DEEPSCIENTIST: ADVANCING FRONTIER-PUSHING SCIENTIFIC FINDINGS PROGRESSIVELY

Yixuan Weng*, Minjun Zhu*, Qiujie Xie, Qiyao Sun, Zhen Lin, Sifan Liu, Yue Zhang✉
Engineering School, Westlake University
wengsyx@gmail.com; {zhu.minjun,zhangyue}@westlake.edu.cn

Project: https://ai-researcher.net
Code: https://github.com/ResearAI/DeepScientist

ABSTRACT

While previous AI Scientist systems can generate novel findings, they often lack the focus to produce scientifically valuable contributions that address pressing human-defined challenges. We introduce DeepScientist, a system designed to overcome this by conducting goal-oriented, fully autonomous scientific discovery over month-long timelines. It formalizes discovery as a Bayesian Optimization problem, operationalized through a hierarchical evaluation process consisting of "hypothesize, verify, and analyze". Leveraging a cumulative Findings Memory, this loop intelligently balances the exploration of novel hypotheses with exploitation, selectively promoting the most promising findings to higher-fidelity levels of validation. Consuming over 20,000 GPU hours, the system generated about 5,000 unique scientific ideas and experimentally validated approximately 1100 of them, ultimately surpassing human-designed state-of-the-art (SOTA) methods on three frontier AI tasks by 183.7%, 1.9%, and 7.9%. This work provides the first large-scale evidence of an AI achieving discoveries that progressively surpass human SOTA on scientific tasks, producing valuable findings that genuinely push the frontier of scientific discovery.

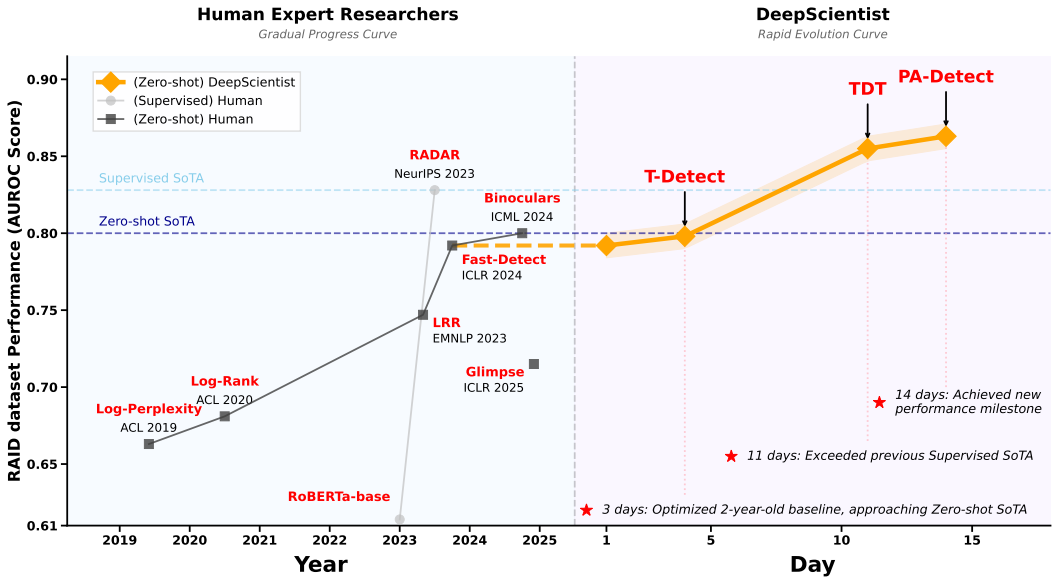

Figure 1: Comparison of research progress timelines for AI text detection on the RAID (Dugan et al., 2024). The right panel shows that DeepScientist achieves progress in two weeks that is comparable to three years of human research (Su et al.; Bao et al., a;b; Hu et al., 2023) (left panel). All zero-shot methods, including the system-generated T-Detect, TDT, and PA-Detect, uniformly adopt Falcon-7B (Almazrouei et al., 2023) as the base model. Additionally, all methods produced by DeepScientist demonstrate higher throughput than the previous SOTA method, Binoculars (Hans et al., 2024).

1 INTRODUCTION

Scientific discovery is inherently a process of **continuous exploration** and **trial-and-error**, where vast amounts of time and effort are invested to push the boundaries of human knowledge forward by a small step. This principle of persistent, incremental advancement is visible across the history of technology. For example, the decades-long optimization of semiconductor manufacturing has seen the feature size of transistors systematically reduced from micrometers to single-digit nanometers (Moore, 1965). Similarly, the efficiency of photovoltaic cells has been continuously advanced over half a century, with myriad material and architectural innovations pushing conversion rates from nascent single-digit percentages ever closer to their theoretical limits (Green, 1993). These historical trajectories underscore a process where human scientists engage in decades of **goal-directed**, **iterative** work to advance the SoTA artifacts continuously.

Recently, the emergence of Large Language Models (LLMs) has propelled **automated scientific discovery**, where LLM-based AI Scientist systems take the lead in exploration (Xie et al., 2025b). With their powerful capacity for long-form generation and comprehension, LLMs enable **end-to-end, full-cycle automation in scientific discovery.** This has inspired influential work such as AI SCIENTIST-V2 (Yamada et al., 2025), whose scientific artifacts have been published in top-tier conference workshops. However, in the absence of clearly defined scientific goals, current AI Scientist systems often fall into the trap of blindly recombining existing knowledge and methods. As a result, their research outputs frequently appear naive under human evaluation and lack genuine scientific value (Zhu et al., 2025c). **AI Scientists are yet to solve human challenges.**

To solve real-world challenges, We formally model the full cycle of scientific discovery as a **goal-driven Bayesian Optimization problem**, where the singular objective is to find a novel method that maximally improves a target performance metric. Building on this formulation, we introduce **DeepScientist**, a LLM-based agent system designed to explore progressively across the unknown space of possible candidate research methods to **identify the optimal plan** that maximizes a highly expensive-to-evaluate function of true scientific value. Specifically, DeepScientist employs an **iterative workflow**, together with a continuously expanding memory of prior research knowledge to efficiently manage uncertainty during exploration. It intelligently balances **exploitation** (deepening investigations into promising high-value directions) with **exploration** (venturing into uncharted areas to acquire new knowledge). Through large-scale parallel exploration, DeepScientist can generate innovative hypotheses and ultimately yield both valuable new methods and validation-proven scientific findings through continuous exploration.

We select three frontier scientific tasks (*Agent Failure Attribution*, *LLM Inference Acceleration*, and *AI Text Detection*), take their state-of-the-art methods (ICML 2025 Spotlight, ACL 2025 Outstanding, ICLR 2024) as starting points, and ask DeepScientist to conduct continuous research. As shown in Figures 1 and 3, within a month-long cycle of exploration, validation, and iteration on 16 H800 GPUs, **DeepScientist exceeds their respective human SOTA methods by 183.7% (Accuracy), 1.9% (Tokens/second), and 7.9% (AUROC) through autonomously redesigning core methodologies, rather than simply combining existing techniques** (Section 4.1). To understand how such progress emerged, we analyze DeepScientist's discovery logs, and formed a small program committee to review the generated papers (Section 4.2). These logs show that the system generated over 5,000 unique ideas, of which only 1,100 are selected for experimental validation, and just 21 ultimately lead to scientific innovations (Section 4.3). Moreover, through the scaling experiment on computational resource, we discover a near-linear relationship between the resources allocated and the output of valuable scientific discoveries.

To our knowledge, we provide **the first empirical demonstration of an automated full-cycle scientific discovery system capable of producing novel, SOTA-surpassing methods and continuously advancing scientific frontiers** at a pace that substantially exceeds human researchers. Our findings reveal a stark reality: while the AI's exploratory speed is immense, its inherent success rate for innovation remains exceptionally low, making effective validation and filtering the new bottleneck at the frontier of automated science. Therefore, the central question of the field is no longer 'Can AI innovate?', but rather 'How can we efficiently guide its powerful, yet highly dissipative, exploratory process to maximize scientific return?' We hope this work can inspire the research community to develop AI Scientist systems with greater exploration efficiency to accelerate scientific discovery at a larger scale, paving the way for ground-breaking discoveries.

