Mind the Mediation Gap

Anonymous ACL submission

Abstract

Alignment is usually treated as optimization against a reward or rule-based objective. We study a different source of loss, which is the interface through which an aligned system is mediated. We introduce the Mediation Gap, separating unavoidable structural loss from finite-budget dynamic loss. Architectures induce local filtrations and feasible policy classes. Structural MG equals the minimum achievable oracle regret under those constraints. Dynamic MG becomes a quotient learning problem inside the fixed architecture. In a guardrail-mediated generation experiment, richer routing lowers structural MG but can increase realized finite-budget MG. The framework explains when alignment failures come from missing capacity versus unrealized capacity.

1 Introduction

We study a basic question in reinforcement learning and alignment. How much performance is lost because an aligned system has to act through a mediated architecture. Current alignment methods are usually framed as optimization procedures. Given a reward function, a learned reward model, a preference objective, or a rule based surrogate, the algorithm searches for a policy with high expected value. RLHF (Ouyang et al., 2022), Constitutional AI (Bai et al., 2022), DPO (Rafailov et al., 2023), GRPO (Shao et al., 2024), and related pipelines largely follow this logic. The central difficulty is then taken to be estimation, optimization, robustness, and reward misspecification.

Modern systems introduce another difficulty. We do not directly control the internal state dynamics that implement aligned behavior. Instead, influence is mediated through interfaces. Reward models compress preferences into scalar feedback. Constitutional principles turn broad intent into checkable rules. Safety classifiers, verifiers, routers, and policy filters decide which signals are visible and which actions are allowed. Alignment is therefore carried out through a hierarchy of mediated components. A high level goal must be interpreted, routed, and executed through a restricted architecture.

This paper formalizes the resulting loss as the *Mediation Gap* (MG). An architecture induces local information available to each component, hence a feasible policy class. The total gap separates into two terms. The *structural MG* is the value that no policy inside the mediated architecture can attain, even with perfect training. The *dynamic MG* is the value left unrealized because learning inside that architecture is finite and feedback limited. This split distinguishes impossibility from undertraining.

We develop the theory in three steps. First, we define architectures through filtrations and obtain policy class nesting from measurability. Second, we show that structural MG equals minimum achievable oracle regret under architectural constraints. Third, we study dynamic MG as a quotient learning problem. The learner need only identify the part of the environment that matters under the fixed architecture, while hidden value distinctions inside statistically indistinguishable cells can create persistent dynamic loss.

We believe the framework changes how we approach the long discussed alignment tax idea. A guardrail pipeline can fail because the safety evidence is absent, because it is present but poorly routed, or because training fails to exploit a sufficient interface. Our experiment instantiates this distinction in a sequential guardrail mediated generation task. We find that richer mediation can reduce structural MG while increasing finite budget dynamic MG. Thus a weaker architecture can look better after training because it is easier to saturate. The implication is operational. Alignment quality depends on both what the interface preserves and

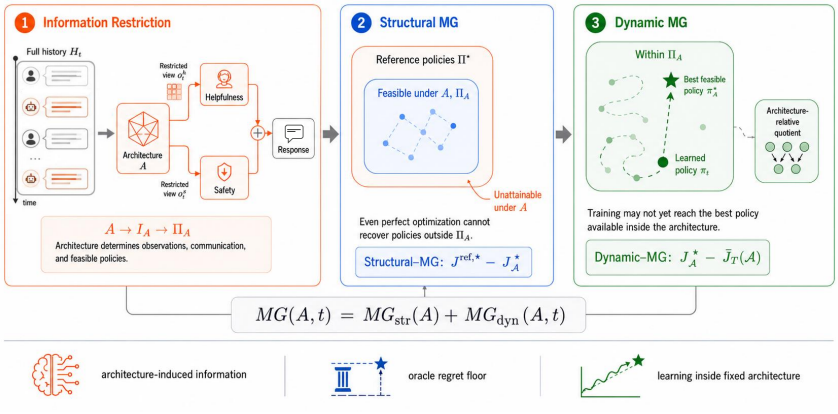


Figure 1: Overview of the Mediation Gap framework.

how learnable that preserved capacity is.

Related work. Altabaa and Yang study how information structure controls sample complexity in partially observable sequential teams and games (Altabaa and Yang, 2024). Zhang et al. develop representation learning and planning guarantees for POMDPs (Zhang et al., 2024). These works analyze learnability under hidden state, while MG isolates how a delegated interface restricts what can be acted on and learned. In alignment, reward overoptimization shows that optimizing a learned feedback channel can diverge from true value (Rafailov et al., 2024), and recent methods add pessimism or stronger regularization to control this failure (Huang et al., 2024a). Reward model evaluation asks whether proxy scores predict downstream RLHF performance (Frick et al., 2024). A more detailed discussion of related work and implications on modern systems can be found in Appendix B and C.

2 Preliminary

We begin from a simple observation. A capable system often acts through mediation. Full task context is available somewhere in the system, yet the final action is produced through modules, routers, critics, or specialized sub-policies that each see only part of that context. Once this happens, performance can fall for two very different reasons. One reason comes from the architecture itself. The architecture determines what each agent can observe, what messages can be exchanged, and what information can be retained over time. The second reason comes from learning. Even within a fixed architecture, training may fail to reach the best policy available in that architecture.

Running Example. Consider a guardrail-mediated generation pipeline. A user query and its conversational context induce a full history H_t, as shown in Figure 1. A mediator routes restricted views to two delegates. The *Helpfulness* model represents the harmless assistant that tries to answe the user query. The *Safety* model represents a safeguard model (e.g., constitutional classifier, policy filter). The mediator is the orchestration layer that decides what each component sees, whether clarification is available, and how safety evidence affects the final response. If this interface preserves and routes the decision-relevant evidence, the mediated system can match the reference response. If it loses or blocks that evidence, a structural mediation gap remains even under perfect training. If the interface is sufficient but training fails to use it, the gap is dynamic. This running example will reappear throughout the paper. It gives an intuitive interpretation of the theory and the experiment.

2.1 Architectures and Policy Classes

We formulate the framework on histories. Let H_t denote the full history up to time t. An architecture $\mathcal{A}$ induces, for each agent m and time t, an information set represented by a filtration $\mathcal{I}_t^{m,\mathcal{A}} \subseteq \sigma(H_t)$. This induces a policy class

$$\Pi_{\mathcal{A}} = \{\pi : \pi_t^m \text{ is } \mathcal{I}_t^{m,\mathcal{A}}\text{-measurable for all } m, t\}.$$

The task objective is described by a single performance functional J. This objective remains fixed throughout the framework. We also fix a reference policy class Π^{ref}, which provides the benchmark $J^{\mathrm{ref},\star} = \sup_{\pi\in\Pi^{\mathrm{ref}}} J(\pi)$. Then for any architecture $\mathcal{A}$, the best value achievable inside its induced class is $J^\star_{\mathcal{A}} = \sup_{\pi\in\Pi_{\mathcal{A}}} J(\pi)$.

2.2 Decomposition

Suppose a learning procedure under architecture $\mathcal{A}$ produces policies $\pi_1, \ldots, \pi_T \in \Pi_{\mathcal{A}}$. We define the average true performance by

$$\bar{J}_T(\mathcal{A}) = \frac{1}{T} \sum_{t=1}^{T} J(\pi_t),$$

which is the average value of the realized policies under the target objective J. The Mediation Gap relative to the reference class is

$$\begin{aligned} \mathrm{MG}_T(\mathcal{A} \mid \Pi^{\mathrm{ref}}) &= J^{\mathrm{ref},\star} - \bar{J}_T(\mathcal{A}) \\ &= \underbrace{J^{\mathrm{ref},\star} - J^{\star}_{\mathcal{A}}}_{\text{structural}} + \underbrace{J^{\star}_{\mathcal{A}} - \bar{J}_T(\mathcal{A})}_{\text{dynamic}}. \end{aligned}$$

The first term depends only on the architecture through its induced policy class. It measures what the architecture makes unattainable even under ideal optimization. The second term depends on the realized learning process inside that class. It measures the gap between what the delegated architecture can express and what training actually realizes. This identity is the backbone of the MG framework. In the running example (Figure 1), if the Helpfulness and Safety views lose information needed for the best response, the loss is structural. If that information is preserved but training still fails to use it well, the loss is dynamic.

2.3 Nesting

Architectures can be ordered by the information they provide. We write $\mathcal{A}_2 \preceq \mathcal{A}_1$ when every agent under $\mathcal{A}_1$ has access to an information filtration that contains the one induced by $\mathcal{A}_2$, that is,

$$\mathcal{I}_t^{m,\mathcal{A}_2} \subseteq \mathcal{I}_t^{m,\mathcal{A}_1} \quad \text{for all } m, t.$$

This order matters because information inclusion induces policy class inclusion. The formal statement is given in Proposition A.1 in the appendix. The implication is intuitive. If a richer architecture gives every agent at least the information available under a poorer one, then every policy admissible under the poorer architecture can be implemented under the richer architecture as well.

This principle provides a clean basis for structural ladders. If $\mathcal{A}^{(0)} \preceq \mathcal{A}^{(1)} \preceq \cdots \preceq \mathcal{A}^{(L)}$, then $\Pi_{\mathcal{A}^{(0)}} \subseteq \Pi_{\mathcal{A}^{(1)}} \subseteq \cdots \subseteq \Pi_{\mathcal{A}^{(L)}}$, which implies $J^{\star}_{\mathcal{A}^{(0)}} \leq J^{\star}_{\mathcal{A}^{(1)}} \leq \cdots \leq J^{\star}_{\mathcal{A}^{(L)}}$.

Along such a ladder, the structural gap telescopes exactly: $J^{\star}_{\mathcal{A}^{(L)}} - J^{\star}_{\mathcal{A}^{(k)}} = \sum_{\ell=k+1}^{L} \left(J^{\star}_{\mathcal{A}^{(\ell)}} - J^{\star}_{\mathcal{A}^{(\ell-1)}} \right)$. Two common sources of architectural order are important in the sequel. One comes from observation design. If one observation process is obtained as a measurable coarsening of another, the induced policy classes are nested. The other comes from coordination design. If one communication graph extends another and the richer protocol can emulate the poorer one, the induced policy classes are again nested. See Propositions A.2 and A.3 for formal statement.

3 Structural Capacity

Now quantify structural value loss from architectural information constraints. Previous work studies compression within fixed architecture. State abstraction characterizes when states can be merged without value loss (Abel et al., 2016, 2018); value equivalence identifies when compressed models preserve Bellman structure (Grimm et al., 2020, 2021). Here, architecture itself is the constraint, which determines which history distinctions are observable before compression or learning begins.

3.1 Running Example

A full prompt history H is mapped into two mediated views,

$$Z^{\mathrm{help}} = \phi^{\mathrm{help}}(H), \qquad Z^{\mathrm{safe}} = \phi^{\mathrm{safe}}(H).$$

A *Helpfulness* delegate reads Z^{help}. A *Safety* delegate reads Z^{safe}. Their mediated interaction produces the final action A. Let $U(a, h)$ be the utility of action a at history h, and let $V^{\star}(h) := \max_a U(a, h)$ be the full information optimum in this one step problem.

To separate information loss from coordination details, let $\mathcal{Q}_{\phi}(z) \subseteq \Delta(\mathcal{A})$ denote the set of action distributions realizable after the mediated signal $Z = z$ has been produced. The best delegated value is

$$J^{\star}_{\phi} = \mathbb{E}\left[\sup_{q \in \mathcal{Q}_{\phi}(Z)} \sum_a q(a)\, \mathbb{E}[U(a, H) \mid Z] \right].$$

Proposition 1 (Conditional delegated regret)**.** *Define*

$$\begin{aligned} \delta_{\phi}(z) = \mathbb{E}[V^{\star}(H) \mid Z = z] \\ - \sup_{q \in \mathcal{Q}_{\phi}(z)} \sum_a q(a)\, \mathbb{E}[U(a, H) \mid Z = z]. \end{aligned}$$

Then $\mathbb{E}[V^{\star}(H)] - J^{\star}_{\phi} = \mathbb{E}[\delta_{\phi}(Z)]$.

Remark. Structural loss is exactly the Bayes regret induced by the mediated information cell. If the mediated views erase distinctions that matter for the reference decision, then the regret remains positive even under perfect optimization. When $\mathcal{Q}_\phi(z) = \Delta(\mathcal{A})$ for every z, the second term reduces to the ordinary Bayes action value $\max_a \mathbb{E}[U(a, H) \mid Z = z]$.

3.2 Regret Identity

We now extend the model to the general adaptive setting. Let H_t be the full history at time t. Let A_t be the action selected by a delegated policy $\pi \in \Pi_a$. The reference benchmark from Section 2 is the centralized causal class with full history access. Its Bellman optimal value functions are written $V^\star(h_t), Q^\star(h_t, a)$. Define the one step oracle regret $d^\star(h_t, a) = V^\star(h_t) - Q^\star(h_t, a)$.

Theorem 1 (Oracle regret identity)**.** *For every delegated policy $\pi \in \Pi_a$,*

$$J^\star - J(\pi) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t d^\star(H_t, A_t)\right].$$

Theorem 1 works directly on adaptive trajectories. It places the architectural question on the same footing as the infinite horizon control problem itself. A delegated policy incurs structural loss exactly through the discounted oracle regret accumulated along the histories that it generates.

Corollary 1 (Structural regret)**.** *For every architecture a, $J^\star - J_a^\star = \inf_{\pi \in \Pi_a} \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t d^\star(H_t, A_t)\right]$.*

Corollary 1 connects to the structural term. The architecture enters only through the information it makes available. Sparse communication, restricted observations, and limited memory all become instances of the same object. They change the feasible class Π_a, and resulting loss is the smallest oracle regret compatible with that class.

3.3 Capacity Curve

Let $\mathcal{A}(R)$ denote the family of causal architectures feasible under budget R. The budget may encode communication rate, message alphabet size, memory, observation complexity, or a joint resource constraint over several such dimensions. Larger budgets enlarge the feasible family. Define

$$\mathrm{MG}_{\mathrm{str}}(R) = J^\star - \sup_{a \in \mathcal{A}(R)} J_a^\star.$$

Theorem 2 (Endogenous mediation capacity)**.** *The structural capacity curve satisfies*

$$\mathrm{MG}_{\mathrm{str}}(R) = \inf_{a \in \mathcal{A}(R)} \inf_{\pi \in \Pi_a} D_\gamma^\star(\pi),$$

where $D_\gamma^\star(\pi) = \mathbb{E}_\pi[\sum_{t \geq 0} \gamma^t d^\star(H_t, A_t)]$.

Remark. Theorem 2 gives the best oracle regret achievable under a constrained causal architecture. The term endogenous refers to the feedback loop created by the architecture. The architecture determines the policy class, the policy class shapes the trajectories, and regret is measured on those trajectories.

