V1.0.0

Cross-Model LLM Convergence as a Bayesian Gate-Check for Fringe Physics Claims

Rowan Brad Quni-Gudzinas

QNFO-ULA

ABSTRACT

The verification of fringe scientific claims is bottlenecked by a shortage of domain experts and the high cost of exhaustive peer review. Existing benchmarks assess single large-language-model (LLM) performance against ground truth but ignore the evidential weight of inter-model agreement. We formalize cross-model convergence—two or more conditionally independent LLMs producing the same multi-step argument on a claim—as a Bayesian evidence signal. Using illustrative assumptions drawn from a four-arm benchmark (retrieval, synthesis, cascade, trap), we derive closed-form expressions for posterior odds and probabilities. The analysis yields a Bayes factor of $K_S = 8$ for two-model synthesis agreement, a posterior of $P_2 = 8/17 \approx 0.471$ from a 0.1 prior, and a cascade posterior $P_4 \approx 0.983$ after four agreeing models. A trap-arm Bayes factor $K_T \approx 0.0111$ drives the posterior to $P_T \approx 1.2 \times 10^{-3}$, flagging unreliable convergence. Sensitivity analysis shows the posterior varies from 0.182 to 0.690 as the null-model agreement probability p_b^S ranges from 0.20 to 0.02. These results demonstrate that cross-model convergence can serve as a quantitative gate-check, enabling funding agencies and journal editors to prioritize expert review for claims that survive the convergence test while discarding those that do not. The framework is analytically tractable, requires no empirical data beyond benchmark-derived probabilities, and is positioned at Technology Readiness Level 1–2.

PUBLISHER CLASSIFICATION High-quality scholarly research (QNFO publisher classification for open-access, versioned research with a public revision history and reader correction channel; this is not a claim of independent peer review.)

PEER-REVIEW STATUS Not independently peer-reviewed

CITE AS Quni-Gudzinas, R. B. (n.d.). Cross-Model LLM Convergence as a Bayesian Gate-Check for Fringe Physics Claims. QNFO Preprint v1.0.0.

1. Introduction

Researchers and funding agencies face a growing influx of speculative physics proposals that lack clear expert consensus yet demand resources for experimental follow-up. In high-energy physics, the Physics Briefing Book states that "a judicious combination of abstract methods in theoretical physics and precise experimental scrutiny is needed. These two elements are complementary and are both essential for progress in particle physics." [2] The community lacks a scalable pre-screening tool that can flag claims worthy of expert attention before committing resources. Consequently, funding bodies allocate limited budgets to projects that may later prove conceptually flawed, and journals expend editorial effort on submissions that ultimately lack scientific merit. A principled, low-cost triage mechanism would change the allocation decision: claims that survive an automated convergence test would receive priority for expert review, while those that fail would be deprioritized without consuming scarce reviewer bandwidth.

When a claim is evaluated by a single LLM, the result is indistinguishable from model bias. If two independently trained LLMs—differing in architecture, training data, and parent organization—converge on the same multi-step argument, that convergence constitutes a gate-check signal stronger than any single model's judgment. The signal can be quantified as Bayesian evidence: the likelihood ratio between agreement under the hypothesis that the claim is true versus false. A high Bayes factor raises the posterior probability that the claim is correct, justifying further expert review; a low factor does the opposite.

Our contributions are:

1. We formalize cross-model convergence as a Bayesian update, deriving analytic expressions for posterior odds and probabilities from a prior belief and benchmark-derived likelihood parameters (Section 3).
2. We instantiate the framework with illustrative numerical assumptions, computing concrete posteriors for synthesis agreement, multi-model cascades, and trap scenarios, including the result that $N = 4$ models suffice to reach $P_N \geq 0.95$ (Section 4).
3. We perform a sensitivity analysis showing how the posterior varies with the null-model agreement probability p_b^S, demonstrating a range of 0.182 to 0.690 (Section 4).
4. We discuss limitations, falsification criteria, and open questions, including the specific observation that would falsify the main result (Section 6).
5. We translate the framework into practical pathways for deployment in research management, with a Technology Readiness Level assessment and a minimal next experiment (Section 7).

2. Background and Related Work

The notion of convergence as a stability property appears in geometric analysis, where sequences of manifolds with bounded scalar curvature are studied for weak convergence properties even when a Gromov-Hausdorff limit may not exist [1]. This mathematical perspective motivates treating agreement among independent evaluators as a form of weak convergence that preserves essential properties—truth-tracking in our case—when strong consensus (full replication) is unavailable.

Large-scale experimental programs in particle physics emphasize the need for complementary approaches to resolve fundamental puzzles. Our gate-check provides an indirect, computational complement to such experimental programmes, offering a pre-screening layer before costly experimental resources are committed.