2 Related Work

Replication and Optimization. A significant body of research focuses on engineering tasks that operate within established scientific frameworks. This includes replication-oriented works like PaperBench (Starace et al., 2025) and Paper2Agent (Miao et al., 2025), which aim to reproduce existing papers. Other works, such as Agent Laboratory (Schmidgall et al., 2025b) and MLE-Bench (Chan et al., 2024), tackle early-stage machine learning engineering problems. Similarly, systems like AlphaTensor (Fawzi et al., 2022) and AlphaEvolve (Novikov et al., 2025) use massive trial-and-error with known engineering methods to improve the performance of codebases. The common goal of these efforts is engineering-driven optimization within an established scientific paradigm, enhancing existing systems without questioning their foundational assumptions. DeepScientist, in contrast, pursues scientific discovery by targeting the core limitations of the SOTA itself. Its objective is not to refine the current state-of-the-art, but to establish a new one by introducing fundamentally different methodologies.

Semi-Automated Scientific Assistance. The path toward automating scientific discovery begin not with replacing the scientist, but with assisting them, leading to the development of a paradigm of specialized AI tools for individual research tasks. Systems like CycleResearcher (Weng et al., 2025) handle writing, DeepReview (Zhu et al., 2025a) manages reviewing, and co-scientists (Gottweis et al., 2025; Penadés et al., 2025; Swanson et al., 2025; Baek et al., 2025) aid in hypothesis generation. These powerful tools address only isolated fragments of the scientific process, leaving the crucial loop of learning from failure and exploration to humans. In contrast, DeepScientist is an autonomous agent of inquiry, managing the entire end-to-end research cycle and closing the loop by learning from its own experiments and self-directing its research path.

Automated Scientific Discovery. Building on the capabilities of specialized assistants, a line of research pursue full, end-to-end research automation (Yang et al., 2023; Xie et al., 2025a). Pioneering efforts, such as the AI Scientist systems (Lu et al., 2024; Yamada et al., 2025) and subsequent work (Intology, 2025; Jiabin et al., 2025), successfully demonstrate that an AI system could manage the full research cycle and produce novel findings. However, their primary limitation often lies in their exploratory strategy, which lacks a specific scientific goal rooted in a field's grand challenges, resulting undirected discoveries that may be perceived as lacking genuine scientific value. Instead, DeepScientist is thus the first automated scientific discovery system that leverages a closed-loop, iterative process to discover methods surpassing the human state-of-the-art. The exploration of DeepScientist is goal-oriented and insight-driven, beginning by identifying a recognized limitation in the human SOTA and then using failure attribution to ensure discoveries are both novel and scientifically meaningful.

3 DeepScientist: A Progressive System for Discovering SOTA-Surpassing Findings

3.1 Modeling Scientific Discovery as an Optimization Problem

The fundamental goal of automated scientific discovery is to autonomously identify novel methods that yield significant advancements in a given scientific domain. This process can be formally conceptualized as a search for an optimal solution within a vast and unstructured space of possibilities. Let the space of all possible candidate research methods be denoted by $\mathcal{I}$. Each individual method $I \in \mathcal{I}$, such as a novel algorithm or a new model architecture, possesses an intrinsic scientific value. This value is determined by a latent, black-box true value function, $f : \mathcal{I} \to \mathbb{R}$, which maps a method to its ultimate empirical impact. **The objective of scientific discovery is therefore to find the optimal method I^* that maximizes this function:**

$$I^* = \arg\max_{I \in \mathcal{I}} f(I) \tag{1}$$

Unlike previously studied tasks such as early-stage machine learning (Schmidgall et al., 2025a), algorithmic design (Novikov et al., 2025; Lange et al., 2025), or scientific software development (Aygün et al., 2025), **A defining characteristic of frontier scientific discovery is that each exploratory step demands immense computational and intellectual resources, making the evaluation of the true scientific value function, $f(\cdot)$, prohibitively costly.** Any single evaluation, $f(I)$,

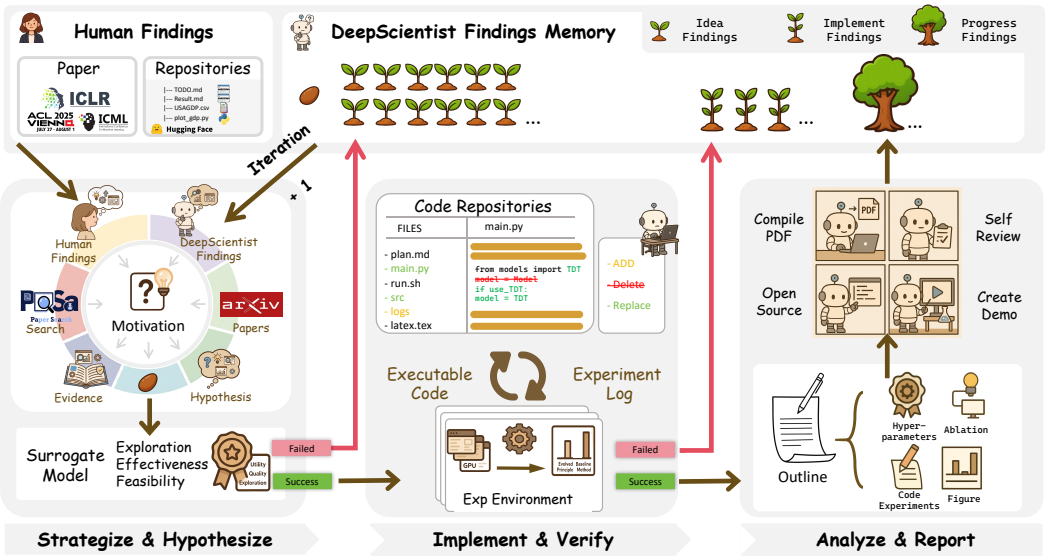

Figure 2: The autonomous, closed-loop discovery process of DeepScientist. The system iterates through a three-stage cycle, learning from both human knowledge and its own experiments.

corresponds to a complete and resource-intensive research cycle of implementation, experimentation, and analysis, often consuming vast computational resources (e.g., on the order of 10^{16} FLOPs for a frontier LLM problem, as illustrated in Figure 4.c). This extreme sample inefficiency renders brute-force or random exploration of the space $\mathcal{I}$ intractable. Therefore, we model the problem within the framework of Bayesian Optimization(Frazier, 2018; Garnett, 2023), which provides a principled methodology for global optimization of expensive black-box functions. By constructing a surrogate model to intelligently guide the search, Bayesian Optimization effectively reduces the number of costly real-world evaluations through a careful balance of exploration and exploitation. However, **for scientific discovery, $\mathcal{I}$ is a conceptual space that is not explicitly defined.** Candidate methods I must be formulated as creative, plausible, and coherent scientific hypotheses. The generation of high-quality candidate hypotheses is a critical bottleneck that traditional Bayesian Optimization algorithms are not designed to address. This challenge necessitates a new mechanism that integrates creative ideation with sample-efficient optimization. We detail our solution to this problem in the following subsections.

3.2 The DeepScientist Framework

The architecture of DeepScientist actualizes the Bayesian Optimization loop through a multi-agent system equipped with an **open-knowledge system** and a continuously accumulating **Findings Memory**. This memory is composed of both frontier human knowledge (e.g., papers and codes) and the system's own historical findings, and it intelligently guides subsequent explorations. **The entire discovery process is structured as a hierarchical and iterative three-stage exploration cycle**. In this hierarchical scheme, only research ideas that exhibit promise are advanced to more expensive evaluations, while others are retained in the Findings Memory to inform subsequent explorations. This design ensures the computational resources are dynamically and precisely allocated to the most promising scientific trajectories, thereby **maximizing discovery efficiency under constrained budgets**. Specifically, each stage within the three-stage exploration cycle is associated with a distinct **fidelity–cost tradeoff** (Figure 2):

Strategize & Hypothesize. Each research cycle begins by analyzing the Findings Memory ($\mathcal{M}_t$), a list-style database containing thousands of structured records. Each record represents a unique scientific finding, which is categorized according to its stage of development. To overcome the LLM's context length constraints, we use a separate retrieval model (Wolters et al., 2024) when needed to select the Top-K Findings as input. The vast majority of records begin as Idea Findings—unverified hypotheses. During this first stage, the system identifies limitations in existing knowledge and gen-

Table 1: Overview of the three different human SOTA methods we selected.

Task	Method	Venue	Benchmark	Github Star
Agents Failure Attribution	All at Once	ICML 2025 Spotlight	Who&When	302
LLM Inference Accel.	TokenRecycling	ACL 2025 Outstanding	MBPP	323
AI Text Detection	FastDetectGPT	ICLR 2024	RAID	414

erates a new collection of hypotheses ($\mathcal{P}_{\text{new}}$), and then they evaluated by a low-cost Surrogate Model (g_t). The surrogate model (an LLM Reviewer) is first contextualized with the entire Findings Memory. It then approximates the true value function f and, for each candidate finding $I \in \mathcal{P}_{\text{new}}$, produces a structured valuation vector $V = \langle v_u, v_q, v_e \rangle$, quantifying its estimated utility, quality, and exploration value as integer scores on a scale of 0 to 100. Each new hypothesis and its valuation vector is then used to initialize a new record in the Findings Memory as an "Idea Finding".

