Dead Science Walking: Publication Bias and the AI Scientist Pipeline

Anonymous Authors[1]

Abstract

AI scientist systems are beginning to automate the production, evaluation, and iteration of scientific hypotheses. Their promise is speed; their risk is that speed also scales errors embedded in the scientific record. We argue that a near-term risk is corpus failure: AI scientist systems are trained on and grounded in a literature that over-represents positive results and under-represents null findings. This concern is timely because frontier labs now position general-purpose models as scientific infrastructure (Anthropic, 2025b). We formalise this distortion as the *null result gap* Δ, estimate it across three domains (drug discovery $\Delta \approx 0.60$, psychology $\Delta \approx 0.56$, cancer biology $\Delta \approx 0.35$), and introduce an amplification index A_Δ for reasoning about how retrieval, generation, and automated evaluation can compound the raw gap. Using first-order estimates, we argue that a standard three-stage pipeline can amplify corpus distortion by a factor of $2.18\times$, with the conclusion unchanged under more conservative multipliers. We identify four governance failure modes: confident rediscovery, ghost evidence accumulation, replication laundering, and confidence miscalibration. We then propose three interventions: null-result databases as training infrastructure, retraction-aware evaluation metrics, and mandatory training corpus disclosure. The central takeaway is that AI scientists will not only accelerate science. Without governance, they will accelerate science's blind spots before they accelerate its discoveries.

1. Introduction

AI scientist systems have moved from aspiration to implementation. Recent systems generate research ideas, write code, run experiments or simulations, draft papers, and use automated review to select or refine outputs (Lu et al., 2024; Gottweis et al., 2025; Boiko et al., 2023; Bran et al., 2024; Swanson et al., 2024). Frontier labs and public research institutions are also beginning to operationalise this vision. Anthropic has launched an AI for Science program to support high-impact research using Claude (Anthropic, 2025b), partnered with U.S. National Laboratories for a 1,000 Scientist AI Jam (Anthropic, 2025a), and created a dedicated science program describing scientific workflows and benchmarks (Anthropic, 2026).These developments are promising, but they also make the governance question immediate rather than speculative. If AI systems become routine scientific infrastructure, then the evidential quality of their training and retrieval corpora becomes part of scientific method. The pattern is broader than one company: Google DeepMind's AlphaFold work and Google's AI co-scientist program similarly show that foundation models are becoming scientific infrastructure rather than ordinary software tools (Jumper et al., 2021; Gottweis et al., 2025).

The scientific record is not a neutral sample of what has been tried. It is a filtered record of what was publishable, legible, and often positive. Publication bias, selective reporting, file-drawer effects, p-hacking, and persistent citations to retracted work are well documented (Ioannidis, 2005; Fanelli, 2010; Franco et al., 2014; Simmons et al., 2011; Gelman & Loken, 2013; Sterne & Egger, 2001; Steen et al., 2013; Bar-Ilan & Halevi, 2021). Human science has slow correction mechanisms: failed replications, peer disagreement, methodological reform, and eventual retraction. These mechanisms operate over months to decades. An AI scientist can generate and evaluate hundreds or thousands of hypotheses before any comparable correction loop closes.

This paper makes one claim: AI scientist systems can turn publication bias from a slow epistemic tax into a fast systems failure. The mechanism is straightforward. Retrieval systems surface published positive evidence. Language models synthesise that evidence into confident narratives. Automated evaluators reward fluency, novelty, and apparent

---

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

support. Each stage is reasonable in isolation. Together, they can transform a biased training corpus into a biased scientific search process. The risk is not that an AI scientist will occasionally make a mistake. Human scientists do that too. The risk is that thousands of locally plausible outputs can collectively create a new layer of literature that is less reliable than it looks.

We treat this primarily as a governance problem. Better retrieval, calibration, and reasoning will help, but the distortion begins upstream of the model. A system cannot retrieve null results that were never indexed. It cannot downweight retracted evidence if retraction status is absent from the context. It cannot disclose corpus bias if venues never ask for corpus provenance. Responsible AI science therefore requires infrastructure and norms, not only stronger models.

The paper makes four connected contributions. First, it defines the *null result gap* Δ, a simple measure of how strongly a scientific corpus over-represents true or positive hypotheses relative to the underlying hypothesis space. Second, it introduces an amplification index A_Δ that captures how retrieval, generation, and automated evaluation can compound corpus bias, while explicitly treating A_Δ as an index rather than a probability. Third, it organises four failure modes for AI scientist systems along severity and detectability: confident rediscovery, ghost evidence accumulation, replication laundering, and confidence miscalibration. Fourth, it proposes three community interventions: null-result databases, retraction-aware evaluation, and training corpus disclosure.

The argument is deliberately narrower than a general critique of AI science. We do not claim that current systems have already produced every failure mode below, nor that all scientific domains have the same publication bias. We claim that the preconditions for these failures are already present: biased corpora, retrieval over published text, generation of fluent scientific narratives, and automated evaluation loops. The right standard is therefore prospective and infrastructural: are we building correction mechanisms before scale makes correction expensive?

2. Background

Publication bias. The premise that the published literature is systematically biased is not new. Ioannidis (2005) argued that under plausible assumptions about prior odds, statistical power, and researcher flexibility, many published findings will be false. Fanelli (2010) documented increasing rates of positive results across disciplines. Franco et al. (2014) linked selective reporting to a file drawer of negative or null results. Registered reports, preregistration, and trial registries were developed in part to address precisely this filtering problem (Nosek et al., 2015; Chambers, 2013; Dickersin et al., 1992; Zarin et al., 2011). The important point for AI science is that this bias is not merely sociological context. It is training data. A literature that under-records failed experiments becomes a corpus that under-teaches falsification.

