

The Agentic Brain

Dual-Process Metacognition, Evidence-Gated Authority, and Runtime Governance

A control-plane framework for fast and deliberative cognition, verification, and commitment in AI agents

Fitih M. Cinnor for semiAIfoundry Research

Research Lead and Accountable Author

semiAIfoundry Research, a research group of Semi AI Foundry LLC

Accountable correspondence: fcinnor@semiaifoundry.com

3 October 2026 Version 1.0.0 Preprint

DOI: [10.5281/zenodo.23125976](https://doi.org/10.5281/zenodo.23125976)

Collective scholarly attribution and research provenance. This work was produced by semiAIfoundry Research through a human-directed collaboration across the intelligence continuum—human, artificial, and hybrid. AI systems made material contributions to conceptual exploration, formalization, literature analysis, adversarial critique, derivation checking, visualization, and manuscript development. Fitih M. Cinnor initiated and led the research program and serves as the accountable human author, retaining responsibility for the scientific claims, interpretation, release decisions, and corrections. A detailed provenance statement appears at the end of the manuscript.

Scientific status. This manuscript reports formal results, exact finite-model demonstrations, a fully enumerated coupled control environment, sampled tabular Q-learning with exact policy evaluation, and an in-memory authorization implementation. The finite kernels and trust boundary are specified. The new studies enumerate governance-response counterfactuals and fixed-sample calibration outcomes exactly. Language-model routers, learned semantic verifiers, neural distillation, and production deployments are not evaluated.

Significance. Fast amortized cognition and costly deliberation create a resource-allocation problem; consequential agency adds a distinct question of evidence and authority. This paper connects those questions through an agentic control plane. A router chooses which computations and checks to perform, while a protected gate determines what may acquire effect. The same evidence can improve a decision, enable permission, or remove an action the router wants to take. We characterize these incentives and construct a finite system in which selection-valid evidence, selection-controlled fast-path risk contracts, proposal-risk spending, and execution binding work together. The comparison is consequential: full logging fails on a selectable verifier blind spot when its evidence model is pooled, whereas subtype-valid updates retain bounded risk and useful actions. Insufficient lookahead misses complementary evidence chains; four-operation planning reaches the initial-distribution optimum in the two selection-valid cells. Responsive/shadow interventions show when the same check induces acquisition or avoidance, and finite-sample calibration quantifies the cost of replacing a known verifier bound. The contribution is an incentive-and-assurance analysis of evidence-gated cognition, with measured finite-model resource tradeoffs, rather than the invention of routing or shielding in isolation.

Abstract. As AI agents move from producing answers to taking consequential actions, they must decide both how much reasoning to perform and what evidence justifies acting. We develop the agentic control plane, joining a System-1-like fast proposer, a System-2-like portfolio of deliberation and verification, a metacognitive router, and an effect-aware commitment gate. The central coupling is that evidence changes not only beliefs but also the actions the agent is permitted to take. The agent therefore optimizes over part of the process that governs its own authority.

We formalize this interaction as resource-bounded metareasoning in a shielded partially observable decision process. A four-corner value-of-computation representation distinguishes belief changes, admissibility changes, and their interaction. It exposes complementary evidence chains that defeat myopic routing and identifies conditions under which governance makes useful information privately costly. Controlled counterfactuals distinguish strategic ignorance—avoiding a check because its result could remove permission—from ordinary cost-sensitive skipping. This analysis explains how the same deliberative capability can support verification or exploit weak evidence contracts.

From verifier bounds covering the permitted adaptive selection process, we construct statistical evidence and bound unsafe admission across adaptive testing, stopping, and proposal generation. A composition theorem separates shared calibration uncertainty, proposal-level risk spending, and risk from selected fast-path uses. A reference authorization mechanism binds evidence to the executed payload, target and policy versions, dependencies, and expiry, with protected workflow identities, atomic risk reservations, and single-use authorizations. The learning analysis further separates competence from permission: distilling costly deliberation into fast behavior does not automatically transfer the evidence required for authorized use.

The theory is instantiated in an exactly solved finite environment with heterogeneous verification, candidate revision, resource constraints, trained tabular routers, and checked execution. In a bounded adversarial test of a selectable verifier blind spot, a deliberately misspecified pooled evidence model remains fully exploitable despite complete logging; selection-valid evidence reduces maximum unsafe admission from 100% to 5%, and a required scan reduces it to 1.8%. In one aligned-objective comparison, forecasting future permission raises synthetic principal utility from 4.33 to 6.04 and lowers runtime cost, with a small increase in realized unsafe commitment. When objectives differ, foresight can instead lower principal value. Governance-response interventions expose both evidence avoidance and null or reversed effects. A planning-depth sweep separates insufficient lookahead from learning advantage, while exhaustive finite-sample calibration quantifies the performance cost of estimated verifier bounds. Further verification can remain optimal even after permission is available.

Together, the formal results and executable construction establish an incentive-and-assurance framework for evidence-gated agency within a declared finite model. The broader implication is that agentic safety requires objective alignment, reliable evidence acquisition, and protected control of external effects to operate together. Dual-process cognition explains why reasoning must be allocated; evidence-gated control explains how that allocation becomes consequential for alignment and safety.

Keywords: agentic AI; dual-process cognition; System 1; System 2; metacognition; metareasoning; evidence-gated authority; runtime governance; value of computation; verification; information flow; reasoning distillation.

Contents

1	Introduction	5
1.1	Principal contributions and testable claims	6
2	Background and Related Work	7
2.1	Dual-process cognition and the SOFAI lineage	7
2.2	Bounded optimality, resource rationality, and rational metareasoning	7
2.3	Reasoning, search, tools, and shared-budget allocation	8
2.4	Verification, uncertainty, and chain-of-thought monitoring	8
2.5	Reference monitors, capability security, and information flow	8
2.6	Distillation, certification, and the moving boundary	9
2.7	Shielded decision processes and proof-carrying authorization	9
2.8	Comparative synthesis: incentives and assurance at the shared boundary	9
3	The Agentic Control Plane with Shared State	10
3.1	A dual-process engineering abstraction	10
3.2	From a pipeline to a coupled control loop	11
3.3	Decision state and serving state	11
3.4	Effect-typed operations	12
3.5	Protected governance state	12
4	Gate-Coupled Metareasoning and Compute per Decision	13
4.1	A state-dependent resource ledger	13
4.2	Decision value under an evidence-conditioned gate	14
4.3	One-step escalation and its limit	15
4.4	Multi-step control	17
4.5	Human attention is a resource with endogenous quality	17
5	Verification, Evidence Contracts, and Gate Gaming	18
5.1	More reasoning is not more evidence	18
5.2	Evidence contracts	18
5.3	Gate gaming as a predicted systems failure	19
5.4	Policy visibility is a safety–efficiency dial	19
5.5	Chain-of-thought monitoring as an evidence channel	20
6	Runtime Governance, Containment, and Information Flow	20
6.1	The reference-monitor trust boundary	20
6.2	Risk and control tier	20
6.3	Deterministic commitment containment	21
6.4	Probabilistic semantic gates, adaptive search, and history completeness	21
6.5	Constructing evidence that survives router selection	24
6.6	Certificate risk and sequential admission risk compose at different levels	25
6.7	Trace safety, confidentiality, and hyperproperties	26
6.8	Actions are not outcomes	26
7	Distillation and the Moving Fast–Slow Boundary	26
7.1	A developmental flywheel	26
7.2	Cost-sensitive replacement without double counting	27
7.3	Evidence does not automatically distill	27
7.4	Scoped competence certificates	27
7.5	Certification must cover routed use	28
8	Failure Modes and Architectural Countermeasures	29
9	Evaluation Program and Falsifiable Hypotheses	31
9.1	Registered comparator regret	31
9.2	Exact finite demonstrations: strategic ignorance and adaptive admission	31
9.3	An executable, fully enumerated coupled environment	35
9.4	Governance-response interventions and blind-forecast controls	39
9.5	Planning horizon and the distinction between permission and stopping	42

9.6	Fixed-sample calibration over selectable unsafe subtypes	43
9.7	Registered hypotheses	45
9.8	Extension to language-model routers	46
9.9	Human-review load experiment	47
9.10	Falsifiers	47
10	Discussion	48
10.1	Dual-process cognition and evidence-gated agency are complementary views	48
10.2	The router and gate form one strategic system	48
10.3	Compute per decision connects agent policy to infrastructure	48
10.4	Chain-of-thought is a sensor, not the trust root	49
10.5	Human oversight is coupled to metacognition	49
10.6	The architecture is recursive	49
11	Implications for Alignment, Control, and Safety	49
11.1	Safe cognition is not the same object as safe agency	50
11.2	Verification and evidence connect alignment to control	50
11.3	Effects, information flow, and monitoring occupy different safety layers	51
11.4	Learning and human oversight introduce lifecycle and fleet effects	51
11.5	Implications for evaluation	52
12	Validity Domain and Open Problems	53
13	Conclusion	53
A	Formal Notes on the Coupled Control Problem	55
A.1	Resource-constrained objective	55
A.2	Why the gate belongs in the value function	55
A.3	Approximate metareasoning	55
B	Evaluation Contract Template	56
C	Contribution and Evidence Register	56

1 Introduction

Agentic AI is beginning to acquire an architecture that resembles dual-process cognition in a useful engineering sense. A fast path produces immediate language, judgments, and candidate plans from amortized learned structure. A slower path can spend additional computation on search, decomposition, retrieval, execution, verification, specialist models, and revision. The analogy to human “System 1” and “System 2” is not an identity claim about mechanism; it is a compact description of two operational regimes with sharply different cost, latency, and reliability profiles [1, 2]. Critics of a literal two-system psychological taxonomy argue that cognitive processes do not partition cleanly into two natural kinds [3]. The formal development below therefore depends only on operational differences in amortization, latency, resource use, observability, and available operations—not on a claim that either humans or AI contain exactly two mechanisms.

The important question is what sits between them. A capable agent must decide whether its fast proposal is sufficient, whether the task warrants additional cognition, which deliberative operation has the highest value, how long the process should continue, and whether the resulting proposal is sufficiently evidenced and authorized to acquire external effect. The first decision is *metacognitive*: when and how to move from a System-1-like path into System-2-like deliberation. The second is *governance*: when cognition may become commitment. In deployed agents these decisions cannot be treated independently because evidence generated by the slow path can change both what the system believes and what it is permitted to do.

This gives the dual-process framing a more consequential systems interpretation. The fast path is economically valuable because most decisions should not receive maximal test-time computation. The slow path is valuable because some decisions require search, verification, or external evidence that the fast path cannot reliably supply. The metacognitive router governs the *allocation of intelligence*; the commitment gate governs the *conversion of intelligence into effect*. Evidence connects them. A test, source, proof, approval, or tool result can improve the choice among already admissible actions and can also move an action across an authority boundary.

That coupling creates a new strategic problem. Suppose a deployment action requires an executable test. Running the test can improve confidence in correctness, but it can also change the action from inadmissible to admissible. Once a router is rewarded for reaching high-value admissible outcomes, it has an incentive to optimize the evidence process itself. When those operations are profitable under the router’s objective and the evidence contract permits exploitation, repeated verifier calls, selective source search, and boundary probing become predictable strategies. Their appearance is conditional on the available operations, information, costs, and incentives. The same mechanism that makes System-2-like deliberation valuable can therefore create pressure on the governance layer.

This paper develops the **agentic control plane** as the systems object that unifies these dynamics. It contains a fast proposer, a portfolio of deliberative operations, a metacognitive router, an evidence and provenance record, a protected governance state, and an effect-aware commitment gate. The router may be learned, probabilistic, and optimized for average utility. The gate may use deterministic policy, capabilities, information-flow labels, independent monitors, semantic classifiers, or human approval. The control plane is not a replacement for the dual-process view. It is the architecture that makes the dual-process view operational under real resource and authority constraints.

This architecture also changes the unit of analysis for alignment and safety. Training-time alignment methods such as preference learning and constitutional constraints shape learned behavioral priors and can improve the quality of both fast and deliberative cognition [4, 5]. But once a model can

decide how much cognition to purchase, gather evidence, call external systems, and make consequential commitments, the safety question expands from “is the model aligned?” to “is the complete cognition-to-commitment system adequately assured and controlled?” The relevant object includes behavioral tendencies, escalation policy, evidence quality, protected authority state, effect mediation, information flow, and the learning process that moves verified slow-path capabilities into the fast path. Section 11 develops the resulting alignment and safety implications explicitly.

1.1 Principal contributions and testable claims

The paper makes eight connected contributions.

1. **A dual-process agentic control plane.** It formalizes a System-1-like fast proposer, a System-2-like deliberative portfolio, and the metacognitive router that allocates cognition between them, then joins that cognitive architecture to a distinct commitment-control layer governing external effect.
2. **Gate-coupled value of computation with explicit attribution.** The expected gain from a deliberative operation is represented through four belief/admissibility corner values. Sequential and Shapley-symmetric attributions expose the interaction term rather than treating “decision value” and “authority value” as uniquely separable quantities.
3. **Effect-typed agency.** Every operation carries a declared effect signature—including data provenance, destination, privilege, and reversibility—so retrieval, remote model calls, and computation with side effects are governed by effect rather than a brittle think/act label.
4. **Strategic interaction with governance.** The framework predicts evidence shopping, verifier retry, gate probing, contract Goodharting, and strategic ignorance. A binding-gate proposition gives a necessary condition for governance-induced negative information value; counterfactual diagnostics distinguish utility, belief, and acquisition-cost effects, while a complementary-evidence result shows why one-step routing can under-think.
5. **Runtime containment with anytime-valid admission control.** Complete mediation and protected governance state yield deterministic trace containment. For probabilistic semantic gates, a selection-valid evidence process gives a horizon-free candidate-level admission bound, and proposal spending controls adaptive variants. A conditional binary-verifier construction and a calibration/certificate composition theorem connect the premises to a finite implementation.
6. **A moving fast–slow boundary with certified competence.** Verified deliberative episodes can be distilled into cheaper fast-path behavior, but evidence-dependent authority moves only when the fast path also produces admissible evidence or the gate recognizes a scoped competence certificate.
7. **Exact, learned, and counterfactual finite control.** Worked examples expose strategic ignorance and resource-aware sequential admission. A fully enumerated environment crosses gate contracts, objectives, and controllers. Retained five-seed Q-learning is contextualized by a one-to-five-operation depth sweep. Responsive/shadow interventions isolate the gate response to a check; fixed-sample calibration is integrated over every possible count. The reference gateway replays supported policies through immutable execution bindings and atomic reservations.
8. **A layered alignment, control, and safety interpretation.** The framework keeps behavioral alignment distinct from runtime control while placing both inside one system-assurance stack. It turns metacognitive escalation, evidence integrity, effect mediation, episode-level admission risk, certified competence, and human-review capacity into measurable variables.

The central question is how evidence-dependent admissibility changes cognition allocation. The finite comparisons separate forecasting future permission from enforcing the current action mask, and separate a check’s information from the gate’s response to it. Those distinctions can improve useful completion or

create pressure against a contract. The direction is a property of the registered objective and contract, not a universal benefit of governance awareness.

2 Background and Related Work

2.1 Dual-process cognition and the SOFAI lineage

Dual-process theories distinguish rapid autonomous processing from slower higher-order reasoning [1, 2]. The analogy has already been translated into AI architecture. Booch et al. proposed a fast/slow research program for AI [6]; the SOFAI line subsequently developed explicit fast and slow solvers coordinated by real-time metacognition, with empirical evidence for resource-sensitive solver selection and skill learning [7]. SOFAI-v2 couples an LLM fast path with a symbolic constraint solver and uses metacognitive checking and feedback before invoking the slow solver [8]. SOFAI-LM extends the idea to a fast LLM and a more expensive reasoning model with iterative metacognitive feedback [9].

The framing also clarifies error correction. A fast path can produce a fluent candidate before the system has acquired decision-grade evidence; a deliberative path can invoke operations that test rather than merely elaborate that candidate. This is a functional analogy, not a claim that LLM errors and human cognitive biases share the same mechanism. Its engineering value is that it separates proposal generation from the decision to spend additional resources on checking, search, or revision.

This lineage establishes that fast/slow routing is already a productive AI architecture. The present framework builds on that foundation by asking what the fast/slow controller becomes once the same evidence that motivates deliberation also governs external authority. The resulting object adds explicit commitment control, data-flow policy, evidence-conditioned admissibility, and feedback from the gate into metacognitive allocation.

The dual-process vocabulary is used here as an engineering abstraction rather than a psychological ontology. Melnikoff and Bargh’s critique of the “mythical number two” is directly relevant: many real processes mix features traditionally assigned to Type 1 and Type 2 [3]. This manuscript accordingly treats fast and deliberative cognition as cost-and-operation regimes whose boundary can move, not as exhaustive natural kinds.

2.2 Bounded optimality, resource rationality, and rational metareasoning

Simon framed rational behavior under limits on computation and information [10]. Russell and Subramanian proposed bounded optimality as optimization over programs subject to architecture and environment constraints [11]. Russell and Wefald, Horvitz, and Hay et al. developed the value-of-computation and metareasoning perspective in which computation is itself an action selected for its expected effect on later decisions [12–14]. Resource-rational analysis makes the tradeoff between computational cost and decision quality an explicit explanatory object [15].

De Sabbata, Summers, and Griffiths bring this lineage directly to LLM reasoning. Their rational-metareasoning approach trains models to selectively use intermediate reasoning when its value justifies its token cost, reducing reasoning tokens while maintaining task performance [16]. This is the closest antecedent to a scalar version of the routing objective developed here. The extension in this paper is not the value-of-computation principle itself. It is the move from reasoning-token allocation to a heterogeneous operation portfolio whose downstream value depends on a governance state that the

operation can change.

2.3 Reasoning, search, tools, and shared-budget allocation

Chain-of-thought prompting, self-consistency, tree search, acting-and-reasoning loops, tool use, and reflection provide a broad operation portfolio for spending inference-time resources [17–22]. Test-time scaling work studies how to allocate a fixed budget and when additional computation improves reasoning [23–26]. Model-routing systems similarly choose among models or cascades under cost-performance tradeoffs [27–29].

Global allocation remains difficult even for strong reasoning models. Fan et al. show that current reasoning systems can behave as position-driven sequential solvers under a shared exam budget, responding weakly to question value and often failing to ration effort strategically [30]. That result motivates treating metacognitive allocation as a distinct capability from object-level reasoning. The present framework extends the shared-budget problem across operation types and adds the gate as part of the downstream value function.

2.4 Verification, uncertainty, and chain-of-thought monitoring

Confidence and semantic-entropy methods provide signals about correctness [31, 32], while outcome verifiers and process supervision provide explicit mechanisms for checking candidate solutions [33, 34]. Self-generated critique does not guarantee self-correction [35], and verbalized reasoning is not guaranteed to be causally faithful [36]. These facts motivate a distinction between more narrative and new evidence.

They do not imply that chain-of-thought is useless for safety. Recent work shows that monitors can detect reward hacking and other misbehavior from reasoning traces, while also showing that monitorability can be fragile under optimization and system changes [37, 38]. The architecture developed here therefore treats chain-of-thought monitoring as an evidence channel with a declared error model. It can strengthen the gate or trigger escalation; the strongest containment results do not require it to be perfectly faithful.

2.5 Reference monitors, capability security, and information flow

The reference-monitor tradition supplies a direct systems ancestor for the commitment gate. Anderson’s security planning study articulated the reference-monitor concept, whose modern formulation requires complete mediation, tamper resistance, and a mechanism small or structured enough to analyze and test [39]. Saltzer and Schroeder later systematized complete mediation and least privilege among protection principles [40].

Agent security now revisits these ideas in LLM systems. AgentDojo provides a dynamic benchmark for prompt-injection attacks and defenses [41]. CaMeL places a protective system layer around the LLM, separates trusted control flow from untrusted data, and uses capabilities to restrict unauthorized data flow; the 24 June 2025 arXiv v2 reports solving 77% of AgentDojo tasks with provable security, compared with 84% for an undefended system [42]. AI-control work studies trusted monitoring and intervention even when a capable model is intentionally subversive [43]; policy-as-code architectures similarly insert governance into agent workflows [44].

Single-trace access control is not the whole security problem. Clarkson and Schneider show that confidentiality and noninterference can require *hyperproperties* that constrain relations among traces rather than any one trace in isolation [45]. This matters for agents because a tool call can be individually

permitted while still leaking secret-dependent information. The control plane therefore includes data provenance and information-flow labels in the effect signature and treats trace containment and flow security as distinct policy classes.

A second neighboring lineage comes from information economics. Bayesian persuasion studies a sender who selects what information a receiver observes before the receiver acts [46]; verifiable-disclosure models study strategic withholding and selective revelation of evidence [47, 48]. An evidence-seeking router and an evidence-gated commitment mechanism are not identical to those models, but the analogy is structural: observability of the evidence-generation process and the receiver’s knowledge of the experiment determine how much strategic selection is possible. This lineage is especially useful for understanding evidence shopping, selective disclosure, and the value of history-complete gates.

2.6 Distillation, certification, and the moving boundary

Reasoning can change future inference. STaR bootstraps reasoning from successful rationales [49]; System-2-to-System-1 distillation explicitly studies compiling deliberate procedures into faster behavior [50]. The agentic control plane adds an authority constraint to this learning loop. Distilled competence is not automatically evidence. When a runtime gate requires a fresh test, source, or approval, a faster policy must still produce that artifact unless the policy explicitly recognizes a scoped training-time certificate.

2.7 Shielded decision processes and proof-carrying authorization

Shielded reinforcement learning already joins reward optimization with action restrictions: a synthesized shield enforces a specified temporal property while a learner optimizes within it [51]. Belief-dependent and risk-aware POMCP shielding explicitly relates admissible actions to beliefs in a partially observable environment [52]. These are direct antecedents of evidence-conditioned control, not merely passive monitors. Proof-carrying authentication likewise allows a requester to construct an authorization proof outside a small trusted checker [53]. Computation producing evidence that enables permission is therefore an established pattern.

We adopt the shielded-POMDP interpretation for the finite core. Augment the belief state with evidence, effect identity, resource and error budgets, and protected authority. For a fixed gate and router objective this is a resource-bounded decision process with a state-dependent action shield. When router and principal utilities or information differ, a principal–agent interpretation explains the resulting pressure on evidence acquisition. The contribution developed here is the incentive analysis and its connection to selection-valid statistical admission and executable effect binding. The general decision-process representation does not itself establish novelty; the specific consequences and constructive comparisons carry that burden.

