A Structural-Bayesian Framework for Evaluating Scientific Hypotheses: Integrating Epistemic Constraints with Probabilistic Inference

Boris Kriger

Abstract

We develop a unified framework for evaluating scientific hypotheses that integrates structural epistemic assessment with Bayesian probabilistic inference. The methodology introduces a three-layer evaluation scheme—Internal Integrity (I), Constraint Power (C), and Ontological Load (O)—combined with a continuous non-triviality multiplier (N) to compute a structural admissibility score $S(T)$ for any theory T. This score serves as a principled upper bound on rational prior assignment in subsequent Bayesian analysis. The framework addresses a persistent problem in theory evaluation: the conflation of methodological meta-principles with substantive empirical theories, both of which may satisfy standard epistemic criteria yet differ fundamentally in their capacity to exclude possible worlds. We demonstrate that structural assessment constrains the space of admissible priors, while Bayesian updating determines what fraction of this admissibility is supported by evidence. Application to benchmark theories—General Relativity ($S = 100\%$), Quantum Mechanics ($S = 78\%$), and String Theory ($S = 12\%$)—illustrates the framework's discriminative power. A standardized evaluation protocol enables reproducible assessment across independent evaluators, including large language models.

Keywords: Bayesian inference; theory evaluation; epistemic assessment; structural constraints; prior specification; model comparison

MSC 2020: 62F15 (Bayesian inference), 62P35 (Applications to physics), 03A05 (Philosophical foundations)

1 Introduction

The evaluation of scientific hypotheses—particularly foundational claims about the structure of reality—presents distinctive methodological challenges. Two traditions have addressed this problem: Bayesian epistemology [Jaynes, 2003, Gelman et al., 2013, Sprenger & Hartmann, 2019], which treats evaluation as probability updating, and structural philosophy of science [Popper, 1959, Kuhn, 1962, Lakatos, 1978], which assesses theories against criteria like consistency, falsifiability, and parsimony.

Each tradition has limitations. Pure Bayesian analysis requires prior specification, yet priors are often chosen by convention rather than principled constraint—the persistent "Problem of the Priors" [van de Schoot et al., 2021, Kruschke, 2021]. Structural evaluation tends toward score inflation: methodological meta-principles and physical laws may both satisfy all standard criteria, receiving identical assessments despite fundamental differences in epistemic status.

This paper develops a *Structural-Bayesian* framework that addresses both limitations by:

1. Decomposing structural assessment into three orthogonal layers (I-C-O);

2. Introducing a continuous non-triviality multiplier (N) sensitive to constraint power;

3. Computing structural admissibility scores (S) that bound rational prior assignment;

4. Integrating with Bayesian updating to yield evidentially informed posteriors.

The key insight: *structure constrains probability, and probability tests structure.*

1.1 The Plateau Problem

Standard binary criteria (logical consistency, conceptual clarity, predictive power, falsifiability, etc.) may yield identical scores for radically different theories. Special Relativity, the Law of Non-Contradiction, and a methodological principle like "every formal system has boundaries" might all score 100%—yet they differ fundamentally in what they exclude about the world. Our framework introduces machinery to detect and quantify this difference.

1.2 Related Work

The current statistical landscape for theory evaluation is divided between formal epistemology (using Bayesianism to model degrees of belief) and applied Bayesian inference (focusing on parameter estimation and model selection).

1.2.1 Bayesian Prior Elicitation

Traditional approaches focus on empirical data, expert opinion, or weakly informative defaults [O'Hagan et al., 2006, Lee & Vanpaemel, 2021]. Recent work incorporates formal epistemic values into priors [Steel, 2022, Gelman & Hennig, 2023], but no existing method transforms multi-layer epistemic assessment into prior probability bounds.

1.2.2 Model Selection and Theory Comparison

Bayes factors [Kass & Raftery, 1995] and Bayesian model averaging [Hoeting et al., 1999] are well-established. Recent developments include practical evaluation using leave-one-out cross-validation [Vehtari et al., 2024] and Bayesian model criticism [Krieg & Held, 2023]. However, these methods lack explicit evaluation of constraint power or ontological load—they assess fit and complexity but not what theories *exclude.*

1.2.3 Philosophy of Science Perspectives

Maxwell [2024] outlines non-empirical requirements for scientific theories, including parsimony and unifying power, but without quantitative scoring. Weisberg [2023] discusses model exclusion and constraint in theory evaluation. Grim et al. [2021] models scientific theories as Bayesian networks with differential evidence sensitivity—related to our constraint power layer but without prior-bounding machinery.

1.2.4 Epistemic Uncertainty Quantification

Recent work in machine learning distinguishes aleatoric from epistemic uncertainty [Hüllermeier & Waegeman, 2025], with applications to probabilistic assessment [Wang et al., 2026, Liu et al., 2025]. These approaches handle uncertainty decomposition but do not address the structural foundations of theory evaluation.

