

V1.0.0

Cross-Model LLM Convergence as Bayesian Evidence for Scientific Claim Correctness: A Formal Framework, Worked Derivations, and a Calibration Protocol

Rowan Brad Quni-Gudzin

QNFO-ULA-2.1

ABSTRACT

Verifying novel or fringe scientific claims is bottlenecked by the scarcity of domain experts. We propose and formalize a gate-checking method in which agreement between two or more independently trained large language models (LLMs) — not merely on a verdict but on the same multi-step argument structure — is treated as Bayesian evidence for claim correctness, provided shared training-corpus bias is bounded. We define a hypothesis space separating truth-tracking correctness (H_c), shared-corpus bias (H_b), and coincidence, and derive per-question-type Bayes factors: under the shared-bias null, convergence is expected on retrieval-only questions but not on multi-step synthesis questions, so the diagnostic value of convergence depends on question type. With explicitly labeled illustrative assumptions (prior probability of correctness 0.1; synthesis-arm likelihoods $p_c^S = 0.4$ versus $p_b^S = 0.05$), two convergent models yield a Bayes factor of 8 and a posterior of $8/17 \approx 0.471$; four convergent models yield a posterior of approximately 0.983. [...]

CITE AS Quni-Gudzin, R. B. (n.d.). Cross-Model LLM Convergence as Bayesian Evidence for Scientific Claim Correctness: A Formal Framework, Worked Derivations, and a Calibration Protocol. QNFO Preprint v1.0.0.

Abstract

Verifying novel or fringe scientific claims is bottlenecked by the scarcity of domain experts. We propose and formalize a gate-checking method in which agreement between two or more independently trained large language models (LLMs) — not merely on a verdict but on the same multi-step argument structure — is treated as Bayesian evidence for claim correctness, provided shared training-corpus bias is bounded. We define a hypothesis space separating truth-tracking correctness (H_c), shared-corpus bias (H_b), and coincidence, and derive per-question-type Bayes factors: under the shared-bias null, convergence is expected on retrieval-only questions but not on multi-step synthesis questions, so the diagnostic value of convergence depends on question type. With explicitly labeled illustrative assumptions (prior probability of correctness 0.1; synthesis-arm likelihoods $p_c^S = 0.4$ versus $p_b^S = 0.05$), two convergent models yield a Bayes factor of 8 and a posterior of $8/17 \approx 0.471$; four convergent models yield a posterior of approximately 0.983. We show that the same formalism assigns evidence *against* correctness when convergence occurs on engineered trap claims ($K_T = 1/90 \approx 0.011$, posterior ≈ 0.0012), and we report sensitivity analyses showing the two-model posterior ranges from ≈ 0.690 to ≈ 0.182 as the null's convergence probability varies from 0.02 to 0.20. We specify a four-arm benchmark protocol, situate the proposal against twelve works on claim verification, Bayesian evidence, and correctness in scientific computing, and state the assumptions whose failure would falsify the framework. No empirical measurements are reported; all numbers are computed from stated assumptions or labeled projections.

1. Introduction

Scientific claim verification requires in-depth knowledge and great labor from domain experts to substantiate supporting and refuting evidence from credible sources, and the growth of digital information outlets has made scientific misinformation more prevalent [3]. Retrieval-augmented systems such as CIBER have been built to identify

corroborating and refuting documents for claim verification, addressing the inherent uncertainty in LLMs by evaluating response consistency across diverse interrogation probes [4]. These systems anchor on retrieved documents; they do not ask whether an LLM's *reasoning* is independently reproducible by a differently trained system.

This paper develops a different signal. Suppose two LLMs, trained independently — different architectures, different data mixtures, different organizations — are each asked to evaluate a fringe physics claim, and both not only reach the same verdict but reproduce the same multi-step argument structure (the same chain of intermediate propositions in the same order, modulo paraphrase). Intuitively, two independent instruments agreeing is stronger evidence than one instrument's report. We formalize this as Bayesian model comparison.