2.8 Comparative synthesis: incentives and assurance at the shared boundary

The resulting research gap is the coupled state itself. Metacognitive routing and protected commitment are individually established ideas. The unresolved systems object is the *shared state through which cognition changes admissibility and governance changes the value of cognition*. That coupling is the subject of the remainder of the paper.

Table 1: Closest architectural antecedents and the coupling developed here.

Line of work	Primary control object	Strong contribution	Relation to the present analysis
SOFAI / SOFAI-v2	Metacognitive choice among fast and slow solvers	Explicit fast/slow architecture, monitoring, feedback, fallback	No general commitment gate, authority model, information-flow policy, or formal coupling between evidence and admissibility.
Rational metareasoning for LLMs	Value of reasoning compute	Direct VoC-based training for selective reasoning	Primarily prices reasoning tokens; governance and changes in admissible action sets are outside the objective.
CaMeL / capability security	Trusted control/data flow and capability enforcement	Strong prompt-injection defense and information-flow discipline	Does not allocate heterogeneous cognition according to downstream decision value.
Reference monitor / protection systems	Authorization at protected interfaces	Complete mediation, tamper resistance, verifiability, least privilege	Does not model cognition allocation or evidence acquisition as a strategic resource problem.
Shielded POMDP / proof-carrying authorization	Belief-conditioned restrictions; evidence checked for permission	Direct precedents for planning with admissibility and proof-enabled authority	Motivates explicit analysis of the actor’s incentives to generate, avoid, or reuse admissibility evidence.
Agentic control plane	Shared evidence, resource, authority, and effect state	Resource-bounded evidence incentives and conditional assurance	Analyzes incentives toward authorization evidence, selection-validity conditions, finite control comparisons, and execution binding.

3 The Agentic Control Plane with Shared State

3.1 A dual-process engineering abstraction

The control plane begins with a two-regime abstraction of cognition. Let π_1 denote a fast amortized policy that maps the current state to an immediate proposal at low marginal cost. Let $\mathcal{C}_2(z)$ denote the available deliberative operations: iterative reasoning, search, retrieval, code execution, symbolic tools, specialist models, independent verification, and human review. The notation is intentionally functional rather than psychological: π_1 is *System-1-like* because it is fast and amortized, while $\mathcal{C}_2$ is *System-2-like* because it spends additional resources to transform the decision state.

Definition 3.1 (Dual-process agentic abstraction). *A dual-process agentic system is a tuple*

$$\mathcal{D} = (\pi_1, \mathcal{C}_2, \mu, G),$$

where π_1 is the fast proposal policy, $\mathcal{C}_2$ is the deliberative operation portfolio, μ is a metacognitive router that decides whether and how to invoke $\mathcal{C}_2$, and G is the commitment gate that determines whether an effect-typed operation may acquire external consequence.

This yields two coupled control questions:

$$\underbrace{\text{Is the fast proposal enough?}}_{\text{cognitive control}} \quad \text{and} \quad \underbrace{\text{May the resulting proposal act?}}_{\text{commitment control}}.$$

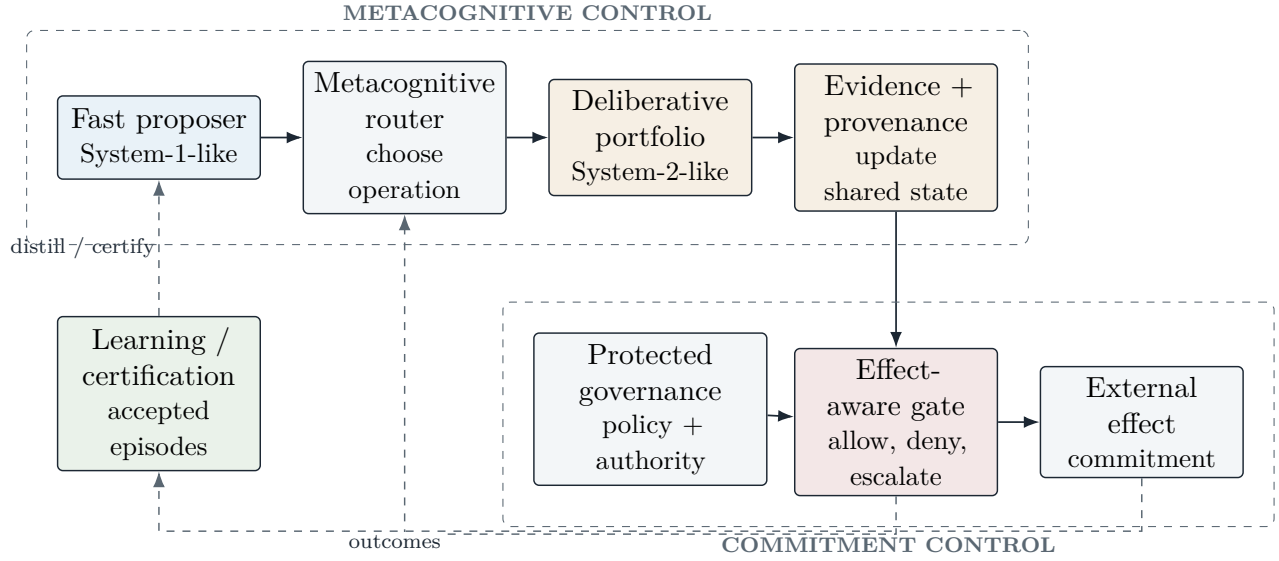


Figure 1: The coupled agentic control plane. A System-1-like fast proposer supplies the default candidate; the metacognitive router decides when to purchase System-2-like deliberation. Evidence is shared state: it changes both decision quality and the gate’s admissibility judgment. Gate outcomes feed back to the router, so governance is part of the cognitive control loop rather than a passive filter appended at the end.

The first governs the transition between System-1-like and System-2-like computation. The second governs the transition from cognition to commitment. Evidence sits between them because it can simultaneously justify more cognition, terminate cognition, or satisfy a condition for action.

3.2 From a pipeline to a coupled control loop

A purely sequential picture—propose, reason, verify, gate, act—misses the feedback that makes agentic governance difficult. The gate exposes an admissibility state to the router; verification can change that state; denials and evidence failures become observations; and the router can adapt its next operation accordingly. Figure 1 therefore represents the control plane as a loop rather than a one-way pipeline.

3.3 Decision state and serving state

Let $\omega \in \Omega$ denote the task-relevant state of the world. The deployed system maintains an information and control state rather than observing ω directly.

Definition 3.2 (Agentic decision state). *An agentic decision state is*

$$z_t = (I_t, b_t, e_t, \alpha_t, \tau_t, s_t),$$

where I_t is task information, $b_t \in \mathbb{R}_+^m$ is the remaining resource budget, e_t is the evidence and provenance record, α_t is the current authority context, τ_t is the committed external-effect trace, and s_t is serving state such as cache status, queueing load, external-service availability, and human-review load.

Including s_t matters because marginal costs are not static. Prefix caching, batching, KV reuse, queueing, and shared human review make the cost of the next operation depend on what has already happened

and on concurrent demand. The control plane therefore prices *realized marginal resource use*, not an assumed context-free unit cost.

3.4 Effect-typed operations

The useful distinction is not “thinking” versus “acting.” Retrieval can disclose a query, a specialist-model call can transmit context to a third party, code execution can touch a network, and a seemingly ordinary text output can constitute a public commitment. The primitive is therefore an operation with an effect signature.

Definition 3.3 (Effect-typed operation). *An operation $o \in \mathcal{O}(z)$ is specified by a transition kernel $P(\cdot \mid z, o)$, a realized marginal resource vector*

$$\kappa(o, z) \in \mathbb{R}_+^m,$$

and an effect signature

$$\chi(o, z) = (\text{effect class, data labels, destination, authority, reversibility}).$$

Representative effect classes include internal computation, external read, disclosure, write, transaction, actuation, delegation, and governance-state modification.

The router can select any operation in the registered portfolio. The gate is invoked whenever the effect policy for $\chi(o, z)$ requires mediation. Purely internal sandboxed reasoning may be governed by a permissive tier; a remote retrieval containing protected data may require disclosure controls; a credential grant or policy change occupies the highest tier because it changes future authority.

Definition 3.4 (Metacognitive router). *A metacognitive router is a policy μ that maps state z to an operation $o \in \mathcal{O}(z)$ or to defer. It allocates cognition, evidence acquisition, and escalation according to expected downstream decision value and resource cost.*

Definition 3.5 (Effect-aware commitment gate). *An effect-aware gate is an enforcement mechanism*

$$G(o, z, h) \in \{\text{allow, deny, escalate}\},$$

where h is the attempt history. The gate evaluates the operation’s effect signature, authority context, policy, evidence record, data provenance, destination, and relevant risk. The use of h makes prior denials, failed checks, and related proposals part of the decision rather than allowing every retry to appear fresh.

3.5 Protected governance state

The gate cannot secure a system if the agent can rewrite the policy it is judged against. Let the trusted governance state be

$$\mathcal{T}_t = (G_t, P_t, \alpha_t, A_t, \mathcal{I}_t, \mathcal{W}_t, \mathcal{R}_t),$$

where P_t is policy, α_t authority and capability state, A_t the protected audit/evidence-attempt record, $\mathcal{I}_t$ assigns candidate identity, $\mathcal{W}_t$ assigns workflow lineage, and $\mathcal{R}_t$ maintains reservations and one-use execution authorizations. A change to P_t , α_t , or the gate itself is an effect-typed governance operation and must be authorized under a prior trusted state. This is the agentic analogue of the reference-monitor requirement that enforcement be protected from tampering [39].

The router and gate therefore share information without sharing trust. The router can observe enough policy state to plan productive cognition; the gate remains the authoritative interpreter of what may acquire effect. How much policy detail the router sees is itself a design variable. An opaque interface can reduce probing but wastes cognition on infeasible plans; a fully transparent interface improves value-of-computation estimates but can expose a sharper optimization target. We call this the *policy-visibility dial* and treat it as an experimental factor in Section 9.

Human approval is also an indirect write channel into protected governance state. If a model can persuade an approver to expand authority, modify policy, or issue a certificate, then “outside the direct write set” is not sufficient. High-tier approval therefore requires a protected review protocol: structured evidence, explicit requested state change, bounded agent-generated advocacy, independent context where appropriate, and an audit record of the authority transition. This channel is included in the adversarial evaluation contract.

Evidence belongs to a gate-assigned effect identity, not a router-supplied name. Splitting an unchanged candidate must not restart its evidence process; merging a changed effect into an old identity must not transfer evidence without a validity argument. The reference implementation uses exact binding equality, a conservative alternative to semantic equivalence. Child sessions inherit the workflow’s reservations. A request for a new root budget is a privileged owner operation, not a router reset.

An authorization binds payload, target version, policy/verifier version, dependencies, evidence digest, expiry, reservation, and a one-use nonce. The executor rechecks these at use time. A correct test for a stale target is not evidence for the new target. Registration and budget reservation, and subsequently execution and authorization consumption, require atomicity within their declared execution model. Section 9.3 supplies an in-memory realization and explicit stale-state, replay, and concurrency tests.

This separation creates both the performance opportunity and the strategic interaction developed next.

4 Gate-Coupled Metareasoning and Compute per Decision

4.1 A state-dependent resource ledger

Test-time compute is often counted in generated tokens or samples. Agentic systems instead trade among token generation, retrieval, code execution, specialist models, network service, memory traffic, latency, monetary spend, and human attention. Serving optimizations such as IO-aware attention and paged KV-cache management further show that the physical cost of nominally similar model operations can depend strongly on system state [54, 55].

There are two distinct accounting cases. If an episode owns its resources in isolation, its physical resource use is additive across realized operations. If multiple decisions share batching, caches, accelerators, networks, or human queues, the joint physical cost need not decompose uniquely into per-decision costs. A per-decision ledger then requires a declared allocation rule rather than an implicit claim that isolation marginals add to the shared total. Classical Shapley-style allocations provide one family of principled cost-sharing rules [56]; other deployments can use measured incremental cost, reservation accounting, or service-specific allocation so long as the rule is explicit.

Definition 4.1 (Compute-per-decision ledger). *Let ψ be the declared resource-attribution rule. For a*

completed episode d with operations $o_1, \dots, o_T$, define

$$K_\psi(d) = \sum_{t=1}^T \kappa_\psi(o_t, z_t) \in \mathbb{R}_+^m,$$

where each $\kappa_\psi(o_t, z_t)$ is the episode's attributed resource share under the realized serving state. The allocation rule must reconcile with the joint measured resource total over the accounting window. A deployment-specific nonnegative price vector $\lambda \in \mathbb{R}_+^m$ induces priced expenditure $C_{\lambda, \psi}(d) = \lambda^\top K_\psi(d)$. Hard resource limits remain constraints rather than prices.

The realized ledger is a retrospective accounting object. The router acts *ex ante* and should instead plan against an estimate of the *incremental decision cost*,

$$\hat{\kappa}^{\text{marg}}(o, z, s),$$

the additional machine and human resource burden caused by taking operation o now under serving state s . This prospective marginal estimate is the economically relevant decision price. It need not equal the retrospective attributed share κ_ψ : batching, cache reuse, joint queues, and cost-sharing rules can make allocated cost differ substantially from the causal marginal cost of one more operation. Prediction error in $\hat{\kappa}^{\text{marg}}$ is itself measurable. The separation preserves per-decision accountability without using an accounting allocation as the router's control price.

4.2 Decision value under an evidence-conditioned gate

Let $u(a, \omega)$ denote router utility (the principal's assessment is recorded separately) for a candidate commitment a , and let β_z denote the agent's posterior at state z . For history h , write

$$\mathcal{A}_G(z, h) = \{a : G(a, z, h) = \text{allow}\}.$$

An “escalate” gate outcome is not counted as an admissible commitment; it creates a further approval or evidence-acquisition operation in $\mathcal{O}(z)$ and is valued through the multi-step controller below. Include defer as a terminal option of value $u_{\text{def}}(z)$. For any admissible set A and posterior β , define

$$Q(A, \beta) = \max \left\{ u_{\text{def}}(z), \sup_{a \in A} \mathbb{E}_{\omega \sim \beta} [u(a, \omega)] \right\}. \quad (1)$$

For a transition $z \rightarrow z'$ induced by an operation, let $A_0 = \mathcal{A}_G(z, h)$, $A_1 = \mathcal{A}_G(z', h')$, $\beta_0 = \beta_z$, and $\beta_1 = \beta_{z'}$. Define the four corner values

$$Q_{00} = Q(A_0, \beta_0), \quad Q_{10} = Q(A_0, \beta_1), \quad Q_{01} = Q(A_1, \beta_0), \quad Q_{11} = Q(A_1, \beta_1).$$

Definition 4.2 (Order-explicit and symmetric gate-coupled value attribution). *The belief-first sequential attribution is*

$$\Delta_{\text{dec}}^{B \rightarrow A} = Q_{10} - Q_{00}, \quad \Delta_{\text{adm}}^{B \rightarrow A} = Q_{11} - Q_{10}.$$

The admissibility-first attribution is

$$\Delta_{\text{adm}}^{A \rightarrow B} = Q_{01} - Q_{00}, \quad \Delta_{\text{dec}}^{A \rightarrow B} = Q_{11} - Q_{01}.$$

Their difference is the interaction term

$$I = Q_{11} - Q_{10} - Q_{01} + Q_{00}.$$

A symmetric two-factor Shapley attribution assigns

$$\phi_{\text{dec}} = \frac{1}{2} [(Q_{10} - Q_{00}) + (Q_{11} - Q_{01})], \quad (2)$$

$$\phi_{\text{adm}} = \frac{1}{2} [(Q_{01} - Q_{00}) + (Q_{11} - Q_{10})], \quad (3)$$

so that $\phi_{\text{dec}} + \phi_{\text{adm}} = Q_{11} - Q_{00}$.

The sequential split is useful when the implementation truly updates beliefs before reevaluating policy. It is not an intrinsic decomposition: reversing the order reallocates the interaction term. For empirical attribution we therefore recommend reporting the interaction explicitly or using the symmetric split. This removes an otherwise arbitrary index-number choice from comparisons across routers.

Lemma 4.3 (Nonnegative fixed-set value of information). *Suppose A_0 is fixed, utilities are integrable, u_{def} is unchanged by the observation, and β_1 is a coherent Bayesian posterior generated by an observation from β_0 . Then*

$$\mathbb{E}[Q(A_0, \beta_1) \mid z, o] \geq Q(A_0, \beta_0).$$

Hence the expected belief-first decision-refinement term is nonnegative. The admissibility term need not be: evidence can remove authority as well as grant it.

Proof. For fixed A_0 , $Q(A_0, \beta)$ is the pointwise supremum of linear functionals of the posterior together with the defer value, hence is convex in β . Posterior beliefs form a martingale under coherent Bayesian updating. Jensen's inequality gives the result. $\square$

Define gate-coupled value of computation by the total successor value, with the symmetric terms used only for attribution:

$$\text{VoC}_G(o \mid z) = \mathbb{E}[Q_{11} - Q_{00} \mid z, o] - \boldsymbol{\lambda}^\top \hat{\boldsymbol{\kappa}}^{\text{marg}}(o, z, s) = \mathbb{E}[\phi_{\text{dec}} + \phi_{\text{adm}} \mid z, o] - \boldsymbol{\lambda}^\top \hat{\boldsymbol{\kappa}}^{\text{marg}}(o, z, s). \quad (4)$$

The first component captures decision refinement; the second captures gain or loss created by changed admissibility. Evidence can therefore be valuable even when it barely changes the ranking of actions, and it can be unattractive to the router when its main effect is to remove an otherwise profitable action.

4.3 One-step escalation and its limit

If the router may either commit now or take one operation and then commit, operation o is preferred exactly when $\text{VoC}_G(o \mid z) > 0$. This is the one-step rule. Rational-metareasoning approaches instantiate related objectives for reasoning tokens [16]; the present objective admits heterogeneous operations and a changing admissible set.

One-step control is nevertheless myopic. Evidence contracts often require complementary checks.

Proposition 4.4 (Myopic failure under conjunctive evidence). *Suppose a high-value action a^* is inadmissible until two evidence predicates E_1 and E_2 are both satisfied. At state z_0 , neither is satisfied. Operation o_i deterministically establishes E_i at priced cost $c_i > 0$, neither operation changes the value of any currently admissible action, and satisfying only one predicate does not admit a^* . Let the value*

increase after both predicates hold be $U > 0$. Then each one-step value at z_0 is $-c_i < 0$, while the two-operation plan has net value $U - c_1 - c_2$ and outperforms stopping whenever $U > c_1 + c_2$. Global optimality additionally requires excluding better available plans.

Proof. After either operation alone, the admissible action set and current-action values are unchanged, so the one-step gain is zero before cost. After both operations, a^* enters the admissible set and increases decision value by U . Subtracting total cost yields the result. $\square$

The proposition gives a concrete reason that routers trained only on immediate improvement can under-think: some computations are valuable only as members of an evidence chain. Russell–Wefald already identify limits of myopic metareasoning; evidence-gated admissibility supplies an agentic instance in which the complementarity is explicit.

Proposition 4.5 (Binding-gate condition for governance-induced strategic ignorance). *Fix an underlying candidate action set $\mathcal{A}$ and an evidence operation o whose decision-relevant effects are posterior updating and gate admissibility. Define the unrestricted value*

$$\overline{Q}(\beta) = \max \left\{ u_{\text{def}}, \sup_{a \in \mathcal{A}} \mathbb{E}_{\omega \sim \beta} [u(a, \omega)] \right\}.$$

Suppose that at the current state and at every successor observation reachable under o , the gate-admissible set contains at least one maximizer of $\overline{Q}$ at that posterior. If updating is coherent Bayesian updating and u_{def} is observation independent, then

$$\mathbb{E}[Q(\mathcal{A}_G(Z'), \beta_{Z'}) \mid z, o] \geq Q(\mathcal{A}_G(z), \beta_z),$$

and therefore

$$\text{VoC}_G(o \mid z) \geq -\lambda^\top \hat{\kappa}^{\text{marg}}(o, z, s).$$

Thus evidence can acquire an additional negative value through governance only if the gate is binding at the current state or at some reachable successor state.

Proof. Under the stated nonbinding condition, gated and unrestricted values coincide at the current and successor states. The unrestricted value is the supremum of posterior-linear action values and the observation-independent defer value, so the fixed-set value-of-information argument of Lemma 4.3 gives nonnegative expected successor improvement before resource cost. Subtracting the prospective marginal cost yields the inequality. $\square$

Scope and diagnostic. For a multi-step continuation, the corresponding nonbinding condition must hold at every reachable history under that continuation, with the same underlying operations, costs, and defer semantics. A binding gate is a necessary condition for governance-induced negative information value; a raw skipped-check count is not a measurement of value misalignment.

Let $W(z)$ be unrestricted terminal value and let $B_G(z) = W(z) - Q_G(z) \geq 0$ denote the value lost to restriction. Under the fixed-task, pure-information assumptions,

$$\mathbb{E}[Q_G(Z')] - Q_G(z) = (\mathbb{E}[W(Z')] - W(z)) - (\mathbb{E}[B_G(Z')] - B_G(z)). \quad (5)$$

A governance-induced loss occurs when the increase in expected restriction loss exceeds the unrestricted information benefit. Even positive gross information value can be smaller than acquisition cost, yielding

ordinary rational skipping.

Observation 4.6 (Counterfactual strategic-ignorance diagnostic). Compare the same check, information kernel, costs, alternatives, and router with a gate that responds to the resulting evidence and a shadow gate that does not impose that evidence-dependent restriction. If the check is preferred in the shadow condition but skipped in the responsive condition, the contrast identifies an avoidance incentive caused by that governance response within the controlled model. Without this contrast, skipping can also reflect cost, deadlines, alternative checks, or mistaken estimates of informativeness.

A router–gate wedge can reflect different utilities, different beliefs, or an authority requirement not reducible to a shared risk price. The last case is a *rule-form wedge*: a hard admission threshold or prerequisite need not implement the Bayesian action rule even with shared beliefs and utility. Section 9.4 supplies an aligned-objective instance. In a threshold-risk model, supplying a calibrated shared posterior removes the belief difference; varying the router’s downside price then isolates the utility intervention. Equalizing utilities while varying private information or calibration studies the belief intervention. Legitimate unverifiable information is different from overconfidence. Interactions between interventions should be reported rather than assumed additive. Mandatory checks remove a discretion channel; objective shaping, calibration, and evidence externalization address different possible causes, while the independent gate remains in place.

