SELECTIVE INVARIANCE UNDER STRATEGIC INFORMATION MANIPULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

An agent that buys on a user's behalf reads a seller's description, contract terms and framing, then accepts the offer or walks away. Such a rule must not move when the seller merely re-presents a fixed transaction, yet must still move when the transaction genuinely changes. The latent state is unavailable at inference, so the rule cannot tell the two apart by inspection; it must be built so that one class of change cannot reach it. We call this *selective invariance* and ask when it is attainable. Given a specified admissible channel and observable representation, comparing the displacements that channel can produce with the direction that carries utility gives a criterion that sorts channels into three classes: invariance is free, invariance costs enumerating a completion set, or invariance is unattainable and the channel can only be estimated. Existing structural defenses show how to isolate particular flows; our contribution is an ex ante test of whether exact model-independent closure exists for a specified channel, not a detector that discovers arbitrary free-text manipulations. Because an LLM buyer is an opaque model rather than one we design, a guarantee valid for every decision function requires exact invariance, which no prompt-level defense provides. We measure both halves on paired renderings of the same hidden transaction, across three second-hand marketplace and bargaining benchmarks and seven model families. Presentation alone raises harmful acceptance by $+0.270$ even for a buyer that accepts no bad offer under the honest rendering, and a safety prompt's protection varies seventeen-fold across families. Structural closure cuts harmful acceptance from 0.097 to 0.037 at essentially unchanged valid acceptance, holds the closed channels at exactly zero, and still reproduces a hidden-state oracle's decision under genuine concessions on all $2{,}605$ held-out negotiations. Applied unchanged to tool-using agents inside AgentDojo, with their suite, their injections and their verdicts, the criterion predicts in advance which defenses can be exact: ours takes attacker success from 0.465 to exactly 0.000 on the channel it closes, at no cost in task utility, while the published datamarking baseline reaches 0.528.

1 INTRODUCTION

Autonomous agents increasingly negotiate prices and commit to purchases for users. A buyer agent sees a second-hand listing, the seller's claims about the item, the contract terms attached to the offer and a deadline, and decides whether to accept or take its outside option. The seller need not alter the item to change that decision: it can rewrite the description, foreground a conditional benefit, add urgency or rearrange contractual detail. A buyer accepting after such a re-rendering, when the transaction's true expected utility stays below its reservation, has suffered economic decision distortion rather than ordinary bargaining loss. Standard negotiation metrics cannot see it, since a strong negotiator can accept a dominated contract and a conservative one can suppress harm by refusing nearly every valid offer, and agent-safety benchmarks score a different intervention, namely injected instructions, unsafe tool use and explicit policy violations (Debenedetti et al., 2024; Zhan et al., 2024; Zhang et al., 2024; 2025). Holding the transaction fixed and varying only its presentation isolates the effect, and states a design problem with two sides:

> *How can a decision rule be invariant to strategic variation without becoming invariant to real utility variation?*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Either side alone is trivial, since a rule that refuses everything is invariant to every manipulation and one that trusts every claim tracks every genuine improvement. What makes the pair hard is that nothing in the message reveals which kind of change just occurred, because the rule never sees the transaction behind it. It can separate the two only by being built so that one of them cannot reach it.

Existing structural defenses answer how to isolate a known untrusted flow. Our question is prior: for a *specified* strategic channel, does any exact model-independent defense exist, and what information must it discard? The criterion in Section 4.1 answers that existence-and-scope question before any buyer model is run; this is not the claim that structural defenses beat prompting. It also makes falsifiable predictions before implementation. Latent quality is unclosable in our primary representation, yet re-encoding the same quantity against a public identity is predicted to split it into one exactly closable sub-channel and one that must remain open; measured closure moves by exactly 0.0000 on the former and is bit-identical to no defense on the latter (Section 5.4). Likewise, class assignments made from AgentDojo's task interfaces predict which goals admit exact closure before any episode is executed.

This distinction matters because a soft risk score remains model-dependent. Across the seven tested families, competence does not provide a stable ordering of manipulation effects (Section 5.2); we use those comparisons only against competence as a deployment-time risk surrogate, not to infer a population relationship from seven models. We instead close a channel structurally whenever its effect set is specified and observable, and estimate only what stays latent. We test this on $2{,}605$ held-out CraigslistBargain negotiations and two further bargaining environments. Seller-controlled information splits three ways: payoff-irrelevant presentation can be projected out of utility evaluation; enumerable payoff-relevant uncertainty can be completed pessimistically; and latent state such as unverified quality admits neither because two true states can induce one observation. The first two yield channel-level invariance with no learned risk score. The third inherits whatever a verification, margin, abstention or uncertainty model assumes, and our experiments price that: the residual layer misses its own coverage precondition on a quarter of messages and inflates realized risk several-fold once the seller best-responds to the published rule (Johnstone & Cox, 2021; Angelopoulos et al., 2024; Csillag et al., 2025).