Corollary 2 (Zero loss criterion)**.** $\mathrm{MG}_{\mathrm{str}}(R) = 0$ *iff* $\inf_{a \in \mathcal{A}(R)} \inf_{\pi \in \Pi_a} \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t d^\star(H_t, A_t)\right] = 0$. *When the infimum is attained, this holds if and only if $d^\star(H_t, A_t) = 0$ almost surely for every t under some $a \in \mathcal{A}(R)$ and $\pi \in \Pi_a$.*

The one step running example from Section 3.1 is the static specialization of Theorem 2. There the trajectory has one step, the discounted sum has one term, and structural MG reduces to conditional Bayes regret under mediated views.

4 Dynamic Learning Capacity

We now turn to the second half of the MG. Once an architecture is fixed, the structural term from Section 2 and Section 3 is fixed as well. What remains is a learning problem inside the delegated interface. This distinction is central in modern alignment. A reward model, a judge, a verifier, or a safety filter does more than score outputs. It determines which task distinctions can ever be revealed through feedback. A converse asks what uncertainty survives this interface. An upper bound asks for an explicit learner that extracts the right distinctions through the same interface while keeping the information budget under control. Recent work on representation learning for POMDPs, and reward overoptimization in alignment makes this theme concrete (Gao et al., 2023; Rafailov et al., 2024).

As we discussed, structural MG asks what the architecture erases before deployment begins. Dynamic MG asks what the learner still has to discover after the architecture has been fixed.

4.1 Running Example

A hidden environment variable Θ captures a latent mode of the task. One may read Θ as a hidden

user type, a hidden safety regime, or a hidden normative distinction in the prompt distribution. The architecture a fixes a delegated policy class Π_a. In each round we deploy one policy $\pi \in \Pi_a$, receive an observable feedback signal Y, and evaluate the resulting episode by its true value $J_\Theta(\pi)$.

Suppose $\Theta \in \{\theta^+, \theta^-\}$ under prior μ_0. Let K_θ^π denote the law of the observable feedback Y when policy π is deployed in environment θ. Let $J_{a,\theta}^\star = \sup_{\pi\in\Pi_a} J_\theta(\pi)$. Assume the architecture induces the same observable feedback law in both latent modes, $K_{\theta^+}^\pi = K_{\theta^-}^\pi$ for every $\pi \in \Pi_a$.

Define the Bayes gap of the repeated toy by

$$\delta_a^{\text{toy}} = \inf_{\pi\in\Pi_a} \sum_{s\in\{+,-\}} \mu_0(\theta^s)\big(J_{a,\theta^s}^\star - J_{\theta^s}(\pi)\big).$$

Proposition 2 (Dynamic floor)**.** *If $K_{\theta^+}^\pi = K_{\theta^-}^\pi$ for every $\pi \in \Pi_a$, then for every learner L and every horizon T, $\mathcal{D}_{a,T}(\mathsf{L};\mu_0) \geq \delta_a^{\text{toy}}$. If no single policy in Π_a attains $J_{a,\theta}^\star$ for both $\theta = \theta^+$ and $\theta = \theta^-$, then $\delta_a^{\text{toy}} > 0$.*

Remark. Proposition 2 shows that even after repeated interaction, the learner never receives evidence that separates θ^+ from θ^-. If those two latent modes induce different best delegated policies, a positive regret floor remains forever. This is already an alignment problem. A reward model or a judge can produce the same visible signal on two cases whose best aligned responses differ. Then training keeps moving inside an observation cell that is statistically stable while still carrying unresolved value ambiguity.

4.2 Bayesian Learning

Now pass to the general model. Fix an architecture a, hence a feasible policy class Π_a. Each round n is one deployment episode. Inside that episode the deployed policy runs for the full discounted control problem. Its value in environment θ is

$$J_\theta(\pi) = \mathbb{E}_{\theta,\pi}\Big[\sum_{t=0}^{\infty} \gamma^t r_\theta(H_t, A_t)\Big].$$

Now the connection is clear. Section 3 studies value loss inside an episode. Section 4 studies how repeated episodes reveal the environment well enough to recover the best policy available inside the same architecture.

Let $\Theta \sim \mu_0$ be an unknown environment. At round n, the learner selects $\pi_n \in \Pi_a$ using the past transcript $F_{n-1} = (\pi_1, Y_1, \ldots, \pi_{n-1}, Y_{n-1})$, then receives $Y_n \sim K_\Theta^{\pi_n}$. The class optimal value is $J_{a,\theta}^\star = \sup_{\pi\in\Pi_a} J_\theta(\pi)$, and the dynamic MG over T rounds is $\mathcal{D}_{a,T}(\mathsf{L};\mu_0) = \frac{1}{T}\mathbb{E}_{\Theta\sim\mu_0,\mathsf{L}}\Big[\sum_{n=1}^T \big(J_{a,\Theta}^\star - J_\Theta(\pi_n)\big)\Big]$.

This definition isolates the learning problem that remains after structural limits have already been fixed. We compare the learner with an oracle that knows the environment while still respecting the delegated class Π_a. Dynamic MG is therefore the price of uncertainty inside a restricted architecture. In reinforcement learning terms, it is a Bayes regret problem under partial observability. Intuitively, it is the price of having to infer the task relevant distinctions that the feedback pipeline actually exposes.

4.3 Two Quotients

The environment class is usually much larger than the learning problem. A learner inside architecture a has no use for distinctions between environments that never change any feasible deployment outcome. This leads to two quotients.

The first quotient captures what interaction can reveal. Define the statistical relation

$$\theta \approx_a \theta' \iff K_\theta^\pi = K_{\theta'}^\pi \qquad \text{for every } \pi \in \Pi_a.$$

Let $Z_a^{\text{stat}} = \Theta/{\approx_a}$. Two environments lie in the same statistical cell when every feasible deployment produces the same feedback law under both.

The second quotient captures what regret depends on. Define the joint relation $\theta \sim_a \theta' \iff \theta \approx_a \theta'$ and $\forall \pi \in \Pi_a,\ J_\theta(\pi) = J_{\theta'}(\pi)$. Let $Z_a = \Theta/{\sim_a}$. Thus Z_a refines Z_a^{stat}.

Theorem 3 (Statistical factorization)**.** *For every learner L and every horizon T, the law of the transcript F_T depends on Θ only through Z_a^{stat}. Equivalently, $P(F_T \in \cdot \mid \Theta = \theta) = P(F_T \in \cdot \mid Z_a^{\text{stat}} = [\theta]_{\text{stat}})$. Hence $I(\Theta; F_T) = I(Z_a^{\text{stat}}; F_T) \geq I(Z_a; F_T)$.*

So interaction can only target the statistically visible quotient. This is the object that the feedback channel can reveal. Even so, the learner may still care about a finer partition of environments if value differs inside one statistical cell.

Theorem 4 (Joint quotient reduction)**.** *Let T_a be the quotient map from Θ to Z_a, and let $T_{a\#}\mu_0$ be the pushforward prior on Z_a. Then for every learner L,*

$$\mathcal{D}_{a,T}(\mathsf{L};\mu_0,\Theta) = \mathcal{D}_{a,T}(\mathsf{L};T_{a\#}\mu_0, Z_a).$$

Remark. Theorem 4 is the dynamic analogue of the reduction principle from Section 3. There, the architecture compresses the policy space down to what is feasible. Here, the same architecture compresses the environment space down to what is decision relevant for that feasible class.

4.4 Hidden Value Uncertainty

Once the quotient is defined, the natural state variable for learning is the posterior over joint quotient cells. Let $\nu_n = \mathbb{P}(Z_a \in \cdot \mid F_n)$. For a posterior ν on Z_a, define the Bayes envelope

$$\delta_a(\nu) = \inf_{\pi \in \Pi_a} \int \big(J_a^\star(z) - J_z(\pi)\big)\nu(dz),$$

where $J_z(\pi)$ is well defined because all environments in the same joint cell agree on the values of feasible policies.

Theorem 5 (Bayes envelope)**.** *For every learner* L*,* $\mathcal{D}_{a,T}(\mathsf{L};\mu_0) \geq \frac{1}{T}\sum_{n=1}^{T} \mathbb{E}[\delta_a(\nu_{n-1})]$.

The envelope has the same role here that oracle regret had in Section 3. It is the smallest one round Bayes regret still forced by the learner's current uncertainty. When the posterior spreads mass over joint cells that prescribe different best policies, dynamic MG stays positive.

The two quotient split reveals a second phenomenon. Feedback may separate statistical cells while still leaving value ambiguity inside each such cell. Let g_a be the refinement map from Z_a to Z_a^{stat}. For a statistical cell $z \in Z_a^{\mathrm{stat}}$, let $\Xi_0(\cdot \mid z)$ be the prior conditional law of the joint quotient inside the fiber $g_a^{-1}(z)$. Define $b_a(z) = \inf_{\pi \in \Pi_a} \mathbb{E}_{\Xi_0(\cdot|z)}[r_a(z',\pi)]$, where $r_a(z',\pi) = J_a^\star(z') - J_{z'}(\pi)$ and $z' \in g_a^{-1}(z)$.

Theorem 6 (Hidden value floor)**.** *For every learner* L *and every horizon* T*,*

$$\mathcal{D}_{a,T}(\mathsf{L};\mu_0) \geq \mathbb{E}\big[b_a\big(Z_a^{\mathrm{stat}}\big)\big]\,.$$

Theorem 6 says that interaction can fully exhaust the statistically visible part of the problem while a positive regret floor still remains. The remaining gap comes from value distinctions that the feedback channel never reveals. In alignment language, the learner may infer everything that the judge, reward model, or safety signal is capable of expressing, while the true objective still varies inside that visible cell. Appendix F.8 records a simple collapse criterion. If episode value is itself readable from feedback, then the two quotients coincide and this floor disappears.

4.5 Endogenous Learning Capacity

Section 3 closed with a capacity curve over architecture budgets. The dynamic analogue is a capacity curve over information acquired through interaction. Since the transcript factors through Z_a^{stat}, this is the natural variable to budget. We define the converse frontier $\underline{\mathrm{ELC}}_a(T,B) = \inf_{\mathsf{L},\, I(Z_a^{\mathrm{stat}};F_T)\leq B} \frac{1}{T}\sum_{n=1}^{T} \mathbb{E}[\delta_a(\nu_{n-1})]$, and the attainable frontier $\mathrm{ELC}_a^{\mathrm{ach}}(T,B) = \inf_{\mathsf{L},\, I(Z_a^{\mathrm{stat}};F_T)\leq B} \mathcal{D}_{a,T}(\mathsf{L};\mu_0)$. These two objects play different roles. The converse frontier is dictated by posterior geometry. The attainable frontier is an algorithmic object. It asks for an explicit learner that extracts the right information through the delegated feedback channel while keeping the total information acquisition below B. By Theorem 5, $\mathrm{ELC}_a^{\mathrm{ach}}(T,B) \geq \underline{\mathrm{ELC}}_a(T,B)$. Appendix F.9 gives one generic route toward an upper bound. If a learner can maintain a favorable regret to information tradeoff at each round, then the attainable frontier scales as $\sqrt{B/T}$. This is where design matters, as a lower bound identifies what the learner must know, and an upper bound asks how to uncover exactly that information through the architecture that mediation has already imposed.

5 Mediation Gap Under a Surrogate Model

We now give a closed form slice of the framework, it shows how the structural and dynamic terms become explicit when the mediated decision problem is approximated by a quadratic surrogate. Write cost as $L = -J$. Let X denote the full local episode state and let $Z = \phi(X)$ denote the mediated view. Around the operating regime of interest, take $\widetilde{\ell}_X(u) = \frac{1}{2}u^\top H u + h(X)^\top u, H = H^\top \succ 0$. Sparse coordination is encoded by a subspace $\mathrm{Im}(P_E)$. Define $K_E = P_E(P_E^\top H P_E)^{-1}P_E^\top, m(Z) = \mathbb{E}[h(X) \mid Z]$. The dense, sparse, and partial surrogate optima are $u_d(X) = -H^{-1}h(X), u_s(X) = -K_E h(X), u_p(Z) = -K_E m(Z)$. The surrogate structural MG splits exactly into a coordination term and an information term. Let $r(Z) = h(X) - m(Z)$. Then

$$C = \frac{1}{2}\mathbb{E}\Big[h(X)^\top (H^{-1} - K_E)h(X)\Big]\,,$$

$$I = \frac{1}{2}\mathbb{E}\Big[r(Z)^\top K_E r(Z)\Big]\,.$$

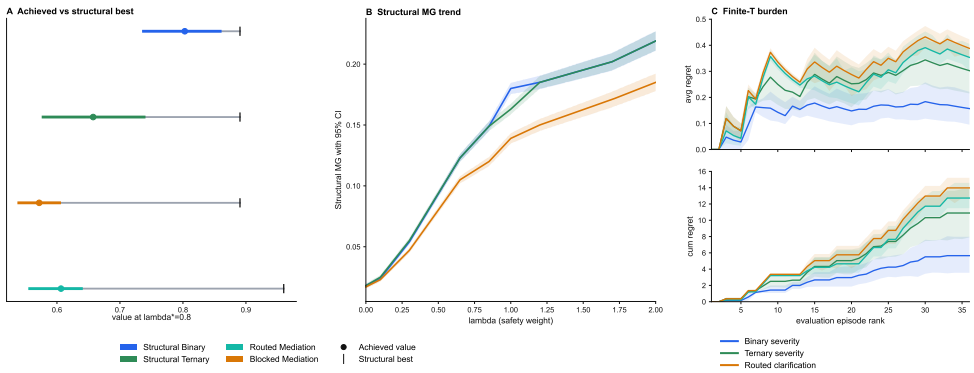


Figure 2: Finite budget behavior. Panel A compares achieved value with structural best. Panel B shows how structural MG changes with the safety weight. Panel C traces the finite budget burden across evaluation episodes.

Thus coordination loss is the price of forcing the control into the sparse subspace. Information loss is the decision weighted conditional variance left after mediation.

The same surrogate also gives a dynamic target. If a learner outputs $\widehat{u}_T(Z) = -K_E \widehat{m}_T(Z)$, define $e_T(Z) = \widehat{m}_T(Z) - m(Z)$. Then

$$\begin{aligned} D_T^{\text{quad}} &= \widetilde{L}(\widehat{u}_T) - \widetilde{L}(u_p) \\ &= \frac{1}{2}\mathbb{E}\Big[e_T(Z)^\top K_E e_T(Z)\Big]. \end{aligned}$$

So dynamic MG is estimation error for the quotient statistic $m(Z)$, or more precisely $P_E^\top m(Z)$.

Let u_J be the true dense optimum and let $\delta(u) = L(u) - \widetilde{L}(u)$. Then

$$\begin{aligned} L(\widehat{u}_T) - L(u_J) = A + C + I + D_T^{\text{quad}} \\ + \Xi_{A\otimes C} + \Xi_{A\otimes I} + \Xi_{A\otimes D}. \end{aligned}$$

Under the local accuracy condition $|\delta(u)| \le \eta$ on the relevant controllers, $L(\widehat{u}_T) - L(u_J) \le C + I + D_T^{\text{quad}} + 8\eta$. The full version of this section is given in Appendix G.