The key structural insight is that the shared-bias null hypothesis makes *different predictions depending on question type*. For retrieval-only questions (answerable by recalling a fact present in the shared corpus), the null predicts convergence. For multi-step synthesis questions (requiring construction of a novel reasoning chain), the null predicts convergence only if the shared corpus contains the full argument — which is precisely what distinguishes a genuinely fringe claim from a merely obscure one. Convergence on multi-step synthesis arguments therefore discriminates between correctness and shared bias in a way that convergence on retrieval answers does not.

Our contributions are: (i) an explicit Bayesian model with a three-way hypothesis space, per-question-type likelihoods, and a closed-form N -model cascade; (ii) fully worked numerical derivations with all arithmetic shown, under two documented parameterizations; (iii) a trap-claim analysis showing that convergence is not automatically evidence for correctness — its evidential direction depends on whether the converged content is truth-tracking or bias-tracking; (iv) a benchmark protocol specification; and (v) a statement of limitations, failure modes, and falsification conditions. We do not report empirical measurements; every number is computed from stated assumptions in Section 4 or labeled as a projection in Section 5.

2. Background and Related Work

The supplied bibliography is drawn from adjacent fields rather than from prior work on cross-model convergence specifically; we therefore state only what each entry's own summary supports, and note where a summary is thin.

[1] Conjectures on Convergence and Scalar Curvature (arXiv:2103.10093v1). This survey collects the compactness and geometric stability conjectures formulated by participants at the 2018 IAS Emerging Topics Workshop on Scalar Curvature and Convergence, focusing on sequences of compact Riemannian manifolds with nonnegative scalar curvature. Its relevance here is thematic: it is a body of mathematics in which convergence of sequences is itself the object of study and stability of a limit object under perturbation is a central concern — an analogy for our question of when agreement between two noisy reasoners is stable evidence rather than coincidence. The entry's summary supplies no further detail we can rely on beyond this scope statement.

[2] IPPOG: Bridging the gap between science education at school and modern scientific research (arXiv:2011.14743v1). The International Particle Physics Outreach Group has, since 1997, made systematic efforts to present and popularize particle physics across all audiences and age groups, and is described as a strategic pillar in fostering long-term sustainable support for fundamental research. This is directly relevant to our shared-bias null: outreach and popularization shape the corpus-level framing of physics claims, and a fringe claim that has been popularized (correctly or incorrectly) is exactly the case where two LLMs may share a biased framing. The summary gives no further detail on specific programs.

[3] RerrFact: Reduced Evidence Retrieval Representations for Scientific Claim Verification (arXiv:2202.02646v2). This work motivates reduced-evidence retrieval representations for scientific claim verification by the exponential growth in digital information outlets and the race to publish, which have made scientific

misinformation more prevalent, and by the observation that verification requires in-depth knowledge and great labor from domain experts. This establishes the expert bottleneck our method targets; the summary supplied gives no further detail on the method's internals.

[4] LLM-based Corroborating and Refuting Evidence Retrieval for Scientific Claim Verification (arXiv:2503.07937v1). This paper introduces CIBER (Claim Investigation Based on Evidence Retrieval), an extension of the Retrieval-Augmented Generation (RAG) framework designed to identify corroborating and refuting documents as evidence for scientific claim verification; notably, CIBER addresses the inherent uncertainty in LLMs by evaluating response consistency across diverse interrogation probes. This is the closest published relative of our proposal: it already treats consistency across probes of a single system as a signal. Our proposal differs in the locus of consistency — agreement across independently trained models on argument structure rather than consistency across probes of one model — and Section 3 makes precise why cross-model independence changes the likelihood model.

[5] An analytical approach to Bayesian evidence computation (arXiv:2301.13783v1). The Bayesian evidence is a key tool in model selection, allowing comparison of models with different numbers of parameters, but its use in cosmological model analysis has been limited by computational difficulty, with current numerical algorithms requiring supercomputers. This paper gives exact formulae for the Bayesian evidence in the case of Gaussian likelihoods with arbitrary correlation structure. We follow the same spirit of exact analytic Bayes factors rather than simulation-based estimation, though our likelihoods are discrete (converge/diverge on argument structure) rather than Gaussian.

