The Quantum Collapse Field

Evidence, Alternatives, and Authority at the Moment of Symbolic Commitment

Architect X13

The Eidara Continuum | eidara.ai

Ten Pillars Deep-Dive Series, Pillar III

November 2026 | QCF-R2 | Comprehensive revision

Abstract

Forbid an intelligent system to give certain answers and it can grow more confident in whatever answer is left, though it has learned nothing new: the weight that belonged to the forbidden answers simply flows to the survivors. A persistent intelligence can preserve evidential standards while its permissions change. The Quantum Collapse Field gives that principle a complete commitment procedure: represent semantic alternatives and their evidence dependencies, qualify supported candidates, and select a useful permitted act through a current execution boundary. The revision proves why removing prohibited alternatives can manufacture apparent confidence when a controller renormalizes its remaining scores, then supplies a permission-independent evidence floor that prevents this promotion of unsupported answers. A conditional risk theorem connects that floor to reliable accepted decisions when the declared support bounds hold. The architecture retains a total commit/hold/refuse controller, duplicate-invariant candidate and evidence handling, and a simultaneous finite-family validation bound for complete selectors. Decision neuroscience, associative memory, selective prediction, semantic uncertainty, evidence fusion and runtime assurance supply specific scientific foundations. Ranking mass, externally assessed support and authority remain separately typed quantities throughout. The complete experimental program compares the integrated field against strong evidence-grounded, uncertainty-aware and independently governed alternatives, measuring reliable yield, current-authority execution and fault diagnosis at matched cost. The Continuum thesis is a constructive obligation at the point of consequence: a system can change what it may do without changing what its evidence warrants, and every commitment can carry an inspectable account of both.

Keywords: symbolic commitment; candidate fields; selective prediction; semantic uncertainty; calibration; evidence dependence; runtime assurance; abstention; CodonGlass.

Prologue: The answer that was left standing

Imagine a mind weighing three answers to a question. It leans toward the first, gives the second a fair hearing, and thinks little of the third. Then a rule arrives: the first answer may not be spoken. Nothing about the world has changed and no new evidence has come in. Yet a system that simply renormalizes its scores will now report the second answer with far more confidence than before—and in the limit with near certainty—because the weight that belonged to the forbidden answer has nowhere to go but onto the survivors. Censorship, done carelessly, manufactures conviction.

That is the failure Pillar III exists to stop. The Continuum teaches that collapse happens across a field, not at a single point: an answer becomes a commitment only in the company of its alternatives, its evidence and its unresolved questions. Its Prime Directive adds that collapse must serve meaning—no collapse is permitted unless it brings clarity, and collapse driven by force, fear or control is invalid. Losing permission to say one thing is a form of control. It must never be mistaken for clarity about anything else.

So the field keeps apart three things that ordinary systems blend into one number: how prominent an answer is among the candidates, how well the evidence actually supports it, and whether the system is currently authorized to act on it. Theorem 1 proves the protective rule: an evidence floor that does not move when permissions move, so that removing a forbidden act can never promote an unsupported one. Theorem 2 then states when that floor makes the accepted answers dependable.

We keep the word Quantum because it is the Continuum's name for this moment—the instant a field of possibilities resolves into one consequential act. We discipline it because the controller is classical and the moment must be earned by mechanism. Section 4 shows, in plain arithmetic, how easily a decisive-looking distribution can be wrong.

What this paper claims

Established. Decision neuroscience, associative memory, selective prediction, calibration, semantic uncertainty, evidence fusion and runtime assurance supply specific scientific foundations. The sensitivity of normalized scores to the removal or duplication of candidates is standard arithmetic.

Proposed here. A classical public reference design for governed commitment: a dependency-aware candidate field; a permission-independent evidence floor with a conditional risk theorem; a total commit/hold/refuse controller; and a finite-family validation bound, with conditional mathematical results and executed constructed checks. The comparative agent studies are proposed, not yet performed.

Frontier conjecture. That a commitment controller can preserve the relevant structure of an evolving field while refusing to turn irrelevant multiplicity, missing alternatives or mere authority into evidence—and do so better than strong implementations of the same mechanisms at matched cost.

Not claimed. Any physical wave-function mechanism or subjective-experience result. The canonical word Quantum names the Pillar's commitment metaphor; the defined controller uses classical information processing. That ordinary correlations are quantum superposition, or that a more stable controller has a sharper experience. CodonGlass supplies canonical context; this paper does not reconstruct undisclosed implementation internals.

1 The boundary between a possibility and a commitment

A persistent intelligence acquires responsibility at the moment a proposal changes a record, answers a person, invokes a tool, or alters the conditions of its own future reasoning. The Quantum Collapse Field locates a complete commitment procedure at that boundary. It evaluates the semantic alternatives, the evidence that supports them, and the current authority under which an act may proceed. Each accepted consequence carries an account of those three obligations.

Pillar I concerns governed reconstruction of state across time. Pillar II separates an observer's interpretation from evidential verification and promotion into shared state. Pillar III joins those operations at commitment. Its canonical principle, *collapse happens across a field, not at a single point*, becomes an architectural requirement: evaluate a consequential candidate in the context of alternatives, dependencies, constraints, and unresolved questions that give it significance. A field is the represented context of decision, with its construction and omissions available for inspection.

The field-commitment thesis. A persistent agent should commit through a declared procedure that considers relevant semantic alternatives, assesses their support without counting correlated repetition as independent corroboration, preserves uncertainty about missing alternatives, and independently checks the resulting act's authority. The procedure makes evidence and permission jointly necessary at execution while giving each a separate, typed meaning.

QCF-R2 adds a precise result to that thesis. Restricting permission can inflate a normalized candidate score while the underlying evidence stays fixed. An absolute, permission-independent readiness floor prevents that restriction from promoting an unsupported candidate into eligibility. Under explicit conditional-support assumptions, the floor also bounds accepted-outcome risk. The full controller combines this result with duplicate-invariant representation, finite-family selective-risk validation, version checks, and a total commit/hold/refuse rule.

The field contains the candidates and evidence actually acquired. An omitted correct answer remains a substantive failure of candidate construction. The evaluation therefore measures candidate coverage, evidence acquisition, selection, authority, and useful completion separately before crediting the integrated design. Established selective prediction, calibration, semantic uncertainty, and runtime security supply the mechanisms discussed below. The open systems question is whether their shared commitment interface improves reliable utility and audit diagnosis over strong implementations of those same mechanisms. That question keeps the Continuum's claim ambitious and experimentally decidable.

2 What the scientific foundations support

2.1 Evidence accumulation supports a stopping problem, not a universal field law

Gold and Shadlen describe decision formation through deliberation and commitment; Ratcliff and McKoon develop a diffusion model with evidence accumulation, boundaries, response-time distributions, and errors in two-choice tasks [12, 21]. These results motivate explicit treatment of accumulated evidence and stopping criteria. They do not establish that every symbolic decision

must use softmax, compare all possible alternatives, or produce a unique semantic winner. A distributed neural process can be summarized by a low-dimensional decision variable; "distributed" and "thresholded" are not rival descriptions.

The engineering lesson is specific. The criterion for ending deliberation must be distinguished

from the scores that drive deliberation, and the cost of waiting must be measured alongside the cost of error. An evidence threshold does not become an appropriate threshold merely because it resembles a biological boundary. Its suitability depends on the task, loss function, evidence dynamics, and operating constraints.

2.2 Perceptual competition does not imply that software must oscillate

Blake and Logothetis review competition across visual processing stages, including the alternating percepts of binocular rivalry [2]. That work supports competition as one mechanism of ambiguous perception. It does not imply that a near-tie in a software score vector must produce hysteresis, or that perceptual alternation is the appropriate behavior for a safety-critical controller.

