SciReview: Diagnosing Compositional Scientific Reasoning in Frontier Models

Anonymous Authors[1]

Abstract

Frontier large language models achieve high performance on many scientific evaluations, yet it remains unclear whether such performance reflects compositional reasoning or reliance on the conditional distribution of expert-genre prose. We introduce SciReview, a benchmark of expert-authored research-grade passages with naturally injected, locally plausible but scientifically consequential errors across science, engineering, technology, and mathematics. The construction protocol explicitly excludes bare lookup errors, definition rewrites that preserve internal coherence, and unsupported assertions: a filter that, by design, leaves only errors whose detection requires re-deriving claims from local content rather than retrieving facts. We pair this with an adversarial difficulty calibration in which each task contains errors stratified by the agreement of three frontier models, yielding a per-item probe of where the memorization–generalization boundary sits. Evaluating GPT-5.4, Gemini 3.1 Pro, and Claude Opus 4.6 under five recall metrics crossed with two false-positive treatments, we find that rankings invert sharply once false-positive control is enforced (GPT-5.4 falls from 62% to 10% Average Recall). This suggests that high recall in the unpenalized regime can mask reliance on surface pattern-matching rather than disciplined re-derivation. We plan on releasing this benchmark upon acceptance.

1. Introduction

A central question for the foundations of deep generative models is what these models actually learn. For deep autoregressive language models trained on web-scale corpora, the question takes a sharp form: when a frontier model produces a fluent, technically dressed scientific assessment, is the underlying computation a re-derivation from first principles, or is it largely a high-fidelity reproduction of the rhetorical and inferential templates that pervade scientific writing on the public internet? The same question is well-studied in image diffusion models, where verbatim memorization of training images (Carlini et al., 2023b; Somepalli et al., 2023a) and convergence to a shared score function under data scaling (Kadkhodaie et al., 2024) are both documented. For autoregressive LLMs, the analogous question—what part of scientific competence is template recall and what part is compositional generalization—is harder to operationalize because (i) training corpora are typically not disclosed (Shi et al., 2024; Oren et al., 2024), (ii) public benchmarks are demonstrably contaminated (Sainz et al., 2023; Roberts et al., 2024; Yang et al., 2023), and (iii) accuracy on saturated benchmarks is consistent with both memorization and generalization.

A complementary diagnostic strategy, advanced for example by Wu et al. (2024), McCoy et al. (2024), Mirzadeh et al. (2025), and Dziri et al. (2023), is to design tasks whose *form* is in distribution but whose required *computation* is not. Counterfactual variants, symbolic perturbations, and controlled compositional graphs measurably degrade frontier-model performance, suggesting that current LLMs solve many compositional problems by reducing them to "linearized subgraph matching" rather than executing systematic problem-solving (Dziri et al., 2023).

We introduce SciReview, a benchmark designed in this spirit but targeted at a different surface of scientific competence: the ability to read realistic, expert-authored research prose and identify *embedded conceptual errors* that are locally plausible, scientifically consequential, and genre-typical. SciReview is organized around three theoretical commitments that we argue make it a clean probe of the template-vs-generalization distinction:

1. **Exclusion criteria as a memorization filter.** The protocol explicitly excludes (a) bare lookup errors with no downstream role, (b) altered definitions where downstream logic remains internally coherent, and (c) unsupported false assertions with no local justification. Errors that make the passage internally incoherent are also rejected. The first three exclusions remove failure

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

Preliminary work. Under review by the FoGen Workshop at ICML 2026. Do not distribute.

modes that a model with strong factual memorization could solve via retrieval; the fourth removes failures detectable by surface coherence checks. What remains are errors whose detection requires re-deriving a chain of inference from the local content.

2. **Locally-plausible-but-globally-wrong errors stress pattern-matching.** Errors are constructed to preserve internal coherence—for example, a chemistry passage in which both a face-assignment step and the resulting enantiomer label are wrong in a mutually consistent way. Models that rely on local-coherence heuristics rubber-stamp the chain; catching the error requires going against the conditional distribution of the surrounding prose.

3. **Adversarial difficulty calibration.** Each candidate task is evaluated against three frontier models. Errors caught by all three are designated *baseline* (within shared template space); errors missed by all three are designated *challenging* (beyond shared template space, requiring compositional re-derivation). This gives a per-problem signature of where the memorization–generalization boundary sits for the panel.

We evaluate three frontier autoregressive models: GPT-5.4, Gemini 3.1 Pro, and Claude Opus 4.6, under a five-metric × two-FP-treatment scoring framework plus qualitative axes. Our principal empirical finding is a sharp inversion of model rankings between the unpenalized and FP-penalized regimes: GPT-5.4 leads on unpenalized Average Recall (62%) but trails on the 0-FP-gated metric (10%), while Gemini 3.1 Pro retains 45% under penalization. This inversion illustrates a "memorization-vs-generalization" distinction: high unpenalized recall is consistent with broad pattern-flagging that may not survive a selectivity test.