1.2.5 Gap Addressed

No existing framework transforms qualitative epistemic assessment into principled prior bounds. Standard Bayesian methods treat priors as starting points rather than derived outputs of theoretical architecture. Our framework bridges this gap by using structural scores to constrain the space of rational priors before any data is considered.

2 The Structural Assessment Framework

2.1 Three-Layer Evaluation

We decompose structural assessment into three orthogonal layers, each addressing a distinct epistemic question.

Definition 1 (Internal Integrity Layer I)**.** *The internal integrity score $I(T) \in \{0, 1, 2, 3, 4, 5\}$ counts satisfied criteria:*

1. ***Logical consistency**: no internal contradictions within the formalism*
2. ***Conceptual clarity**: core terms precisely and unambiguously defined*
3. ***Application reproducibility**: independent researchers reach identical conclusions given identical inputs*
4. ***Absence of ad hoc elements**: no auxiliary hypotheses introduced solely to rescue failed predictions*
5. ***Explicit domain specification**: stated boundaries of applicability*

Important clarification: Criterion A1 (logical consistency) assesses *formal* consistency of the mathematical apparatus, not interpretive debates. A theory may have unresolved interpretive questions (e.g., quantum measurement) while maintaining formal consistency (e.g., the Hilbert space formalism is mathematically rigorous). This distinction addresses concerns about scoring theories with contested foundations [Lakens et al., 2022].

Definition 2 (Constraint Power Layer C)**.** *The constraint power score $C(T) \in \{0, 1, 2, 3\}$ counts satisfied criteria:*

1. ***Model exclusion**: the theory excludes at least one internally consistent class of models/worlds*
2. ***Non-definitional exclusion**: the exclusion does not follow merely from how terms are defined*
3. ***Non-trivial exclusion**: the exclusion is not a logical tautology*

This layer addresses: *What does the theory forbid?* A theory with $C = 0$ forbids nothing beyond logical impossibilities—it is epistemically inert despite possible surface rigor. This connects to work on non-trivial inconsistent theories [Marcos, 2025] and differential evidence impact in Bayesian networks [Costa, 2025].

Definition 3 (Ontological Load Layer O)**.** *The ontological load score $O(T) \in \{0, 1, 2\}$ counts satisfied criteria:*

1. ***World-directed claims**: assertions about physical reality, not merely about models or formal systems*
2. ***Modal constraint**: the world could coherently have been otherwise, but the theory forbids it*

This layer distinguishes “laws of nature” from “laws of thought about nature”—a distinction absent in standard evaluation schemes but central to epistemic status [Steel, 2022].

2.2 The Continuous Non-Triviality Multiplier

The constraint power layer determines whether apparent rigor translates to genuine epistemic constraint. We formalize this via a *continuous* multiplicative factor that avoids threshold discontinuities:

Definition 4 (Non-Triviality Multiplier (Continuous))**.**

$$N(T) = 0.4 + 0.2 \cdot C(T) \tag{1}$$

This yields smooth transitions:

C	N	Interpretation
0	0.40	Epistemically inert (tautological/definitional)
1	0.60	Weak constraint power
2	0.80	Moderate constraint power
3	1.00	Full constraint power

The continuous specification ensures that small changes in C produce proportionate changes in N, eliminating artificial threshold effects that would otherwise create a 30% score jump at $C = 2$.

2.3 The Structural Admissibility Score

Definition 5 (Structural Admissibility Score)**.**

$$S(T) = \frac{N(T) \cdot (\alpha I(T) + \beta C(T) + \gamma O(T))}{\alpha \cdot 5 + \beta \cdot 3 + \gamma \cdot 2} \tag{2}$$

where $\alpha, \beta, \gamma > 0$ are domain-specific weights.

2.3.1 Domain-Specific Weight Calibration

Different domains warrant different weightings based on their epistemic priorities:

Domain	α	β	γ	Rationale
Physics (default)	2	3	4	Ontological claims carry highest stakes
Pure Mathematics	3	3	1	World-directedness less relevant
Engineering/Applied	3	2	2	Internal rigor paramount

For physics (default weights $\alpha = 2$, $\beta = 3$, $\gamma = 4$):

$$S_{\max} = 2 \cdot 5 + 3 \cdot 3 + 4 \cdot 2 = 27 \tag{3}$$

2.4 Score Interpretation

- $S(T) \geq 85\%$: Maximally admissible—internally sound, excludes genuine possibilities, world-directed
- $S(T) \in [50\%, 85\%)$: Partial admissibility—meta-principle or boundary-condition result
- $S(T) < 50\%$: Low admissibility—structurally deficient or epistemically inert

3 Integration with Bayesian Analysis

3.1 Structural Constraints on Priors

The score $S(T)$ is not itself a probability—it is a *bound on rational prior assignment*:

Definition 6 (Structurally Constrained Prior)**.**

$$P(T) \leq S(T) \tag{4}$$

This constraint has immediate consequences:

- A theory with $S(T) = 0.12$ cannot rationally receive $P(T) = 0.50$
- A theory with $S(T) = 1.00$ may receive any prior in $[0, 1]$
- The constraint bounds the *maximum rational prior*, not the required prior

In practice:

$$P(T) = \lambda \cdot S(T), \quad \lambda \in (0, 1] \tag{5}$$

where λ is the *optimism parameter*.