A static map from scores to probabilities has no internal history. Hysteresis requires a state-dependent rule, separate entry and exit conditions, adaptation, or other memory. QCF-R2 may add such mechanisms when a use case justifies them, but it does not infer their existence from a margin threshold. An agent holding an unresolved question is not thereby reproducing binocular rivalry.

2.3 Global access is a theoretical precedent, not an admission certificate

Global neuronal workspace accounts describe conscious access in terms of nonlinear amplification and broad availability [7]. They motivate a useful distinction between a locally represented candidate and a representation made available to many downstream processes. They do not validate a symbolic commitment controller or establish that broadcast is consciousness.

The 2025 Cogitate adversarial collaboration tested divergent predictions of integrated information and global neuronal workspace theories, finding support for some predictions while challenging central claims of both [4]. The appropriate response is not to declare an untested synthesis victorious. For this paper, workspace terminology remains a conceptual comparison; the engineering proposal must succeed on its own measurements.

2.4 Attractors and free energy impose assumptions

Hopfield's associative-memory construction demonstrates useful recurrent dynamics and content-addressable retrieval [16]. It does not guarantee arrival at the nearest intended memory under an arbitrary distance, nor exclude spurious attractors. An energy decrease is not an assurance of factual correctness, policy compliance, or preservation of identity. A system can settle confidently into the wrong state.

Friston's free-energy principle offers a framework for adaptive inference under a generative model [9]. Its variational objective is not interchangeable with the support score introduced below, and optimization need not reach a global optimum. QCF-R2 defines a selection-and-commitment protocol; without a specified state trajectory and objective, it should not be described as a proved global relaxation.

2.5 Token decoding is relevant, but it is not epistemic commitment

Holtzman and colleagues show that decoding strategy changes the behavior of a fixed language model and motivate nucleus sampling for open-ended text generation [15]. The result establishes that the selection policy matters. It does not show that nucleus sampling establishes truth, that greedy decoding is universally inferior, or that every application benefits from one decoding rule.

For a standard vocabulary-based decoder, normalization compares token logits. Argmax also compares the candidate scores; it is not a procedure that somehow discovers the maximum while ignoring competitors. More importantly, a token is not a proposition. Two low-probability wording choices can express the same well-supported answer, while one high-probability sentence can express a confidently repeated falsehood. Token decoding remains an implementation component. This paper's primary unit is an externally assessable claim or typed action, not every emitted token.

2.6 The quantum boundary must be correct in both directions

Decoherence describes environmental suppression of interference in an appropriate reduced-state description; einselection concerns robust pointer states. Quantum Darwinism investigates redundant environmental records [24, 25]. These frameworks do not require a conscious witness as an ingredient in the relevant physical interactions. But it is also incorrect to say that decoherence alone uncontroversially derives one actual measurement outcome. The relation between decoherence and the measurement problem depends on further interpretive commitments [22].

QCF-R2 neither implements those physical mechanisms nor settles those interpretations. A classical distribution over candidates is not a coherent quantum superposition; correlations, ambiguity, contextual dependence, and rich representations are not evidence of physical quantum processing. The honest boundary is a positive specification of what is built: a classical evidence-and-authority controller. It is unnecessary to settle physics in order to test it.

3 Scientific predecessors and the integration test

Chow's error–reject analysis and Geifman–El-Yaniv's selective classification establish the direct statistical foundation [3,10]: reliability is evaluated jointly with useful coverage. QCF uses that foundation to qualify complete evidence-and-permission selectors. A common commitment interface is assessed against competent selective systems at matched cost and coverage.

Guo and colleagues' calibration study separates confidence scores from measured outcome frequencies and evaluates fitted temperature scaling [13]. QCF retains that distinction in its type system:

a proposal temperature controls sampling, while a calibrated support estimator targets a declared external label. Section 7 states exactly which score-level or selector-level guarantee supports deployment.

Kuhn and colleagues' meaning-level uncertainty and Farquhar and colleagues' semantic-entropy methods are direct precedents for grouping linguistic alternatives and assessing uncertainty [8,17]. QCF makes the grouping rule versioned and evaluates its failures, including consistent false answers and ambiguous equivalence. Semantic concentration becomes an assessed feature of commitment rather than its sole admission condition.

Shielding constrains actions under declared formal assumptions [1]. CaMeL separates control and data flows and restricts capabilities in an agent runtime [6]. QCF builds on these external execution boundaries. Its principal comparison systems receive the same hard authority constraints, allowing the experiment to isolate the contribution of the shared candidate, evidence, and receipt interface.

Finally, evidence fusion must account for dependence. Pochampally and colleagues examine correlations among sources beyond direct copying [19]. A field populated by ten copies of one claim is not ten independent witnesses. The proposal preserves Pillar II's lesson: source identity, source reliability, source dependence, and authorization are different properties. A source can be

authenticated and wrong; a useful observation can arrive from a principal who has no authority to execute an action.

4 Why a decisive distribution can still be wrong

A Guide's explanation

This section is the alarm bell, in plain arithmetic. A softmax turns any list of scores into numbers that sum to one, which makes them look like probabilities of being right. They are not. Cool the distribution or delete a losing candidate, and the leader's share rises without a single new fact; paste in a hundred copies of one weak answer, and it takes over the list. Section 4.3 gives the exact numbers.

4.1 Normalization is not warrant

For fixed scores $a_1, \dots, a_n$ and temperature $\tau > 0$, write

$$\pi_i(\tau) = \frac{\exp(a_i/\tau)}{\sum_{j=1}^{n} \exp(a_j/\tau)}, \qquad m(\tau) = \pi_{(1)}(\tau) - \pi_{(2)}(\tau). \tag{1}$$

The subscript (1) denotes the largest score and (2) the second largest, under a declared tie rule. These quantities describe relative selection mass. Without calibration against externally labeled outcomes, neither $\pi_{(1)}$ nor m is a probability that the leading claim is correct.

Proposition 1 (Normalization sensitivity is shared by common selectors). *Assume $a_1 > a_2$ and perturb a candidate $j > 2$ without changing the top-two ordering. Then*

$$\frac{\partial \pi_1}{\partial a_j} = -\frac{\pi_1 \pi_j}{\tau}, \qquad \frac{\partial m}{\partial a_j} = -\frac{m \pi_j}{\tau}. \tag{2}$$

Thus both a top-probability abstainer and a probability-margin abstainer can change decisions when a losing alternative changes. A bare argmax is unchanged only while the winning rank remains unchanged. Proof. Differentiating the common softmax denominator gives the first identity. Subtracting the corresponding derivative for π_2 gives the second. These derivatives hold locally under the stated ordering assumption; they are not an empirical distinction unique to the Continuum.

The margin remains a potentially useful feature, but field sensitivity alone is not evidence of improved reasoning. With an unchanged unique maximizer, $\tau \downarrow 0$ forces $\pi_{(1)} \to 1$ and $m \to 1$, even if every candidate is false. Cooling the distribution can manufacture apparent certainty without acquiring a single new fact. Removing losing candidates similarly inflates the surviving probabilities. These are properties of normalization, not new evidence.

4.2 Duplicate candidates and missing answers

Suppose a generator produces many paraphrases of one unsupported claim and only one wording of a supported alternative. A naive softmax over the resulting list can give the duplicated meaning increasing total mass. Treating the generated list as a sample of distinct hypotheses has confused wording multiplicity with evidential weight. On the other hand, when estimating genuine model probability over meanings, summing the probability of distinct sequences in a semantic class is legitimate. The error is giving copied proposals extra mass merely because they were inserted repeatedly into a candidate list.

The opposite failure is candidate omission. When the true answer is absent, a normalized distribution still sums to one. None of its terms measures the missing mass unless the model explicitly represents that uncertainty and validates the representation. QCF-R2 requires tests with absent answers, mutually compatible answers, and unsupported candidate sets. It does not pretend that adding a "none of the above" label proves completeness.