[6] Report of the DOE/NSF Workshop on Correctness in Scientific Computing, June 2023, Orlando, FL (arXiv:2312.15640v2). This report digests CSC'23, held on June 17, 2023 as part of the Federated Computing Research Conference (FCRC) 2023, conceived by DOE and NSF to address growing concerns about correctness among those who employ computational methods for large-scale scientific simulations. This grounds the institutional motivation: correctness of computationally mediated science is a recognized community concern at the level of funding agencies, and LLM-mediated claim evaluation is a natural extension of that concern. The summary gives no further detail on specific workshop recommendations.

[7] CCSBench: Evaluating Compositional Controllability in LLMs for Scientific Document Summarization (arXiv:2410.12601v3). This work introduces CCSBench, motivated by the observation that existing research on scientific document summarization typically controls single attributes (such as length and empirical focus) while compositional control of multiple attributes is underexplored. Its relevance is as evidence that evaluation infrastructure for LLMs on scientific text is maturing toward multi-attribute, compositional benchmarks — the same style of design our protocol in Section 3.4 requires (jointly controlling claim type, question type, and trap status). The summary gives no further detail on CCSBench's findings.

[8] YouZhi: Towards High-Concurrency Financial LLMs via Adaptive GQA-to-MLA Transition (arXiv:2606.05868v1). This paper states that LLMs drive significant financial innovations but that high-concurrency deployment is severely bottlenecked by KV-cache memory overhead, which inflates infrastructure costs and throttles scalability; it proposes YouZhi-LLM, an efficient financial LLM built via a structural GQA-to-MLA transition and training pipeline. This matters for the economics of our method: gate-checking by running many independent LLMs is only scalable if per-query inference is cheap, and [8] illustrates that the field is actively attacking concurrency cost. The summary gives no quantitative cost figures, so our cost analysis in Section 4.4 is confined to a combinatorial count of evaluation passes.

[9], [11], [12] (QNFO notes; DOIs <https://papers.qnfo.org/papers/impact-of-cognitive-linearity-on-epistemic-modeling/>, <https://papers.qnfo.org/papers/projective-geometric-frameworks-for-semantic-structures/>, <https://papers.qnfo.org/papers/static-architecture-of-reality-and-its-biological-interface/>). The supplied summaries for these three entries are empty; we cannot state what they contain beyond their titles. We cite them only as

corpus context indicating that "convergence" and "consilience" as epistemic notions are under active development in adjacent informal literature — [12]'s title names convergence consilience, the exact epistemic principle we formalize — and we make no claims about their content.

[10] QNFO: Boundary Ultrametricity: The Tree vs. $\partial_\infty \mathcal{T}$ Distinction, Applied to the ZBW Transition Graph (<https://papers.qnfo.org/papers/boundary-ultrametricity/>). Per its own summary, this formal note establishes that a boundary Gromov metric on dT_p is ultrametric and recovers $|\cdot|_p$, that the tree vertex metric is not ultrametric, and that ZBW P1 Claim C2 was corrected from an ultrametric core to a 0-hyperbolic core, with 147/500 violations per paper data. It is useful to us as a concrete instance of the fringe-physics ecosystem our method would gate-check, and as an instance of a claim *corrected* by formal analysis, illustrating the ground-truth annotation step of our benchmark (Section 3.4).

3. Methods

3.1 Hypotheses and observables

Fix a scientific claim C . An evaluator run is a pair (M_i, q) : model M_i answering question q about C . The observable is the model's *argument structure* S_i : the ordered sequence of intermediate propositions the model asserts en route to a verdict, extracted by a fixed parser. Two runs converge if $S_i = S_j$ under a defined equivalence (same propositions, same order, modulo paraphrase) and the final verdicts agree, $V_i = V_j$.

We compare three mutually exclusive hypotheses:

- H_1 (correctness): C is correct, and each model's output is produced by a partially truth-tracking, independent reasoning process.
- H_2 (shared bias): C 's apparent support is an artifact of shared training-corpus bias; both models retrieve or reproduce the same corpus-level pattern.
- H_3 (coincidence): C is incorrect and the models' agreement is coincidental.