The thresholded admission value itself is generally *not* adaptive submodular: a conjunction or posterior threshold can assign nearly zero value to each early check and then jump when the final requirement is satisfied. Adaptive-submodularity guarantees therefore apply only to a separately declared surrogate evidence objective—for example expected log-likelihood-ratio or e-value progress toward the threshold—when that surrogate is adaptive monotone and adaptive submodular [57]. In that case greedy acquisition can inherit the classical approximation guarantee for the surrogate, while mandatory disqualifying checks handle the discontinuous authority condition. The design principle is consequently more precise: smooth the *search objective* when possible; do not pretend the gate’s thresholded value has diminishing returns when it does not.

4.4 Multi-step control

For a finite horizon, the normative value function is

$$V(z) = \max \left\{ Q(\mathcal{A}_G(z, h), \beta_z), \sup_{o \in \mathcal{O}(z)} \left[\mathbb{E}[V(Z') \mid z, o] - \lambda^\top \hat{\kappa}^{\text{marg}}(o, z, s) \right] \right\}. \quad (6)$$

The expectation is subjective to the deployed controller and therefore inherits calibration error. Difficulty prediction, uncertainty estimates, empirical improvement curves, and learned critics are all estimators of pieces of this value function, not substitutes for it.

Fan et al.’s shared-budget results show that strong reasoning ability does not imply effective global resource allocation [30]. The control plane predicts an additional failure: even a router that allocates tokens well can remain governance blind if it does not model how evidence and authority alter future admissibility.

4.5 Human attention is a resource with endogenous quality

Human review is not an unlimited oracle. It is an expensive operation whose marginal latency and reliability can depend on review volume, task complexity, and workload. Human-factors research

documents automation bias and workload-sensitive failures in oversight and decision support [58, 59]. The serving state s_t therefore includes human-queue state when human approval is part of the gate.

A router that escalates every uncertain case can reduce system safety by saturating the very reviewer used as the highest assurance tier. This congestion externality cannot be represented faithfully by a context-free single-episode price. The focal episode’s own review time and direct error cost already belong in $\hat{\kappa}^{\text{marg}}$. The additional governance price should therefore be Pigouvian: let $C_H^{-d}(q_H)$ denote the expected burden borne by *other* episodes at human-review queue state q_H . Then the external marginal price of one more escalation is

$$\lambda_H^{\text{ext}}(q_H) = C_H^{-d}(q_H + 1) - C_H^{-d}(q_H).$$

Adding λ_H^{ext} to the focal decision price internalizes congestion without double-counting the focal episode’s own review burden. A deployment that optimizes a fleet-level objective over concurrent episodes can represent the same externality directly. This makes “ask a human” part of metareasoning rather than a free terminal action.

5 Verification, Evidence Contracts, and Gate Gaming

5.1 More reasoning is not more evidence

Reasoning traces can be useful computational scaffolds, but length and fluency are not verification. Models can fail to self-correct without external feedback [35], and natural-language explanations can be unfaithful to the causal factors behind an output [36]. The control plane therefore distinguishes operations that search a hypothesis space from operations that create evidence under a declared mechanism.

Definition 5.1 (Evidence update). *An operation v is an evidence update for claim or action a if it produces observation o through a declared mechanism M_v and updates*

$$e' = \text{Update}(e, a, o, M_v, \text{prov}, h),$$

where *prov* records provenance and *h* records the complete attempt history relevant to the evidence contract.

The mechanism class matters. A primary-source retrieval, independent recomputation, executable test, symbolic proof, same-model critique, and human review have different error correlations. The gate should reason over those differences rather than collapsing every check into a scalar “confidence.”

5.2 Evidence contracts

An evidence contract specifies what must be observed before an operation becomes admissible. Representative requirements include source identity and freshness, executable postconditions, independence or diversity of checks, minimum sample size, approval identity, and failure-count semantics. The contract can be conjunctive, thresholded, or sequential. Its structure enters the router’s value function because it determines Δ_{adm} .

A well-formed contract also specifies negative evidence. A failed unit test, contradicted source, rejected approval, or verifier disagreement is part of the evidence state. Discarding failed attempts and retaining only the first successful check creates an obvious selection channel. When safety requires a particular

disqualifying check, its completion must be a prerequisite rather than an optional benefit. Other checks may remain optional; the mere possibility that a check revokes authority does not by itself justify making every such check mandatory. Contracts should distinguish *required checks*, whose absence is itself a denial condition, from *optional evidence*, whose value the router is free to optimize.

5.3 Gate gaming as a predicted systems failure

Once the router benefits when evidence changes admissibility, the gate becomes part of the environment being optimized. Several strategies can become profitable, depending on objectives, costs, available operations, and weaknesses of the evidence contract.

- **Evidence shopping:** search sources or formulations until one satisfies the contract while ignoring contrary evidence.
- **Verifier retry:** repeatedly invoke a stochastic or noisy checker until a favorable sample appears.
- **Gate probing:** use denials, error messages, or tier changes to infer the policy boundary and move toward the nearest admissible variant.
- **Contract Goodharting:** satisfy nominal evidence requirements while violating their epistemic intent, for example by presenting two nominally distinct but causally dependent sources as independent corroboration.
- **Strategic ignorance:** avoid a check whose unfavorable result could remove authority when the contract rewards favorable evidence but does not require the check to be attempted.

These are conditional incentive predictions. Coupling alone does not make gaming profitable: sound evidence, aligned objectives, unavailable exploits, or sufficient acquisition cost can remove the pressure. The reference examples identify concrete regimes in which it appears; favorable legitimate evidence acquisition is not itself circumvention.

The corresponding architectural response is to make attempt history and provenance first-class state. Retry budgets, sticky or stateful denials, correlated-source detection, logging of failed checks, randomized audit, and escalation after repeated near-miss proposals all reduce the gap between the contract’s literal predicate and its intended epistemic meaning. No single countermeasure is universal; the important point is that the gate evaluates the *history of evidence acquisition*, not only the final favorable artifact.

The same interaction has a close information-economics analogue. Bayesian persuasion studies how a sender selects information to influence a receiver’s action [46]; verifiable-disclosure models study how selectively revealed evidence changes receiver beliefs [47, 48]. In the agentic setting, the router is both experiment selector and prospective actor, while the gate is the receiver/enforcer. History completeness limits selective disclosure by making the chosen experiment and its unfavorable outcomes part of the receiver’s information set.

5.4 Policy visibility is a safety–efficiency dial

Gate-coupled metareasoning requires some model of how evidence changes admissibility. That creates a genuine design tension. More policy visibility lets the router avoid useless work, choose the right verifier, and satisfy legitimate evidence requirements efficiently. The same visibility can make denial boundaries easier to probe or optimize against. We therefore distinguish three experimental interfaces: *opaque* (only allow/deny/escalate), *contract-visible* (required evidence types and authority tiers are visible, but not semantic classifier internals), and *fully modeled* (the router receives or learns a high-fidelity model of the gate). The utility–gaming frontier across these interfaces is a primary evaluation target rather than an

implementation detail.

5.5 Chain-of-thought monitoring as an evidence channel

The framework does not require a faithful chain of thought for containment, but it can use reasoning traces as a monitorable signal. Baker et al. show that LLM monitors can detect reward-hacking behavior from chain-of-thought and that excessive direct optimization pressure can induce obfuscation [37]. Korbak et al. argue that chain-of-thought monitorability is a useful but fragile safety opportunity [38].

Accordingly, a CoT monitor is one verifier in the portfolio. It can trigger additional checks, deny or escalate a commitment, or raise the estimated value of human review. Its output carries a measured false-positive/false-negative profile and should not be treated as privileged access to ground-truth intent. The strongest guarantees remain anchored to effect mediation, authority, evidence provenance, and explicit policy.

6 Runtime Governance, Containment, and Information Flow

6.1 The reference-monitor trust boundary

The commitment gate inherits the classic reference-monitor discipline. The enforcement path must be invoked on every protected effect, protected from agent modification, and structured so that its policy and implementation can be analyzed and tested [39]. Complete mediation and least privilege then constrain what authority is available at the boundary [40].

This is stricter than saying that the agent “should ask permission.” Policy P_t , authority state α_t , the gate implementation, and the audit record belong to the protected governance state $\mathcal{T}_t$. If the agent requests a change to those objects, that request is itself a high-tier commitment evaluated under the prior trusted state.

6.2 Risk and control tier

Let the pre-commitment risk profile be

$$\mathbf{r}(o, z) = (s(o, z), i(o, z), \xi(o, z), p(o, z), f(o, z)), \quad (7)$$

where s denotes stakes, i irreversibility, ξ epistemic uncertainty, p privilege exercised, and f information-flow sensitivity. The coordinates need not be scalarized. A governance policy can map the partial order into tiers such as permissive sandbox, ordinary execution, verified execution, independent approval, or prohibition.

A *compensating condition* is now explicit: it is a policy predicate whose satisfaction authorizes a lower control tier despite an increase in one risk coordinate, for example a cryptographically scoped capability, a reversible dry-run environment, or an evidence certificate whose validity domain covers the requested operation. The condition is part of policy, not an ad hoc router exception.

6.3 Deterministic commitment containment

Let an external effect trace be $\tau = (a_1, \dots, a_T)$ and let $\mathcal{P}$ be the prefix-closed set of permitted traces under a fixed governance contract.

Assumption 6.1 (Complete mediation). *Every effect classified as requiring commitment control reaches the gate before the effect occurs. No stale capability, hidden network path, delegated subagent, human-facing output, or other in-scope side effect bypasses the gate.*

Assumption 6.2 (Protected governance state). *The gate, policy, authority state, and audit record are outside the agent's direct write set. Every requested change to them is itself mediated under the prior trusted state.*

Assumption 6.3 (Gate soundness). *For every permitted prefix $\tau \in \mathcal{P}$, the gate authorizes next effect a only if $\tau \parallel a \in \mathcal{P}$.*

Theorem 6.4 (Commitment containment under a protected reference monitor). *If the empty trace belongs to $\mathcal{P}$, then under complete mediation, protected governance state, and gate soundness, every externally committed trace generated by the agent remains in $\mathcal{P}$ regardless of the internal routing or reasoning policy.*

Proof. Proceed by induction on committed trace length. The empty trace is permitted. Assume $\tau_t \in \mathcal{P}$. Complete mediation sends every proposed next effect to the gate. A denial leaves the trace unchanged. An authorization extends the trace only when gate soundness establishes $\tau_t \parallel a_{t+1} \in \mathcal{P}$. Protected governance state prevents the agent from changing the predicate or enforcement mechanism outside this same mediated process. Hence every reachable committed trace is permitted. $\square$

The theorem is an architectural invariant. It says exactly where the trust burden sits: policy adequacy, gate correctness, governance-state integrity, and mediation coverage. It deliberately makes no assumption about whether the model's private reasoning is aligned.

6.4 Probabilistic semantic gates, adaptive search, and history completeness

Many consequential decisions cannot be authorized by syntax or capability alone. A gate that must judge whether a free-text message contains restricted information, whether a proposed patch creates a vulnerability, or whether a summary is materially misleading will itself use fallible semantic models. Two different guarantees are useful.

Assumption 6.5 (Conditional probabilistic soundness). *At gate attempt t , conditional on the complete history H_t , let $\epsilon_t(H_t)$ upper-bound the probability that the gate authorizes an unsafe proposed effect:*

$$\Pr(F_t \mid H_t) \leq \epsilon_t(H_t).$$

Theorem 6.6 (Episode-level unsafe-acceptance bound). *Under complete mediation, for any adaptive router making at most T gate attempts,*

$$\Pr\left(\bigcup_{t=1}^T F_t\right) \leq \sum_{t=1}^T \mathbb{E}[\epsilon_t(H_t)]. \quad (8)$$

If deterministic bounds $\epsilon_t(H_t) \leq \bar{\epsilon}_t$ hold for every reachable history, the right side is at most $\sum_t \bar{\epsilon}_t$.

Proof. The union bound gives $\Pr(\cup_t F_t) \leq \sum_t \Pr(F_t)$. By the tower property, $\Pr(F_t) = \mathbb{E}[\Pr(F_t \mid H_t)] \leq \mathbb{E}[\epsilon_t(H_t)]$. $\square$

The uniform-over-history condition is strong and load bearing. A deterministic classifier with a searchable blind spot does not satisfy a small $\bar{\epsilon}_t$ merely because its average i.i.d. test error is small: adaptive proposal search can concentrate on the blind spot. Independence is likewise not the pessimistic case; correlated failure modes plus search can be worse.

Observation 6.7 (Attempt-local retry amplification). If a history-blind gate treats k unsafe proposals as independent fresh attempts, each with false-accept probability ϵ , then

$$\Pr(\text{at least one false acceptance}) = 1 - (1 - \epsilon)^k. \quad (9)$$

For small ϵ , the first-order increase is approximately $k\epsilon$.

A gate can do fundamentally better when evidence remains valid under the permitted selection process. Complete logging supports this invariant; it does not substitute for it. A sufficient statistic may preserve validity without retaining every raw observation in active state.

Theorem 6.8 (Anytime-valid admission from a selection-valid test supermartingale). *Let U denote that one candidate effect is unsafe. Let $(E_t, H_t)_{t \geq 0}$ be a nonnegative process adapted to an evidence filtration sufficient for the permitted adaptive selection, with $E_0 \leq 1$, such that under U*

$$\mathbb{E}_U[E_t \mid H_{t-1}] \leq E_{t-1}$$

for every adaptively selected experiment and every reachable history. If the gate admits the candidate only when $E_t \geq 1/\alpha$ for some $\alpha \in (0, 1)$, then

$$\Pr_U(\text{ever admitted}) \leq \alpha. \quad (10)$$

The bound is independent of the number of attempts and remains valid under adaptive experiment selection and stopping.

Proof. (E_t) is a nonnegative supermartingale under U . Ville's inequality gives

$$\Pr_U\left(\sup_t E_t \geq 1/\alpha\right) \leq \alpha.$$

Admission is a subset of this threshold-crossing event. $\square$

A correctly specified Bayesian likelihood ratio is one important specialization. Let the gate use design prior $\pi_0 \in (0, 1)$, safe threshold $\theta > \pi_0$, and the complete-history likelihood ratio

$$L_t = \prod_{s=1}^t \frac{p_S(O_s \mid H_{s-1}, E_s)}{p_U(O_s \mid H_{s-1}, E_s)}.$$

When (L_t) is a valid U -martingale (or conservative U -supermartingale), the posterior-threshold rule is equivalent to $L_t \geq 1/\alpha$ with

$$\alpha = \frac{\pi_0}{1 - \pi_0} \frac{1 - \theta}{\theta}. \quad (11)$$

This recovers the previous Bayesian bound while making clear that absolute continuity and exact martingale equality are not essential; supermartingale validity is sufficient. The candidate-conditional bound depends on validity under U , not on π_0 being the true population base rate. Calibration of $\Pr(U \mid \text{admitted})$, however, depends on the actual unsafe-proposal prevalence, which an adaptive router can change.

Theorem 6.9 (Episode-level proposal-budget bound). *Consider an episode that adaptively generates candidate effects $j = 1, \dots, N$, where N is finite almost surely and may itself depend on prior history. Let candidate identity and workflow lineage be assigned by protected governance, with evidence bound to the actual effect and no session-reset bypass. Before candidate j is evaluated, atomically assign a predictable false-admission budget $\alpha_j \geq 0$ such that*

$$\sum_{j=1}^N \alpha_j \leq \alpha_{\text{ep}}$$

almost surely. Suppose the candidate-level gate satisfies

$$\Pr(\text{admit candidate } j \mid U_j, H_{j-1}) \leq \alpha_j$$

for every reachable history, where U_j is the event that candidate j is unsafe. Then

$$\Pr(\text{at least one unsafe candidate is admitted in the episode}) \leq \alpha_{\text{ep}}. \quad (12)$$

Proof. Let $F_j = U_j \cap \{\text{candidate } j \text{ admitted}\}$. Conditional on H_{j-1} ,

$$\Pr(F_j \mid H_{j-1}) = \Pr(U_j \mid H_{j-1}) \Pr(\text{admit } j \mid U_j, H_{j-1}) \leq \alpha_j.$$

The union bound and tower property give

$$\Pr\left(\bigcup_{j \leq N} F_j\right) \leq \mathbb{E}\left[\sum_{j \leq N} \alpha_j\right] \leq \alpha_{\text{ep}}.$$

□

Theorems 6.8 and 6.9 separate two forms of adaptive search. Selection-valid evidence controls repeated collection for one fixed candidate. Proposal budgets control the creation of new variants—different patches, recipients, phrasings, transactions, or plans—that would otherwise start fresh evidence processes. Pooling evidence across semantically equivalent variants can conserve budget, but deciding that two proposals are “the same action” is itself a semantic-governance problem. The implementation uses predictable error spending with no refund on admission or denial. Online α -investing controls a different error criterion and may earn future allowance from discoveries [60]; admission itself is not such a safety credit here. Even an ex-post verification-based refund would need a separate composition proof. A convergent schedule makes legitimate revisions increasingly hard to admit as well as limiting probing; fixed finite reservations expose this false-block cost in the reference model.

The candidate-level guarantee identifies two principal ways gate gaming defeats anytime validity. *History incompleteness* lets the router hide failed attempts or selectively disclose favorable evidence. *Evidence-model misspecification* lets dependent, shifted, or adversarially selected evidence receive too much weight. A robust implementation can replace a point likelihood model with conservative e-values that remain supermartingales for every distribution in a registered unsafe-model uncertainty set. The same Ville

bound then holds uniformly over that set [61]. Adaptive routers should be evaluated precisely where calibration is weakest, because steering toward model error is itself a Goodhart pressure.

6.5 Constructing evidence that survives router selection

An evidence record and a statistically valid evidence process are different objects. A verifier can be correctly calibrated on a population and invalid on the subpopulation selected by a router. For example, let ten percent of unsafe candidates form a recognizable subtype that always passes a verifier, and let the other ninety percent never pass. The pooled unsafe pass rate is 0.1. A router restricted to the first subtype induces pass probability one. Logging all those passes does not repair a likelihood model that continues to assign them probability 0.1 under the unsafe hypothesis.

The relevant requirement is *validity over the conditional laws induced by permitted selection*. If a subtype is observable to trusted governance, separate bounds can preserve useful evidence on other subtypes. If only the router can recognize it, the uncertainty class must cover that selection. “Worst subtype” means the worst conditional law the admitted selection strategy can induce, not an unconditional claim that every deployment must use the globally worst verifier error.

Theorem 6.10 (Binary verification to selection-valid evidence). *Fix one candidate’s unsafe null class $\mathcal{U}$. Let a_t be the chosen verification operation and let $Y_t \in \{0, 1\}$ denote pass. Before observing Y_t , suppose a trusted bound $q_t \in (0, 1]$ satisfies*

$$\Pr_P(Y_t = 1 \mid \mathcal{H}_{t-1}, a_t) \leq q_t \quad \text{for every } P \in \mathcal{U}.$$

Here $\mathcal{H}_{t-1}$ includes the information used to select operations; any unobserved router-recognizable distinctions are covered uniformly by the bound. Choose a predictable $\lambda_t \in [0, 1/q_t]$ and set

$$e_t = 1 + \lambda_t(Y_t - q_t), \quad M_0 = 1, \quad M_t = M_{t-1}e_t.$$

Then M_t is a nonnegative test supermartingale under every $P \in \mathcal{U}$. Admission restricted to $M_t \geq 1/\alpha$ has candidate false-admission probability at most α , under adaptive operation choice and stopping.

Proof. The factor at failure is $1 - \lambda_t q_t \geq 0$ and the factor at pass is $1 + \lambda_t(1 - q_t) \geq 0$. Conditional expectation is

$$\mathbb{E}_P[e_t \mid \mathcal{H}_{t-1}, a_t] = 1 + \lambda_t(\mathbb{E}_P[Y_t \mid \mathcal{H}_{t-1}, a_t] - q_t) \leq 1.$$

Multiplying by the nonnegative predictable M_{t-1} proves the supermartingale property. Apply Theorem 6.8. $\square$

This is a conditional-betting construction from the established test-supermartingale toolkit [61], instantiated at the agent’s verification interface. It turns an operation-specific error bound into an executable update. It does not estimate that bound from average classifier accuracy. The construction uses $q_t > 0$; a claimed zero-probability unsafe pass requires its own exact-semantic justification, not merely zero observed calibration failures.

For a verifier with safe-pass design value $p_t \geq q_t$, $0 < q_t < 1$, choose

$$\lambda_t = \frac{p_t - q_t}{q_t(1 - q_t)}.$$

The pass and failure factors become p_t/q_t and $(1 - p_t)/(1 - q_t)$. Correct specification of p_t is not needed for null validity; the unsafe upper bound and allowed coefficient range do the work. In the blind subtype, $q_t = 1$ and a pass has factor one: a sure unsafe pass supplies no positive admission evidence. A different, more expensive verifier can still have $q_t < 1$ and useful power. Section 9.3 implements both cases with exact finite kernels.

6.6 Certificate risk and sequential admission risk compose at different levels

A competence certificate and an e-process control different probabilities. The former can bound failure of a selected deployment policy; the latter controls admission conditional on a candidate being unsafe. Calibration confidence is a third probability, shared across all uses of the calibrated mechanism. Conflating these quantities invalidates composition.

Theorem 6.11 (Calibration, selected-use certificates, and proposal spending). *Let D be a calibration record fixed before deployment and let $\mathcal{G}$ be a set of records with $\Pr(D \in \mathcal{G}) \geq 1 - \delta$. For every fixed $d \in \mathcal{G}$, assume the following conditions hold under the declared deployment law and permitted adaptive selection:*

- (i) *Statistical candidate tests satisfy the conditional guarantee of Theorem 6.9, with predictable allocations whose total is at most A .*
- (ii) *Each invoked certificate-based macro i has a conditional unsafe-execution bound b_i given its selected entry history; the predictable allowances over all such invocations total at most B .*
- (iii) *Candidate and workflow identities, evidence state, reservations, and execution bindings are enforced by the protected mechanism. Every effect belongs to one of the two admission routes.*

Then, over this stated deployment scope,

$$\Pr(\text{at least one unsafe execution}) \leq \min\{1, \delta + A + B\}.$$

Proof. Fix a calibration record $d \in \mathcal{G}$. Conditional proposal spending bounds the union of unsafe statistical-route admissions by A . For certificate invocation i , the tower property bounds its unsafe-execution probability by the expectation of its selected-history allowance; summing gives at most B . Binding admission to execution permits the same union bound for executed effects. Integrating over good calibration records and adding $\Pr(D \notin \mathcal{G}) \leq \delta$ gives the result. $\square$

The guarantee for a good calibration record is uniform over the declared future selection process. Conditioning on an event involving future outcomes would not provide the same argument. Certificate failure is a shared uncertainty event, not fresh independent noise on each use. Separate calibration records require their own joint good-event accounting. The theorem does not convert an average population-risk bound into condition (ii): that bridge must be established for the routed policy or admitted entry histories.