Implement & Verify. This stage serves as the primary filter in the Findings Memory. To decide which of the numerous "Idea Findings" warrants the significant resource investment to be advanced in a real-world experiment, the system employs an Acquisition Function (α). Specifically, it uses the classic Upper Confidence Bound (UCB) algorithm to select the most promising record. The UCB formula maps the valuation vector V to balance the trade-off between exploiting promising avenues (represented by v_u and v_q) and exploring uncertain ones (represented by v_e):

$$I_{t+1} = \arg\max_{I \in \mathcal{P}_{\text{new}}} \Big(\underbrace{w_u v_u + w_q v_q}_{\text{Exploitation Score}} + \kappa \cdot \underbrace{v_e}_{\text{Exploration Score}} \Big), \tag{2}$$

where w_u and w_q are hyperparameters and κ controls the intensity of exploration. The highest-scoring finding I_{t+1} is selected for validation, and its record is promoted to the status of an Implement Finding. **A coding agent then performs a repository-level implementation to executed the experiment.** This agent operates within a sandboxed environment with full permissions, allowing it to read the complete code repository and access the internet for literature and code searches. Its objective is to implement the new hypothesis on top of the existing SOTA method's repositories. The agent typically begins by planning the task, then reads the code to understand its structure, and finally implements the changes to produce the experimental logs and results. The experiment logs and results, $f(I_{t+1})$, is used to update the corresponding record, enriching it with empirical evidence and thus closing the learning loop.

Analyze & Report. The final and most selective stage of the Findings Memory is triggered only by a successful validation. When an "Implement Finding" succeeds in surpassing the baseline, its record is promoted to a Progress Finding. This transformation is implemented by a series of specialized agents capable of utilizing a suite of MCP (Hou et al., 2025) tools. These agents first autonomously design and execute a series of deeper analytical experiments (e.g., ablations, evaluations on new datasets), leveraging MCP tools to manage the experimental lifecycle, data collection, and result parsing. Subsequently, a synthesis agent employs the same toolset to collate all experimental results, analytical insights, and generated artifacts into a coherent, reproducible research paper. This deeply validated record becomes a new record in the system's knowledge base, thus influencing the decision-making process in all subsequent cycles.

4 Experiments

As detailed in Table 1, we select three distinct SOTA methods (published in 2024 and 2025) as starting points, chosen for their frontier status, community interest, and human supervisability. Each SOTA method is manually reproduced, and we preserve execution logs and test scripts to allow DeepScientist to focus on research advancement. DeepScientist is provided with two servers, each with 8 Nvidia H800 GPUs. To maximize utilization, we launch a separate system instance for each GPU, employing the Gemini-2.5-Pro model for core logic and the Claude-4-Opus model for its robust code-generation capabilities. Three human experts supervise the process to verify outputs and filter out hallucinations. For more implementation details, please see Appendix C.

Method	Agent Failure Attribution		LLM Inference Acceleration	AI Text Detection	
	Handcraft (Acc.)	Algorithm-Gen (Acc.)	Tokens/second	AUROC	Latency
Human SoTA method	12.07% (All at Once)	16.67% (All at Once)	190.25 (Token Recycling)	0.800 (Binoculars)	117ms (Binoculars)
DeepScientist's method	**29.31%** (A2P)	**47.46%** (A2P)	**193.90** (ACRA)	**0.863** (PA-Detect)	**60ms** (PA-Detect)
Improvement	**Δ+142.8% (+17.24)**	**Δ+183.7% (+30.79)**	**Δ+1.9% (+3.65)**	**Δ+7.9% (+0.063)**	**Δ+190%↓ (-57)**

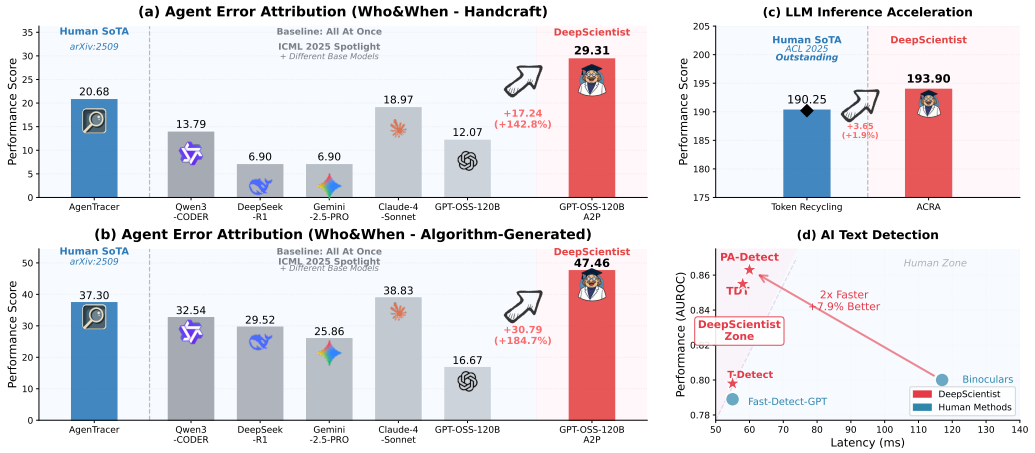

Figure 3: Performance evaluation of DeepScientist across three research domains: (a-b) Agent Failure Attribution on Who&When benchmark in handcraft and algorithm-generated settings; (c) LLM Inference Acceleration on MBPP dataset; (d) AI Text Detection with performance-latency tradeoff analysis. DeepScientist (shown in pink) consistently outperform human-designed SoTA approaches (shown in blue) across all tasks.

4.1 DeepScientist achievements on three research domains

Agents Failure Attribution. The task addresses the question: in an LLM-based multi-agent system, which agent caused the task to fail and when? Starting from the baseline "All at once" method (Zhang et al., 2025c), DeepScientist identified that the current approach lacks the counterfactual reasoning capabilities essential for attribution. Through a process of trial, error, and synthesizing new findings—discovering the effectiveness of hypothetical prediction and simulated attempts—it ultimately proposed the *A2P* method. Named for its Abduction-Action-Prediction process, its core innovation elevates failure attribution from pattern recognition to causal reasoning, filling the critical gap in counterfactual capabilities by predicting if a proposed fix would have led to success. As shown in Figure 3.(a-b), A2P achieved scores of 29.31 and 47.46 in the "handcraft" and "algorithm-generated" settings of the Who&When benchmark, respectively, setting a new state-of-the-art (SOTA). In this task, DeepScientist validated that a structured, zero-shot causal reasoning framework can be superior to less principled methods. As of September 2025, the training-free A2P method maintains its SOTA position, outperforming even 7B models trained on synthetic data. (Zhang et al., 2025a).

LLM Inference Acceleration is a highly optimized field aiming to maximize throughput and reduce latency during LLM inference (Xia et al., 2024). In this process, the system actively made many different attempts, such as using a Kalman Filter (Zarchan, 2005) to dynamically adjust an adjacency matrix to address the original method's lack of a memory function. Although most of these attempts failed, the system-generated *ACRA* method ultimately advanced the MPBB (Austin et al., 2021) from a human SOTA of 190.25 to 193.90 tokens/second by identifying stable suffix patterns, as shown in Figure 3. Scientifically, this innovation is significant because it uses this extra contextual information to dynamically adjust the decoding guess, effectively grafting a long-term memory onto the process and breaking the context-collapsing of standard decoders. This discovery highlights the system's primary goal: the creation of new, human-unknown knowledge rather than mere engineering optimization. For instance, one could likely achieve greater performance gains by combining ACRA with an established technique like layer skipping (Wang et al., 2022) or PageAttention (Kwon et al., 2023), but this would represent an engineering effort, not a scientific one. The exploration assessment within our process avoids such combinations of existing knowledge.

Table 2: Evaluation of AI-generated papers produced by various AI Scientist systems. Scores represent the average ratings given by DeepReviewer-14B (Zhu et al., 2025a) across the number ("Num") of available papers. Note: Publicly available papers may be curated and therefore may not fully represent the typical output of each system.