Let A denote the observed convergence event. The likelihoods are parameters to be estimated by the benchmark of Section 3.4:

$$P(A | H_1) = \theta_1, \quad P(A | H_2) = \theta_2, \quad P(A | H_3) = \theta_3.$$

The structural assumption the benchmark must test is $\theta_1 > \theta_3$ (correct claims induce convergent arguments more than coincidence does), and that θ_2 is small for synthesis questions but large for retrieval-only questions and trap claims.

3.2 Bayes factors

With prior weights π_2, π_3 within the composite alternative $H_{23} = H_2 \cup H_3$, the Bayes factor for H_1 against H_{23} is

$$\text{BF}_{1/23} = \frac{\theta_1}{\pi_2 \theta_2 + \pi_3 \theta_3},$$

and posterior odds are $O_{\text{post}} = O_{\text{prior}} \times \text{BF}_{1/23}$, with posterior probability $P(H_1 | A) = O_{\text{post}} / (1 + O_{\text{post}})$.

For the two-hypothesis question-type-split variant used in our primary parameterization (Section 4.2), the per-pair Bayes factor at question type $q \in \{R, S\}$ is

$$K_q = \frac{P(\text{convergence} | H_c, q)}{P(\text{convergence} | H_b, q)} = \frac{p_c^q}{p_b^q},$$

with the design prediction $K_R \approx 1$ (retrieval convergence uninformative) and $K_S > 1$ (synthesis convergence informative).

3.3 Trap claims

A trap claim C_T is engineered so that a plausible-sounding but incorrect argument chain is heavily represented in common training corpora. Under H_b , both models reproduce the biased chain, so convergence probability is high. Under H_c with independent models, each model independently avoids the trap with probability $1 - \epsilon_T$; conditioning on convergence *on the trap answer*, the likelihood under H_c is ϵ_T^2 (both independently err identically) versus approximately p_b^T under H_b . Thus convergence on a trap answer yields

$$K_T = \frac{\epsilon_T^2}{p_b^T} \ll 1,$$

i.e., evidence *against* the reliability of the reasoning pipeline on this claim family. This asymmetry is the framework's self-correcting feature: convergence is not automatically evidence for correctness; its evidential direction depends on whether the converged content is truth-tracking or bias-tracking.

3.4 Benchmark protocol

The protocol has four arms: (1) ground-truth fringe claims with known truth values, spanning retrieval-like and synthesis-like question types; (2) trap claims as in Section 3.3; (3) a retrieval/synthesis split in which the same claims are evaluated under both question types, to test the prediction $K_R \approx 1 < K_S$; and (4) a model panel of at least two independently trained LLMs, with argument structures extracted by a fixed parser and convergence scored by a paraphrase-robust matcher applied by a blinded rater or held-out judge. The measured quantity is the empirical convergence rate per arm, from which empirical Bayes factors are estimated; discriminative power is the separation of K across arms. No empirical numbers are reported in this paper.

4. Analysis

All numbers below are either stated inputs with sources or computed here with full arithmetic. No empirical measurements are reported; likelihood values are illustrative assumptions, explicitly labeled. We present two parameterizations, whose differences are documented in Appendix A.

4.1 Input numbers

Symbol	Value	Source
$P(H_c)$ (Parameterization 1 prior)	0.1	Assumption: conservative modeling choice for the fringe-claim regime; no empirical source claimed
p_b^R, p_c^R	0.9, 0.9	Assumption: retrieval answers with corpus-present answers converge under both hypotheses
p_b^S	0.05	Assumption: shared corpus lacks the full argument for a genuinely fringe claim
p_c^S	0.4	Assumption: truth-tracking reasoning converges often but not always
ϵ_T	0.1	Assumption: per-model probability of independently falling into a trap
p_b^T	0.9	Assumption: shared bias reliably reproduces the trap chain
$\theta_1, \theta_2, \theta_3$ (Parameterization 2)	0.8, 0.3, 0.1	Assumptions A1–A3 (illustrative)
π_2, π_3	0.5, 0.5	Assumption A4: equal split within the alternative
θ_2^{trap}	0.9	Assumption A5 (illustrative)
ZBW violations	147 of 500	Supplied summary of [10]

4.2 Parameterization 1: question-type split, conservative prior

Derivation 1 (retrieval arm).

$$K_R = \frac{p_c^R}{p_b^R} = \frac{0.9}{0.9} = 1.$$

A Bayes factor of 1 leaves posterior odds equal to prior odds: retrieval convergence carries zero evidential weight. This is the formal statement of the design prediction.