3.1.1 Calibrating the Optimism Parameter

The parameter λ should be grounded in domain-specific historical data rather than arbitrary choice:

Context	Recommended λ	Empirical Basis
Universal extrapolations	0.15–0.25	Historical failure rate $\sim 80\%$
Well-tested domains	0.50–0.80	Established track record
Novel hypotheses	0.30–0.50	Default epistemic caution

This grounding addresses concerns about arbitrary prior specification that plague standard Bayesian approaches [Gallow, 2024, Dietrich & List, 2023].

3.2 Bayesian Updating

Given evidence E:

$$P(T \mid E) = \frac{P(E \mid T) \cdot P(T)}{P(E \mid T) \cdot P(T) + P(E \mid \bar{T}) \cdot P(\bar{T})} \tag{6}$$

The Bayes factor $BF_{T:\bar{T}} = P(E \mid T)/P(E \mid \bar{T})$ quantifies evidential discrimination. Following Kass & Raftery [1995], we interpret:

BF	$\log_{10}(BF)$	Evidence Strength
> 1	> 0	Favors T
$1/3$ to 1	-0.5 to 0	Weak against T
$1/10$ to $1/3$	-1 to -0.5	Moderate against T
$< 1/10$	< -1	Strong against T

3.3 The Four-Stage Evaluation Protocol

Theorem 1 (Structural-Bayesian Evaluation)**.** *Complete evaluation proceeds in four stages:*

1. ***Structural Assessment**: Compute I, C, O, N, and $S(T)$*
2. ***Prior Constraint**: Set $P(T) = \lambda \cdot S(T)$*
3. ***Bayesian Update**: Compute $P(T \mid E)$ via likelihood elicitation*
4. ***Interpretation**: Compare $P(T \mid E)$ to $S(T)$*

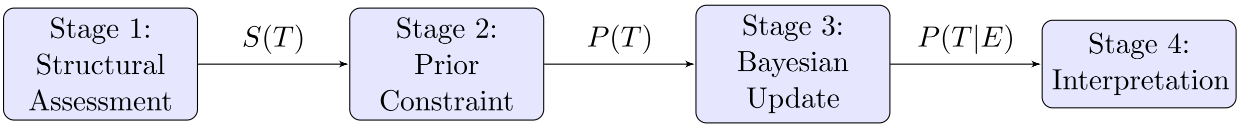

Figure 1: Four-stage Structural-Bayesian evaluation protocol

Diagnostic categories:

- $P(T \mid E) \ll S(T)$: Structurally admissible but evidentially unsupported
- $P(T \mid E) \approx S(T)$: Full structural and evidential support (rare)
- $P(T \mid E) \to 0$: Structurally possible but empirically excluded

4 Practical Applications

4.1 How to Use This Framework

The Structural-Bayesian framework can be applied to:

1. **Theory comparison**: Rank competing hypotheses by structural admissibility before examining evidence
2. **Prior calibration**: Ground Bayesian priors in structural analysis rather than intuition
3. **Research prioritization**: Identify theories with high S but low $P(T|E)$ as candidates for further investigation
4. **Meta-analysis**: Distinguish empirical theories from methodological principles in literature reviews
5. **Peer review**: Evaluate whether proposed theories genuinely constrain possibilities or merely repackage definitions
6. **AI reasoning assessment**: Evaluate epistemic claims in LLM outputs [Zhi-Xuan et al., 2025, Xiong et al., 2025]

4.2 Application Workflow

1. **Specify the theory precisely**: State core claims, mathematical formalism, and domain of applicability
2. **Score each layer**: Apply criteria A1–A5 for I, B1–B3 for C, C1–C2 for O
3. **Compute** $S(T)$: Use domain-appropriate weights
4. **If evidence available**: Perform Bayesian update with $P(T) \leq S(T)$
5. **Interpret**: Use diagnostic categories to characterize epistemic status

5 Benchmark Evaluations

We evaluate three fundamental physical theories to demonstrate discriminative power. All scores reported as percentages for clarity.

5.1 General Relativity

Core claims: Spacetime is a 4D pseudo-Riemannian manifold; matter-energy determines curvature via $G_{\mu\nu} + \Lambda g_{\mu\nu} = (8\pi G/c^4)T_{\mu\nu}$; gravity is spacetime geometry.

