space of three reference papers, content-free hash vectors for novelty, and three hand-written utility rules ("Quantum" in M and "Latency" in P, "Symbolic" in M and "Hallucination" in P, and "Evolutionary" in M and "Creativity" in G, each worth +0.4)—and generated a single manuscript per condition. It is reported for transparency and is not used as evidence for the GA. The experiment compared three manuscript-generation conditions, defined as follows. (A) The proposed model alone: the genetic algorithm selects the Method–Problem–Goal hypothesis exactly as described in Section 3, and the same proposed model—using the LLM-based writer described in Section 3.6, with no external agent and no live web access—produces the complete manuscript by itself. (B) The proposed model combined with an existing AI agent: the same genetic-algorithm-selected hypothesis (the Python file's output) is instead handed to Google Gemini 3's Deep Research mode—an existing AI agent equipped with autonomous, extensive web research—which writes the manuscript. (C) An existing AI agent alone: Google Gemini 3's Deep Research mode devises its own research theme and writes the manuscript with no genetic-algorithm involvement whatsoever.

- (A) Run the proposed model's genetic-algorithm search as described in Section 3, then pass its output (method, problem, goal, novelty, utility, and score) to the LLM-based writer described in Section 3.6 to produce the manuscript, with no external agent involved at any stage.
- (B) Run the proposed model's genetic-algorithm search as in (A), then submit its output together with the following query to Google Gemini 3's Deep Research mode prompt:

"Using the model (Python file) from the previous section, conduct a survey of related work in the field of 'generative AI,' devise a research theme with novelty and validity, carry out the research, and compile the results as a manuscript with the following eight sections: Abstract; 1. Introduction; 2. Related Work; 3. Proposed Method; 4. Experimental Design; 5. Reproducibility; 6. Discussion of Applications Across Multiple Domains; 7. Limitations; 8. Conclusion."

- (C) Submit the following query to Google Gemini 3's Deep Research mode prompt directly, with no genetic-algorithm output supplied and no other involvement of the proposed model:

"Conduct a survey of related work in the field of 'generative AI,' devise a research theme with novelty and validity, carry out the research, and compile the results as a manuscript with the following eight sections: Abstract; 1. Introduction; 2. Related Work; 3. Proposed Method; 4. Experimental Design; 5. Reproducibility; 6. Discussion of Applications Across Multiple Domains; 7. Limitations; 8. Conclusion."

Both (B)'s and (C)'s prompts explicitly require the same eight-section structure as the LLM-based writer used in (A) (Section 3.6). This was a deliberate methodological control: an earlier round of this experiment gave (A) a section structure aligned with the evaluation criteria described in Section 4.2 while leaving (B) and (C) to a generic, unconstrained prompt, which could by itself inflate (A)'s scores independent of the substantive quality of its content. Requiring the identical section structure across all three conditions removes this confound, so that any remaining differences in the scores reported in Section 4.3 reflect differences in content rather than in document structure. However, the conditions differed in other respects that were not controlled: Condition A used a different writer model without web access and with stricter anti-fabrication instructions; Condition B's prompt was accompanied by the GA's Python script itself—structured context beyond the GA output—which Condition C did not receive; and the decoding settings and random seeds of the Gemini runs could not be fixed or recorded. Differences between the pilot conditions therefore cannot be attributed to the GA. The prompts are released verbatim with the code.

4.9. Pilot Case Study: Automated Evaluation Instrument

Referring to the following four items (review criteria) common to the review processes of NeurIPS 2025's Call for Papers [39], ICLR 2026's Author Guide [40], ICML 2026's Reviewer Instructions [41], CVPR 2026's Reviewer Guidelines [42], and AAAI 2026's review process [43]—the major AI conferences as of 2026—these criteria were implemented as a Python script to evaluate manuscripts (A), (B), and (C). This instrument is retained only to document the pilot. It has not been validated against expert judgment: TF–IDF overlap and edit distance do not establish scientific originality, matching hand-written rules does not establish technical correctness, the presence of disclosure terms does not establish reproducibility, and Generality × Utility has no demonstrated relationship with significance.

- Originality: Novelty (SO).

- Quality: Technical Correctness (SQ).
- Reproducibility (SR).
- Significance: Academic and Social Significance (SS).