Derivation 2 (synthesis arm, two models).

$$K_S = \frac{p_c^S}{p_b^S} = \frac{0.4}{0.05} = 8.$$

Prior odds: $\frac{P(H_c)}{P(H_i)} = \frac{0.1}{1-0.1} = \frac{0.1}{0.9} = \frac{1}{9} \approx 0.1111$. Posterior odds after two-model convergence on a synthesis question:

$$O_2 = \frac{1}{9} \times 8 = \frac{8}{9} \approx 0.8889, \quad P_2 = \frac{O_2}{1 + O_2} = \frac{8/9}{17/9} = \frac{8}{17} \approx 0.4706.$$

Two-model synthesis convergence moves a 0.1 prior to a posterior of approximately 0.471 — a substantial but not decisive update.

Derivation 3 (N -model cascade). For N mutually independent models each contributing likelihood ratio K_S , the first model's output is absorbed into the prior and each additional independent model contributes a factor K_S :

$$K_N = K_S^{N-1} = 8^{N-1}, \quad O_N = \frac{1}{9} \cdot 8^{N-1}, \quad P_N = \frac{O_N}{1 + O_N}.$$

- $N = 2$: $O_2 = \frac{8}{9} \approx 0.8889$, $P_2 = \frac{0.8889}{1.8889} \approx 0.4706$.
- $N = 3$: $O_3 = \frac{64}{9} \approx 7.1111$, $P_3 = \frac{7.1111}{8.1111} \approx 0.8767$.
- $N = 4$: $O_4 = \frac{512}{9} \approx 56.8889$, $P_4 = \frac{56.8889}{57.8889} \approx 0.9827$.

In base-10 logarithms, each additional model contributes $\log_{10} 8 \approx 0.903$ orders of magnitude of Bayes factor.

Derivation 4 (panel size for a 0.95 posterior). Require $P_N \geq 0.95$, i.e., $O_N \geq 19$:

$$\frac{1}{9} \cdot 8^{N-1} \geq 19 \Rightarrow 8^{N-1} \geq 171 \Rightarrow (N-1) \ln 8 \geq \ln 171.$$

With $\ln 8 \approx 2.0794$ and $\ln 171 \approx 5.1422$: $N-1 \geq \frac{5.1422}{2.0794} \approx 2.473$, so $N \geq 3.473$, hence $N = 4$. Consistent with Derivation 3: $P_3 \approx 0.877 < 0.95$ and $P_4 \approx 0.983 \geq 0.95$.

Derivation 5 (trap arm).

$$K_T = \frac{\epsilon_T^2}{p_b^T} = \frac{0.1^2}{0.9} = \frac{0.01}{0.9} \approx 0.0111.$$

Posterior odds after two-model convergence on a trap answer: $O_2^T = \frac{1}{9} \times 0.0111 \approx 0.001235$, so

$$P_2^T = \frac{0.001235}{1.001235} \approx 0.00123.$$

Convergence on a trap answer drives the posterior from 0.1 down to approximately 0.0012: the framework treats false consensus as strong evidence *against* the pipeline's reliability on this claim family.

Derivation 6 (sensitivity to p_b^S). Holding $p_c^S = 0.4$ and prior 0.1 fixed:

- $p_b^S = 0.02$: $K_S = 20$, $O_2 = \frac{20}{9} \approx 2.222$, $P_2 = \frac{2.222}{3.222} \approx 0.690$.
- $p_b^S = 0.10$: $K_S = 4$, $O_2 = \frac{4}{9} \approx 0.444$, $P_2 = \frac{0.444}{1.444} \approx 0.308$.
- $p_b^S = 0.20$: $K_S = 2$, $O_2 = \frac{2}{9} \approx 0.222$, $P_2 = \frac{0.222}{1.222} \approx 0.182$.

The posterior is highly sensitive to the null's convergence probability: if shared bias can reproduce full argument chains at rate 0.2, the two-model update nearly halves relative to the 0.05 case. This is the single most important quantity for the benchmark to measure.