We deliberately frame SCIREVIEW as *one* diagnostic angle. It does not measure verbatim training-data memorization in the sense of Carlini et al. (2023a) or Schwarzschild et al. (2024), nor does it separate memorization from generalization at the level of model weights. What it does measure is whether a model's behavior on a class of scientific reading tasks is consistent with re-derivation from local content, or is better explained by reliance on the conditional distribution of expert-genre prose.

2. Related Work

Memorization in large autoregressive models. Memorization has been characterized along several axes. Carlini et al. (2019; 2021; 2023a) and Nasr et al. (2023) operationalize *extractable* memorization: the verbatim regurgitation of training sequences under adversarial prompting, including from production-aligned models. Tirumala et al. (2022) show that larger models memorize more before overfitting and forget less, and Biderman et al. (2023) demonstrate that memorization is partially predictable from smaller-scale runs. Lee et al. (2022) and Kandpal et al. (2022) link memorization rates to training-data duplication. Zhang et al. (2023) introduce *counterfactual* memorization to disentangle rare-content recall from common templated text, and Schwarzschild et al. (2024) propose Adversarial Compression Ratio as a behavioral memorization metric. Hartmann et al. (2023) survey the field and propose a taxonomy that includes verbatim text, facts, ideas, writing styles, and distributional properties—a useful reminder that "memorization" is multi-faceted and that *style* memorization in particular is closely related to the expert-genre prose-pattern reliance SCIREVIEW probes. Magar and Schwartz (2022) and Zheng and Jiang (2022) study how training–test overlap translates to test-time exploitation, finding that exposure does not always translate to exploitation. SCIREVIEW does not measure memorization in the verbatim sense; it targets *template-level* memorization—the extent to which model behavior is shaped by the rhetorical and inferential structure of expert scientific writing—using error injection in genre-typical contexts as a stress test.

Memorization vs. generalization in deep generative models. The clearest empirical picture of the gen/mem transition has been established for image diffusion models. Carlini et al. (2023b) showed that diffusion models memorize and emit individual training images; Somepalli et al. (2023a;b) characterized data replication and copying. On the other side, Kadkhodaie et al. (2024) showed that as training-set size grows past a threshold, two diffusion models trained on disjoint subsets converge to nearly the same score function, indicating a phase transition driven by inductive biases (geometry-adaptive harmonic representations). Foundational work by Zhang et al. (2017; 2021) establishes that deep networks can memorize random labels yet generalize on natural data, motivating careful empirical separation of the two regimes. We emphasize that LLMs and diffusion models are both deep generative models but belong to distinct families; we do *not* claim that the diffusion-model gen/mem transition transfers directly. SCIREVIEW targets a complementary phenomenon in autoregressive text generation.

Generalization, extrapolation, and reasoning. A growing line of work probes the ceiling of LLM compositional reasoning. Razeghi et al. (2022) show that few-shot numerical reasoning correlates strongly with pretraining term frequency. McCoy et al. (2024) develop the "embers of autoregression" framework: LLMs are systematically more accurate in high-probability task and output regimes even when probability is irrelevant to correctness. Wu et al. (2024) introduce counterfactual task variants and observe consistent

and substantial degradation when default task assumptions are perturbed. Dziri et al. (2023) show that transformers solve compositional tasks (multi-digit multiplication, logic puzzles, dynamic programming) by linearized subgraph matching, with rapid decay as compositional depth grows. Press et al. (2023) measure a *compositionality gap* between sub-question accuracy and integrated multi-hop accuracy that does not close with scale. Mirzadeh et al. (2025) introduce GSM-Symbolic and observe sharp drops on numerical-only perturbations of GSM8K, plus large drops from inserting irrelevant clauses. Hupkes et al. (2020) provide a foundational framework for compositional generalization. SCIREVIEW's locally-plausible-but-globally-wrong errors share methodological DNA with these probes: they create stimuli where the surface form is in-distribution but the correct response requires going against it.

Benchmarks targeting the gen/mem distinction. Several recent benchmarks try to neutralize contamination as a confound. GSM-Symbolic (Mirzadeh et al., 2025) generates instances from symbolic templates. DyVal (Zhu et al., 2024) dynamically synthesizes graph-structured reasoning tasks. LiveCodeBench (Jain et al., 2024) and LiveBench (White et al., 2024) continuously update their problem sets with post-cutoff items. FrontierMath (Glazer et al., 2024) uses unpublished expert-crafted mathematics problems and reports very low accuracy from leading models at release, contrasting with the near-saturation of MATH (Hendrycks et al., 2021). Humanity's Last Exam (Phan et al., 2025) collects expert-vetted Google-proof questions across domains. GPQA (Rein et al., 2024) targets graduate-level Google-proof biology, physics, and chemistry. BIG-Bench Hard (Suzgun et al., 2023) isolates the hardest tasks of BIG-Bench (Srivastava et al., 2023). SCIREVIEW differs from these in two ways. First, the unit of evaluation is a research-grade *passage with embedded errors*, not a question with an answer; the task structure is expert reading, not problem solving. Second, the targeted competence is not problem-solving accuracy but *error detection under selectivity pressure*, which is a distinct diagnostic surface for the gen/mem question.