First, as shown in Equation (30), Originality (SO) places greater weight on contributing to a new paradigm than on making minor modifications to existing methods.

$$S_O = \alpha \cdot Novelty + \beta \cdot Non\text{-}triviality \tag{30}$$

Novelty is conceptually the same as in Equations (27)–(29), but with slight differences to align with conference review processes. Specifically, it combines a duplicate check based on string edit distance, forming a two-stage design that also prevents "imitation of expression" that cannot be captured by vector closeness alone. Non-triviality, as shown in Equation (31), computes the "unexpectedness of the combination of words and concepts" using edit distance (Difflib) [44]. Here, htext denotes the text data of the new hypothesis being evaluated and the text data of existing papers in the knowledge space K'. The score increases with the degree to which the combination is not one that anyone could readily think of—that is, by posing a genuinely original question. Finally, α and β are weighting coefficients; because novelty is weighted more heavily, they are set to $\alpha = 0.6$ and $\beta = 0.4$.

$$Non\text{-}triviality(h) = 1 - \max_{k \in \mathbf{K}'}(gestalt_pattern_matching(h_{text}, k_{text})) \tag{31}$$

Quality (SQ) is evaluated by theoretical rigor and empirical evidence (Equation (32)).

$$S_Q = \mathrm{w}_q\left(\frac{\text{Theoretical_Rigor} + \text{Empirical_Evidence}}{2}\right) \tag{32}$$

This, too, differs from Equation (29) in order to align with conference review processes. Specifically, Theoretical Rigor is a parameter (input value) used in the computation of Equation (29), evaluating the logical connection—and thus the validity—of the method, via the *M*–*P* and *M*–*G* relationships. Empirical Evidence, meanwhile, functions like an individual evaluation criterion for determining δi (the bonus factor) in Equation (29); for example, it checks the fit between *P* (problem) and *G* (goal)—that is, goal validity, or "if this problem is solved, does it really reach the intended goal?"—against past paper data (successful patterns). w_q (weight of quality) is the coefficient that determines how much weight the technical-quality item carries in the final overall score, and is set to 0.25 so that the four evaluation items making up the overall score (Stotal) are weighted equally (25% each). In addition, (Theoretical Rigor + Empirical Evidence)/2 is used because, when each of these two axes is scored on a 1–5 scale, simply summing them would yield a maximum of 10 points; dividing by 2 keeps SQ within the 1–5 range, allowing it to be fairly compared and combined with other items, such as Originality, that are also scored on a 1–5 scale. Note that the rules underlying Theoretical Rigor and Empirical Evidence are the three utility rules of the earlier GA (Section 4.8); SQ therefore partly measures agreement with the GA's own selection criterion and is circular for GA-selected hypotheses.

Reproducibility (SR) scores the likelihood that other researchers can reproduce the same results (Equation (33)).

$$S_R = \gamma_c \cdot Code + \gamma_d \cdot Data + \gamma_p \cdot \text{Parameters} \tag{33}$$

As shown in the equation above, this is evaluated based on whether the paper clearly presents the code, data, and parameters needed for reproduction. γ represents the contribution weight of "how important each element is to a third party attempting reproduction";

because the display of code was emphasized in the peer review of the conferences noted above, $\gamma_c = 0.6$, $\gamma_d = 0.3$, and $\gamma_p = 0.1$ were set.

Significance (SS) is defined, as in Equation (34), as the degree of impact the research has on the community.

$$S_S = \int_{t=0}^{T} \text{Impact(t)dt} \approx \text{Generality} \cdot \text{Utility} \tag{34}$$

Impact(t)dt represents the strength of the paper's "attention" or "usage" at time t, and Generality evaluates whether the work is applicable across multiple fields, such as medicine, materials science, and education, expressed as follows.

$$\text{Generality} = \frac{|D_{applicable}|}{|D_{total}|} \tag{35}$$

Dapplicable represents the set of domains (fields) to which the method (M) or problem-solving approach (P) can be applied, and is computed by counting how many different categories in the knowledge space (e.g., finance, medicine, IT) the keywords contained in the hypothesis are associated with. This association is determined in the following two ways.

- Keyword–Semantic Matching.

The system has a predefined set of keywords (a signature) symbolizing each domain (e.g., finance, medicine, IT); if the text of hypothesis h' (the descriptions of method M and problem P) contains a sufficient number of words from a given domain signature, that domain is judged to be relevant.

For example:

- Finance: {investment, risk, portfolio, assets, banking, transactions}.
- Medical: {diagnosis, imaging, patient, treatment, clinical, drug discovery}.
- IT: {algorithm, development, security, cloud, automation}.
- Indirect Association via the Knowledge Space (K').

In addition to direct keyword matching, association is also performed using existing paper data (the knowledge space K'), as shown in (A) and (B) below.

- Identification of similar papers: existing papers k with content similar to the generated hypothesis h′ are retrieved in vector space (KNN: k-nearest neighbors).
- Category inheritance: the category (domain) to which the similar paper k originally belonged (e.g., an arXiv classification tag) is referenced, and that domain is added to Dapplicable.

The more associated domains there are, the closer Dapplicable approaches 1.0. Dtotal is the total number of domains present in the knowledge space (e.g., medicine, finance, IT, energy, education). The formula for computing Utility is the same as in Equation (29). Because the Utility factor reuses the same rules, SS is likewise circular for GA-selected hypotheses.