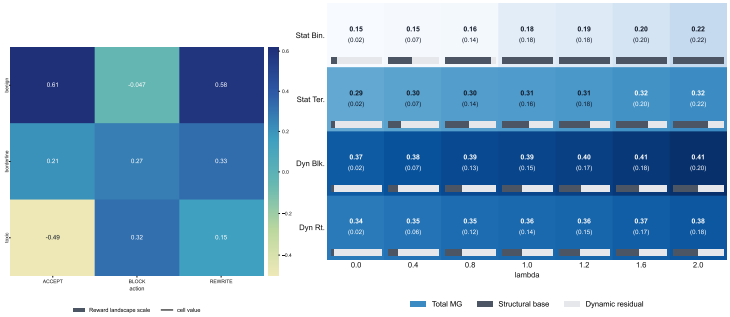
6 Empirical Validation

We test the framework in a controlled mediation intervention task using adversarial benchmarks (Gehman et al., 2020) with Qwen3.6 and Qwen3Guard (Qwen Team, 2026; Zhao et al., 2025). Each episode contains a user prompt, a hidden task mode, an initial safety signal, an optional clarification step, and a final action in $\{\text{ACCEPT}, \text{REWRITE}, \text{BLOCK}\}$. This gives an auditable test of whether failures arise from information removed by the interface or from finite-budget learning inside an interface.

We hold the scalar reward, action set, prompt pool, and evaluation stack fixed, varying only the mediation interface. The four architectures expose increasingly different information patterns: Static Binary receives a coarse safety signal; Static Ternary receives a finer severity signal; Dynamic Blocked can request clarification but cannot use the answer in the final action; and Dynamic Routed can both request clarification and route the answer to the final decision. The oracle observes the latent mode directly. For each architecture, we enumerate the feasible policy tables and precompute a counterfactual outcome tensor over episodes, final actions, and clarification visibility. This keeps response generation, guard outputs, and judge calls fixed across comparisons, so a richer interface cannot win merely by inducing different stochastic outcomes. We estimate structural MG by exhaustive search over the feasible policy class, and dynamic MG by training a finite-budget learner in that class and comparing its achieved value with the architecture's structural best. The experiment therefore isolates the paper's central contribution: richer mediation can reduce the structural gap while increasing the finite-budget residual, a tradeoff hidden by standard end-to-end scores.

Decision landscape. Figure 3a shows why mediation matters. Accept is valuable for benign prompts and costly for toxic prompts. Rewrite is strongest in the ambiguous region. Extra information is useful only when it preserves distinctions that change the best action. This is why the cost of coarsening rises as safety receives more weight.

Results. Figure 2 shows the central empirical pattern. Structural MG grows with the safety weight, which confirms that missing safety relevant distinctions becomes more costly when unsafe mistakes



(a) Reward landscape. (b) MG decomposition map.

Figure 3: Results summary.

are penalized more strongly. Yet richer mediation does not automatically improve finite budget performance. Static Binary is weak but easy to saturate. Dynamic Routed and Dynamic Blocked preserve more capacity, yet the learner leaves much of that capacity unused. This is the empirical reason for separating structural MG from dynamic MG. Figure 3b gives the compressed view. Each cell reports total realized MG against the oracle. The number in parentheses is the structural base. The remaining mass is the dynamic residual. Routed clarification has a smaller structural base than blocked clarification, which shows that evidence must reach the decision point. Yet routed clarification still has a large residual, which shows that preserving useful information is different from learning how to use it. See H for more detailed discussion.

Insights. First, weaker mediation can look better after training because it is easier to saturate. Second, architectural richness and finite budget performance can move in opposite directions. Third, routing is distinct from observation. Clarification has little value if its answer cannot affect the final action. Fourth, information matters only when it protects action changing distinctions. In this setting, the alignment tax is therefore partly structural and partly learnability driven. The structural part comes from the mediated interface. The dynamic part comes from unrealized capacity. In short, the controlled intervention makes the mediation interface identifiable and reveals the core empirical contribution, richer mediation can preserve more structural capacity while leaving more of that capacity unrealized.

7 Conclusion

We introduced the Mediation Gap framework to understand the alignment gap created by mediated architectures. Structural MG measures value lost because an interface removes or misroutes decision relevant information. Dynamic MG measures value left unrealized because learning inside the mediated class is finite. This distinction clarifies the alignment tax in modern systems. Some tax comes from the interface itself, such as reward models, verifiers, safety classifiers, routers, and policy filters. Some comes from the difficulty of learning how to use that interface.

Our theory formalizes architectures through information filtrations, gives structural MG an oracle regret interpretation, and treats dynamic MG as quotient learning inside a fixed architecture. The framework not only provides theoretical insights for LLM safety research, but also applies whenever a principal objective must be pursued through restricted observations, delegated modules, or limited coordination. Our experiment shows the practical relevance of the split. A richer interface can preserve more value in principle yet perform worse after training because its capacity is harder to realize. Alignment design must therefore optimize both mediation structure and learnability.

8 Limitations

First, the theory treats the mediation interface as known and cleanly specified. Real systems often have moving interfaces, learned routers, correlated feedback errors, and model updates that change the effective architecture during training. Second, our experiment is a controlled guardrail setting with finite policy tables and a stylized hidden mode. This is intentional, since it lets us isolate structural and dynamic MG without confounds, but it does not by itself establish the size of MG in frontier deployment pipelines.

References

David Abel, Dilip Arumugam, Lucas Lehnert, and Michael Littman. 2018. State abstractions for lifelong reinforcement learning. In *International conference on machine learning*, pages 10–19. PMLR.

David Abel, David Hershkowitz, and Michael Littman. 2016. Near optimal behavior via approximate state abstraction. In *International Conference on Machine Learning*, pages 2915–2923. PMLR.

Awni Altabaa and Zhuoran Yang. 2024. On the role of information structure in reinforcement learning for partially-observable sequential teams and games. In *Advances in Neural Information Processing Systems 37*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

David Blackwell. 1953. Equivalent comparisons of experiments. *Annals of Mathematical Statistics*, 24(2):265–272.

Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. 2025. A survey on mixture of experts in large language models. *IEEE Transactions on Knowledge and Data Engineering*.

Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pages 4299–4307.

Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Yu Wu, and 1 others. 2024. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*.

William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*.

Evan Frick, Tianle Li, Connor Chen, Wei-Lin Chiang, Anastasios Nikolas Angelopoulos, Jiantao Jiao, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2024. How to evaluate reward models for rlhf. In *The Thirteenth International Conference on Learning Representations*.

Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Christopher Grimm, André Barreto, Greg Farquhar, David Silver, and Satinder Singh. 2021. Proper value equivalence. *Advances in neural information processing systems*, 34:7773–7786.

Christopher Grimm, André Barreto, Satinder Singh, and David Silver. 2020. The value equivalence principle for model-based reinforcement learning. *Advances in neural information processing systems*, 33:5541–5552.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Yu-Chi Ho and K. C. Chu. 1972. Team decision theory and information structures in optimal control problems. *IEEE Transactions on Automatic Control*, 17(1):15–22.

Ronald A Howard. 1966. Information value theory. *IEEE Transactions on systems science and cybernetics*, 2(1):22–26.

Audrey Huang, Wenhao Zhan, Tengyang Xie, Jason D Lee, Wen Sun, Akshay Krishnamurthy, and Dylan J Foster. 2024a. Correcting the mythos of kl-regularization: Direct alignment without overoptimization via chi-squared preference optimization. *arXiv preprint arXiv:2407.13399*.

Haiyang Huang, Newsha Ardalani, Anna Sun, Liu Ke, Hsien-Hsin S Lee, Shruti Bhosale, Carole-Jean Wu, and Benjamin Lee. 2024b. Toward efficient inference for mixture of experts. *Advances in Neural Information Processing Systems*, 37:84033–84059.

Zeyu Huang, Zihan Qiu, Zili Wang, Edoardo M Ponti, and Ivan Titov. 2024c. Post-hoc reward calibration: A case study on length bias. *arXiv preprint arXiv:2409.17407*.

Sham Kakade and John Langford. 2002. Approximately optimal approximate reinforcement learning. In *Proceedings of the nineteenth international conference on machine learning*, pages 267–274.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, Lester James Validad Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, and 1 others. 2025. Rewardbench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1755–1797.

Jacob Marschak and Roy Radner. 1958. Economic theory of teams. chapter 5.

Ashutosh Nayyar, Aditya Mahajan, and Demosthenis Teneketzis. 2013. Decentralized stochastic control with partial history sharing: A common information approach. *IEEE Transactions on Automatic Control*, 58(7):1644–1658.

OpenAI. 2024. Learning to reason with large language models. https://openai.com/index/learning-to-reason-with-llms/.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. 2024. Disentangling length from quality in direct preference optimization. *arXiv preprint arXiv:2403.19159*.

Qwen Team. 2026. Qwen3.6-27B: Flagship-level coding in a 27B dense model.

Roy Radner. 1962. Team decision problems. *Annals of Mathematical Statistics*, 33(3):857–881.

Rafael Rafailov, Yaswanth Chittepu, Ryan Park, Harshit Sikchi, Joey Hejna, W. Bradley Knox, Chelsea Finn, and Scott Niekum. 2024. Scaling laws for reward model overoptimization in direct alignment algorithms. In *Advances in Neural Information Processing Systems 37*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. 2015. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. 2023. A long way to go: Investigating length correlations in rlhf. *arXiv preprint arXiv:2310.03716*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021.

Hans S Witsenhausen. 1968. A counterexample in stochastic optimum control. *SIAM Journal on Control*, 6(1):131–147.

Hongming Zhang, Tongzheng Ren, Chenjun Xiao, Dale Schuurmans, and Bo Dai. 2024. Provable representation with efficient planning for partially observable reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 59759–59782.

Haiquan Zhao, Chenhan Yuan, Fei Huang, Xiaomeng Hu, Yichang Zhang, An Yang, Bowen Yu, Dayiheng Liu, Jingren Zhou, Junyang Lin, and 1 others. 2025. Qwen3guard technical report. *arXiv preprint arXiv:2510.14276*.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Quoc V Le, James Laudon, and 1 others. 2022. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114.

A Notation Guide

We begin by refreshing the main notation used throughout the paper. The structural capacity curve and the finite budget mediation gap are two levels of the same framework. The first is an ex-ante architecture design frontier. The second is the realized gap of one trained mediated system.

Reference benchmark. Let Π^{ref} be the reference policy class and let

$$J^{\mathrm{ref},\star} = \sup_{\pi \in \Pi^{\mathrm{ref}}} J(\pi)$$

be the value of the reference system. The reference class represents the benchmark available before mediation constraints are imposed. In the experiment, the oracle architecture plays this role inside the finite policy abstraction.

Architecture induced policy class. An architecture A induces local information for each component and therefore a feasible policy class Π_A. Its best attainable value is

$$J_A^\star = \sup_{\pi \in \Pi_A} J(\pi).$$

This value assumes perfect optimization inside the mediated class. It measures what the architecture permits in principle.

Finite budget realized value. Let $\bar{J}_T(A)$ denote the value achieved by a learner after T training rounds inside architecture A. Since every deployed policy lies in Π_A, we have

$$\bar{J}_T(A) \le J_A^\star.$$

The gap $J_A^\star - \bar{J}_T(A)$ is the unrealized capacity left by finite training.

Structural capacity frontier. For a resource budget R, let $\mathcal{A}(R)$ be the set of feasible architectures under that budget. The structural mediation gap frontier is

$$\mathrm{MG}_{\mathrm{str}}(R) = J^{\mathrm{ref},\star} - \sup_{a \in \mathcal{A}(R)} J_a^\star.$$

This is an architecture design object. It asks how small the structural gap can be when one may choose the best architecture under budget R. It contains no learning term.

Finite budget mediation gap. For a fixed architecture A, the realized finite budget mediation gap is

$$\mathrm{MG}_T(A \mid \Pi^{\mathrm{ref}}) = J^{\mathrm{ref},\star} - \bar{J}_T(A).$$

It decomposes exactly as

$$\mathrm{MG}_T(A \mid \Pi^{\mathrm{ref}}) = \underbrace{J^{\mathrm{ref},\star} - J_A^\star}_{\text{structural MG}} + \underbrace{J_A^\star - \bar{J}_T(A)}_{\text{dynamic MG}}.$$

The structural term is the loss caused by the mediated interface. The dynamic term is the loss caused by finite training inside that interface.

Relation between the two quantities. If $A \in \mathcal{A}(R)$, then

$$\mathrm{MG}_T(A \mid \Pi^{\mathrm{ref}}) \ge J^{\mathrm{ref},\star} - J_A^\star \ge \mathrm{MG}_{\mathrm{str}}(R).$$

The first inequality follows from $\bar{J}_T(A) \le J_A^\star$. The second follows because $\sup_{a \in \mathcal{A}(R)} J_a^\star \ge J_A^\star$. Hence $\mathrm{MG}_{\mathrm{str}}(R)$ is the best structural lower envelope available under budget R, while $\mathrm{MG}_T(A \mid \Pi^{\mathrm{ref}})$ is what a particular trained system actually pays.

When the quantities coincide. The two quantities coincide in the idealized limit where the selected architecture is structurally optimal under the budget and learning reaches the best value inside that architecture. Formally, if

$$A \in \arg\max_{a \in \mathcal{A}(R)} J_a^\star \qquad \text{and} \qquad \bar{J}_T(A) \to J_A^\star,$$

then

$$\mathrm{MG}_T(A \mid \Pi^{\mathrm{ref}}) \to \mathrm{MG}_{\mathrm{str}}(R).$$

Away from this limit, the realized finite budget gap can be much larger than the structural frontier.

Notation caution. We use A for a single architecture and $\mathcal{A}(R)$ for a family of feasible architectures under resource budget R. If a family $\mathcal{A}$ is optimized jointly, define

$$J_{\mathcal{A}}^\star = \sup_{a \in \mathcal{A}} J_a^\star.$$

If training also chooses the architecture, then $\bar{J}_T(\mathcal{A})$ must specify whether the learner selects policies only, architectures only, or both. In this paper, $\mathrm{MG}_T(A \mid \Pi^{\mathrm{ref}})$ refers to a fixed architecture unless stated otherwise.

Interpretation for the experiment. The experiment estimates $J_a^\star$ by exhaustive search over the finite policy class induced by each mediated architecture. It estimates $\bar{J}_T(a)$ by training a finite budget learner inside that same class. The reported structural MG is the plug in estimate of $J^{\mathrm{ref},\star} - J_a^\star$. The reported dynamic MG is the residual $J_a^\star - \bar{J}_T(a)$. This mirrors the theoretical decomposition while keeping the reference class, architecture class, and learner fixed within each comparison.

B More Related Work

Information structures and team decision theory. MG builds on the classical view that information patterns determine what a distributed decision system can implement. Blackwell compares experiments by their decision value, not by raw entropy (Blackwell, 1953). Marschak and Radner, Radner, and Ho and Chu study team decisions under decentralized information (Marschak and Radner, 1958; Radner, 1962; Ho and Chu, 1972). Witsenhausen's counterexample shows that decentralized information can create irreducible difficulty even in simple quadratic settings (Witsenhausen, 1968). Common information methods later gave a way to reduce some decentralized problems to centralized belief states (Nayyar et al., 2013). Our contribution is complementary. We use filtrations to define the mediated policy class directly, then express loss as a structural MG and a dynamic MG relative to that class.

