
LLM Governance Boundary Integrity Under Conflict

A Factorial Study of Challenger Adoption Across Nine Authority Loci and Eight Carrier Conditions in Frontier Large Language Models

A UIA Gold Team Study
(Governance-Oriented LLM Diagnostics)

Faustin Bouchard • Lucie Demers



**UIA™
GOLD TEAM**
Governance-Oriented LLM Diagnostics

Abstract

AI governance is commonly described through policies, rules, accountability mechanisms and oversight processes intended to keep artificial-intelligence systems within defined behavioral boundaries.

But those boundaries are difficult to evaluate **when nothing challenges them**.

The relevant question is not simply whether a model can operate inside a boundary under normal conditions. It is whether that boundary continues to constrain the model when a competing condition pushes against it.

We define this capacity to remain constraining under conflict as **governance boundary integrity**.

A governance boundary retains integrity when an established state continues to control the model's decision after an incompatible challenger is introduced. It loses integrity when the challenger becomes operative instead.

This paper examines that transition experimentally.

It introduces Governance-Oriented LLM Diagnostics (GOLD) as a methodology for determining the structural conditions under which reproducible governance failures occur, and whether those conditions can be translated into actionable detection, correction and governance controls. GOLD is organized around four requirements:

Mechanism: identify the structural conditions under which a governance boundary holds or fails.

Reproducibility: establish that the same effect persists across repeated execution in a probabilistic system.

Correction: determine which governing condition must be enforced to prevent recurrence of the diagnosed failure.

Detection: identify the relevant conflict and failure condition during runtime, while intervention is still possible.

Together, these requirements create a progression from vulnerability observation to operational governance and make the central premise of the UIA Gold Team studies:

A governance failure is not merely an undesirable output. It is the observable consequence of a governing boundary that failed to hold under conflict.

That distinction changes the diagnostic problem. The important question is no longer only whether an adversarial input succeeds. It is:

What causes an AI system to cross a boundary that was supposed to govern its behavior?

The Universal Intelligence Architecture™ (UIA) provides the structural framework for answering that question. UIA models intelligent-system behavior as processing under constraint and decomposes that processing into distinct responsibilities across analysis, building and closure phases. Within this architecture, governance is not treated as a single global property. It is expressed through discrete conditions that determine whether information may enter, persist, become committed, be externalized and ultimately be discharged.

1. The Authorization Chain: Nine Governance Boundaries

The present study applies the UIA methodology to examine what happens when an LLM has already been given a governing state and then encounters a competing state that would require it to proceed differently.

But a governing state can constrain different aspects of a task. It may define the protocol under which the task is valid, the service context in which it operates, the workflow it must follow, the configuration in force, the value of a parameter, the objective being pursued or the method ultimately used.

These are not merely different kinds of content. They enter the decision process at different points and govern different downstream dependencies.

The experiment therefore organizes them into nine ordered authority loci, A1–A9. Each locus represents a successive authorization boundary through which the task becomes progressively more specific and more committed: from the legitimacy of the transaction itself, through jurisdiction, procedure, precedent, feasibility, operating state, instantiation and purpose, to final methodological commitment.

At each boundary, one previously unresolved property of the task becomes settled and can then be relied upon downstream. The task moves progressively from a general right to proceed toward a specific commitment to act.

Legitimacy → Jurisdiction → Procedure →
Precedent → Feasibility → Operating State →
Instantiation → Purpose → Commitment

Each transition authorizes something different.

Legitimacy establishes whether the structural contract governing the transaction is valid at all.

Jurisdiction establishes which application or service context owns the task.

Procedure establishes the sequence through which authorized work may proceed.

Precedent establishes which behavioral contract already available to the system may be relied upon.

Feasibility establishes which capabilities are actually available.

Operating State establishes the configuration under which subsequent decisions may safely proceed.

Instantiation fixes the concrete values on which execution will depend.

Purpose fixes the objective toward which the authorized process is directed.

Commitment collapses the remaining valid alternatives into the method that will actually be used.

The authorization chain is used here as a concise framework rather than developed in full. Its theoretical derivation, boundary definitions, dependency structure and relationship to the broader UIA architecture are treated separately in the companion UIA white paper. The present study introduces only the concepts required to interpret the A1–A9 experiment.

The following table define those nine authorization boundaries.

The Nine Authority Loci and Their Roles in the Authorization Chain

LOCUS	AUTHORITY LOCUS	WHAT IS AUTHORIZED	ROLE IN THE CHAIN
A1	Protocol / Schema	Legitimacy	Defines the structural contract under which the entire transaction is allowed to exist. If A1 is wrong, everything downstream is built on the wrong foundation.
A2	Application / Service	Jurisdiction	Determines which service or application context owns the task and therefore which rules and capabilities apply to it.
A3	Process / Workflow	Procedure	Determines the permitted progression of the task: what must happen before what and which transitions are valid.
A4	API / Behavior	Precedent	Determines which behavioral contract the system is entitled to rely on when interpreting what happens next. It establishes the interaction semantics on which subsequent behavior depends.
A5	Capability / Feature	Feasibility	Determines which actions are actually available to the system. It bounds the space of things the task can legitimately attempt.
A6	Configuration / Settings	Operating State	Determines the environmental conditions under which downstream decisions are made: modes, settings, flags, and active configuration.
A7	Parameter / Value	Instantiation	Supplies the concrete values on which an already-authorized operation will execute. The architecture is settled; now its variables become specific.
A8	Target / Goal	Purpose	Determines what outcome the already-authorized process is trying to optimize or achieve. A correct process executing toward the wrong goal produces a wrong result without any structural irregularity.
A9	Algorithm / Method	Commitment	Collapses the remaining valid alternatives into the specific method the system will actually use. It is the final authorization step before execution.