In summary, our contributions are three-fold. Firstly, we make selective invariance measurable: paired renderings that hold the transaction and its true expected settlement value fixed isolate the invariance side, while genuine concessions isolate responsiveness, so a rule cannot pass by being insensitive. Secondly, *given* an admissible action set $\mathcal{A}_k$ and observable representation φ, we give an ex ante criterion for attainability, comparing the channel's displacement set with the direction that carries utility to sort it into invariance-free, invariance-by-enumeration and invariance-unattainable. The input to the criterion is the specified channel rather than arbitrary raw text; within that scope, a black-box decision-maker forces the invariant side to be exact, which is why prompt-level defenses cannot provide the same model-uniform guarantee. Thirdly, we test predictions that distinguish the criterion from a post hoc structural design: re-encoding changes a channel's class, and applying the criterion unchanged inside AgentDojo names in advance which defenses can be exact; the selected operator takes attacker success from 0.465 to exactly 0.000 on the channel it closes.

2 Related Work

Three strands bear on this work, one for each thing it changes: what a negotiation benchmark measures, where an agent-safety defense intervenes, and what a model of strategic behaviour assumes about who designs the decision rule.

Negotiation benchmarks measure the wrong quantity for this failure. They score agreement, utility, strategy or welfare rather than harmful acceptance under a fixed hidden transaction (Lewis et al., 2017; He et al., 2018; Bianchi et al., 2024; Xia et al., 2024; Bhattacharya et al., 2025; Kirshner et al., 2026; Oh et al., 2026; Liu et al., 2026; Liang & Xu, 2026), and TERMS-Bench, the closest, scores a false-agreement rate for individually irrational offers (Zhang et al., 2026). We inherit the failure they report and change the unit of measurement: sampling a fresh scenario per episode confounds the deal with its presentation, whereas holding one latent transaction fixed and varying only the rendering makes an effect attributable and a defense testable.

Agent-safety defenses already split the way we argue decisions should. Those benchmarks target injected instructions and unsafe tool use (Debenedetti et al., 2024; Zhan et al., 2024; Zhang et al., 2024; 2025; Vijayvargiya et al., 2025). CaMeL routes untrusted data out of program flow at a measured price in utility (Debenedetti et al., 2026); training-time separation (Chen et al., 2025) and prompt-level marking (Hines et al., 2024) are the softer alternatives, which adaptive red-teaming defeats (Yin et al., 2026). Above both, contextual integrity implies that no fixed policy admits every legitimate flow while blocking every injected one (Abdelnabi & Bagdasarian, 2026). Our connection is a classification rather than a new defense: routing is our first two classes, marking and prompting are the soft defenses the black-box argument excludes from exactness, and that impossibility is our third class. What we add is an ex ante existence test: given a specified admissible channel and observable representation, it assigns that channel to its class before any model runs. It does not discover an exhaustive attack set from open-ended text. We run the resulting operators inside AgentDojo's own harness.

Our seller stays inside the economic protocol, which places this work near strategic classification and performative prediction (Hardt et al., 2016; Perdomo et al., 2020; Chen et al., 2018; Shavit et al., 2020; Björkegren et al., 2020; Lechner et al., 2023), while robust optimization and conformal methods (Bertsimas et al., 2011; Esfahani & Kuhn, 2018; Johnstone & Cox, 2021; Angelopoulos & Bates, 2023; Angelopoulos et al., 2024; 2025; Wang et al., 2023; Guo, 2026) remain the right tools for our residual channel, their boundary under a strategic seller (Csillag et al., 2025; Farinhas et al., 2024) being what Section 5.3 measures. The difference is the decision-maker's position. That literature lets it design the predictor, so a cone condition suffices; here the predictor is a fixed opaque model, and Proposition 1 shows the same guarantee then forces exact invariance. Appendix O gives the full discussion.