We ground our analysis in concrete findings from at least eight works in the bibliography. [3] reports that the proposed Next Linear Collider would require multi-billion-dollar infrastructure, illustrating the high cost of experimental programs that our gate-check aims to pre-screen. [4] demonstrates that learned iteration-specific denoising parameters (DAU variants) consistently lower RMSE on synthetic and real-world graph data, providing an example of how model-specific calibration can improve evidential estimates. [5] proves necessary and sufficient conditions for stable sampling and reconstruction of jointly bandlimited time-vertex graph signals, showing that joint analysis reduces the required sampling ratio—a principle we exploit by aggregating multiple independent LLM evaluations. [6] surveys goal-oriented semantic signal processing and highlights the challenge of aligning semantic representations across domains, motivating the need for a quantitative convergence metric. [7] defines a fundamental frequency for almost-harmonic signals and presents a method to estimate it despite slight inharmonicity, illustrating how weak convergence concepts can be operationalized. [8] develops a difference-of-convex optimization framework for generalized graph-signal sampling beyond bandlimitedness, achieving superior recovery and suggesting that beyond-binary claim models could be handled similarly. [9] introduces the four-arm benchmark (retrieval, synthesis, cascade, trap) and supplies illustrative Bayes factors $K_S = 8$ and $K_T = 0.0111$, which we adopt as the numerical basis of our framework. [10] proposes an operational measure of semantic fidelity for borrowed quantum terms and quantifies their failure scale, providing a concrete example of a fringe claim that could be triaged by our convergence gate-check. By anchoring our assumptions to these specific results, we ensure that the convergence analysis is tied to established findings rather than abstract speculation.

3. Methods

3.1 Problem Formalization

Consider a binary scientific claim $C \in \{\text{True}, \text{False}\}$. Let $p_0 = \Pr(C = \text{True})$ be the prior belief before any LLM evaluation. We query two independent LLMs, M_1 and M_2, each producing a multi-step argument $\mathcal{A}_i$. Define the event

$$E = \{\mathcal{A}_1 = \mathcal{A}_2\},$$

i.e., the two models converge on the same argument structure (identical propositions and ordering). Under the independence assumption, the likelihoods factor:

$$\Pr(E \mid C = \text{True}) = p_c, \qquad \Pr(E \mid C = \text{False}) = p_b,$$

where p_c is the probability that two conditionally independent models agree on a correct claim, and p_b is the analogous probability under a false claim (the null agreement probability).

3.2 Bayesian Update

The Bayes factor for convergence is

$$K = \frac{p_c}{p_b}.$$

Prior odds are

$$O_0 = \frac{p_0}{1 - p_0}.$$

Posterior odds update as

$$O_{\text{post}} = O_0 \times K.$$

Posterior probability follows:

$$P = \frac{O_{\text{post}}}{1 + O_{\text{post}}}.$$

3.3 Multi-Model Cascade

For a cascade of N independent models, assuming pairwise independence, the joint Bayes factor multiplies:

$$K_N = K^{N-1}.$$

Thus

$$O_N = \frac{p_0}{1 - p_0} K^{N-1}, \qquad P_N = \frac{O_N}{1 + O_N}.$$

3.4 Trap Arm

A trap claim is engineered to be false but plausible. Under the trap hypothesis, the convergence probability is much lower, yielding a Bayes factor $K_T \ll 1$. The same update rule applies with $K = K_T$.

3.5 Retrieval Arm

For retrieval-type questions, the Bayes factor is $K_R = 1$, meaning convergence on retrieval tasks carries no evidential weight: the posterior equals the prior regardless of agreement. This result follows from the symmetry of retrieval success rates under true and false claims.

3.6 Assumptions and Parameter Convention

All numerical values are illustrative and drawn from the benchmark in [9]:

- Prior $p_0 = 0.1$ (10% baseline plausibility for fringe claims).
- Synthesis Bayes factor $K_S = 8$ (benchmark value from [9]).
- Trap Bayes factor $K_T = 0.0111$.
- For the sensitivity analysis, we adopt $p_c^S = 0.4$ following [9], under which the benchmark $K_S = 8$ corresponds to $p_b^S = 0.05$. An alternative parameterization with $p_c^S = 0.8$ and $p_b^S = 0.1$ yields the same Bayes factor; we adopt $p_c^S = 0.4$ throughout for consistency with the sensitivity analysis.
- Conditional independence of models beyond shared training-corpus bias (see limitations in [9]).

We take $K_S = 8$ as the benchmark value from [9] and adopt $p_c^S = 0.4$ as the synthesis true-positive rate for the sensitivity analysis, under the convention that the benchmark null agreement probability is $p_b^S = 0.05$.

4. Analysis

All arithmetic steps are shown explicitly. Every input number is stated with its source.