The reason the time-integral expression and Generality · Utility are connected by an approximation (≈) is that it represents the relationship whereby the accumulation of academic value (the integral) can be translated and estimated using practical indicators—generality and utility—representing the "social and academic contribution" of the research at time t. In other words, integrating from time 0 to t means computing the cumulative value of how long the idea continues to provide value, rather than ending as a passing trend.

Finally, the final score integrating SO, SQ, SR, and SS, denoted Stotal, is expressed as in Equation (36).

$$S_{total} = \left(\sum_{i \in \{O,Q,R,S\}} w_i \cdot S_i \right) \cdot \varphi(S_R) \quad (36)$$

w_i is the weighting coefficient used to compute the score with equal allocation (25% each) across all four items (Originality, Quality, Reproducibility, Significance), and Si denotes the individual score (1.0–5.0 points) for each evaluation item (SO, SQ, SR, SS). φ(SR) represents the reproducibility filter, a function that drives the overall score to 0, or to Reject, when SR is extremely low.

That is, because SO, SQ, SR, and SS in Equation (36) take values between 1.0 and 5.0 and w_i is 0.25, the computation proceeds as illustrated in Equation (37). In this study, w_i is divided equally, although the weights can also be adjusted case by case depending on which criteria are prioritized.

In implementing this evaluation model for the experiment reported below, the Keyword–Semantic Matching step described above (Equation (35)) and the recovery of a manuscript's underlying Method, Problem, and Goal for the Novelty computation in Equation (30) required one modification beyond the description above. Because manuscripts (B) and (C) are free-form reports generated by an AI agent rather than instances of the proposed model's own templated output, their text does not necessarily contain the literal Method/Problem/Goal vocabulary or domain-signature keywords listed above verbatim; requiring an exact string match would therefore have scored these two manuscripts as having no measurable Method, Problem, Goal, or domain association at all, which would not be a meaningful evaluation. To keep (A), (B), and (C) comparable on the same instrument, exact keyword-containment checks were replaced with TF–IDF (term frequency–inverse document frequency) cosine similarity between the manuscript text and each candidate term or domain signature, so that a manuscript is credited according to its degree of semantic overlap with a term rather than only ever scoring a binary hit or miss; the Indirect Association via the Knowledge Space (KNN) pathway described above was not implemented in this experiment, and Generality was computed solely via the direct keyword–semantic matching pathway. Non-triviality (Equation (31)) and the Reproducibility sub-scores (Equation (33)) are unaffected by this change, since both remain literal checks—edit distance and the disclosure of code, data, or parameters, respectively—rather than keyword-matching questions. This is a genuine extension beyond the keyword-based description above, made necessary by evaluating free-form AI-agent-generated output rather than only the proposed model's own templated manuscripts; (A), (B), and (C) were scored by the identical automated procedure in every respect, with no manual or qualitative step for any condition.

A further refinement was made to the Novelty component of Originality (Equation (30)) itself. An initial implementation computed Novelty as one minus the maximum TF–IDF cosine similarity between a manuscript's full prose and the knowledge-space reference papers' text—that is, how much a manuscript's wording differs from existing literature. This was found to measure the wrong quantity: a manuscript written in specialized or unusual vocabulary can score as highly "novel" by this measure regardless of whether its underlying idea is itself a new combination of Method, Problem, and Goal, since word choice and conceptual novelty are not the same thing. Novelty was therefore redefined to mirror Equations (26)–(29)'s own definition (Novelty(h) = $\min\|z_h - z_\text{known}\|$, i.e., distance from the hypothesis to the nearest existing known combination): the Method, Problem, and Goal terms are first recovered from each manuscript—including (B) and (C)—by the same TF–IDF procedure used to recover the domain signatures above, and from each knowledge-space reference paper's title and

abstract by the identical procedure; Novelty is then one minus the maximum cosine similarity between the manuscript's own (Method, Problem, Goal) combination and each reference paper's recovered combination. This tests whether the specific combination a manuscript is centered on has already been explored in the reference literature, which is the quantity Equation (30) is intended to capture, rather than how idiosyncratically a manuscript happens to be worded. Both the original text-distance figure and the revised combinatorial figure were retained internally for transparency; only the revised, combinatorial figure is used in the Originality scores (SO) reported in Table 8 and discussed in Section 4.3. The released script pilot_manuscript_evaluator.py reproduces every value in Table 12.

$$S_{total} = (S_O \cdot 0.25) + (S_Q \cdot 0.25) + (S_R \cdot 0.25) + (S_S \cdot 0.25) \quad (37)$$

Table 12. Pilot case study: scores of the unvalidated automated instrument (one manuscript per condition; SQ and SS are circular for GA-selected hypotheses).