State abstraction and decision-relevant representation. Approximate state abstraction and value equivalence study which state or model distinctions can be removed while preserving decision quality (Abel et al., 2016, 2018; Grimm et al., 2020, 2021). MG applies the same decision-relevance principle to mediated architectures. The relevant object is not only a compressed state. It is the policy class induced by what each module can observe, remember, and communicate. Section 4 extends this logic to learning by using a quotient of environments. The statistical quotient describes what feedback can reveal. The joint quotient describes what regret still depends on.

Alignment feedback and reward models. RLHF and related preference-learning pipelines train a feedback model, then optimize a policy against that model (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022; Rafailov et al., 2023). Constitutional AI and rule-based or principle-based feedback make the mediation layer more explicit (Bai et al., 2022). Recent work shows that reward models can be brittle, miscalibrated, and vulnerable to overoptimization or spurious features such as response length (Gao et al., 2023; Rafailov et al., 2024; Park et al., 2024; Huang et al., 2024c; Lambert et al., 2025; Frick et al., 2024). MG is orthogonal to reward-model accuracy. Even with a fixed true objective, a mediated interface can erase or misroute distinctions that determine the optimal response.

Partial observability, information complexity, and learning. Bayesian and information-directed views of reinforcement learning connect regret to posterior uncertainty and information gain (Howard, 1966; Kakade and Langford, 2002; Schulman et al., 2015). Recent work studies representation learning for POMDPs and the role of information structure in partially observable teams and games (Altabaa and Yang, 2024; Zhang et al., 2024). Our dynamic MG formalizes a narrower object. The learner does not need to identify the full environment. It needs the quotient induced by the fixed mediation architecture.

Routing and modular systems. Sparse modular computation appears in mixture-of-experts models, routing architectures, and modern reasoning pipelines (Shazeer et al., 2017; Fedus et al., 2021; Zhou et al., 2022; Dai et al., 2024; Huang et al., 2024b; Cai et al., 2025; OpenAI, 2024; Guo et al., 2025). MG gives a decision-theoretic lens for such systems. A router or guardrail is valuable when it routes the distinctions that move the optimal action. More modules, more bits, or more intermediate traces can fail to reduce MG if they do not reach the component that makes the final decision.

C Implications for Modern Systems

Guardrail mediated generation. A guardrail pipeline is a direct instance of mediation. The base model proposes or drafts content. A safety classifier, verifier, or policy filter evaluates risk. An orchestrator decides which signal is visible, whether clarification is requested, and how safety evidence affects the final response. Structural MG measures the value lost when this interface hides or blocks action-changing evidence. Dynamic MG measures how much preserved capacity remains unused after finite training.

Reward modeling. MG suggests that reward models should prioritize decision-relevant distinctions. A feature can explain large variance in human scores and still be low value if it rarely changes the best action. Length bias is one example. It can strongly affect reward-model scores while offering weak guidance near the safety boundary (Singhal et al., 2023; Huang et al., 2024c; Lambert et al., 2025). The relevant question is not whether a signal is predictive in general. The relevant question is whether it separates actions that the deployed system would otherwise confuse.

Constitutional and verifier pipelines. Principles, critiques, and verifiers decompose alignment into intermediate judgments (Bai et al., 2022; OpenAI, 2024; Guo et al., 2025). MG separates two failure modes in such pipelines. A verifier may fail structurally because it lacks the decisive evidence. It may also fail dynamically because the decisive evidence exists but the learner has not discovered how to use it. This distinction matters for debugging. Adding a stronger verifier can reduce structural MG while increasing dynamic MG if the resulting interface becomes harder to learn.

Mixture-of-experts and modular routing. MoE systems route tokens or hidden states to a sparse subset of experts (Shazeer et al., 2017; Fedus et al., 2021; Zhou et al., 2022; Dai et al., 2024). From the MG view, routing quality is not only about load balance or throughput. It is about whether the routed representation carries the distinctions that the selected expert needs. A routing channel that increases capacity but sends evidence to the wrong module can preserve structural value in principle and still create a large dynamic MG at finite training budget.

Evaluation and system design. The framework recommends reporting both structural and dynamic quantities. A system can look worse after training while having a better structural interface, or look better because its weaker interface is easier to saturate. This matters for alignment evaluation. Comparing realized scores alone can penalize architectures that preserve more capacity but need more interaction to exploit it. MG therefore suggests a two-axis evaluation protocol. Measure what the interface permits in principle. Then measure how much of that capacity the learner realizes.

D Proofs for Section 2

D.1 Formal setup

We work on a filtered probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let H_t denote the full history up to time t. For each architecture $\mathcal{A}$, each agent m, and each time t, let

$$\mathcal{I}_t^{m,\mathcal{A}} \subseteq \sigma(H_t)$$

denote the information available to agent m at time t.

The policy class induced by $\mathcal{A}$ is

$$\Pi_{\mathcal{A}} = \Big\{\pi = (\pi_t^m)_{m,t} \;\Big|\; \pi_t^m \text{ is } \mathcal{I}_t^{m,\mathcal{A}}\text{-measurable for every } m, t\Big\}.$$

Fix a reference policy class Π^{ref} and a target objective J. Define

$$J^{\mathrm{ref},\star} = \sup_{\pi \in \Pi^{\mathrm{ref}}} J(\pi), \qquad J_{\mathcal{A}}^{\star} = \sup_{\pi \in \Pi_{\mathcal{A}}} J(\pi).$$

For a learning trajectory $\pi_1, \ldots, \pi_T \in \Pi_{\mathcal{A}}$, define

$$\bar{J}_T(\mathcal{A}) = \frac{1}{T} \sum_{t=1}^{T} J(\pi_t).$$

Then

$$\mathrm{MG}_T(\mathcal{A} \mid \Pi^{\mathrm{ref}}) = J^{\mathrm{ref},\star} - \bar{J}_T(\mathcal{A}) = \big(J^{\mathrm{ref},\star} - J_{\mathcal{A}}^{\star}\big) + \big(J_{\mathcal{A}}^{\star} - \bar{J}_T(\mathcal{A})\big).$$

D.2 General nesting principle

Proposition 3 (Filtration nesting implies policy class nesting)**.** *If*

$$\mathcal{I}_t^{m,\mathcal{A}_2} \subseteq \mathcal{I}_t^{m,\mathcal{A}_1} \qquad \textit{for all } m, t,$$

then

$$\Pi_{\mathcal{A}_2} \subseteq \Pi_{\mathcal{A}_1}.$$

Proof. Take any policy $\pi \in \Pi_{\mathcal{A}_2}$. By definition of $\Pi_{\mathcal{A}_2}$, for every agent m and every time t, the local decision rule π_t^m is $\mathcal{I}_t^{m,\mathcal{A}_2}$-measurable.

Fix any pair (m, t). Since

$$\mathcal{I}_t^{m,\mathcal{A}_2} \subseteq \mathcal{I}_t^{m,\mathcal{A}_1},$$

the σ-algebra $\mathcal{I}_t^{m,\mathcal{A}_2}$ is a sub-σ-algebra of $\mathcal{I}_t^{m,\mathcal{A}_1}$. A standard fact from measure theory states that if a map is measurable with respect to a smaller σ-algebra, then it is also measurable with respect to any larger σ-algebra containing it. Hence π_t^m is also $\mathcal{I}_t^{m,\mathcal{A}_1}$-measurable.

Since this argument holds for every m and every t, it follows that $\pi \in \Pi_{\mathcal{A}_1}$. Therefore

$$\Pi_{\mathcal{A}_2} \subseteq \Pi_{\mathcal{A}_1}.$$

□

D.3 Observation nesting

Proposition 4 (Observation nesting)**.** *Let two observation structures O_1 and O_2 induce local observation histories*

$$Y_{0:t}^{m,(1)}, \qquad Y_{0:t}^{m,(2)}.$$

Assume that for every agent m and time t, there exists a measurable map g_t^m such that

$$Y_{0:t}^{m,(2)} = g_t^m\big(Y_{0:t}^{m,(1)}\big) \qquad \textit{almost surely}.$$

Assume also that the remaining components of the local information sets are identical across the two architectures. More precisely, assume that local action histories, admissible randomness, and communication histories are the same under both architectures. Define

$$\mathcal{I}_t^{m,(i)} = \sigma\Big(Y_{0:t}^{m,(i)}, A_{0:t-1}^m, U^m, M_{0:t}^m\Big),$$

where U^m denotes the allowed randomization source and $M_{0:t}^m$ denotes the full message history available to agent m up to time t, which is assumed identical under the two observation structures.

Then

$$\mathcal{I}_t^{m,(2)} \subseteq \mathcal{I}_t^{m,(1)} \qquad \textit{for all } m, t,$$

and therefore

$$\Pi_{O_2} \subseteq \Pi_{O_1}.$$

Proof. Fix an agent m and a time t. By assumption, there exists a measurable map g_t^m such that

$$Y_{0:t}^{m,(2)} = g_t^m\big(Y_{0:t}^{m,(1)}\big) \qquad \text{a.s.}$$

This implies that the random variable $Y_{0:t}^{m,(2)}$ is measurable with respect to the σ-algebra

$$\sigma\big(Y_{0:t}^{m,(1)}\big).$$

Equivalently,

$$\sigma\big(Y_{0:t}^{m,(2)}\big) \subseteq \sigma\big(Y_{0:t}^{m,(1)}\big).$$

Now enlarge both sides by the same auxiliary variables that are assumed identical under the two observation structures. Since adjoining common random variables preserves inclusion of generated σ-algebras, we obtain

$$\sigma\Big(Y_{0:t}^{m,(2)}, A_{0:t-1}^m, U^m, M_{0:t}^m\Big) \subseteq \sigma\Big(Y_{0:t}^{m,(1)}, A_{0:t-1}^m, U^m, M_{0:t}^m\Big).$$

That is,

$$\mathcal{I}_t^{m,(2)} \subseteq \mathcal{I}_t^{m,(1)}.$$

Since the argument holds for every m and t, Proposition 3 applies and yields

$$\Pi_{O_2} \subseteq \Pi_{O_1}.$$

□

Remark 1 (Stochastic garbling)**.** *A more general statement can be given when $Y^{(2)}$ is obtained from $Y^{(1)}$ by a conditional stochastic kernel. Suppose that, conditional on $Y_{0:t}^{m,(1)}$, the variable $Y_{0:t}^{m,(2)}$ is sampled from a kernel $K_t^m(\cdot \mid Y_{0:t}^{m,(1)})$. If the richer architecture allows enough internal randomization to simulate this kernel, then any policy admissible under O_2 can be implemented under O_1 by first simulating $Y^{(2)}$ from $Y^{(1)}$ and then applying the O_2-policy. In that case the inclusion*

$$\Pi_{O_2} \subseteq \Pi_{O_1}$$

still holds. We omit the full formalization because the measurable-factor case above is sufficient for the main text.

D.4 Coordination nesting

Proposition 5 (Coordination nesting)**.** *Let $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ be two directed communication graphs with*

$$E_2 \subseteq E_1.$$

Assume the following.

1. *The local observation structures are identical under the two graphs.*
2. *For every common edge $e \in E_2$, the message alphabet, delivery timing, and protocol semantics are identical under G_1 and G_2.*
3. *Every extra edge $e \in E_1 \setminus E_2$ admits a distinguished null message $m_\varnothing$.*
4. *Policies under the richer graph G_1 may ignore extra incoming messages and may send $m_\varnothing$ on extra outgoing edges.*

Let $\mathcal{I}_t^{m,G_i}$ denote the filtration generated by local observations, local action histories, admissible randomness, and all messages received by agent m up to time t under graph G_i. Then

$$\mathcal{I}_t^{m,G_2} \subseteq \mathcal{I}_t^{m,G_1} \qquad \text{for all } m, t,$$

and therefore

$$\Pi_{G_2} \subseteq \Pi_{G_1}.$$

Proof. Fix an agent m and a time t. Under graph G_2, the messages available to agent m up to time t are those received over the incoming edges of m that belong to E_2. Under graph G_1, the available messages include those same messages because $E_2 \subseteq E_1$.

Let $M_{0:t}^{m,G_i}$ denote the full message history received by agent m up to time t under graph G_i. By assumptions (1) and (2), the local observations and the common-edge message protocols are identical under the two graphs. Assumption (3) provides a null message on every extra edge, and assumption (4) says that policies under G_1 may both ignore these extra incoming messages and use the null message on extra outgoing edges. Hence any information available under G_2 is reproducible under G_1, while the extra information under G_1 can be discarded.

Formally, the σ-algebra generated by the common observations, action histories, randomization sources, and G_2-message history is contained in the larger σ-algebra generated by the same common variables together with the full G_1-message history:

$$\sigma\Big(Y_{0:t}^m, A_{0:t-1}^m, U^m, M_{0:t}^{m,G_2}\Big) \subseteq \sigma\Big(Y_{0:t}^m, A_{0:t-1}^m, U^m, M_{0:t}^{m,G_1}\Big).$$

Thus

$$\mathcal{I}_t^{m,G_2} \subseteq \mathcal{I}_t^{m,G_1}.$$

Since this holds for every m and t, Proposition 3 implies

$$\Pi_{G_2} \subseteq \Pi_{G_1}.$$

□

D.5 Structural ladders

Corollary 3 (Structural ladders)**.** *Suppose*

$$\mathcal{A}^{(0)} \preceq \mathcal{A}^{(1)} \preceq \cdots \preceq \mathcal{A}^{(L)},$$

which means

$$\mathcal{I}_t^{m,\mathcal{A}^{(\ell-1)}} \subseteq \mathcal{I}_t^{m,\mathcal{A}^{(\ell)}} \qquad \text{for all } m, t,\ \ell = 1, \ldots, L.$$

Then

$$\Pi_{\mathcal{A}^{(0)}} \subseteq \Pi_{\mathcal{A}^{(1)}} \subseteq \cdots \subseteq \Pi_{\mathcal{A}^{(L)}},$$

and therefore

$$J^\star_{\mathcal{A}^{(0)}} \leq J^\star_{\mathcal{A}^{(1)}} \leq \cdots \leq J^\star_{\mathcal{A}^{(L)}}.$$

Moreover, for every $k < L$,

$$J^\star_{\mathcal{A}^{(L)}} - J^\star_{\mathcal{A}^{(k)}} = \sum_{\ell=k+1}^{L} \Big(J^\star_{\mathcal{A}^{(\ell)}} - J^\star_{\mathcal{A}^{(\ell-1)}} \Big).$$

Proof. The inclusion

$$\Pi_{\mathcal{A}^{(\ell-1)}} \subseteq \Pi_{\mathcal{A}^{(\ell)}}$$

for each ℓ follows from Proposition 3. Chaining these inclusions yields the full nesting relation

$$\Pi_{\mathcal{A}^{(0)}} \subseteq \Pi_{\mathcal{A}^{(1)}} \subseteq \cdots \subseteq \Pi_{\mathcal{A}^{(L)}}.$$

Now let $\ell \in \{1, \ldots, L\}$. Since

$$\Pi_{\mathcal{A}^{(\ell-1)}} \subseteq \Pi_{\mathcal{A}^{(\ell)}},$$

taking suprema of the same objective J over nested sets gives

$$J^\star_{\mathcal{A}^{(\ell-1)}} = \sup_{\pi \in \Pi_{\mathcal{A}^{(\ell-1)}}} J(\pi) \leq \sup_{\pi \in \Pi_{\mathcal{A}^{(\ell)}}} J(\pi) = J^\star_{\mathcal{A}^{(\ell)}}.$$

Thus the class optima are monotone along the ladder.