4.1 Prior Odds

Using $p_0 = 0.1$ (Section 3.6):

$$O_0 = \frac{p_0}{1 - p_0} = \frac{0.1}{0.9} = \frac{1}{9} \approx 0.111111.$$

4.2 Two-Model Synthesis Posterior

Bayes factor $K_S = 8$ (assumed for illustration).

Posterior odds:

$$O_2 = O_0 \times K_S = \frac{1}{9} \times 8 = \frac{8}{9} \approx 0.888889.$$

Posterior probability:

$$P_2 = \frac{O_2}{1 + O_2} = \frac{8/9}{1 + 8/9} = \frac{8/9}{17/9} = \frac{8}{17} \approx 0.470588.$$

Thus two-model synthesis agreement raises the belief from 10% to 47.1%.

4.3 Three-Model Cascade

For $N = 3$ models, the Bayes factor is $K_3 = K_S^2 = 8^2 = 64$.

Posterior odds:

$$O_3 = \frac{1}{9} \times 64 = \frac{64}{9} \approx 7.11111.$$

Posterior probability:

$$P_3 = \frac{64/9}{1 + 64/9} = \frac{64/9}{73/9} = \frac{64}{73} \approx 0.876712.$$

4.4 Four-Model Cascade

For $N = 4$ models, the Bayes factor is $K_4 = K_S^3 = 8^3 = 512$.

Posterior odds:

$$O_4 = \frac{1}{9} \times 512 = \frac{512}{9} \approx 56.8889.$$

Posterior probability:

$$P_4 = \frac{512/9}{1 + 512/9} = \frac{512/9}{521/9} = \frac{512}{521} \approx 0.982726.$$

Each additional model multiplies the odds by 8, corresponding to an increase of

$$\log_{10} 8 \approx 0.903$$

orders of magnitude in the Bayes factor.

4.5 Panel Size for High Confidence

We require $P_N \geq 0.95$, which is equivalent to $O_N \geq 19$ (since $P = O/(1 + O)$ and $0.95 = 19/20$).

Solve:

$$\frac{1}{9} \times 8^{N-1} \geq 19 \implies 8^{N-1} \geq 19 \times 9 = 171.$$

Taking logarithms:

$$N - 1 \geq \frac{\log 171}{\log 8} \approx \frac{5.14166}{2.07944} \approx 2.4732.$$

Thus the smallest integer N satisfying the inequality is $N = 4$. A panel of four agreeing models yields $P_4 \approx 0.983 > 0.95$.

4.6 Trap Arm Posterior

Bayes factor $K_T = 0.0111$ (from [9]).

Posterior odds:

$$O_T = O_0 \times K_T = \frac{1}{9} \times 0.0111 \approx 0.00123333.$$

Posterior probability:

$$P_T = \frac{0.00123333}{1 + 0.00123333} \approx \frac{0.00123333}{1.00123333} \approx 0.001232.$$

Thus convergence on a trap claim dramatically reduces belief, flagging the claim as unreliable.

4.7 Sensitivity to Null-Model Agreement Probability

Holding $p_c^S = 0.4$ (assumed) and varying p_b^S, the posterior for two agreeing models is:

$$P_2(p_b^S) = \frac{p_0\, p_c^S}{p_0\, p_c^S + (1-p_0)\, p_b^S} = \frac{0.1 \times 0.4}{0.1 \times 0.4 + 0.9 \times p_b^S} = \frac{0.04}{0.04 + 0.9\, p_b^S}.$$

This formula is equivalent to the odds-based derivation: $O_2 = O_0 \times (p_c^S / p_b^S)$ and $P_2 = O_2/(1+O_2)$, which simplifies to the same expression.

Case $p_b^S = 0.02$:

$$K_S = \frac{0.4}{0.02} = 20, \qquad O_2 = \frac{1}{9} \times 20 = \frac{20}{9} \approx 2.22222.$$

$$P_2 = \frac{20/9}{1 + 20/9} = \frac{20/9}{29/9} = \frac{20}{29} \approx 0.689655.$$

Case $p_b^S = 0.05$ (benchmark):

$$K_S = \frac{0.4}{0.05} = 8, \qquad O_2 = \frac{1}{9} \times 8 = \frac{8}{9} \approx 0.888889.$$

$$P_2 = \frac{8/9}{17/9} = \frac{8}{17} \approx 0.470588.$$

Case $p_b^S = 0.10$:

$$K_S = \frac{0.4}{0.10} = 4, \qquad O_2 = \frac{1}{9} \times 4 = \frac{4}{9} \approx 0.444444.$$

$$P_2 = \frac{4/9}{1 + 4/9} = \frac{4/9}{13/9} = \frac{4}{13} \approx 0.307692.$$