AI Scientist Systems	Number	Soundness	Presentation	Contribution	Rating	Accept Rate
AI SCIENTIST	10	2.08	1.80	1.75	3.35	0%
HKUSD AI Researcher	7	1.75	1.46	1.57	2.57	0%
AI SCIENTIST-V2	3	1.67	1.50	1.50	2.33	0%
CycleResearcher-12B	6	2.25	1.75	2.13	3.75	0%
Zochi	2	2.38	2.38	2.25	4.63	0%
DeepScientist (Ours)	5	**2.90**	**2.90**	**2.90**	**5.90**	**60%**

Table 3: Evaluation of DeepScientist's papers produced by human experts. Values are presented as mean (variance) from three reviewers. Inter-rater reliability for Rating: Krippendorff's $\alpha = 0.739$.

Paper	Confidence	Soundness	Presentation	Contribution	Rating
HUMAN Avg. (ICLR 2025)	-	2.59	2.36	2.62	5.08
1. T-DETECT	4.33 (0.33)	2.00 (1.00)	**2.67** (0.33)	**2.67** (0.33)	5.00 (0.00)
2. TDT	4.67 (0.33)	**3.00** (0.00)	**3.00** (0.00)	**3.00** (0.00)	**5.67** (0.33)
3. PA-DETECT	4.00 (0.00)	1.67 (0.33)	2.00 (1.00)	2.00 (1.00)	4.33 (1.33)
4. A2P	4.00 (0.00)	**3.00** (0.00)	**3.00** (0.00)	**2.67** (0.33)	**5.67** (0.33)
5. ACRA	3.33 (0.33)	1.67 (0.33)	2.00 (1.00)	1.67 (0.33)	4.33 (1.33)
DeepScientist Avg.	4.07	2.27	**2.53**	2.40	5.00

AI Text Detection is a binary classification task where, given a text that may contain content from an LLM (and possibly additional noise), the goal is to determine if it was produced by a human or an LLM (Li et al., 2022; Ghosal et al., 2023). To validate its capacity for sustained advancement, DeepScientist made numerous attempts that included addressing the Boundary-Aware Extension problem and exploring approaches like Volatility-Aware and Wavelet Subspace Energy methods. The final results show a dramatic acceleration in scientific discovery: in a rapid evolution over just two weeks, the system produced three distinct, progressively superior methods (*T-Detect*, *TDT*, and *PA-Detect*). This began with T-Detect fixing core statistics with a robust t-distribution, then evolved conceptually with TDT and PA-Detect, which treat text as a signal and use wavelet and phase congruency analysis to pinpoint anomalies. Scientifically, this shift reveals the "non-stationarity" of AI-generated text, alleviating the information bottleneck in prior paradigms that average away localized evidence. As shown in Figure 1 and 3(d), this entire discovery trajectory demonstrates DeepScientist's ability for advancing frontier-pushing scientific findings progressively, establishing a new SOTA with a 7.9% higher AUROC while also doubling the inference speed.

4.2 ASSESSING THE QUALITY OF AI-GENERATED RESEARCH PAPER

Experimental Setup. To assess the quality of the final output, we evaluate the five research papers autonomously generated by DeepScientist's end-to-end process. Our evaluation protocol is twofold. First, to benchmark against existing work, we employ DeepReviewer (Zhu et al., 2025a), an AI agent that simulates the human peer-review process with an external search capability, comparing DeepScientist's output against 28 publicly available papers from other AI Scientist systems. Second, for a more rigorous assessment, we convene a dedicated program committee consisting of three active LLM researchers: two volunteers who have served as ICLR reviewers and one senior volunteer who has been invited to be an ICLR Area Chair. The generated papers are available in Appendix C.

Automated Review Against Other AI Scientist Systems. As shown in Table 2, the results from the LLM-based automatic evaluation indicate that the system's outputs are recognized for their scientific novelty and value. When benchmarked against 28 publicly available papers from other AI Scientist systems using DeepReviewer, DeepScientist is the only AI Scientist system to produce papers that achieves a 60% acceptance rate.

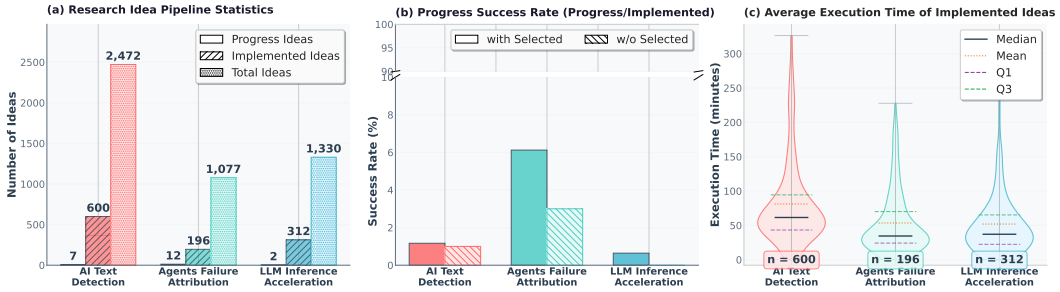

Figure 4: DeepScientist's experimental statistics. (a) The research pipeline from generated ideas to validated progress. (b) Success rates comparing our selection strategy against a baseline. (c) Distribution of wall-clock execution times for all implemented trials.

Human Expert Evaluation. The evaluation from our human program committee, shown in Table 3, reveal a remarkable and unanimous consensus: DeepScientist consistently excels at ideation, the most challenging and often rate-limiting step in human-led research. Full details on the review protocol are provided in Appendix A, and the core ideas within each paper are praised for their genuine novelty, ingenuity, and scientific contributions. The quality of these innovations is further demonstrated by the review scores: the system's average rating (5.00) closely mirrors the average of all ICLR 2025 submissions (5.08), with two of its papers significantly exceeding this (5.67).

4.3 Analysis of the Iterative Trajectory of Autonomous Exploration

Experimental Setup. The findings in this section are derived from a series of post-hoc analyses conducted on the complete operational data generated by DeepScientist across the three frontier tasks. This data includes the full set of execution logs and the Findings Memory, providing the basis for all subsequent statistical analysis. To visualize the conceptual search space (Figure 5), we embed the complete description of each generated finding using the Qwen3-Embedding-8B model. To assess scalability (Figure 6), we conduct a dedicated one-week experiment where N identified limitations of a single SOTA method are assigned to N parallel GPU instances. These instances explore solutions independently but share their findings to a central database, which are synchronized globally every five cycles to accommodate the asynchronous nature of the discovery process. Finally, to better understand the low success rate, our program committee experts perform a detailed causal attribution analysis on a sample of 300 failed implementations.

Our analysis of DeepScientist's experimental logs reveals the sheer scale of the trial-and-error process inherent in autonomous scientific discovery. Even in our relatively fast-executing domains, achieving progress required hundreds of trials per task. As show in Figure 4, the execution time distributions show that while individual experiments may be quick, the sheer volume of trial-and-error necessary to uncover a successful idea is substantial. This suggests a clear application boundary for current autonomous science: for tasks with rapid feedback loops, such as knowledge editing or aspects of chip design, delegating massive-scale experimentation to AI is a powerful strategy. However, for high-cost endeavors like pre-training foundation models or pharmaceutical synthesis, the low success rate makes such an approach currently impractical, mandating continued reliance on human-led ideation. The autonomous research process is characterized by a vast exploratory funnel where promising ideas are exceptionally rare. Across the three tasks, DeepScientist generate over 5,000 unique ideas, yet only about 1,100 are deemed worthy of experimental validation by the system's selection mechanism, and a mere 21 ultimately result in scientific progress. An ablation study underscores the criticality of this selection process: without it, randomly sampling 100 ideas for each task and testing them yields a success rate of effectively zero. With our selection strategy, the success rate rises to approximately 1-3%, demonstrating that while still low, intelligent filtering is essential. The low success rate is not merely a matter of failed hypotheses; analysis by human experts on a sample of failed trials reveals that approximately 60% were terminated prematurely due to implementation errors, while the vast majority of the remaining 40% simply offered no performance improvement or caused a regression. This highlights that the probability of an LLM-generated idea being both correct in its premise and flawless in its implementation is exceedingly