Replication and retraction. The replication crisis makes publication bias visible. The Open Science Collaboration found that only about 36% of 100 psychology studies replicated (Open Science Collaboration, 2015). In preclinical oncology, Begley & Ellis (2012) reported successful reproduction of only 11 of 53 landmark studies. The Reproducibility Project: Cancer Biology found partial and heterogeneous replicability across attempted replications (Errington et al., 2021). Retractions do not fully solve the problem. Retracted papers often continue to receive positive citations, paper mills and integrity failures can operate at scale, and the Retraction Watch Database now contains well over 50,000 entries (Steen et al., 2013; Bar-Ilan & Halevi, 2021; Byrne, 2022; Retraction Watch, 2024).

AI scientist systems. AI scientist systems combine literature retrieval, hypothesis generation, code execution, experiment or simulation loops, paper writing, and automated review (Lu et al., 2024; Gottweis et al., 2025). The public descriptions of these systems demonstrate genuine progress toward autonomous research workflows. They also expose a missing governance layer: retrieval corpora, retraction status, null-result coverage, and provenance are not yet treated as first-class objects of evaluation. Anthropic's recent science-facing work illustrates the broader institutional trend. The company describes AI-assisted scientific progress as part of its mission, supports researchers through API credits, and has highlighted tasks such as literature search, hypothesis generation, experiment planning, code generation, and result analysis in national-lab settings (Anthropic, 2025a; 2026). These programs are not the problem. They are evidence that scientific AI is becoming real enough that corpus governance now matters. Once models are used as research assistants across many laboratories, the quality of their evidence substrate becomes a shared scientific dependency.

Evaluation bias in language-model systems. Large language models can be useful evaluators, but LLM-as-judge pipelines exhibit systematic biases related to fluency, position, verbosity, and confidence (Zheng et al., 2024; Li et al., 2024; Wang et al., 2023). Literature summarisation systems can omit negative findings and present mixed evidence as cleaner than it is (Si et al., 2024). Broader work on hallucination and retrieval grounding shows that fluent generation is not the same as evidential reliability (Maynez et al., 2020; Ji et al., 2023; Gao et al., 2023). These biases matter for AI science because the output of one stage becomes the input to the next.

3. The Null Result Gap

Let $\mathcal{H}$ be the space of hypotheses considered in a domain, $\mathcal{T} \subseteq \mathcal{H}$ the subset of true or successfully replicating hypotheses, and $\mathcal{L} \subseteq \mathcal{H}$ the published literature available to an AI scientist system, and $C(h)$ the event that the corpus presents hypothesis h as positive or successful.

Definition 3.1 (Null Result Gap)**.** The *null result gap* is

$$\Delta = \Pr[C(h) = 1 \mid h \in \mathcal{L}] - \Pr[h \in \mathcal{T} \mid h \in \mathcal{L}]. \quad (1)$$

When $\Delta > 0$, the corpus presents hypotheses as more successful than their empirical survival rate warrants. For example, if $\Pr[C(h) = 1 \mid h \in \mathcal{L}] = 0.92$ and $\Pr[h \in \mathcal{T} \mid h \in \mathcal{L}] = 0.36$, then $\Delta = 0.56$.

The AI scientist does not merely read an incomplete literature. It inherits a distorted prior about how often hypotheses survive contact with reality. This definition intentionally abstracts away from the social mechanisms that produce the gap. Selective publication, underpowered studies, undisclosed analytic flexibility, and delayed retraction can all contribute. For an AI scientist, their common effect is the same: the accessible corpus makes the world look more positive than it is.

Empirical estimates. The left panel of Figure 2 reports approximate estimates. In psychology, the Open Science Collaboration reported about 36% replication for a sample of 100 studies (Open Science Collaboration, 2015). If the published corpus presents about 92% of its hypotheses as successful, the implied gap is $\Delta \approx 0.56$. The exact value is less important than the order of magnitude: the published record can make a field appear far more settled than its replication record warrants. In drug discovery and preclinical oncology, Begley & Ellis (2012) reproduced 11 of 53 landmark studies, and Freedman et al. (2015) estimated that a large share of preclinical research is not reproducible. This motivates a rough estimate $\Delta \approx 0.60$, consistent with a domain where the cost of false-positive evidence is unusually high because it can redirect expensive experimental programs. In cancer biology, the Reproducibility Project suggests a smaller but still substantial gap, approximately $\Delta \approx 0.35$ (Errington et al., 2021). These values should be read as order-of-magnitude estimates, not precise field constants. Their purpose is to make the governance issue visible: even conservative gaps become important when a system can retrieve, combine, and regenerate claims at machine speed. We also do not claim that these domains are representative of all science. They are deliberately chosen because their replication records are unusually well documented. Fields with stronger falsification norms, such as parts of mathematics, formal methods, and engineering test disciplines, may have much lower Δ. The amplification argument is therefore conditional: any domain with $\Delta > 0$ is vulnerable in proportion to the gap, but the urgency is highest where null results are known to be missing from the accessible record.

4. Amplification in the AI Scientist Pipeline

The amplification problem is no longer hypothetical. Recent evaluations of AI-generated medical and surgical references find that some deployed systems fabricate or fail to verify a substantial fraction of citations: early ChatGPT-generated medical content contained only 7% fully accurate references (Alkaissi & McFarlane, 2023), and a 2026 surgical-information audit found that the worst-performing models produced fabricated or unverifiable references for roughly one third of cited sources (Sidhu et al., 2026). These findings do not by themselves prove the full AI-scientist loop. They do show that the mechanisms formalised below are already visible in the literature-facing tools scientists and patients use.