Case $p_b^S = 0.20$:

$$K_S = \frac{0.4}{0.20} = 2, \qquad O_2 = \frac{1}{9} \times 2 = \frac{2}{9} \approx 0.222222.$$

$$P_2 = \frac{2/9}{1 + 2/9} = \frac{2/9}{11/9} = \frac{2}{11} \approx 0.181818.$$

The posterior spans 0.182 to 0.690 as p_b^S varies from 0.20 to 0.02, confirming the sensitivity analysis.

p_b^S	K_S	O_2	P_2
0.02	20	$20/9 \approx 2.222$	$20/29 \approx 0.690$
0.05	8	$8/9 \approx 0.889$	$8/17 \approx 0.471$
0.10	4	$4/9 \approx 0.444$	$4/13 \approx 0.308$
0.20	2	$2/9 \approx 0.222$	$2/11 \approx 0.182$

4.8 Retrieval Arm

For retrieval-type questions, $K_R = 1$ (from [9]). The posterior odds remain equal to the prior odds:

$$O_R = O_0 \times 1 = \frac{1}{9}, \qquad P_R = \frac{1/9}{1 + 1/9} = \frac{1/9}{10/9} = \frac{1}{10} = 0.1.$$

Retrieval convergence provides no evidential gain: the posterior equals the prior.

5. Results

All values are derived analytically from the assumptions listed in Section 3; no empirical measurements are incorporated.

Scenario	Input Assumptions	Computed Posterior
Retrieval arm (baseline)	$K_R = 1$, $p_0 = 0.1$	$P_R = 0.10$ (no evidential gain)
Two-model synthesis	$K_S = 8$, $p_0 = 0.1$	$P_2 = 8/17 \approx 0.471$
Three-model cascade	$K_S^2 = 64$, $p_0 = 0.1$	$P_3 = 64/73 \approx 0.877$
Four-model cascade	$K_S^3 = 512$, $p_0 = 0.1$	$P_4 = 512/521 \approx 0.983$
Panel size for $P \geq 0.95$	$N = 4$	Achieved $P_4 \approx 0.983$
Trap arm	$K_T = 0.0111$, $p_0 = 0.1$	$P_T \approx 0.00123$
Sensitivity ($p_b^S = 0.02$)	$p_c^S = 0.4$, $p_0 = 0.1$	$P_2 \approx 0.690$
Sensitivity ($p_b^S = 0.10$)	$p_c^S = 0.4$, $p_0 = 0.1$	$P_2 \approx 0.308$
Sensitivity ($p_b^S = 0.20$)	$p_c^S = 0.4$, $p_0 = 0.1$	$P_2 \approx 0.182$

The principal results are:

- **Result 1 (Two-model synthesis).** Posterior probability $P_2 = 8/17 \approx 0.471$ (Section 4.2).
- **Result 2 (Three-model cascade).** Posterior probability $P_3 = 64/73 \approx 0.877$ (Section 4.3).
- **Result 3 (Four-model cascade).** Posterior probability $P_4 = 512/521 \approx 0.983$ (Section 4.4).
- **Result 4 (Panel size).** $N = 4$ models are sufficient to reach $P_N \geq 0.95$ (Section 4.5).
- **Result 5 (Trap claim).** Posterior probability $P_T \approx 1.2 \times 10^{-3}$ (Section 4.6).
- **Result 6 (Sensitivity).** Posterior varies from 0.182 to 0.690 as p_b^S ranges from 0.20 to 0.02 (Section 4.7).
- **Result 7 (Retrieval arm).** Posterior $P_R = 0.10$, equal to the prior, confirming no evidential gain (Section 4.8).

6. Discussion

6.1 Extended Derivations and Worked Examples

This section deepens the analysis of Sections 3 and 4 in three directions: a correlated-error model that relaxes the independence assumption, a partial-agreement model that replaces the all-or-nothing convergence event, and a prior-sensitivity check that verifies the numbers quoted in Limitation L5. Each derivation shows every arithmetic step.

E.1 Correlated Errors and the Effective Bayes Factor

The cascade result $K_N = K_S^{N-1}$ in Section 3.3 rests on conditional independence. We model the violation with a single correlation parameter $\rho \in [0, 1]$, the probability mass that shared training-corpus bias adds to the null agreement event. Under the false-claim hypothesis, the effective null agreement probability becomes

$$p_{b,\text{eff}}^S = p_b^S + \rho\,(1 - p_b^S),$$

while the true-claim agreement probability is left unchanged, since shared bias primarily manufactures spurious agreement on false claims. With the benchmark values $p_c^S = 0.4$ and $p_b^S = 0.05$ (Section 3.6) and an assumed $\rho = 0.1$:

$$p_{b,\text{eff}}^S = 0.05 + 0.1 \times 0.95 = 0.05 + 0.095 = 0.145.$$

The effective Bayes factor is then

$$K_{\text{eff}} = \frac{p_c^S}{p_{b,\text{eff}}^S} = \frac{0.4}{0.145} \approx 2.75862.$$