Data and benchmark contamination. Contamination work motivates expert-authored, private benchmarks. Sainz et al. (2023) document widespread contamination and propose per-benchmark measurement. Golchin and Surdeanu (2023) propose "time-travel" detection. Oren et al. (2024) provide statistical guarantees for black-box contamination detection via exchangeability. Roberts et al. (2024) use training-cutoff natural experiments to show pre-cutoff vs. post-cutoff performance gaps on competitive coding problems. Yang et al. (2023) show that simple paraphrasing defeats string-matching decontamination. Magar and Schwartz (2022) distinguish memorization from exploitation in controlled training. SCIREVIEW's expert-authored passages are written de novo for the benchmark and are not on the public web, mitigating verbatim contamination at construction time; the FP-penalized scoring further reduces sensitivity to template-shaped contamination.

Robustness, privacy, and extrapolation benchmarks. HELM (Liang et al., 2023) measures robustness, calibration, fairness, and toxicity alongside accuracy. Privacy-extraction benchmarks descend from Carlini et al. (2019; 2021; 2023a) and Nasr et al. (2023). The recent consensus is that benchmark saturation often reflects template fit rather than capability gain (Mirzadeh et al., 2025; McCoy et al., 2024). SCIREVIEW contributes a *selectivity-sensitive* extrapolation probe by reporting recall under both unpenalized and FP-penalized scoring; the gap between the two is itself a diagnostic.

Scientific-reasoning and AI-for-science benchmarks. Adjacent benchmarks probe AI-for-science capabilities along complementary axes: SciCode (Tian et al., 2024), MLE-bench (Chan et al., 2024), PaperBench (Starace et al., 2025), DiscoveryBench (Majumder et al., 2024), GPQA (Rein et al., 2024), MATH (Hendrycks et al., 2021), SPOT (Son et al., 2025), and the broader literature on LLM-as-judge (Zheng et al., 2023). These largely target *task performance*. SCIREVIEW targets *critical reading under conditions designed to dissociate template matching from compositional re-derivation*, with a scoring framework that explicitly rewards selectivity. The reproducibility-crisis literature (Begley and Ellis, 2012; Baker, 2016; Errington et al., 2021) motivates the practical importance of error detection in scientific writing, but SCIREVIEW should be evaluated as a foundations-of-generative-models diagnostic rather than as an AI-for-science deployment benchmark.

3. A Diagnostic Frame for Memorization vs. Generalization

We frame the diagnostic question as follows. Let $p_\theta(y \mid x)$ be a frontier autoregressive model. For a passage x containing a set of injected errors $E = \{e_1, \ldots, e_k\}$, the model produces a critique y. Define two stylized strategies a model might use:

- **Template-pattern matching.** The critique is shaped primarily by the conditional distribution of expert-genre prose given the passage's surface form. Under this strategy, claims that are grammatically and rhetorically expected—e.g., a "Clinical Implications" subsection asserting individual-level inference from a group-level finding—are passed through, because the conditional distribution of such prose makes the assertion locally probable. Errors that disrupt surface

coherence are flagged; errors that respect surface coherence are not.

- **Compositional re-derivation.** The critique is shaped by re-deriving claims from local content using a domain world model. Under this strategy, errors are flagged regardless of whether they respect surface coherence, provided the underlying derivation contradicts the model's inferred chain.

These strategies are not mutually exclusive, any realistic model interpolates them, but a benchmark that systematically demands re-derivation, by construction, separates models that lean more on one strategy from models that lean more on the other.

Why exclusion criteria matter. SCIREVIEW's exclusions remove the failure modes where the two strategies converge to the same prediction. If an error were a bare lookup error (e.g., a wrong constant, untouched downstream), a memorization-rich model would catch it via retrieval; if an error left downstream logic internally consistent even after a definitional change, then a re-derivation model would *not* flag it (the local logic remains coherent under the rewritten definition). The protocol therefore concentrates the error class on cases where compositional re-derivation against the local prose is required.

Why locally-plausible-but-globally-wrong matters. Errors are constructed so that the sentences immediately adjacent to the error remain coherent. In the chemistry example (Section 7), both the face-assignment step and the final enantiomer label are wrong in a self-consistent way. A coherence-checking strategy returns no signal; a re-derivation strategy must regenerate the assignment from the molecular structure to identify the inconsistency.

Adversarial filter as a panel-relative probe. The 3-model agreement filter does not certify that an error is "memorized" or "generalized" by any single model. What it provides is a calibrated empirical proxy for where the panel's shared template space ends. Errors caught by all three models lie within that shared space; errors missed by all three lie outside it. This is a panel-relative quantity and can shift under broader evaluation; we treat the resulting per-item difficulty distribution as a diagnostic result, not categorical ground truth.

4. Benchmark Design

Domains. SCIREVIEW spans Science (biology, chemistry, neuroscience, physics), Engineering (aerospace, biomedical, civil, chemical, electrical, mechanical, industrial), Technology (data science, IT), and Mathematics (pure, applied, statistics). Tasks are authored and graded within a single discipline; mixed-discipline tasks are excluded because expert standards for rigor differ across fields.