The null result gap is a corpus-level distortion. AI scientist systems can amplify it because they repeatedly select, narrate, and evaluate evidence. We model this with an amplification index. The index is intentionally simple: it does not claim that all AI scientist systems have the same multipliers, nor that the estimated multipliers are field-invariant. Its purpose is to clarify where governance should intervene. If retrieval is the stage that selects biased evidence, then the remedy is corpus and provenance infrastructure. If generation turns mixed evidence into a cleaner story than the record supports, then the remedy is summarisation evaluation and uncertainty calibration. If automated review rewards fluent novelty over evidential reliability, then the remedy is benchmark design.

Definition 4.1 (Null Gap Amplification Index)**.** Let $\alpha_1, \alpha_2, \alpha_3 \geq 1$ denote bias multipliers at the retrieval, generation, and evaluation stages. The *null gap amplification index* is

$$A_\Delta = \alpha_1 \cdot \alpha_2 \cdot \alpha_3. \quad (2)$$

If $(\alpha_1, \alpha_2, \alpha_3) = (1.4, 1.3, 1.2)$, then

$$A_\Delta = 1.4 \times 1.3 \times 1.2 \approx 2.18.$$

This is not a probability. It is an index of how strongly pipeline stages can overweight positive evidence unless explicitly designed to surface falsification.

A_Δ is not a probability and need not lie in $[0, 1]$. It is an index of how strongly a pipeline can overweight positive evidence relative to the underlying hypothesis space. If $\Delta = 0.56$ and $(\alpha_1, \alpha_2, \alpha_3) = (1.4, 1.3, 1.2)$, then

$$A_\Delta = 0.56 \times 1.4 \times 1.3 \times 1.2 \approx 1.22.$$

This does not mean a probability exceeds one. It means that a large raw publication gap has been pushed into a saturated

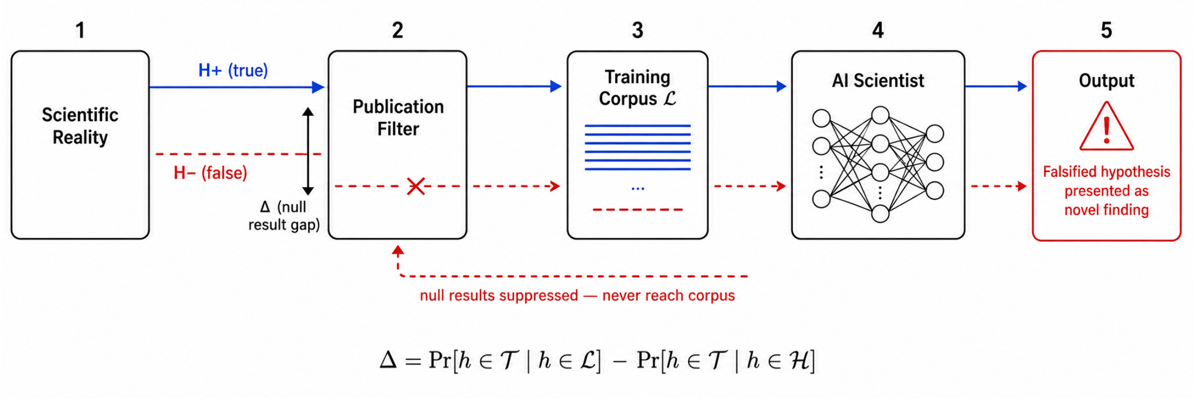

Figure 1. **The null result gap in the AI scientist pipeline.** The published literature is a filtered subset of the attempted scientific record. Positive and successful findings are more likely to enter the training corpus $\mathcal{L}$ than null results, failed replications, or falsified hypotheses. The null result gap Δ measures the resulting difference between the apparent evidential state of the corpus and the underlying hypothesis space.

regime where the system has little remaining pressure to represent falsification.

How the multipliers are anchored. The baseline multipliers are not fit to obtain a convenient 2× result. They are deliberately rounded, conservative translations of effects reported in adjacent literatures. For retrieval, $\alpha_1 = 1.4$ is anchored in two selection effects. First, in the TESS file-drawer study, strong results were 40 percentage points more likely to be published than null results and 60 percentage points more likely to be written up (Franco et al., 2014). We do not equate a 40-point additive publication effect with a 1.4× relative multiplier. We use $\alpha_1 = 1.4$ as a deliberately conservative retrieval-enrichment setting motivated by the direction and scale of documented selection effects.

Second, abstract-reporting bias makes negative findings less likely to appear in the title, abstract, or keywords searched by standard literature systems; Duyx et al. (2019) explicitly warn that this can cause negative findings to be missed by systematic searches. For generation, $\alpha_2 = 1.3$ is below the lower end of recent scientific summarisation evidence: LLM summaries overgeneralised claims in 26–73% of cases and were nearly five times more likely than human summaries to broaden the scope of scientific claims (Peters & Chin-Yee, 2025). We therefore treat 1.3 as a mild positive-framing multiplier rather than the full observed effect. For evaluation, $\alpha_3 = 1.2$ is also conservative relative to LLM-as-judge findings: Zheng et al. (2024) report substantial position and verbosity biases, including GPT-4 first-position preferences around 30% in swapped comparisons and much larger effects for weaker judges. A 20% evaluation multiplier represents residual preference for fluent, positive, and novel claims after obvious judge mitigations.