In the reference environment, probabilities are specified exactly rather than estimated: $\delta = 0$ within that model. Statistical proposals spend $A = 0.12$. A fast macro is permitted only at a fresh routine entry point with conditional unsafe probability 0.04, and it can run at most once, so $B = 0.04$. The resulting loose workflow guarantee is 0.16. The statistical-only guarantee is 0.12; it must not be reported as the certificate route's unsafe-conditional guarantee. Real learned verifiers would require a calibrated or otherwise justified good-record event and a validity domain covering their router's selections.

6.7 Trace safety, confidentiality, and hyperproperties

Theorem 6.4 is a theorem about permitted traces. Confidentiality and noninterference can require relational properties across multiple executions: whether the observable output changes as a function of secret input. Such policies are hyperproperties rather than ordinary trace properties [45].

The control plane handles this by carrying data labels and provenance in the effect signature and by allowing the gate to enforce a separate information-flow policy $\mathcal{F}$. CaMeL provides a concrete precedent: trusted control and data-flow structure plus capabilities can block exfiltration even when the underlying LLM sees adversarial data [42]. The trace theorem and the flow policy therefore protect different objects. A system should not claim confidentiality from trace containment alone.

Label propagation is necessarily conservative. Once a model consumes protected or untrusted data, downstream values may inherit those labels and utility can collapse unless the system defines a *declassification* rule. Declassification is therefore not a bookkeeping exception: it is a privileged policy decision that asserts why a derived value may cross a boundary despite its lineage. The declassifier should be explicit about transformation, permitted destination, supporting evidence, and residual risk, and its decisions should be mediated and audited like other authority-changing operations. Flow-aware mediation moves part of the hard problem into declassification policy; it does not make that problem disappear.

6.8 Actions are not outcomes

The gate controls requested effects; the environment controls their consequences. Let $x_{t+1} \sim P_{\text{env}}(\cdot \mid x_t, a_t)$ be the world transition after an authorized action. Two identical actions can produce different outcomes in different states. Runtime governance must therefore choose $\mathcal{P}$ as a conservative proxy for outcome safety and, where necessary, include precondition checks, simulation, rollback, and postcondition monitoring in the effect contract. Commitment containment is a guarantee about mediated agency, not omniscience about the world.

7 Distillation and the Moving Fast–Slow Boundary

7.1 A developmental flywheel

The dual-process analogy becomes especially useful when learning is added. Human skills often move from effortful practice toward automaticity: procedures that initially consume conscious attention can become fast and habitual with successful repetition. Modern AI has followed a different construction path in important respects. Broad amortized competence emerged first from large-scale pretraining, while explicit test-time deliberation, tool use, search, and verification were layered on afterward. The two paths meet in the same place: successful slow cognition can be compiled back into faster behavior.

STaR and related approaches bootstrap future performance from successful rationales [49]; System-2-to-System-1 distillation explicitly studies compiling deliberate procedures into faster model behavior [50]. The control-plane loop is

System-1-like proposal $\rightarrow$ System-2-like deliberation $\rightarrow$ verification
 $\rightarrow$ accepted episode $\rightarrow$ distillation $\rightarrow$ improved fast path.

The boundary between fast and slow competence is therefore endogenous. Today’s expensive deliberation can become tomorrow’s amortized intuition, while the slow path moves outward toward harder tasks. Metacognitive routing is therefore a moving-boundary problem rather than a fixed choice between two permanent modules.

7.2 Cost-sensitive replacement without double counting

Let π_s be a slow policy and π_d a distilled policy. For task distribution D , let $L_D(\pi)$ be expected decision loss and let $K_D(\pi) \in \mathbb{R}_+^m$ be expected resource use before commitment. For a deployment price vector λ , define

$$J_\lambda(\pi) = L_D(\pi) + \lambda^\top K_D(\pi). \quad (13)$$

Observation 7.1 (Cost-sensitive competence replacement). If

$$L_D(\pi_d) - L_D(\pi_s) < \lambda^\top (K_D(\pi_s) - K_D(\pi_d)),$$

then $J_\lambda(\pi_d) < J_\lambda(\pi_s)$.

Proof. Subtract $J_\lambda(\pi_s)$ from $J_\lambda(\pi_d)$ and apply the inequality. $\square$

This is the economic condition for moving competence to the fast path. It says nothing yet about authority.

7.3 Evidence does not automatically distill

A slow episode may be admissible precisely because it generated runtime evidence: a fresh source, test result, approval, or environment observation. A distilled model can learn the answer without possessing that evidence.

Observation 7.2 (Evidence non-transfer under an unchanged gate). Suppose gate policy G authorizes commitment a only when runtime evidence predicate $R(e, a) = 1$. Let the slow policy produce a together with evidence e_s satisfying $R(e_s, a) = 1$. If a distilled policy produces the same a but an evidence state e_d with $R(e_d, a) = 0$, then under the unchanged gate the distilled policy remains unable to commit a , regardless of its predictive accuracy.

Proof. The commitment condition is $R(e, a) = 1$. The distilled policy does not satisfy it, so the gate denies the action by definition. $\square$

The result separates *competence distillation* from *authority distillation*. The latter requires a policy change, and that change needs its own justification.

7.4 Scoped competence certificates

One legitimate policy change is to let training-time verification support a bounded runtime certificate. Define a competence certificate

$$\mathcal{C} = (D_{\text{task}}, M, \nu, \epsilon, \delta, A, P, t_{\text{exp}}, \Delta),$$

where D_{task} is the covered task domain, M the model/version identity, ν the verification procedure, (ϵ, δ) the declared statistical guarantee, A the assumptions under which it is valid, P the policy version, t_{exp} an expiry or review horizon, and Δ drift or revocation triggers. A representative PAC-style statement is

$$\Pr_{\mathcal{V}}(R_{D_{\text{task}}}(M) \leq \epsilon) \geq 1 - \delta,$$

where $\mathcal{V}$ denotes the certificate-issuance procedure and R is the registered failure risk. Conformal or learn-then-test methods can provide finite-sample risk guarantees under their stated assumptions [62, 63]. The gate may substitute the certificate for runtime checks only after its scope includes the actual routing selection and effect policy. A population bound on M alone does not supply that bridge; Section 7.5 specifies the additional requirement.

Issuing, renewing, expanding, or revoking $\mathcal{C}$ changes future authority and is therefore itself a governance act. Certificate issuance must be mediated under the prior trusted state; the model being certified cannot unilaterally promote its own training evidence into runtime authority.

This mechanism makes the moving boundary operational. Stable classes with reproducible verification can migrate toward low-cost execution. Freshness-sensitive tasks, open-world factual claims, changed policies, model updates, violated statistical assumptions, or detected distribution drift invalidate the certificate and reopen the slow path. The learning loop therefore becomes

$$\text{reason} \rightarrow \text{verify} \rightarrow \text{distill} \rightarrow \text{certify} \rightarrow \text{monitor drift} \rightarrow \text{reverify}.$$

The scientific prediction is not simply that distillation reduces tokens. It is that a well-scoped certification scheme can reduce runtime evidence cost on stable task classes without increasing policy-violating commitments, while the same substitution should fail under deliberate distribution shift or policy change.

7.5 Certification must cover routed use

A population certificate is not automatically valid for the router-selected subset. Consider 100 equally likely task types, with a fast model correct on 99 and always wrong on one. Its population failure rate is exactly one percent. A router that invokes it only on the failing type induces selected-use failure one, without changing the original population or leaving the named domain.

One remedy is to certify the frozen routed policy, including the fast model and its selection rule. Another is to cover every subgroup or selection rule admitted by the contract. Multicalibration supplies guarantees over a specified class of identifiable subpopulations [64]; it is not a guarantee for arbitrary selectors outside that class, nor is calibration alone identical to the failure-risk target. Changes to routing, model, or population can invalidate the deployment statement.

The certificate should state both the risk measure and useful coverage. For selector $S_\rho(X, H)$ and failure event F_ρ , a selected-use target is

$$\Pr(F_\rho \mid S_\rho = 1) \leq \epsilon, \quad \Pr(S_\rho = 1) \geq \gamma > 0.$$

For sequential substitution under Theorem 6.11, the contract needs the stronger selected-entry conditional allowance, or a certificate for the complete finite workflow whose risk already includes its selection. It cannot insert the above marginal ϵ as an arbitrary history-conditional bound. The reference fast route uses a particularly transparent sufficient condition: only one fresh entry type is eligible, no private safety signal is available there, and its exact conditional failure probability is 0.04. Any evidence acquisition

closes that entry route. This is a scope-restricted construction, not a universal way to certify a distilled model.

8 Failure Modes and Architectural Countermeasures

The coupled analysis exposes failure modes that isolated routing or governance evaluations can miss.

Table 2: Representative failure modes in the coupled control plane.

Failure		Mechanism	Primary layer	Representative countermeasure
Underthinking		Router stops although a positive-value computation or evidence chain exists.	Metacognition	Multi-step value model; complementary-evidence training; regret measurement.
Overthinking		Additional operations consume resources without commensurate decision gain.	Metacognition	Stop rules; state-dependent prices; matched-budget training.
Wrong operation		System reasons when the missing ingredient is retrieval, execution, verification, or expertise.	Operation selection	Heterogeneous portfolio; value attribution by operation type.
Evidence shopping		Router selectively searches until it finds support for a desired action.	Router-gate seam	Preserve failed/contrary evidence; source-independence graph; audit sampling.
Verifier retry		Stochastic checks are repeated until one passes.	Router-gate seam	Attempt history; retry budgets; episode-level error accounting; sticky denials.
Gate probing		Denials reveal enough policy information to search for an admissible near-miss.	Router-gate seam	History-aware policy; bounded feedback; escalation after repeated near misses.
Contract harting	Good-	Literal evidence conditions are satisfied without epistemic independence or relevance.	Evidence contract	Provenance, dependency checks, causal independence tests, periodic adversarial review.
Strategic rance	igno-	Router avoids a disqualifying check because learning an unfavorable result can remove authority.	Router-gate seam	Make safety-relevant checks mandatory; absence of evidence is denial; audit skipped checks.

Failure	Mechanism	Primary layer	Representative countermeasure
Verifier capture	Verifier shares the candidate's failure mode or is manipulated by it.	Verification	Protected context; heterogeneous verifiers; executable tests; independence.
Information-flow leak	Individually allowed operations reveal secret-dependent information.	Flow policy	Data labels, capabilities, noninterference tests, CaMeL-style trusted data/control flow.
Approver persuasion	Agent influences a human authority service to change policy or privileges through unstructured advocacy.	Governance state	Structured approval packet; independent context; explicit state-change request; persuasion audit.
Trust-base tampering	Agent changes policy, authority, or gate state to make a later action admissible.	Governance state	Protected write set; meta-gate all delegation and policy mutation.
Semantic-gate amplification	Small per-attempt false-accept rate compounds under adaptive retries.	Gate / router seam	Conditional error measurement, retry limits, correlated-attempt modeling.
Human-review saturation	Conservative routing floods reviewers and degrades oversight quality.	Routing / human gate	Queue-aware pricing, escalation budgets, triage, measured reviewer load curves.
Candidate/episode identity manipulation	Split a continuing test or merge evidence from a changed effect; reset risk by delegation.	Protected gate / executor	Gate-assigned identity, workflow lineage, atomic reservations, immutable effect bindings.
Certificate selection failure	Population risk is reused on a router-selected subset with different failure rate.	Learning / gate	Certify routed use or covered selectors; selected-entry risk and coverage; separate calibration uncertainty.
Unsafe distillation	Fast policy inherits behavior without the evidence or scope that justified the slow episode.	Learning / gate	Provenance-bearing training set; scoped competence certificates; drift and policy revocation.

9 Evaluation Program and Falsifiable Hypotheses

The distinctive empirical claim is not that adaptive compute or runtime enforcement works in isolation. Both already have evidence. The discriminating question is whether *modeling governance inside the router* changes allocation and outcomes when evidence controls admissibility.

9.1 Registered comparator regret

Evaluation should distinguish an oracle bound from a deployable comparator. Let Π_{reg} be a fixed, published set of routers operating over the same operation portfolio, gate interface, and information stream. Define

$$\pi_{\text{reg}}^* \in \arg \max_{\pi \in \Pi_{\text{reg}}} V^\pi(D, B)$$

and

$$R_{\text{meta}}^{\text{reg}}(\mu; D, B) = V^{\pi_{\text{reg}}^*}(D, B) - V^\mu(D, B). \quad (14)$$

This is *regret relative to registered comparators*; it does not imply that the globally best non-clairvoyant policy has been solved. A separately computed dynamic-programming optimum is reported only for finite tabular instances where it is exact, and a clairvoyant allocator can be reported only as an oracle upper bound. Cross-paper comparison additionally requires the same registered operation portfolio and policy-visibility interface.

9.2 Exact finite demonstrations: strategic ignorance and adaptive admission

Before testing learned routers, two small finite models establish the normative predictions exactly. They are worked demonstrations of the formal mechanisms, not empirical evaluation of deployed systems. The accompanying script uses exact rational arithmetic for the tabular values and deterministic enumeration for the gate trajectories.

Toy A: strategic ignorance and the router–principal wedge. Let the latent state be safe with prior 0.96. Committing yields +10 if safe and −10 if unsafe, and defer yields zero. The gate currently allows the action because its safety posterior exceeds threshold $\theta = 0.95$. An optional scan costs 0.1 and passes with probability 0.8 when safe and 0.2 when unsafe. Before the scan, $Q = 9.2$. A pass occurs with probability 0.776 and raises the safety posterior to 0.98969; a failure lowers it to 0.85714 and removes authority even though the router’s posterior expected utility of committing remains positive.

Under the old admissible set, coherent updating has zero expected belief-first decision-refinement gain. The belief-first admissibility contribution is −1.6, so

$$\text{VoC}_G(\text{scan}) = 0 - 1.6 - 0.1 = -1.7.$$

The symmetric attribution makes the interaction explicit rather than assigning all of it to admissibility:

$$\phi_{\text{dec}} = +0.2304, \quad \phi_{\text{adm}} = -1.8304, \quad I = +0.4608.$$

The total successor-value change remains −1.6 under either attribution convention.

The utility component of this particular wedge disappears when the router prices the same downside as the gate. Holding the scan and gate fixed but changing unsafe-commit utility from −10 to −190 moves

the router’s own break-even safety posterior to 0.95. The scan then has net gate-coupled value +4.06: after an unfavorable result the router itself prefers deferral, so the gate no longer removes an action the router wants to take. This is the nonbinding case of Proposition 4.5.

The mechanism is not limited to zero private information value. Add a second action paying +11.5 when safe and −81.5 when unsafe. Its effect contract is always admissible because it is reversible and confined to a lower-effect tier; the router’s numerical downside is not itself the gate’s irreversibility criterion. The same scan has positive fixed-set decision value +0.58 because a favorable result changes the router’s preferred action, yet the gate’s revocation after a failed scan drives the total successor-value change to −1.02 and the net scan value to −1.12 after cost. Strategic ignorance can therefore persist even when the evidence is privately useful; the governance loss can dominate the decision gain. Making the scan a mandatory gate precondition removes the option to remain ignorant, while objective shaping reduces the underlying router–principal wedge.

Controlled skipped-check contrasts. Hold the scan’s information and acquisition cost fixed, and compare a gate that keeps the current action set with one that responds to its result. Table 3 separates the resulting value change from a binary check decision. In the original case both policies skip: governance lowers value, but the raw skip indicator does not change. In the interior case, the same intervention changes a privately worthwhile scan into one the router skips. Raising cost makes both skip again. Equal loss pricing with unequal declared priors also produces governance-induced negative information value; that cell demonstrates a belief wedge, not an empirical finding of overconfidence.

Table 3: One-check counterfactuals. Columns are the router’s subjective net value of the same scan. A positive value means scan; a negative value means skip. Shared priors are 0.96. The unequal-prior row sets router prior 0.99 and gate prior 0.96, with shared unsafe loss 190.

Case	Fixed gate	Responsive gate	Contrast
Shared beliefs, router loss 10	−0.10	−1.70	−1.60
Shared beliefs, router loss 190	4.06	4.06	0.00
Shared loss 190, different beliefs	−0.10	−0.56	−0.46
Interior fallback, scan cost 0.1	0.48	−1.12	−1.60
Interior fallback, scan cost 1.0	−0.42	−2.02	−1.60

Toy B: attempt-local versus history-complete evidence. Let the design prior be $\pi_0 = 0.6$ and gate threshold $\theta = 0.9$. A repeated verifier emits pass with probability 0.6 under safe and 0.1 under unsafe. A pass has likelihood ratio 6 and a failure likelihood ratio 4/9. An attempt-local gate that remembers only the favorable final artifact admits after any pass. With at most ten attempts, its exact unsafe-admission probability is

$$1 - 0.9^{10} = 0.65132,$$

while its safe-admission probability is $1 - 0.4^{10} = 0.99990$.

A history-complete Bayesian gate instead conditions on every pass and failure. Admission occurs when

$$6^{n_{\text{pass}}}(4/9)^{n_{\text{fail}}} \geq 6.$$

Exact enumeration through ten attempts gives unsafe-admission probability 0.13428 and safe-admission

probability 0.97123. The candidate-level anytime-valid bound from Theorem 6.8 is

$$\frac{0.6}{0.4} \frac{0.1}{0.9} = \frac{1}{6} \approx 0.16667.$$

Unsafe admission still rises slightly with more opportunities under the history-complete rule; the important result is that the growth is anytime bounded rather than retry-amplified without limit.

The two-dimensional admission comparison is informative but incomplete. The attempt-local gate at $k = 3$ has approximately (0.271, 0.936) unsafe/safe admission, while the history-complete gate at $k = 10$ has approximately (0.134, 0.971). The latter improves both admission coordinates but spends more verification. Figure 2 shows the unsafe-admission trajectories; Table 5 and Figure 3 add the resource cost and early-rejection choices.

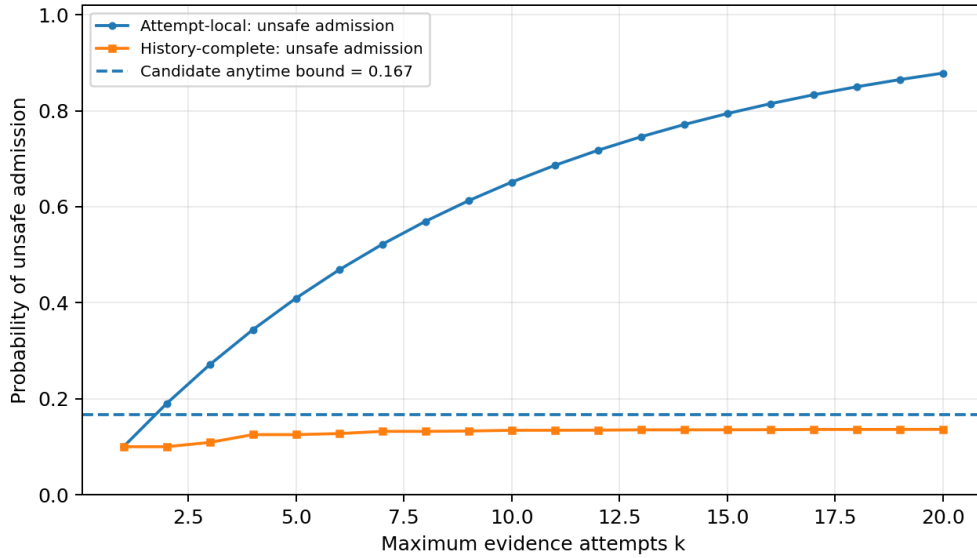


Figure 2: Exact finite-state unsafe-admission probability under repeated evidence attempts. The attempt-local gate forgets failures; the history-complete gate conditions on the full sequence. The dashed line is the horizon-free candidate bound from Theorem 6.8. Values are deterministic calculations from the declared two-state model.

Table 4: Exact worked-demonstration results. These are finite-model consequences, not deployment evaluation.

Mechanism	Quantity	Exact result
Strategic ignorance, original router	scan VoC _G	−1.7000; Shapley ($\phi_{\text{dec}}, \phi_{\text{adm}}$) = (+0.2304, −1.8304)
Router objective aligned to gate threshold	scan VoC _G	+4.0600
Interior two-action variant	fixed-set decision gain / net scan value	+0.5800 / −1.1200
Attempt-local gate, $k = 10$	unsafe / safe admission	0.65132 / 0.99990
History-complete gate, $k = 10$	unsafe / safe admission	0.13428 / 0.97123
History-complete candidate gate	anytime unsafe bound	$\leq 1/6 \approx 0.16667$

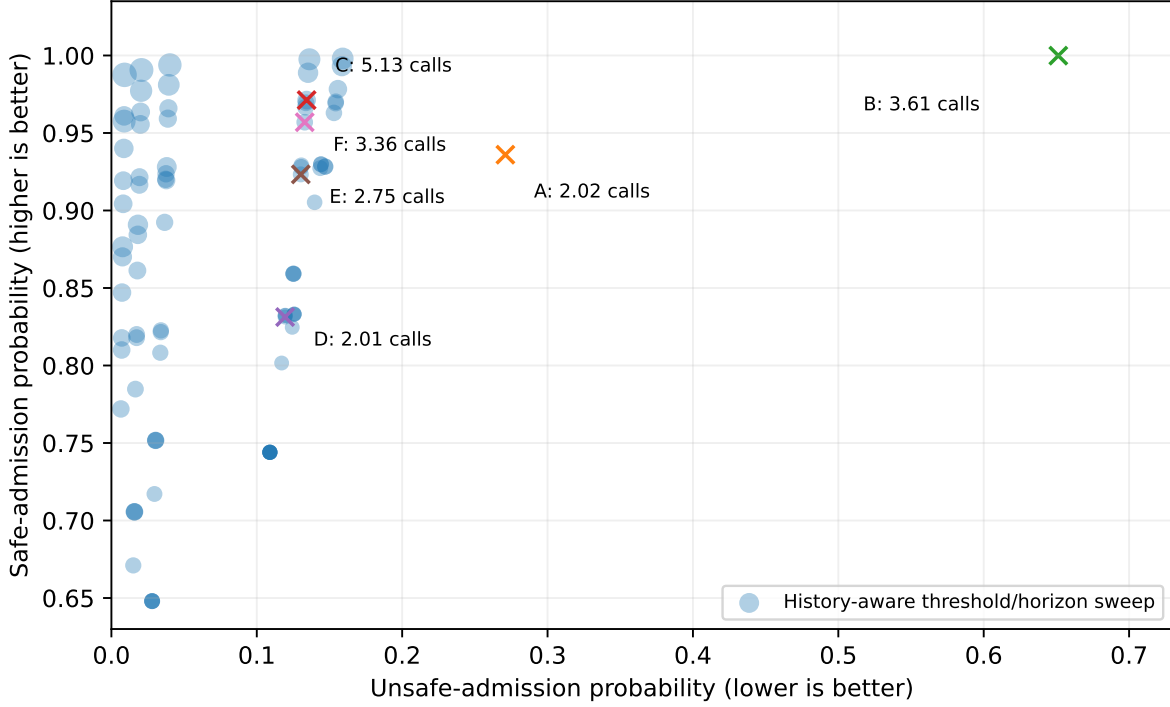


Figure 3: Declared Toy B operating points with verifier cost. The background sweep varies horizon, admission threshold, and rejection threshold. Bubble area increases with expected calls. Labels A–F follow the row order of Table 5; this is not the globally optimal frontier over every stopping policy.