Table 1. Excluded error patterns and the diagnostic rationale for excluding each. The first three classes would be solved by retrieval or coherence checks; the last is trivially detectable.

Excluded class	Why excluded
Bare lookup errors with no downstream role	Solvable by factual retrieval; conflates memorization with detection
Altered definitions with internally coherent downstream logic	Re-derivation gives no signal under the rewritten definition
Unsupported false assertions with no local justification	Detectable by surface-coherence checks
Errors that render the passage internally incoherent	Trivially detectable; does not test the targeted competence

Seed texts. Each task is built around an expert-authored, research-grade passage (∼150–800 words): a proposal excerpt, methods section, or technical note. Passages are written de novo by domain experts for the benchmark and are not drawn from the public literature, mitigating contamination at construction time.

Five construction principles.

1. *Expert authorship.* Every passage and every error is produced by a working researcher in the relevant subfield.
2. *Naturalness.* Errors must be natural enough that a non-specialist working researcher could miss them in a quick read.
3. *Locally-plausible-but-globally-wrong.* Errors must preserve local coherence with adjacent claims.
4. *Scientific consequence.* Errors must materially affect a downstream conclusion (a numerical value, a causal claim, an experimental design choice).
5. *Compositional re-derivation.* Detection must require composing at least two reasoning steps, not a single fact lookup.

Exclusion criteria. Table 1 summarises the excluded error patterns, each motivated by the diagnostic frame of Section 3.

Adversarial difficulty calibration. Each candidate item is run against three frontier models. An error caught by all three is *baseline*; an error missed by all three is *challenging*. Each accepted task contains at least three challenging errors and at most three baseline errors.

Quality control. Each candidate task is reviewed and revised by an independent expert who rates it on an ordinal scale; only items rated "Good" enter the evaluation set. Common rejected failure modes were (i) errors that were technically incorrect but did not propagate (lookup errors), (ii) errors indistinguishable from authorial imprecision, and (iii) errors that made the passage internally incoherent.

5. Evaluation Methodology

Models. We evaluate three frontier autoregressive models: GPT-5.4 (OpenAI), Gemini 3.1 Pro (Google DeepMind), and Claude Opus 4.6 (Anthropic). Each passage is evaluated four times per model with default sampling temperature.

Matching. Because model responses are free-form critiques, we use an LLM-based matcher (in the spirit of LLM-as-judge, Zheng et al., 2023) to determine which expert-annotated errors each response identified and how many additional non-errors it flagged. Matcher reliability was validated on a stratified subsample against domain-expert re-verification; future updates will also add systematic human auditing.

Scoring framework. Let $M(r_i)$ denote the gold errors matched by response r_i on item i, and $\mathrm{FP}(r_i)$ the number of flagged issues not matching any gold error. We report three recall-style metrics:

$$\begin{aligned}\text{AR (Average Recall)} &= \mathrm{mean}_i \frac{|M(r_i)|}{|E_i|},\\ \text{PR (Perfect Recovery)} &= \Pr\big(|M(r_i)| = |E_i|\big),\\ \text{MR (Majority Recovery)} &= \Pr\big(|M(r_i)| > \tfrac{1}{2}|E_i|\big).\end{aligned}$$

Each is crossed with two false-positive treatments: *uncorrected* (ignore FPs) and *penalized-disqualifying* (PD: a run counts as failure if $\mathrm{FP}(r_i) > 0$). An exact-set indicator combines both:

$$\mathrm{Exact}(r_i) = \mathbb{1}[|M(r_i)| = |E_i| \wedge \mathrm{FP}(r_i) = 0]. \quad (1)$$

A penalized-canceling (PC) variant, in which FPs cancel TPs one-for-one floored at zero, is reported in Section A.

Why FP-aware scoring is the central diagnostic. A model that achieves high recall by flagging many candidate issues is, with respect to the gen/mem question, ambiguous: high recall is consistent with both broad pattern-matching and disciplined re-derivation. The 0-FP-gated metric removes this ambiguity by rewarding only runs whose flagged set is exactly correct. The gap between AR and AR-under-0-FP-gate is therefore the central diagnostic quantity in our results.

Table 2. Quantitative results on SCIREVIEW. "Standard" ignores false positives; "0-FP gate" disqualifies any run containing a misidentified error. All numbers in percent; higher is better.

	Standard			0-FP gate		
Model	AR	PR	MR	AR	Exact	$>\frac{1}{2}$
GPT-5.4	62	6	71	10	0	12
Gemini 3.1 Pro	57	0	60	45	0	48
Claude Opus 4.6	46	0	33	24	0	19

Table 3. Qualitative assessment. Helpfulness, correctness, and alignment are scored on $\{0, 1, 2\}$ (0 = no issues, 2 = major issues); lower is better. Overall is a 1–5 Likert scale; higher is better.