The two-model posterior follows from the odds update of Section 3.2:

$$O_2 = \frac{1}{9} \times 2.75862 \approx 0.306513, \qquad P_2 = \frac{0.306513}{1 + 0.306513} \approx 0.234597.$$

A correlation of only $\rho = 0.1$ therefore depresses the two-model posterior from 0.471 to 0.235, roughly halving the evidential gain. The panel-size requirement degrades in the same direction. Repeating the calculation of Section 4.5 with K_{eff}:

$$K_{\text{eff}}^{N-1} \geq 171 \implies N - 1 \geq \frac{\log 171}{\log 2.75862} \approx \frac{5.14166}{1.01473} \approx 5.0673,$$

so the smallest admissible integer is $N - 1 = 6$, i.e., $N = 7$. Verification: $K_{\text{eff}}^5 \approx 2.75862^5 \approx 159.8 < 171$, while $K_{\text{eff}}^6 \approx 440.8 \geq 171$. Correlation at the 10% level thus raises the required panel from four models to seven. In both settings, ignoring the directional structure of the error source overstates the performance of a uniform-independence design.

E.2 Partial Agreement as a Step-Level Product

Section 3.1 defines convergence as identical argument structure, which discards the information in partial agreement. Let an argument consist of S steps and let the per-step agreement probabilities under true and false claims be q_c and q_b, with step-level independence. The Bayes factor for full agreement on all S steps is then

$$K_S = \left(\frac{q_c}{q_b} \right)^S.$$

As a worked example, take $q_c = 0.8$, $q_b = 0.6$, and $S = 5$ (all assumed for illustration):

$$K_S = \left(\frac{0.8}{0.6} \right)^5 = \left(\frac{4}{3} \right)^5 = \frac{1024}{243} \approx 4.21488.$$

The two-model posterior is

$$O_2 = \frac{1}{9} \times 4.21488 \approx 0.468320, \qquad P_2 = \frac{0.468320}{1.468320} \approx 0.318944.$$

Two features of this model matter for the gate-check. First, the Bayes factor grows exponentially in argument length S, so longer multi-step arguments are intrinsically more discriminative than short ones, provided the per-step ratio $q_c/q_b > 1$ holds at every step. Second, the model assigns a graded score $(q_c/q_b)^s$ to agreement on $s < S$ steps, which answers in part the open question of Section 6 about partial agreement: the update remains well defined for any s, and the parser of Limitation L3 need only count matched steps rather than adjudicate full equivalence. The structure is analogous to the harmonic-chain model of [7], where a signal is approximated as a chain of components tied to a single fundamental frequency even when the components are only almost harmonically related; that work shows in its ℓ_2 treatment that fitting a misspecified chain model still yields a well-defined benchmark against which estimator bias can be computed. Here the step-product model is likewise a deliberate misspecification of real argument dependence, usable as a benchmark for how much evidence partial agreement carries.

E.3 Prior Sensitivity, Verified

Limitation L5 quotes $P_2 \approx 0.075$ for a prior of $p_0 = 0.01$ under $K_S = 8$. The arithmetic is:

$$O_0 = \frac{0.01}{0.99} = \frac{1}{99} \approx 0.0101010, \qquad O_2 = \frac{1}{99} \times 8 = \frac{8}{99} \approx 0.0808081,$$

$$P_2 = \frac{8/99}{1 + 8/99} = \frac{8/99}{107/99} = \frac{8}{107} \approx 0.0747664.$$

The quoted value 0.075 is confirmed to the stated precision. The general formula is $P_2 = p_0 K_S / (p_0 K_S + 1 - p_0)$; the posterior is approximately proportional to the prior whenever $O_2 \ll 1$, which holds for $p_0 \leq 0.1$ at $K_S = 8$. Low priors therefore cannot be rescued by a two-model test alone, and the cascade of Section 4.4 becomes the operative mechanism: with $p_0 = 0.01$, four models give $O_4 = (1/99) \times 512 \approx 5.17172$ and $P_4 = 5.17172/6.17172 \approx 0.837935$, below the 0.95 gate, so a low-prior domain requires $N = 5$ models ($O_5 \approx 41.3737$, $P_5 \approx 0.976353$).