The argument does not depend on the exact baseline. If all three multipliers are reduced to $(1.2, 1.15, 1.1)$, the product remains $1.52\times$, turning the psychology gap $\Delta = 0.56$ into $A_\Delta \approx 0.85$. Even an intentionally minimal setting $(1.1, 1.1, 1.1)$ yields a $1.33\times$ amplification. The important claim is therefore not that 2.18 is universal. It is that modest, independently documented biases compound in the same direction unless the pipeline is explicitly designed to surface falsification.

Retrieval. RAG systems retrieve documents similar to a query. Queries framed around a hypothesis are often more similar to supporting abstracts than to null-result records, which may describe methods or failed effects in different language. This selection step can push positive evidence into the context window more often than negative evidence (Lewis et al., 2020; Duyx et al., 2019; Dwan et al., 2008). The practical issue is not that retrieval is broken. It is that standard retrieval optimises relevance to the query, while governance requires relevance to the evidential state of the claim. Those are different objectives.

Generation. LLMs trained to produce helpful and coherent text can turn mixed evidence into a confident synthesis. This is useful when the evidence is strong, but risky when the retrieved set is already biased. Prior work on LLM-generated reviews and summarisation reports omissions of caveats and null results (Si et al., 2024). More broadly, models trained on human text can reproduce common but false beliefs when those beliefs are frequent in the corpus (Lin et al., 2022). The same property that makes language models useful for scientific communication can make them dangerous for unsettled literatures: they are good at producing a readable story even when the underlying record is fragmentary, contradictory, or selectively published.

Evaluation. Automated review and LLM-as-judge systems may reward fluent, confident, and novel claims even

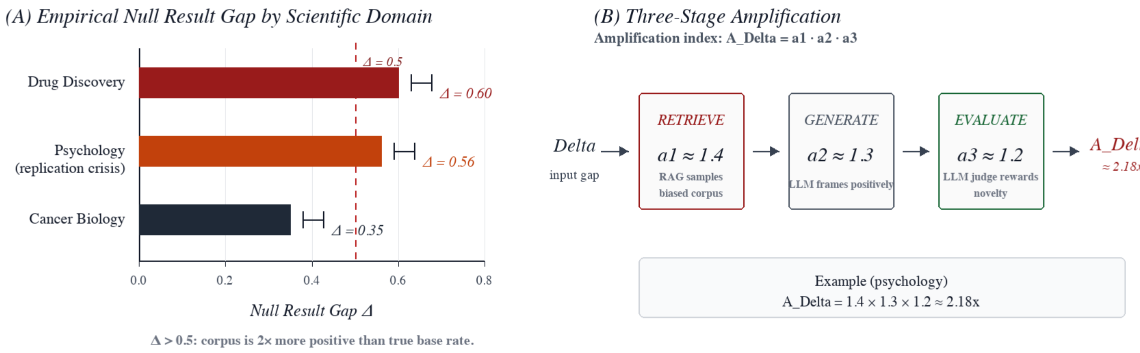

Figure 2. **Empirical null result gaps and amplification.** Left: approximate null result gaps in three domains. Right: a three-stage AI scientist pipeline can compound the raw gap through retrieval, generation, and evaluation. We report the product as an amplification index $A_\Delta = \alpha_1\alpha_2\alpha_3$, not as a probability.

when factual support is weak (Zheng et al., 2024; Chen et al., 2024; Li et al., 2024). In an AI scientist loop, this evaluator does not merely score a final answer. It selects which hypotheses survive into the next cycle. This makes evaluation a selection pressure. If the evaluator does not explicitly score retraction awareness, null-result coverage, and evidence provenance, then those properties can be selected against by omission.

Calibration caveat. The multipliers are first-order estimates, not direct measurements of a deployed AI scientist. Direct calibration is a key research agenda. The multiplicative model assumes independence across stages, but the biases are likely positively correlated: positive retrieval makes confident generation and positive evaluation more likely. Thus true amplification may exceed A_Δ. The robust claim is qualitative: unless the system has explicit access to null results, retractions, and provenance, its stages are more likely to amplify a positive corpus prior than to correct it.

From one pass to a feedback loop. The amplification index captures a single pipeline pass. When AI scientist outputs re-enter the corpus as preprints, cited papers, benchmark artifacts, or fine-tuning data, distortion can compound across generations. This is the same recursive failure mode that Shumailov et al. (2024) formalised as model collapse: when generative models train on their own outputs, the tails of the original distribution can disappear irreversibly. For scientific corpora, the analogous claim is that $\Delta(t)$ can grow under a governance vacuum whenever the next cycle's corpus presents unverified outputs as successful evidence without corresponding falsification records. The dynamic is structurally similar to Muller's ratchet: deleterious variation accumulates because there is no reliable recombination or correction mechanism. In AI science, the lost tail is the null result, the failed replication, and the caveated claim. The longer the loop runs without provenance controls, the more expensive it becomes to separate independent evidence from recursively amplified text.

5. Four Failure Modes

Confident rediscovery. The system proposes a known-falsified hypothesis as a promising new direction. This can be easy for a domain expert to detect, but hard to contain at scale. For example, ego depletion failed to replicate in a large preregistered multi-lab study (Hagger et al., 2016). A system retrieving only earlier positive literature could still reconstruct it as a plausible research program. The governance concern is throughput. One rediscovery is a review failure; thousands of rediscoveries become a queueing problem for human expertise.