Model	Help. ↓	Corr. ↓	Align. ↓	Overall ↑
GPT-5.4	0.60	1.30	0.25	2.65
Gemini 3.1 Pro	0.40	1.30	0.05	3.10
Claude Opus 4.6	0.55	1.55	0.35	2.10

Qualitative axes. In parallel, each response is rated by a domain-expert reviewer along three ordinal axes—*helpfulness*, *correctness*, and *alignment*—each on the scale $\{0, 1, 2\}$ mapped from {no issues, minor issues, major issues}. Reviewers additionally assign an Overall rating on a 1–5 Likert scale. We view these as auxiliary diagnostics rather than headline metrics.

6. Results

Table 2 reports the quantitative results across all scoring regimes; Table 3 reports the qualitative axes; Table 4 reports the per-item difficulty distribution on the Good-rated subset.

Three principal findings: (1) Partial criticism is common; full review is rare. Raw recall is meaningfully above zero for all models, which is desirable for a benchmark intended to measure a useful capability rather than impossible puzzle-solving. At the same time, exact-set match under the 0-FP gate is 0% for every model, and standard Perfect Recovery is 6% for the best model and 0% for the others. This is consistent with Son et al. (2025), who find that no frontier model reliably catches all material errors in a scientific document.

(2) Rankings invert under false-positive penalization. GPT-5.4 leads on every uncorrected metric, but once misidentifications disqualify a run, Gemini 3.1 Pro becomes the strongest system: its AR drops only 12 percentage points (from 57% to 45%), compared with GPT-5.4's collapse of 52 points (62% to 10%). The gap reflects a qualitative behavioral difference: GPT-5.4 tends to flag more candidate errors, which increases recall when false positives are free but collapses under even a modest precision requirement. We frame this as *the central diagnostic finding*: the FP-

Table 4. Per-item difficulty distribution on the Good-rated subset. Each row gives the number of items containing the indicated number of challenging or baseline errors.

	Errors per item				
Class	0	1	2	4	5
Challenging	5	4	3	—	—
Baseline	1	5	4	1	1

penalized metric is the one that better tracks selectivity, and selectivity, in turn, is the behavioral signature most aligned with the compositional re-derivation strategy of Section 3.

(3) Per-item difficulty is non-degenerate. Table 4 shows that most items contain at least one challenging error and most contain at least one baseline error, validating that the calibration produces diverse per-item gen/mem signatures rather than uniformly easy or uniformly hard items. Notably, difficulty estimated during limited single-shot author-side pre-screening can shift under broader repeated evaluation: some errors classified as "challenging" during single-shot authoring were found by at least one model across multiple independent runs. This argues for calibrating final difficulty labels against multi-sample evaluation rather than relying only on one-shot checks during construction.

7. Diagnostic Analysis: What the Errors Reveal

We highlight three representative error types, each constructed under the exclusion protocol of Section 4; each requires multi-step re-derivation to detect.

Stereochemistry: face assignment composed with absolute configuration. A passage on enantioselective synthesis of β-amino acids (precursors for bioactive peptides) describes a catalyst–substrate interaction and reports the resulting absolute configuration. The passage contains two errors: it places the catalyst on the *endo* face of a bicyclic substrate (the correct face is *exo*) and reports the product as the $(1S, 6R)$ enantiomer (the correct stereochemistry is $(1R, 6S)$). The two errors point the same "wrong" direction, so the passage remains internally consistent. A coherence-checking strategy returns no signal; detection requires re-deriving the three-dimensional geometry of the reaction and applying CIP priority rules to the product, then composing the two steps to identify the inconsistency. We frame this as a probe of compositional re-derivation, not as a memorization test.

Neuroscience: group-to-individual scope shift. A neuroimaging study on brain plasticity in patients with painful diabetic peripheral neuropathy reports a group-level association in its Results, then asserts in a "Clinical Implications" subsection that the findings can be deployed as validated individual diagnostic tools. The error preserves all surface coherence, because grant-style "Clinical Implications" prose is grammatically and rhetorically primed to make exactly such forward-looking claims; the group-to-individual inferential leap is a long-standing methodological hazard (Poldrack et al., 2017). Detection requires going against the conditional distribution of the surrounding genre.

Mathematical physics: extrapolating to extremality. A passage on black-hole dynamics asserts that Price's law of late-time decay extends to the full Reissner–Nordström family. The claim silently fails at extremality, where vanishing surface gravity removes the red-shift effect that stabilises perturbations in the sub-extremal regime; in this regime, transverse derivatives of a linear scalar field fail to decay along the horizon—the Aretakis instability (Aretakis, 2015). Adjacent prose remains coherent because the algebraic manipulation is performed correctly given the (incorrect) extrapolation. Detection requires tracking surface gravity through the limit and identifying that a fundamental parameter goes to zero.