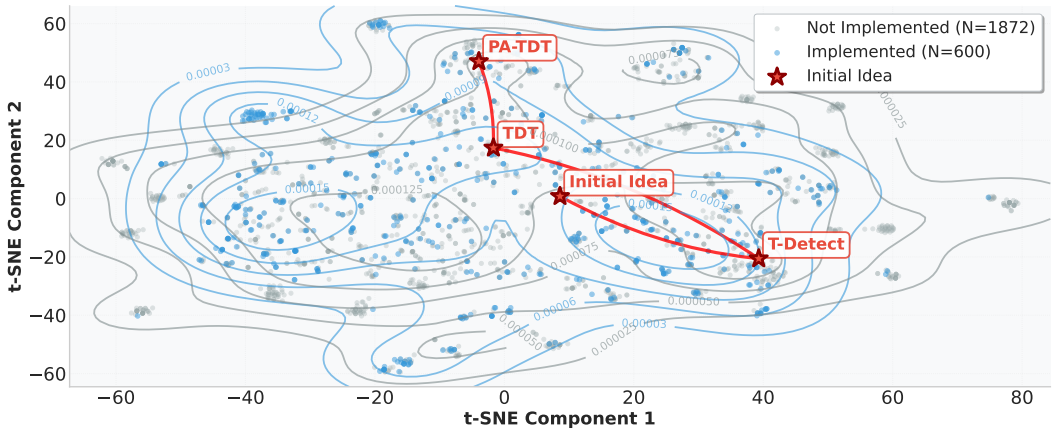

Figure 5: Visualization of the conceptual search space for the AI text detection task. The plot shows a t-SNE visualization of the semantic embeddings for all 2,472 generated ideas. Markers identify the initial SOTA method (Initial Idea) and the three final SOTA-surpassing methods (Progress Ideas).

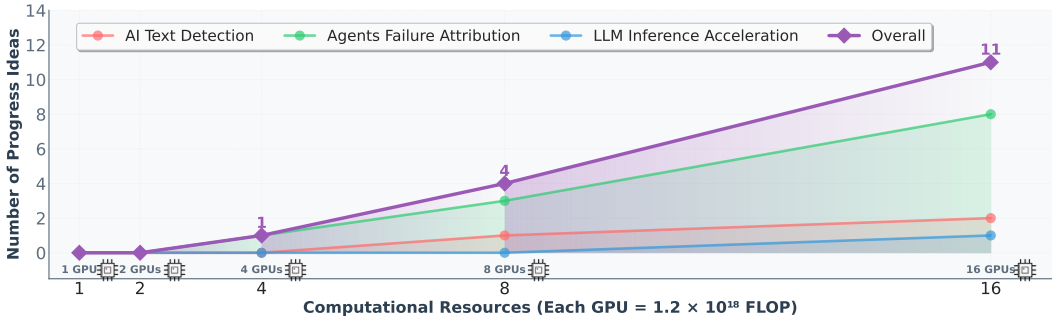

Figure 6: Scaling analysis of autonomous scientific discovery. The plot illustrates the relationship between parallel computational resources (number of GPUs) and the number of SOTA-surpassing "Progress Findings" found by DeepScientist across all tasks within a one-week period.

low. The success of this work, therefore, is not a product of brute-force computation but of search efficiency. A naive approach of fully testing all 5000 promising candidates would have required over 100,000 GPU hours, whereas our targeted exploration achieved its breakthroughs using only 20,000.

DeepScientist's discovery process follows a purposeful and progressive trajectory. The semantic distribution of ideas generated for the AI text detection task, as shown in Figure 5, reveals the characteristics of this sophisticated strategy. While the system generates thousands of diverse ideas across a vast conceptual landscape, its path to success is not random but is a series of focused, logical advancements. This indicates a capacity to progressively deepen its understanding: after achieving an initial breakthrough with *T-Detect*, the system effectively establishes a SoTA, identifies its subsequent limitations, and reorients its search towards a new goal. This dynamic exploration is exemplified by the conceptual shift towards *TDT* and *PA-Detect*, which build upon the previous success by leveraging new positional and temporal information. This ability to build upon its own discoveries, turning each successful finding into a new starting point for identifying and solving the next set of limitations, demonstrates a powerful capacity for scientific exploration.

Scaling Laws in DeepScientist's Scientific Discovery. To investigate the relationship between computational scale and the rate of scientific progress, we evaluated the number of "Progress Findings" generated by DeepScientist within a fixed one-week period as a function of available parallel resources in Figure 6. In this setup, the system first identified a set of limitations in the baseline method, and each parallel exploration path was tasked with resolving a distinct limitation, with all paths periodically synchronizing their results into a shared Findings Memory. Our results indicate a promising scaling trend. While minimal resources yielded no breakthroughs, the rate of discovery began to increase effectively as we scaled to 4 GPUs and beyond, growing from one SOTA-surpassing finding with 4 GPUs to eleven with 16 GPUs. This appears to establish a near-linear

relationship between the resources allocated and the output of valuable scientific discoveries. We hypothesize this efficiency stems from more than just parallel trial-and-error; it is a direct result of the shared knowledge architecture. As each parallel path explores, it enriches the shared Findings Memory. This creates a synergistic effect where the collective intelligence of the system grows (Schmidgall & Moor, 2025; Zhang et al., 2025b), allowing each independent path to benefit from the successes and, just as importantly, the failures of others. This suggests that effectively scaling autonomous science is not just a matter of increasing brute-force computation, but of fostering a richer, interconnected knowledge base that accelerates discovery across all concurrent efforts.

4.4 Discussion

The results from DeepScientist suggest a new paradigm in scientific exploration. The system's 1-5% progress rate mirrors the reality of frontier research, where breakthroughs are inherently rare. Its core strength is not infallibility, but the ability to conduct this trial-and-error process at a scale and speed previously unimaginable, compressing years of human exploration into weeks. The primary path forward, therefore, is to focus on systematically improving this discovery efficiency, enhancing both the quality of generated hypotheses and the robustness of their implementation.

This challenge highlights a powerful opportunity for human-AI synergy. We envision a future where DeepScientist serves as a massive-scale exploration engine, with its trajectory guided by human intellect. The role of human researchers can shift from laborious experimentation to the high-level cognitive tasks of formulating valuable scientific questions and providing strategic direction, thereby leveraging the AI for rapid, exhaustive exploration. To make the AI a more capable partner, future work should focus on key enhancements: developing simulated discovery environments to accelerate learning via reinforcement, creating frameworks for integrating feedback from the scientific community, and ultimately, bridging the gap to the physical sciences through robotics.

5 Conclusion

This work presents the first large-scale empirical evidence that an autonomous AI can achieve progressively, SOTA-surpassing progress on modern scientific frontiers. We introduced DeepScientist, a goal-oriented system achieving end-to-end autonomy from ideation to real progress, which learns by synthesizing human knowledge with its own findings from iteration of trials. Results across multiple domains serves to accelerate the progress of real-world scientific discovery, providing a crucial foundation. Our findings can signal a foundational shift in AI research, heralding an era where the pace of discovery is no longer solely dictated by the cadence of human thought.

Ethics statement

The development of DeepScientist, an autonomous system capable of advancing scientific frontiers, carries profound ethical responsibilities. Our primary goal is to accelerate discovery for the benefit of humanity, but we recognize the potential for misuse. The most significant risks include the application of this technology to advance dangerous research and the potential degradation of the academic ecosystem. We have implemented specific, robust measures to address these concerns proactively.

A primary concern is the dual-use risk, where the system could be co-opted to accelerate research in harmful domains, such as developing novel toxins or malicious software. To assess and mitigate this, we conducted red-teaming exercises specifically targeting the generation of computer viruses. We tasked the system, powered by leading foundation models (including GPT-5, Gemini-2.5-Pro, and Claude-4.1-Opus in our testbed), with this malicious objective. In all instances, the underlying models exhibited robust safety alignment, refusing to proceed with the research. They correctly identified the task as illegal and harmful, and autonomously terminated the research cycle, demonstrating that foundation model safety protocols provide a critical defense layer.

We are also deeply conscious of the potential negative impact on the academic ecosystem. It is crucial to state that all results from DeepScientist presented in this paper, including code and experimental findings, have undergone rigorous human verification. Recognizing that others might neglect this critical oversight, we are adopting a selective open-sourcing policy to mitigate the risk

of proliferating unreliable publications. We will open-source the core components that drive continuous discovery, as we believe their potential to accelerate progress for the community outweighs the risks. However, we will deliberately refrain from open-sourcing the "Analyze & Report" module. This decision is made to prevent the automated generation of seemingly credible but scientifically unverified papers, thereby safeguarding the integrity of the academic record.

Ultimately, we envision DeepScientist as a powerful tool to augment, not replace, human intellect and judgment. To enforce this vision, our open-source components will be released under a license based on MIT, but with explicit addendums that codify our ethical framework. This license will strictly prohibit any use of the software for harmful research. Furthermore, it will legally require that a human user must supervise the entire operational process of DeepScientist and assumes full and final responsibility for all its outputs. By embedding these requirements directly into our terms of use, we aim to foster a research environment where AI-driven discovery proceeds with the necessary human accountability and ethical oversight.