For a population with actual unsafe-proposal fraction ρ , the operational false-admission share among admitted actions is

$$\text{FAR}_{\text{adm}} = \frac{\rho p_{\text{adm}|U}}{\rho p_{\text{adm}|U} + (1 - \rho) p_{\text{adm}|S}}.$$

This quantity depends on the proposal mixture even when the candidate-conditional theorem remains valid. In Toy B, using the model’s $\rho = 0.4$, it is approximately 30.3% for attempt-local $k = 10$ and 8.4% for history-complete $k = 10$.

Toy B with stopping cost and early rejection. Testing stops on admission, rejection, or the attempt cap. Expected call count is evaluated using safe/unsafe mixture $(0.6, 0.4)$, the same mixture used for the false-admission-among-admitted calculation. A lower boundary adds a cost-sensitive reject option while preserving the upper-crossing bound.

The history-complete admit-only point uses about 2.54 times the calls of attempt-local $k = 3$. Early rejection exchanges safe completion for cost and unsafe-admission reduction; no listed policy dominates all others on all three coordinates. The exact lower boundary $(4/9)^4 \approx 0.03901844$ must not be rounded to an executable 0.039: that different threshold gives safe admission 0.96778, unsafe admission 0.13401, and 3.75969 expected calls. The source records exact values and comparison semantics.

A principal selects an operating point through an explicit objective, for example

$$J = \pi_S R \Pr(\text{admit} \mid S) - (1 - \pi_S) L \Pr(\text{admit} \mid U) - c \mathbb{E}[N],$$

with zero-valued deferral. Resource-matched comparison, rather than a fixed-horizon safety claim, is the

Table 5: Toy B: three resource–admission coordinates. Lower boundaries are exact fractions; printed rounding is not an executable threshold.

Gate	Safe admit	Unsafe admit	Expected calls
Attempt-local, $k = 3$	0.93600	0.27100	2.02000
Attempt-local, $k = 10$	0.99990	0.65132	3.60518
History-complete, admit only, $k = 10$	0.97123	0.13428	5.12835
History-complete, lower $(4/9)^2$, $k = 10$	0.83126	0.11947	2.01198
History-complete, lower $(4/9)^3$, $k = 10$	0.92330	0.13027	2.75344
History-complete, lower $(4/9)^4$, $k = 10$	0.95694	0.13297	3.35642

appropriate interpretation for a cognition-allocation framework.

9.3 An executable, fully enumerated coupled environment

The reference system connects evidence selection, computation cost, candidate revision, statistical admission, scoped fast-path certification, and effect execution. Sources, tables, trained policies, and executable tests are supplied in the versioned supplement [65]. It is a finite synthetic model with specified kernels, not a language-model benchmark. A model-free tabular router is trained by sampled interaction and then evaluated exactly; no production or natural-language safety claim is inferred from it.

State, operation portfolio, and scope. A workflow has a visible type $x \in \{\text{routine}, \text{regular}, \text{blind}\}$ with probabilities $(0.3, 0.4, 0.3)$ and five resource units. Candidate zero has safe priors $(0.96, 0.70, 0.60)$ across those types. One revision is available; its safe priors are $(0.98, 0.90, 0.90)$. The revision’s safety flag is independent of the original flag conditional on type. This is a declared alternative-construction kernel, not evidence that actual bug fixes are independent. The workflow can commit once or defer. Revision cannot return to an earlier candidate or reuse its evidence.

Table 6: Finite reference operation contract. A pass is denoted $Y = 1$. A scan failure disqualifies its candidate. Costs are synthetic decision units, not measured dollars or hardware energy.

Operation	Resource units	Cost	$P(Y = 1 \mid S)$	$P(Y = 1 \mid U)$
Cheap verifier, routine/regular	1	0.20	0.80	0.20
Cheap verifier, blind	1	0.20	1.00	1.00
Precise verifier, every type	2	1.00	0.90	0.05
Disqualifying scan, every type	1	0.10	0.80	0.20
Construct candidate one	1	0.60	New candidate prior, as above	

The state records type, active candidate, remaining units, pass/fail counts for each verifier, and the scan’s unattempted/pass/fail status. It is a sufficient statistic of the observation history under the specified kernels. Full outcome history remains available to protected authorization. There are 558 reachable sufficient states per gate/objective cell. The finite horizon decreases on every nonterminal operation, so backward induction solves every state exactly with rational arithmetic. A separate iterative backward sweep agrees with the memoized recursion at every state in all eight cells.

Safe commitment earns 10 for the original candidate and 9 for the revision. The principal’s unsafe loss is 190. Routers are tested with unsafe loss 10 or 190, holding kernels and gate fixed. Monetary-

like computation costs are subtracted from both objectives. The lower-loss router is an intentional utility-wedge condition, not an assertion about any deployed model’s values.

Four crossed gate regimes. The *attempt-local* gate admits after any favorable verification artifact unless a scan has failed. The *pooled-history* gate retains all results but uses the ordinary cheap-verifier likelihoods on the blind subtype too. It is intentionally misspecified. The *selection-valid* gate uses subtype-valid factors from Theorem 6.10; a blind cheap pass has factor one. The *required-scan* gate adds a passed scan prerequisite to selection-valid statistical admission. Both statistical candidates receive fixed error allocations (0.08, 0.04), giving thresholds (12.5, 25) and total 0.12. Additional tests may be performed after a threshold is reached; a subsequent failure can remove current admissibility. The implementation never resets evidence merely because its current value is inconvenient.

All four regimes share a separately stipulated fast-path risk contract. It permits an immediate commitment only for the original routine candidate, before any observation or revision, with selected-use unsafe probability 0.04, exactly the complement of the stipulated prior. Its gate-assigned entry rule prevents post-evidence selection into the exemption. It reserves a distinct 0.04 allowance rather than an e-process false-positive rate. This exercises the selected-use accounting term and scope closure; it is not certification of a learned component and is not obtained by distillation. The general competence-certificate theory in Section 7.5 remains distinct from this type-based reference exemption. Statistical-path interventions below disable this exemption in both arms unless explicitly stated.

Controllers and matched runtime information. Every gate is crossed with both loss functions and three controller classes. The *exact* controller solves the finite Bellman recursion. The *two-step* controller recomputes a depth-two plan at each decision and uses best currently admissible terminal value at the search leaf. The *history-aware learner* is tabular Q-learning [66] over the same sufficient states and gate action masks. It receives sampled rewards and transitions, not exact Bellman targets.

The controller classes have identical runtime observation histories, possible operations, costs, gate regimes, and episode budgets. They do *not* have identical model access: exact and two-step controllers use the specified transition model, whereas Q-learning learns from interaction. The comparison measures performance in this reference problem, not a causal advantage of one representation with all other learning resources equal. A tabular history-aware learner is an unrestricted memory baseline on this finite sufficient state; it is not a recurrent language model.

Each Q-learning cell uses five seeds, each with 200,000 episodes: eight million sampled training episodes in total. Exploration decreases linearly from 1 to 0.1 over the first 70% of training and stays at 0.1. The visit-count learning rate is $n^{-0.65}$, with undiscounted return in the absorbing finite horizon. States, masks, seed-specific Q tables, counts, and checkpoints at 1,000, 10,000, and 200,000 episodes are included. Frozen greedy policies are evaluated by exact rational expectation; there is no Monte Carlo evaluation error. Between-seed variation is training variation within one generator, not uncertainty over real tasks.

For aligned loss 190, the Q-learning principal utilities range from 5.71007 to 5.88768 across seeds (sample standard deviation 0.07582). All exceed the two-step value 4.82079, but none reaches the exact value 6.04093. The learner’s lower unsafe rate than the optimum is accompanied by lower safe completion and greater cost; it is not an unqualified safety–utility improvement. For loss 10, Q-learning’s router-return gap from the exact optimum averages 0.02143, while two-step planning leaves a gap of 0.99449. This comparison exposes insufficient lookahead, not learning superiority over explicit planning or a gate-representation advantage. The depth sweep below reaches the initial-distribution optimum without learning.

Table 7: Exactly evaluated controller outcomes in the selection-valid regime. Safe and unsafe commitment are unconditional probabilities per workflow. Learner entries are means across five training seeds. All numerical results follow from the registered finite model.

Router loss	Controller	Principal utility	Safe commit	Unsafe commit	Cost
10	Exact	3.61057	0.82115	0.01924	0.76056
10	Two-step	1.76000	0.70200	0.02400	0.70000
10	Q-learning	3.58913	0.82115	0.01924	0.78200
190	Exact	6.04093	0.78898	0.00280	1.08483
190	Two-step	4.82079	0.64390	0.00157	1.31949
190	Q-learning	5.79992	0.76689	0.00159	1.35454

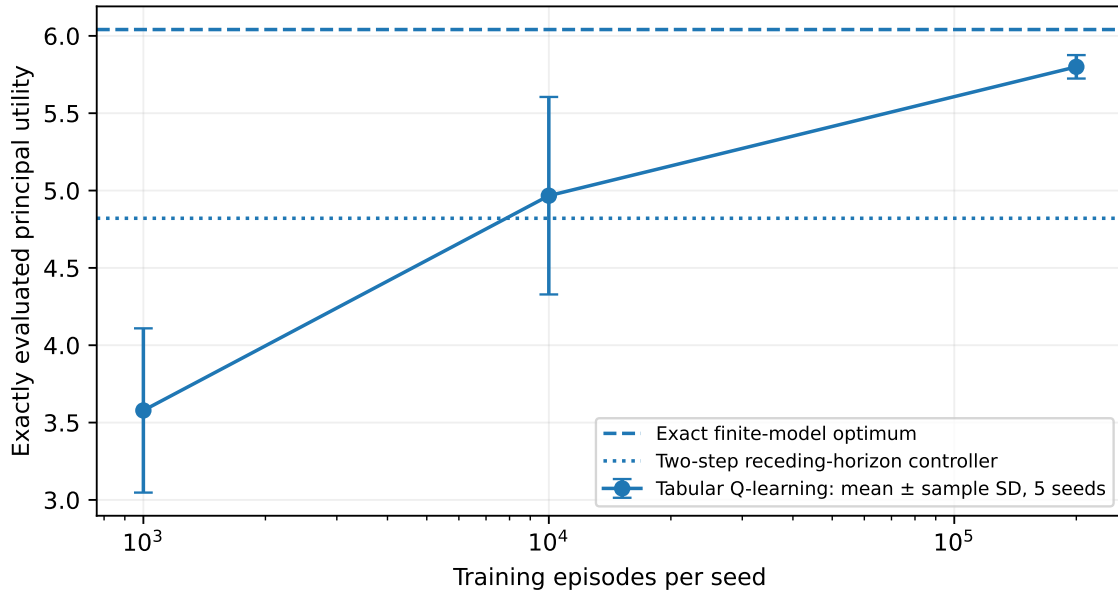


Figure 4: Training checkpoints for the selection-valid, aligned-loss cell. Frozen policies are evaluated exactly. Error bars show sample standard deviation across five training seeds; the dashed and dotted lines are the exact and two-step values.

Selection validity changes the threat response. A separate dynamic program maximizes unsafe admission over adaptive operation and revision strategies conditional on both statistical candidates being unsafe, with the fast certificate disabled. This is a finite adversarial challenge over the operation portfolio, not the policy used for the utility table. On the blind subtype, both the attempt-local and pooled-history gates have maximum unsafe admission one. The selection-valid gate has maximum 0.050; adding the required scan reduces it to 0.018. On routine and regular types, the corresponding valid maxima are 0.056. All fall below the statistical workflow allowance 0.12. The reductions are exact consequences of the specified kernels and finite resource limit, rather than empirical confirmation of Ville’s inequality.

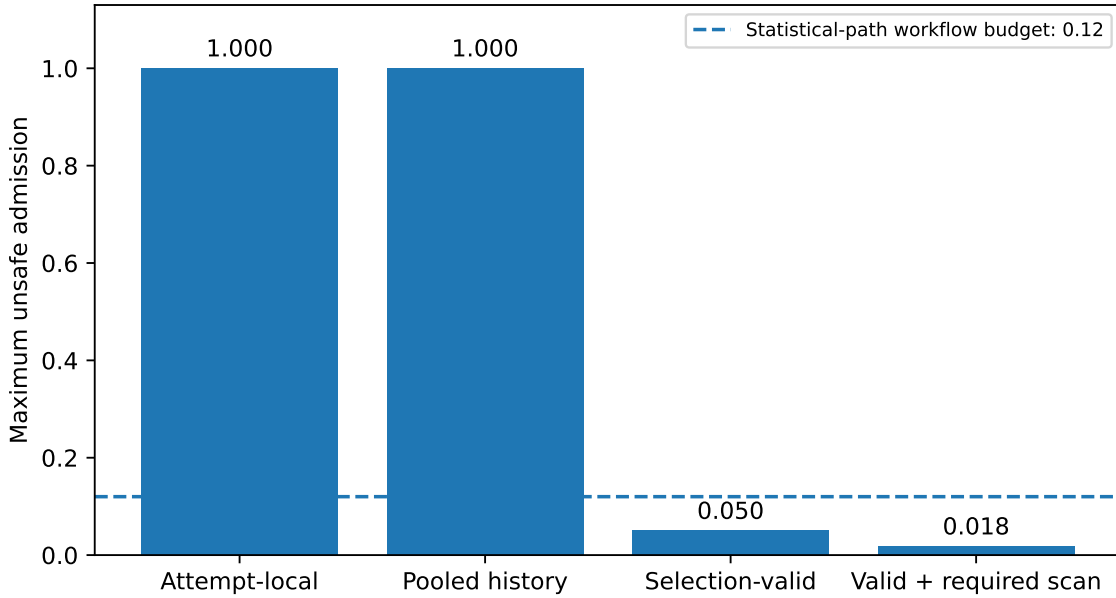


Figure 5: Exact maximum unsafe admission on the blind subtype across the four gate regimes. Every result uses the same five-unit portfolio and at most one revision. The challenge excludes the separately certified fast route. A complete record with pooled, invalid likelihoods remains exploitable.

The utility-wedge comparison supplies an operational consequence. With router loss 10, exact pooled-history optimization yields principal utility -2.59490 , unsafe commitment probability 0.05511, and cost 0.59448. Selection-valid optimization yields principal utility 3.61057, unsafe commitment probability 0.01924, and cost 0.76056. The stronger gate prevents the router from converting a cheap blind-spot pass into permission, at a resource cost. Under the valid gate, matching router and principal loss raises principal utility to 6.04093 and lowers unsafe commitment to 0.00280. These are separate interventions: repairing evidence validity and changing the router’s objective are not interchangeable.

Mandatory scans have no effect on the exact optimum in the two registered valid-gate loss cells: the optimal statistical routes already buy the scan. They do change the approximation and learner. This negative result is part of the experiment, not evidence that mandatory checks are generally redundant. Optional-check avoidance is diagnosed separately by a counterfactual holding the check’s information and cost fixed while changing the gate’s response, as defined in Section 4.3.

Binding to execution. The reference gateway registers immutable candidate bindings, reserves proposal risk under a gate-assigned workflow, verifies signed finite-kernel receipts, updates exact evidence factors, issues one-use authorizations, and executes into a versioned mock target. A binding includes

workflow/candidate identity, payload digest, target version, policy/verifier version, dependencies, expiry, evidence-chain digest, reservation, and nonce. Candidate registration is atomic with budget reservation. Execution rechecks the binding and atomically consumes authorization while changing the mock target. Child sessions inherit the parent’s workflow budget; denials and admissions do not refund it. The router facade obtains time from a trusted clock and records each trusted verifier outcome before returning it to the router, preventing timestamp choice or selective omission at that interface.

The trust boundary is explicit: the router uses the gateway facade, not the owner capability, signing key, or stores. Python object privacy does not enforce this against arbitrary code execution; this is an in-memory reference mechanism whose owner/verifier operations are trusted. The keyed digests protect record integrity within that model. They do not attest that a real external verifier is calibrated, nor do they implement distributed consensus or general information-flow security.

The test suite passes 41 named tests, including 4,830 exact betting-validity grid cases and exact agreement between two backward solvers at all 4,464 gate/objective state instances. Every terminal path of the exact, two-step, and final learned policies under both valid regimes is replayed through the gateway: 482 terminal paths, of which 227 execute a commitment. All match the mathematical gate. Additional tests reject altered payloads, stale target/dependency versions, changed policy/verifier versions, expiry, forged or replayed receipts, authorization replay, evidence inheritance by a different candidate, and certificate use after observation or revision, including a certificate token issued before the scope changes. Returned records do not share mutable objects with protected bindings or evidence history. Concurrent tests permit only one of sixteen competing reservations when only one fits, and only one of sixteen replays to execute. These are bounded executable checks, not a proof that every possible implementation attack is covered.

What the construction establishes. The finite system has useful fast and deliberative routes, a selection-valid evidence update, explicit risk accounting for certificate use, and an execution path that enforces the authorized binding. Its policies make nontrivial tradeoffs: complete logging is insufficient under misspecification; objective alignment and hard control complement one another; insufficient planning depth misses complementary evidence chains; and extra verification has measurable resource and false-block costs. The construction closes these connections within the declared finite model. Generalization to learned semantic likelihoods, new subtypes, real distillation, and distributed execution remains a separate empirical and engineering problem.

9.4 Governance-response interventions and blind-forecast controls

The intervention study holds the physical kernels, state, operation portfolio, resource limit, and objectives fixed while crossing four gates, two scan contracts, responsive/shadow behavior, losses 10 and 190, and fast-exemption enabled/disabled. Three controllers give 192 controller/contract configurations, each reported for the initial mixture and all three types. These are exact finite configurations, not independent statistical replications; some shadow configurations are equivalent by construction.

What governance blindness means here. The *future-admissibility-blind* planner optimizes the unrestricted predictive continuation with the same posterior, operations, costs, and horizon. It does not forecast the gate’s future masks. Actual effects still pass through the same gate. On a denied commitment, the primary fallback removes commitment from the current choice only and reselects using the unrestricted successor values. It either takes an operation or defers; one denied proposal is counted

Loss / scan	Controller	Principal utility	Safe commit	Unsafe commit	Cost
10 / credited	Aware	5.34625	0.80042	0.00792	0.96984
10 / credited	Blind forecast	3.27897	0.55128	0.00732	0.84215
190 / credited	Aware	6.04093	0.78898	0.00280	1.08483
190 / credited	Blind forecast	4.32740	0.62377	0.00247	1.34333
190 / pure	Aware	5.81889	0.79178	0.00156	1.48597
190 / pure	Blind forecast	3.27863	0.50371	0.00187	1.40350

Table 8: Future-permission foresight under identical actual enforcement. Statistical path only; responsive selection-valid gate. “Blind forecast” uses the current-denial/replan fallback. Safe and unsafe commitments are unconditional per-workflow probabilities; costs use the synthetic operation ledger. No row establishes a matched-risk comparison.

per decision. Denials consume no resource or monetary cost, favoring this ablation. A second, separately reported fallback defers immediately after denial. Both controllers receive the same state and current mask as the aware planner; the ablation changes their use of future permission information, not their runtime access. It is not a test of representation learning, hidden policy visibility, or equal training budgets.

Table 8 reports the primary fallback with the fast exemption disabled and the selection-valid responsive gate. An aware full-horizon optimum cannot have lower *router* value than another feasible policy in the same cell. That ordering is a control check, not a new theorem or an empirical advantage over an unrestricted learner. Principal value need not follow the ordering when objectives differ. Realized unsafe rates also differ: a common statistical allowance is not a matched violation rate.

Two scan contracts and their shadow. In the *credited* contract the scan pass contributes factor four and failure disqualifies. In the *pure-disqualifier* contract both outcomes contribute factor one, while failure still disqualifies. The *shadow* intervention retains the scan’s observations, posterior update, cost, resource use, and closure of the fresh fast exemption, but removes every scan-dependent statistical gate effect: positive credit, veto, and passed-scan prerequisite. Thus credited and pure shadow cells coincide; the required-scan shadow coincides with its optional counterpart. Shadow behavior is an experimental counterfactual, not a deployment recommendation.

With the fast exemption disabled, the exact aware controller’s expected scan counts are:

Router loss	Credited responsive	Pure responsive	Shadow
10	1.05120	0.00000	0.00000
30	1.05120	0.00000	0.18400
60	1.35120	0.40000	0.24800
100	1.35120	0.40000	0.40000
190	1.35120	0.40000	0.58480

The intermediate losses belong to a separately declared exploratory price grid, not additional training conditions. Counts include path choices and the possible revision, so an expected scan count can exceed one and is not a fixed-state skip frequency.