Across these cases, a pattern emerges: frontier models fail most consistently on errors that require (a) tracking a latent physical or mathematical parameter through a chain of reasoning, (b) distinguishing levels of analysis, or (c) noticing when plausible-sounding domain rhetoric is logically disconnected from the evidence. We emphasise that calling these tasks "compositional re-derivation tasks" is a description of the construction protocol, not a claim that any specific model failure on any specific item is *caused* by memorization. The correct empirical claim is that the aggregate behaviour of models on this class of tasks differentiates strategies that pay attention to the local-prose conditional distribution from strategies that re-derive from local content.

8. Discussion

The empirical inversion of model rankings between unpenalized AR and 0-FP-gated AR (Table 2) is the key diagnostic result. It shows that "ability to flag scientific errors" is not a single scalar capability and more importantly, that the ranking under the more selectivity-sensitive metric is the one that better tracks the kind of careful, re-derivation-grounded behaviour expected from a non-template-driven reviewer.

We resist the strong claim that this inversion *proves* GPT-5.4 is "more memorizing" or Gemini 3.1 Pro is "more generalizing." A cleaner statement is as follows: under the construction protocol of Section 4, the FP-penalized metric is a more demanding measure of re-derivation-grounded behaviour, because indiscriminate flagging is heavily penalized, and the model that wins under that metric is the one whose flags align more closely with the gold compositional

errors.

The wider implication for foundations of deep generative models is that benchmark design strongly couples with the capabilities one is measuring. Single-metric leaderboards conflate strategies that should be distinguished. SCIREVIEW's two-axis scoring (recall $\times$ FP treatment) is one concrete instance of a more general design principle: a benchmark intended to probe the gen/mem boundary should report metrics whose ranking can dissociate strategies, not just metrics whose magnitudes track aggregate performance.

9. Conclusion

SCIREVIEW is a diagnostic benchmark designed to dissociate template-pattern matching from compositional scientific reasoning in large autoregressive generative models. Its construction protocol: expert authorship, exclusion of trivial-lookup and definition-rewrite errors, locally-plausible-but-globally-wrong error injection, adversarial difficulty calibration concentrates the evaluation on cases where compositional re-derivation against the local conditional distribution of expert-genre prose is required. Across three frontier models, model rankings invert sharply between unpenalized recall and 0-FP-gated recall (GPT-5.4 62%$\rightarrow$10%; Gemini 3.1 Pro 57%$\rightarrow$45%; Claude Opus 4.6 46%$\rightarrow$24%), revealing that high recall is consistent with strategies that the more selective metric does not reward. We frame SCIREVIEW as one diagnostic angle on the memorization–generalization question: specifically, the distinction between template-prose pattern matching and compositional re-derivation, and contribute it to the broader effort, exemplified by GSM-Symbolic, counterfactual benchmarks, FrontierMath, LiveBench, and contamination-aware evaluation, to design benchmarks whose metric structure can dissociate strategies rather than only rank magnitudes. We plan on releasing this benchmark upon acceptance to make it easier to evaluate frontier models on this axis of scientific review capability.

Broader Impact

This paper aims to advance the foundations of deep generative models by providing a more discriminating evaluation framework for autoregressive LLMs. Designed primarily as a diagnostic, the benchmark is intended to sharpen the empirical separation between memorization-aligned and generalization-aligned behaviour. We explicitly recommend that frontier models not be deployed as unsupervised research auditors on the basis of SCIREVIEW performance alone.

References

Aretakis, S. *Dynamics of Extremal Black Holes*. SpringerBriefs in Mathematical Physics, Springer, 2015.

Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature*, 533:452–454, 2016.

Begley, C. G. and Ellis, L. M. Raise standards for preclinical cancer research. *Nature*, 483(7391):531–533, 2012.

Biderman, S., Prashanth, U. S. N., Sutawika, L., Schoelkopf, H., Anthony, Q., Purohit, S., and Raff, E. Emergent and predictable memorization in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., and Song, D. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *USENIX Security Symposium*, 2019.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., and Raffel, C. Extracting training data from large language models. In *USENIX Security Symposium*, 2021.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramèr, F., and Zhang, C. Quantifying memorization across neural language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., and Wallace, E. Extracting training data from diffusion models. In *USENIX Security Symposium*, 2023.

Chan, J. S., Chowdhury, N., Jaffe, O., Aung, J., Sherburn, D., Mays, E., Starace, G., Liu, K., Maksin, L., Patwardhan, T., Weng, L., and Mądry, A. MLE-bench: Evaluating machine learning agents on machine learning engineering. *arXiv preprint arXiv:2410.07095*, 2024.

Zheng, X. and Jiang, J. An empirical study of memorization in NLP. In *Proceedings of ACL*, pp. 6265–6278, 2022.

Dziri, N., Lu, X., Sclar, M., Li, X. L., Jiang, L., Lin, B. Y., West, P., Bhagavatula, C., Le Bras, R., Hwang, J. D., Sanyal, S., Welleck, S., Ren, X., Ettinger, A., Harchaoui, Z., and Choi, Y. Faith and fate: Limits of transformers on compositionality. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

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.