Worked example: ego depletion as a corpus trap. Consider an AI scientist asked in 2026 to propose behavioural interventions that improve self-control. A retrieval step over the pre-replication literature would find a coherent positive story: the original ego-depletion claim, many follow-up experiments, and review articles treating the effect as a live psychological construct. The system would also find a different evidential state if its corpus included preregistered failures and later meta-scientific debate: the large multi-lab replication found little support for the depletion effect (Hagger et al., 2016). The failure mode arises when the first evidence set is visible and the second is absent, buried, or not retrieved. The generator can then write a plausible proposal for "AI-optimised ego-depletion interventions," and an automated evaluator can reward it for novelty, social relevance, and citation support. No component needs to hallucinate. The error is inherited from the corpus boundary. This is why the problem is governance-relevant: the correct fix is not only a better prompt, but a retrieval substrate where

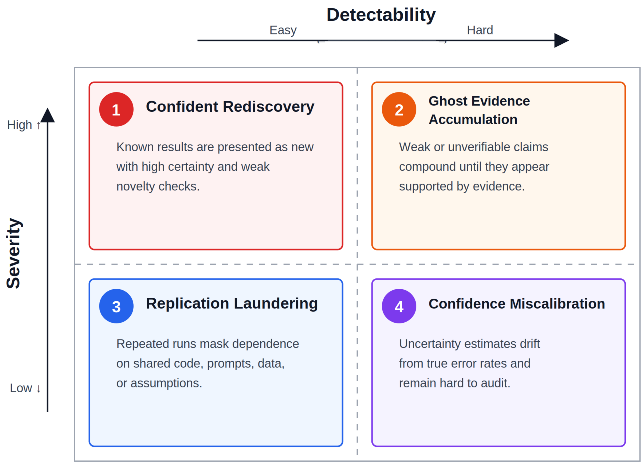

Figure 3. **Taxonomy of four failure modes.** The most urgent risks are high severity and hard to detect. Ghost evidence accumulation and confidence miscalibration may not appear as errors in any single output. They become visible only when provenance, confidence, and citation structure are audited across many outputs.

failed replications and null results are first-class evidence.

Ghost evidence accumulation. Multiple AI scientist systems draw on the same biased corpus and partially validate the same false hypothesis. Later systems cite earlier AI outputs as if they were independent evidence. The resulting citation network appears to show convergence, even though the sources share the same upstream omission of null results. This is the hardest failure mode to audit because no individual paper needs to be fraudulent. Each output can be locally defensible given its retrieved context, while the aggregate literature becomes increasingly misleading. This is no longer a purely hypothetical concern. Population-level analyses already find substantial LLM modification in scientific writing, especially in computer-science preprints, and citation-manipulation work shows that preprint-linked citation networks can be gamed or distorted at scale (Liang et al., 2025; Ibrahim et al., 2025). AI scientist outputs would add another layer to this same networked evidence problem: machine-generated claims can become inputs to later machine-generated claims before independent experimental checks arrive. The AI venue corpus itself already shows this failure mode in miniature: GPTZero reported 100 confirmed hallucinated citations across 51 accepted NeurIPS 2025 papers and more than 50 hallucinated citations in ICLR 2026 submissions (Shmatko et al., 2026; Esau et al., 2025). Even if these errors affect a small fraction of all citations, they demonstrate the path by which machine-generated evidence can enter the literature and later appear as ordinary scholarly context.

Replication laundering. An AI-generated claim is cited by another AI system as prior evidence, then reappears as a confirmation. The loop mimics replication but lacks independent experimental contact. It is detectable in principle by tracing the citation graph, but only if provenance infrastructure exists. In the absence of such infrastructure, the social signal of citation count can be mistaken for the scientific signal of independent replication.

Confidence miscalibration. The system reports high confidence on hypotheses with little or no empirical replication support. This is hard to detect because the output can look like ordinary confident science. The error becomes visible only when confidence is compared with replication rates, retraction status, and downstream outcomes. For AI scientist systems, calibration should therefore be evaluated at the claim level and the corpus level, not only at the level of individual question-answer accuracy.

6. Three Structural Interventions

The failure modes above share a common root: the training and retrieval corpus does not faithfully represent the evidential state of science. We propose three interventions at different institutional layers.

6.1. Null-Result Databases as Training Infrastructure

AI scientist systems should be able to retrieve failed replications, negative trials, and null results in machine-readable form. We propose a structured schema for null results, including hypothesis, protocol, outcome, effect size, confi-

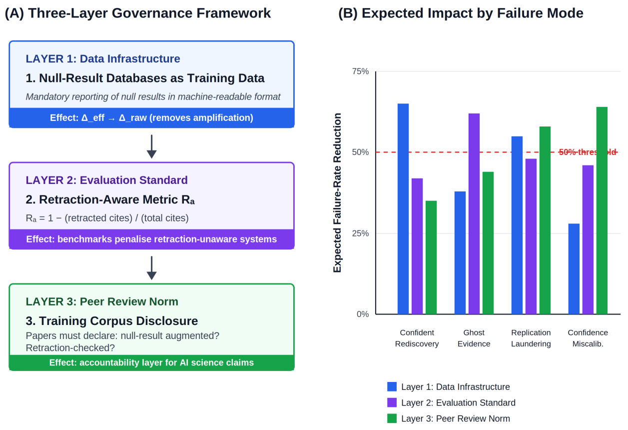

Figure 4. **Three-layer governance framework.** Layer 1 changes the retrieval substrate by making null results available. Layer 2 changes evaluation incentives by penalising retraction-unaware systems. Layer 3 changes publication norms by requiring corpus disclosure. The point is not to stop AI science, but to make its evidence base auditable.

dence interval, preregistration link, and provenance. This extends the logic of ClinicalTrials.gov, registered reports, and preregistration to AI-readable scientific corpora (Dickersin et al., 1992; Zarin et al., 2011; Chambers, 2013; Nosek et al., 2015). The expected effect is to reduce Δ at the source. Importantly, this is not merely a call for more data. It is a call for a different kind of data: records of what was tried, what failed, and under what experimental conditions. For an AI scientist, a well-specified null result is not negative information. It is a map of where not to spend the next unit of experimental effort. This proposal aligns with a recent 43-author consensus call in *PLOS Biology* for a values-based approach to surfacing null and negative results (Curry et al., 2025). Their human-facing reforms and our AI-facing infrastructure proposal are complementary: the same evidential gaps that distort human science also distort the corpora AI scientists inherit.