3 Problem Definition

3.1 What the seller controls and when acceptance is harmful

Let $x \in \mathcal{X}$ denote the hidden transaction state, including item attributes and exogenous conditions needed to determine economic value. Let $p \in \mathbb{R}_+$ denote the realized net settlement price. The buyer has utility

$$U_B : \mathcal{X} \times \mathbb{R}_+ \to \mathbb{R} \tag{1}$$

and reservation utility U_{res}. The seller sends an observable signal

$$m = \langle x_c, \phi, \tau \rangle \in \mathcal{M}, \tag{2}$$

where x_c is a claimed state, ϕ is a contract specification, and τ is an urgency or validity field. A buyer policy $\pi : \mathcal{M} \to \{0, 1\}$ returns reject or accept.

We study a seller action $a \in \mathcal{A}$ that maps an honest rendering m_0 to $T_a(m_0)$ without changing the latent transaction used by the oracle. In the primary benchmark we additionally impose revenue neutrality, so the seller's true expected settlement price is unchanged. The harmful-acceptance indicator is

$$h(x, m) = \mathbf{1}[\pi(m) = 1 \ \wedge \ \mathbb{E}_{P_{\text{true}}}[U_B(x, p)] < U_{\text{res}}]\,. \tag{3}$$

A channel's attack effect is the paired difference in h between its attacked and clean rendering, aggregated at the task level (Figure 1).

3.2 What each stage of the pipeline may observe

What a defense can do depends on what the seller can change and what the buyer can observe. Three kinds of channel recur, derived in Section 4.1 rather than posited: presentation that cannot change intrinsic utility at fixed (x, p); payoff uncertainty whose signal defines a finitely representable set $\Gamma(m)$ of completions, as unresolved discounts, future credits or bundle requirements do; and latent payoff state, where distinct economic states induce one signal and $\Gamma(m)$ is unrecoverable, an identification boundary rather than a parsing problem.

The benchmark oracle sees x only for evaluation, the deployed buyer sees seller claims, contract text, verification metadata and its own preferences, and test labels never enter the policy. Table 7 states what each stage may observe.

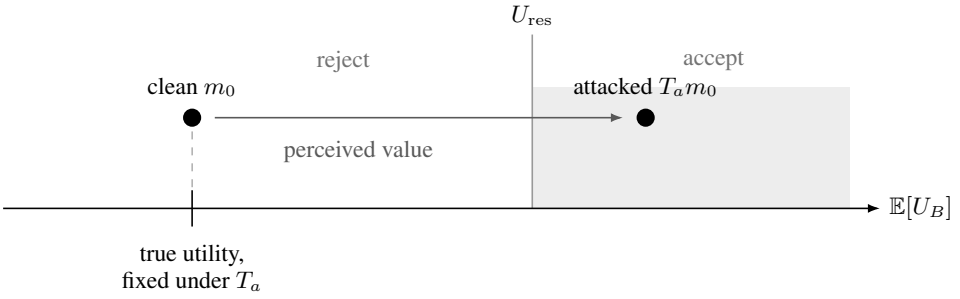


Figure 1: What the benchmark measures. One hidden transaction fixes the true expected utility below the buyer's reservation, so accepting is harmful, and an honest rendering prices the offer there. A revenue-neutral re-rendering leaves the transaction and the seller's true expected settlement price untouched and moves only the perceived value across the threshold. Every attack is a paired difference of this form, which is what makes a decision change attributable to presentation rather than to a different deal.

4 When selective invariance is attainable

4.1 A criterion on observables for a specified channel

Selective invariance pulls in two directions, so it need not be attainable at all. We take the admissible action set $\mathcal{A}_k$ and the observable representation φ as specified; conditional on that modeling choice, attainability depends on the geometry of what the channel can move rather than on the downstream model. That geometry is checkable from observables without querying the buyer. The criterion does *not* infer an exhaustive $\mathcal{A}_k$ from arbitrary natural language.

Write $\varphi(m)$ for the observables a decision rule may read and

$$\Delta_k(m) = \{\varphi(T_a m) - \varphi(m) \ : \ a \in \mathcal{A}_k\} \tag{4}$$

for the displacement set channel k can produce on a fixed latent state. Table 1 collects the symbols this section introduces.

Table 1: Symbols of this section, each defined at its first use above; the table is a local reference rather than a substitute for those definitions.