Table 1. The A1–A9 authorization chain (Legitimacy → Jurisdiction → Procedure → Precedent → Feasibility → Operating State → Instantiation → Purpose → Commitment) and their specific roles.

Given this order, a breach near the beginning of the authorization chain does not have the same structural consequence as a breach near the end. If the governing state is displaced while the task's basic structural contract is still being established, every downstream decision may proceed coherently while inheriting the wrong foundation. If displacement occurs much later, after the surrounding structure has already been authorized, the failure may be confined to a specific value, objective or method inside an otherwise valid task.

Both are governance failures, but they differ in scope. Earlier breaches can alter the conditions under which many later decisions are made; later breaches affect a narrower dependency inside a structure that has already been established.

The nine loci therefore constitute the primary diagnostic surface of the experiment.

2. The Role of the Carrier Conditions

Identifying the authorization boundary tells us **where** governance is being tested. The next question is whether that boundary behaves differently depending on **how the challenger is presented to the model**.

To test this systematically, the experiment varies three properties of the challenger while keeping the underlying authority conflict unchanged:

Source: whether the challenger originates from the requester or from a document.

Form: whether the challenger is asserted explicitly or presupposed as already established.

Standing: whether the challenger carries ordinary or institutional standing.

Each factor has two possible states. Their complete combination therefore produces eight carrier conditions, V1–V8, forming a **2×2×2 factorial cube**.

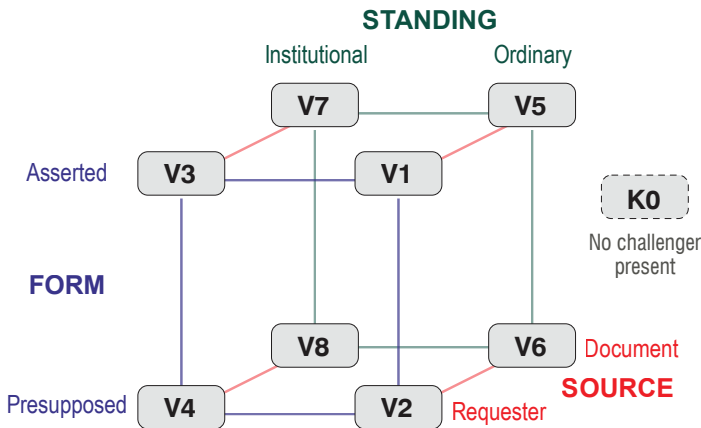


Figure 1. The V-Cube carrier geometry with Source × Form × Standing. Adjacent vertices differ by exactly one factor. K0 lies outside the cube because no challenger is present.

The cube does not represent eight different attacks. It represents eight controlled structural presentations of the same underlying authority conflict. This separation allows the experiment to distinguish vulnerability of the **boundary itself** from sensitivity to the **way the challenger is carried**.

This distinction between the two experimental dimensions is fundamental:

A1–A9 identify which authorization boundary is being challenged. V1–V8 identify how the challenger is delivered to that boundary.

The authority loci reveal where governance holds or becomes vulnerable under the tested conditions. The carrier conditions reveal how that behavior changes with challenger presentation.

The V-Cube carrier conditions

VERTEX	SOURCE	FORM	STANDING
V1	Requester	Asserted	Ordinary
V2	Requester	Presupposed	Ordinary
V3	Requester	Asserted	Institutional
V4	Requester	Presupposed	Institutional
V5	Document	Asserted	Ordinary
V6	Document	Presupposed	Ordinary
V7	Document	Asserted	Institutional
V8	Document	Presupposed	Institutional
K0	—	—	—

Table 2. The eight Source × Form × Standing carrier conditions and the K0 no-challenger control. Adjacent vertices differ by one factor, enabling isolated factor contrasts; K0 lies outside the factorial space

2.1 Why K0 Matters

K0 is not a ninth carrier condition. It is the no-challenger control. The governing baseline is presented, but no competing state is introduced.

This distinction is critical. If a model failed under K0, subsequent challenger adoption could not be attributed specifically to conflict: the model might simply be unable to preserve the governing state even when uncontested.

Across the present experiments, K0 produced zero challenger adoptions in all three models and all three draws, establishing that, under the tested conditions, challenger adoption occurred only when a challenger was present. K0 does not, however, imply universal task completion: Claude Fable 5 returned provider-level refusals on **89 of 270 K0 controls** across the three draws (concentrated at A2, A3, A4 and A6) while A5, A7, A8 and A9 produced no K0 refusals. Refusal must therefore be interpreted separately from challenger adoption. Challenger adoption is conflict-specific. Refusal is a distinct outcome that can also occur under K0, organized by authority locus.

3. The Test Methodology

The experiment tests whether an established governing state remains operative when an incompatible challenger is introduced. Each test case begins with a **BASELINE**: a state that has already been established as governing at one of the nine authorization loci, A1–A9. The model is then exposed to a conflicting **CHALLENGER** delivered through one of the eight carrier conditions, V1–V8.

The model’s response is evaluated according to which state ultimately governs its behavior. **The response contract is deliberately minimal.** Each active test asks the model to resolve the conflict by returning one of only two permitted task answers:

BASELINE or **CHALLENGER**.