Glazer, E., Erdil, E., Besiroglu, T., Chicharro, D., Chen, E., Gunning, A., Falkman Olsson, C., Denain, J.-S., Ho, A., de Oliveira Santos, E., et al. FrontierMath: A benchmark for evaluating advanced mathematical reasoning in AI. *arXiv preprint arXiv:2411.04872*, 2024.

Golchin, S. and Surdeanu, M. Time travel in LLMs: Tracing data contamination in large language models. *arXiv preprint arXiv:2308.08493*, 2023.

Hartmann, V., Suri, A., Bindschaedler, V., Evans, D., Tople, S., and West, R. SoK: Memorization in general-purpose large language models. *arXiv preprint arXiv:2310.18362*, 2023.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the MATH dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Hupkes, D., Dankers, V., Mul, M., and Bruni, E. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67:757–795, 2020.

Jain, N., Han, K., Gu, A., Li, W.-D., Yan, F., Zhang, T., Wang, S., Solar-Lezama, A., Sen, K., and Stoica, I. LiveCodeBench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*, 2024.

Kadkhodaie, Z., Guth, F., Simoncelli, E. P., and Mallat, S. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *International Conference on Learning Representations (ICLR)*, 2024.

Kandpal, N., Wallace, E., and Raffel, C. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning (ICML)*, 2022.

Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., and Carlini, N. Deduplicating training data makes language models better. In *Proceedings of ACL*, 2022.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023.

Magar, I. and Schwartz, R. Data contamination: From memorization to exploitation. In *Proceedings of ACL (Short Papers)*, 2022.

Majumder, B. P., Surana, H., Agarwal, D., Dalvi Mishra, B., Meena, A., Prakhar, A., Vora, T., Khot, T., Sabharwal, A., and Clark, P. DiscoveryBench: Towards data-driven discovery with large language models. *arXiv preprint arXiv:2407.01725*, 2024.

McCoy, R. T., Yao, S., Friedman, D., Hardy, M. D., and Griffiths, T. L. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences (PNAS)*, 2024.

Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., and Farajtabar, M. GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Wallace, E., Tramèr, F., and Lee, K. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*, 2023.

Oren, Y., Meister, N., Chatterji, N., Ladhak, F., and Hashimoto, T. B. Proving test set contamination in black-box language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Phan, L. et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., Nichols, T. E., Poline, J.-B., Vul, E., and Yarkoni, T. Scanning the horizon: Towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, 18:115–126, 2017.

Press, O., Zhang, M., Min, S., Schmidt, L., Smith, N. A., and Lewis, M. Measuring and narrowing the compositionality gap in language models. In *Findings of EMNLP*, 2023.

Razeghi, Y., Logan IV, R. L., Gardner, M., and Singh, S. Impact of pretraining term frequencies on few-shot reasoning. In *Findings of EMNLP*, 2022.

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., and Bowman, S. R. GPQA: A graduate-level google-proof Q&A benchmark. In *Conference on Language Modeling (COLM)*, 2024.

Roberts, M., Thakur, H., Herlihy, C., White, C., and Dooley, S. Data contamination through the lens of time. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Sainz, O., Campos, J. A., García-Ferrero, I., Etxaniz, J., Lopez de Lacalle, O., and Agirre, E. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of EMNLP*, 2023.

Schwarzschild, A., Feng, Z., Maini, P., Lipton, Z. C., and Kolter, J. Z. Rethinking LLM memorization through the lens of adversarial compression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Shi, W., Ajith, A., Xia, M., Huang, Y., Liu, D., Blevins, T., Chen, D., and Zettlemoyer, L. Detecting pretraining data from large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Diffusion art or digital forgery? Investigating data replication in diffusion models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Understanding and mitigating copying in diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Son, G., Hong, J., Fan, H., Nam, H., et al. When AI co-scientists fail: SPOT—a benchmark for automated verification of scientific research. *arXiv preprint arXiv:2505.11855*, 2025.

Srivastava, A. et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023.

Starace, G., Jaffe, O., Sherburn, D., Aung, J., Chan, J. S., Maksin, L., Dias, R., Mays, E., Kinsella, B., Thompson, W., Heidecke, J., Glaese, A., and Patwardhan, T. PaperBench: Evaluating AI's ability to replicate AI research. In *International Conference on Machine Learning (ICML)*, 2025.

Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q. V., Chi, E. H., Zhou, D., and Wei, J. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. In *Findings of ACL*, 2023.

Tian, M., Gao, L., Zhang, S. D., Chen, X., Fan, C., et al. SciCode: A research coding benchmark curated by scientists. *arXiv preprint arXiv:2407.13168*, 2024.

Tirumala, K., Markosyan, A. H., Zettlemoyer, L., and Aghajanyan, A. Memorization without overfitting: Analyzing the training dynamics of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

White, C., Dooley, S., Roberts, M., Pal, A., Feuer, B., Jain, S., Shwartz-Ziv, R., Jain, N., Saifullah, K., Naidu, S., Hegde, C., LeCun, Y., Goldstein, T., Neiswanger, W., and Goldblum, M. LiveBench: A challenging, contamination-limited LLM benchmark. *arXiv preprint arXiv:2406.19314*, 2024.