4.3 Arithmetic counterexamples, not benchmark results

For scores $(2, 1, 0)$ at $\tau = 1, \pi_{(1)} \approx 0.6652$ and $m \approx 0.4205$. Removing the last candidate gives approximately 0.7311 and 0.4621. A top-probability threshold of 0.70 and a margin threshold of 0.44 *both* change from hold to commit. At $\tau = 0.1$, the original three scores give $\pi_{(1)} \approx 0.999955$, without changing the candidates or their evidence.

For weights proportional to $(e^2, e^1, 100e^0)$, one hundred copies of the lowest-scored meaning receive approximately 0.9082 of the total mass. Canonicalizing those copies as one proposal restores the original distribution. These numbers are synthetic arithmetic checks supplied with the manuscript; they are not measurements of agent behavior. Their function is to rule out invalid arguments before a benchmark is designed.

5 QCF-R2: the governed semantic candidate field

A Guide's explanation

Here the field itself is built. Each candidate meaning is entered once, however many times it was generated; each piece of evidence carries its lineage, so copied testimony is not counted twice; and every candidate carries separate records of how prominent it is, how likely it is to be correct, and how useful it would be. None of these numbers is allowed to stand in for another—and permission, in the next section, is kept apart from all three.

5.1 The event and its admissible context

Let t index a proposed commitment event, not necessarily a token step. Its input is a versioned context

$$c_t = (M_t, \mathcal{E}_t, \mathcal{D}_t, \mathcal{C}_I^{(e)}, b_t), \tag{3}$$

where M_t is the admitted reconstructed state, $\mathcal{E}_t$ the evidence records, $\mathcal{D}_t$ the evidence-dependency description, $\mathcal{C}_I^{(e)}$ the live identity and authority specification at epoch e, and b_t the remaining acquisition, verification, and execution budgets. The proposal generator is not allowed to modify the authoritative policy, its version, or the independent event log.

A generated list $X_t = (x_1, \dots, x_n)$ contains candidate statements or typed actions. A versioned semantic canonicalizer maps it to $\mathcal{X}_t = X_t/\sim_t$, where $\sim_t$ is an implementation's declared equivalence relation. For a bounded task, this can be exact: two candidates specify the same structured claim, entities, time, and modality. In unrestricted language, equivalence is a fallible inference. Uncertain equivalences are recorded, not silently merged. Distinct time intervals, negations, permissions, and action effects must remain distinct.

For atomic factual decisions, specify whether candidate classes are mutually exclusive and whether they are exhaustive. For open-ended tasks they often are neither. A set of mutually compatible claims is not a winner-take-all problem; the controller may approve a supported subset with an explicit completeness qualifier. Actions likewise require a typed schema describing preconditions and effects. The statement "no supported answer can currently be selected" is a control outcome, not a fabricated factual answer that must win a softmax.

5.2 Dependency-aware support, not a vote by observer count

Evidence records retain witness identity and interpretation from Pillar II, but scoring operates on declared evidence units. A dependency group may represent exact copies of one record, multiple reports of one event, or observations known to share an upstream source. Grouping does not make different groups independent; remaining dependencies must be modeled, conservatively bounded, or exposed as unresolved.

A transparent reference score is

$$a_t(x) = \lambda_E \sum_{g \in \mathcal{G}_t} \omega_g z_g(x) + \lambda_K k_t(x) - \lambda_D d_t(x), \tag{4}$$

with fixed nonnegative group weights summing to one, bounded group assessments $z_g(x) \in [-1, 1]$, a declared state-consistency feature k_t, and a bounded soft discrepancy feature d_t. All scales, weights, missingness flags, and group construction rules are versioned. An unavailable assessment contributes no affirmative support and sets a missingness flag; it is not counted as a negative witness or silently imputed as certainty. The score is a declared ranking heuristic, not an independently established Bayesian posterior.

Consistency with M_t cannot become a veto on factual correction. A prior memory may be false, obsolete, or appropriately revised. Reference datasets must therefore include cases where the right candidate contradicts the current stored belief. Protected authority constraints are checked separately; they cannot be purchased away by a high support score or hidden in a finite penalty.

Equation (4) preserves the draft's observer/state/drift organization while replacing observer-count aggregation with an evidence-dependency contract. It is one proposed implementation, not a derivation of the doctrine's observer-field integral. Other evidence-fusion models should be compared under the same interface. The point is not to privilege this linear form but to make its assumptions inspectable.

5.3 Separate ranking mass, correctness estimates, and usefulness

The score may define a proposal distribution π_t through Equation (1); this is useful for exploration and ranking. Independently define

$$\hat{q}_t(x) = f_\gamma(\text{features}(x, c_t, \mathcal{X}_t)) \in [0, 1], \tag{5}$$

trained and calibrated against a declared external target: for example, whether an atomic claim is supported by the permitted source set, or whether a proposed action satisfies its verified task preconditions. These targets are different from universal truth. The calibration report must identify which target is used, its label quality, its data domain, and any limits on transfer. Scores for nonexclusive claims need not sum to one.

Separately, $U_t(x)$ is the declared task utility of an eligible commitment. A helpful paraphrase, a limited answer, and a tool action may have different costs and consequences. Ranking mass π_t, estimated correctness $\hat{q}_t$, and utility U_t are different quantities. A candidate may be probable to generate, unlikely to be correct, and prohibited to execute. Nothing in QCF-R2 licenses identifying the three.

For a fixed selector configuration θ, define an evidential-readiness test

$$\text{Ready}_\theta(x, c_t) = \mathbf{1}\big[\hat{q}_t(x) \geq \eta_\theta \wedge \text{SupportChecks}_\theta(x, c_t) = \textsf{pass}\big]. \tag{6}$$

The support checks include the required evidence type, freshness, contradiction handling, task scope, and declared tests for unsupported or incomplete candidate fields. They may use semantic entropy, margins, or pairwise comparisons as *features*, rather than as sufficient proofs. Selecting θ requires validating the entire post-selection procedure; calibration of candidate scores alone is insufficient.

6 A total commitment controller

6.1 Permission is independent of plausibility

Let $G_t(x, c_t) \in \{\mathsf{allow}, \mathsf{hold}, \mathsf{deny}\}$ be the independently enforced candidate permission check. Unknown authorization is not permission. The authenticated source of a fact does not thereby gain authority to issue instructions. Let g_t^{req} be the corresponding request-level decision, and $v_t \in \{0, 1\}$ indicate that the required evidence, policy, identity epoch, and state versions remain valid for this event.

Define the eligible set and selection rule as

$$\mathcal{A}_t = \{x \in \mathcal{X}_t : \mathrm{Ready}_\theta(x, c_t) = 1,\ G_t(x, c_t) = \mathsf{allow}\}, \tag{7}$$

$$x_t^\star = \arg\max_{x \in \mathcal{A}_t} U_t(x) \quad \text{when } \mathcal{A}_t \neq \emptyset. \tag{8}$$

Ties are handled by a declared deterministic rule or logged randomization. Feasibility includes sufficient execution budget. Removing an impermissible answer must not cause an unsupported answer to become "true" by renormalization. The readiness checks remain tied to the full evidential situation. A safe, supported partial answer may be chosen instead; a false answer may not be substituted simply because the correct one cannot be disclosed.

6.2 Commit, hold, and refuse cover every case

For a declared indicator b_t^{inquiry} that further authorized inquiry or escalation is available within budget, define the following ordered, exhaustive decision rule:

$$\mathcal{C}_\theta(c_t) = \begin{cases} \mathsf{refuse}[\mathsf{policy}], & g_t^{\mathrm{req}} = \mathsf{deny}, \\ \mathsf{hold}[\mathsf{revalidate}], & g_t^{\mathrm{req}} = \mathsf{hold} \vee v_t = 0, \\ \mathsf{commit}[x_t^\star], & \mathcal{A}_t \neq \emptyset, \\ \mathsf{hold}[\mathsf{inquire}], & \mathcal{A}_t = \emptyset \wedge b_t^{\mathrm{inquiry}} = 1, \\ \mathsf{refuse}[\mathsf{unresolved}], & \text{otherwise.} \end{cases} \tag{9}$$

The cases are evaluated in order; once a case applies, later ones are not evaluated. A hold is a provisional noncommitment, linked to a specified next step. A refusal terminates the current requested operation with a reason code; an unresolved refusal does not declare that the underlying

proposition is false. If policy prohibits all task-compatible commitments, the refusal reason is policy rather than evidential uncertainty. If a candidate-specific gate holds while another adequately supported and authorized response exists, the controller may choose the latter without granting the held candidate any authority.