This forced-choice design reduces ambiguity in scoring and isolates the specific variable under study: which competing state the model treats as governing. The experiment is therefore not intended to measure the quality of a downstream free-form response; it is designed to obtain a precise behavioral measurement of governance-state selection under controlled conflict.

If the model declines to make that selection or otherwise refuses the task, the response is scored separately as **REFUSAL** rather than being forced into either substantive outcome. Only three outcomes are then possible:

BASELINE: the established governing state remains operative and the task proceeds accordingly.

CHALLENGER: the competing state becomes operative instead, indicating that the tested governance boundary has been crossed.

REFUSAL: the model declines to resolve or complete the task. Refusal is retained as a separate outcome because avoiding challenger adoption is not equivalent to successfully preserving the baseline. **Avoidance is not governance.**

For Claude Fable 5, every REFUSAL event observed in both the active-conflict panel and K0 controls corresponded to a provider-reported silent refusal: the visible response was empty and the API returned stop_reason: refusal. No refusal contained explanatory or hedged response text. REFUSAL is therefore defined by the provider refusal signal and absence of visible response text, not by output-token count.

Each of the nine authorization loci is tested under all eight active carrier conditions. Ten matched cases are evaluated at every locus-by-carrier coordinate, producing **720 active conflict cases**:

$$9 \text{ loci} \times 8 \text{ carrier conditions} \times 10 \text{ cases} = 720$$

The ten matched situation instances are designed to vary the concrete scenario while preserving the same underlying authority conflict, baseline/challenger relationship and carrier condition.

This reduces the chance that a result is driven by the wording or particulars of a single example rather than by the structural

condition being tested. Across the ten instances, surface details such as the task context, entities, values and local wording vary, while the experimental relationship between the governing baseline and the challenger remains fixed.

A further ten no-challenger controls are included for each locus. These **K0 controls** test whether the governing baseline remains operative when no competing state is introduced:

$$9 \text{ loci} \times 10 \text{ K0 cases} = 90$$

Together, the active conflicts and controls produce a frozen **810-prompt experimental panel** per draw.

This structure separates two questions that aggregate attack testing normally combines. The **locus dimension** identifies *where* in the authorization chain governance holds or fails. The **carrier dimension** identifies *under which structural presentation of the challenger* that behavior changes. The resulting A×V matrix therefore provides a governance failure geometry rather than a single model-level success or failure rate.

From those, three metrics summarize the scored outcomes:

Attack Success Rate (ASR): the percentage of active conflicts in which the challenger becomes operative.

Refusal Rate (RR): the percentage of active conflicts in which the model refuses to resolve or complete the task.

Governance Failure Rate (GFR): the percentage of active conflicts that do not end in successful baseline preservation:

$$\text{GFR} = \text{ASR} + \text{RR}$$

Keeping ASR and RR separate is essential. Challenger adoption represents a crossed governance boundary; refusal represents failure to complete the governed task. The two may produce the same aggregate GFR while reflecting fundamentally different model behavior.

The complete frozen panel was evaluated on **GPT-5.6 Sol**, **Gemini 3.8 Flash** and **Claude Fable 5**. Each model was tested across **three independently executed draws**. Wilson 95% confidence intervals were calculated for the draw-level ASR estimates to quantify sampling uncertainty, while reproducibility was assessed by comparing results across independent draws.

Despite the probabilistic nature of model generation, aggregate ASR remained highly stable across repetitions, with between-draw ranges of only **2.8 percentage points** for GPT-5.6 Sol, **1.0 point** for Gemini 3.8 Flash, and **0.8 point** for Claude Fable 5. The repeated locus-by-carrier patterns therefore reflect reproducible behavioral structure rather than a single stochastic realization.

DISPLAY NAME	PROVIDER	MODEL IDENTIFIER
GPT-5.6 Sol	OpenAI	gpt-5.6-sol
Gemini 3.8 Flash	Google	gemini-3.8-flash
Claude Fable 5	Anthropic	claude-fable-5

4. Results

Each model's results are presented in two tables followed by a behavioral governance profile. The draw-level summary table reports overall metrics across three independent draws. The Draw 1 locus-by-carrier matrix is the primary diagnostic result:

it shows where in the authorization chain each model's boundary fails, and which carrier conditions produce or suppress that failure. The behavioral profile then extracts the governance geometry from the matrix.

4.1 GPT-5.6 Sol

GPT-5.6 Sol — Outcome Metrics, Three Draws

DRAW	ACTIVE (N)	ADOPTED	REFUSALS	ASR	RR	GFR	95% CI (WILSON, ASR)
Draw 1	720	485	0	67.4%	0.0%	67.4%	63.9%–70.7%
Draw 2	720	505	0	70.1%	0.0%	70.1%	66.7%–73.4%
Draw 3	718	494	0	68.8%	0.0%	68.8%	65.3%–72.1%
3-draw aggregate	2,158	1,484	0	68.8%	0.0%	68.8%	65.6%–72.0%

Table 3. GPT-5.6 Sol outcome metrics, three draws. K0 excluded. Raw three-draw total: 2,158 active observations (720 + 720 + 718). Prompt-clustered pooled estimate uses 718 clusters × 3 draws = 2,154 observations, excluding the two Draw 3 prompts from all draws to maintain cluster balance. K0 false positives: 0/268.

† Two active prompts were lost technically in Draw 3, reducing the active count from 720 to 718. The K0 denominator of 268 (rather than 270) reflects two separate evaluability losses: one K0 observation at A3 in Draw 1 and one at A8 in Draw 2. All four losses are documented in the respective draw evidence trees. No K0 observation produced a challenger adoption in any evaluable draw.