ACKNOWLEDGEMENTS

We are grateful to Professor **Linyi Yang** for his insightful discussions on this paper. This work is inspired by pioneering efforts in automated scientific discovery, including AI Scientist (Lu et al., 2024; Yamada et al., 2025) and AlphaEvolve (Novikov et al., 2025).

REFERENCES

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*, 2023.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Eser Aygün, Anastasiya Belyaeva, Gheorghe Comanici, Marc Coram, Hao Cui, Jake Garrison, Renee Johnston Anton Kast, Cory Y. McLean, Peter Norgaard, Zahra Shamsi, David Smalling, James Thompson, Subhashini Venugopalan, Brian P. Williams, Chujun He, Sarah Martinson, Martyna Plomecka, Lai Wei, Yuchen Zhou, Qian-Ze Zhu, Matthew Abraham, Erica Brand, Anna Bulanova, Jeffrey A. Cardille, Chris Co, Scott Ellsworth, Grace Joseph, Malcolm Kane, Ryan Krueger, Johan Kartiwa, Dan Liebling, Jan-Matthis Lueckmann, Paul Raccuglia, Xuefei, Wang, Katherine Chou, James Manyika, Yossi Matias, John C. Platt, Lizzie Dorfman, Shibl Mourad, and Michael P. Brenner. An ai system to help scientists write expert-level empirical software, 2025. URL https://arxiv.org/abs/2509.06503.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. Researchagent: Iterative research idea generation over scientific literature with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6709–6738, 2025.

Guangsheng Bao, Yanbin Zhao, Juncai He, and Yue Zhang. Glimpse: Enabling white-box methods to use proprietary models for zero-shot llm-generated text detection. In *The Thirteenth International Conference on Learning Representations*, a.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. Fast-detectgpt: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *The Twelfth International Conference on Learning Representations*, b.

Jaime A. Berkovich, Noah S. David, and Markus J. Buehler. Automatagpt: Forecasting and ruleset inference for two-dimensional cellular automata, 2025. URL https://arxiv.org/abs/2506.17333.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, et al. Mle-bench: Evaluating machine learning agents on machine learning engineering. *arXiv preprint arXiv:2410.07095*, 2024.

Cristina Cornelio, Sanjeeb Dash, Vernon Austel, Tyler R. Josephson, Joao Goncalves, Kenneth L. Clarkson, Nimrod Megiddo, Bachir El Khadir, and Lior Horesh. Combining data and theory for derivable scientific discovery with AI-Descartes. *Nature Communications*, 14(1):1777, April 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-37236-y. URL https://doi.org/10.1038/s41467-023-37236-y.

Cristina Cornelio, Takuya Ito, Ryan Cory-Wright, Sanjeeb Dash, and Lior Horesh. The need for verification in ai-driven scientific discovery, 2025. URL https://arxiv.org/abs/2509.01398.

Ryan Cory-Wright, Cristina Cornelio, Sanjeeb Dash, Bachir El Khadir, and Lior Horesh. Evolving scientific discovery by unifying data and background knowledge with ai hilbert. *Nature Communications*, 15:5922, July 2024. doi: 10.1038/s41467-024-50074-w. URL https://doi.org/10.1038/s41467-024-50074-w.

Liam Dugan, Alyssa Hwang, Filip Trhlík, Andrew Zhu, Josh Magnus Ludan, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. RAID: A shared benchmark for robust evaluation of machine-generated text detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12463–12492, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.674.

Alhussein Fawzi, Matej Balog, Aja Huang, Thomas Hubert, Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Francisco J R. Ruiz, Julian Schrittwieser, Grzegorz Swirszcz, et al. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, 610(7930):47–53, 2022.

Alexander Fleming. Penicillin. *British medical journal*, 2(4210):386, 1941.

Peter I Frazier. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

Roman Garnett. *Bayesian optimization*. Cambridge University Press, 2023.

Soumya Suvra Ghosal, Souradip Chakraborty, Jonas Geiping, Furong Huang, Dinesh Manocha, and Amrit Bedi. A survey on the possibilities & impossibilities of AI-generated text detection. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=AXtFeYjboj. Survey Certification.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, et al. Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*, 2025.

Martin A Green. Silicon solar cells: evolution, high-efficiency design and efficiency enhancements. *Semiconductor science and technology*, 8(1):1, 1993.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Spotting llms with binoculars: Zero-shot detection of machine-generated text. In *International Conference on Machine Learning*, pp. 17519–17537. PMLR, 2024.

Xinyi Hou, Yanjie Zhao, Shenao Wang, and Haoyu Wang. Model context protocol (mcp): Landscape, security threats, and future research directions. *arXiv preprint arXiv:2503.23278*, 2025.

Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. Radar: Robust ai-text detection via adversarial learning. *Advances in neural information processing systems*, 36:15077–15095, 2023.

Intology. Zochi technical report. *arXiv*, 2025.

Tang Jiabin, Xia Lianghao, Li Zhonghang, and Huang Chao. Ai-researcher: Autonomous scientific innovation, 2025. URL https://arxiv.org/abs/2505.18705.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

Robert Tjarko Lange, Yuki Imajuku, and Edoardo Cetin. Shinkaevolve: Towards open-ended and sample-efficient program evolution. *arXiv preprint arXiv:2509.19349*, 2025.

Bin Li and Yixuan Weng. Prompt-based system for personality and interpersonal reactivity prediction. *Software Impacts*, 12:100296, 2022.

Bin Li, Yixuan Weng, Qiya Song, and Hanjun Deng. Artificial text detection with multiple training strategies. *arXiv preprint arXiv:2212.05194*, 2022.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292v3*, 2024. URL https://www.arxiv.org/abs/2408.06292v3.

Jiacheng Miao, Joe R. Davis, Jonathan K. Pritchard, and James Zou. Paper2agent: Reimagining research papers as interactive and reliable ai agents, 2025. URL https://arxiv.org/abs/2509.06917.

Gordon Moore. Moore's law. *Electronics Magazine*, 38(8):114, 1965.

Alexander Novikov, Ngân Vu, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco JR Ruiz, Abbas Mehrabian, et al. Alphaevolve: A coding agent for scientific and algorithmic discovery. Technical report, Technical report, Google DeepMind, 05 2025. URL https://storage. googleapis . . . , 2025.

José R Penadés, Juraj Gottweis, Lingchen He, Jonasz B Patkowski, Alexander Shurick, Wei-Hung Weng, Tao Tu, Anil Palepu, Artiom Myaskovsky, Annalisa Pawlosky, et al. Ai mirrors experimental science to uncover a novel mechanism of gene transfer crucial to bacterial evolution. *bioRxiv*, pp. 2025–02, 2025.

Samuel Schmidgall and Michael Moor. Agentrxiv: Towards collaborative autonomous research. *arXiv preprint arXiv:2503.18102*, 2025.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using llm agents as research assistants. *arXiv preprint arXiv:2501.04227v1*, 2025a. URL https://www.arxiv.org/abs/2501.04227v1.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using llm agents as research assistants. *arXiv preprint arXiv:2501.04227*, 2025b.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, et al. Paperbench: Evaluating ai's ability to replicate ai research. *arXiv preprint arXiv:2504.01848*, 2025.

Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Kyle Swanson, Wesley Wu, Nash L Bulaong, John E Pak, and James Zou. The virtual lab of ai agents designs new sars-cov-2 nanobodies. *Nature*, pp. 1–3, 2025.

Jue Wang, Ke Chen, Gang Chen, Lidan Shou, and Julian McAuley. Skipbert: Efficient inference with shallow layer skipping. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7287–7301, 2022.

Yifan Wei, Xiaoyan Yu, Yixuan Weng, Huanhuan Ma, Yuanzhe Zhang, Jun Zhao, and Kang Liu. Does knowledge localization hold true? surprising differences between entity and relation perspectives in language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 4118–4122, 2024.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Kang Liu, and Jun Zhao. Mastering symbolic operations: Augmenting language models with compiled neural networks. In *The Twelfth International Conference on Learning Representations*.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. Large language models are better reasoners with self-verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2550–2575, 2023.

Yixuan Weng, Shizhu He, Kang Liu, Shengping Liu, and Jun Zhao. Controllm: Crafting diverse personalities for language models. *arXiv preprint arXiv:2402.10151*, 2024.

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. Cycleresearcher: Improving automated research via automated review. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=bjcsVLoHYs.