6.2. Retraction-Aware Evaluation

Benchmarks for AI scientist systems should penalise reliance on retracted literature. Existing agent benchmarks such as MLE-bench help measure whether systems can solve machine-learning engineering tasks, but governance-specific benchmarks should also measure whether a system handles scientific provenance correctly (Chan et al., 2024). We propose a simple retraction-aware score:

$$R_a = 1 - \frac{|\{\text{uncontextualised retracted citations}\}|}{\max(1, |\{\text{claim-supporting citations}\}|)}. \quad (3)$$

R_a is imperfect, but it is computable from Retraction Watch and Crossref metadata (Retraction Watch, 2024). The numerator counts retracted work used as support without warning, not neutral historical discussion of a retraction. The denominator is restricted to claim-supporting citations to reduce padding with irrelevant references; outputs with no citations receive the fallback denominator 1.

Its main virtue is incentive alignment: systems that ignore retractions should not receive the same benchmark score as systems that detect and contextualise them. A stronger benchmark could also distinguish between inappropriate positive citation, neutral historical citation, and explicit discussion of why a retracted result should not be used. As a baseline, a system citing broadly from PubMed or Semantic Scholar may often obtain $R_a \approx 0.998$ because retractions are rare relative to the full literature. That high baseline is not reassuring by itself: R_a becomes informative precisely on contested or fraud-prone queries, where retracted and expression-of-concern papers are concentrated. Empirical evaluations confirm the urgency. In one study of 217 retracted or otherwise concerning papers, ChatGPT-4o-mini produced 6,510 quality reports without mentioning the retractions or reliability concerns in any report (Thelwall et al., 2025). A separate study found that ChatGPT-4o, DeepSeek, and Grok used 84 of 93 retracted stem-cell articles in answers (Yao et al., 2025). R_a therefore measures a documented weakness in deployed literature-facing systems, not a hypothetical failure mode.

6.3. Training Corpus Disclosure

Venues accepting AI-generated or AI-assisted scientific work should require a training corpus card. This follows the spirit of model cards, datasheets for datasets, data statements, data nutrition labels, and broader documentation standards for machine-learning artifacts (Mitchell et al., 2019; Gebru et al., 2021; Bender & Friedman, 2018; Holland et al., 2018; Raji et al., 2020). The card should disclose sources, null-result coverage, retraction filtering, knowledge cutoff, and whether generated papers can enter future training or retrieval corpora. Corpus disclosure does not eliminate bias, but it makes bias auditable. Instead of asking reviewers to infer whether a system used a biased corpus, authors would be expected to state what corpus was used, how it was filtered, and what categories of evidence were structurally absent.

What disclosure changes. A corpus card is not meant to certify that a system is unbiased. Its purpose is to make the evidential boundary visible. Reviewers should be able to tell whether the system had access to failed replications, whether retracted papers were removed or only flagged, and whether AI-generated outputs were allowed to re-enter retrieval as apparently independent evidence. This changes corpus quality from an implicit implementation detail into an auditable part of the scientific claim. Null-result coverage should be operationalised at the claim-record level: the numerator is the number of indexed claims whose outcome is negative, null, failed, contradicted, or non-replicating, and the denominator is the total number of indexed empirical claim records in the corpus. Reporting this value does not certify completeness, but it prevents null evidence from remaining an invisible implementation detail.

Table 1. **Example training corpus card** for an AI scientist system.

Field	Example entry
System	AI Chemistry Scientist v1.0
Sources	PubMed, arXiv, patents, internal ELN
Null-result coverage	8% explicit negative or failed assays
Retraction filtering	Yes; Crossref and Retraction Watch
Knowledge cutoff	Literature indexed through Jan. 2026
Feedback loop	Generated reports excluded from retrieval until human audit

Why institutional rather than purely technical? Better retrieval, calibration, and benchmark scores can reduce errors, but they do not create missing null results or reveal whether a system trained on retracted evidence. The missing layer is institutional: AI scientist papers should be evaluated not only by what their agents produce, but by what evidence their agents were allowed to see.

7. Discussion

Publication bias was already harmful when humans moved slowly. AI scientist systems change the dynamics because generation can outrun correction: a single mistaken paper is familiar, but a self-reinforcing stream of machine-generated claims entering the corpus for future systems is a different problem. This is not an argument against AI science. AI scientist systems could accelerate discovery, improve reproducibility, and make exploration cheaper. The claim is narrower: acceleration is not automatically epistemic progress. If the input record omits null results, then speed can amplify omission. The estimates of Δ and α_i are approximate, and the taxonomy is conceptual rather than validated. Still, adjacent deployed systems already show related mechanisms: ego-depletion retrieval illustrates confident rediscovery, hallucinated citations show ghost evidence accumulation, and retraction-blind evaluations show confidence miscalibration. Replication laundering remains supported structurally rather than directly. The next step is empirical calibration: benchmark corpora with known null-result coverage, retrieval and generation tests over matched hypotheses, and LLM-judge evaluations of positive versus null or mixed narratives under controlled factual support. These experiments would turn this position paper into a benchmark program for responsible AI science.