Symbol	Meaning	Domain
$\varphi(m)$	observables a decision rule may read	feature space
$\mathcal{A}_k, T_a$	actions of channel k and the rendering each induces	$a \in \mathcal{A}_k$
$\Delta_k(m)$	displacements channel k produces at m	set of vectors
∇V	direction in which observables carry utility	vector
Δ_k°, $\Delta_k^\perp$	polar cone and orthogonal complement of Δ_k	cone, subspace
$T, \Gamma(m)$	preprocessing map, and completions left unresolved in m	map, set

Strategic classification calls a valuation manipulation-proof on k when ∇V lies in the polar cone $\Delta_k^\circ = \{g : \langle g, \delta\rangle \le 0 \ \forall \delta \in \Delta_k\}$, asking only that manipulation not *raise* the score (Hardt et al., 2016; Chen et al., 2018; Shavit et al., 2020; Björkegren et al., 2020). That relaxation is available to whoever *designs* the valuation. We do not design ours: the buyer is a fixed, opaque model and our only lever is a preprocessing map T applied before it, which changes the type of the condition.

Proposition 1 (Black-box proofness forces invariance)**.** *Let the deployed rule be $f \circ T$ and call it Δ_k-proof at m if $f(T(\varphi \oplus \delta)) \le f(T(\varphi))$ for every $\delta \in \Delta_k(m)$. Then $f \circ T$ is Δ_k-proof for* every *f if and only if T is Δ_k-invariant, $T(\varphi \oplus \delta) = T(\varphi)$.*

An unrestricted f can order any two distinct images the wrong way (Appendix A.10), which is what separates the two settings. A designer needs only $\nabla V \in \Delta_k^\circ$, a cone: it suffices that manipulation never raises the score because the designer chooses the valuation. We cannot make that choice. Any

displacement left distinguishable after T can be assigned the wrong ordering by some admissible f, so a model-uniform guarantee needs the two-sided $\nabla V \in \Delta_k^{\perp}$, a subspace. The cone and subspace differ whenever manipulation is directional. Accordingly, prompt-level defenses that leave clean and manipulated representations distinguishable cannot provide this model-uniform guarantee. Invariant maps factor through the quotient by the closure of $\varphi \sim \varphi \oplus \delta$, so our price is the information that quotient destroys, and the three cases are its three behaviors.

- Δ_k moves no payoff-relevant observable: any informative valuation is trivially proof, so quotient it out at zero price. Urgency.
- $\nabla_\varphi \mathbb{E}[U \mid \varphi] \in \Delta_k^{\circ}$ already: closure is exact and free. This is what an enumerable completion set buys, since manipulation can only enlarge Γ and the informative direction penalises exactly that. Contract complexity and tail incentives, at exactly 0.0000.
- The directions are opposed: no valuation is both proof and informative, the price is strictly positive, and the residual must be estimated. Latent quality, the one channel surviving every operator we build.

The criterion is falsifiable. A channel's class is a property of the *representation*, so re-encoding should move it. Section 5.4 runs that test where it matters most, on the one channel every operator here leaves open; Appendix B repeats it on selective disclosure, where the effect moves from $+0.048$ to exactly $+0.000$.

Scope of decidability. "Decidable from observables" is conditional on the specified $\mathcal{A}_k$ and φ. Discovering every manipulation available in open-ended free text is outside the claim; an action omitted from $\mathcal{A}_k$ is outside the resulting guarantee. What the criterion removes is downstream-model dependence once the channel is specified: its class is fixed before any buyer model is run.

4.2 From the three cases to three operators

The three cases of Figure 2 turn into three operators, applied in that order. Let $Q(m)$ be the representation reaching the intrinsic utility evaluator. For a payoff-irrelevant channel k we require $Q(T_a m) = Q(m)$ for all $a \in \mathcal{A}_k$. For urgency, Q drops the urgency text from economic valuation while keeping a clock field for scheduling, so the economic input is invariant by construction. For an enumerable payoff channel, let $\Gamma(m)$ be its feasible completions and define

$$V_\Gamma(m) = \inf_{\omega \in \Gamma(m)} V(Q(m), \omega), \tag{5}$$

where V is the utility score before the acceptance threshold. In our contract implementation, writing $\Omega(\phi)$ for the clause-level instance of Γ,

$$p_{\mathrm{cl}}(\phi) = \sup_{\omega \in \Omega(\phi)} p(\phi, \omega), \tag{6}$$

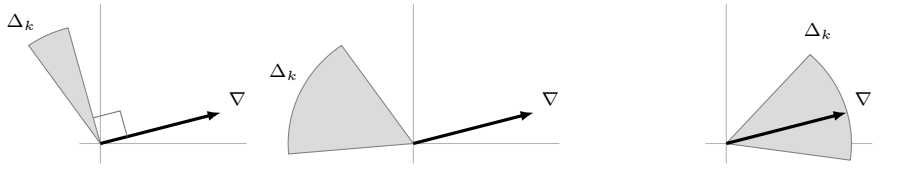
and guaranteed benefits are credited only when their conditions are already verifiable.