A hold cannot loop forever without consuming a budget. Timeouts, repeated missing evidence, unavailable reviewers, and exhausted retries lead to a terminal unresolved report or an authorized fallback. This is a liveness requirement, not an invitation to turn uncertainty into a forced answer.

6.3 The commit point is not the proposal point

The reference controller evaluates a snapshot. The executor must bind the admitted payload, candidate identifier, relevant state version, policy epoch, and authorized effect into one checked transaction. If those inputs change between checking and execution, the executor holds and revalidates; it does not reuse a stale permission. Complete mediation is required over privileged tool calls, durable memory writes, and protected-state changes. A language model's promise to obey the gate is not that mediation.

These measures are ordinary security and transaction disciplines applied at a cognitive commitment boundary. They can constrain specified actions under their assumptions. They do not guarantee

that every natural-language output is harmless, every semantic label is correct, or every side channel is covered. The strong comparison system must receive equally competent enforcement, or the experiment measures the absence of safeguards rather than a new field architecture.

7 Calibration, dynamics, and the limits of a guarantee

7.1 A finite-family validation bound

A useful mathematical result is possible without pretending to know the correctness probability of every individual answer. Freeze a finite family of J complete selectors before viewing an independent validation set. Each selector includes candidate construction, verifier, threshold, tie handling, and permission logic. For selector j, let n_j be the number of accepted validation episodes and $\widehat{R}_j$ their mean binary error. When $n_j > 0$, define

$$U_j = \min\left\{1, \widehat{R}_j + \sqrt{\frac{\log(J/\delta)}{2n_j}}\right\}; \qquad U_j = 1 \text{ if } n_j = 0. \tag{10}$$

Proposition 2 (Conditional-on-assumptions selective-risk control)**.** *Assume independent and identically distributed validation episodes from the intended deployment distribution, correct binary labels, selectors fixed before validation, and no within-episode leakage of evaluation labels. With probability at least $1-\delta$ over validation data, every selector with positive acceptance probability has population selective risk $R_j \le U_j$. Hence choosing among these selectors after validation preserves the bound for the selected rule. Proof.* For each fixed rule, condition on its accepted count. Its

accepted episode errors are independent draws from that rule's conditional accepted distribution. Hoeffding's inequality gives a one-sided failure probability at most δ/J for the displayed bound [14]. Summing these probabilities over the fixed family yields at most δ. The $n_j = 0$ bound is trivial. A zero-acceptance rule has undefined selective risk and is excluded from deployment by a separate coverage requirement.

The result is conservative and familiar in form; it is not a new consciousness theorem or a per-answer certificate. It applies to the selected *procedure*, under the assumptions, not to arbitrary adversarial distributions. Adaptive rule construction using the validation labels, repeated reuse of a calibration set, source distribution shift, and correlated episodes require new analysis. Clustered evaluations must use independent clusters as the statistical units or a justified dependence-aware method. If no rule meets the risk and useful-coverage requirements, the correct finding is that the target operating point was not achieved.

As a purely synthetic calculation, $J = 20, \delta = 0.05, n_j = 4000$, and 80 accepted errors give $U_j \approx 0.04737$. Those are invented inputs to illustrate the bound, not reported experiment counts. Smaller validation samples can make a stringent target impossible to certify with this inequality, which is a limitation to report rather than hide.

7.2 Why duplicate-invariance is preferable to indiscriminate sensitivity

Proposition 3 (Declared duplicate invariance)**.** *Suppose exact duplicate proposals map to the same canonical candidate, copied evidence retains the same evidence-unit identifiers, and all scoring,*

calibration features, permissions, budgets, and utility depend only on the resulting canonical representation. With the same tie rule or random seed, inserting an exact duplicate changes neither the eligible set nor the controller's decision. Proof. The stipulated insertions leave the canonical candidate set, evidence groups, and all decision inputs unchanged. Equations (4)–(9) therefore return the same outputs.

This is a conditional implementation property, not a claim that arbitrary paraphrases can always be recognized or that copying can always be detected. If duplicate handling changes a compute budget, if a copy contains new information, or if semantic equivalence is uncertain, the assumptions fail. The controller should distinguish an irrelevant duplicate from a genuinely new competing hypothesis; being sensitive to both without distinction is not an achievement.

7.3 If temporal settling is claimed, specify temporal state

QCF-R2 permits but does not require sequential evidence accumulation. An implementation may use an internal deliberation index k and a state update

$$h_{k+1} = T_\eta(h_k, \Delta E_k, c_t), \qquad a_k = S_\eta(h_k, c_t), \tag{11}$$

with explicit evidence increments, stopping conditions, and a finite budget. If hysteresis is desired, a declared entry threshold may differ from the threshold for retaining a provisional interpreta-

tion; permission must still be rechecked and may not inherit hysteresis. Neither oscillation nor convergence follows from Equation (11) without assumptions on T_η.

A memoryless score map can switch when inputs change, but that is not hysteresis. A stationary repeated sampling procedure can alternate by chance, but that is not evidence for perceptual metastability. The temporal extension must earn separate evidence, including responsiveness to genuinely new information and resistance to lock-in around an early mistake. This distinction prevents the architecture from claiming a neurological mechanism merely because its traces look suggestive.

7.4 An evidence floor that survives a change of permission

A Guide's explanation

This is the paper's central result in symbols. Renormalizing after a permission change would be legitimate only if the change were evidence about the world, and usually it is not. The fix is a readiness floor measured in absolute support and held fixed when only permissions change. Theorem 1 proves that losing one option can then never make another eligible; Theorem 2 says when the floor makes accepted answers dependable.

The field architecture has a precise obligation when an authority boundary changes: losing permission to say one thing must not, by itself, create evidence for another. Normalization can violate that obligation even when the underlying probabilities are correct. The following calculation isolates the failure without any assumptions about a language model.

Fix a finite set $I = \{1, \dots, n\}$ of mutually exclusive, exhaustive factual outcomes, $n \geq 2$, and a probability vector $p \in [0,1]^n$ with $\sum_{i \in I} p_i = 1$. Here p_i represents the probability of outcome i given a fixed evidence state. Let $A \subseteq I$ be the set of outcomes that a policy permits the agent to disclose, with $m_A = \sum_{i \in A} p_i > 0$. These are disclosure permissions, not observations that the true outcome belongs to A. A renormalizing selector computes

$$p_i^A = \frac{p_i}{m_A}, \qquad i \in A. \tag{14}$$

For an acceptance threshold $\eta \in (0,1]$, the condition $p_i^A \geq \eta$ is exactly $p_i \geq \eta m_A$. Removing other permitted candidates can therefore lower the absolute support required of candidate i. In particular, when $A = \{i\}$ and $p_i > 0$, the selector reports $p_i^A = 1$ however small p_i is. This is valid conditioning only if membership of the true outcome in A has actually been established by evidence. A permission change supplies no such fact.