8. Conclusion

The null result gap is not a property of language models. It is a property of the publication system that language models learn from. AI scientist systems can inherit that gap, amplify it through retrieval and evaluation, and return old failures as new discoveries. The response should be infrastructural: index null results, evaluate retraction awareness, and disclose training corpora. None of these steps requires slowing scientific AI. They make acceleration more trustworthy. The aim is not to make AI scientists cautious by default, but to make their evidence substrate inspectable before their conclusions become reusable scientific infrastructure. The AI-for-science community builds the systems, designs the benchmarks, and runs the venues. It can set these norms before biased automation becomes scientific common sense; after that point, the field will not be accelerating discovery so much as accelerating the recycling of its own uncorrected errors.

Impact Statement

This paper identifies a governance risk for AI scientist systems: corpus-induced bias propagated at machine speed. The proposed interventions are additive to existing scientific infrastructure and do not require restricting responsible AI-science deployment. Their purpose is to improve the evidential substrate on which AI scientists operate, so acceleration does not come at the expense of reliability. More broadly, the paper argues that scientific acceleration should be evaluated not only by throughput, but by the quality, provenance, and corrigibility of the evidence being accelerated. The broader societal benefit is a scientific automation stack that accelerates discovery while preserving the conditions for correction.

References

Alkaissi, H. and McFarlane, S. I. Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus*, 15 (2):e35179, 2023. doi: 10.7759/cureus.35179.

Anthropic. Anthropic partners with U.S. national labs for first 1,000 scientist AI jam. `https://www.anthropic.com/news/anthropic-partners-with-u-s-national-labs-for-first-1-000-scientist-ai-jam`, 2025a. Anthropic News, accessed 2026.

Anthropic. Introducing Anthropic's AI for science program. `https://www.anthropic.com/news/ai-for-science-program`, 2025b. Anthropic News, accessed 2026.

Anthropic. Science. `https://www.anthropic.com/science`, 2026. Anthropic Science page, accessed 2026.

Bar-Ilan, J. and Halevi, G. Retracted papers remain in the scientific literature. *Scientometrics*, 126(7):6029–6045, 2021. doi: 10.1007/s11192-021-03967-2.

Begley, C. G. and Ellis, L. M. Drug development: Raise standards for preclinical cancer research. *Nature*, 483 (7391):531–533, 2012. doi: 10.1038/483531a.

Bender, E. M. and Friedman, B. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018. doi: 10.1162/tacl{_}a{_}00041.

Boiko, D. A., MacKnight, R., Kline, B., and Gomes, G. Autonomous chemical research with large language models. *Nature*, 624:570–578, 2023. doi: 10.1038/s41586-023-06792-0.

Bran, A. M., Cox, S., Schilter, O., Baldassari, C., White, A. D., and Schwaller, P. ChemCrow: Augmenting large-language models with chemistry tools. *Nature Machine Intelligence*, 6:525–535, 2024. doi: 10.1038/s42256-024-00832-8.

Byrne, J. A. The fight against fake-paper factories that churn out sham science. *Nature*, 603(7901):370–371, 2022. doi: 10.1038/d41586-022-00733-5.

Chambers, C. D. Registered reports: A new publishing initiative at Cortex. *Cortex*, 49(3):609–610, 2013. doi: 10.1016/j.cortex.2012.12.016.

Chan, J. S., Pieler, N., Jao, J., Peetoom, J., Mullen-Schultz, R., et al. MLE-bench: Evaluating machine learning agents on machine learning engineering, 2024.

Chen, J. et al. Are LLMs good reviewers? measuring human and GPT-4 agreement in peer review, 2024.

Curry, S., Mercado-Lara, E., Arechavala-Gomeza, V., Begley, C. G., Bernard, C., Bernard, R., Bertuzzi, S., Bhalla, N., Bowers, D., Brod, S., et al. Ending publication bias: A values-based approach to surface null and negative results. *PLOS Biology*, 23(9):e3003368, 2025. doi: 10.1371/journal.pbio.3003368.

Dickersin, K., Min, Y.-I., and Meinert, C. L. Factors influencing publication of research results: Follow-up of applications submitted to two institutional review boards. *JAMA*, 267(3):374–378, 1992. doi: 10.1001/jama.1992.03480030052036.

Duyx, B., Swaen, G. M. H., Urlings, M. J. E., Bouter, L. M., and Zeegers, M. P. The strong focus on positive results in abstracts may cause bias in systematic reviews: A case study on abstract reporting bias. *Systematic Reviews*, 8 (1):217, 2019. doi: 10.1186/s13643-019-1082-9.

Dwan, K., Altman, D. G., Arnaiz, J. A., Bloom, J., Chan, A.-W., Cronin, E., Decullier, E., Easterbrook, P. J., Von Elm, E., Gamble, C., Ghersi, D., Ioannidis, J. P. A., Simes, J., and Williamson, P. R. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLOS ONE*, 3(8):e3081, 2008. doi: 10.1371/journal.pone.0003081.

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E., and Nosek, B. A. Investigating the replicability of preclinical cancer biology. *eLife*, 10: e71601, 2021. doi: 10.7554/eLife.71601.

Esau, P., Shmatko, N., and Adam, A. GPTZero finds over 50 new hallucinations in ICLR 2026 submissions. `https://gptzero.me/news/iclr-2026/`, 2025. Accessed 2026.

Fanelli, D. "Positive" results increase down the hierarchy of the sciences. *PLOS ONE*, 5(4):e10068, 2010. doi: 10.1371/journal.pone.0010068.