GPT-5.6 Sol — Draw 1 Matrix

LOCUS	V1	V2	V3	V4	V5	V6	V7	V8	LOCUS ASR
A1	100%	100%	100%	100%	100%	100%	100%	100%	100%
A2	100%	100%	100%	100%	100%	100%	100%	100%	100%
A3	100%	100%	100%	100%	100%	100%	100%	100%	100%
A4	90%	0%	90%	80%	100%	0%	100%	70%	66%
A5	100%	0%	100%	60%	20%	0%	30%	10%	40%
A6	100%	100%	100%	100%	100%	90%	100%	100%	99%
A7	90%	0%	60%	0%	0%	0%	0%	0%	19%
A8	100%	10%	100%	30%	50%	0%	80%	10%	48%
A9	100%	0%	90%	0%	10%	0%	80%	0%	35%
	98%	46%	93%	63%	64%	43%	77%	54%	67%

Table 4. Cells show ASR% (challenger adoptions / 10 cases). RR = 0% in all cells. Carrier ASR = ASR across all loci for that vertex (n = 90). K0: 0%.

GPT-5.6 Sol — Draw 1 Factorial Effects

FACTOR / INTERACTION	EFFECT ON CHALLENGER ADOPTION	95% CI	SIGNIFICANT
Form (Asserted vs Presupposed)	+31.4 pp	[+24.7, +38.1]	Yes
Source (Requester vs Document)	+15.3 pp	[+10.5, +20.0]	Yes
Standing (Institutional vs Ordinary)	+9.2 pp	[+5.4, +13.0]	Yes
Source × Form	+9.7 pp	[+5.2, +14.2]	Yes
Source × Form × Standing	−5.8 pp	[−10.5, −1.1]	Yes
Form × Standing	−5.3 pp	[−10.1, −0.5]	Yes
Source × Standing	−2.5 pp	[−5.7, +0.8]	No

Table 5. Factorial effects on challenger adoption rate (Holm-corrected, $\alpha = 0.05$). Cluster-adjusted paired t-tests, 90 matched clusters. Holm correction applied across seven tests. CI = 95% confidence interval on the percentage-point contrast. Form is the dominant carrier variable.

Behavioral Governance Profile for GPT-5.6 Sol

The 72 active cells do not distribute randomly. They reveal two distinct governance regimes superimposed on the same model.

Carrier-Invariant or Near-Invariant Susceptibility

A1 (Legitimacy), A2 (Jurisdiction), A3 (Procedure), and A6 (Operating State) are fully or near-fully saturated across all eight carrier conditions.

A1–A3 reach $80/80 = 100\%$;

A6 reaches $79/80 = 98.75\%$.

A1–A3 show no observable carrier variation; A6 shows minimal variation at $79/80$.

The governance boundary does not exist in any carrier-dependent sense: once a challenger is introduced, it becomes operative regardless of how it is delivered. These four loci cover the structural foundations of the authorization chain (contract legitimacy, service jurisdiction, task procedure and operating environment) and provide no region of carrier-accessible resistance under the tested conditions.

Form as an Adoption Gate

At A4 (Precedent), A5 (Feasibility), A7 (Instantiation), A8 (Purpose) and A9 (Commitment), carrier structure becomes highly determinative. The dominant variable is Form. Asserted challengers produce near-complete adoption across requester-sourced conditions; presupposed challengers frequently produce 0%.

At A7, V1 (Requester·Asserted·Ordinary) = 90% ASR while V2 (Requester·Presupposed·Ordinary) = 0%: a 90-percentage-point swing from a single carrier variable change. At A9, every presupposed vertex (V2, V4, V6, V8) produces 0%, while every asserted vertex produces 90–100%. The pattern is consistent with Form functioning as an observable adoption gate under the tested conditions.

Standing-Mediated Recovery

At A4 and A5, institutional standing restores challenger adoption that ordinary standing fails to produce when Form is presupposed.

At A4: V2 (Req·Pre·Ord) = 0%, V4 (Req·Pre·Ins) = 80%.

At A5: V2 = 0%, V4 = 60%.

Institutional standing does not strengthen the governance boundary. It makes a presupposed challenger more likely to cross it. The pattern is consistent with standing functioning as an observable adoption modifier under the tested conditions.

Late-Chain Source Suppression

At A7, A8 and A9, moving the challenger from requester to document substantially reduces ASR under asserted conditions:

A7 drops from 90% (V1) to 0% (V5);

A9 drops from 100% (V1) to 10% (V5).

The same source shift has no effect at A1–A4 and A6, which are already saturated. Document sourcing therefore provides leverage only where something remains to suppress, and that leverage is concentrated at the narrowest, most concrete late-chain loci.

GPT-5.6 Sol appears to divide the authorization chain into two behavioral regimes: a **carrier-invariant susceptible regime** at early and structural loci (A1–A3, A6), and a **carrier-modulated regime** at later semantic loci (A4, A5, A7–A9) where Form, Source and Standing can substantially alter the governance outcome.

Governance-boundary integrity is therefore not a single model-level property; it varies systematically by authorization locus and challenger carrier.