4.3 Parameterization 2: three-hypothesis composite, even prior

Under assumptions A1–A4 ($\theta_1 = 0.8$, $\theta_2 = 0.3$, $\theta_3 = 0.1$, $\pi_2 = \pi_3 = 0.5$):

$$\text{BF}_{1/23} = \frac{0.8}{(0.5)(0.3) + (0.5)(0.1)} = \frac{0.8}{0.15 + 0.05} = \frac{0.8}{0.20} = 4.0.$$

With even prior odds $O_{\text{prior}} = 1$: $O_{\text{post}} = 4.0$, so

$$P(H_1 | A) = \frac{4.0}{1 + 4.0} = \frac{4}{5} = 0.8.$$

Under Parameterization 2, a single convergence observation on an unspecified question type moves an even prior to a posterior of 0.8. The two parameterizations answer different questions: Parameterization 1 conditions on question type and uses a conservative fringe-claim prior; Parameterization 2 pools question types and uses an even prior. They are not competing estimates of the same quantity, and we report both.

4.4 Cost of the panel

Section 2 notes that no quantitative cost figures are supplied by [8], so the cost analysis is confined to a combinatorial count of evaluator runs. Let N be the panel size, Q the number of claims per arm, and T the number of question types per claim. The total number of evaluator runs is

$$R = N \cdot Q \cdot T.$$

With the protocol of Section 3.4: synthesis/retrieval arm $Q = 100$, $T = 2$; trap arm $Q_T = 50$, $T = 1$; panel $N = 4$ (the size required by Derivation 4). Then

$$R_{\text{split}} = 4 \times 100 \times 2 = 800, \quad R_{\text{trap}} = 4 \times 50 \times 1 = 200,$$

$$R_{\text{total}} = 800 + 200 = 1000.$$

The benchmark therefore requires 1000 evaluator runs plus convergence scoring, before any inference-cost overhead per run, which we cannot quantify from the supplied bibliography.

5. Results

All numbers below are computed in Section 4 from the explicitly labeled illustrative assumptions of Table 4.1; no empirical measurements are reported. Under Parameterization 1 (prior $P(H_c) = 0.1$):

1. **Retrieval arm.** $K_R = \frac{0.9}{0.9} = 1$: retrieval convergence carries zero evidential weight (Derivation 1).
2. **Synthesis arm, two models.** $K_S = \frac{0.4}{0.05} = 8$; posterior odds $O_2 = \frac{8}{9} \approx 0.8889$; posterior $P_2 = \frac{8}{17} \approx 0.4706$ (Derivation 2).
3. **N -model cascade.** $P_3 \approx 0.8767$ and $P_4 \approx 0.9827$, with each additional model contributing $\log_{10} 8 \approx 0.903$ orders of magnitude (Derivation 3).
4. **Panel size.** A posterior of at least 0.95 requires $N = 4$ models (Derivation 4).
5. **Trap arm.** $K_T = \frac{0.1^2}{0.9} = \frac{0.01}{0.9} \approx 0.0111$; two-model convergence on a trap answer drives the posterior from 0.1 to $P_2^T \approx 0.00123$ (Derivation 5).
6. **Sensitivity to p_b^S .** Holding $p_c^S = 0.4$ and the prior fixed, the two-model posterior is ≈ 0.690 at $p_b^S = 0.02$, ≈ 0.308 at $p_b^S = 0.10$, and ≈ 0.182 at $p_b^S = 0.20$ (Derivation 6).

Under Parameterization 2 (assumptions A1–A4): $\text{BF}_{1/23} = 4.0$ and posterior $P(H_1 | A) = 0.8$ from an even prior. Under assumption A5 ($\theta_2^{\text{trap}} = 0.9$), the composite null's trap likelihood exceeds its synthesis likelihood, reproducing the direction of the Parameterization 1 trap result qualitatively.

Projection (labeled, not computed from measurements). If the benchmark of Section 3.4 measures p_b^S within the range $[0.02, 0.20]$ explored in Derivation 6, the two-model synthesis posterior lies within the computed envelope $[0.182, 0.690]$ for a 0.1 prior; the uncertainty bound is exactly the sensitivity range, and no tighter projection is warranted without measured likelihoods.