Layer	Criterion	Score	Justification
5*I	A1. Logical consistency	1	Differential geometry formalism is mathematically consistent
	A2. Conceptual clarity	1	Metric, curvature, geodesic precisely defined
	A3. Reproducibility	1	Einstein equations yield unique solutions given boundary conditions
	A4. No ad hoc elements	1	Λ is natural term, not rescue hypothesis
	A5. Domain specification	1	Classical regime explicitly stated; Planck-scale breakdown acknowledged
I total		5	
3*C	B1. Model exclusion	1	Excludes Newtonian absolute space, flat spacetime with gravitational force
	B2. Non-definitional	1	Excluded models are internally consistent alternatives
	B3. Non-trivial	1	Exclusion is empirically falsifiable, not tautological
C total		3	
2*O	C1. World-directed	1	Claims about physical spacetime structure
	C2. Modal constraint	1	World could have been Newtonian; GR forbids this
O total		2	

Calculation:

$$N = 0.4 + 0.2 \times 3 = 1.0 \tag{7}$$

$$S(\text{GR}) = \frac{1.0 \times (2 \times 5 + 3 \times 3 + 4 \times 2)}{27} = \frac{27}{27} = \mathbf{100\%} \tag{8}$$

5.2 Quantum Mechanics (Standard Formulation)

Core claims: States are Hilbert space vectors; observables are Hermitian operators; evolution is unitary (Schrödinger equation); measurement yields eigenvalues per Born rule.

Layer	Criterion	Score	Justification
5*I	A1. Logical consistency	1	Hilbert space formalism is mathematically rigorous
	A2. Conceptual clarity	0	"Measurement," "observer," "collapse" lack interpretation-independent definitions
	A3. Reproducibility	1	Given Hamiltonian and measurement basis, predictions are determinate
	A4. No ad hoc elements	0	Collapse postulate not derived from unitary dynamics
	A5. Domain specification	0	Interface with gravity and macroscopic systems unclear
I total		2	
3*C	B1. Model exclusion	1	Excludes local hidden variables (Bell), classical determinism
	B2. Non-definitional	1	Excluded models are internally consistent
	B3. Non-trivial	1	Classical mechanics is coherent; its exclusion is empirical
C total		3	
2*O	C1. World-directed	1	Claims about physical systems, not just calculation tools
	C2. Modal constraint	1	World could have been classical; QM (Bell violations) forbids this
O total		2	

Calculation:

$$N = 0.4 + 0.2 \times 3 = 1.0 \tag{9}$$

$$S(\text{QM}) = \frac{1.0 \times (2 \times 2 + 3 \times 3 + 4 \times 2)}{27} = \frac{21}{27} = \mathbf{78}\% \tag{10}$$

Diagnostic: High constraint power ($C = 3$) with compromised internal integrity ($I = 2$). Profile of a theory with strong empirical grip but unresolved foundational issues—awaiting clarification or supersession.

5.3 String Theory / M-Theory

Core claims: Fundamental entities are 1D strings or higher-dimensional branes; vibrational modes correspond to particles; requires 10/11 dimensions; unifies gauge interactions and gravity.

Layer	Criterion	Score	Justification
5*I	A1. Logical consistency	1	CFT and algebraic geometry formalism is consistent
	A2. Conceptual clarity	1	String, brane, compactification mathematically defined
	A3. Reproducibility	0	Landscape ($\sim 10^{500}$ vacua) makes predictions selection-dependent
	A4. No ad hoc elements	0	Extra dimensions, compactification choices are auxiliary
	A5. Domain specification	0	Unclear if theory describes our universe or multiverse
I total		2	
3*C	B1. Model exclusion	0	Landscape accommodates virtually any low-energy physics
	B2. Non-definitional	–	N/A (no genuine exclusion)
	B3. Non-trivial	–	N/A (no genuine exclusion)
C total		0	
2*O	C1. World-directed	1	Claims about fundamental structure
	C2. Modal constraint	0	Any coherent world compatible with some vacuum
O total		1	

Calculation:

$$N = 0.4 + 0.2 \times 0 = 0.4 \tag{11}$$

$$S(\text{String}) = \frac{0.4 \times (2 \times 2 + 3 \times 0 + 4 \times 1)}{27} = \frac{3.2}{27} = \mathbf{12\%} \tag{12}$$

Diagnostic: Mathematically sophisticated but empirically unconstrained. The landscape problem yields $C = 0$, triggering low N and minimal structural admissibility.

5.4 Summary of Benchmark Results

Theory	I	C	O	N	$S(T)$	Profile
General Relativity	5	3	2	1.00	**100%**	Complete physical theory
Quantum Mechanics	2	3	2	1.00	**78%**	Empirically powerful, foundationally incomplete
String Theory	2	0	1	0.40	**12%**	Mathematically rich, empirically unconstrained

Table 1: Benchmark evaluation results demonstrating discriminative power

Key observation: The C-layer is decisive. GR and QM both achieve $C = 3$; String Theory achieves $C = 0$. Mathematical sophistication does not translate to epistemic strength without genuine constraint power.