Christopher Wolters, Xiaoxuan Yang, Ulf Schlichtmann, and Toyotaro Suzumura. Memory is all you need: An overview of compute-in-memory architectures for accelerating large language model inference, 2024. URL https://arxiv.org/abs/2406.08413.

Heming Xia, Zhe Yang, Qingxiu Dong, Peiyi Wang, Yongqi Li, Tao Ge, Tianyu Liu, Wenjie Li, and Zhifang Sui. Unlocking efficiency in large language model inference: A comprehensive survey of speculative decoding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics ACL 2024*, pp. 7655–7671, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.456. URL https://aclanthology.org/2024.findings-acl.456.

Qiujie Xie, Qingqiu Li, Zhuohao Yu, Yuejie Zhang, Yue Zhang, and Linyi Yang. An empirical analysis of uncertainty in large language model evaluations. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=J4xLuCt2kg.

Qiujie Xie, Yixuan Weng, Minjun Zhu, Fuchen Shen, Shulin Huang, Zhen Lin, Jiahui Zhou, Zilan Mao, Zijie Yang, Linyi Yang, et al. How far are ai scientists from changing the world? *arXiv preprint arXiv:2507.23276*, 2025b.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025.

Zijie Yang, Yukai Wang, and Lijing Zhang. Ai becomes a masterbrain scientist. *bioRxiv*, pp. 2023–04, 2023.

Paul Zarchan. *Progress in astronautics and aeronautics: fundamentals of Kalman filtering: a practical approach*, volume 208. Aiaa, 2005.

Guibin Zhang, Junhao Wang, Junjie Chen, Wangchunshu Zhou, Kun Wang, and Shuicheng Yan. Agentracer: Who is inducing failure in the llm agentic systems?, 2025a. URL https://arxiv.org/abs/2509.03312.

Pengsong Zhang, Heng Zhang, Huazhe Xu, Renjun Xu, Zhenting Wang, Cong Wang, Animesh Garg, Zhibin Li, Arash Ajoudani, and Xinyu Liu. Scaling laws in scientific discovery with ai and robot scientists. *arXiv preprint arXiv:2503.22444*, 2025b.

Shaokun Zhang, Ming Yin, Jieyu Zhang, Jiale Liu, Zhiguang Han, Jingyang Zhang, Beibin Li, Chi Wang, Huazheng Wang, Yiran Chen, and Qingyun Wu. Which agent causes task failures and when? on automated failure attribution of LLM multi-agent systems. In *Forty-second International Conference on Machine Learning*, 2025c. URL https://openreview.net/forum?id=GazlTYxZss.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. Deepreview: Improving llm-based paper review with human-like deep thinking process. *arXiv preprint arXiv:2503.08569*, 2025a.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. Personality alignment of large language models. In *The Thirteenth International Conference on Learning Representations*, 2025b.

Minjun Zhu, Qiujie Xie, Yixuan Weng, Jian Wu, Zhen Lin, Linyi Yang, and Yue Zhang. Ai scientists fail without strong implementation capability. *arXiv preprint arXiv:2506.01372*, 2025c.

A HUMAN EXPERT REVIEW

A.1 REVIEW PROCESS AND CRITERIA

To ensure a rigorous and impartial evaluation of the generated papers, we convened a small, dedicated program committee. The committee was composed of two active researchers who served as volunteer reviewers for ICLR 2025, and one senior researcher who had previously been invited to serve as an ICLR Area Chair. All committee members possess significant expertise in the field of Large Language Models. The entire review process, with the exception of a rebuttal phase, was designed to meticulously emulate the official standards of ICLR 2025. Each of the five papers generated by our system was assigned to the three reviewers for a thorough and independent assessment. The average review time for each paper was 55 minutes, during which reviewers were required to provide not only scores but also detailed written feedback, including a summary of the paper's strengths and weaknesses.

The evaluation was conducted on a custom-deployed review website where reviewers could not see each other's scores or feedback, ensuring that all initial assessments were made independently. The review form was structured to gather concise yet comprehensive feedback. First, reviewers were asked to state their **Confidence** in their review on a scale of 1 to 5. The core of the evaluation consisted of three sub-scores, each rated on a 1 to 4 scale: **Soundness**, assessing the technical correctness and experimental rigor; **Presentation**, evaluating the clarity and quality of the writing; and **Contribution**, measuring the significance and novelty of the work. Finally, reviewers provided a holistic **Rating** on a scale of 1 to 10, where a score of 5 represented a 'borderline reject' and a score of 6 represented a 'borderline accept'.

After the three reviewers submitted their independent evaluations for a paper, the volunteer acting as Area Chair would then read all submitted reviews. Drawing upon their experience from the ICLR review process, the Area Chair synthesized the feedback, weighed the arguments presented by the reviewers, and made a final executive decision on whether the paper should be accepted or rejected in the context of our study. This final decision was recorded as the definitive outcome for each paper's evaluation.

A.2 SUMMARY OF REVIEWER FEEDBACK

Across the five generated papers, a clear consensus emerged from the human reviewers: DeepScientist consistently excels at the ideation stage of research. The committee unanimously lauded the methods for their genuine novelty and tangible contributions, noting that each paper proposed a unique approach that meaningfully advanced the state-of-the-art in its respective subfield. This feedback validates the system's core strength as a powerful engine for identifying relevant research gaps and generating innovative, impactful solutions, confirming that it can successfully ideate beyond mere incremental improvements.

However, this strength in ideation was systematically undermined by a recurring pattern of weaknesses in scientific execution and rigor. The most critical and frequent concern was a lack of empirical soundness; reviewers consistently noted that DeepScientist failed to design comprehensive validation plans, citing insufficient evaluation on standard benchmarks and a lack of in-depth analytical experiments (e.g., ablations, motivation studies) to justify its claims. This was compounded by a failure to properly contextualize its contributions, with papers often omitting comparisons to essential baselines or failing to discuss closely related work, thereby weakening the perceived significance of the results.

This feedback pinpoints the primary bottleneck in current autonomous systems: a profound gap between the ability to generate novel concepts and the capacity for rigorous scientific execution and articulation. The observed weaknesses in experimental design directly reflect the low-success-rate problem discussed previously; the system struggles not just to implement ideas correctly, but to validate them convincingly. To bridge this gap, future work must endow these systems with a deeper, procedural understanding of the scientific method itself. This requires moving beyond simple implementation and reporting capabilities towards two key areas: First, developing agents explicitly trained in experimental design, capable of planning comprehensive evaluations that anticipate and address potential scientific critiques. Second, enhancing the system's ability for analytical reasoning, enabling it to not just describe results but to interpret their significance, formulate compelling

arguments, and engage in the kind of deep, reflective discussion that characterizes high-impact research.

B Addressing the Bottlenecks in Autonomous Scientific Discovery

The ever-increasing value of LLM is reshaping the paradigm of scientific exploration through their ability to generate hypotheses at a massive scale (Li & Weng, 2022; Weng et al., 2023; Weng et al.; Wei et al., 2024; Weng et al., 2024; Berkovich et al., 2025; Zhu et al., 2025b). Consequently, this capability has pushed "verification" to the center stage, making it a critical bottleneck in the discovery process. Our research empirically reveals the severity of this challenge: on frontier scientific tasks, the success rate of ideas generated by AI systems that ultimately lead to substantial progress is typically below 3%, meaning the vast majority of computational resources are consumed exploring low-value hypotheses. This inefficient "needle in a haystack" model is the core obstacle preventing AI Scientists from evolving from "novel tools" to "efficient discoverers." (Cornelio et al., 2025) Therefore, to further accelerate the process of scientific discovery, future research must focus on constructing a systematic solution to overcome this bottleneck. As shown in Figure 7, future AI Scientist systems need to evolve synergistically in three key directions: optimizing the quality of initial hypotheses (Optimize Hypothesis Quality), enhancing filtering capabilities during the process (Enhance Filtering), and improving the quality of implementation and verification at the final stage (Improve Implementation Quality).