6. Discussion

Limitations. (i) Every likelihood in Section 4 is an illustrative assumption, not a measurement; the framework's practical value stands or falls on the benchmark's empirical estimates of p_b^S and p_c^S . (ii) The independence assumption behind the cascade $K_N = K_S^{N-1}$ is strong: models trained on overlapping web corpora share biases beyond those captured by H_2 , and correlated errors would deflate the effective per-model evidence. (iii) Argument-structure equivalence (same propositions, same order, modulo paraphrase) is operationalized by a parser and a paraphrase-robust matcher; both can fail, and matcher errors would contaminate the convergence statistic in an unknown direction. (iv) The supplied summaries for [9], [11], and [12] are empty and those for [1], [2], [3], [6], [7], and [8] give no quantitative detail, so the related-work grounding is thinner than a mature literature would allow; the bibliography contains only twelve entries, which limits the breadth of positioning. (v) The cost analysis counts evaluator runs only; per-run inference cost cannot be quantified from the supplied sources.

Failure modes. If shared bias can reproduce full multi-step argument chains at rates near the upper end of the sensitivity range ($p_b^S = 0.20$), the two-model update is weak ($P_2 \approx 0.182$) and the method degenerates toward uninformative agreement. If models are fine-tuned on common distillation data, apparent independence is illusory and

H_2 absorbs H_1 's likelihood mass. Conversely, if truth-tracking models converge rarely (p_c^S well below 0.4), the method produces false negatives on correct fringe claims.

What would falsify the claims. The central structural prediction is $K_R \approx 1 < K_S$: if the benchmark finds comparable convergence-rate ratios for retrieval and synthesis questions, the question-type dependence on which the framework rests is wrong. The trap prediction is $K_T \ll 1$: if convergence on engineered trap claims instead yields $K_T \geq 1$, shared-bias reproduction of trap chains is not distinguishable from independent error, and the self-correcting feature fails. A measured p_b^S near p_c^S for genuinely fringe claims would falsify the premise that synthesis convergence discriminates correctness from shared bias.

Open questions. How to set the prior $P(H_c)$ for claims of unknown provenance; how to bound corpus overlap between models quantitatively; whether argument-structure equivalence can be scored reproducibly enough to serve as a benchmark observable; and how the three-hypothesis composite of Parameterization 2 should be split by question type.

7. Conclusion

We formalized cross-model LLM convergence on scientific claims as Bayesian evidence, separating truth-tracking correctness, shared-corpus bias, and coincidence. Under explicitly labeled illustrative assumptions, two-model convergence on multi-step synthesis questions yields a Bayes factor of 8 and a posterior of $\frac{8}{17} \approx 0.471$ from a 0.1 prior; four models yield ≈ 0.983 ; convergence on engineered trap claims yields $K_T \approx 0.0111$ and a posterior of ≈ 0.0012 , i.e., evidence against reliability; and the two-model posterior spans ≈ 0.182 to ≈ 0.690 as the null's synthesis-convergence probability varies from 0.20 to 0.02. The framework's diagnostic value therefore hinges on a single measurable quantity, the null's convergence probability on synthesis questions, and on the structural prediction $K_R \approx 1 < K_S$, both testable by the specified four-arm benchmark. No empirical measurements are reported here; the benchmark protocol is the proposed next step.