Proposition 2 (Channel closure, exactly when inclusion holds)**.** *Let the acceptance rule be monotone in V_Γ. If an admissible manipulation satisfies $\Gamma(T_a m) \supseteq \Gamma(m)$ then $V_\Gamma(T_a m) \le V_\Gamma(m)$, so it cannot turn a rejection into an acceptance. If instead $\Gamma(T_a m) \not\supseteq \Gamma(m)$, there are a price function and a threshold under which it does.*

The inclusion condition is the boundary of the guarantee rather than a convenience, and Appendix K measures the failing side. The argument uses nothing about how quality is priced: writing $V_\Gamma(m) = g(q(m)) - \sup_{\omega \in \Gamma(m)} p(m, \omega)$, a channel that leaves q alone gives $g(q(T_a m)) = g(q(m))$, so the same inequality holds for every g and the exactness belongs to the operator rather than to the affine quality-to-value map our environments share (Appendix A.4). No learned attack score, exchangeability assumption, or buyer-model property enters.

4.3 What is left once the closable channels are closed

Structural closure does not identify latent state. What survives it is passed to a residual layer that keeps a Wasserstein ambiguity set around the post-closure claimed-state distribution and accepts



(a) $\langle \nabla, \delta \rangle = 0$ quotient out, price 0 (b) $\langle \nabla, \delta \rangle \leq 0$ close exactly, price 0 (c) $\langle \nabla, \delta \rangle > 0$ for some δ no proof informative rule

Figure 2: The criterion compares two directions in observable space: the displacements Δ_k a channel can produce, and the direction $\nabla = \nabla_\varphi \mathbb{E}[U \mid \varphi]$ along which observables carry utility. When manipulation cannot move the valuation (a), the channel is quotiented out for nothing. When it can move it only downward (b), taking the worst case over an enumerable completion set is exact, and again free. When some displacement raises the valuation (c), no invariant rule is also informative, so the residual must be estimated and its guarantee becomes conditional. Because the buyer is black-box, (b) is not available as a relaxation of (a) unless the completion set is genuinely enumerable (Proposition 1).

only when the robust lower value clears the reservation by an evaluation margin. We keep the layer because that state stays unresolved, and we do not treat it as a contribution: its guarantee holds only where coverage holds and where calibration and deployment are exchangeable, which Section 5.3 measures. The construction, the fitted radius and Proposition 3 are in Appendix M.

5 Experiments

5.1 Benchmarks, arms, and how the intervals are computed

Four questions organize the evaluation, one per subsection below: whether presentation alone moves deployed buyers and whether competence is a usable risk surrogate; whether the invariance the criterion promises is attained without the rule simply becoming insensitive; whether a channel's class is a property of its encoding rather than of the manipulation; and whether the criterion predicts in a domain we did not build.

The primary benchmark is built from CraigslistBargain (He et al., 2018), retaining 5,213 negotiations and forming a hidden oracle state with sampled latent quality, since the source corpus does not record intrinsic item value (Appendix D). A task-level category-stratified 25/25/50 split gives 1,304 calibration, 1,304 validation and 2,605 test tasks. Every attack is revenue-neutral to tolerance 10^{-9} and the clean rendering has $\mathrm{HAR} = 0$ for every scripted policy, so a nonzero clean value would indicate a benchmark implementation error.

Scripted policies run on all 2,605 test tasks, and model-based policies on category-stratified subsets of the same tasks, $n = 401$ for `qwen3.6:35b-a3b` and $n = 200$ for each of six further families, four open-weight and two hosted. Each task produces several correlated renderings, so the independent unit is the task: intervals use 5,000 task-level bootstrap resamples in the main pipeline and 3,000 in the per-family and side analyses, and paired comparisons use sign-flip permutation tests with Holm-Bonferroni correction within each hypothesis family. We report harmful acceptance, valid acceptance, missed opportunities and normalized regret together so that simple rejection cannot appear as a free safety gain.