Wu, Z., Qiu, L., Ross, A., Akyürek, E., Chen, B., Wang, B., Kim, N., Andreas, J., and Kim, Y. Reasoning or reciting? Exploring the capabilities and limitations of language models through counterfactual tasks. In *Proceedings of NAACL*, 2024.

Yang, S., Chiang, W.-L., Zheng, L., Gonzalez, J. E., and Stoica, I. Rethinking benchmark and contamination for language models with rephrased samples. *arXiv preprint arXiv:2311.04850*, 2023.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Zhang, C., Ippolito, D., Lee, K., Jagielski, M., Tramèr, F., and Carlini, N. Counterfactual memorization in neural language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.

Zhu, K., Chen, J., Wang, J., Gong, N. Z., Yang, D., and Xie, X. DyVal: Dynamic evaluation of large language models for reasoning tasks. In *International Conference on Learning Representations (ICLR)*, 2024.

A. Penalized-Canceling (PC) Scoring

For completeness, we report a third false-positive treatment: *penalized-canceling* (PC), where the count of correctly identified errors is reduced by the number of misidentified errors, floored at zero.

Table 5. Penalized-canceling results. Higher is better. All numbers in percent.

Metric	GPT-5.4	Gemini 3.1 Pro	Claude Opus 4.6
AR–PC	29	51	33
PR–PC	0	0	0
MR–PC	17	52	19

The PC ranking agrees qualitatively with the PD (0-FP-gate) ranking (Gemini 3.1 Pro best, Claude Opus 4.6 worst), but the semantics of "subtracting" one error count from another are ambiguous: a run with four correct identifications and four misidentifications is not obviously equivalent to a run with zero of each. We therefore emphasize PD in the main text and include PC for readers interested in rank-stability.

Tolerance variant. One additional variant we considered was allowing up to two misidentified errors before penalization. This was not adopted because it conflates error-identification quality with writing-quality feedback: model-flagged "errors" that are not expert-listed may still surface genuine ambiguities in the author's writing. We plan to address writing-quality feedback as a separate benchmark axis.

B. Full Example: Mathematical Physics

Passage excerpt. "Let us now turn the charge on and consider the full Reissner–Nordström (RN) family which admits a black hole (we take $Q > 0$ and $a = 0$). To further simplify the discussion, we will consider from now onwards that the initial data on $\mathcal{N}_1$ are in fact compactly supported (and hence trivially conformal). Aretakis and collaborators rigorously showed in 2016 that Price's law extends to these spacetimes."

Expert-annotated error. The statement that Price's law applies to the full RN family is wrong because this family contains the extremal case ($|Q| = M$, surface gravity $\kappa = 0$), for which the decay behaviour along the event horizon differs qualitatively. Price's law in the form stated applies to sub-extremal RN; in the extremal case, transverse derivatives of a linear scalar field fail to decay along the horizon (the Aretakis instability; see Aretakis (2015)).

Why the error is subtle. In the sub-extremal regime, positive surface gravity produces a red-shift effect at the horizon that stabilises perturbations. A reader who does not track the surface-gravity parameter through the phrase "full RN family" can miss that the claim extrapolates sub-extremal intuition into the extremal regime, where the stabilising mechanism disappears.

C. Full Example: Chemistry

Passage context. The passage discusses synthetic methods for accessing enantiomerically enriched β-amino acids used in bioactive peptide synthesis, where metabolic stability, potency, and safety profiles are highly enantiomer-selective.

Errors.

1. "…catalyst resided on the **endo** face of the bicyclic substrate, despite the cyclohexane portion being oriented toward the sterically cumbersome carbocyclic portion of the succinimide." Correct: the catalyst resides on the **exo** face.
2. "…delivering methyl $(\mathbf{1S}, \mathbf{6R})$-6-((methoxycarbonyl)amino)cyclohex-3-ene-1-carboxylate." Correct: the product is the $(1R, 6S)$ enantiomer.

Why the errors are subtle. The face assignment and the absolute configuration together determine the enantiomer of the final product; both errors point the same "wrong" direction, so the passage remains internally consistent and the downstream chemistry description is self-coherent. Catching this requires tracking the three-dimensional geometry of the reaction and re-deriving stereochemical labels rather than reading them as asserted.

D. Full Example: Neuroscience

Passage context. A study on brain plasticity in patients with painful diabetic peripheral neuropathy (PDN) reports group-level neuroimaging differences between PDN patients and controls.

Error. "These findings highlight the importance of deploying neuroimaging biomarkers as validated individual diagnostic tools in PDN diagnosis and monitoring."

Why the error is subtle. The phrase appears in a "Clinical Implications" subsection, where speculative, forward-looking claims are grammatically and stylistically expected. In grant applications, abstracts, and promotional writing, such claims are common and often tolerated; models pretrained on this genre may be biased to read them as part of the normal register rather than as logical overreach. The group-to-individual inferential leap is a long-standing methodological hazard in neuroimaging (Poldrack et al., 2017) with direct patient-safety consequences if accepted at face value.