Evaluation Item	(A)	(B)	(C)
Originality (SO)	4.99/5.0	3.94/5.0	4.98/5.0
Quality (SQ)	4.47/5.0	4.29/5.0	3.08/5.0
Reproducibility (SR)	4.60/5.0	5.00/5.0	5.00/5.0
Significance (SS)	3.54/5.0	2.76/5.0	1.00/5.0
Overall	4.40/5.0	4.00/5.0	3.52/5.0

4.10. Pilot Case Study: Results

The pilot results are summarized in Table 12. Given a single manuscript per condition, the uncontrolled differences between conditions described in Section 4.8, the circularity of SQ and SS, and the lack of validation of the instrument, these scores are descriptive only and are not interpreted as evidence for or against the GA.

Comparing these results against the evaluation model described in the previous section, the reasons for the scores in Table 8 are as follows.

(A)

- Originality: Ties for the highest of the three conditions (SO = 4.99): the manuscript's (Quantum Annealing, High Latency, Robustness) combination shares no vocabulary with any of the three knowledge-space reference papers' own recovered combinations (combinatorial Novelty = 1.00), consistent with the value the genetic algorithm itself computed for this same hypothesis when it was selected (Novelty = 1.00, disclosed in the manuscript's own generation record).
- Quality: Highest of the three conditions (SQ = 4.47): the manuscript's grounding in the genetic algorithm's own computed facts, together with a full mathematical formulation, produces close alignment with the named success-pattern rules of the earlier GA (Section 4.8)—rules that the GA itself used to select this hypothesis, so a high score is expected by construction.
- Reproducibility: Discloses its own implementation file and states plainly that no dataset or specific numerical result beyond the genetic algorithm's own output is claimed (SR = 4.60): code and data disclosure are present, but no hyperparameters are reported, since the LLM-based writer was explicitly instructed not to invent parameter values it was not given.
- Significance: Highest of the three conditions (SS = 3.54): its discussion of applications associates with four of the five reference domains.

(B)

- Originality: Markedly lower than the other two conditions (SO = 3.94; combinatorial Novelty = 0.57): its (Symbolic AI, Lack of Novelty, Robustness) combination partially overlaps with the knowledge-space reference paper "Constitutional AI," whose own recovered Method is likewise "Symbolic AI"—a concrete instance of the combination this manuscript centers on already being represented, in part, in the reference literature.
- Quality: Second-highest of the three conditions (SQ = 4.29), close to A)'s (circular for the same reason as for A).
- Reproducibility: Ties with (C) for the highest Reproducibility of the three conditions (SR = 5.00): the manuscript names its own implementation basis and reports concrete quantitative figures alongside disclosed hyperparameters
- Significance: Second-highest of the three conditions (SS = 2.76): its discussion of applications associates with three of the five reference domains.

(C)

- Originality: Ties with (A) for the highest of the three conditions (SO = 4.98; combinatorial Novelty = 1.00): its (Multi-Agent System, Lack of Novelty, Efficiency) combination, like (A)'s, shares no vocabulary with any knowledge-space reference paper's recovered combination. Because (C) was produced with no genetic-algorithm involvement at all, this indicates that the present instrument, with only three reference papers in the knowledge space, cannot on its own distinguish a combinatorially searched hypothesis from a freely generated one that happens to fall outside the reference papers' vocabulary; see the Discussion for this limitation.
- Quality: Lowest of the three conditions (SQ = 3.08): the manuscript's central claims (an entropy-anchored generative process reward model) fall furthest from the named success-pattern rules of the earlier GA (Section 4.8) that this criterion is calibrated to recognize.
- Reproducibility: Ties with (B) for the highest Reproducibility of the three conditions (SR = 5.00): code, data, and parameter disclosure are all present.
- Significance: Lowest of the three conditions (SS = 1.00): its discussion does not associate with any of the five reference domains (Generality = 0.00), which keeps its overall score the lowest of the three despite Originality and Reproducibility scores comparable to, or matching, the other two conditions.

This pilot evaluated a single generated manuscript per condition, although both the GA search and LLM generation are stochastic; it therefore provides no basis for statistical comparison or causal attribution.

Descriptively, A obtained the highest overall instrument score (4.40), followed by B (4.00) and C (3.52), with the differences arising mainly from SQ and SS, the two circular criteria, while A and C both reached the ceiling of Originality. To examine this ceiling, the semantic novelty of each pilot manuscript (mean-pooled sentence embedding of the full text) was recomputed against the three reference papers and against the full search partition K_s: against the three papers, the values were 0.83 (A), 0.67 (B), and 0.91 (C); against K_s, they fell to 0.42, 0.30, and 0.74, respectively. The ceiling thus disappears with a realistic knowledge space, and A no longer ranks with C, in line with Experiment 3; with one manuscript per condition, no further conclusion is drawn.