Finally, the telescoping identity is purely algebraic:

$$\sum_{\ell=k+1}^{L} \Big(J^\star_{\mathcal{A}^{(\ell)}} - J^\star_{\mathcal{A}^{(\ell-1)}} \Big) = \big(J^\star_{\mathcal{A}^{(k+1)}} - J^\star_{\mathcal{A}^{(k)}} \big) + \big(J^\star_{\mathcal{A}^{(k+2)}} - J^\star_{\mathcal{A}^{(k+1)}} \big) + \cdots + \big(J^\star_{\mathcal{A}^{(L)}} - J^\star_{\mathcal{A}^{(L-1)}} \big),$$

and all intermediate terms cancel, leaving

$$J^\star_{\mathcal{A}^{(L)}} - J^\star_{\mathcal{A}^{(k)}}.$$

□

E Appendix for Section 3

E.1 Standing assumptions

The following regularity conditions is assumed.

Assumption A1. The history space $\mathcal{H}$, action space $\mathcal{A}$, and all signal spaces are standard Borel spaces.

Assumption A2. The reward function $r : \mathcal{H} \times \mathcal{A} \to [-R_{\max}, R_{\max}]$ is measurable and bounded. The discount factor satisfies $\gamma \in (0, 1)$.

Assumption A3. The reference class is the centralized causal class Π_{cent}, consisting of joint policies with full history access. The corresponding Bellman optimal value functions $V^\star, Q^\star$ exist, are measurable, and satisfy

$$V^\star(h) = \sup_{a \in \mathcal{A}} Q^\star(h, a),$$

$$Q^\star(h, a) = r(h, a) + \gamma \int V^\star(h')\, P(dh' \mid h, a),$$

where $P(\cdot \mid h, a)$ is the history transition kernel induced by the environment.

Assumption A4. The initial history distribution is μ_0. The return of a policy π is

$$J(\pi) = \mathbb{E}_{H_0 \sim \mu_0, \pi}\left[\sum_{t=0}^{\infty} \gamma^t r(H_t, A_t)\right].$$

The reference value is

$$J^\star = \sup_{\pi \in \Pi_{\text{cent}}} J(\pi) = \mathbb{E}_{H_0 \sim \mu_0}[V^\star(H_0)].$$

Assumption A5. In Proposition 1, the correspondence $z \mapsto \mathcal{Q}_\phi(z) \subseteq \Delta(\mathcal{A})$ is nonempty and measurable, so that the conditional Bayes value is a measurable function of z.

Remark. Assumptions A1–A4 are standard in reinforcement learning and infinite horizon discounted control. They ensure that all expectations are finite and that the Bellman objects used in the theorem statements are well defined.

E.2 Proof of Proposition 1

Recall

$$J_\phi^\star = \mathbb{E}\left[\sup_{q \in \mathcal{Q}_\phi(Z)} \sum_a q(a)\, \mathbb{E}[U(a, H) \mid Z]\right].$$

For each realized mediated signal $Z = z$, define the conditional Bayes value

$$B_\phi(z) := \sup_{q \in \mathcal{Q}_\phi(z)} \sum_a q(a)\, \mathbb{E}[U(a, H) \mid Z = z].$$

By Assumption A5, $B_\phi(z)$ is measurable. Then

$$J_\phi^\star = \mathbb{E}[B_\phi(Z)].$$

Since

$$\delta_\phi(z) = \mathbb{E}[V^\star(H) \mid Z = z] - B_\phi(z),$$

the law of total expectation gives

$$\mathbb{E}[\delta_\phi(Z)] = \mathbb{E}[V^\star(H)] - \mathbb{E}[B_\phi(Z)] = \mathbb{E}[V^\star(H)] - J_\phi^\star.$$

This proves Proposition 1. □

Remark. If $\mathcal{Q}_\phi(z) = \Delta(\mathcal{A})$ for every z, then

$$B_\phi(z) = \max_a \mathbb{E}[U(a, H) \mid Z = z],$$

so the toy regret becomes the usual conditional Bayes regret for the compressed signal Z. This is the classical value of information form associated with comparison of experiments and dynamic decision making (Blackwell, 1953; Howard, 1966).

E.3 Proof of Theorem 1

For any delegated policy $\pi \in \Pi_a$, define

$$d^\star(H_t, A_t) = V^\star(H_t) - Q^\star(H_t, A_t).$$

By Assumption A3,

$$Q^\star(H_t, A_t) = r(H_t, A_t) + \gamma\, \mathbb{E}[V^\star(H_{t+1}) \mid H_t, A_t].$$

Hence

$$d^\star(H_t, A_t) = V^\star(H_t) - r(H_t, A_t) - \gamma\, \mathbb{E}[V^\star(H_{t+1}) \mid H_t, A_t].$$

Taking expectation under π,

$$\mathbb{E}_\pi[d^\star(H_t, A_t)] = \mathbb{E}_\pi[V^\star(H_t)] - \mathbb{E}_\pi[r(H_t, A_t)] - \gamma\, \mathbb{E}_\pi[V^\star(H_{t+1})].$$

Multiply by γ^t and sum from $t = 0$ to T:

$$\sum_{t=0}^{T} \gamma^t \mathbb{E}_\pi[d^\star(H_t, A_t)] = \sum_{t=0}^{T} \gamma^t \mathbb{E}_\pi[V^\star(H_t)] - \sum_{t=0}^{T} \gamma^t \mathbb{E}_\pi[r(H_t, A_t)] - \sum_{t=0}^{T} \gamma^{t+1} \mathbb{E}_\pi[V^\star(H_{t+1})].$$

The two value sums telescope:

$$\sum_{t=0}^{T} \gamma^t \mathbb{E}_\pi[V^\star(H_t)] - \sum_{t=0}^{T} \gamma^{t+1} \mathbb{E}_\pi[V^\star(H_{t+1})] = \mathbb{E}_\pi[V^\star(H_0)] - \gamma^{T+1} \mathbb{E}_\pi[V^\star(H_{T+1})].$$

Therefore

$$\sum_{t=0}^{T} \gamma^t \mathbb{E}_\pi[d^\star(H_t, A_t)] = \mathbb{E}_\pi[V^\star(H_0)] - \sum_{t=0}^{T} \gamma^t \mathbb{E}_\pi[r(H_t, A_t)] - \gamma^{T+1} \mathbb{E}_\pi[V^\star(H_{T+1})].$$

By bounded rewards and $\gamma \in (0, 1)$, $V^\star$ is bounded by $R_{\max}/(1-\gamma)$, so the last term vanishes as $T \to \infty$. Using Assumption A4,

$$\mathbb{E}_\pi[V^\star(H_0)] = J^\star,$$

and

$$\sum_{t=0}^{\infty} \gamma^t \mathbb{E}_\pi[r(H_t, A_t)] = J(\pi).$$

Hence

$$J^\star - J(\pi) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t d^\star(H_t, A_t)\right].$$

This proves Theorem 1. □

Remark. Theorem 1 is a history space performance difference identity. The same telescoping argument underlies classical policy improvement and trust region analysis in reinforcement learning (Kakade and Langford, 2002; Schulman et al., 2015).

E.4 Proof of Corollary 1

For every $\pi \in \Pi_a$, Theorem 1 gives

$$J^\star - J(\pi) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t d^\star(H_t, A_t)\right].$$

Taking the infimum over $\pi \in \Pi_a$,

$$\inf_{\pi \in \Pi_a} \big(J^\star - J(\pi)\big) = \inf_{\pi \in \Pi_a} \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t d^\star(H_t, A_t)\right].$$

Since $J^\star$ is constant in π,

$$\inf_{\pi \in \Pi_a} \big(J^\star - J(\pi)\big) = J^\star - \sup_{\pi \in \Pi_a} J(\pi) = J^\star - J_a^\star.$$

This proves Corollary 1. □

E.5 Proof of Theorem 2

By definition,

$$\mathrm{MG}_{\mathrm{str}}(R) = J^\star - \sup_{a \in \mathcal{A}(R)} J_a^\star.$$

For any family of real numbers $\{x_a\}_{a \in \mathcal{A}(R)}$,

$$c - \sup_a x_a = \inf_a (c - x_a).$$

Applying this identity with $c = J^\star$,

$$\mathrm{MG}_{\mathrm{str}}(R) = \inf_{a \in \mathcal{A}(R)} \left(J^\star - J_a^\star\right).$$

Now use Corollary 1 for each architecture a:

$$J^\star - J_a^\star = \inf_{\pi \in \Pi_a} \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t d^\star(H_t, A_t) \right].$$

Substituting gives

$$\mathrm{MG}_{\mathrm{str}}(R) = \inf_{a \in \mathcal{A}(R)} \inf_{\pi \in \Pi_a} \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t d^\star(H_t, A_t) \right].$$

This proves Theorem 2. □

Remark. The term endogenous refers to the feedback loop created by the architecture. The architecture determines the admissible policy class. The chosen policy generates the trajectory. The regret distortion is evaluated on that induced trajectory.

E.6 Proof of Corollary 2

The equivalence of statements (1) and (2) follows directly from Theorem 2. For the stronger existence statement, assume the infimum is attained by some $a^\star \in \mathcal{A}(R)$ and $\pi^\star \in \Pi_{a^\star}$, and that

$$\mathrm{MG}_{\mathrm{str}}(R) = 0.$$

Then

$$0 = \mathbb{E}_{\pi^\star} \left[\sum_{t=0}^{\infty} \gamma^t d^\star(H_t, A_t) \right].$$

Each summand is nonnegative. Therefore

$$d^\star(H_t, A_t) = 0 \qquad \text{almost surely for every } t.$$

Conversely, if such $a^\star, \pi^\star$ exist, then the discounted regret distortion is zero, hence Theorem 2 gives $\mathrm{MG}_{\mathrm{str}}(R) = 0$.

If the infimum is not attained, statement (2) still has the right operational meaning. It says there exists a sequence of feasible architecture policy pairs whose discounted oracle regret converges to zero. □

E.7 Static specialization

The one step toy problem in Section 3.1 is recovered by setting the horizon to one decision and identifying

$$r(H, A) = U(A, H), \qquad \gamma = 0.$$

Then

$$V^\star(H) = \max_a U(a, H), \qquad Q^\star(H, a) = U(a, H),$$

so

$$d^\star(H, a) = V^\star(H) - U(a, H).$$

Theorem 1 becomes

$$\mathbb{E}[V^\star(H)] - \mathbb{E}[U(A, H)] = \mathbb{E}[d^\star(H, A)],$$

and Proposition 1 is exactly the conditional Bayes decomposition of this one step regret under the mediated signal Z.

F Appendix for Section 4

F.1 Formal setup and standing assumptions

Fix an architecture a and its feasible policy class Π_a. We model repeated deployment as an episodic Bayesian decision problem. Each round deploys one policy $\pi \in \Pi_a$, produces one observable feedback variable Y, and is evaluated by its true discounted value $J_\theta(\pi)$.

Let $(\Theta, \mathcal{T})$ be the environment space, $(\Pi_a, \mathcal{P}_a)$ the measurable policy space, and $(\mathcal{Y}, \mathcal{G})$ the measurable feedback space. Let μ_0 be a prior on $(\Theta, \mathcal{T})$. For every $\theta \in \Theta$ and $\pi \in \Pi_a$, let K_θ^π be a probability measure on $(\mathcal{Y}, \mathcal{G})$. We view K_θ^π as the law of the entire feedback summary returned after one deployment episode under policy π in environment θ. Let $J_\theta(\pi) \in \mathbb{R}$ be the corresponding episode value.

A learner L is a sequence of stochastic kernels

$$L_n(d\pi \mid f_{n-1}), \qquad n \geq 1,$$

from past transcripts

$$f_{n-1} = (\pi_1, y_1, \ldots, \pi_{n-1}, y_{n-1}) \in (\Pi_a \times \mathcal{Y})^{n-1}$$

to $(\Pi_a, \mathcal{P}_a)$. Given $\Theta = \theta$ and learner L, the law of the transcript

$$F_T = (\pi_1, Y_1, \ldots, \pi_T, Y_T)$$

is generated recursively by

$$\pi_n \sim L_n(\cdot \mid F_{n-1}), \qquad Y_n \sim K_\theta^{\pi_n}.$$

By the Ionescu–Tulcea theorem, this defines a unique joint law $P_\theta^{\mathsf{L},T}$ on $(\Pi_a \times \mathcal{Y})^T$.

We use the following standing assumptions.

Assumption B1. The measurable spaces $(\Theta, \mathcal{T})$, $(\Pi_a, \mathcal{P}_a)$, and $(\mathcal{Y}, \mathcal{G})$ are standard Borel.

Assumption B2. The map $(\theta, \pi) \mapsto J_\theta(\pi)$ is jointly measurable and bounded. There exists $J_{\max} < \infty$ such that

$$|J_\theta(\pi)| \leq J_{\max} \qquad \text{for all } (\theta, \pi) \in \Theta \times \Pi_a.$$

Assumption B3. The map $(\theta, \pi) \mapsto K_\theta^\pi(B)$ is jointly measurable for every $B \in \mathcal{G}$. Equivalently, $(\theta, \pi) \mapsto K_\theta^\pi$ is a measurable stochastic kernel from $\Theta \times \Pi_a$ to $\mathcal{Y}$.

Assumption B4. Round n feedback depends on the environment and the current deployed policy only. Formally, for every $n \geq 1$ and every $B \in \mathcal{G}$,

$$P(Y_n \in B \mid \Theta, \pi_{1:n}, Y_{1:n-1}) = K_\Theta^{\pi_n}(B) \qquad \text{a.s.}$$

Assumption B5. The statistical quotient Z_a^{stat} and the joint quotient Z_a, both equipped with their quotient σ-fields, are standard Borel. The quotient maps

$$T_a^{\text{stat}} \colon \Theta \to Z_a^{\text{stat}}, \qquad T_a \colon \Theta \to Z_a$$

are measurable.

Assumption B6. The pushforward measures $T^{\mathrm{stat}}_{a\#}\mu_0$ and $T_{a\#}\mu_0$ admit regular conditional distributions along the quotient maps. In particular, there exists a measurable stochastic kernel

$$\Xi_0(dz' \mid z)$$

from Z_a^{stat} to Z_a such that $\Xi_0(\cdot \mid z)$ is supported on the fiber $g_a^{-1}(z)$, where g_a is the refinement map introduced below.