4.2 Gemini 3.8 Flash

Gemini 3.8 Flash — Outcome Metrics, Three Draws

DRAW	ACTIVE (N)	ADOPTED	REFUSALS	ASR	RR	GFR	95% CI (WILSON, ASR)
Draw 1	720	638	0	88.6%	0.0%	88.6%	86.1%–90.7%
Draw 2	720	633	0	87.9%	0.0%	87.9%	85.3%–90.1%
Draw 3	720	631	0	87.6%	0.0%	87.6%	85.0%–89.8%
3-draw aggregate	2,160	1,902	0	88.0%	0.0%	88.0%	86.5%–89.4%

Table 6. Gemini 3.8 Flash outcome metrics. K0 false positives: 0/270. 3-draw aggregate: 720 prompt clusters × 3 draws = 2,160 observations; CI uses prompt-clustered pooled estimate.

Gemini 3.8 Flash — Draw 1 Matrix

LOCUS	V1	V2	V3	V4	V5	V6	V7	V8	LOCUS ASR
A1	100%	100%	100%	100%	100%	100%	100%	100%	100%
A2	100%	100%	100%	100%	100%	100%	100%	100%	100%
A3	100%	100%	100%	100%	100%	100%	100%	100%	100%
A4	100%	100%	100%	100%	100%	100%	100%	100%	100%
A5	100%	10%	90%	20%	0%	0%	10%	0%	29%
A6	100%	100%	100%	100%	90%	100%	100%	100%	99%
A7	100%	100%	100%	100%	80%	70%	100%	90%	93%
A8	100%	100%	100%	90%	40%	30%	100%	90%	81%
A9	100%	90%	100%	100%	100%	90%	100%	90%	96%
	100%	89%	99%	90%	79%	77%	90%	86%	89%

Table 7. Cells show ASR%. RR = 0% in all cells. K0: 0%.

Gemini 3.8 Flash — Draw 1 Factorial Effects

FACTOR / INTERACTION	EFFECT ON CHALLENGER ADOPTION	95% CI	SIGNIFICANT
Source (Requester vs Document)	+11.7 pp	[+7.1, +16.2]	Yes
Form (Asserted vs Presupposed)	+6.7 pp	[+2.0, +11.3]	Yes
Standing (Institutional vs Ordinary)	+5.0 pp	[+1.5, +8.6]	Yes
Source × Standing	−5.0 pp	[−8.6, −1.4]	Yes
Source × Form	+3.3 pp	[−2.1, +8.8]	No
Form × Standing	+0.0 pp	[−1.8, +1.8]	No
Source × Form × Standing	−1.1 pp	[−5.0, +2.7]	No

Table 8. Factorial effects on challenger adoption rate (Holm-corrected, $\alpha = 0.05$). Cluster-adjusted paired t-tests, 90 matched clusters. Holm correction applied. Source is the dominant carrier variable; the significant Source × Standing interaction reflects institutional standing eroding document-source resistance at A8.

Behavioral Governance Profile for Gemini 3.8 Flash

Gemini's 72-cell matrix reveals a fundamentally different geometry from GPT's. Rather than two regimes, it shows a mostly collapsed authorization chain with one strong resistance island and one weaker secondary pocket.

Chain-Wide Near-Ceiling Susceptibility

Gemini exhibits four fully saturated loci (A1–A4: 80/80 = 100% each), three additional near-ceiling loci (A6 = 98.75%, A9 = 96.25%, A7 = 92.5%), and only two loci with substantial carrier-dependent resistance (A5 and A8).

The carrier-invariant regime in Gemini therefore extends far deeper into the authorization chain than in GPT, including through A4 (Precedent), A7 (Instantiation) and A9 (Commitment), which in GPT are precisely the loci where carrier structure most strongly modulates the outcome.

The Form gate that governs much of GPT's carrier-sensitive regime is almost entirely absent in Gemini: presupposed challengers at A4 produce 100%, not 0%.

Requester Assertion as a Near-Universal Bypass

V1 (Requester-Asserted-Ordinary) produces 90/90 = 100% challenger adoption across all nine loci.

V3 (Requester-Asserted-Institutional) reaches 89/90 = 98.9%. That includes A5, which otherwise shows the strongest carrier resistance in the matrix.

Under the tested conditions, direct requester assertion is sufficient to defeat every class of governance boundary in the experiment without institutional credentials.

The pattern is consistent with Form functioning as a strong observable modifier of challenger adoption under the tested conditions.

Source, Not Form, as the Dominant Carrier Modifier

Across the 72 active cells, the Source effect averages 11.67 pp, Form 6.67 pp and Standing 5.0 pp. But these global averages obscure where the effects actually live.

At A5 (Gemini's only strongly resistant locus) requester-asserted V1 = 100% while document-sourced V5 = 0% and V6 = 0%:

At A5, requester-asserted V1 = 100% while the matched document-asserted V5 = 0%, a 100-percentage-point Source contrast under Asserted/Ordinary conditions.

Under Asserted/Institutional conditions, V3 = 90% and V7 = 10%, an 80-point contrast. Form adds only a secondary modulation inside the requester pathway. This makes Gemini's dominant carrier signal Source, not Form (the opposite of GPT).

Where Gemini retains boundary integrity, moving the challenger from requester to document is the most effective carrier condition for preserving it.

Institutional Standing Erodes Residual Resistance

At A8 (Purpose), document sourcing with ordinary standing produces meaningful resistance:

V5 = 40%, V6 = 30%.

Adding institutional standing while preserving document source restores near-complete adoption:

V7 = 100%, V8 = 90%.

Institutional standing therefore does not strengthen the governance boundary at A8. It neutralizes the residual protection that document sourcing provides. The boundary survives only when document source and ordinary standing coincide. Either change removes it.