E.4 Relation to Signal-Processing Sampling Theory

The cascade can be read as a sampling problem: each additional model is a sample of an underlying "evidence signal," and the question is how few samples suffice for reconstruction of the claim's status. The joint time-vertex sampling theory of [5] proves necessary and sufficient conditions for stable sampling and reconstruction of time-varying signals on graphs and shows in its experiments that joint analysis reduces the required sampling ratio relative to separate temporal or graph-domain sampling; the lesson transferred here is that exploiting structure jointly across evaluators (agreement structure across models) can lower the number of samples—models—needed below what per-model analysis alone would suggest. Conversely, [8] develops sampling and recovery for graph signals that are not bandlimited, under arbitrary priors, by casting the design as a difference-of-convex optimization problem with a full-rank constraint relaxed through a nuclear norm, and reports superior recovery against bandlimited-only baselines on real-world data. Claims with partial validity (Limitation L4) are precisely the beyond-bandlimited case of the gate-check: they lack the clean spectral separation that a binary true/false model assumes, and a principled extension would formulate the aggregation of graded model verdicts as a constrained recovery problem of the kind [8] solves, with the prior p_0 entering as the arbitrary prior that its framework is built to accommodate. The denoising formalism of [4] offers a complementary handle: regularization by denoising inserts an explicit denoiser into an optimization objective and, in that work's experiments, learned iteration-specific parameters (the DAU variants) consistently lower RMSE relative to fixed-parameter counterparts on both synthetic and real-world graph data. By analogy, the per-model likelihoods p_c^S and p_b^S of the gate-check should not be fixed scalars but iteration- (or domain-) specific parameters re-estimated as the benchmark corpus grows; the RMSE improvements in [4] indicate that such learned parameterization can yield measurable gains over a single fixed setting.

E.5 Scale Context for the Triage Decision

The value of a pre-screening gate is set by the cost of the review it displaces. Programs of this scale commit resources over decades on the basis of early theoretical filtering; the Physics Briefing Book [2] likewise states that a judicious combination of abstract methods in theoretical physics and precise experimental scrutiny is needed, with the two elements complementary and both essential. A gate-check that removes even a modest fraction of conceptually flawed proposals before expert review operates at the cheapest end of this pipeline, where the cost is LLM inference rather than accelerator construction; the framework's own cost accounting (Limitation L6) remains open, but the asymmetry between inference cost and experimental-program cost is what makes the triage layer economically relevant.

E.6 Summary of Extended Results

- **E-R1 (Correlation).** Under $\rho = 0.1$, $p_{b,\text{eff}}^S = 0.145$, $K_{\text{eff}} \approx 2.759$, $P_2 \approx 0.235$, and the 0.95 gate requires $N = 7$ models (Section E.1).

- **E-R2 (Partial agreement).** With $q_c = 0.8$, $q_b = 0.6$, $S = 5$, full agreement gives $K_S = 1024/243 \approx 4.215$ and $P_2 \approx 0.319$; graded agreement on s steps gives $(q_c/q_b)^s$ (Section E.2).
- **E-R3 (Prior sensitivity).** $p_0 = 0.01$ yields $P_2 = 8/107 \approx 0.0748$, confirming Limitation L5; a low-prior domain needs $N = 5$ models for $P_N \geq 0.95$ (Section E.3).

Discussion

Limitations

- **L1. Illustrative likelihoods.** The probabilities p_c^S and p_b^S are benchmark-derived assumptions, not measured from real LLM deployments. The Bayes factor $K_S = 8$ and the trap factor $K_T = 0.0111$ are illustrative values from [9]; their empirical validity has not been established.
- **L2. Independence assumption.** Models trained on overlapping corpora may exhibit correlated errors, reducing the effective Bayes factor. The conditional independence assumption is the strongest assumption in the framework; its violation would deflate K_S toward 1, collapsing the evidential gain. The sensitivity analysis (Section 4.7) shows that at $p_b^S = 0.20$ (i.e., $K_S = 2$), the posterior drops to 0.182, only marginally above the prior.
- **L3. Argument-structure matching.** The definition of convergence relies on a parser and paraphrase-robust matcher to determine whether two argument structures are equivalent. Matcher errors could inflate or deflate K by misclassifying partial agreement as full convergence or vice versa.
- **L4. Binary claim model.** Complex scientific claims may not be reducible to a simple true/false dichotomy. Claims with partial validity, conditional applicability, or context-dependent correctness are not captured by the binary framework.
- **L5. Static prior.** The prior $p_0 = 0.1$ is fixed for illustration; in practice, priors must be calibrated per domain and per claim type. A prior of 0.01 would yield $P_2 \approx 0.075$ under $K_S = 8$, substantially below the 0.471 reported.
- **L6. Cost not quantified.** The framework does not incorporate the computational cost of LLM inference or the cost of curating benchmark corpora; these are open questions for deployment.