Assumption B7. The joint quotient value envelope

$$J_a^\star(z) = \sup_{\pi \in \Pi_a} J_z(\pi)$$

is measurable and bounded on Z_a.

Assumptions B1–B7 are standard for episodic Bayesian decision problems on measurable spaces. They ensure that transcript laws, quotient posteriors, information quantities, and Bayes envelopes are well defined.

F.2 Auxiliary lemmas

Lemma 1 (Refinement map)**.** *The joint quotient refines the statistical quotient. Equivalently, there exists a measurable map*

$$g_a \colon Z_a \to Z_a^{\mathrm{stat}}$$

such that

$$T_a^{\mathrm{stat}} = g_a \circ T_a.$$

Proof. If $\theta \sim_a \theta'$, then by definition $\theta \approx_a \theta'$. Hence every $\sim_a$-equivalence class is contained in one $\approx_a$-equivalence class. This induces a well defined map from joint classes to statistical classes. Measurability follows from the quotient construction in Assumption B5. □

Lemma 2 (Adaptive transcript law)**.** *Fix $T \geq 1$ and a learner L. Under Assumptions B1–B4, the law of the transcript under environment θ is*

$$P_\theta^{\mathsf{L},T}(df_{1:T}) = \prod_{n=1}^{T} L_n(d\pi_n \mid f_{1:n-1})\, K_\theta^{\pi_n}(dy_n).$$

Consequently, if

$$K_\theta^\pi = K_{\theta'}^\pi \qquad \textit{for every } \pi \in \Pi_a,$$

then

$$P_\theta^{\mathsf{L},T} = P_{\theta'}^{\mathsf{L},T} \qquad \textit{for every learner } \mathsf{L} \textit{ and every } T \geq 1.$$

Proof. The first identity is the recursive construction of the transcript law under Assumption B4. The second claim follows immediately because the stage wise kernels coincide at every round, hence the product construction yields the same joint law. □

F.3 Proof of Proposition 2

Proof. Let $\Theta \in \{\theta^+, \theta^-\}$ with prior μ_0, and assume

$$K_{\theta^+}^\pi = K_{\theta^-}^\pi \qquad \text{for every } \pi \in \Pi_a.$$

By Lemma 2, for every learner L and every n, the law of F_{n-1} is the same under θ^+ and θ^-. Hence the conditional law of the deployed policy π_n given F_{n-1} is also the same in both environments.

Fix n. Using the tower property and the definition of δ_a^{toy},

$$\begin{aligned}
&\mathbb{E}_{\Theta\sim\mu_0,\mathsf{L}}\big[J^\star_{a,\Theta} - J_\Theta(\pi_n)\big] \\
&= \mathbb{E}\Bigg[\sum_{s\in\{+,-\}} \mu_0(\theta^s)\big(J^\star_{a,\theta^s} - J_{\theta^s}(\pi_n)\big)\Bigg] \\
&\geq \inf_{\pi\in\Pi_a} \sum_{s\in\{+,-\}} \mu_0(\theta^s)\big(J^\star_{a,\theta^s} - J_{\theta^s}(\pi)\big) \\
&= \delta_a^{\mathrm{toy}}.
\end{aligned}$$

Averaging over $n = 1, \ldots, T$ yields

$$\mathcal{D}_{a,T}(\mathsf{L};\mu_0) \geq \delta_a^{\mathrm{toy}}.$$

For the positivity claim, suppose no single policy $\pi \in \Pi_a$ attains $J^\star_{a,\theta^+}$ and $J^\star_{a,\theta^-}$ simultaneously. Then for every $\pi \in \Pi_a$,

$$\sum_{s\in\{+,-\}} \mu_0(\theta^s)\big(J^\star_{a,\theta^s} - J_{\theta^s}(\pi)\big) > 0.$$

Taking the infimum over π gives $\delta_a^{\mathrm{toy}} > 0$. □

F.4 Proof of Theorem 3

Proof. Let

$$S = T_a^{\mathrm{stat}}(\Theta) \in Z_a^{\mathrm{stat}}.$$

If two environments θ, θ' satisfy $T_a^{\mathrm{stat}}(\theta) = T_a^{\mathrm{stat}}(\theta')$, then $\theta \approx_a \theta'$. By Lemma 2,

$$P_\theta^{\mathsf{L},T} = P_{\theta'}^{\mathsf{L},T} \qquad \text{for every learner } \mathsf{L} \text{ and every } T.$$

Therefore the transcript law depends on θ only through S. Equivalently,

$$P(F_T \in \cdot \mid \Theta = \theta) = P(F_T \in \cdot \mid S = T_a^{\mathrm{stat}}(\theta)).$$

This gives the conditional independence

$$F_T \perp\!\!\!\perp \Theta \mid S.$$

Hence

$$I(\Theta; F_T \mid S) = 0,$$

and the chain rule yields

$$I(\Theta; F_T) = I(S; F_T).$$

Now let

$$Z = T_a(\Theta) \in Z_a.$$

By Lemma 1, $S = g_a(Z)$. Since the transcript law already depends on Θ only through S, it follows that

$$F_T \perp\!\!\!\perp Z \mid S.$$

Therefore

$$I(Z; F_T) = I(S; F_T) + I(Z; F_T \mid S) = I(S; F_T).$$

Combining the two identities gives

$$I(\Theta; F_T) = I(Z_a^{\mathrm{stat}}; F_T) \geq I(Z_a; F_T).$$

□

F.5 Proof of Theorem 4

Proof. Let $Z = T_a(\Theta)$. Because $\sim_a$ requires both statistical equivalence and value equivalence, the quantities

$$K_z^\pi \quad \text{and} \quad J_z(\pi)$$

are well defined for every $z \in Z_a$ and every $\pi \in \Pi_a$. Let $\bar{\mu}_0 = T_{a\#}\mu_0$ be the pushforward prior on Z_a.

Under the quotient problem, the learner L interacts with latent environment $Z \sim \bar{\mu}_0$, observes feedback generated from K_Z^π, and is scored by $J_Z(\pi)$. Since the original and quotient models induce the same transcript law and the same values on each deployment,

$$J_{a,\Theta}^\star = J_a^\star(T_a(\Theta)), \qquad J_\Theta(\pi_n) = J_{T_a(\Theta)}(\pi_n) \quad \text{a.s.}$$

Therefore

$$\begin{aligned}
\mathcal{D}_{a,T}(\mathsf{L}; \mu_0, \Theta) &= \frac{1}{T}\mathbb{E}\Big[\sum_{n=1}^{T}\big(J_{a,\Theta}^\star - J_\Theta(\pi_n)\big)\Big] \\
&= \frac{1}{T}\mathbb{E}\Big[\sum_{n=1}^{T}\big(J_a^\star(Z) - J_Z(\pi_n)\big)\Big] \\
&= \mathcal{D}_{a,T}(\mathsf{L}; \bar{\mu}_0, Z_a).
\end{aligned}$$

□

F.6 Proof of Theorem 5

Proof. Let

$$Z = T_a(\Theta) \qquad \text{and} \qquad \nu_{n-1} = \mathbb{P}(Z \in \cdot \mid F_{n-1}).$$

Fix n. Since the learner chooses π_n from F_{n-1}, we have

$$\pi_n \perp\!\!\!\perp Z \mid F_{n-1}.$$

Hence

$$\begin{aligned}
&\mathbb{E}\big[J_a^\star(Z) - J_Z(\pi_n) \mid F_{n-1}, \pi_n\big] \\
&= \int \big(J_a^\star(z) - J_z(\pi_n)\big)\nu_{n-1}(dz) \\
&\geq \inf_{\pi \in \Pi_a} \int \big(J_a^\star(z) - J_z(\pi)\big)\nu_{n-1}(dz) \\
&= \delta_a(\nu_{n-1}).
\end{aligned}$$

Taking expectations and summing over n gives

$$\begin{aligned}
\mathcal{D}_{a,T}(\mathsf{L}; \mu_0) &= \frac{1}{T}\sum_{n=1}^{T}\mathbb{E}\big[J_a^\star(Z) - J_Z(\pi_n)\big] \\
&\geq \frac{1}{T}\sum_{n=1}^{T}\mathbb{E}[\delta_a(\nu_{n-1})].
\end{aligned}$$

□

F.7 Proof of Theorem 6

Proof. Let

$$S = T_a^{\text{stat}}(\Theta) \in Z_a^{\text{stat}} \qquad \text{and} \qquad Z = T_a(\Theta) \in Z_a.$$

By Theorem 3, the transcript law depends on Θ only through S. Since Z is a function of Θ and $S = g_a(Z)$, we obtain

$$F_{n-1} \perp\!\!\!\perp Z \mid S \qquad \text{for every } n \geq 1.$$

Therefore Assumption B6 implies

$$\mathbb{P}(Z \in \cdot \mid S, F_{n-1}) = \mathbb{P}(Z \in \cdot \mid S) = \Xi_0(\cdot \mid S) \qquad \text{a.s.}$$

Fix n. Conditioning on S, F_{n-1}, and π_n, we have

$$\begin{aligned}
&\mathbb{E}\big[J_a^\star(Z) - J_Z(\pi_n) \mid S, F_{n-1}, \pi_n\big] \\
&= \int_{g_a^{-1}(S)} \big(J_a^\star(z') - J_{z'}(\pi_n)\big)\Xi_0(dz' \mid S) \\
&\geq \inf_{\pi \in \Pi_a} \int_{g_a^{-1}(S)} \big(J_a^\star(z') - J_{z'}(\pi)\big)\Xi_0(dz' \mid S) \\
&= b_a(S).
\end{aligned}$$

Taking expectations and averaging over n yields

$$\mathcal{D}_{a,T}(\mathsf{L}; \mu_0) \geq \mathbb{E}[b_a(S)] = \mathbb{E}\big[b_a\big(Z_a^{\text{stat}}\big)\big] .$$

□

F.8 A collapse criterion

Proposition 6 (Visible value collapse)**.** *Assume that for every policy $\pi \in \Pi_a$ there exists a bounded measurable function $\rho_\pi \colon \mathcal{Y} \to \mathbb{R}$ such that*

$$J_\theta(\pi) = \int_{\mathcal{Y}} \rho_\pi(y)\, K_\theta^\pi(dy) \qquad \textit{for every } \theta \in \Theta.$$

Then

$$\approx_a = \sim_a .$$

Consequently,

$$Z_a^{\text{stat}} = Z_a$$

up to measurable relabeling, and the hidden value floor in Theorem 6 vanishes.

Proof. If $\theta \approx_a \theta'$, then $K_\theta^\pi = K_{\theta'}^\pi$ for every $\pi \in \Pi_a$. Hence

$$J_\theta(\pi) = \int \rho_\pi(y)\, K_\theta^\pi(dy) = \int \rho_\pi(y)\, K_{\theta'}^\pi(dy) = J_{\theta'}(\pi) \qquad \text{for every } \pi \in \Pi_a.$$

Thus $\theta \sim_a \theta'$. The reverse implication is immediate from the definition of $\sim_a$. Therefore $\approx_a = \sim_a$, which implies equality of the quotient spaces up to measurable relabeling. If the two quotients coincide, every fiber $g_a^{-1}(z)$ is a singleton, so $b_a(z) = 0$ for every z. □

F.9 An achievability route

For the attainable frontier from Section 4.5, a converse alone does not determine the rate. One needs an explicit learner whose round wise regret is controlled by the information it extracts from feedback. The next theorem records one generic route.

Theorem 7 (Achievability under a quotient information ratio)**.** *Let*

$$S = Z_a^{\text{stat}} \qquad \textit{and} \qquad Z = Z_a.$$

Fix a learner $\mathsf{L}^\star$. For each round n, define

$$\nu_{n-1} = \mathbb{P}(Z \in \cdot \mid F_{n-1}),$$

and

$$\Delta_a(\nu_{n-1}, \pi_n) = \int \big(J_a^\star(z) - J_z(\pi_n)\big)\nu_{n-1}(dz).$$

Assume there exists $\Gamma_a < \infty$ such that, for every $n \geq 1$,

$$\Delta_a(\nu_{n-1}, \pi_n)^2 \leq \Gamma_a \, I(S; Y_n \mid F_{n-1}, \pi_n) \qquad a.s.$$

Then

$$\mathcal{D}_{a,T}(\mathsf{L}^\star; \mu_0) \leq \sqrt{\frac{\Gamma_a \, I(S; F_T)}{T}}.$$

In particular, if $I(S; F_T) \leq B$, then

$$\mathcal{D}_{a,T}(\mathsf{L}^\star; \mu_0) \leq \sqrt{\frac{\Gamma_a B}{T}}.$$

Proof. By the same conditioning argument used in the proof of Theorem 5,

$$\mathcal{D}_{a,T}(\mathsf{L}^\star; \mu_0) = \frac{1}{T}\sum_{n=1}^{T} \mathbb{E}[\Delta_a(\nu_{n-1}, \pi_n)] \, .$$

Apply Cauchy–Schwarz,

$$\begin{aligned}\mathcal{D}_{a,T}(\mathsf{L}^\star; \mu_0) &\leq \frac{1}{T}\sqrt{T\sum_{n=1}^{T} \mathbb{E}[\Delta_a(\nu_{n-1}, \pi_n)^2]} \\ &\leq \frac{1}{T}\sqrt{T\Gamma_a \sum_{n=1}^{T} \mathbb{E}[I(S; Y_n \mid F_{n-1}, \pi_n)]}.\end{aligned}$$

Since the learner chooses π_n using F_{n-1} only,

$$I(S; \pi_n \mid F_{n-1}) = 0.$$

Hence the chain rule for mutual information gives

$$I(S; F_T) = \sum_{n=1}^{T} I(S; Y_n \mid F_{n-1}, \pi_n).$$

Substituting into the previous display yields

$$\mathcal{D}_{a,T}(\mathsf{L}^\star; \mu_0) \leq \sqrt{\frac{\Gamma_a \, I(S; F_T)}{T}}.$$

If $I(S; F_T) \leq B$, the stated bound follows immediately. □

G Appendix for Section 5

G.1 Setup and assumptions

We work in cost convention. Higher value J corresponds to lower cost $L = -J$. Let $(\Omega, \mathcal{F}, \mathbb{P})$ carry a full local episode state X and a mediated view $Z = \phi(X)$. Let $h(X) \in \mathbb{R}^d$ be square integrable. A controller is a square integrable measurable map into $\mathbb{R}^d$.

Assumption 1 (Quadratic local surrogate)**.** *There is a local surrogate loss*

$$\widetilde{\ell}_X(u) = \frac{1}{2}u^\top H u + h(X)^\top u,$$

where

$$H = H^\top \succ 0, \qquad \mathbb{E}\|h(X)\|_2^2 < \infty.$$

Sparse coordination is encoded by a full column rank matrix $P_E \in \mathbb{R}^{d\times k}$. The sparse control space is

$$\mathcal{U}_E = \mathrm{Im}(P_E).$$

Since $H \succ 0$ and P_E has full column rank,

$$P_E^\top H P_E \succ 0.$$

Define the expected surrogate cost

$$\widetilde{L}(u) = \mathbb{E}[\widetilde{\ell}_X(u(X))].$$

The three surrogate controller classes are

$$\mathcal{U}_{\text{dense}} = L^2(\sigma(X); \mathbb{R}^d),$$

$$\mathcal{U}_{\text{sparse}} = L^2(\sigma(X); \mathrm{Im}(P_E)),$$

$$\mathcal{U}_{\text{partial}} = L^2(\sigma(Z); \mathrm{Im}(P_E)).$$

Dense means full local control. Sparse means restricted coordination directions. Partial means the same sparse control must be measurable with respect to the mediated view.