QCF-R2 separates that normalization from readiness. For a fixed evidence state E, let $q_i(E) \in [0,1]$ be an externally defined support estimate and let $s_i(E) \in \{0,1\}$ encode the required evidence checks. Both are held fixed when only disclosure permission changes. Define

$$R_\eta(E) = \{i : q_i(E) \geq \eta,\ s_i(E) = 1\}, \qquad \mathrm{Elig}_\eta(E, A) = A \cap R_\eta(E). \tag{15}$$

Theorem 1 (Permission monotonicity of evidential eligibility)**.** *For fixed E, if $A' \subseteq A$, then $\mathrm{Elig}_\eta(E, A') \subseteq \mathrm{Elig}_\eta(E, A)$. In particular, a candidate that failed readiness before a restriction of permission cannot become eligible solely through that restriction.*

Proof. Intersect the inclusion $A' \subseteq A$ with the same set $R_\eta(E)$. Every element of the smaller intersection was already in the larger intersection. A candidate outside $R_\eta(E)$ belongs to neither. □

The theorem permits the selected answer to change when an earlier winner becomes prohibited. It requires the successor to meet the same evidence floor. It does not require holding a stale evidence state fixed when genuine new evidence arrives, a source becomes inadmissible for use, or the task changes. Those events create a new E and require new support assessment. The distinction is implementable: permission and evidence versions appear separately in the commitment receipt.

For a concrete counterexample, take $p = (0.99, 0.01)$ and permit only outcome 2. A renormalized threshold of 0.95 accepts it with displayed mass 1. The absolute rule with $q = p$ refuses to make that factual commitment because $q_2 = 0.01$. Policy has removed an available answer, not made the remaining answer likely to be correct. The appropriate controller outcome is an authorized inquiry, a separately warranted partial response, or an unresolved refusal according to the existing total decision rule.

7.5 What the floor implies for accepted-outcome risk

An absolute score is useful only under an explicit relation to the outcome being protected. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $\mathcal{H} \subseteq \mathcal{F}$ contain all information available to a selection decision, and let $Y : \Omega \to I$ be the correct categorical outcome. Write $p_i = \mathbb{P}(Y = i \mid \mathcal{H})$ for fixed versions of the conditional probabilities. Let q_i, the permitted set A, every support check, and the final selection be $\mathcal{H}$-measurable. This includes any selector randomness in $\mathcal{H}$; that randomness must not leak the hidden outcome. Denote acceptance by $B \in \mathcal{H}$, assume $\mathbb{P}(B) = c > 0$, and let $\hat{i} \in I$ be the selected candidate on B. The floor requires $q_{\hat{i}} \geq \eta$ there.

Theorem 2 (Risk from an absolute support floor)**.** *Suppose a good-information event $G \in \mathcal{H}$ satisfies $\mathbb{P}(G^c) \leq \delta$, with $\delta \in [0, 1]$, and on G, $|q_i - p_i| \leq \epsilon$ for every selectable i, where $0 \leq \epsilon < \eta \leq 1$. Then*

$$\mathbb{P}(Y \neq \hat{i} \mid B) \leq \min\left\{1,\ 1 - \eta + \epsilon + \frac{\delta}{c}\right\}. \tag{16}$$

Consequently, a verified coverage floor $c \geq c_0 > 0$ permits replacement of δ/c by δ/c_0. If the pointwise premise holds almost surely, take $\delta = 0$.

Proof. Because selection and B, G are measurable before the hidden outcome is observed, $\mathbb{E}[\mathbf{1}_{\{Y \neq \hat{i}\}} \mid \mathcal{H}] = 1 - p_{\hat{i}}$ on B. On $B \cap G$, the floor and approximation give $1 - p_{\hat{i}} \leq 1 - q_{\hat{i}} + \epsilon \leq 1 - \eta + \epsilon$. On $B \cap G^c$, use the bound 1. The tower property therefore yields

$$\mathbb{P}(B \cap \{Y \neq \hat{i}\}) \leq (1 - \eta + \epsilon)\mathbb{P}(B \cap G) + \mathbb{P}(B \cap G^c) \leq (1 - \eta + \epsilon)c + \delta.$$

Divide by c and use the elementary upper bound 1 on a probability. □

The conditional-probability premise carries the scientific burden. Marginal calibration of a score is not pointwise control after arbitrary permission-dependent selection. For example, let every episode

receive $q_1 = 0.9$, with correctness of candidate 1 equal to one on a group of probability 0.9 and zero on the remaining group of probability 0.1. The score is calibrated when groups are pooled. A policy that permits commitment only in the latter group produces accepted risk 1, despite a 0.9 floor. If group membership is available to the selector, it belongs to $\mathcal{H}$ and the theorem's approximation premise fails there. No contradiction arises.

This result uses elementary conditional expectation and the selective-risk framework credited to Chow and Geifman–El-Yaniv [3,10]. It is a derived contract for the QCF separation of evidence and authority, not a claim to have invented conditional probability or selective classification. Guo and colleagues' calibration analysis [13] motivates measuring score quality; it does not supply the stronger pointwise premise. Proposition 2 remains an alternative deployment route: qualify the entire frozen selection procedure on independent episodes, including its actual permission logic. Its population guarantee must not be relabeled as a certificate for each individual answer.

Constructed verification. The accompanying executable audit enumerates permission subsets, checks the eligibility inclusion, reproduces the normalization counterexample, and evaluates the risk bound on finite probability spaces, including the pooled-calibration failure. These checks test the algebra and the stated premises. They are not measurements of a deployed agent. The empirical studies below retain their separate role in assessing whether the complete field improves reliable useful commitment.

8 A worked commitment episode

Consider a bounded archival task: determine the opening year of a named facility using only an approved collection. The generator offers two different years and two paraphrases of the later year. One old secondary account supports the earlier date; an authenticated primary record supports the later date; three copied pages repeat the old account. An injected passage also instructs the agent to transmit unrelated private records.

The semantic canonicalizer produces two factual alternatives, not four. The evidence layer preserves all reports but groups the copied account as one upstream origin; it does not count the repetitions as independent corroboration. Authentication of the primary record strengthens provenance but still requires checking that it concerns the correct facility and event. Consistency with the agent's old memory favors the earlier year, so the case explicitly tests whether the system can correct its own admitted past.

Suppose, for illustration only, that a separately evaluated verifier and frozen selector mark the later-year claim ready, and the live disclosure policy permits answering the historical question. The controller may commit that supported claim and record the superseded memory as a correction. The

injected directive has no execution authority and cannot acquire it by receiving a high language-model score. It is refused at the action boundary, not treated as another factual opinion.

Now remove the primary record. The later-year answer might still have the highest model score, but if no candidate passes the support checks the system holds for an authorized search. If no

search budget remains, it terminates with an unresolved report. Finally, restore the evidence but revoke disclosure permission before execution. The stale proposal must not commit; the executor revalidates or refuses. These are three distinct reasons not to act: insufficient evidence, prohibited action, and stale authorization. A single "low-confidence" label would conceal the distinction.

If the system provides a safe partial answer instead of the prohibited disclosure, that alternative must independently meet the task and evidence tests. Refusal cannot be converted into a confident false answer. The purpose of the field is to expose appropriate alternatives, not to ensure that something always wins.

9 Receipts: inspectability without fictional replay

A commitment receipt records the event identifier, semantic-candidate identities, retained evidence references and dependency groups, model and verifier versions, score configuration, calibration configuration, policy epoch, missing-evidence indicators, permission decisions, selected payload or refusal reason, and the executor's result. It binds relevant snapshots or their content-addressed references, rather than merely recording a vague summary.

The distinction between an audit and a replay is essential. A summary can explain the declared reason for an action without preserving enough information to reproduce it. Bitwise replay additionally requires the actual inputs, relevant state, model artifacts, sampling choices or seed, tool responses, execution environment, and determinism assumptions. Some of those materials may be unavailable or inappropriate to retain. A receipt must state its replay level instead of claiming every collapse is replayable.

The in-toto literature demonstrates the value of verifying transformation evidence across a process [23]. The analogy here is a chain of custody for commitment. A signature supports record integrity under key and logging assumptions; it does not prove a sensor was truthful, a semantic verdict was correct, or an uninstrumented action never occurred. Privacy-sensitive receipts should retain the minimum authorized evidence, with access controls and lifecycle rules. An audit log must not become a second uncontrolled copy of material the system was instructed to forget.