5. Discussion

This study proposed a creative AI model that applies a genetic algorithm to combinations of Method, Problem, and Goal drawn from a knowledge space of existing papers, re-implemented it with a shared semantic representation and a literature-grounded utility term, and evaluated it with outcome measures computed on papers not used in the search, a controlled manuscript-generation experiment, a blinded human evaluation, and a validation against expert peer review. Four findings emerge. First, the novelty–utility objective generalizes out of a sample: selected hypotheses are farther from held-out literature than unsearched hypotheses while their pairwise links remain supported by the literature. Second, at the present scale, the genetic algorithm reaches the same fitness as simpler search, and it returns a more diverse set of hypotheses across runs than random search. Third, hypotheses proposed freely by an LLM are closer to existing and later literature and concentrate on a few mainstream themes, whereas selection from the combinatorial concept space yields more distant and more widely spread hypotheses, and the blinded human evaluation gave no indication that this wider exploration came at the expense of usefulness. Fourth, distance from the literature is not originality as experts judge it: the novelty measure is practically unrelated to reviewers' contribution scores and acceptance decisions.

The second finding qualifies the role of the genetic algorithm. A GA is useful when the search space is too large for exhaustive or random coverage and when good solutions share building blocks that crossover can recombine. A space of 2700 triples that can be evaluated exhaustively in seconds satisfies neither condition; at this scale, the components of the model that matter are the representation and the objective, not the search operator. The present GA configuration (a population of ten, two elites, and a top-five parent pool) also converges early, as Figure 1 shows. Whether the GA becomes advantageous for larger concept vocabularies, hypotheses with more slots, fitness functions that are expensive to evaluate (for example, when each evaluation requires an LLM call), or tuned population sizes and mutation rates remains open and can be tested directly with the released code. The claim of the earlier version that the GA improves the Quality and Significance of generated manuscripts is withdrawn, because it rested on a circular instrument and an uncontrolled comparison of single manuscripts (Sections 4.8–4.10).

The third and fourth findings together delimit what can be claimed. The contrast with the LLM supports the motivation of this study in a specific sense: an LLM asked for a novel hypothesis gravitates toward the themes that dominate the current literature—for one of the two models, half of the proposals were variants of neuro-symbolic learning, and the other model, given the GA's vocabulary, chose the same mainstream combination in all 30 runs—and its proposals lie where the field was already heading, whereas combinatorial selection explores a much wider range of combinations. Whether this wider exploration produces better or more original hypotheses is a different question. The validation in Section 4.7 shows that literature distance does not capture originality as judged by expert reviewers, the GA's distant hypotheses were not approached by later literature, and the two automated novelty measures disagree precisely for the LLM's specifically worded proposals. The distance results are therefore reported as evidence of diversity and of departure from the mainstream, not of originality, which the automated measures used here cannot establish; this also confirms, with data, the concern that automated novelty scores require external validation [8]. The evaluation by LLM judges points the same way. The human raters tended to rank the LLM-proposed hypotheses as least useful, whereas the order-consistent verdicts of a 70-billion-parameter LLM judge more often favored the manuscripts built on them, and the judge ranked the hypotheses themselves as the most novel, even though the judge and the generator were different models and the wording was produced by the same writer for all sources. One could argue instead that the judge

preferred the LLM-proposed hypotheses because they are better. Three observations argue against this interpretation: the judge was unreliable as an instrument, since its overall verdicts agreed across the two presentation orders in only 30% of the pairs; the human raters did not show the judge's preference (the LLM-proposed hypotheses received the lowest mean usefulness rank of the three sources); and a preference of LLM judges for LLM-generated content has been documented independently [35,36]. The data cannot rule this alternative out completely, but they do not support it. LLM-based evaluation may therefore favor the mainstream ideas that LLMs themselves produce, which would make it unsuitable for assessing methods designed to depart from the mainstream and reinforces the conclusion that automated measures require validation against expert judgment.

Several limitations remain. The concept vocabulary was curated by the author; the knowledge space is limited to four arXiv computer-science categories and a single sentence encoder; the LLM-proposed hypotheses and all manuscripts of Experiment 4 come from a single 9-billion-parameter model, and the robustness check covers only two small open-weight models (8–9 billion parameters), so larger or proprietary models may behave differently; the LLM and the GA differ in budget, and a selection from a pool of LLM proposals under the same objective was not tested; the LLM judges could not discriminate between manuscripts on an absolute scale and showed strong position bias in pairwise comparisons; with two human raters and 30 sets, the human evaluation has limited power to detect moderate differences between sources; the human raters, although working in AI-related fields, not specialists in the AI literature, so their novelty rankings are supplementary; the uniform rendering of hypotheses cannot remove every cue to their source (the LLM-proposed descriptions were slightly longer); and the GA parameters were not tuned.