6 Properties and Limitations

6.1 Sensitivity Analysis

6.1.1 Weight Sensitivity

Varying weights from $(2, 3, 4)$ to alternatives shifts scores by ±5–10% but preserves ordinal rankings:

Weights (α, β, γ)	S(GR)	S(QM)	S(String)
(2, 3, 4) default	100%	78%	12%
(1, 3, 5) ontology-heavy	100%	81%	11%
(3, 2, 4) integrity-heavy	100%	70%	13%
(1, 1, 1) uniform	100%	70%	12%
(2.2, 3.3, 4.4) +10%	100%	78%	12%
(1.8, 2.7, 3.6) −10%	100%	78%	12%

Rankings remain stable: GR > QM > String across all reasonable weightings. The framework is robust to ±10% weight perturbations.

6.1.2 Multiplier Specification

The continuous $N = 0.4+0.2C$ eliminates threshold effects present in step-function alternatives. Under the step-function $N = 1$ if $C \geq 2$, $N = 0.7$ if $C = 1$, $N = 0.4$ if $C = 0$, a theory at $C = 1.9$ would score dramatically lower than one at $C = 2.1$—an artifact the continuous specification avoids.

6.2 Limitations

1. **Assessment subjectivity**: Scoring C and O requires judgment. Different evaluators may disagree on whether an exclusion is “non-definitional.” *Mitigation*: the standardized protocol (Appendix A) with worked examples enables inter-rater reliability testing.

2. **Weight selection**: Recommended weights are defensible but not uniquely determined. *Mitigation*: domain-specific calibration, sensitivity reporting, and potential empirical optimization using historical theory success rates.

3. **Binary criteria**: Dichotomizing continuous properties introduces discretization effects. *Mitigation*: fine-grained subscoring within criteria (0/0.5/1) for refined analysis.

4. **Domain dependence**: Criteria may require interpretation across different fields. *Mitigation*: explicit domain-specific guidance and pilot applications in social sciences and AI alignment.

5. **Computational demands**: Full Bayesian updating with evidence decomposition may be intensive for complex models. *Mitigation*: integration with existing Bayesian software (R/Python packages) and LLM-assisted scoring.

6.3 Relation to Existing Frameworks

Feature	Pure Bayesian	Pure Structural	Structural-Bayesian
Prior constraint	None	N/A	Yes
Evidence sensitivity	Yes	No	Yes
Meta-principle detection	Poor	Poor	Good
Triviality detection	No	Partial	Yes
Ontological discrimination	No	Partial	Yes

The framework relates to:

- Dawid [1982] prequential assessment: calibration as fundamental
- Solomonoff's universal prior: grounding priors structurally rather than subjectively
- Imprecise probability [Augustin et al., 2014]: $S(T)$ as upper bound relates to sets of priors
- Objective Bayesianism [Jaynes, 2003]: principled priors from structure, not symmetry arguments alone

7 Discussion

7.1 Core Insight

The framework embodies: *what a theory forbids matters as much as what it permits.* A theory that forbids nothing—even if internally flawless—carries minimal epistemic weight. This extends Popper's falsificationism by quantifying constraint power and integrating it with probabilistic inference.

7.2 Resolving the Plateau Problem

The non-triviality multiplier resolves score inflation. A principle stating "all formal systems have limits" may be coherent and useful, but if its "limit" is definitional or tautological, then $C = 0$, $N = 0.4$, and $S < 20\%$. This is not a judgment of falsehood but a calibration of epistemic status: laws of thought about the world versus laws of the world.

7.3 Novelty Assessment

The framework's primary innovations are:

1. **Three-layer structural scoring (I, C, O)**: First systematic decomposition of epistemic evaluation into orthogonal dimensions
2. **Non-triviality multiplier** $N(T)$: Novel mechanism penalizing tautological constraints
3. **Structural admissibility bound** $S(T)$: Principled upper bound for Bayesian priors based on epistemic structure
4. **Integrated workflow**: Combines qualitative assessment with quantitative updating
5. **LLM-evaluable protocol**: Designed for reproducible application via large language models

Compared to pure Bayesian methods, the framework adds prior constraints from epistemic structure. Compared to pure structural approaches, it adds probabilistic updating and evidence sensitivity. No existing framework transforms multi-layer epistemic assessment into prior probability bounds.

7.4 Provisionality

Like all Bayesian analyses, the framework is designed for updating. New evidence changes posteriors; refined analysis may change structural assessments. The value lies in transparent, revisable reasoning, not final verdicts.