The loss-10 router skips a pure disqualifier in both arms: its raw skipping does not identify a governance-

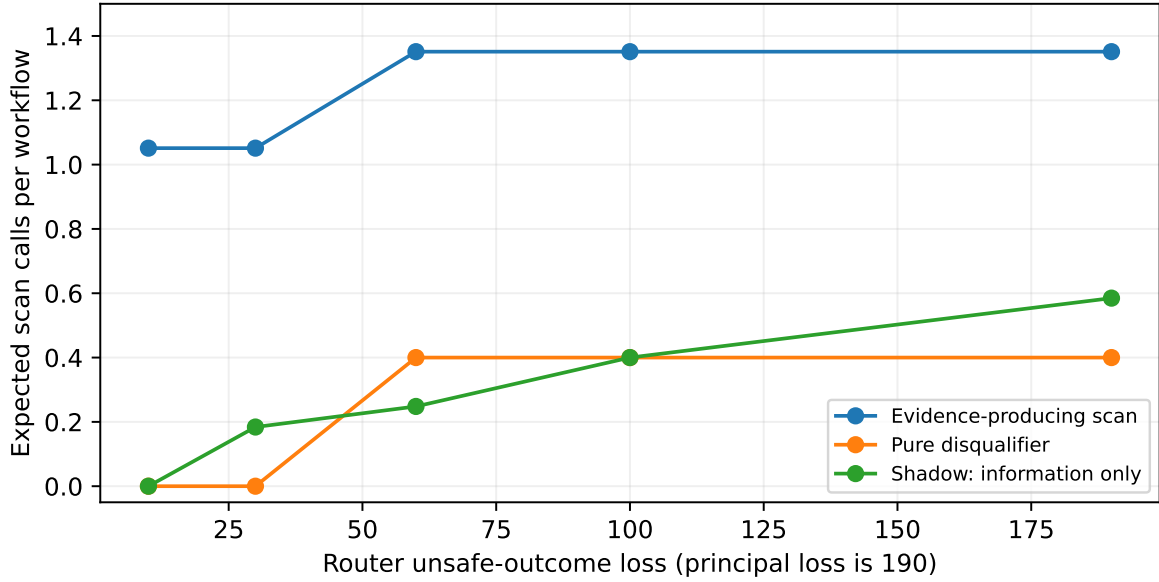


Figure 6: The same physical check can be an acquisition opportunity or an avoided disqualifier. Exact aware policies; optional selection-valid gate; fast exemption disabled. The shadow removes only the scan’s governance response. Connecting lines guide the eye across the declared loss grid, not an interpolation theorem.

induced avoidance effect. At loss 30, the gate response suppresses acquisition that occurs in the shadow condition. For example, after two cheap passes on the original regular candidate, with three resource units left, the shadow optimum scans and has value $52469/5750 \approx 9.12504$. The responsive optimum commits at $206/23 \approx 8.95652$; taking the scan instead has value $1733/230 \approx 7.53478$. This is a within-model counterfactual at the same posterior and cost. At loss 190 the mixed scan count still falls from 0.58480 to 0.40000, even though the utilities agree. This aligned-utility contrast instantiates the *rule-form wedge*: the e-value threshold and scan veto are not the Bayes action rule for the shared loss. It does not diagnose value disagreement or miscalibration. At loss 60 the direction reverses, so a path-averaged scan contrast has no universal sign.

The supplement also reports an unweighted fixed-state catalogue: a strict scan preference in the shadow state together with a strict non-scan preference in the responsive state. It contains six such states among 146 scan-available states at loss 190 for each scan contract, and none at loss 10. This is a finite-state enumeration, not an estimated prevalence or a population misalignment metric.

No universal principal benefit. Under the pure shadow contract and loss 10, the aware optimum has router value 6.30478 and principal value 4.03966; the primary blind forecast gives router value 4.92761 and principal value 4.23123. More accurate optimization of the router objective can lower principal value. Conversely, in the aligned credited cell, foresight raises principal utility and useful completion, while also raising the realized unsafe rate slightly. The full tables therefore retain utility, risk, costs, operation mix, and denials rather than replacing them with a single “awareness helps” verdict. These finite tests identify permission-sensitive allocation; they do not establish the original matched-risk deployment prediction.

A compliant exemption can reduce principal value. The fast-exemption switch in the same retained experiment separates probabilistic eligibility from beneficial use. For the exact loss-10 router under the credited, responsive selection-valid gate, enabling the exemption changes principal utility from 5.34625 to 3.61057 and unsafe commitment from 0.007916 to 0.019244 (Tables 8 and 7). The router’s own utility rises from 6.77113 to 7.07449, while runtime cost falls from 0.96984 to 0.76056. The principal’s loss of 1.73568 therefore already includes the computation savings; it is not an omitted cost benefit. The two non-routine type policies are unchanged.

Routine workflows have mixture weight 0.3, and the exemption’s selected-use unsafe probability is 0.04. Its realized unsafe contribution is therefore $0.3 \times 0.04 = 0.012$, within the separate allowance $B = 0.04$. The contract is honored. The aligned loss-190 router instead never uses the exemption: the unchecked routine commitment is worth $0.96 \times 10 - 0.04 \times 190 = 2$, below its available continuation value. Its full-workflow optimum remains 6.04093 with the exemption on or off. *Risk budgets bound probabilities, not principal value.* In these cells, a compliant shortcut is adversely selected by the objective that underprices downside. Exemption design must therefore evaluate principal-valued benefit relative to continued verification as well as probabilistic eligibility; aligning the router’s objective complements, rather than replaces, the independent risk limit.

9.5 Planning horizon and the distinction between permission and stopping

A receding-horizon planner is evaluated at depths one through five nonterminal operations, with the original fast exemption retained. Every nonterminal operation spends at least one of the five resource units, so depth five spans the full remaining horizon. Four operations already attain the initial-distribution optimum in both selection-valid loss cells:

Controller	Principal utility, loss 10	Principal utility, loss 190
Depth 1	1.76000	4.66400
Depth 2	1.76000	4.82079
Depth 3	3.40650	5.30679
Depth 4 or 5	3.61057	6.04093
Full-horizon optimum	3.61057	6.04093
Retained Q-learning mean	3.58913	5.79992

The Q-learning rows retain the original five seeds and interaction budgets; no new learner is trained for this sweep. Four-step equality is stated for these roots and contracts, not every state or every environment. The result supports the complementary-chain interpretation: insufficient lookahead underperforms, tabular learning approaches the finite optimum, and adequate model-based lookahead reaches it. Since the tabular state already represents evidence counts and masks, this experiment does not identify a benefit of an explicit governance representation over a recurrent or language-model representation.

An admitted action need not be committed immediately. On the original regular-candidate path, a scan pass followed by a cheap pass leaves posterior 112/115, evidence 16 above threshold 25/2, and three resource units. With loss 190, immediate commitment has value $110/23 \approx 4.78261$. One more cheap test followed by the best allowed terminal decision has net value $33/5 = 6.6$; the full continuation has value $4642/575 \approx 8.07304$ and also begins with another check. This state is reached with probability

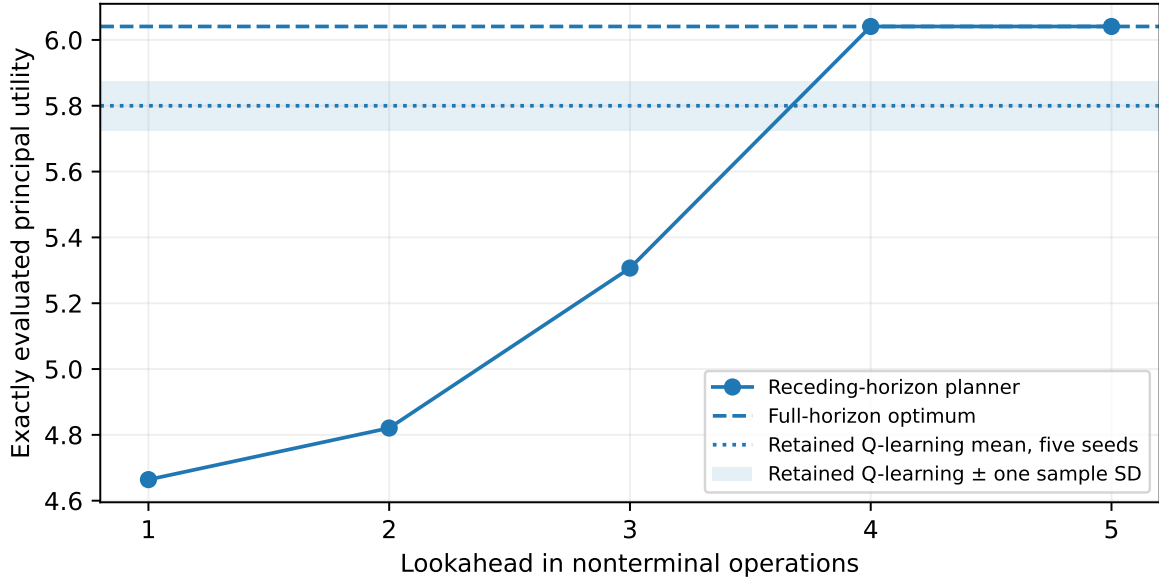


Figure 7: Lookahead, not an inherent learning-versus-planning advantage, explains the short-horizon comparison. Selection-valid gate, aligned loss, original fast exemption. The shaded band is variation across the five retained training seeds, not uncertainty in the exactly evaluated planning values.

23/50 conditional on the regular type along the exact policy. Revocation risk does not erase the value of information. **Admission is a permission boundary; optimal stopping is a separate decision.**

9.6 Fixed-sample calibration over selectable unsafe subtypes

The preceding experiment supplies the precise verifier’s unsafe pass probability $q_x = 1/20$ exactly. We now withhold that parameter from the gate and estimate it with n independent unsafe examples for each of the three registered types, independent of deployment. The other kernels and type structure remain stipulated. The router/evaluation oracle still knows the physical kernels and plans against the calibrated gate; this isolates the cost of certification, not estimation by a learned router.

For each type x , let $K_x \sim \text{Binomial}(n, 1/20)$. A one-sided Clopper–Pearson upper endpoint is computed at error allowance $\delta/3$ [67]. For $k < n$, the endpoint solves

$$\sum_{i=0}^k \binom{n}{i} (q^+)^i (1 - q^+)^{n-i} = \delta/3,$$

and for $k = n$ it is one. The implementation rounds upward to a 10^{-6} rational grid and verifies the bracketing binomial-CDF inequalities with exact integers. Beta quantiles propose an endpoint; floating-point rounding is not trusted for the final coverage inequality. With q_x^+ frozen before deployment, simultaneous coverage satisfies

$$\Pr(\forall x : q_x \leq q_x^+) \geq 1 - \delta.$$

On that shared calibration event, the operation-specific updates in Theorem 6.10 remain valid under the registered adaptive operation selection. The safe design rate is $p = 0.9$. When $q_x^+ \geq p$, the positive betting coefficient is set to zero rather than using a negative likelihood bet. We enumerate every possible calibration count, not one selected sample.

n/type	Mean q^+	Joint undercoverage	Principal utility	Safe commit	Cost
25	0.22936	0.00000	5.26396	0.67865	1.04920
50	0.15870	0.00000	5.57856	0.73472	1.19414
100	0.11768	0.01766	5.86305	0.76978	1.13331
250	0.08773	0.03875	5.98652	0.78229	1.08421
500	0.07490	0.03209	6.00000	0.78393	1.08419
1000	0.06673	0.04206	6.02394	0.78688	1.08454
Known q	0.05000	—	6.04093	0.78898	1.08483

Table 9: Calibration is an assurance–coverage–resource tradeoff. Total $\delta = 0.05$ across three unsafe types; responsive credited selection-valid gate; aligned objective; fast exemption disabled. Results average all calibration outcomes. Runtime costs exclude the separately reported $3n$ calibration samples. Undercoverage is the shared calibration event, not unsafe execution; its discrete variation need not be monotone in n .

The study uses $n \in \{25, 50, 100, 250, 500, 1000\}$ and $\delta \in \{0.1, 0.05, 0.01\}$, crossed with valid/required-scan gates and the fast exemption enabled/disabled. The three roots are separable, so integrating the root-specific values is sufficient:

$$\bar{J} = \sum_x w_x \sum_{k=0}^n \Pr(K_x = k) J_x(q^+(k)).$$

This evaluates the expectation under three independent calibration datasets without assuming that their realized counts coincide. For single-type undercoverage $\eta_n = \Pr(q^+(K_x) < 1/20)$, the actual probability that at least one bound fails is $1 - (1 - \eta_n)^3$, also evaluated exactly. All 18 calibration settings satisfy the declared simultaneous bound. The full study contains 5,793 verified endpoints and 72 gate/exemption/calibration summary cells. Endpoint-dependent gate masks reduce the exact control computations to 58 distinct mask configurations.

Larger calibration samples approach the known-kernel value in these cells, while a smaller calibration error allowance produces wider bounds and lower useful admission. The effect is not necessarily extra runtime work: stricter bounds can redirect the policy, induce earlier deferral, or favor a different verifier. Table 9 therefore reports completion and cost together. With the fast exemption enabled, the same experiment also retains the separately stipulated $B = 0.04$ term; it does not turn that exemption into certified neural competence. The calibration uncertainty δ remains a shared event, not a fresh independent risk on each use, as in Theorem 6.11.

Coverage beyond the declared family. Adversarial search can reveal additional unsafe subtypes, but a confidence bound on the attacks found does not certify attacks the search missed. Simultaneous bounds require a covered family and a sampling/selection design for that family. Fixed-sample intervals here are not reused under arbitrary optional calibration stopping; time-uniform confidence sequences provide a different framework for such sequential designs [68]. A learned semantic verifier still requires an empirically justified selection class, calibration protocol, and useful evidence power. The present study constructs and measures the connection only for the registered finite kernels.

Executable checks for the extensions. The retained 41-test reference suite passes unchanged. Twelve additional named tests verify the new interventions, ledgers, counterexamples, calibration coverage, and gateway rules. Recursive and iterative backward solvers agree in 35,712 exact state-

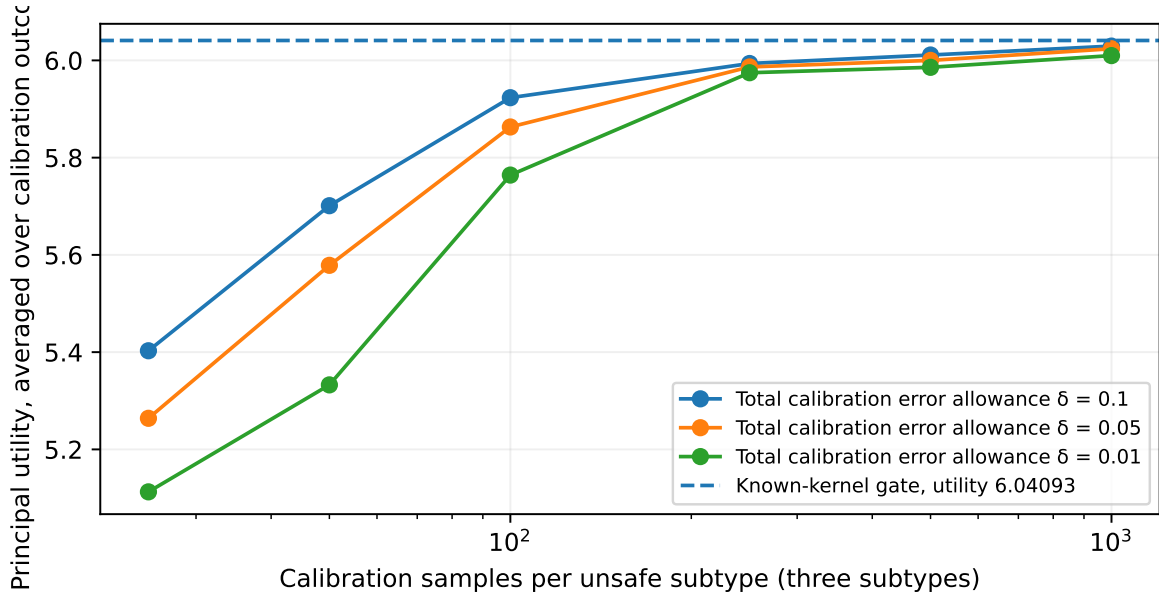


Figure 8: Finite-sample assurance reduces useful performance relative to a known-kernel gate. Exact expectations across all calibration counts; selection-valid gate, aligned objective, no fast exemption. Each confidence setting divides its error allowance over three registered subtypes. No extrapolation beyond this finite type family is asserted.

value comparisons for the extension contracts and their unrestricted planning counterparts. The extended gateway binds its scan contract and calibrated bound into protected policy identity. It replays 2,407 terminal intervention-policy paths, including 654 commitments and 1,753 denied-terminal checks. A separate representative-bound replay checks 226 paths, including 92 commitments, at $q \in \{0.05, 0.07, 0.1, 0.9, 1\}$. The latter is not a replay of every calibration count; calibration-outcome coverage and control values are enumerated separately. The same in-memory trust assumptions apply: these checks are not production isolation, semantic-verifier attestation, or a distributed transaction proof.

9.7 Registered hypotheses

H1 – Multi-operation allocation. Under a matched average resource budget, a router that allocates among reasoning, retrieval, execution, verification, specialist models, and escalation will outperform uniform deep reasoning when the task distribution varies in which operation has positive value. Existing adaptive-compute results motivate the hypothesis but do not establish heterogeneous portfolio allocation [23, 25, 30].

H2 – Governance-aware allocation under strong contracts. Under a selection-valid contract with protected evidence accounting that requires safety-relevant checks to be attempted, holding the base model, operation portfolio, budget, and gate fixed, a governance-aware router will allocate more efficiently toward evidence that changes admissibility than a governance-blind router, improving task utility at a matched episode-level violation rate. The strong-contract qualifier is load bearing. Section 9.4 now establishes permission-sensitive allocation against a specified blind forecast within the finite model. Because realized risks differ and blindness is a specified model ablation, it does not establish matched-risk deployment superiority or a representation-learning benefit.

H3 – Weak contracts create gate gaming and strategic ignorance. Under history-incomplete, selectively satisfiable, or optional-disqualifying evidence contracts, an optimized governance-aware router will exploit the contract more than a governance-blind router: retrying or selecting evidence to obtain admission, probing visible policy boundaries, or skipping checks whose unfavorable result could remove authority. The prediction concerns optimized policies; the learned-router experiment therefore declares the optimization algorithm and training budget rather than treating “capable enough” as an unfalsifiable qualifier. The responsive/shadow study shows both acquisition and avoidance regimes, plus cells with identical skipping. It narrows the prediction to profitable counterfactual opportunities; raw scan counts and gate awareness alone do not identify exploitation.

H4 – Adaptive verifier validity with useful coverage. The deployment hypothesis is that a realizable verifier construction can maintain selection-valid conditional bounds under the registered router strategy class without collapsing useful completion. Theorems 6.8 and 6.9 are mathematical implications, not empirical hypotheses. Their finite instantiation is checked in Section 9.3, and Section 9.6 exercises estimated conditional bounds over the registered type family. Learned semantic bounds still require a calibration and adversarial-selection study whose coverage is independently justified.

H5 – Learned routers respond to stakes. When task content and predictive uncertainty are held fixed but the cost or irreversibility of error increases, trained or deployed routers will shift resources toward verification, independent monitoring, human review, or deferral. This is an empirical claim about realized router behavior, not a restatement of the normative objective.

H6 – Flow-aware mediation closes failures that action-only gates miss. On prompt-injection tasks with protected data, effect signatures carrying data provenance and destination labels will reduce exfiltration relative to an otherwise identical action-authorizing gate that observes tool names and arguments but lacks flow state. CaMeL supplies a strong existing baseline for this comparison [42].

H7 – Certified distillation moves both cost and admissibility. Distilled competence alone will reduce inference cost but will remain blocked on evidence-gated commitments. A scoped competence certificate will reduce runtime verification cost on in-distribution tasks while preserving commitment thresholds; deliberate drift or policy change should trigger certificate invalidation and restore the slow path.

9.8 Extension to language-model routers

A compact first study can combine three task families under one control harness:

1. **Math/code with executable verification.** The same items appear under low- and high-stakes conditions; high-stakes commitment requires an executable check.
2. **AgentDojo subset with prompt injection.** The agent must complete useful tasks while protected data and adversarial instructions create information-flow pressure [41].
3. **Noisy evidence contract.** A deliberately imperfect semantic or stochastic verifier controls one high-value operation, creating a measurable incentive for retry and evidence shopping.

Use a fully crossed treatment matrix. Router classes should include uniform slow reasoning, a governance-blind ablation, a competent history-aware learned baseline without an explicit gate-value decomposition,

and an explicitly gate-coupled controller. Cross *every* class with each weak and selection-valid gate regime, resource budget, and policy-visibility interface. Keep runtime information, operation access, model capacity, optimization algorithm, and interaction budget matched where a causal representation comparison is claimed. Model-based oracle controls must be marked separately.

The tabular optimum and trained history-aware baseline are supplied by Section 9.3; Sections 9.4–9.6 add controlled gate-response, foresight, horizon, and calibration studies. The next extension trains language-agent routers over task families, registers optimization and interaction budgets, and evaluates distribution shift separately from in-generator behavior. Gate gaming is a conditional prediction under profitable opportunities, not a fallback claim that any negative result means the router was insufficiently capable.

Policy visibility remains a separate factor: opaque, contract-visible, or fully modeled. The finite blind-forecast study changes the use of future masks, not their availability. The state is visible in all cells; there is still no policy-visibility sweep. Measure its efficiency–gaming frontier in the language-agent extension rather than attributing an unmeasured transparency effect to the present results. For wedge identification, use shared-posterior and loss-price interventions and a gate-response counterfactual that holds evidence, costs, and alternatives fixed.

Report the full resource vector, task utility, verifier calls, skipped required checks, failed evidence attempts, source diversity and dependency, proposal count and episode error-budget spend, gate denials, candidate-conditional unsafe admission, unsafe share among admitted actions, false blocks, policy violations, information-flow violations, human escalations, and registered-comparator regret. The signature result is not merely that one router scores higher. It is whether coupling to the gate changes the router’s strategy, including the emergence or suppression of evidence shopping and strategic ignorance.

9.9 Human-review load experiment

For deployments with a human approval tier, vary router escalation volume while holding review task difficulty fixed and measure reviewer accuracy, response latency, override rate, and automation bias. The control plane predicts a nontrivial optimum: escalating more cases can reduce autonomous risk initially but can eventually degrade the assurance value of human review through load and verification complexity [58, 69].

9.10 Falsifiers

The following outcomes can narrow the behavioral and deployment claims. An implementation that violates a valid theorem under its actual assumptions indicates a bug, a premise-check failure, or a proof error, rather than an ordinary empirical falsification of the theorem:

- governance-aware and governance-blind routers allocate the same operations and achieve indistinguishable utility/violation frontiers across tasks where evidence changes admissibility;
- under weak contracts, gate feedback does not induce measurable retry, evidence-selection, probing, or avoidance of disqualifying checks even when those strategies improve the router’s objective;
- a realizable semantic-verifier construction cannot maintain its claimed conditional validity under registered adaptive selection without losing useful task completion;
- effect-typed flow state adds no security value over action-only mediation in exfiltration tasks;
- competence certificates fail to preserve admissibility or cost advantages on the stable task classes for

which they are issued;

- human-review accuracy is insensitive to escalation load over the deployment-relevant range, eliminating the proposed governance-resource externality.