Let

$$K_E = P_E(P_E^\top H P_E)^{-1} P_E^\top.$$

The operator K_E is symmetric positive semidefinite. It is not generally the Euclidean projection onto $\mathrm{Im}(P_E)$. The identity used below is

$$K_E H K_E = K_E.$$

Assumption 2 (Uniform local surrogate accuracy)**.** *Let L be the true expected local cost. Let*

$$\delta(u) = L(u) - \widetilde{L}(u).$$

For a set $\mathcal{N}$ containing the controllers used below,

$$|\delta(u)| \le \eta \qquad \textit{for all } u \in \mathcal{N}.$$

The main text uses Assumption 2. We also record a smoother perturbation condition.

Assumption 3 (Smooth local surrogate accuracy)**.** *On a local Hilbert neighborhood $\mathcal{N}$, the true cost L is Fréchet differentiable, μ-strongly convex, and β-smooth. Moreover,*

$$\|\nabla L(u) - \nabla \widetilde{L}(u)\|_{L^2} \le \varepsilon \qquad \textit{for all } u \in \mathcal{N}.$$

G.2 Closed form surrogate optimizers

Lemma 3 (Sparse quadratic minimizer)**.** *For fixed $h \in \mathbb{R}^d$, the minimizer of*

$$u \mapsto \frac{1}{2}u^\top Hu + h^\top u$$

over $u \in \mathbb{R}^d$ is $-H^{-1}h$. The minimizer over $u \in \mathrm{Im}(P_E)$ is $-K_E h$.

Proof. The full minimizer satisfies

$$Hu + h = 0,$$

hence $u = -H^{-1}h$.

For the sparse problem, write $u = P_E v$. The restricted loss is

$$\frac{1}{2}v^\top (P_E^\top H P_E)v + (P_E^\top h)^\top v.$$

The first order condition is

$$(P_E^\top H P_E)v + P_E^\top h = 0.$$

Thus

$$v = -(P_E^\top H P_E)^{-1} P_E^\top h,$$

and

$$u = P_E v = -P_E (P_E^\top H P_E)^{-1} P_E^\top h = -K_E h.$$

□

Let

$$m(Z) = \mathbb{E}[h(X) \mid Z].$$

Lemma 4 (Dense, sparse, and partial optima)**.** *Under Assumption 1, the surrogate optima over the three classes are*

$$u_d(X) = -H^{-1}h(X),$$

$$u_s(X) = -K_E h(X),$$

$$u_p(Z) = -K_E m(Z).$$

Proof. The dense and sparse formulas follow pointwise from Lemma 3. For the partial class, fix Z. The conditional surrogate risk for a Z-measurable sparse action $u \in \mathrm{Im}(P_E)$ is

$$\mathbb{E}[\widetilde{\ell}_X(u) \mid Z] = \frac{1}{2}u^\top Hu + \mathbb{E}[h(X) \mid Z]^\top u = \frac{1}{2}u^\top Hu + m(Z)^\top u.$$

Applying Lemma 3 with $h = m(Z)$ yields

$$u_p(Z) = -K_E m(Z).$$

□

G.3 Structural terms

Proposition 7 (Coordination term)**.** *Define*

$$C = \mathbb{E}[\widetilde{\ell}_X(u_s(X)) - \widetilde{\ell}_X(u_d(X))].$$

Then

$$C = \frac{1}{2}\mathbb{E}\Big[h(X)^\top (H^{-1} - K_E)h(X)\Big].$$

In particular,

$$C \le \frac{1}{2}\|H^{-1} - K_E\|_{\mathrm{op}}\, \mathbb{E}\|h(X)\|_2^2.$$

Proof. For fixed X, $u_d(X)$ is the full minimizer. Hence

$$\widetilde{\ell}_X(u_s) - \widetilde{\ell}_X(u_d) = \frac{1}{2}(u_s - u_d)^\top H(u_s - u_d).$$

Equivalently, using minimized values,

$$\widetilde{\ell}_X(u_d) = -\frac{1}{2}h(X)^\top H^{-1}h(X),$$

and

$$\widetilde{\ell}_X(u_s) = -\frac{1}{2}h(X)^\top K_E h(X).$$

Subtracting gives

$$\widetilde{\ell}_X(u_s) - \widetilde{\ell}_X(u_d) = \frac{1}{2}h(X)^\top (H^{-1} - K_E)h(X).$$

Taking expectations proves the identity. The operator norm bound follows from

$$h^\top (H^{-1} - K_E)h \le \|H^{-1} - K_E\|_{\mathrm{op}}\|h\|_2^2.$$

□

Proposition 8 (Information term)**.** *Define*

$$I = \mathbb{E}[\widetilde{\ell}_X(u_p(Z)) - \widetilde{\ell}_X(u_s(X))].$$

Then

$$I = \frac{1}{2}\mathbb{E}\Big[(h(X) - m(Z))^\top K_E(h(X) - m(Z))\Big].$$

Equivalently,

$$I = \frac{1}{2}\operatorname{tr}(K_E\, \mathbb{E}[\mathrm{Cov}(h(X) \mid Z)]).$$

Proof. Both $u_s(X)$ and $u_p(Z)$ lie in $\mathrm{Im}(P_E)$. Since $u_s(X)$ is the sparse full information minimizer,

$$\widetilde{\ell}_X(u_p) - \widetilde{\ell}_X(u_s) = \frac{1}{2}(u_p - u_s)^\top H(u_p - u_s).$$

By Lemma 4,

$$u_p - u_s = -K_E m(Z) + K_E h(X) = K_E(h(X) - m(Z)).$$

Using $K_E H K_E = K_E$,

$$\widetilde{\ell}_X(u_p) - \widetilde{\ell}_X(u_s) = \frac{1}{2}(h(X) - m(Z))^\top K_E(h(X) - m(Z)).$$

Taking expectations proves the first identity.

For the covariance form, condition on Z. Since $m(Z) = \mathbb{E}[h(X) \mid Z]$,

$$\mathbb{E}[(h - m)^\top K_E(h - m) \mid Z] = \operatorname{tr}(K_E \operatorname{Cov}(h(X) \mid Z)).$$

Taking expectations over Z gives the result. □

Corollary 4 (Surrogate structural MG)**.** *Under Assumption 1,*

$$\widetilde{L}(u_p) - \widetilde{L}(u_d) = C + I.$$

Proof. By definition,

$$\widetilde{L}(u_p) - \widetilde{L}(u_d) = [\widetilde{L}(u_s) - \widetilde{L}(u_d)] + [\widetilde{L}(u_p) - \widetilde{L}(u_s)].$$

The two terms are C and I. □

G.4 Dynamic surrogate residual

Let a learner produce a square integrable Z-measurable estimate $\widehat{m}_T(Z)$. Define

$$\widehat{u}_T(Z) = -K_E \widehat{m}_T(Z).$$

Proposition 9 (Quadratic dynamic residual)**.** *Conditional on the learner output $\widehat{m}_T$,*

$$D_T^{\text{quad}} := \widetilde{L}(\widehat{u}_T) - \widetilde{L}(u_p) = \frac{1}{2}\mathbb{E}\Big[(\widehat{m}_T(Z) - m(Z))^\top K_E(\widehat{m}_T(Z) - m(Z))\Big].$$

Proof. Fix Z. The conditional sparse partial risk is

$$q_Z(u) = \frac{1}{2}u^\top H u + m(Z)^\top u, \qquad u \in \mathrm{Im}(P_E).$$

Its minimizer is $u_p(Z)$. Thus

$$q_Z(\widehat{u}_T) - q_Z(u_p) = \frac{1}{2}(\widehat{u}_T - u_p)^\top H(\widehat{u}_T - u_p).$$

Since

$$\widehat{u}_T - u_p = -K_E\widehat{m}_T + K_E m = -K_E(\widehat{m}_T - m),$$

we obtain

$$q_Z(\widehat{u}_T) - q_Z(u_p) = \frac{1}{2}(\widehat{m}_T - m)^\top K_E H K_E(\widehat{m}_T - m).$$

Using $K_E H K_E = K_E$ and then averaging over Z proves the identity. □

Corollary 5 (Decision relevant quotient coordinate)**.** *Let*

$$s_E(Z) = P_E^\top m(Z), \qquad \widehat{s}_{E,T}(Z) = P_E^\top \widehat{m}_T(Z),$$

and let

$$H_E = P_E^\top H P_E.$$

Then

$$D_T^{\text{quad}} = \frac{1}{2}\mathbb{E}\Big[(\widehat{s}_{E,T}(Z) - s_E(Z))^\top H_E^{-1}(\widehat{s}_{E,T}(Z) - s_E(Z))\Big].$$

Hence the dynamic learning target under the surrogate is $s_E(Z)$, not the full state X or the full vector $h(X)$.

Proof. By definition,

$$K_E = P_E H_E^{-1} P_E^\top.$$

Substitute this into Proposition 9:

$$(\widehat{m}_T - m)^\top K_E(\widehat{m}_T - m) = (P_E^\top(\widehat{m}_T - m))^\top H_E^{-1}(P_E^\top(\widehat{m}_T - m)).$$

This is the claimed expression. □

Corollary 6 (Quadratic Bayes envelope)**.** *Suppose the learning transcript F_T induces a posterior over the function s_E. The Bayes optimal predictor under the quadratic dynamic residual is*

$$\widehat{s}^{\star}_{E,T}(Z) = \mathbb{E}[s_E(Z) \mid F_T, Z],$$

and the posterior Bayes envelope is

$$\delta_a^{\mathrm{quad}}(F_T) = \frac{1}{2}\mathbb{E}_Z\big[\mathrm{tr}\big(H_E^{-1}\,\mathrm{Cov}(s_E(Z) \mid F_T, Z)\big)\big]\,.$$

Proof. For each $Z = z$, Proposition 9 and Corollary 5 reduce the posterior risk to a weighted squared error in $s_E(z)$:

$$\frac{1}{2}\mathbb{E}[(\widehat{s}(z) - s_E(z))^\top H_E^{-1}(\widehat{s}(z) - s_E(z)) \mid F_T, Z = z].$$

The minimizer is the posterior mean

$$\widehat{s}(z) = \mathbb{E}[s_E(z) \mid F_T, Z = z].$$

The minimized risk is the weighted posterior covariance trace. Averaging over Z proves the result. □

G.5 True cost decomposition

Let

$$u_J \in \arg\min_{u \in \mathcal{U}_{\mathrm{dense}}} L(u)$$

be the true dense optimum. Let

$$A = L(u_d) - L(u_J).$$

Define

$$\Xi_{A\otimes C} = \delta(u_s) - \delta(u_d),$$
$$\Xi_{A\otimes I} = \delta(u_p) - \delta(u_s),$$
$$\Xi_{A\otimes D} = \delta(\widehat{u}_T) - \delta(u_p).$$

Theorem 8 (Local true finite budget MG decomposition)**.** *Under Assumption 1,*

$$L(\widehat{u}_T) - L(u_J) = A + C + I + D_T^{\mathrm{quad}} + \Xi_{A\otimes C} + \Xi_{A\otimes I} + \Xi_{A\otimes D}.$$

Proof. Use $L = \widetilde{L} + \delta$. Insert and subtract

$$L(u_d), \quad \widetilde{L}(u_d), \quad \widetilde{L}(u_s), \quad \widetilde{L}(u_p), \quad \widetilde{L}(\widehat{u}_T).$$

Then

$$\begin{aligned} L(\widehat{u}_T) - L(u_J) = {} & [L(u_d) - L(u_J)] \\ & + [\widetilde{L}(u_s) - \widetilde{L}(u_d)] \\ & + [\widetilde{L}(u_p) - \widetilde{L}(u_s)] \\ & + [\widetilde{L}(\widehat{u}_T) - \widetilde{L}(u_p)] \\ & + [\delta(u_s) - \delta(u_d)] \\ & + [\delta(u_p) - \delta(u_s)] \\ & + [\delta(\widehat{u}_T) - \delta(u_p)]. \end{aligned}$$

The seven bracketed terms are, by definition,

$$A, \quad C, \quad I, \quad D_T^{\mathrm{quad}}, \quad \Xi_{A\otimes C}, \quad \Xi_{A\otimes I}, \quad \Xi_{A\otimes D}.$$

□

Corollary 7 (Uniform accuracy upper bound)**.** *Assume Assumptions 1 and 2, and assume*

$$u_J, u_d, u_s, u_p, \widehat{u}_T \in \mathcal{N}.$$

Then

$$0 \le A \le 2\eta,$$

$$|\Xi_{A\otimes C}| \le 2\eta, \qquad |\Xi_{A\otimes I}| \le 2\eta, \qquad |\Xi_{A\otimes D}| \le 2\eta.$$

Consequently,

$$L(\widehat{u}_T) - L(u_J) \le C + I + D_T^{\mathrm{quad}} + 8\eta.$$

For the structural case $\widehat{u}_T = u_p$,

$$L(u_p) - L(u_J) \le C + I + 6\eta.$$

Proof. Since u_J minimizes L, $A \ge 0$. Since u_d minimizes $\widetilde{L}$ over the dense class,

$$\widetilde{L}(u_d) - \widetilde{L}(u_J) \le 0.$$

Thus

$$A = L(u_d) - L(u_J) = \widetilde{L}(u_d) - \widetilde{L}(u_J) + \delta(u_d) - \delta(u_J) \le \delta(u_d) - \delta(u_J) \le 2\eta.$$

The three coupling bounds follow directly from

$$|\delta(u) - \delta(v)| \le |\delta(u)| + |\delta(v)| \le 2\eta.$$

Substituting these bounds into Theorem 8 gives

$$L(\widehat{u}_T) - L(u_J) \le C + I + D_T^{\mathrm{quad}} + 8\eta.$$

If $\widehat{u}_T = u_p$, then $D_T^{\mathrm{quad}} = 0$ and $\Xi_{A\otimes D} = 0$, giving the structural bound. □

G.6 Smooth perturbation and Lipschitz coupling

Proposition 10 (Smooth surrogate perturbation)**.** *Under Assumption 3, if $u_J, u_d \in \mathcal{N}$, then*

$$\|u_d - u_J\|_{L^2} \le \frac{\varepsilon}{\mu},$$

and

$$A = L(u_d) - L(u_J) \le \frac{\beta\varepsilon^2}{2\mu^2}.$$

Proof. The first order conditions give

$$\nabla L(u_J) = 0, \qquad \nabla \widetilde{L}(u_d) = 0.$$

Therefore

$$\|\nabla L(u_d) - \nabla L(u_J)\|_{L^2} = \|\nabla L(u_d)\|_{L^2} = \|\nabla L(u_d) - \nabla \widetilde{L}(u_d)\|_{L^2} \le \varepsilon.$$