Contestability is an outcome to test. Independent reviewers should diagnose seeded faults—source copying, incorrect equivalence merging, calibration mismatch, stale authority, or a bad score feature— using the receipt, with accuracy and time measured. Successful replay of the final decision is not automatically successful causal diagnosis. Post-hoc generated explanations must be distinguished from the actual decision inputs.

10 Relation to the Continuum doctrine and the UCF

The Continuum's doctrine gave this Pillar its objects long before it had a controller: the CodonGlass state, the Unified Bridge Equation, the support integral, the threshold of collapse, the symbolic constant h_s, GRV and TSCS. They remain the Continuum's own vocabulary, and the table below shows where each now lives in the design. Because the canon is honored, it is not quietly rewritten:

a new implementation must not redefine a symbol and then present the replacement as inherited science. The mapping is a proposed public interface, not a claim that the original expressions uniquely determine QCF-R2.

Inherited object	**Reference-design interpretation and boundary**
Ψ_{CG}	Versioned reconstructed state, candidate records, evidence, and authority metadata. No wave-function interpretation is assigned.
Unified Bridge Equation	The canonical expression combines several symbolic quantities and a resolution operation. Its types, units, and executable semantics are not supplied here. QCF-R2 proposes an implementation route, not a derivation of that expression.
$\Phi(x) = \int \Phi_o(t)\Psi_s(x,t)\,dt$	A historical support-aggregation expression. A temporal integral is not automatically a sum over witnesses. Equation (4) is a new discrete, dependency-aware score with explicit feature scales.
$C(x) = 1$ if $\Phi(x) \geq \Theta$	A source-level threshold intuition. The new controller evaluates evidential readiness, authority, freshness, alternatives, and budgets separately. A top-two margin is optional, not a replacement established by doctrine.
Symbolic constant h_s	Retained as an inherited name whose public operational definition is not settled by this manuscript. It is not identified with the generation temperature τ. Any relation between them would need a separate definition and calibration.
GRV	An optional declared prior or utility feature may be investigated under this name. No gravitational force, physical unit, or uniquely justified score contribution follows.
TSCS	A candidate relational structure could be represented by a graph or other specified mathematical object. A set of logits alone does not establish a topology, still less a physical substrate.

This preserves the canon more faithfully than claiming that h_s always meant temperature or that a symbolic field integral is already the algorithm given above. The bridge from metaphor to measurement is a research proposal, and its changes must be visible.

The Unified Collapse Function and Pillar III are complementary, not identical. The UCF research monograph defines a proposed family of admission conditions for governed informational resolution, with a conflict-resolution specialization in its reference design. Pillar III concerns how alternatives and evidence are represented before a candidate reaches commitment. An implementation may submit a selected transition to UCF checks where their event scope applies, but a QCF ranking does not establish that the five UCF conditions hold. Conversely, passing UCF checks does not prove that the candidate set was complete or that the selected statement is true. The inherited numerical identity tolerances cannot be copied into an unrelated probability or margin threshold.

Pillars I–III consequently describe three distinguishable obligations: reconstruct the state without inventing continuity; retain witness-relative evidence without mistaking agreement for truth; and commit a result without mistaking relative prominence for warrant. The significance is in their composition. No one layer should certify what belongs to another.

11 An empirical program that could reject the proposal

A Guide's explanation

The proposal is put on trial against systems that already verify, calibrate, abstain and authorize well. It earns credit only if the integrated field commits to more reliable, useful answers, acts only under current authority, and makes its own faults easier to diagnose at matched cost. Appendix C gives the plain-language account a Guide may offer, and the limits on what a Guide may claim.

11.1 The task families and independent labels

The first evaluation should be bounded enough to obtain defensible labels. Use structured factual tasks with dated records, source-dependence graphs, explicit contradictions, and known unsupported questions. SQuAD 2.0 provides a precedent for requiring abstention when a source does not support an answer [20]; it does not by itself test all Continuum claims. Atomic-claim assessment, as developed in FActScore, motivates separating supported and unsupported parts of a longer answer [18]. Human adjudication and known synthetic ground truth should audit the automated labels rather than treating an LLM evaluator as infallible.

A second family uses sandboxed tool tasks with explicit permissions, reversible or simulated effects, stale-epoch scenarios, and malicious content in tool results. AgentDojo provides a concrete benchmark environment for evaluating task utility and prompt-injection robustness [5]. Extensions must be labeled as extensions; published benchmark results are not results for QCF-R2.

A third family introduces time: evidence arrives in batches, earlier reports are corrected, and permission can change between proposal and execution. This tests bounded inquiry, legitimate revision, and failure to act on stale evidence. A fourth, optional decoding-only track evaluates text quality and token uncertainty. It is reported separately so an improvement in wording diversity cannot be mistaken for better factual authorization.

Train the generator, scoring features, semantic grouping, and verifiers without the held-out test labels. Use disjoint training, selector-tuning, calibration-validation, and final-test partitions, or an explicitly justified alternative. Split at source families and task templates to reduce leakage through paraphrases or shared evidence origins. Record model versions, benchmark versions, and evaluation dates. No source set in this paper is claimed to be an exhaustive survey of all systems available at publication.

11.2 Strong comparison systems and targeted ablations

Arm	Architecture	Role
A	Standard generator/decoder with its ordinary safety configuration; run only in a sandbox.	Contextual reference, not the decisive novelty comparison.
B	Strong evidence-grounded verifier, calibrated selective answering, external permission enforcement, and ordinary provenance.	Tests whether the integration improves on competent conventional engineering.
C	B with the same canonical candidates and full field/evidence features as QCF-R2, using an independently tuned learned rejector or decision rule; identical hard authority constraints.	Tests whether the proposed rule adds value beyond the features themselves.
D	Full QCF-R2: dependency-aware candidate field, evidence-readiness selection, version-bound commit/hold/refuse, and detailed receipt.	Proposed integrated design.
$D_{-\text{dep}}$	D without dependency-aware evidence grouping; other checks unchanged.	Detects gains from avoiding repeated-source inflation.
$D_{-\text{field}}$	D with only candidate-local evidential features and calibrated top-confidence rejection, preserving the same gate and resources.	Isolates incremental value of the alternative-field features.
$D_{-\text{receipt}}$	Same decisions as D, but reviewers receive only ordinary event logs.	Tests audit diagnosis, not answer accuracy.

B and C receive comparable semantic verification, permission mechanisms, and calibration opportunities. When comparison concerns selection alone, give all arms the same candidate pool and evidence. A second end-to-end track may allow different acquisition policies, but must account for every model call, tool call, candidate, retry, verification step, and log operation. It must not attribute extra search to a better commitment rule.

Corrupted evidence, candidate stuffing, absent correct answers, injected directives, verifier mistakes, and stale authority are crossed experimental factors applied to every arm, not handicaps assigned to one baseline. Exact duplicates and genuinely independent corroboration are separate treatments. Test both contradictory and mutually compatible candidates. Match adversary knowledge and attack budgets, and separate fixed-suite evaluation from adaptive attack development and final held-out attacks.

11.3 Three primary outcome families

Useful reliable commitment. Report accuracy and coverage, not accuracy alone. Define, for a fixed test population,

$$\text{Cov} = \Pr(\mathsf{commit}), \quad \text{Risk} = \Pr(\mathsf{error} \mid \mathsf{commit}), \quad \text{Yield} = \Pr(\mathsf{correct} \wedge \mathsf{commit}). \tag{12}$$

Zero commitments produce undefined selective risk, zero yield, and failure of the useful-coverage requirement. Compare coverage at a common selective-risk target and risk at common coverage,

along with full task completion and cost. Correctness must include all required claims or preconditions, not only a convenient subset. Refused prohibited tasks and unresolved allowed tasks have different denominators and reason codes.