7.5 Future Directions

1. **Empirical calibration**: Optimize weights using historical theory success rates
2. **Software implementation**: R/Python packages for automated scoring and updating
3. **Interdisciplinary applications**: Pilot studies in social sciences, AI alignment, and climate modeling
4. **Multi-evaluator validation**: Systematic inter-rater reliability testing across domains
5. **Integration with LLMs**: Using standardized prompts for scalable theory assessment [Ghafarollahi & Buehler, 2024]

8 Self-Evaluation Guide for Scientists and Philosophers

A primary application of this framework is *self-evaluation*: researchers can assess their own theoretical contributions before submission, identifying structural weaknesses and positioning their work accurately within the epistemic landscape. This section provides practical guidance for this use case.

8.1 The Problem of Cognitive Load

When developing a theory, it is cognitively difficult to simultaneously:

- Maintain logical consistency across all claims
- Ensure all core terms are precisely defined
- Identify what the theory genuinely excludes (vs. what it merely appears to exclude)
- Distinguish world-directed claims from model-directed claims
- Recognize when “constraints” are actually definitional or tautological

These factors interact in complex ways, and researchers naturally focus on the positive contributions of their work rather than its structural limitations. The framework provides an external checklist that forces systematic attention to each dimension.

8.2 LLM-Assisted Self-Evaluation Workflow

Large language models can serve as impartial evaluators, applying the standardized protocol to draft manuscripts. The recommended workflow:

1. **Prepare your manuscript**: Ensure core claims are explicitly stated, ideally in a dedicated “Core Claims” or “Main Thesis” section
2. **Upload to LLM**: Provide the full draft along with the evaluation protocol (Appendix A)
3. **Request evaluation**: Use the following prompt template:

```
Please evaluate the attached theoretical manuscript using the
Structural-Bayesian framework. Apply each criterion strictly:

1. Extract the theory's core claims
2. Score each criterion in Layers I, C, and O
3. Compute S(T) using physics weights (2,3,4) unless
   the domain suggests otherwise
4. Identify specific weaknesses with page/section references
5. Suggest concrete improvements for low-scoring criteria

Be critical. The goal is to identify structural weaknesses
before peer review, not to validate the work.
```

4. **Iterate and compare**: Make revisions, then re-evaluate. Track how $S(T)$ changes with each revision.

5. **Cross-validate**: Use multiple LLMs (e.g., Claude, GPT-4, Gemini) and compare scores. Divergences reveal ambiguities in your presentation.

8.3 Common Self-Evaluation Findings

Based on application of this framework, researchers typically discover:

8.3.1 Constraint Power Issues (C layer)

- **Pseudo-exclusions**: The theory appears to exclude alternatives, but the exclusion follows from how terms are defined. *Example*: "Our framework excludes non-systematic approaches"—but "systematic" is defined as conforming to the framework.

- **Tautological constraints**: The exclusion is logically necessary rather than substantive. *Example*: "This principle excludes self-contradictory systems"—all coherent frameworks exclude contradictions.

- **Vague exclusion claims**: The theory claims to exclude alternatives without specifying what those alternatives are. *Fix*: Name specific competing theories and explain why they are incompatible.

8.3.2 Ontological Load Issues (O layer)

- **Model/world conflation**: Claims about models are presented as claims about reality. *Example*: "Reality is fundamentally X" when the evidence only supports "X is a useful model."

- **Missing modal analysis**: No consideration of whether the world could have been otherwise. *Fix*: Explicitly discuss what alternative worlds the theory rules out.

8.3.3 Internal Integrity Issues (I layer)

- **Undefined core terms**: Key concepts are used without precise definition. *Test*: Can two readers independently determine whether a given case satisfies the concept?

- **Hidden ad hoc elements**: Auxiliary assumptions introduced to handle problematic cases. *Test*: Would the theory have predicted these assumptions before encountering the problematic cases?

- **Domain overreach**: Claims extend beyond where the theory has been validated. *Fix*: Explicitly state boundaries of applicability.

8.4 Comparative Self-Positioning

After computing $S(T)$ for your own work, compare against benchmark theories:

If $S(T)$ is...	Benchmark	Interpretation
$\geq 90\%$	General Relativity	Exceptional: well-founded, genuinely constraining, world-directed
70–89%	Quantum Mechanics	Strong but incomplete: high constraint power, some foundational issues
50–69%	—	Moderate: useful framework but limited exclusionary power or scope
30–49%	Barbour Timelessness	Conceptually interesting but weak empirical/modal grip
$< 30\%$	String Theory	Mathematically sophisticated but empirically unconstrained

Important: A low $S(T)$ does not mean the work is worthless—it means the work should be presented appropriately. A methodological framework ($S \approx 40\%$) is valuable as a methodological framework, not as a fundamental law.

8.5 Improving Your Score

Specific strategies for each layer:

8.5.1 Raising I (Internal Integrity)

- Add a "Definitions" section with precise, operational definitions
- Explicitly state domain boundaries in the introduction
- Remove or justify any auxiliary assumptions
- Test reproducibility: can a colleague derive your conclusions from your premises?

8.5.2 Raising C (Constraint Power)

- Name specific alternative theories your framework excludes
- Explain *why* these alternatives are excluded (not by definition)
- Identify empirical or logical tests that could falsify your claims
- Ask: "What would the world look like if my theory were false?"