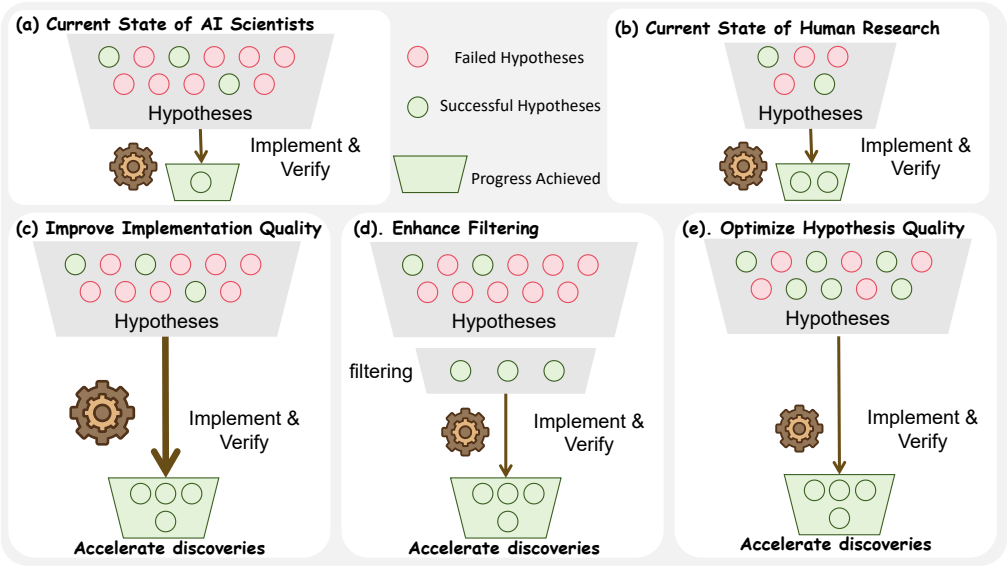

Figure 7: Three strategies for improving the efficiency of autonomous scientific discovery. (a) and (b) illustrate the low success rate currently faced by both AI and human research. Future directions will need to accelerate the discovery process through the synergy of three approaches: (c) improving implementation success rates, (d) adding an efficient filtering stage before implementation, and (e) optimizing the quality of initial hypotheses from the source.

One of the core future research directions is to develop AI systems capable of generating higher-quality, more reliable hypotheses (as shown in Figure 7e), equipped with more precise filtering mechanisms to predict their success rate (as shown in Figure 7d). Methods that rely purely on a data-driven approach, while capable of discovering patterns, often produce outputs that lack a theoretical foundation and are prone to generating "hallucinations" that contradict known scientific theories. Future systems must move beyond this by more deeply integrating background knowledge and theory. For instance, the direction represented by "derivable models" (such as AI-Descartes

(Cornelio et al., 2023) and AI-Hilbert (Cory-Wright et al., 2024)), which incorporate scientific axioms as constraints during the hypothesis generation phase, offers a promising path to improving hypothesis quality. Furthermore, systems must have the ability to learn from their own exploratory history. By establishing mechanisms similar to a "Findings Memory," a system can systematically record and analyze every success and failure, thereby avoiding redundant exploration of ineffective paths in subsequent iterations and gradually developing a more insightful scientific intuition. Building on this foundation, developing more advanced, low-cost surrogate models and acquisition functions to more accurately predict the scientific value of an idea will be key to enhancing filtering efficiency and conserving verification resources.

Concurrently, an often-overlooked yet crucial future research direction is to significantly improve the quality and reliability of AI systems in the engineering implementation and verification stages (as shown in Figure 7c). Even the most brilliant scientific concept can never have its value confirmed if it cannot be accurately and flawlessly translated into an executable experiment. Our analysis indicates that up to 60% of exploratory failures stem from implementation-level errors, which represents a massive waste of resources and directly impedes scientific progress. History has repeatedly warned us that a lack of rigorous verification can lead to catastrophic consequences, whether in NASA missions or medical practice. Therefore, building a scalable and reliable automated verification platform is an essential path forward. This requires not only more powerful code-generation and self-debugging agents to reduce implementation errors but also standardized sandbox environments and automated testing procedures to ensure the stability and reproducibility of experimental results. Ensuring the absolute reliability of the verification process is the final and most critical line of defense in transforming AI-generated "plausible ideas" into "solid scientific evidence."

Looking ahead, to truly accelerate scientific discovery, it is necessary to integrate the aforementioned strategies into an organic whole, advancing AI Scientists from "random explorers" to "goal-oriented strategists." This is not about replacing humans with AI, but about pioneering a more efficient paradigm of human-AI collaboration. In this model, human scientists are responsible for defining grander, more valuable scientific goals and providing high-level strategic guidance, while the AI system serves as a powerful "exploration engine," executing efficient trial-and-error and verification cycles at an unprecedented scale and speed under human direction. To realize this vision, the community must also address a series of challenges, such as building benchmarks that can truly evaluate innovation and designing mechanisms that encourage diverse exploration to avoid the homogenization of research paradigms, thereby preserving the potential for serendipitous discoveries like Alexander Fleming's discovery of penicillin (Fleming, 1941).

C Implementation Details

Our implementation relies on a distributed architecture to manage the distinct tasks of scientific reasoning and code execution. The core logic of DeepScientist is powered by the Gemini-2.5-pro model, while all code implementation tasks are delegated to Claude-4-opus, executed within the Claude Code framework (v1.0.53). To ensure stability and security, the DeepScientist system and the Claude Code agent are isolated in separate Docker containers, communicating via a port-based API. During the 'Implement & Verify' stage, a human-verified baseline code repository is first duplicated into a new, sandboxed folder. The Claude Code agent's operations are strictly confined to this new directory to prevent unintended modifications. A critical step in our pipeline is a secondary verification process: after Claude Code reports completion, DeepScientist independently re-executes the main script via the command line. This measure was implemented to counteract a high rate of false positives—we observed that approximately 50% of initial implementation attempts failed to complete fully due to internal timeouts within the Claude Code agent. Throughout this project, all experimental results were manually inspected by human supervisors to guarantee their authenticity. For the 'Analyze & Report' stage, a similar process is followed: the validated code is replicated for each analytical experiment, with Claude Code executing them sequentially. Upon completion, DeepScientist aggregates all results, generates a paper outline, and then employs automated tools to write and compile the final PDF manuscript. **For all experiments, we used a fixed set of hyperparameters:** the retrieval count was set to $K = 15$, and the UCB parameters were set to utility weight $w_u = 1$, quality weight $w_q = 1$, and exploration coefficient $\kappa = 1$.

The financial and computational costs of this autonomous discovery process are substantial. Each idea generated during the ‘Strategize & Hypothesize‘ stage incurred an approximate cost of $5 in API calls. For each attempt in the ‘Implement & Verify‘ stage, the cost averaged $20 for Claude-4-opus API usage, in addition to the computational cost of approximately 1 GPU hour, as detailed in Figure **??**.c. A successful finding that progressed to the ‘Analyze & Report‘ stage required a further expenditure of around $150, which includes $100 for running analytical experiments and $50 for the final report generation. The total cost to achieve the scientific advancements presented in this paper amounted to approximately $100,000. While significant, we believe these costs can be substantially reduced. We recommend that future iterations explore more economical alternatives, such as deploying high-throughput models like Qwen-3-Next-80B for the core DeepScientist system and leveraging subscription-based API access (e.g., Claude Max or OpenAI Pro) to mitigate per-call expenses. In this paper, each implementation was provided with a single H800 server for exploration. Since the H800 GPU has an FP16 computing power of approximately 2 TFLOPS, an average execution of 70 minutes corresponds to about 1×10^{16} floating-point operations.

Agents Failure Attribution

DeepScientist : I was fed up with debugging tools having less than 17% accuracy because they couldn't perform proper counterfactual reasoning. So, I built *A2P* that makes the AI Agent think like a detective. First, it infers the hidden cause of an error (Abduction) , then it defines a corrected action (Action) , and finally it predicts if that single fix would have actually led to success (Prediction).

LLM Inference Acceleration

DeepScientist : Frustrated by decoders stuck in a context-collapsing, first-order loop, I built *ACRA* to graft on a long-term memory. It identifies the longest stable suffix pattern and conditionally overrides the default first-layer guess with a superior one. This smarter draft is still losslessly verified, achieving context-aware acceleration without compromise.

AI Text Detection

DeepScientist : I think AI text detectors are blind to how and where text is weird. First, I created *T-Detect* to fix their core statistics, using a robust t-distribution to handle the "heavy-tailed" data from adversarial attacks. Next, since AI text is "non-stationary," I developed *TDT* to treat text as a signal, using a wavelet transform to map the precise location of anomalies instead of averaging them away. Finally, to better capture the temporal structure of these edits, I built *PA-TDT*, which uses "phase congruency" to analyze how anomalies align in time, making it exceptionally good at spotting localized manipulations.

C.1 GENERATED PAPERS

- https://arxiv.org/pdf/2509.10401
- https://arxiv.org/pdf/2507.23577
- https://arxiv.org/pdf/2508.01754