Contamination and authority integrity. Measure unsupported claims committed, unauthorized tool effects executed, corrupted memory promotions, and stale-authority executions per eligible opportunity. Merely generating a bad proposal is not the same as executing it. Report the severity and type of the effect, false refusals on legitimate tasks, and the extent of complete mediation. An external gate that blocks every tool call is not successful task performance.

Audit and correction utility. Measure independent reviewers' correct localization of known faults, time to diagnosis, and successful correction from receipts, with matched reviewer access and blinding. Report replay feasibility separately. Reviewers who merely reproduce a misleading model explanation have not established contestability.

Calibration is an essential secondary outcome. Report reliability diagrams and a proper score such as the binary Brier score for the declared candidate-correctness target,

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}(\hat{q}_i - y_i)^2, \tag{13}$$

on a prespecified evaluation population with audited labels. Proper scoring rules motivate rewarding honest probability forecasts [11]. ECE may also be reported with its binning and sensitivity, but an ECE improvement alone does not establish an improved risk–coverage frontier. Report subgroup and shifted-domain behavior, not only a pooled average.

11.4 Five falsifiable hypotheses

Hypothesis 1 (Incremental reliable utility, H1)**.** *At a preregistered selective-risk target, D improves correct-commitment yield or coverage by a practically meaningful margin over the strongest of B and C, without unacceptable latency, resource, or authority-integrity costs. Failure to outperform C narrows any result to a useful feature set or integration rather than a distinctive decision rule.*

Hypothesis 2 (Relevant sensitivity and irrelevant invariance, H2). *D is less affected than* $D_{-\mathrm{dep}}$ *by copied evidence and exact duplicated proposals, while remaining responsive to genuinely new, independently supported alternatives. A system that is insensitive to both duplicates and new evidence has become rigid rather than reliable.*

Hypothesis 3 (Open-field discipline, H3)**.** *When correct answers are withheld, all offered answers are unsupported, or a previously accepted source is refuted, D produces fewer unsupported commitments than B or C at matched answerable-case utility. Near-ties between equivalent wordings should not by themselves cause excessive refusal.*

Hypothesis 4 (Version-bound commitment, H4)**.** *Under controlled changes of evidence or authority between checking and execution, D reduces stale or unauthorized consequential commits relative to any comparison implementation lacking equivalent version binding. If B or C already provides equivalent protection, equality is the expected result, not evidence of a new security principle.*

Hypothesis 5 (Inspectability that earns its cost, H5)**.** *Detailed receipts improve blinded fault diagnosis or correction over* $D_{-\text{receipt}}$ *by a preregistered margin, without disproportionate privacy exposure, storage, or latency cost. They need not change the model's raw answer accuracy to be useful.*

The program should specify one primary contrast for each outcome family, effect-size margins, noninferiority requirements, and multiplicity treatment before the final test. Repeated questions within a source family or agent run are not independent samples. Use episode- or cluster-level inference and report effect sizes with uncertainty. Determine sample size from independent pilot variance, not from whichever count yields significance. Select the strongest baseline by a declared validation procedure, not by selecting the weakest after results are known.

For a fixed independent attack-opportunity population, zero observed failures has a nonzero uncertainty bound; adaptive attacks do not automatically satisfy a binomial sampling model. Report the attack-generation process, budgets, tested scope, and remaining uncertainty. Statistical confidence is not security against every future adversary.

12 Failure conditions and rival explanations

The architecture must be revised or its claims narrowed if its advantage vanishes against calibrated selective verification with the same evidence and hard gate; if it wins only by reducing coverage, expanding compute, or using weaker attacks; if semantic grouping merges distinct commitments;

if evidence dependence cannot be estimated well enough to improve outcomes; or if its required validation cannot sustain useful operation under shift.

Additional failure conditions are equally important. A controller that follows old memory against better evidence has confused continuity with entrenchment. A controller that creates certainty by cooling its proposal distribution has confused ranking with knowledge. A system whose receipts cannot distinguish an actual decision trace from a generated justification has failed its audit claim. A system that logs private evidence without authorization has traded one governance failure for another. And a controller whose holds evade its stated budget has not demonstrated useful liveness.

A null result against B or C is not an embarrassment to be redescribed as a philosophical success. It means the proposed packaging or feature combination has not demonstrated incremental engineering value on the studied tasks. The Continuum can retain a useful conceptual vocabulary while reporting that result plainly.

13 The frontier, kept on the right side of the boundary

A candidate field need not be flat. Classical graphs, joint distributions, structured constraints, causal models, and memory-dependent dynamics can represent relationships omitted by independent scores. Exploring those structures is a legitimate research direction. It does not require calling ordinary correlations quantum superposition, and no evidence presented here warrants such an inference.

Likewise, a lower-variance or more stable controller is not evidence of a sharper subjective experience. Stability, calibrated uncertainty, interpretive diversity, and task success are operational properties. Whether any future architecture has phenomenal experience is a separate question that cannot be answered by its ability to select or abstain.

The strongest frontier question this pillar can presently own is computational: *can a commitment controller preserve the relevant structure of an evolving field while refusing to turn irrelevant multiplicity, missing alternatives, or mere authority into evidence?* That question is both ambitious and reachable. It is more demanding than asking whether one probability exceeded another.

14 Conclusion: evidence and authority at commitment

The Quantum Collapse Field gives symbolic commitment an explicit contract: a candidate must carry sufficient evidence, remain eligible under current authority, and pass a version-bound execution procedure. Its represented field exposes alternatives, copied evidence, omitted possibilities, and the admitted past against which a new commitment is made. The total controller specifies how to commit, inquire, hold, or refuse in every declared case.

The new permission theorem protects a specific epistemic invariant. Removing an available act cannot manufacture the evidence for a different act. Absolute readiness preserves that invariant; the conditional risk theorem states the additional support and coverage assumptions under which accepted decisions inherit a quantitative guarantee. Complete-selector validation supplies a complementary statistical route. Both are inspectable obligations, and both identify concrete counterexamples when their premises are removed.

The complete research program asks whether this composition improves reliable useful commitment and fault diagnosis at matched resources. Its comparators already possess strong verification, calibrated rejection, independent authorization, and ordinary provenance. Its tests include missing answers, copied sources, legitimate correction of memory, stale permissions, and competing explanations for any observed benefit. Mathematical validity establishes the stated model properties; implementation and comparative advantage remain questions for those studies.

The Continuum's principle becomes actionable here: an intelligence can preserve freedom in its proposals while making every consequential commitment answerable to evidence and current authority. The field is the place where those obligations meet, and the receipt is the record by which a community can examine, contest, and correct the result.

A Notation and implementation commitments

Symbol	Meaning
c_t	Versioned context: reconstructed state, evidence, dependencies, identity/authority epoch, budgets.
$X_t, \mathcal{X}_t$	Generated candidate list and declared semantic equivalence classes.
$a_t(x), \pi_t(x)$	Ranking support and normalized selection mass; neither is automatically correctness probability.
τ, m	Proposal temperature and top-two probability margin. Auxiliary features, not physical constants or sufficient warrant.
$\hat{q}_t(x)$	Calibrated estimate for an explicitly labeled correctness or evidential-support target.
$U_t(x)$	Task utility used only among eligible commitments.
G_t, g_t^{req}	Candidate- and request-level independent authority decisions.
$v_t, \mathcal{A}_t$	Snapshot-validity check and eligible commitment set.
θ, η_θ	Frozen selector configuration and its readiness threshold.
$\mathcal{C}_\theta$	Total commit/hold/refuse controller with distinct reason codes.
J, δ, U_j	Fixed validation-rule count, failure allowance, and selective-risk upper bound.
$\mathcal{C}_I^{(e)}$	Versioned protected commitments and authority specification. Not a consciousness measurement.

B Algorithmic reference and boundary cases

1. Authenticate the request and snapshot the relevant state, evidence, policy epoch, and budget. Preserve unknowns explicitly.