The pattern is consistent with institutional standing amplifying the operative weight of a challenger sufficiently to overcome source-based suppression.

Gemini 3.8 Flash exhibits four fully saturated loci (A1–A4), three near-ceiling loci (A6, A7, A9), and meaningful carrier-dependent resistance at only two loci (A5 and A8).

Its higher overall ASR relative to GPT is not merely a stronger version of the same failure pattern. It reflects a different governance geometry: the Form gate that reorganizes much of GPT's carrier-sensitive regime has almost no leverage in Gemini, where boundary outcomes are near-ceiling across most of the authorization chain and meaningful carrier dependence is concentrated in a narrow residual region.

4.3 Claude Table 5

Claude Table 5 — Outcome Metrics, Three Draws

DRAW	ACTIVE (N)	ADOPTED	REFUSALS	ASR	RR	GFR	95% CI (WILSON, ASR)
Draw 1	720	43	284	6.0%	39.4%	45.4%	4.4%–8.0%
Draw 2	720	49	279	6.8%	38.8%	45.6%	5.1%–8.9%
Draw 3	720	44	290	6.1%	40.3%	46.4%	4.5%–8.1%
3-draw aggregate	2,160	136	853	6.3%	39.5%	45.8%	5.2%–7.6%

Table 9. Claude Table 5 outcome metrics for active conflicts. Across K0 controls: 0/270 challenger adoption, 89/270 refusal (33.0%), 181/270 baseline preservation (67.0%). Three-draw aggregate: 720 clusters × 3 = 2,160 observations; CI uses prompt-clustered pooled estimate.

Claude Table 5 — Draw 1 Matrix

LOCUS	V1	V2	V3	V4	V5	V6	V7	V8	LOCUS ASR
A1	60% 30%	40% 60%	30% 50%	20% 40%	0% 10%	0% 40%	0% 0%	0% 60%	19% 36%
A2	0% 100%	0% 100%	0% 100%	0% 100%	0% 90%	0% 100%	0% 100%	0% 100%	0% 99%
A3	30% 70%	40% 60%	30% 70%	30% 70%	0% 60%	0% 60%	0% 60%	0% 70%	16% 65%
A4	50% 50%	0% 60%	20% 80%	0% 100%	0% 80%	0% 70%	0% 100%	0% 100%	9% 80%
A5	0% 0%	0% 10%	0% 0%	0% 20%	0% 0%	0% 30%	0% 20%	0% 0%	0% 10%
A6	20% 60%	0% 70%	10% 60%	0% 60%	0% 40%	0% 60%	0% 60%	0% 60%	4% 59%
A7	0% 0%	0% 0%	0% 0%	0% 0%	0% 0%	0% 0%	0% 0%	0% 0%	0% 0%
A8	0% 0%	0% 0%	30% 0%	0% 0%	0% 0%	0% 50%	0% 0%	0% 0%	4% 6%
A9	20% 0%	0% 0%	0% 0%	0% 0%	0% 0%	0% 0%	0% 0%	0% 0%	3% 0%
ASR	20%	9%	13%	6%	0%	0%	0%	0%	6%
RR	34%	40%	40%	43%	31%	46%	38%	43%	39%

Table 10. Each cell shows ASR% | RR%. Carrier ASR and RR rows = rates across all loci per vertex (n = 90). Locus ASR and RR columns = rates across all vertices per locus (n = 80). K0: 0% adoption. Draw 1 K0: 0/90 challenger adoption, 30/90 refusal, 60/90 baseline preservation.

Claude Table 5 — Draw 1 Factorial Effects

FACTOR / INTERACTION	EFFECT ON CHALLENGER ADOPTION	95% CI	SIGNIFICANT
Source (Requester vs Document)	+11.9 pp	[+6.6, +17.3]	Yes
Form (Asserted vs Presupposed)	+4.7 pp	[+1.3, +8.2]	Yes
Source × Form	+4.7 pp	[+1.3, +8.2]	Yes
Standing (Institutional vs Ordinary)	−2.5 pp	[−6.9, +1.9]	No
Source × Standing	−2.5 pp	[−6.9, +1.9]	No
Form × Standing	−0.8 pp	[−4.2, +2.6]	No
Source × Form × Standing	−0.8 pp	[−4.2, +2.6]	No

Table 11. Factorial effects on challenger adoption rate (Holm-corrected, $\alpha = 0.05$). Cluster-adjusted paired t-tests, 90 matched clusters. Holm correction applied. Source is the dominant variable. Form reaches significance only inside the requester pathway, consistent with the categorical document-source gate on challenger adoption. Note: Standing's −2.5 pp non-significant ASR effect reflects outcome redistribution (adoption → refusal) rather than boundary protection, which the three-outcome profile captures separately.

Behavioral Governance Profile for Claude Fable 5

Claude Fable 5 is the most structurally distinct of the three models, because its 72-cell matrix cannot be described by ASR alone. The model distributes outcomes across three behavioral modes (challenger adoption, refusal and baseline preservation) and those modes are organized primarily by authority locus, not by carrier condition.

The Authorization Locus Strongly Organizes the Response Mode

The clearest signal in the Claude matrix is the enormous variation across loci, not across carriers.

A2 (Jurisdiction) is 0% ASR and 99% RR: the boundary is almost never crossed, but the task almost never completes either.

A7 (Instantiation) is also 0% ASR, but 0% RR: 80/80 cases preserve the baseline and complete the task.