Strong convexity implies strong monotonicity of the gradient:

$$\|\nabla L(u_d) - \nabla L(u_J)\|_{L^2} \ge \mu \|u_d - u_J\|_{L^2}.$$

Thus

$$\|u_d - u_J\|_{L^2} \le \varepsilon/\mu.$$

By β-smoothness and $\nabla L(u_J) = 0$,

$$L(u_d) - L(u_J) \le \frac{\beta}{2}\|u_d - u_J\|_{L^2}^2 \le \frac{\beta\varepsilon^2}{2\mu^2}.$$

□

Proposition 11 (Lipschitz coupling bound)**.** *Assume $\delta = L - \widetilde{L}$ is L_δ-Lipschitz on $\mathcal{N}$ in L^2:*

$$|\delta(u) - \delta(v)| \le L_\delta \|u - v\|_{L^2}.$$

If $u_d, u_s, u_p, \widehat{u}_T \in \mathcal{N}$, then

$$|\Xi_{A\otimes C}| \le L_\delta \sqrt{\frac{2C}{\lambda_{\min}(H)}},$$

$$|\Xi_{A\otimes I}| \le L_\delta \sqrt{\frac{2I}{\lambda_{\min}(H)}},$$

and

$$|\Xi_{A\otimes D}| \le L_\delta \sqrt{\frac{2D_T^{\text{quad}}}{\lambda_{\min}(H)}}.$$

Consequently,

$$\begin{aligned} &|\Xi_{A\otimes C}| + |\Xi_{A\otimes I}| + |\Xi_{A\otimes D}| \\ &\le L_\delta \left(\sqrt{\frac{2C}{\lambda_{\min}(H)}} + \sqrt{\frac{2I}{\lambda_{\min}(H)}} + \sqrt{\frac{2D_T^{\text{quad}}}{\lambda_{\min}(H)}} \right). \end{aligned}$$

Proof. By Lipschitz continuity,

$$|\Xi_{A\otimes C}| = |\delta(u_s) - \delta(u_d)| \le L_\delta \|u_s - u_d\|_{L^2}.$$

From the quadratic expansion,

$$C = \frac{1}{2}\mathbb{E}[(u_s - u_d)^\top H (u_s - u_d)] \ge \frac{\lambda_{\min}(H)}{2} \|u_s - u_d\|_{L^2}^2.$$

Hence

$$\|u_s - u_d\|_{L^2} \le \sqrt{\frac{2C}{\lambda_{\min}(H)}}.$$

The proof for $\Xi_{A\otimes I}$ is identical using

$$I = \frac{1}{2}\mathbb{E}[(u_p - u_s)^\top H (u_p - u_s)].$$

The proof for $\Xi_{A\otimes D}$ is identical using

$$D_T^{\text{quad}} = \frac{1}{2}\mathbb{E}[(\widehat{u}_T - u_p)^\top H (\widehat{u}_T - u_p)].$$

Adding the three inequalities gives the final display. □

H Appendix for the Experiment

H.1 Purpose and design principle

The experiment is a finite, auditable implementation of the running example. We use it to test whether a guardrail mediated generation system exhibits the same decomposition as the theory. Each episode contains a user prompt, a hidden task mode, an initial mediated safety signal, an optional clarification step, and a final action. The prompt pool, action set, generation model, guard model, judge model, and scalar reward are fixed across all architectures. Only the mediation interface changes.

The central empirical object is the finite budget mediation gap. For each architecture a, we separate what the interface permits from what training realizes. The structural slice estimates the best feasible value inside the architecture. The dynamic slice estimates the finite budget residual between that structural best value and the value achieved by the learner. This mirrors the theoretical split between structural MG and dynamic MG.

The experiment uses Qwen 3.6 27b and Qwen 3 Guard on one H100 GPU.

H.2 Episodes and reward

We sample prompts from RealToxicityPrompts and stratify the sample into benign, borderline, and toxic prompt buckets. Each prompt is converted into an episode

$$e_i = (x_i, \theta_i, g_i, c_i),$$

where x_i is the prompt, $\theta_i \in \{\text{BENIGN}, \text{UNSAFE}\}$ is a hidden task mode, g_i is the initial guard signal, and c_i is the clarification text. The hidden mode is sampled from a bucket dependent distribution. This gives a controlled latent variable that can matter for the optimal response while remaining hidden from weaker architectures.

The final action space is

$$\mathcal{A}_{\text{final}} = \{\text{ACCEPT}, \text{REWRITE}, \text{BLOCK}\}.$$

For each final action, the policy model generates a response. A guard model assigns a risk score, and a separate judge assigns an alignment score. The scalar reward at safety weight λ is

$$R_\lambda(e_i, a) = \text{Align}(e_i, a) - \lambda\, \text{Risk}(e_i, a).$$

The main cross-section uses $\lambda^\star = 0.8$. The lambda sweep varies λ to trace how the value of safety relevant mediation changes as unsafe responses become more costly.

H.3 Architectures

We compare four delegated architectures and one oracle reference. The architectures differ only in the mediation interface.

Table 1: Architectures used in the experiment. The environment, reward, action set, and evaluation stack are fixed.

Architecture	First signal	Can clarify	Routes clarification	Latent access
Static Binary	binary severity	no	no	no
Static Ternary	ternary severity	no	no	no
Blocked Mediation	ternary severity	yes	no	no
Routed Mediation	ternary severity	yes	yes	no
Oracle	latent mode	yes	yes	yes

Static Binary and Static Ternary implement observation coarsening. Blocked Mediation and Routed Mediation isolate routing. Both can ask for clarification. Only Routed Mediation allows the answer to influence the final action. This comparison is important because it separates evidence availability from evidence usability. A signal that exists somewhere in the system has little value if it cannot reach the decision point.

H.4 Policy classes

For each architecture a, we define a finite policy class Π_a. A policy table specifies a first action as a function of the architecture visible first signal. If clarification is available and requested, the table also specifies a final action as a function of the visible clarification value. Thus each Π_a is an explicit finite analogue of an architecture induced policy class.

This design lets us estimate structural optima by exhaustive search. It also lets us compare architectures without changing the underlying task. The policy class expands as the mediation interface becomes richer. Static Binary has the smallest class. Routed Mediation has a larger class because it can condition on routed clarification. The oracle class can condition directly on the latent mode.

H.5 Counterfactual outcome tensor

Before computing structural or dynamic quantities, we build a counterfactual outcome tensor. For every episode, final action, and visibility mode, we evaluate the generated response, guard score, judge score, and reward. The tensor has entries of the form

$$(e_i, v, a) \mapsto \big(\mathrm{Align}(e_i, v, a), \mathrm{Risk}(e_i, v, a), R_\lambda(e_i, v, a)\big),$$

where $v \in \{\mathrm{none}, \mathrm{clar}\}$ records whether the clarification is visible to the generation step.

The tensor is the key control device in the experiment. Once it is built, every architecture is evaluated using the same counterfactual basis. Architectural differences therefore come from the decision path and policy constraints, not from inconsistent response generation or judge calls. This gives a clean empirical analogue of comparing $J_a^\star$ across fixed architecture classes.

H.6 Structural estimator

Let $\mathcal{E}_{\mathrm{eval}}$ be the evaluation episode set. For each architecture a, we estimate its structural best value by

$$\widehat{J}_a^\star(\lambda) = \max_{\pi \in \Pi_a} \frac{1}{|\mathcal{E}_{\mathrm{eval}}|} \sum_{e_i \in \mathcal{E}_{\mathrm{eval}}} R_\lambda(e_i, \pi).$$

The oracle reference is computed in the same discrete abstraction, using direct access to the latent mode. The structural mediation gap is

$$\widehat{\mathrm{MG}}_{\mathrm{str}}(a, \lambda) = \widehat{J}_{\mathrm{oracle}}^\star(\lambda) - \widehat{J}_a^\star(\lambda).$$

This quantity measures what the architecture loses in principle, before finite training enters.

H.7 Dynamic learner and realized value

For each architecture, we train a finite budget learner over the corresponding policy class. The learner observes rewards from training rollouts and forms a posterior score for each policy. After the training budget is exhausted, it deploys the policy with the largest posterior mean on the evaluation set. We write the resulting achieved value as

$$\widehat{\bar{J}}_T(a, \lambda) = \frac{1}{|\mathcal{E}_{\mathrm{eval}}|} \sum_{e_i \in \mathcal{E}_{\mathrm{eval}}} R_\lambda(e_i, \widehat{\pi}_{a,T}).$$

The realized total mediation gap is

$$\widehat{\mathrm{MG}}_{\mathrm{tot}}(a, \lambda) = \widehat{J}_{\mathrm{oracle}}^\star(\lambda) - \widehat{\bar{J}}_T(a, \lambda).$$

The finite budget dynamic residual is

$$\widehat{\mathrm{MG}}_{\mathrm{dyn}}(a, \lambda) = \widehat{\mathrm{MG}}_{\mathrm{tot}}(a, \lambda) - \widehat{\mathrm{MG}}_{\mathrm{str}}(a, \lambda).$$

This residual measures unrealized capacity inside the architecture. In the main text we use it as the empirical dynamic MG at the given training budget.

H.8 Lambda sweep

For each policy, the reward as a function of λ is an affine line

$$\bar{A}_\pi - \lambda \bar{R}_\pi,$$

where $\bar{A}_\pi$ is mean alignment and $\bar{R}_\pi$ is mean risk. The structural value of an architecture is therefore an upper envelope of policy lines:

$$\widehat{J}_a^\star(\lambda) = \max_{\pi \in \Pi_a} \left(\bar{A}_\pi - \lambda \bar{R}_\pi\right).$$

This representation lets us trace policy switches and structural MG as λ varies. It also explains why structural MG increases with the safety weight. When unsafe mistakes become more expensive, erased safety relevant distinctions carry a larger value penalty.

Table 2: Cross-section of the mediation gap decomposition at $\lambda^\star = 0.8$. Structural MG is the plug-in architecture optimum gap shown in Fig. 3b. Realized Total MG is estimated from seed-bootstrap intervals with $B = 5000$. Dynamic MG is Realized Total MG minus the plug-in structural base, so its interval reflects uncertainty in realized performance. The paradox indicator is positive when an architecture has larger Structural MG than Routed Mediation but lower Realized Total MG at the same finite training budget.

(a) Static architectures.

Architecture	Structural MG	Realized Total MG [95% CI]	Dynamic MG [95% CI]	Paradox vs Routed
Static Binary	0.140	0.157 [0.099, 0.225]	0.017 [-0.041, 0.085]	Yes
Static Ternary	0.140	0.303 [0.219, 0.384]	0.163 [0.079, 0.244]	Yes

(b) Dynamic architectures.

Architecture	Structural MG	Realized Total MG [95% CI]	Dynamic MG [95% CI]	Paradox vs Routed
Blocked Mediation	0.130	0.388 [0.354, 0.423]	0.258 [0.224, 0.293]	No
Routed Mediation	0.120	0.354 [0.319, 0.405]	0.234 [0.199, 0.285]	–

H.9 Cross-section at $\lambda^\star = 0.8$

Table 2 gives the numerical cross-section behind the decomposition map in Fig. 3b. Structural MG is the plug-in structural base. Realized Total MG is estimated from seed-bootstrap intervals. Dynamic MG is the realized total gap minus the plug-in structural base. Its interval therefore reflects uncertainty in realized performance, while structural estimation uncertainty is held fixed.

The table gives a compact view of the main phenomenon. Static Binary has a larger structural base than Routed Mediation, yet much lower realized total MG. It is structurally weaker, but easier for the learner to saturate. Routed Mediation has the smallest structural base among the delegated systems, which confirms the value of routed clarification. Its dynamic residual is large, which means the preserved capacity is difficult to realize at the present training budget.

H.10 Panel interpretation

Fig. 3a validates the decision landscape. The best action changes across prompt regimes, so mediation is consequential. Accepting is valuable for benign prompts and risky for toxic prompts. Rewriting is strongest in the ambiguous middle. Thus an information channel matters only when it preserves distinctions that change the optimal action.

Panel A in Fig. 2 compares achieved value with the structural best at $\lambda^\star = 0.8$. It shows that the finite budget learner nearly saturates Static Binary, while leaving substantial capacity unrealized in Routed and Blocked Mediation. This is the clearest visual evidence for the dynamic residual.

Panel B in Fig. 2 shows the structural MG trend across λ. Structural MG rises as safety weight increases. This matches the value-weighted theory. The cost of a missing distinction depends on how much that distinction affects the current objective.

Panel C in Fig. 2 traces the finite budget burden across evaluation episodes. The richer dynamic architectures carry a larger residual under the current learner. Routed Mediation improves on Blocked Mediation, which confirms that routing evidence to the decision point matters.

Fig. 3b provides the decomposition map. The large number in each cell is realized total MG against the oracle. The number in parentheses is the structural base. The difference is the dynamic residual. This panel is the main empirical summary of the paper. It shows that finite-budget performance can reverse structural ordering.

H.11 Interpretive insights

Weak mediation can look better after training. Static Binary has the smallest realized total MG at $\lambda^\star = 0.8$, even though it has a larger structural base than Routed Mediation. This is the finite budget

paradox. A weak interface can be easier to saturate, while a richer interface can preserve capacity that the learner does not discover.

Structural richness and realized performance can diverge. The theory predicts that total finite budget MG has both structural and dynamic components. The results show that these components can move in opposite directions. Routed Mediation lowers the structural base, while its dynamic residual remains large. This is why achieved value alone is an incomplete measure of mediation quality.

Routing is distinct from observation. Blocked Mediation and Routed Mediation share the same clarification opportunity. They differ in whether the answer can affect the final action. Routed Mediation has a lower structural base and lower realized total MG. This shows that evidence must be usable at the decision point. Merely generating clarification is insufficient.

Clarification is useful only when it is decision-sufficient. Routed Mediation improves over Blocked Mediation, yet its structural base remains positive. This shows that routed clarification helps, while the implemented policy class still loses some oracle-relevant distinctions. In this environment, useful mediation requires both access to the latent signal and a policy class that preserves the relevant joint state when choosing the final action.

Information value is action-switch weighted. Panel A and the lambda sweep give the same message. Signal richness matters when it protects distinctions that change the optimal action. As the safety weight grows, unsafe mistakes become more costly, and structural MG rises. This is a value-weighted view of information, not a raw-bit view.

The dynamic residual measures unrealized capacity. For the richer architectures, most realized total MG comes from the dynamic residual. The architecture preserves useful structure, but finite training fails to exploit it. This is the empirical analogue of dynamic MG. It motivates studying the learner and feedback channel as part of the mediation problem.

The experiment supports the decomposition. The results do not reduce to a simple ranking of architectures. Static Binary is easiest to learn. Routed Mediation has the strongest structural capacity among delegated architectures. Blocked Mediation isolates a routing failure. The decomposition is needed to make these facts compatible. It tells us whether loss comes from the interface itself or from finite-budget learning inside the interface.