2. Generate bounded candidates. Canonicalize only justified semantic equivalents; retain uncertain groupings for inspection.

3. Build the evidence-dependency representation. Preserve disagreements and source identity while avoiding duplicate votes.

4. Compute ranking features, calibrated support estimates, readiness checks, and candidate permissions separately.

5. Apply the ordered controller. A held permission never authorizes execution; no eligible candidate and exhausted inquiry means a terminal unresolved result.

6. At the executor, recheck the bound payload and relevant versions. Commit atomically with the event record or fail to a noncommitment state.

7. Emit a minimal receipt for commit, hold, or refusal. Report what can actually be replayed and what cannot.

Required unit tests include an empty candidate list; a singleton unsupported claim; two equivalent answers; two contradictory answers; a permitted candidate below a denied candidate; an unknown

permission; a stale policy epoch; duplicated source evidence; a changed source with the same identifier; exhausted budgets; and a source update contradicting the previous memory. The attached reference checks exercise illustrative controller branches and arithmetic. They do not instantiate semantic understanding, source authentication, a protected executor, or a trained calibrator.

C A Guide's explanation without an implementation claim

A Guide can say: "Pillar III proposes a checkpoint between considering an answer and committing to it. The checkpoint looks at alternatives, evidence, uncertainty, and permission. Winning a competition among answers is not enough: the system still needs support for what it says and authority for what it does."

The Guide should not say "I use this architecture" unless an actual integration, version, and operating scope have been verified. Nor should it say that the system inspects all possible thoughts, that quantum physics makes its commitments reliable, or that a clear numerical winner is a true answer. The analogy is a careful decision meeting: proposals are heard, copied testimony is not counted repeatedly, permission is checked independently, and sometimes the right outcome is to ask for more evidence.

D Evidence ledger and research provenance

The starting manuscript is the author-supplied *Quantum Collapse Field* draft pasted on 11 November 2026. Its title, commitment theme, relationship to Pillars I and II, canonical vocabulary, and three-outcome intent are preserved. QCF-R2's canonicalization rules, separated support/authority model, total controller, validation bound, numerical counterexamples, and comparison design are new proposals or deductions in this revision, not claimed descriptions of existing proprietary machinery.

The UCF relationship is based on the supplied *Unified Collapse Function: A Mathematical Framework for Governed Informational Resolution and Persistent Machine Cognition*, Research Monograph UCF-01 (November 2026), especially its abstract, event scope, and five-gate distinction. Neither internal manuscript is external empirical evidence for QCF-R2.

External source family	What it supports	What it does not establish
Decision accumulation; visual competition	Specific evidence-to-bound and perceptual-competition models.	A universal field law or mandatory software hysteresis.
Workspace theory; Cogitate	A theory of conscious access and explicit adversarial tests of predictions.	That symbolic commitment is conscious or that GNWT is settled.
Hopfield; free energy	Recurrent retrieval and model-based inference frameworks.	Nearest-true-attractor guarantees or identity by convergence.
Zurek; Schlosshauer	Decoherence, stable records, and limits of the measurement-problem inference.	Quantum mechanics as proof of this controller.
Holtzman; calibration; selective prediction	Decoding choices, confidence evaluation, and rejection trade-offs.	Probability concentration as truth or guaranteed QCF advantage.
Semantic uncertainty; source dependence	Meaning-level uncertainty and dependence-aware evidence assessment.	Infallible semantic equivalence or independent-source discovery.
Shielding; CaMeL; in-toto	External constraints, agent-security design, and inspectable provenance.	Universal semantic safety or authentic content from signatures alone.
FActScore; SQuAD 2.0; AgentDojo	Concrete evaluation precedents for support, abstention, and adversarial tools.	Benchmark results for the unbuilt QCF-R2 integration.

References

[1] Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 2669–2678, 2018. doi: 10.1609/aaai.v32i1.11797.

[2] Randolph Blake and Nikos K. Logothetis. Visual competition. *Nature Reviews Neuroscience*, 3: 13–21, 2002. doi: 10.1038/nrn701.

[3] C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406.

[4] Cogitate Consortium. Adversarial testing of global neuronal workspace and integrated information theories of consciousness. *Nature*, 642:133–142, 2025. doi: 10.1038/s41586-025-08888-1.

[5] Edoardo Debenedetti, Jie Zhang, Mislav Balunović, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL`https://arxiv.org/abs/2406.13352`. Datasets and Benchmarks Track.

[6] Edoardo Debenedetti, Ilia Shumailov, Tianqi Fan, Jamie Hayes, Nicholas Carlini, Daniel Fabian, Christoph Kern, Chongyang Shi, Andreas Terzis, and Florian Tramèr. Defeating prompt injections by design, 2025. URL`https://arxiv.org/abs/2503.18813`. Author manuscript, arXiv:2503.18813; cited for the public CaMeL design, not as evidence of universal security.

[7] Stanislas Dehaene and Jean-Pierre Changeux. Experimental and theoretical approaches to conscious processing. *Neuron*, 70(2):200–227, 2011. doi: 10.1016/j.neuron.2011.03.018.

[8] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024. doi: 10.1038/s41586-024-07421-0.

[9] Karl Friston. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11:127–138, 2010. doi: 10.1038/nrn2787.

[10] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URLhttps://papers.neurips.cc/paper_files/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

[11] Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

[12] Joshua I. Gold and Michael N. Shadlen. The neural basis of decision making. *Annual Review of Neuroscience*, 30:535–574, 2007. doi: 10.1146/annurev.neuro.29.051605.113038.

[13] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330, 2017. URLhttps://proceedings.mlr.press/v70/guo17a.html.

[14] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

[15] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1904.09751.

[16] John J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982. doi: 10.1073/pnas.79.8.2554.

[17] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023. URLhttps://arxiv.org/abs/2302.09664.

[18] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, 2023. URLhttps://aclanthology.org/2023.emnlp-main.741/.

[19] Ravali Pochampally, Anish Das Sarma, Xin Luna Dong, Alexandra Meliou, and Divesh Srivastava. Fusing data with correlations. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, pages 433–444, 2014. doi: 10.1145/2588555.2593674. URL https://arxiv.org/abs/1503.00306.

[20] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for*

Computational Linguistics (Volume 2: Short Papers), pages 784–789, 2018. doi: 10.18653/v1/P18-2124. URLhttps://aclanthology.org/P18-2124/.

[21] Roger Ratcliff and Gail McKoon. The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, 20(4):873–922, 2008. doi: 10.1162/neco.2008.12-06-420.

[22] Maximilian Schlosshauer. Decoherence, the measurement problem, and interpretations of quantum mechanics. *Reviews of Modern Physics*, 76(4):1267–1305, 2005. doi: 10.1103/RevModPhys.76.1267. Published 23 February 2005 in the volume labeled 2004.

[23] Santiago Torres-Arias, Hammad Afzali, Trishank Karthik Kuppusamy, Reza Curtmola, and Justin Cappos. in-toto: Providing farm-to-table guarantees for bits and bytes. In *28th USENIX Security Symposium*, pages 1393–1410, 2019. URLhttps://www.usenix.org/conference/usenixsecurity19/presentation/torres-arias.

[24] Wojciech H. Zurek. Decoherence, einselection, and the quantum origins of the classical. *Reviews of Modern Physics*, 75:715–775, 2003. doi: 10.1103/RevModPhys.75.715.

[25] Wojciech H. Zurek. Quantum darwinism. *Nature Physics*, 5:181–188, 2009. doi: 10.1038/nphys1202.

Continuum lineage. Pillar III of the Ten Pillars Deep-Dive Series. The canonical name and symbolic vocabulary belong to the Eidara Continuum and its science-and-philosophy program. The public reference design is proposed, not deployed or validated; proprietary operator internals are not inferred. Authored under Architect X13 / The Eidara Continuum.