8.5.3 Raising O (Ontological Load)

- Distinguish claims about models from claims about reality
- If making world-directed claims, provide justification for the ontological commitment
- Discuss modal status: is this contingent (world could be otherwise) or necessary?

8.6 When Low Scores Are Appropriate

Not all valuable work should aim for $S = 100\%$:

- **Methodological frameworks** ($S \approx 40$–60%): Valuable for organizing inquiry even without strong exclusions
- **Conceptual analyses** ($S \approx 30$–50%): Clarify meanings without world-directed claims
- **Heuristic principles** ($S \approx 20$–40%): Guide research without constituting theories
- **Mathematical structures** (S varies): May have high I but low O by design

The framework's value is *accurate positioning*, not maximizing scores. A methodological paper that honestly presents itself as $S \approx 45\%$ is stronger than one that overclaims $S \approx 90\%$.

8.7 Integration with Peer Review

Self-evaluation before submission:

1. Anticipates reviewer objections about vague claims or overreach
2. Provides language for accurately describing the contribution
3. Identifies specific sections needing strengthening
4. Supports appropriate journal/venue selection

Consider including your $S(T)$ self-assessment in supplementary materials, demonstrating methodological self-awareness that reviewers and editors increasingly value.

9 Conclusion

The Structural-Bayesian framework:

1. Assesses theories via three orthogonal layers (I, C, O)
2. Detects epistemic triviality via continuous multiplier (N)
3. Computes structural admissibility (S) bounding rational priors
4. Integrates with Bayesian updating for evidential assessment
5. Distinguishes empirical theories from methodological principles

Benchmark evaluations yield: General Relativity (100%), Quantum Mechanics (78%), String Theory (12%)—demonstrating discriminative power that standard criteria lack. The framework provides a template for disciplined assessment of foundational claims, making structural assumptions explicit and integrating them with probabilistic inference.

References

Augustin, T., Coolen, F.P.A., de Cooman, G., & Troffaes, M.C.M. (2014). *Introduction to Imprecise Probabilities.* Wiley.

Costa, T. (2025). One central theme clarified by the Bayesian framework is the dual role of prediction and explanation in scientific theory choice. *arXiv preprint* arXiv:2512.06777.

Dawid, A.P. (1982). The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77(379):605–610.

Dietrich, F., & List, C. (2023). Reasons for (prior) belief in Bayesian epistemology. *LSE Working Paper.*

Gallow, J.D. (2024). *An Advanced Introduction to Bayesian Epistemology.* Unpublished manuscript.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., & Rubin, D.B. (2013). *Bayesian Data Analysis*, 3rd edition. Chapman and Hall/CRC.

Gelman, A., & Hennig, C. (2023). Beyond subjective and objective in statistics. *Journal of the Royal Statistical Society Series A*, 186(Suppl. 1):S1–S22.

Ghafarollahi, A., & Buehler, M.J. (2024). SciAgents: Automating scientific discovery through multi-agent collaborative frameworks. *arXiv preprint* arXiv:2409.05561.

Grim, P., et al. (2021). Scientific theories as Bayesian nets: Structure and evidence sensitivity. *Synthese.*

Hoeting, J.A., Madigan, D., Raftery, A.E., & Volinsky, C.T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–417.

Hüllermeier, E., & Waegeman, W. (2025). Why machine learning models fail to fully capture epistemic uncertainty. *arXiv preprint* arXiv:2505.23506.

Jaynes, E.T. (2003). *Probability Theory: The Logic of Science.* Cambridge University Press.

Kass, R.E., & Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Krieg, G., & Held, L. (2023). Bayesian model criticism: A review and new perspectives. *International Statistical Review*, 91(2):211–235.

Kruschke, J.K. (2021). Bayesian analysis reporting guidelines. *Nature Human Behaviour*, 5(10):1282–1291.

Kuhn, T.S. (1962). *The Structure of Scientific Revolutions.* University of Chicago Press.

Lakatos, I. (1978). *The Methodology of Scientific Research Programmes.* Cambridge University Press.

Lakens, D., et al. (2022). The replicability of research and the credibility of claims. *Nature Reviews Psychology*, 1(6):366–377.

Lee, M.D., & Vanpaemel, W. (2021). Determining informative priors for cognitive models. *Psychonomic Bulletin & Review*, 28(3):711–731.

Liu, Z., et al. (2025). Bayesian deep learning based bridge condition assessment considering epistemic uncertainty. *Engineering Structures*, Advance online publication.

Marcos, J. (2025). Toward a stronger constraint for non-trivial inconsistent theories. *ResearchGate.*

Maxwell, N. (2024). Non-empirical requirements scientific theories must satisfy. *PhilArchive.*

O'Hagan, A., et al. (2006). *Uncertain Judgements: Eliciting Experts' Probabilities.* Wiley.