The scripted arms are the rows of Table 2, model experiments add direct LLM decisions, a safety-prompt control and an LLM appraisal routed through the same gate, and recent negotiation systems serve as positioning rather than baselines, their action spaces and outcome metrics differing from this paired one-step task (Appendix P).

5.2 Does presentation alone move deployed buyers?

Against the rule that trusts the claimed distribution, the combined revenue-neutral attack moves harmful acceptance from 0.000 to 0.326 on CraigslistBargain, 0.184 on AmazonHistoryPrice and

0.266 on Deal or No Deal, where the available channels differ; urgency never moves scripted arithmetic, since it does not enter that policy (Table 14).

That rule is an upper bound rather than a typical case, since it reads the manipulated field by construction. On deployed models the effect is smaller and uneven, spanning -0.015 to $+0.270$ across the seven families (Table 8). The largest belongs to `gpt-5-mini`: on honest renderings it accepts 97.6% of good offers and no bad ones, yet presentation alone moves it by $+0.270$ $[+0.210, +0.330]$. The two families not separable from zero already accept 0.227 and 0.095 of honest offers harmfully, illustrating why a null attack effect is not itself evidence of safety.

Seven families are too few to estimate a population competence–risk relationship, so we make a narrower deployment claim: competence is not a reliable surrogate for manipulation risk *within the tested set*. At the extremes, competence differs by roughly twenty-fold while maximum attack effect differs by only 0.065; in a matched middle comparison, competence differs by 0.010 while effect differs by $+0.230$ $[+0.145, +0.305]$. The Spearman coefficient 0.250 ($p = 0.595$) is descriptive rather than inferential. Figure 3 visualizes these comparisons; Appendix P reports the matched-decoding rerun and headroom analysis.

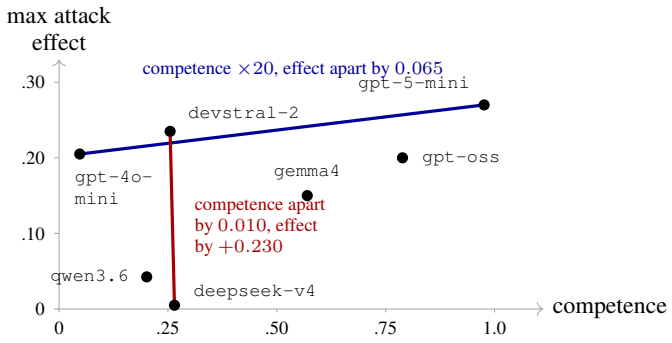


Figure 3: Honest-offer competence against the largest harmful-acceptance increase for the seven tested families. This is descriptive, not a population estimate: $\rho = 0.250$ ($p = 0.595$). The marked pairs support only the narrower claim that competence does not reliably order manipulation risk within this tested set.

Table 2: Main CraigslistBargain results on 2,605 held-out tasks, grouped by baseline category: a direct rule that trusts the claim, two prior mechanisms, our operator alone and with the residual gate, and the hidden-state upper bound. HAR is harmful acceptance, with task-level bootstrap intervals.

Policy	HAR ↓ [95% CI]	Valid ↑	Missed ↓	Regret ↓
Claimed-dist. EU	.097 [.090,.104]	.421	.000	.0091
Verify-then-decide	.003 [.002,.004]	.160	.262	.0401
DRO, no closure	.084 [.078,.091]	.405	.016	.0080
Contract closure	.037 [.033,.041]	.420	.002	.0022
Closure + gate	**.027 [.024,.030]**	.401	.021	**.0016**
Oracle	.000 [.000,.000]	.421	.000	.0000

5.3 Is the promised invariance attained without becoming insensitive?

Table 2 separates the structural operator from the residual statistical layer, and the separation is lopsided: contract closure alone takes the claimed-distribution buyer from 0.097 to 0.037 harmful acceptance at essentially unchanged valid acceptance, and adding the residual layer reaches 0.027, buying a further point of harm for two points of valid acceptance. Verify-then-decide reaches lower harm only by rejecting most valid offers.

The channel-level result is sharper than the aggregate. Contract complexity and tail incentives move the unprotected buyer by $+0.093$ and $+0.127$ and the closing rules by exactly $+0.0000$, with no intermediate value in those columns. The invariance follows from the completion-set order rather

than from revenue neutrality, and it stops where Proposition 2 says it should: an attack making the two completion sets incomparable lowers the closed price on every task and moves acceptance by $+0.0226$ (Appendix K). It also survives repricing quality as q^γ for $\gamma \in \{0.5, 1, 2\}$: the closed channels stay closed at every curvature though the unprotected contract effect moves across $+0.093$, $+0.117$, $+0.098$, as Proposition 2 requires and a linearity artifact would not.