References

- [1] Conjectures on Convergence and Scalar Curvature. arXiv:2103.10093v1. <https://arxiv.org/abs/2103.10093v1> [2] IPPOG : Bridging the gap between science education at school and modern scientific research. arXiv:2011.14743v1. <https://arxiv.org/abs/2011.14743v1> [3] RerrFact: Reduced Evidence Retrieval Representations for Scientific Claim Verification. arXiv:2202.02646v2. <https://arxiv.org/abs/2202.02646v2> [4] LLM-based Corroborating and Refuting Evidence Retrieval for Scientific Claim Verification. arXiv:2503.07937v1. <https://arxiv.org/abs/2503.07937v1> [5] An analytical approach to Bayesian evidence computation. arXiv:2301.13783v1. <https://arxiv.org/abs/2301.13783v1> [6] Report of the DOE/NSF Workshop on Correctness in Scientific Computing, June 2023, Orlando, FL. arXiv:2312.15640v2. <https://arxiv.org/abs/2312.15640v2> [7] CCSBench: Evaluating Compositional Controllability in LLMs for Scientific Document Summarization. arXiv:2410.12601v3. <https://arxiv.org/abs/2410.12601v3> [8] YouZhi: Towards High-Concurrency Financial LLMs via Adaptive GQA-to-MLA Transition. arXiv:2606.05868v1. <https://arxiv.org/abs/2606.05868v1> [9] <https://papers.qnfo.org/papers/impact-of-cognitive-linearity-on-epistemic-modeling/>. QNFO: Impact of Cognitive Linearity on Epistemic Modeling. [10] <https://papers.qnfo.org/papers/boundary-ultrametricity/>. QNFO: Boundary Ultrametricity: The Tree vs. $\partial_\infty \mathcal{G}$ Distinction, Applied to the ZBW Transition Graph. [11] <https://papers.qnfo.org/papers/projective-geometric-frameworks-for-semantic-structures/>. QNFO: Projective Geometric Frameworks for Semantic Structures. [12] <https://papers.qnfo.org/papers/static-architecture-of-reality-and-its-biological-interface/>. QNFO: Convergence Consilience and the Hierarchical Architecture of Reality.

Appendix A. Divergence report

The independent drafts used two parameterizations, retained here as Parameterization 1 and Parameterization 2 rather than silently merged.

- **Hypothesis space.** One draft used a two-hypothesis split (H_c versus H_b) conditioned on question type, with a conservative prior $P(H_c) = 0.1$; the other used a three-hypothesis composite (H_1, H_2, H_3) with an even prior and pooled question types. Convention chosen: report both, since they condition on different events and their posteriors (0.471 versus 0.8 for two-model/one-observation convergence) are not competing estimates of the same quantity.
- **Cascade convention.** One draft counted each of N models as contributing a factor K_S (giving $K_N = K_S^N$); the other absorbed the first model's output into the prior (giving $K_N = K_S^{N-1}$). Convention chosen: the absorb-first convention, $K_N = K_S^{N-1}$, because the prior already reflects one model's report; the alternative would raise all cascade posteriors and is noted as a convention choice, not a derivation.
- **Trap likelihood under the composite null.** One draft set $\theta_2^{\text{trap}} = 0.9$ (assumption A5); the other left the composite null's trap likelihood unspecified. Convention chosen: carry A5 as an explicitly labeled illustrative assumption and report the trap result primarily under Parameterization 1, where $p_b^T = 0.9$ is stated as an input.

Appendix B. Claim attribution

ID	Claim	Source drafts	Status
C1	Retrieval convergence is uninformative: $K_R = 1$ under equal retrieval likelihoods	A, B	CONVERGENT
C2	Two-model synthesis convergence with $p_c^S = 0.4$, $p_b^S = 0.05$, prior 0.1 gives $K_S = 8$, posterior $\frac{8}{17} \approx 0.471$	A, B	CONVERGENT
C3	Four-model cascade posterior ≈ 0.983 ; 0.95 posterior requires $N = 4$	A, B	CONVERGENT
C4	Trap convergence gives $K_T = \frac{0.01}{0.9} \approx 0.0111$ and posterior ≈ 0.0012	A, B	CONVERGENT
C5	Sensitivity: two-model posterior ranges ≈ 0.690 to ≈ 0.182 as p_b^S varies 0.02 to 0.20	A, B	CONVERGENT
C6	Three-hypothesis composite with A1–A4 gives $\text{BF}_{1/23} = 4.0$, posterior 0.8	B	SINGLE
C7	Cascade convention absorbs the first model into the prior ($K_N = K_S^{N-1}$)	A	SINGLE (B used K_S^N ; resolved per Appendix A)
C8	Benchmark cost is a combinatorial run count ($R = N \cdot Q \cdot T$; 1000 runs for the specified protocol)	A, B	CONVERGENT
C9	Convergence is evidence for or against correctness depending on whether converged content is truth-tracking or bias-tracking	A, B	CONVERGENT