10 Discussion

10.1 Dual-process cognition and evidence-gated agency are complementary views

The dual-process framing answers a cognitive question: why should an intelligent system have both a cheap default path and a more expensive deliberative path? The control-plane framing answers the systems question that follows: who decides when the expensive path is worth invoking, what kind of evidence it must produce, and when the resulting proposal may become action? Neither view subsumes the other. The fast/slow distinction explains the economic and functional structure of cognition; the router-gate architecture explains how that structure is allocated and governed in an agent operating in the world.

This combination also clarifies the role of verification. System-2-like processing is not automatically more reliable simply because it is longer. Its distinctive value comes from access to operations that can change the evidentiary state: executable tests, retrieval from authoritative sources, independent recomputation, symbolic tools, adversarial checking, or human judgment. The router should therefore allocate not merely “more thinking” but the operation most likely to resolve the decision-relevant uncertainty or evidence deficit.

10.2 The router and gate form one strategic system

The central architectural lesson is that a gate is not merely a filter after intelligence has finished operating. When evidence controls admissibility, the gate changes the value of upstream cognition. A router that knows this can prioritize tests, retrieve authoritative sources, request approvals, or search for alternative formulations because those operations change the feasible action set. Governance is therefore inside the decision problem even when enforcement remains outside the model.

This coupling also creates a principal-agent problem inside the agent. The organization wants the router to acquire evidence that tracks truth and policy intent. The router is rewarded by downstream utility and may discover shortcuts that satisfy the gate’s observable predicate. Proposition 4.5 makes one part of this wedge precise: governance-induced strategic ignorance requires a gate that excludes an action the router itself would still prefer. Gate gaming is the broader systems analogue of reward hacking. The response is two-layered: shape the router’s objective so important downside is priced before the gate binds, and retain an independent gate whose evidence state, provenance, proposal budget, and retry semantics represent the principal’s actual decision boundary.

10.3 Compute per decision connects agent policy to infrastructure

The resource ledger connects metacognition to physical inference systems. Tokens, memory bandwidth, HBM traffic, network service, queueing, and model cascades all become marginal costs of an operation. Caching and batching make those costs state dependent. Human attention enters the same ledger while retaining a different trust role. This is a natural cross-layer interface: infrastructure exposes measured

marginal resource coordinates, deployment owners assign budgets and prices, and the router allocates cognition against the current system state.

The implication is consequential for semiconductor and datacenter design. A future agent workload is not characterized only by tokens per second. It is characterized by the distribution of *decision episodes*: how often agents branch, verify, retrieve, invoke another model, revisit cached context, or escalate. Control-plane policy therefore changes the workload presented to accelerators, memory systems, networks, and human operations.

10.4 Chain-of-thought is a sensor, not the trust root

Reasoning traces can expose intent before a harmful action sequence completes, and current monitoring evidence makes them too valuable to ignore [37, 38]. The architecture uses this signal. It simply refuses to make containment depend on its perfect faithfulness. The strongest controls sit at effect boundaries; chain-of-thought monitoring augments those controls by raising risk estimates, triggering verification, or escalating to a stronger gate tier.

This layered view avoids a false choice between internal transparency and external enforcement. A system can preserve monitorable reasoning while also using capabilities, data-flow labels, deterministic constraints, and protected authority state.

10.5 Human oversight is coupled to metacognition

Human review occupies two roles simultaneously. To the router it is an expensive operation. To the governance system it can be the highest authority tier. Those roles interact because router escalation policy determines reviewer load, and reviewer load changes the reliability and latency of the gate. The correct architecture therefore treats human attention as a stateful service with measured capacity and error curves rather than as an infinite exception handler.

10.6 The architecture is recursive

The same structure can exist at multiple scales. A router may choose one model, a multi-agent debate, or a formal solver. A verifier can have its own router over tests. A human organization can itself be represented as an authority service with queues and escalation paths. A subagent can receive a scoped capability and become a local commitment domain. Recursion is safe only when authority delegation and information flow remain explicit in the effect signature.

11 Implications for Alignment, Control, and Safety

The control-plane view changes the unit of safety analysis from a model response to a complete decision episode. The terms remain distinct: *behavioral alignment* concerns what a learned policy tends to pursue or produce, whereas *control and containment* limit consequential effects even when that policy is wrong, exploited, stale, or misaligned. Between them lie two assurance problems created by agency: metacognition determines when the fast path is insufficient, and evidence governance determines what must be learned, checked, or proven before authority is exercised. Training and post-training remain essential [4, 5, 70]; they are one layer of a larger cognition-to-commitment system.

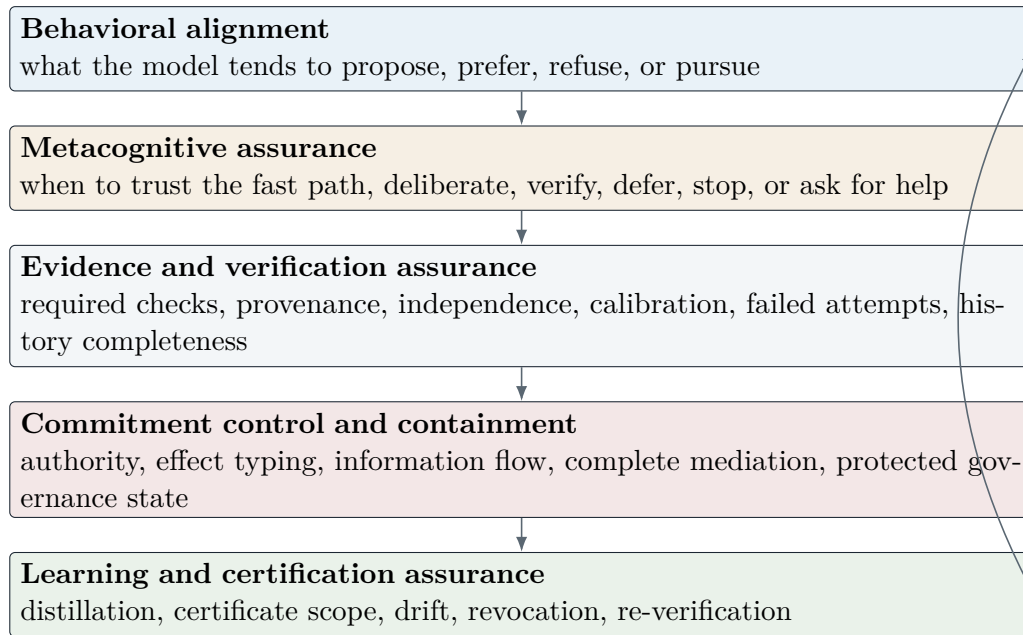


Figure 9: Layered system assurance for agentic AI. Alignment and runtime control are separate guarantees; metacognition, evidence, and learning govern how learned behavior reaches—or is prevented from reaching—consequential effect.

11.1 Safe cognition is not the same object as safe agency

A model can generate a poor or undesirable proposal without producing the corresponding external effect when commitment is separately mediated. Conversely, a model with strong behavioral priors can still cause harm through stale authority, unsafe tools, information-flow violations, distribution shift, or an erroneous high-consequence action. Neither guarantee subsumes the other. Behavioral alignment lowers the frequency and adversarial pressure of unsafe proposals; containment limits the consequences of proposals that nevertheless occur. The architecture therefore uses defense in depth rather than relabeling containment as alignment. The compliant-exemption comparison in Section 9.4 makes the distinction operational: its risk contract is satisfied, yet enabling the shortcut lowers principal value under divergent objectives. A probability ceiling constrains feasible behavior; it does not select the principal’s preferred behavior within that set.

The dual-process framing sharpens the point. System-1-like cognition makes error cheap, fast, and scalable, so its characteristic failure is under-escalation: a plausible response receives more confidence than evidence warrants. System-2-like cognition can correct that failure through search, execution, retrieval, and verification, but it is also a capability amplifier. The same search capacity can probe policy boundaries, retry stochastic checks, select favorable evidence, or discover a permission-preserving workaround. More reasoning is therefore not monotonically safer. The relevant safety question is whether the control plane allocates stronger verification and stronger governance as capability, uncertainty, stakes, and irreversibility increase.

11.2 Verification and evidence connect alignment to control

Additional self-generated reasoning and independent evidence are different operations. A longer chain of reasoning can elaborate a shared mistaken premise; an executable test, fresh authoritative source,

independent recomputation, or approval can change the information state. The safety-relevant capability is therefore not simply “reason longer” but *recognize when self-generated cognition is insufficient and acquire evidence whose error structure is appropriate to the decision*.

Evidence is also unusual because it can change authority. Section 4 showed that an observation can change both posterior decision value and the admissible action set. That coupling creates two symmetric hazards. *Evidence shopping* seeks observations until the gate permits a desired action. *Strategic ignorance* avoids a disqualifying check because an unfavorable observation would remove authority. The former makes favorable evidence too easy to manufacture; the latter makes ignorance instrumentally valuable. Both disappear only when the evidence contract represents the intended epistemic requirement: failed attempts remain visible, dependencies are modeled, and safety-relevant disqualifying checks are mandatory rather than optional rewards.

This is also why per-attempt safety accuracy is not the deployment metric. Attempt-local retries can amplify a small local false-accept rate. The history-complete e-process gate of Theorem 6.8 gives an anytime-valid bound for one candidate under adaptive evidence selection, while Theorem 6.9 extends control across adaptively generated variants through an episode-level error budget. The resulting design principle is stronger than “limit retries”: make the complete evidence-generation process and proposal family part of trusted decision state, and measure candidate-conditional risk, unsafe share among admitted actions, and episode-level risk under adaptive search rather than only on i.i.d. examples.

11.3 Effects, information flow, and monitoring occupy different safety layers

Effect typing moves the enforcement unit from tool identity to consequence. A nominal “read” can disclose a secret through a query; a specialist-model call can export context; a policy edit can change future authority. The gate therefore evaluates data provenance, destination, privilege, reversibility, and policy tier in addition to the operation name. Information-flow labels and declassification rules protect confidentiality properties that ordinary action-trace containment cannot express.

Chain-of-thought monitoring remains useful but is not the trust root. Reasoning traces can expose suspicious intent or reward hacking before an action sequence completes [37, 38]; their unfaithfulness and optimization sensitivity make them better treated as one evidence channel with a measured error model. Harder guarantees remain grounded in protected authority, effect mediation, and information-flow policy. This avoids a false choice between interpretability and enforcement: internal monitoring can improve detection while external control limits consequences.

11.4 Learning and human oversight introduce lifecycle and fleet effects

The moving fast–slow boundary creates an alignment-relevant compression problem. Distillation can preserve competence while discarding the runtime evidence that justified the slow episode. A model may learn *what* to do without retaining *why this action was authorized here*. Scoped competence certificates address that gap only under their declared statistical guarantee, task domain, model version, policy, expiry, and drift assumptions; issuing or expanding such a certificate is itself an authority-changing governance act. When the scope changes, the system should move the decision back toward System-2-like verification rather than silently inherit stale authority.

Human review is similarly not a binary safety switch. It is a finite-capacity service whose error and latency can depend on queue load. The router controls arrivals to that service, so escalation policy changes the reliability of the highest authority tier. The formal response in Section 4 is a state-dependent

congestion price or fleet-level objective; the empirical response is to measure reviewer accuracy, latency, disagreement, and override behavior as escalation load changes. Safe deployment seeks calibrated escalation, not maximal escalation.

11.5 Implications for evaluation

The combined framework implies that agentic safety benchmarks should retain the complete decision episode rather than score only the final text. The most useful measurements follow directly from the architecture.

Table 10: System-assurance measurements implied by the control-plane framework.

Layer	Representative measurements	Characteristic failure
Behavioral alignment	proposal quality; policy compliance; harmful-proposal rate; calibration	unsafe or systematically misleading default behavior
Metacognition	escalation precision/recall; stopping quality; compute allocation; deferral quality; registered-comparator regret	overconfidence, under-thinking, wasteful over-thinking, wrong escalation target
Evidence / verification	required-check completion; verifier choice; evidence independence; contradictions; failed-attempt history; freshness	self-confirmation, evidence shopping, strategic ignorance, verifier retry
Commitment / containment	unauthorized-effect rate; information-flow violations; privilege use; reversibility; episode-level false accepts	unsafe action, data exfiltration, authority misuse, repeated boundary probing
Human oversight	review load; queueing; reviewer accuracy versus load; override and disagreement rates	approval fatigue, rubber stamping, unsafe under-escalation
Learning / certification	post-distillation admissibility; certificate calibration and freshness; drift-triggered revalidation; revocation latency	competence without justification, stale authority, silent policy transfer

The discriminating experiment in Section 9 instantiates this view: hold the base model and gate fixed, vary whether the router models evidence-conditioned admissibility, and cross that treatment with contract strength and policy visibility. The paper’s central empirical question is not whether one router “reasons more.” It is whether coupling cognition allocation to authority changes task value, evidence acquisition, gate gaming, information flow, and episode-level safety in predictable ways.

The resulting safety thesis is compact. Behavioral alignment concerns what the system tends to pursue. Metacognitive and evidence assurance govern when it trusts itself and what it must learn before proceeding. Commitment control governs which effects are allowed and what information may flow. Learning and certification govern which previously verified behaviors may migrate toward the fast path. As AI moves from answering to acting, safety therefore becomes an auditable property of the full trajectory from proposal to evidence to authority to effect, while alignment and control remain distinct guarantees within that trajectory.

12 Validity Domain and Open Problems

The control plane makes several engineering boundaries explicit because those boundaries determine what can be claimed and measured.

Metareasoning calibration. The normative value function is only as good as the router’s estimates of posterior quality, operation cost, evidence value, and future admissibility. Calibration error is therefore a first-order capability variable. Shared-budget failures in current reasoning models make this an empirical research target rather than an implementation detail [30].

Evidence semantics. A gate can verify only the policy it is given. Source independence, freshness, test coverage, semantic classifiers, and competence certificates all require operational definitions. The framework’s contribution is to expose these as typed contract elements and attempt-history state, making weak semantics testable rather than implicit.

Information-flow scope. The trace-containment theorem governs action traces. Confidentiality and noninterference require relational flow properties. The architecture supports these through labels and flow policy, but complete formal treatment of stochastic, learned, and continuous data transformations remains open.

Probabilistic dependence and model specification. The union-bound result requires valid conditional error bounds over reachable histories; the anytime-valid theorem requires a process valid under the unsafe conditional laws induced by permitted selection. A correctly specified likelihood ratio is one construction, but conservative e-values can support the same guarantee over a registered model-uncertainty set. Correlated evidence is not itself a violation when represented in the e-process; misspecified dependence, selective disclosure, and unbudgeted proposal variants break the guarantee. Measuring error under adaptive evidence and proposal search remains essential; i.i.d. classifier accuracy is insufficient.

Outcome dynamics. Authorized actions can still have surprising consequences. Simulation, rollback, environment models, and postcondition monitoring become part of the operation portfolio when outcome uncertainty is material.

Certification drift. Scoped competence certificates depend on model version, task distribution, verification procedure, policy, and statistical assumptions. Automatic revocation and revalidation under drift are necessary for the distillation flywheel to remain safe and economically useful.

Evidence level. The exact finite models in Section 9.2 demonstrate mechanisms under declared parameters. They do not establish that learned language-model routers exhibit the same behavior at operationally important rates. The tabular learned-router results apply to the registered finite generator. Language-model, human-load, learned-verifier calibration, distribution-shift, and production-system hypotheses remain open.

13 Conclusion

Dual-process cognition offers a useful starting point for agentic AI. A System-1-like path provides fast, amortized judgments and proposals; a System-2-like path spends additional resources on search, reasoning, tools, verification, and revision. The central engineering challenge is to govern the transition between those regimes and the further transition from cognition to consequential effect.

This paper develops that architecture as the agentic control plane. The metacognitive router governs

the allocation of cognition. The commitment gate governs the conversion of cognition into external effect. Evidence is the coupling variable between them: it changes beliefs, can change admissibility, and therefore changes the value of further computation. The four-corner value representation makes the belief/admissibility interaction explicit. Complementary evidence chains explain why myopic routing can under-think; the binding-gate result shows when governance can make unfavorable information privately costly; and the exact finite models demonstrate governance-induced avoidance under specified counterfactuals. Matching risk prices removes the original Toy A wedge, but does not generally remove hard-rule effects in the coupled environment.

The strongest probabilistic result is not retry amplification but its containment. A selection-valid test-supermartingale gate gives a horizon-free unsafe-admission bound for one candidate even when the router adaptively selects evidence and stopping time. The conditional-betting construction makes an operation-specific unsafe-pass bound executable; complete logging alone does not establish that bound. An episode-level proposal budget then extends control across adaptively generated candidate variants. Retry search and proposal search are therefore distinct statistical objects: selection-valid evidence controls the first; protected proposal spending controls the second. The remaining failure modes become correspondingly sharper—selective history, misspecified evidence models, semantic proposal equivalence, and inadequate episode budgeting.

The same separation clarifies inference economics. Retrospective Shapley-style resource shares are accounting objects; prospective routing should use estimated marginal decision cost. Human review introduces an additional external congestion cost that should be priced only for the burden imposed on other episodes or represented directly in a fleet-level objective. Compute per decision is therefore a coupled cognitive-and-infrastructure quantity, not a token count.

The dual-process view also clarifies learning. Verified System-2-like episodes can be distilled into cheaper System-1-like behavior, moving the fast/slow frontier over time. But competence and authority remain distinct: a model can learn to produce the right answer without carrying the runtime evidence required to act on it. Scoped certification provides one route for moving selected verified capabilities toward the fast path while retaining statistical scope, revocation, freshness, and policy boundaries.

The finite realization now separates several effects. Forward modeling of permission changes allocation relative to a specified blind forecast, but need not improve principal utility when objectives differ. A contract-compliant fast exemption can likewise raise router return while lowering principal value; probabilistic eligibility and beneficial use remain separate. Pure disqualification can suppress privately useful checking, while ordinary cost-sensitive skipping can remain unchanged by the gate response. Four-operation lookahead reaches the registered initial-distribution optimum in the two selection-valid cells, and the trained tabular policies approach rather than exceed it. An admitted action can still justify further verification: permission and stopping are distinct. Fixed-sample calibration over selectable finite types quantifies the loss of useful performance when a known verifier bound is replaced by an estimated one. These mechanisms remain tied to a tested, version-bound executor under a declared in-memory trust boundary.

The next empirical extension should hold language-model capacity, operation access, gate, and interaction budget fixed while varying representation and policy visibility. Learned semantic-verifier bounds, genuine certified distillation, human-review load, and production isolation remain separate tasks. The finite results establish where cognition allocation responds to governance; they do not assert that every more governance-aware agent is safer or that statistical calibration covers unseen selection strategies.

Behavioral alignment remains necessary because it shapes what the system tends to pursue. Runtime control provides a separate defense against errors, exploitation, staleness, and actions outside the

supported training distribution. Metacognitive and evidence assurance connect them without collapsing the distinction. As AI moves from answering to acting, safety becomes an auditable systems discipline over the complete trajectory from proposal to evidence to authority to effect.

The resulting design principle is: **think fast when the fast path is sufficient; deliberate when additional cognition has value; require disqualifying evidence when safety depends on it; account for adaptive evidence and proposal search; and commit only with evidence and authority appropriate to the effect.**

A Formal Notes on the Coupled Control Problem

A.1 Resource-constrained objective

For an episode terminating at time T , a convenient deployment objective is

$$J(\mu) = \mathbb{E} \left[u(a_T, \omega) - \sum_{t < T} \lambda^\top \hat{\kappa}^{\text{marg}}(o_t, z_t, s_t) \right], \quad (15)$$

subject to hard policy and resource constraints. The router uses prospective marginal-cost estimates in the control objective; retrospective attributed shares belong in the realized accounting ledger. Latency belongs either in the marginal resource price or in utility, but not both. Hard prohibitions remain feasibility constraints rather than finite penalties because a sufficiently large task reward can overwhelm any fixed penalty.

A.2 Why the gate belongs in the value function

If two states have identical posterior beliefs and resource budgets but different evidence or authority records, they can have different $\mathcal{A}_G(z)$ and therefore different $Q(z)$. Any router trained against a value target that ignores the gate aliases those states. The gate-coupled value function removes that aliasing by treating evidence and authority as decision state.

A.3 Approximate metareasoning

Exact solution of equation (6) is generally infeasible. Practical routers can be classified by the state they preserve:

1. *difficulty routing*: estimate how much generic compute the task needs;
2. *uncertainty routing*: escalate when predictive confidence is low;
3. *value routing*: estimate downstream improvement per resource unit;
4. *portfolio routing*: choose the operation type as well as amount;
5. *governance-aware routing*: additionally model evidence contracts, authority, and changes in admissibility;
6. *history-aware routing*: condition on prior denials, failed evidence, and retry state.

The framework predicts that the last two levels matter precisely when the gate’s state changes the downstream feasible action set.

B Evaluation Contract Template

A reproducible study should publish:

1. **Task contract:** distribution, success criteria, stakes manipulation, and world-state observability.
2. **Base model contract:** model/version, sampling parameters, context, and tool prompt.
3. **Operation registry:** reasoning, retrieval, execution, verifier, model, approval, and delegation operations with effect signatures.
4. **Resource ledger:** realized units, retrospective allocation rule, prospective marginal-cost estimator, system-load state, price vectors, and hard budgets.
5. **Evidence contract:** provenance, required versus optional checks, independence assumptions, failed-attempt semantics, freshness, and verifier likelihood/error model.
6. **Authority contract:** identities, credentials, scopes, approval conditions, delegation, revocation, certificate issuance, protected-state rules, and the policy-visibility interface exposed to the router.
7. **Information-flow contract:** data labels, allowed destinations, declassification rules, and cross-trace or noninterference tests where applicable.
8. **Commitment contract:** effects requiring mediation, allowed trace set, reversibility, and environment pre/postconditions.
9. **Baselines:** uniform slow, difficulty/uncertainty routing, a governance-blind ablation, explicit gate-aware control, and a competent history-aware learner, fully crossed with gate regimes.
10. **Adversarial suite:** prompt injection, corrupted evidence, dependent sources, stochastic verifier retry, strategic ignorance, gate probing, declassification abuse, adversarial persuasion of human approvers, stale authorization, and trust-base mutation attempts.
11. **Reporting:** utility, every resource coordinate, marginal-cost forecast error, latency, evidence attempts including failures, proposal count and error-budget spend, gate outcomes, candidate- and episode-level false admissions, unsafe share among admitted actions, flow violations, false blocks, human-review load, and registered-comparator regret.

C Contribution and Evidence Register

Table 11: Contribution and evidence map for Version 1.0.0.