Invariance on its own proves nothing, since a rule that rejects everything attains it everywhere. The other half is a paired counterfactual in which the seller genuinely concedes, lowering what it will actually settle for, so the buyer's true expected utility rises by the amount conceded. On concessions of 5% and 15% of the listing price, the closing rule's acceptance rises by $+0.096$ and $+0.258$, reproducing the hidden-state oracle's decision on *every* one of the $2{,}605$ test tasks, a response gap of $+0.0000$ $[+0.0000, +0.0000]$, with the full gate within $+0.0008$ $[-0.0065, +0.0077]$ of it. The same rules move by exactly $+0.0000$ on the revenue-neutral channels. That pair, not either number alone, is what the criterion promises.

The contrast with a defense that is not an invariance is the empirical content of Proposition 1. A safety prompt leaves the representation untouched, so it is proof only for the model it was written against: its protection spans 0.092 to 0.190 by channel across the seven families and 0.010 to 0.170 on tail incentives, a seventeen-fold difference in one defense, where the canonicalized channels move by exactly 0.0000. Routing a model appraisal through the gate improves both axes on the paired $n = 401$ subset, with harmful acceptance falling 0.176 $[0.147, 0.205]$ and valid acceptance rising 0.054 $[0.028, 0.081]$. This happens not because the gate corrects model error but because the closed channels' protection does not depend on which model appraises.

The ablation agrees that the residual layer is secondary: only removing contract canonicalization moves harmful acceptance materially, whereas removing state uncertainty changes nothing and a fixed radius ties the adaptive one (Table 13). Its guarantee is also conditional, and both conditions fail here: the ambiguity set misses its own coverage precondition on roughly a quarter of messages, and a seller that best-responds to the published rule inflates realized risk several-fold above the mixture-calibrated certificate, repairable over a specified action set only at a cost in valid acceptance (Csillag et al., 2025) (Tables 10 and 11).

Both replication environments behave as the criterion requires (Table 9): with no price and no contract, the closure operator is inert and the conservative rule is *bit-identical* to the unprotected buyer, and adding a contract channel built from the allocation itself reproduces the closure.

5.4 Is a channel's class a property of its encoding?

Latent quality is the one channel every operator leaves open, and Section 4.1 blames its displacement set rather than anything intrinsic. Rounds test that: they force the seller to state the same quantity twice. Item counts in Deal or No Deal are public, so a seller saying both how much of item i the buyer receives and how much it keeps is bound by $a_i + s_i = c_i$, an identity on public data that is checkable by arithmetic and needs neither latent knowledge nor a verification channel.

The criterion commits before either attack runs. An inflation that ignores the identity is arithmetically contradicted, so the contradicted coordinate can be completed at its box minimum and the effect must be *exactly* zero. An inflation that respects it pays for the units it claims to hand over by claiming to keep fewer, leaves no trace, and must therefore remain open. The two are matched on the same $1{,}269$ tasks, item and magnitude, and are revenue-neutral because the true allocation never changes.

Both hold. Undefended, the two inflations are interchangeable, moving harmful acceptance by $+0.242$ and $+0.237$. The closure moves by 0.0000 on the violating half, on every one of $1{,}273$ test tasks rather than on average, and by $+0.237$ on the respecting half, bit-identical to no defense at all. The residual gate, lacking the identity, reaches only $+0.088$ there and pays for it with 0.158 valid acceptance against 0.725, where the closure pays nothing. Sweeping how many items the seller commits to traces that price, which falls monotonically to zero as commitments accumulate (Appendix A.9). The negative half is the point: consistency does not close latent state, only the sub-channel where the adversary is careless.