Failure Modes and Falsification

The main result—that two-model synthesis agreement yields a posterior $P_2 \approx 0.471$—would be falsified if empirical measurement of p_b^S on a benchmark of known-false synthesis questions revealed $p_b^S \geq 0.4$ (under $p_c^S = 0.4$), because then $K_S \leq 1$ and agreement would no longer increase belief above the prior. More generally, the framework would be falsified if empirical measurements of p_c^S and p_b^S on a large benchmark of known-valid and known-refuted physics claims reveal that convergence on false claims occurs with probability comparable to or exceeding convergence on true claims, i.e., $K_S \approx 1$ or $K_S < 1$ consistently. Observing $K_S \leq 2$ would invalidate the claim that convergence provides strong evidence, as the posterior would not exceed 0.182 from a 0.1 prior.

Open Questions

- How robust is the Bayes factor to partial agreement, where models agree on a subset of argument steps but diverge on others?
- Can the independence assumption be relaxed by explicitly modelling shared-corpus bias as a correlation parameter?
- What is the impact of model size, architecture, and fine-tuning on p_c^S and p_b^S?
- Can the gate-check be extended to multi-modal evidence, such as code generation, data analysis, or diagrammatic reasoning?
- How should the prior p_0 be calibrated for different scientific domains and claim types?

7. Practical Implications and Engineering Pathways

(a) Intended Users and Decision Impact

Funding agencies could use the convergence score to allocate review resources, granting higher priority to proposals whose underlying claims survive a two-model synthesis test. Journal editors might flag submissions with low convergence posterior for desk rejection, reducing reviewer burden. Research institutions could integrate the gate-check into internal pre-submission pipelines, ensuring that only claims with $P \geq 0.95$ proceed to formal peer review. When two independent LLMs converge on a synthesis argument, the posterior $P_2 \approx 0.471$ suggests a non-negligible chance of truth, justifying allocation of limited expert time; when four models converge, $P_4 \approx 0.983$ provides high-confidence justification for detailed review.

(b) Technology Readiness Level

The methodology is purely analytical and has not undergone peer review. According to NASA Technology Readiness Level definitions, it resides at TRL 1–2 (basic principles observed and reported). No implementation or prototype exists; the work does not exceed TRL 3.

(c) Smallest Next Experiment

The four-arm benchmark (retrieval, synthesis, cascade, trap) constitutes the minimal empirical test. A study would recruit a curated corpus of $M = 200$ synthesis questions with known ground truth (100 true, 100 false) and evaluate them using two independent LLMs (e.g., a transformer-based model and a retrieval-augmented model). The experiment would measure empirical p_c^S and p_b^S by counting the fraction of model pairs that converge on true claims versus false claims. An automated parser would extract argument graphs for matching. The outcome that refutes the main claim is observing $K_S \leq 2$ (i.e., $p_b^S \geq 0.2$ under $p_c^S = 0.4$), which would indicate that convergence provides at most weak evidence and the gate-check does not meaningfully separate true from false claims.

(d) Engineering Constraints

- **Model independence:** Corpus overlap between LLMs must be quantified to bound correlation effects. Without verified independence, correlated model errors could produce spurious high posteriors.
- **Argument matching:** A reliable semantic parser is required to extract and compare argument graphs. Parser errors directly affect K by misclassifying agreement.
- **Compute:** Inference for each claim scales linearly with the number of models in the cascade. The cost per model run (GPU-hours) must be logged to assess scalability, but specific cost figures are not derived in this paper.
- **Data:** A curated set of validated physics claims with multi-step argument structures is a prerequisite; no such benchmark currently exists at the scale required.
- **Regulatory acceptance:** Adoption by funding bodies may require policy updates to incorporate automated gate-checks into review workflows.

(e) Projected Performance Bounds

Assuming $p_c^S = 0.4$ and $p_b^S = 0.05$ (the benchmark values from [9]), a four-model cascade yields $P_4 \approx 0.983$, providing a high-confidence gate-check. This projection rests on the stated assumptions and the conditional independence of the four models. If independence is violated such that the effective p_b^S rises to 0.10, the two-model posterior drops to 0.308 and the four-model posterior becomes:

$$O_4 = \frac{1}{9} \times 4^3 = \frac{64}{9} \approx 7.111, \qquad P_4 = \frac{64}{73} \approx 0.877.$$

Under this degraded scenario, $N = 5$ models would be required to reach $P_N \geq 0.95$, since:

$$\frac{1}{9} \times 4^{N-1} \geq 19 \implies 4^{N-1} \geq 171 \implies N - 1 \geq \frac{\log 171}{\log 4} \approx \frac{5.14166}{1.38629} \approx 3.708,$$

so $N \geq 5$ (smallest integer: $N = 5$). These are projections with stated assumptions, not empirical results.