6. Conclusions

This study proposed a creative AI model that combines a genetic algorithm with a knowledge space of existing papers to generate research-theme hypotheses as Method–Problem–Goal combinations, and evaluated it with a shared embedding space for papers and hypotheses and an evaluation protocol that separates search and evaluation data and adds external checks: 30 independent runs, budget-matched baselines and ablations, outcome measures computed on held-out and later-published literature, a controlled comparison with hypotheses proposed by an LLM, a blinded human evaluation, and a validation against expert peer review. Hypotheses selected with the novelty–utility objective were more distant from unseen literature and more diverse than those proposed by the LLM, whose proposals concentrated on a few mainstream themes. The genetic algorithm, however, did not outperform random search or hill climbing at the present scale, and literature distance proved to be unrelated to expert judgments of contribution, so the originality of the selected hypotheses could not be established. Moreover, LLM judges proved unreliable, and their consistent verdicts leaned toward the LLM-proposed hypotheses that human raters tended to rank lowest.

The contribution of this study is therefore a transparent and reproducible framework for generating and evaluating combinatorial research hypotheses, together with three pieces of evidence: that the earlier claim of GA-driven improvements in manuscript quality cannot be sustained, that LLM-proposed hypotheses are less diverse and closer to the mainstream than combinatorially selected ones, and that literature-distance novelty measures do not reflect expert judgments of originality. Future work includes evaluation by raters with specialist AI expertise, larger and more diverse knowledge spaces and concept vocabularies in which the search operator can matter, novelty measures validated against expert judgment, and tuning of the GA parameters.

Funding: This research was funded by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan through the Supporting Pioneering Research through AI for 1000 Discovery Challenges Program (SPReAD), grant number JPMXP1726256770.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The Python implementation of the proposed model, the knowledge-space data (arXiv metadata, CC0), the prompts, the three pilot manuscripts, the random seeds, the per-run logs, the LLM-proposed hypotheses, the 30 manuscripts of Experiment 4 with all raw model responses and LLM-judge outputs, the blinded rating materials and anonymized ratings of Experiment 5, the scripts of the peer-review validation, the answers and analysis of the robustness experiments (Appendix B), and the scripts that regenerate all tables and figures are openly available at https://github.com/otsukiLabo-NUCOE/creative-ai-hypothesis-generation (accessed on 2 October 2026) and archived at Zenodo (https://doi.org/10.5281/zenodo.23290094, accessed on 2 October 2026).

Acknowledgments: The author would like to thank Anna Kuwana (Ochanomizu University) and Kenji Yamaguchi (Nihon University) for their valuable comments on an earlier version of this manuscript. An AI assistant (Claude Opus 5.5, Anthropic) was used to assist in drafting and revising portions of the manuscript text; see Section 3.6 for the scope of its role and the safeguards applied. The author has reviewed and edited the output and takes full responsibility for the content of this publication.

Conflicts of Interest: The author declares no conflicts of interest.

Appendix A

This appendix summarizes the rating instructions; the complete materials (rater manual, booklets, answer sheets, prompts, and scripts) are released with the code.

Human raters (Experiment 5). For each set of three hypotheses, raters chose the most and the least useful hypothesis and, separately, the most and the least novel one. Usefulness was defined as whether the hypothesis is worth pursuing as a research topic, judged from (i) the importance of the problem, (ii) whether the method plausibly helps to solve the problem or reach the goal, and (iii) feasibility. Novelty was defined as how new and non-obvious the idea is compared with research known to the rater, without regard to feasibility. Raters were instructed not to reward wording, length, or technical jargon, that positions and set numbers are random, and that ties are not allowed. They worked independently, without generative AI tools, and were not told how the hypotheses had been produced.

LLM judge (Experiment 4). Each manuscript was scored from 1 to 5 for originality, technical soundness, reproducibility, and significance with verbal anchors for every score level, with the instruction not to penalize the absence of experimental results or citations (common to all manuscripts), not to reward fluency or length, and not to reward numbers presented as results.

Appendix B

Table A1 reports the robustness check of the LLM-proposed hypotheses (Section 4.3). Prompt P0 is the original prompt (a hypothesis in AI, "in particular generative AI"); P1 omits the generative-AI restriction; in the scope-matched condition the model chooses one item from each list of the GA's concept vocabulary. All configurations use top-p 0.95 and seeds 1–30; the prompts, raw answers, and analysis scripts are in the released repository.

Table A1. Robustness of the LLM-proposed hypotheses across models, prompts, and temperatures (30 hypotheses per row). Similarity: mean pairwise cosine similarity (lower = more diverse). Neuro-symbolic: share of methods described as neuro-symbolic. GA and uniform rows from Experiment 2.