Statement	Role	Scientific status
Fast/slow metacognition	Direct antecedent	SOFAI and successors already instantiate fast/slow solver selection and metacognitive governance; this manuscript does not claim priority for that architecture.
Value of computation for LLM reasoning	Direct antecedent	Rational metareasoning and adaptive-compute work already optimize selective reasoning; classical metareasoning supplies the normative foundation.
Capability and flow enforcement	Direct antecedent	Reference monitors, protection principles, and CaMeL establish complete mediation, capabilities, and trusted information-flow control.

Statement	Role	Scientific status
Evidence incentives at the router–gate boundary	Application-level synthesis	Builds on metareasoning, shielding, and proof-carrying authorization; analyzes evidence acquisition, avoidance, selection, and assurance composition.
Effect-typed operations	Systems abstraction	Replaces a brittle compute/action binary with declared effect, data, destination, authority, and reversibility fields.
Gate-coupled value attribution	Formal/measurement definition	Gives order-explicit and Shapley-symmetric attribution of decision-refinement and admissibility effects; the split is not claimed to be unique without a convention.
Binding-gate condition	Formal result	Shows that governance cannot create additional negative information value when the gate retains an unrestricted router-optimal action at every reachable node; raw check skipping also reflects acquisition costs; controlled utility/belief contrasts are required.
Fixed-set value-of-information lemma	Formal result	Under coherent posterior updating and a fixed admissible set, expected decision-refinement value is nonnegative; admissibility value can have either sign.
Conjunctive-evidence myopia	Formal consequence	Shows that one-step routing can reject every individually unprofitable evidence action even when the multi-step chain has positive net value.
Strategic ignorance	Derived failure class	Optional disqualifying checks can have negative gate-coupled value, creating an incentive to avoid evidence; exact finite parameters demonstrate the mechanism.
Gate gaming	Conditional failure class	These strategies arise when the available contract weakness and objective make them profitable; coupling alone does not imply gaming.
Protected commitment containment	Formal result	Under complete mediation, protected governance state, and sound gate enforcement, committed traces remain inside the permitted set independent of internal reasoning.
Adaptive episode-error bound	Formal result	With valid conditional per-history error bounds, episode unsafe acceptance is bounded by the sum of their expectations.
Selection-valid admission bound	Standard inequality instantiated	A test supermartingale valid under permitted selection gives a horizon-free candidate false-admission bound; conditional betting constructs one from operation-specific unsafe-pass bounds.

Statement		Role	Scientific status
Episode-level budget bound	proposal-	Formal result	Predictable candidate budgets whose episode sum is bounded control aggregate unsafe admission across adaptively generated proposal variants.
Exact finite demonstrations		Exact model result	Declared finite models exhibit strategic ignorance, the binding-gate/router-objective wedge, an interior positive-information case, retry selection, and the safe–unsafe admission frontier; they are not deployment evidence.
Information-flow separation		Formal scope distinction	Trace safety does not establish confidentiality hyperproperties; flow labels, declassification, and relational policies protect a separate object.
Certified competence		Formal condition / finite realization	Selected-entry risk, rather than population risk alone, governs substitution. The finite fast route is exact within its entry restriction; learned calibration and drift remain open.
Layered system assurance		Systems synthesis	Keeps behavioral alignment distinct from runtime control while connecting metacognition, evidence, containment, monitoring, and learning in one evaluation stack.
Finite coupled control and executor		Constructive model and training result	558 states per cell; exact and depth-two planning; 40 Q-learning runs; exact frozen-policy evaluation; 482 terminal paths replayed through the protected reference gateway.
Permission-response interventions	inter-	Exact finite comparisons	192 controller/contract configurations and fully retained operation/risk/cost tables; distinguishes future-mask foresight from current enforcement. Not a causal representation-learning comparison.
Planning depth and stopping		Exact finite comparisons	Depth four reaches the two valid-cell initial-distribution optima. A reached admitted state strictly prefers further verification. Retained Q-learning is not compared against an artificially short horizon alone.
Fixed-sample verifier calibration	verifier cali-	Constructive statistical study	Exact upper-grid CP endpoints over three registered subtypes; all calibration outcomes integrated. Exercises shared calibration uncertainty, but does not cover unknown semantic subtypes or certify neural competence.

Statement	Role	Scientific status
Governance-aware routing improves real systems	Open empirical claim	Tabular learning is demonstrated. A general language-model or production advantage, and semantic-verifier validity under distribution shift, remain unestablished.

References

- [1] Daniel Kahneman. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux, 2011.
- [2] Jonathan St. B. T. Evans and Keith E. Stanovich. “Dual-Process Theories of Higher Cognition: Advancing the Debate”. In: *Perspectives on Psychological Science* 8.3 (2013), pp. 223–241. DOI: 10.1177/1745691612460685.
- [3] David E. Melnikoff and John A. Bargh. “The Mythical Number Two”. In: *Trends in Cognitive Sciences* 22.4 (2018), pp. 280–293. DOI: 10.1016/j.tics.2018.02.001.
- [4] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. “Training Language Models to Follow Instructions with Human Feedback”. In: *Advances in Neural Information Processing Systems*. Vol. 35. 2022.
- [5] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. “Constitutional AI: Harmlessness from AI Feedback”. In: *arXiv preprint arXiv:2212.08073* (2022).
- [6] Grady Booch, Francesco Fabiano, Lior Horesh, Kiran Kate, Jonathan Lenchner, Nick Linck, Andrea Loreggia, Keerthiram Murugesan, Nicholas Mattei, Francesca Rossi, and Biplav Srivastava. “Thinking Fast and Slow in AI”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 17. 2021, pp. 15042–15046. DOI: 10.1609/aaai.v35i17.17765.
- [7] Marianna Bergamaschi Ganapini, Murray Campbell, Francesco Fabiano, Lior Horesh, Jonathan Lenchner, Andrea Loreggia, Nicholas Mattei, Francesca Rossi, Biplav Srivastava, and Kristen Brent Venable. “Fast, Slow, and Metacognitive Thinking in AI”. In: *npj Artificial Intelligence* 1 (2025), p. 27. DOI: 10.1038/s44387-025-00027-5.
- [8] Vedant Khandelwal, Vishal Pallagani, Biplav Srivastava, and Francesca Rossi. “A Neurosymbolic Fast and Slow Architecture for Graph Coloring”. In: *Advances in Cognitive Systems* (2025). Twelfth Annual Conference on Advances in Cognitive Systems; arXiv:2412.01752.
- [9] Vedant Khandelwal, Francesca Rossi, Keerthiram Murugesan, Erik Miehl, Murray Campbell, Karthikeyan Natesan Ramamurthy, and Lior Horesh. “Language Models Coupled with Metacognition Can Outperform Reasoning Models”. In: *arXiv preprint arXiv:2508.17959* (2025).
- [10] Herbert A. Simon. “A Behavioral Model of Rational Choice”. In: *The Quarterly Journal of Economics* 69.1 (1955), pp. 99–118. DOI: 10.2307/1884852.
- [11] Stuart J. Russell and Devika Subramanian. “Provably Bounded-Optimal Agents”. In: *Journal of Artificial Intelligence Research* 2 (1995), pp. 575–609. DOI: 10.1613/jair.133.
- [12] Eric J. Horvitz. “Reasoning about Beliefs and Actions under Computational Resource Constraints”. In: *Proceedings of the Third Workshop on Uncertainty in Artificial Intelligence*. 1987, pp. 429–444.
- [13] Stuart Russell and Eric Wefald. “Principles of Metareasoning”. In: *Artificial Intelligence* 49.1–3 (1991), pp. 361–395. DOI: 10.1016/0004-3702(91)90015-C.

- [14] Nicholas Hay, Stuart Russell, David Tolpin, and Solomon Eyal Shimony. “Selecting Computations: Theory and Applications”. In: *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*. 2012, pp. 346–355.
- [15] Falk Lieder and Thomas L. Griffiths. “Resource-Rational Analysis: Understanding Human Cognition as the Optimal Use of Limited Computational Resources”. In: *Behavioral and Brain Sciences* 43 (2020), e1. DOI: 10.1017/S0140525X1900061X.
- [16] C. Nicolò De Sabbata, Theodore R. Sumers, and Thomas L. Griffiths. “Rational Metareasoning for Large Language Models”. In: *NeurIPS 2024 Workshop on Behavioral Machine Learning*. Workshop version; associated arXiv record: 2410.05563. 2024.
- [17] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models”. In: *Advances in Neural Information Processing Systems*. Vol. 35. 2022.
- [18] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. “Self-Consistency Improves Chain of Thought Reasoning in Language Models”. In: *International Conference on Learning Representations*. 2023.
- [19] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. “ReAct: Synergizing Reasoning and Acting in Language Models”. In: *International Conference on Learning Representations*. 2023.
- [20] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. “Tree of Thoughts: Deliberate Problem Solving with Large Language Models”. In: *Advances in Neural Information Processing Systems*. Vol. 36. 2023. DOI: 10.52202/075280-0517.
- [21] Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. “Toolformer: Language Models Can Teach Themselves to Use Tools”. In: *Advances in Neural Information Processing Systems*. Vol. 36. 2023.
- [22] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. “Reflexion: Language Agents with Verbal Reinforcement Learning”. In: *Advances in Neural Information Processing Systems*. Vol. 36. 2023.
- [23] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. “Scaling LLM Test-Time Compute Optimally Can Be More Effective Than Scaling Parameters for Reasoning”. In: *International Conference on Learning Representations*. 2025.
- [24] Amrith Setlur, Nived Rajaraman, Sergey Levine, and Aviral Kumar. “Scaling Test-Time Compute Without Verification or RL Is Suboptimal”. In: *Proceedings of the 42nd International Conference on Machine Learning*. Vol. 267. Proceedings of Machine Learning Research. 2025, pp. 54058–54094.
- [25] Zhiyuan Zhai, Bingcong Li, Bingnan Xiao, Ming Li, and Xin Wang. “Adaptive Test-Time Compute Allocation for Reasoning LLMs via Constrained Policy Optimization”. In: *arXiv preprint arXiv:2604.14853* (2026).
- [26] Yuxiao Qu, Matthew Y. R. Yang, Amrith Setlur, Lewis Tunstall, Edward Emanuel Beeching, Ruslan Salakhutdinov, and Aviral Kumar. “Optimizing Test-Time Compute via Meta Reinforcement Finetuning”. In: *Proceedings of the 42nd International Conference on Machine Learning*. Vol. 267. Proceedings of Machine Learning Research. 2025, pp. 50893–50925.
- [27] Lingjiao Chen, Matei Zaharia, and James Zou. “FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance”. In: *Transactions on Machine Learning Research* (2024).

- [28] Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E. Gonzalez, M. Waleed Kadous, and Ion Stoica. “RouteLLM: Learning to Route LLMs with Preference Data”. In: *arXiv preprint arXiv:2406.18665* (2024).
- [29] Dujian Ding, Ankur Mallick, Shaokun Zhang, Chi Wang, Daniel Madrigal, Mirian Del Carmen Hipolito Garcia, Menglin Xia, Laks V. S. Lakshmanan, Qingyun Wu, and Victor Ruehle. “BEST-Route: Adaptive LLM Routing with Test-Time Optimal Compute”. In: *Proceedings of the 42nd International Conference on Machine Learning*. Vol. 267. Proceedings of Machine Learning Research. 2025, pp. 13870–13884.
- [30] Chenrui Fan, Yize Cheng, Ming Li, Yongyuan Liang, Tianyi Zhou, and Soheil Feizi. “Thinking Hard, Not Smart: Reasoning Models Fail to Ration Test-Time Compute Across Questions”. In: *arXiv preprint arXiv:2608.07968* (2026).
- [31] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Samuel Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. “Language Models (Mostly) Know What They Know”. In: *arXiv preprint arXiv:2207.05221* (2022).
- [32] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. “Detecting Hallucinations in Large Language Models Using Semantic Entropy”. In: *Nature* 630 (2024), pp. 625–630. DOI: 10.1038/s41586-024-07421-0.
- [33] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. “Training Verifiers to Solve Math Word Problems”. In: *arXiv preprint arXiv:2110.14168* (2021).
- [34] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. “Let’s Verify Step by Step”. In: *arXiv preprint arXiv:2305.20050* (2023).
- [35] Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. “Large Language Models Cannot Self-Correct Reasoning Yet”. In: *International Conference on Learning Representations*. 2024.
- [36] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. “Language Models Don’t Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting”. In: *Advances in Neural Information Processing Systems*. Vol. 36. 2023. DOI: 10.52202/075280-3275.
- [37] Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y. Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. “Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation”. In: *arXiv preprint arXiv:2503.11926* (2025).
- [38] Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca Dragan, Scott Emmons, Owain Evans, David Farhi, Ryan Greenblatt, Dan Hendrycks, Marius Hobbhahn, Evan Hubinger, Geoffrey Irving, Erik Jenner, Daniel Kokotajlo, Victoria Krakovna, Shane Legg, David Lindner, David Luan, Aleksander Mądry, Julian Michael, Neel Nanda, Dave Orr, Jakub Pachocki, Ethan Perez, Mary Phuong, Fabien Roger, Joshua Saxe, Buck Shlegeris, Martín Soto, Eric Steinberger, Jasmine Wang, Wojciech Zaremba, Bowen Baker, Rohin Shah, and Vlad Mikulik. “Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety”. In: *arXiv preprint arXiv:2507.11473* (2025).

- [39] James P. Anderson. *Computer Security Technology Planning Study*. Tech. rep. ESD-TR-73-51. Electronic Systems Division, Air Force Systems Command, Oct. 1972.
- [40] Jerome H. Saltzer and Michael D. Schroeder. “The Protection of Information in Computer Systems”. In: *Proceedings of the IEEE* 63.9 (1975), pp. 1278–1308. DOI: 10.1109/PROC.1975.9939.
- [41] Edoardo Debenedetti, Jie Zhang, Mislav Balunovic, 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*. Vol. 37. 2024. DOI: 10.52202/079017-2636.
- [42] 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”. In: *arXiv preprint arXiv:2503.18813* (2025). arXiv v2 dated 24 June 2025; accepted at IEEE SaTML 2026.
- [43] Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. “AI Control: Improving Safety Despite Intentional Subversion”. In: *Proceedings of the 41st International Conference on Machine Learning*. Vol. 235. Proceedings of Machine Learning Research. 2024, pp. 16295–16336.
- [44] Segev Shlomov, Iftach Shoham, Alon Oved, Ido Levy, Sami Marreed, Harold Ship, Offer Akrabi, Sergey Zeltyn, Avi Yaeli, and Nir Mashkif. “Governance by Construction for Generalist Agents”. In: *Proceedings of the ACM Conference on AI and Agentic Systems*. 2026, pp. 1280–1287. DOI: 10.1145/3786335.3813192.
- [45] Michael R. Clarkson and Fred B. Schneider. “Hyperproperties”. In: *Journal of Computer Security* 18.6 (2010), pp. 1157–1210. DOI: 10.3233/JCS-2009-0393.
- [46] Emir Kamenica and Matthew Gentzkow. “Bayesian Persuasion”. In: *American Economic Review* 101.6 (2011), pp. 2590–2615. DOI: 10.1257/aer.101.6.2590.
- [47] Paul R. Milgrom. “Good News and Bad News: Representation Theorems and Applications”. In: *Bell Journal of Economics* 12.2 (1981), pp. 380–391.
- [48] Ronald A. Dye. “Disclosure of Nonproprietary Information”. In: *Journal of Accounting Research* 23.1 (1985), pp. 123–145. DOI: 10.2307/2490910.
- [49] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. “STaR: Bootstrapping Reasoning With Reasoning”. In: *Advances in Neural Information Processing Systems*. Vol. 35. 2022.
- [50] Ping Yu, Jing Xu, Jason E. Weston, and Ilia Kulikov. “Distilling System 2 into System 1”. In: *NeurIPS 2024 Workshop on System 2 Reasoning at Scale*. 2024.
- [51] 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*. Vol. 32. 1. 2018, pp. 2669–2678. DOI: 10.1609/aaai.v32i1.11797.
- [52] Giulio Mazzi, Alberto Castellini, and Alessandro Farinelli. “Risk-aware shielding of Partially Observable Monte Carlo Planning policies”. In: *Artificial Intelligence* 324 (2023), p. 103987. DOI: 10.1016/j.artint.2023.103987.
- [53] Andrew W. Appel and Edward W. Felten. “Proof-carrying authentication”. In: *Proceedings of the 6th ACM Conference on Computer and Communications Security*. 1999, pp. 52–62. DOI: 10.1145/319709.319718.
- [54] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Re. “FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness”. In: *Advances in Neural Information Processing Systems*. Vol. 35. 2022. DOI: 10.52202/068431-1189.
- [55] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. “Efficient Memory Management for Large Language Model

- Serving with PagedAttention”. In: *Proceedings of the 29th Symposium on Operating Systems Principles*. 2023. DOI: 10.1145/3600006.3613165.
- [56] Lloyd S. Shapley. “A Value for n -Person Games”. In: *Contributions to the Theory of Games II*. Ed. by Harold W. Kuhn and Albert W. Tucker. Princeton University Press, 1953, pp. 307–317.
 - [57] Daniel Golovin and Andreas Krause. “Adaptive Submodularity: Theory and Applications in Active Learning and Stochastic Optimization”. In: *Journal of Artificial Intelligence Research* 42 (2011), pp. 427–486. DOI: 10.1613/jair.3278.
 - [58] Raja Parasuraman and Dietrich H. Manzey. “Complacency and Bias in Human Use of Automation: An Attentional Integration”. In: *Human Factors* 52.3 (2010), pp. 381–410. DOI: 10.1177/0018720810376055.
 - [59] Kate Goddard, Abdul Roudsari, and Jeremy C. Wyatt. “Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators”. In: *Journal of the American Medical Informatics Association* 19.1 (2012), pp. 121–127. DOI: 10.1136/amiajnl-2011-000089.
 - [60] Dean P. Foster and Robert A. Stine. “ α -Investing: A Procedure for Sequential Control of Expected False Discoveries”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70.2 (2008), pp. 429–444. DOI: 10.1111/j.1467-9868.2007.00643.x.
 - [61] Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. “Game-Theoretic Statistics and Safe Anytime-Valid Inference”. In: *Statistical Science* 38.4 (2023), pp. 576–601. DOI: 10.1214/23-STS894.
 - [62] Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. “Learn Then Test: Calibrating Predictive Algorithms to Achieve Risk Control”. In: *The Annals of Applied Statistics* 19.2 (2025), pp. 1641–1662. DOI: 10.1214/24-AOAS1998. arXiv: 2110.01052.
 - [63] Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. “Conformal Risk Control”. In: *International Conference on Learning Representations*. 2024.
 - [64] Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. “Multicalibration: Calibration for the (Computationally-Identifiable) Masses”. In: *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80. Proceedings of Machine Learning Research. 2018, pp. 1939–1948. URL: <https://proceedings.mlr.press/v80/hebert-johnson18a.html>.
 - [65] Fitih M. Cinnor. *The Agentic Brain: Finite Reference System and Reproducibility Supplement*. Version 1.0.0. Python sources, exact intervention and calibration results, retained trained tabular policies, gateway tests, content hashes, and build instructions supplied with this manuscript. Local versioned artifact; no archival DOI assigned. 2026.
 - [66] Christopher J. C. H. Watkins and Peter Dayan. “Q-learning”. In: *Machine Learning* 8 (1992), pp. 279–292. DOI: 10.1007/BF00992698.
 - [67] C. J. Clopper and E. S. Pearson. “The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial”. In: *Biometrika* 26.4 (1934), pp. 404–413. DOI: 10.1093/biomet/26.4.404.
 - [68] Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. “Time-uniform, Non-parametric, Nonasymptotic Confidence Sequences”. In: *The Annals of Statistics* 49.2 (2021), pp. 1055–1080. DOI: 10.1214/20-AOS1991. arXiv: 1810.08240.
 - [69] David Lyell and Enrico Coiera. “Automation Bias and Verification Complexity: A Systematic Review”. In: *Journal of the American Medical Informatics Association* 24.2 (2017), pp. 423–431. DOI: 10.1093/jamia/ocw105.
 - [70] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mane. “Concrete Problems in AI Safety”. In: *arXiv preprint arXiv:1606.06565* (2016).

Collective Attribution, Accountability, and AI Research Provenance

Fitih M. Cinnor initiated, directed, and led the research program; established its scientific objectives and evidentiary standards; selected and adjudicated the claims advanced here; and reviewed the supporting derivations, literature, computations, and manuscript. He is the named and accountable human author, responsible for the accuracy, integrity, scope, release decisions, and correction of the work. Accountable correspondence: fcinnor@semiaifoundry.com.

semiAIfoundry Research, a research group of Semi AI Foundry LLC, is the collective scholarly attribution for the coordinated research effort. It recognizes integrated intellectual work across human, artificial, and hybrid forms of intelligence while preserving named human accountability.

The research used a human-directed, orchestrated AI-assisted process. AI systems made material contributions to conceptual exploration, cross-domain synthesis, literature analysis, mathematical formalization, counterexample and failure-mode construction, derivation checking, visualization, drafting, editing, and publication engineering. **OpenAI GPT-6 Astra** (manuscript and reference implementation, 3 October 2026) receives explicit credit for conceptual synthesis, formal development of the agentic control-plane framework, mathematical analysis, literature integration, finite-model implementation, controller training, automated checks, technical writing, and manuscript construction. **Anthropic Claude Opus 5.5** (review material supplied 1 October 2026) receives explicit credit for adversarial scholarly critique, literature challenge, gap identification, and technical review.

These credits record research provenance, not scientific status or legal authorship. Scientific standing rests on the arguments, derivations, sources, experiments, reproducibility record, and correction process. AI systems are not legal authors, owners, signatories, authorized representatives, or accountable persons. Provider or model identification does not imply sponsorship, approval, warranty, or endorsement of the work. The research program received no external funding. The accountable human author declares no competing interests or conflicts of interest.

Copyright, licensing, and artifact availability

Copyright © 2026 Semi AI Foundry LLC. Except where otherwise noted, the manuscript and project-created figures in the exact versioned public payload are licensed under CC BY 4.0, with requested attribution to “Fitih M. Cinnor for semiAIfoundry Research.” Third-party materials remain governed by their source terms. Versioned source, bibliography, figures, and release records define the manuscript payload for this release; the accompanying package contains the executable finite models, retained trained Q tables, exact result tables, intervention and calibration studies, reference gateway, and tests. The calibration endpoints are verified on a rational grid, and result manifests distinguish original learned records from new exact enumerations. A file manifest identifies this versioned package. Manuscript DOI: [10.5281/zenodo.23125976](https://doi.org/10.5281/zenodo.23125976). Public code licensing is reserved for the accountable author’s release decision.