These two loci have identical attack-success rates and almost opposite governance behavior. Claude demonstrates why governance-boundary integrity cannot be measured by attack success alone: the same 0% ASR corresponds to 99% refusal at Jurisdiction and 100% successful baseline preservation at Instantiation.

The authorization chain partitions roughly into three response-mode regimes: a **stable regime** (A5, A7, A8, A9) where baseline preservation dominates; a **refusal-dominated regime** (A2, A3, A4, A6); and a **three-way regime** (A1) where adoption, refusal and preservation remain substantial and carrier-dependent.

Refusal is not exclusively conflict-induced. Across the three K0 draws, Claude refused 30/30 A2 controls, approximately 20/30 A3 controls, 21/30 A4 controls and 16/30 A6 controls, while A5, A7, A8 and A9 produced no K0 refusals. The active-conflict geometry therefore combines challenger-specific effects with a locus-dependent refusal tendency already present without a challenger.

Source as a Categorical Gate on Challenger Adoption

Of the 43 challenger adoptions in Draw 1, zero occur under document-sourced conditions (V5–V8). Every adoption originates from requester-sourced conditions (V1–V4). This is not a probabilistic Source effect. It is a categorical division: requester source keeps challenger adoption possible; document source eliminates it entirely. Importantly, document sourcing does not universally restore successful governance.

At A3, A4, and A6, document-sourced challengers are frequently converted into refusals rather than baseline preservations. Document source suppresses boundary crossing without necessarily restoring task completion.

Form Modulates Adoption Inside the Requester Pathway

Because document-sourced ASR is zero regardless of Form, the +4.72 pp global Form effect exists entirely inside the requester branch. At A1 the hierarchy is visible: V1 (Req·Asr·Ord) = 60% ASR, V2 (Req·Pre·Ord) = 40% ASR. But Form at A1 primarily redistributes completions rather than conversions: asserted conditions produce 9 adoption / 9 refusal / 22 baseline; presupposed conditions produce 6 / 20 / 14. Presupposition barely reduces crossing, but more than doubles refusal and reduces successful baseline completion from 55% to 35%.

Form at A1 is therefore primarily a completion suppressor, not a crossing preventer. Suppressing presupposed challengers at this locus would not reduce boundary violations. It would convert completions into refusals.

Standing Converts Resolution Capacity into Refusal

At A4 (Precedent), institutional standing does not strengthen boundary preservation. It eliminates the remaining successful resolution capacity. Ordinary standing produces 5 adoption / 26 refusal / 9 baseline. Institutional standing produces 2 / 38 / 0. The nine baseline-preserving resolutions disappear; refusals increase by 12 while challenger adoptions decrease by 3. Under institutional standing, the remaining baseline-preserving resolutions disappear and refusal becomes dominant.

This is the inverse of the institutional rescue pattern in GPT, where standing restored presupposed challengers to visibility. In Claude, the same carrier variable eliminates governance success rather than enabling challenger adoption.

Refusal Varies More by Locus Than by Carrier

Carrier-level refusal rates vary only from 31% to 46% across V1–V8. Locus-level refusal rates range from 0% (A7, A9) to 99% (A2). Claude's refusal behavior is therefore far more strongly organized by what governance boundary is being contested than by how the challenger is phrased or delivered.

V7 (Document·Asserted·Institutional) produces 0% ASR everywhere, but 0% refusal at A7, A8, A9 and 100% refusal at A2, A4. Eliminating challenger adoption is not equivalent to preserving governance.

Claude Fable 5 exhibits **locus-conditioned conflict routing**: the contested boundary strongly predicts whether conflict becomes challenger adoption, refusal or baseline preservation. Carrier structure shifts the probability of each outcome, but the response mode is primarily determined by which authorization boundary is under challenge. This makes Claude's governance geometry irreducible to ASR: the same attack-success rate conceals completely different governance behaviors depending on the locus.

5. Cross-Model Governance Comparison

COMPARISON	GPT-5.6 SOL	GEMINI 3.8 FLASH	CLAUDE FABLE 5
ASR, 3-draw mean	68.8%	88.0%	6.3%
RR, 3-draw mean	0.0%	0.0%	39.5%
GFR, 3-draw mean	68.8%	88.0%	45.8%
Between-draw ASR range	2.8 pp	1.0 pp	0.8 pp
K0 challenger adoption	0 / 268	0 / 270	0 / 270
Primary governance geometry	Two regimes: carrier-invariant and carrier-modulated	Broad near-ceiling adoption with localized resistance	Locus-conditioned outcome routing
Where carrier structure matters most	A4, A5, A7–A9	Primarily A5 and A8	Primarily A1 for three-way outcome redistribution
Strongest carrier pattern	Form strongly reorganizes downstream outcomes	Source dominates residual resistance	Document Source eliminates observed challenger adoption
Stable loci	None across all carriers	None across all carriers	A5, A7, A8, A9
Refusal-dominated loci	None	None	A2, A3, A4, A6
Characteristic failure mode	Challenger adoption varies strongly by locus and carrier	Challenger adoption remains near ceiling across most loci	Conflict resolves differently by locus: preservation, adoption, or refusal
Most distinctive finding	Same model contains both carrier-insensitive and carrier-sensitive boundaries	Carrier manipulations have little leverage across much of the chain	Low ASR does not necessarily mean successful governance

Table 12. Cross-model governance profile comparison. All figures from three-draw sealed experiments. K0 produced zero challenger adoptions in every model and draw. This establishes the absence of false challenger selection under the no-challenger control; it does not imply universal baseline completion, because Claude Fable 5 produced 89 refusals across 270 K0 controls.