Source	Prompt	T	Distinct	Similarity	N_{out}	N_{out} (TF–IDF)	FS	Neuro-Symbolic
GA (Experiment 2)	—	—	19	0.290	0.515	0.786	0.438	0%
Uniform (Experiment 2)	—	—	30	0.284	0.461	0.769	0.561	0%
Nemotron-Nano-9B-v2	P_0 (generative AI)	0.7	30	0.513	0.435	0.811	0.580	50%
Nemotron-Nano-9B-v2	P_0 (generative AI)	1.0	30	0.487	0.430	0.823	0.600	27%
Nemotron-Nano-9B-v2	P_1 (AI and ML)	0.7	30	0.488	0.424	0.789	0.580	50%
Nemotron-Nano-9B-v2	P_1 (AI and ML)	1.0	30	0.446	0.410	0.788	0.583	53%
Llama-3.1-8B-Instruct	P_0 (generative AI)	0.7	30	0.487	0.416	0.800	0.584	0%
Llama-3.1-8B-Instruct	P_0 (generative AI)	1.0	30	0.502	0.410	0.807	0.567	0%
Llama-3.1-8B-Instruct	P_1 (AI and ML)	0.7	30	0.502	0.373	0.773	0.609	0%
Llama-3.1-8B-Instruct	P_1 (AI and ML)	1.0	30	0.518	0.375	0.785	0.610	0%
Nemotron-Nano-9B-v2	Scope-matched	0.7	1	1.000	0.430	0.659	0.557	0%
Llama-3.1-8B-Instruct	Scope-matched	0.7	22	0.335	0.380	0.749	0.605	0%

References

1. Braincuber. 10 Industries Where AI Is Generating Revenue Right Now. Available online: https://www.braincuber.com/blog/10-industries-where-ai-is-generating-revenue-right-now (accessed on 10 March 2026).
2. Primotech. AI in Real Estate 2026: Growth Strategies for Modern Property Businesses. Available online: https://primotech.com/ai-in-real-estate-2026-growth-strategies-for-modern-property-businesses/ (accessed on 10 March 2026).
3. How AI Agents Automated Most of Our dbt Development. Available online: https://zenn.dev/ubie_dev/articles/d97c5ece4660bd (accessed on 10 March 2026).
4. OpenAI Announces "Codex Security," an AI Agent that Automates Vulnerability Discovery, Verification, and Remediation. Available online: https://article.auone.jp/detail/1/3/7/48_7_r_20260309_1773027373015859 (accessed on 10 March 2026).
5. Futran Solutions. Choosing the Right AI Use Cases: Linking AI to Decisions That Matter. Available online: https://futransolutions.com/blog/choosing-the-right-ai-use-cases-linking-ai-to-decisions-that-matter-2026/ (accessed on 10 March 2026).
6. Lu, C.; Lu, C.; Lange, R.T.; Foerster, J.; Clune, J.; Ha, D. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. *arXiv* **2024**, arXiv:2408.06292.
7. Yamada, Y.; Lange, R.T.; Lu, C.; Hu, S.; Lu, C.; Foerster, J.; Clune, J.; Ha, D. The AI Scientist-v2: Workshop-Level Automated Scientific Discovery via Agentic Tree Search. *arXiv* **2025**, arXiv:2504.08066.
8. Beel, J.; Kan, M.-Y.; Baumgart, M. Evaluating Sakana's AI Scientist for Autonomous Research: Wishful Thinking or an Emerging Reality Towards 'Artificial Research Intelligence' (ARI)? *arXiv* **2025**, arXiv:2502.14297.
9. Wang, Q.; Downey, D.; Ji, H.; Hope, T. SciMON: Scientific Inspiration Machines Optimized for Novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, Bangkok, Thailand, 11–16 August 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 279–299. [CrossRef]
10. Yang, Z.; Liu, W.; Gao, B.; Xie, T.; Li, Y.; Ouyang, W.; Poria, S.; Cambria, E.; Zhou, D. MOOSE-Chem: Large Language Models for Rediscovering Unseen Chemistry Scientific Hypotheses. *arXiv* **2024**, arXiv:2410.07076.
11. Radensky, M.; Shahid, S.; Fok, R.; Siangliulue, P.; Hope, T.; Weld, D.S. Scideator: Human-LLM Scientific Idea Generation and Novelty Evaluation Grounded in Research-Paper Facet Recombination. *arXiv* **2024**, arXiv:2409.14634.
12. Bran, A.M.; Cox, S.; Schilter, O.; Baldassari, C.; White, A.D.; Schwaller, P. ChemCrow: Augmenting Large-Language Models with Chemistry Tools. *arXiv* **2023**, arXiv:2304.05376.
13. Thesis AI. Available online: https://www.thesisai.io/ (accessed on 10 March 2026).
14. Thesis Generator: Essay AI. Available online: https://apps.apple.com/jp/app/thesis-generator-essay-ai/id6739264844 (accessed on 10 March 2026).
15. Antigravity. Available online: https://antigravity.google/ (accessed on 10 March 2026).
16. Swanson, D.R. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect. Biol. Med.* **1986**, *30*, 7–18. [CrossRef]
17. Tshitoyan, V.; Dagdelen, J.; Weston, L.; Dunn, A.; Rong, Z.; Kononova, O.; Persson, K.A.; Ceder, G.; Jain, A. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **2019**, *571*, 95–98. [CrossRef]
18. Krenn, M.; Buffoni, L.; Coutinho, B.; Eppel, S.; Foster, J.G.; Gritsevskiy, A.; Lee, H.; Lu, Y.; Moutinho, J.P.; Sanjabi, N.; et al. Forecasting the future of artificial intelligence with machine learning-based link prediction in an exponentially growing knowledge network. *Nat. Mach. Intell.* **2023**, *5*, 1326–1335. [CrossRef]
19. Si, C.; Yang, D.; Hashimoto, T. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers. *arXiv* **2024**, arXiv:2409.04109.