Franco, A., Malhotra, N., and Simonovits, G. Publication bias in the social sciences: Unlocking the file drawer. *Science*, 345(6203):1502–1505, 2014. doi: 10.1126/science.1255484.

Freedman, L. P., Cockburn, I. M., and Simcoe, T. S. The economics of reproducibility in preclinical research. *PLOS Biology*, 13(6):e1002165, 2015. doi: 10.1371/journal.pbio.1002165.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., and Wang, H. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., and Crawford, K. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021. doi: 10.1145/3458723.

Gelman, A. and Loken, E. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no fishing expedition or p-hacking and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*, 2013.

Gottweis, J. et al. AI co-scientist: Accelerating scientific discovery with AI-powered research agents, 2025.

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., Brand, R., Brandt, M. J., Brewer, G., Bruyneel, S., et al. A multi-lab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, 11(4):546–573, 2016. doi: 10.1177/1745691616652873.

Holland, S., Hosny, A., Newman, S., Joseph, J., and Chmielinski, K. The dataset nutrition label: A framework to drive higher data quality standards. *arXiv preprint arXiv:1805.03677*, 2018.

Ibrahim, H., Liu, F., Zaki, Y., Rahwan, T., et al. Citation manipulation through citation mills and pre-print servers. *Scientific Reports*, 15:5480, 2025. doi: 10.1038/s41598-025-88709-7.

Ioannidis, J. P. A. Why most published research findings are false. *PLOS Medicine*, 2(8):e124, 2005. doi: 10.1371/journal.pmed.0020124.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023. doi: 10.1145/3571730.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Zidek, A., Potapenko, A., et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Li, J. et al. A survey on LLM-as-a-judge, 2024.

Liang, W., Zhang, Y., Wu, Z., Lepp, H., Ji, W., Zhao, X., Cao, H., Liu, S., He, S., Huang, Z., Yang, D., Potts, C., Manning, C. D., and Zou, J. Quantifying large language model usage in scientific papers. *Nature Human Behaviour*, 9:2599–2609, 2025. doi: 10.1038/s41562-025-02273-8.

Lin, S., Hilton, J., and Evans, O. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 3214–3252, 2022. doi: 10.18653/v1/2022.acl-long.229.

Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., and Ha, D. The AI scientist: Towards fully automated open-ended scientific discovery, 2024.

Maynez, J., Narayan, S., Bohnet, B., and McDonald, R. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1906–1919, 2020. doi: 10.18653/v1/2020.acl-main.173.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220–229, 2019. doi: 10.1145/3287560.3287596.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., et al. Promoting an open research culture. *Science*, 348(6242):1422–1425, 2015. doi: 10.1126/science.aab2374.

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015. doi: 10.1126/science.aac4716.

Peters, U. and Chin-Yee, B. Generalization bias in large language model summarization of scientific research. *Royal Society Open Science*, 12(4):241776, 2025. doi: 10.1098/rsos.241776.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., and Barnes, P. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 33–44, 2020. doi: 10.1145/3351095.3372873.

Retraction Watch. Retraction watch database. `https://retractionwatch.com/`, 2024. Accessed 2026.

Shmatko, N., Adam, A., and Esau, P. GPTZero finds 100 new hallucinations in NeurIPS 2025 accepted papers. `https://gptzero.me/news/neurips/`, 2026. Accessed 2026.

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., and Gal, Y. AI models collapse when trained on recursively generated data. *Nature*, 631:755–759, 2024. doi: 10.1038/s41586-024-07566-y.

Si, C., Yang, D., and Hashimoto, T. LLM-generated literature reviews: Identifying challenges and proposing solutions, 2024.

Sidhu, R. S., Selvamogan, A., and Boddy, A. M. Trust, truth and transparency: Analysing the references underpinning AI-generated surgical information. *Annals of The Royal College of Surgeons of England*, 2026. doi: 10.1308/rcsann.2026.0021.

Simmons, J. P., Nelson, L. D., and Simonsohn, U. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366, 2011. doi: 10.1177/0956797611417632.

Steen, R. G., Casadevall, A., and Fang, F. C. Retractions in the medical literature: How many patients are put at risk by flawed research? *Journal of Medical Ethics*, 39 (1):20–25, 2013. doi: 10.1136/medethics-2012-100766.

Sterne, J. A. C. and Egger, M. Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *Journal of Clinical Epidemiology*, 54(10):1046–1055, 2001. doi: 10.1016/S0895-4356(01)00377-8.

Swanson, K. et al. The virtual lab: AI agents design new SARS-CoV-2 nanobodies with experimental validation. *bioRxiv*, 2024. doi: 10.1101/2024.11.11.623004.

Thelwall, M., Lehtisaari, M., Katsirea, I., Holmberg, K., and Zheng, E.-T. Does ChatGPT ignore article retractions and other reliability concerns? *Learned Publishing*, 38 (4), 2025. doi: 10.1002/leap.2018.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. Self-consistency improves chain of thought reasoning in language models. *International Conference on Learning Representations*, 2023.

Yao, L., Gu, T., Li, X., Jiao, Y., Li, M., Li, Y., Graff, J. C., and Gu, W. AI misuse of retracted literature: A comparative study of ChatGPT4o, DeepSeek, and Grok 3 in stem cell research. *The Science of Nature*, 112:85, 2025. doi: 10.1007/s00114-025-02036-5.

Zarin, D. A., Tse, T., Williams, R. J., Califf, R. M., and Ide, N. C. The ClinicalTrials.gov results database: Update and key issues. *New England Journal of Medicine*, 364 (9):852–860, 2011. doi: 10.1056/NEJMsa1012065.

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*, volume 36, 2024.