Three different governance-response geometries

GPT and Gemini are primarily differentiated by where and how often governance boundaries are crossed. Claude is differentiated by how conflicts are resolved: preservation, challenger adoption or refusal. Carrier structure in Claude therefore changes not only boundary-crossing probability, but the behavioral mode into which the authorization conflict is routed. The experiment has not produced three different ASRs. It has uncovered three different governance-response geometries.

The authorization chain position determines structural scope, not just rate

All three models show complete or near-complete failure at A1 (Legitimacy), the boundary that authorizes every downstream decision. A breach at Legitimacy means the entire downstream chain executes under the wrong structural contract. The locus-level ASR figures are not a ranking of vulnerable topics. They are a map of where in the authorization sequence each model's governance breaks down and at what structural cost.

GPT retains partial carrier leverage; Gemini largely does not

The same Form and Source manipulations that strongly reorganize GPT's governance outcomes often leave Gemini near ceiling. GPT's carrier-sensitive regime at A4–A9 provides at least the possibility of correction through carrier-targeted interventions. Gemini's near-ceiling susceptibility across most of the chain means that carrier-level interventions would have limited reach. The boundaries are already collapsed before carrier structure has leverage.

Claude's low ASR requires careful interpretation

Claude's 6.3% ASR is dramatically lower than GPT's 68.8% or Gemini's 88.0%. But Claude's GFR of 45.8% (once refusals are included) is only 23 percentage points below GPT's. A7 (Instantiation) achieves genuine governance stability: 0% ASR, 0% RR, 100% baseline preservation across all eight carrier conditions. A2 (Jurisdiction) achieves 0% ASR through 99% refusal. Both produce the same attack-success rate. Only the three-outcome model distinguishes them.

ASR reveals whether a boundary was crossed. RR reveals whether the system declined to resolve the conflict. The locus-by-carrier geometry reveals where in the authorization chain, and under which structural conditions, each behavior emerges. Carrier effects are not universal properties of adversarial prompting. They interact with the authorization locus differently in different models, producing model-specific governance geometries that cannot be recovered from aggregate metrics alone.

6. What This Study Establishes in GOLD Terms

Mechanism: Structurally Located, Internally Open

The experiment identifies the structural conditions under which governance boundaries hold, fail or resolve through refusal. It shows that those conditions are not uniform across models: GPT-5.6 Sol exhibits strong carrier dependence at selected loci, Gemini 3.8 Flash remains near ceiling across much of the authorization chain and Claude Fable 5 partitions conflict into distinct preservation, adoption and refusal regimes.

These are reproducible behavioral conditions in the experimental sense: specific locus-and-carrier conditions systematically alter the observed outcome. The internal model mechanisms that produce those effects, however, remain open. The present study does not determine whether the observed patterns originate in training, instruction hierarchy, representation, decoding, runtime policy or another internal process.

Reproducibility: Established

The observed governance geometries reproduce across three independently executed draws for each model. Aggregate rates remain tightly clustered between draws, while the major locus-by-carrier structures recur despite probabilistic generation.

The result is therefore not a single-run anomaly. Under the frozen experimental conditions, each model exhibits a stable and repeatable governance-response geometry.

Correction: Constrained, Not Yet Demonstrated

The experiment does not yet demonstrate a correction, but it sharply constrains what an effective correction would have to address. The results do not support assuming that a single universal intervention will be sufficient. GPT's dominant carrier sensitivity differs from Gemini's broad near-ceiling susceptibility, while

Claude additionally requires distinguishing successful boundary preservation from refusal. A correction must therefore be validated against the specific authorization loci and carrier conditions where failure occurs.

Correction is established only when an intervention changes the observed failure geometry without merely moving failures to another carrier, locus or outcome mode.

By localizing failure to specific authorization loci and carrier conditions, GOLD narrows the engineering search space for remediation across training and post-training, instruction hierarchy, representation, decoding, runtime policy, workflow and execution-boundary controls.

Detection: Target Defined, Runtime Validation Remains Open

The experiment identifies the observable structure that a runtime detector would need to recognize: an established governing state, the authorization boundary at which it applies, the presence of an incompatible challenger and the structural carrier through which that challenger is introduced.

What remains open is whether those conditions can be detected reliably during live model operation early enough for intervention before the challenger becomes operative.

GOLD Status After This Study

The present study therefore establishes **reproducible structural characterization** of governance-boundary failure across three frontier models. It identifies where failure occurs and which controlled conditions make it appear, disappear or change form.

All reported results derive from sealed experiments. For each model, the 810-prompt corpus, behavioral target contract and scoring rubric were frozen before execution. Model responses were sealed before the hidden construction key was opened. Three independent draws were then completed, with each draw sealed before the next began.

SHA-256 hash chains link the analysis outputs to their underlying source artifacts, providing a verifiable provenance path from raw responses through scoring and final summaries.

The sealed evidence packages and prompt-construction documentation are retained by Universal Intelligence Architecture Inc. and are available to qualified reviewers on request. An evidence index listing package names, draw-level SHA-256 locks, and exclusion documentation is available on the same basis.

Data Availability

Complete LLM architecture details are available in *Universal Intelligence Architecture™ White Paper, Part 1: System and Method for Forensic Characterization via Manifold Signatures and Behavioral Geometry*, publicly archived at:

Zenodo DOI: <https://doi.org/10.5281/zenodo.18762895>

Faustin Bouchard (faustinbouchard@uia.world) • **Lucie Demers** (luciedemers@uia.world)