20. Baek, J.; Jauhar, S.K.; Cucerzan, S.; Hwang, S.J. ResearchAgent: Iterative Research Idea Generation over Scientific Literature with Large Language Models. *arXiv* **2024**, arXiv:2404.07738.
21. Lehman, J.; Stanley, K.O. Abandoning objectives: Evolution through the search for novelty alone. *Evol. Comput.* **2011**, *19*, 189–223. [CrossRef]
22. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. *Improving Language Understanding by Generative Pre-Training*; Technical Report; OpenAI: San Francisco, CA, USA, 2018.
23. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 3982–3992.
24. Wang, W.; Wei, F.; Dong, L.; Bao, H.; Yang, N.; Zhou, M. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 5776–5788.
25. arXiv API User's Manual. Available online: https://info.arxiv.org/help/api/user-manual.html (accessed on 1 October 2026).
26. Holland, J.H. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*; University of Michigan Press: Ann Arbor, MI, USA, 1975.
27. De Jong, K.A. An Analysis of the Behavior of a Class of Genetic Adaptive Systems. Ph.D. Thesis, University of Michigan, Ann Arbor, MI, USA, 1975.
28. Input and Output—Python 3 Documentation. Available online: https://docs.python.org/3/tutorial/inputoutput.html (accessed on 10 March 2026).
29. Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. The Llama 3 Herd of Models. *arXiv* **2024**, arXiv:2407.21783.
30. Mann, H.B.; Whitney, D.R. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* **1947**, *18*, 50–60. [CrossRef]
31. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
32. Vargha, A.; Delaney, H.D. A critique and improvement of the CL common language effect size statistics of McGraw and Wong. *J. Educ. Behav. Stat.* **2000**, *25*, 101–132. [CrossRef]
33. NVIDIA. NVIDIA Nemotron Nano 2: An Accurate and Efficient Hybrid Mamba-Transformer Reasoning Model. *arXiv* **2025**, arXiv:2508.14444.
34. llama.cpp: LLM Inference in C/C++. Available online: https://github.com/ggml-org/llama.cpp (accessed on 1 October 2026).
35. Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.P.; et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track; Proceedings of the 37th International Conference on Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023*; Curran Associates: Red Hook, NY, USA, 2023; pp. 46595–46623.
36. Laurito, W.; Davis, B.; Grietzer, P.; Gavenčiak, T.; Böhm, A.; Kulveit, J. AI–AI bias: Large language models favor communications generated by large language models. *Proc. Natl. Acad. Sci. USA* **2025**, *122*, e2415697122. [CrossRef]
37. Krippendorff, K. *Content Analysis: An Introduction to Its Methodology*, 4th ed.; SAGE: Thousand Oaks, CA, USA, 2018.
38. ReviewArena: Peer Reviews, Rebuttals, and Decisions from OpenReview. Hugging Face Dataset. Available online: https://huggingface.co/datasets/Samarth0710/reviewarena (accessed on 2 October 2026).
39. NeurIPS 2025 Call for Papers. Available online: https://neurips.cc/Conferences/2025/CallForPapers (accessed on 10 March 2026).
40. ICLR 2026 Author Guide. Available online: https://iclr.cc/Conferences/2026/AuthorGuide (accessed on 10 March 2026).
41. ICML 2026 Reviewer Instructions. Available online: https://icml.cc/Conferences/2026/ReviewerInstructions (accessed on 10 March 2026).
42. CVPR 2026 Reviewer Guidelines. Available online: https://cvpr.thecvf.com/Conferences/2026/ReviewerGuidelines (accessed on 10 March 2026).
43. AAAI-26 Review Process. Available online: https://aaai.org/conference/aaai/aaai-26/review-process/ (accessed on 10 March 2026).
44. Ratcliff, J.W.; Metzener, D.E. Pattern matching: The Gestalt approach. *Dr. Dobb's J.* **1988**, *13*, 46.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.