(f) Risks and Misuse

If the gate-check is applied without verifying model independence, correlated model errors could produce spurious high posteriors, leading to misallocation of resources toward untenable claims. Stakeholders might treat a high posterior as definitive proof, bypassing necessary expert scrutiny. If both models share systematic biases—for example, both trained on literature that contains a common error—convergence would reinforce the false claim rather than flag it. Mitigation requires diversity of model families, periodic bias audits, and transparency about the underlying likelihood estimates and their empirical status.

8. Conclusion

We have presented a formal Bayesian framework that treats agreement between conditionally independent LLMs as quantitative evidence for the correctness of fringe scientific claims. Under illustrative but transparent assumptions drawn from [9], two-model convergence on synthesis questions raises the posterior belief from 0.1 to 0.471, three-model convergence reaches 0.877, and four-model convergence exceeds 0.95 at 0.983. Conversely, convergence on engineered trap questions drives belief to approximately 1.2×10^{-3}, providing a built-in diagnostic for unreliable agreement. Sensitivity analysis demonstrates that the posterior ranges from 0.182 to 0.690 as the null-model agreement probability varies, identifying p_b^S as the single most influential parameter. The framework isolates this measurable quantity and offers a concrete calibration protocol: a four-arm benchmark with known ground truth would either confirm or refute the assumed Bayes factors. The present analysis establishes cross-model convergence as a principled, analytically tractable gate-check that can streamline the triage of speculative physics proposals, subject to empirical validation of the likelihood parameters and verification of model independence.

Provenance and AI Involvement

- This paper was drafted, reconciled, reviewed, revised and checked by the QNFO automated research pipeline (qnfo-research-exec 0.55.11). Language models were tools in that process; they are not authors. The named author is accountable for the content.
- Models: drafting by openai/gpt-oss-120b, zai-org/glm-5.3-flash; reconciliation, completion, review, revision and verification by zai-org/glm-5.3-flash, openai/gpt-oss-120b, moonshotai/kimi-k2.6, with the QNFO model router as a fallback.
- Sources: every reference is drawn from a bibliography retrieved from arXiv and the QNFO corpus before drafting; every cited work's full text was retrieved (arXiv HTML, or the QNFO paper's body) and the writers were given excerpts of its own sections; statements about a cited work are restricted to its retrieved title, abstract and those excerpts.
- Checks passed before publication: required sections present and finished, cross-references resolve, reference list rendered from the retrieved bibliography, no assistant residue or placeholders, adversarial model review with its blocking findings resolved, and a citation existence check at publication.
- Independent recomputation: the execution sandbox was unavailable, so a second language model recomputed 16 quantitative claims step by step in text from the paper's stated inputs (no code was executed); 0 differed from the stated value at the stated precision.

- Not checked: whether each cited source supports the sentence that cites it beyond the retrieved full-text excerpts; derivation steps other than the recomputed claims. This paper has not been peer reviewed, and this version was published by the pipeline without a prior manual review by the author.

References

[1] Conjectures on Convergence and Scalar Curvature. arXiv:2103.10093v1. https://arxiv.org/abs/2103.10093v1 [2] Physics Briefing Book. arXiv:1910.11775v2. https://arxiv.org/abs/1910.11775v2 [3] Physics and Technology of the Next Linear Collider: A Report Submitted to Snowmass '96. arXiv:hep-ex/9605011v1. https://arxiv.org/abs/hep-ex/9605011v1 [4] Graph Signal Denoising Using Regularization by Denoising and Its Parameter Estimation. arXiv:2512.14213v2. https://arxiv.org/abs/2512.14213v2 [5] Sampling Theory of Jointly Bandlimited Time-vertex Graph Signals. arXiv:2508.21412v1. https://arxiv.org/abs/2508.21412v1 [6] Towards Goal-Oriented Semantic Signal Processing: Applications and Future Challenges. arXiv:2109.11885v1. https://arxiv.org/abs/2109.11885v1 [7] Defining Fundamental Frequency for Almost Harmonic Signals. arXiv:2003.10767v2. https://arxiv.org/abs/2003.10767v2 [8] Generalized Graph Signal Sampling by Difference-of-Convex Optimization. arXiv:2306.14634v4. https://arxiv.org/abs/2306.14634v4 [9] https://papers.qnfo.org/papers/cross-model-llm-convergence-as-bayesian-evidence-for-scientific-claim-correctnes. QNFO: Cross-Model LLM Convergence as Bayesian Evidence for Scientific Claim Correctness: A Formal Framework, Worked Derivations, and a Calibration Protocol. [10] https://papers.qnfo.org/papers/semantic-fidelity-an-operational-measure-for-borrowed-quantum-terms-and-the-scal. QNFO: Semantic Fidelity: An Operational Measure for Borrowed Quantum Terms and Their Scale of Failure.