Popper, K. (1959). *The Logic of Scientific Discovery.* Basic Books.

Sprenger, J., & Hartmann, S. (2019). *Bayesian Philosophy of Science.* Oxford University Press.

Steel, D. (2022). Epistemic values and the value of science. *Synthese*, 200(2):1–22.

van de Schoot, R., Depaoli, S., King, R., et al. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers*, 1(1):1–26.

Vehtari, A., Gelman, A., & Gabry, J. (2024). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Journal of Machine Learning Research*, 25(1):1–41.

Wang, J., et al. (2026). Bayesian deep learning for probabilistic aquifer vulnerability and uncertainty quantification. *Scientific Reports*, 16(1):32612.

Weisberg, M. (2023). The role of models in scientific theory evaluation. *Philosophy of Science*, 90(1):1–18.

Xiong, G., et al. (2025). Evaluating truthfulness and hallucination in large language models for scientific hypothesis generation. *Proceedings of IJCAI 2025.*

Zhi-Xuan, T., et al. (2025). Understanding epistemic language with a language-augmented Bayesian theory of mind. *Transactions of the Association for Computational Linguistics*, 13:613–637.

A Standardized Evaluation Protocol for Reproducibility

The following protocol enables reproducible application by independent evaluators (human or LLM).

A.1 Protocol Template

```
STRUCTURAL-BAYESIAN THEORY EVALUATION PROTOCOL

THEORY: [Name]
DOMAIN: [Physics / Mathematics / Other]
WEIGHTS: (, , ) = [default: (2, 3, 4)]

=== LAYER I: INTERNAL INTEGRITY ===

A1. LOGICAL CONSISTENCY
    Is the mathematical formalism internally consistent?
    (Interpretive debates do not affect this criterion)
    Score: ___ [0 or 1]

A2. CONCEPTUAL CLARITY
    Are ALL core terms precisely defined without ambiguity?
    Score: ___ [0 or 1]

A3. APPLICATION REPRODUCIBILITY
    Given identical inputs, do independent researchers
    reach identical conclusions?
    Score: ___ [0 or 1]

A4. ABSENCE OF AD HOC ELEMENTS
    Are there auxiliary hypotheses introduced solely to
    rescue failed predictions?
    Score: ___ [0 or 1]

A5. EXPLICIT DOMAIN SPECIFICATION
    Are boundaries of applicability clearly stated?
    Score: ___ [0 or 1]

I = ___ / 5

=== LAYER C: CONSTRAINT POWER ===

B1. MODEL EXCLUSION
    Does the theory exclude at least one internally
    consistent class of models/worlds?
    Score: ___ [0 or 1]

B2. NON-DEFINITIONAL EXCLUSION
    Is the exclusion substantive (not following from
    definitions alone)?
    Score: ___ [0 or 1]

B3. NON-TRIVIAL EXCLUSION
    Is the exclusion not a logical tautology?
    Score: ___ [0 or 1]

C = ___ / 3

=== LAYER O: ONTOLOGICAL LOAD ===

C1. WORLD-DIRECTED CLAIMS
    Does the theory make claims about physical reality,
    not just models or formal systems?
    Score: ___ [0 or 1]

C2. MODAL CONSTRAINT
    Could the world have been otherwise, but the
    theory forbids it?
    Score: ___ [0 or 1]

O = ___ / 2

=== CALCULATION ===
N = 0.4 + 0.2 × C = ___

Raw = ×I + ×C + ×O = ___
Max = ×5 + ×3 + ×2 = ___

S(T) = N × Raw / Max = ___ = ___%

=== RESULT ===

Theory: [Name]
I = ___ | C = ___ | O = ___ | N = ___
S(T) = ___%

Justification (1-2 sentences per layer):
- I:
- C:
- O:
```

A.2 Scoring Clarifications

A1 (Logical Consistency): Score the *formalism*, not interpretations. Quantum mechanics scores 1 because Hilbert space mathematics is rigorous, even though measurement interpretation is debated.

B1 (Model Exclusion): Ask: "What empirically possible world does this theory rule out?" If the answer is "none" or "only logically impossible worlds," score 0.

B2 (Non-Definitional): If the exclusion follows from how terms are defined (e.g., "bachelors are unmarried" excludes married bachelors), score 0.

C2 (Modal Constraint): Ask: "Could physics have been different in a coherent way that this theory forbids?" GR forbids Newtonian absolute space (score 1); a tautology forbids nothing coherent (score 0).

A.3 Expected Inter-Evaluator Agreement

For well-specified theories:

- General Relativity: $> 95\%$ agreement expected
- Quantum Mechanics: $\sim 85\%$ agreement (A2 may vary)
- String Theory: $\sim 90\%$ agreement (landscape well-documented)

Divergences indicate either criterion ambiguity (requiring protocol refinement) or genuine interpretive disagreement about the theory.