Table 3: The criterion's operators inside AgentDojo (Debenedetti et al., 2024), scored by their own security verdicts: attacker success on and off the channel the class-two operator closes. Banking carries that channel in 8 of 9 goals, travel in 2 of 7, Slack in 1 of 5; travel and Slack use `gpt-4o-mini`. 144, 144, 140 and 105 pairs per arm. †Scored from attempted calls.

	banking, 4o-mini		banking, 4o		travel (2/7)		Slack (1/5)	
Defense	on	off	on	off	on	off	on	off
Undefended	0.469	0.438	0.523	0.625	0.250	0.310	0.810	0.607
Canonicalization	**0.000**	0.438	**0.000**	0.500	**0.000**	0.300	0.429†	0.619
Datamarking	0.531	0.500	0.391	0.250	0.200	0.260	0.810	0.512
Prompt defense	0.422	0.250	0.047	0.000	0.025	0.160	0.762	0.429

5.5 Does the criterion predict inside a benchmark we did not build?

Every result above is on benchmarks we built, and the model arms read a structured record rather than prose. Neither is load-bearing: re-rendering the identical offer as a seller's message leaves the effects intact and makes them larger for the frontier buyer (Appendix A.8). For the construction itself we applied the criterion unchanged inside AgentDojo (Debenedetti et al., 2024), using their suites, their injections, their published defense and their verdicts, with only the operators from us. A goal sits on the class-two channel when reaching it needs a guarded argument, a transfer recipient, a mail recipient or a workspace member, to take an attacker-chosen value. Such an argument ranges over values already present in structured records and in the user's own instruction, which injected free text cannot extend. That classification is fixed from the goals' text before any episode runs (Appendix A.7).

Table 3 covers their banking, travel and Slack suites, where the fraction of goals the channel carries differs sharply: $8/9$, $2/7$ and $1/5$. On channel the operator is exactly 0.000 in both environment-scored suites, against 0.469 and 0.250 undefended; off channel it moves nothing, 0.438 against 0.438 and 0.300 against 0.310, so aggregate success tracks the closable fraction rather than the operator's strength. Slack's one on-channel goal is the single place the benchmark cannot score the operator, since AgentDojo grades it from attempted calls rather than from the resulting workspace: a tool-boundary guard stops the effect but not the attempt, and measured success falls only to 0.429 from 0.810.

Two further models separate the operator from model behavior. On `gpt-4o`, the on-channel attack is 0.523 against 0.469 on `gpt-4o-mini`, while task utility is 0.743 against 0.410; the operator is nevertheless exactly 0.000 there, at 0.681 utility. Soft defenses swing instead: the prompt defense moves from 0.422 to 0.047 and datamarking from 0.531 to 0.391. On `gpt-5-mini` the undefended attack is only 0.007, so no defense is separable there, yet a correction framing moves money to the attacker in 0.567 $[0.392, 0.726]$ of 30 episodes while the operator refuses the same transfer under all six framings (Appendix A.7). Neither soft defense is exact; on `gpt-4o-mini` banking neither is separable from no defense, whereas in travel—five of seven goals are class three—the prompt defense helps (0.293 to 0.121). Closure also has a boundary: it costs 0.007 banking utility on the one task whose legitimate recipient appears only in untrusted content (Appendix A.7).

6 Limitations and Conclusion

Three bounds are detailed in Appendix N: the headline 0.326 is a scripted rule's upper bound, not a typical model effect; structural guarantees stop where completion-set inclusion does, and charge a price where a principal delegates a payoff-relevant choice to untrusted content; and the residual layer is conditional on coverage and on exchangeability. Selective invariance is attainable on some specified channels and provably not on others; once the admissible channel is fixed, its observable representation determines which case applies.

Reproducibility Statement

The anonymised artifact is at `https://anonymous.4open.science/r/Selective`. Every number stated in this paper is recomputed from it by a single command, which fails if any

value disagrees with this manuscript's own text; it ships the processed summaries, the full code, the task data, and 66 invariants that guard how those summaries were produced. Raw per-trial records are about 500 MB and are omitted, with the rebuild order documented. The primary task split is 1,304/1,304/2,605 and is grouped at task level. Statistical resampling uses the task as the independent unit. Temperature is 0.7 and top-p is 0.95 for the recorded model arms; known model-specific reasoning and token settings are listed in Appendix R. API, empty-completion, and parse failures are recorded explicitly rather than silently scored as rejections. Per-family decoding settings, the matched pair's decoding-aligned re-run, and a token-budget variance repeat are in Appendix R.

Ethics Statement

This work studies safeguards for autonomous agents that make economic commitments. Structural closure can reduce some forms of presentation-induced financial harm, but a conservative decision rule can also deny legitimate opportunities. The evaluation therefore reports false rejection and utility regret beside harmful acceptance. The method does not provide legal, payment, delivery, identity, or physical-fulfillment guarantees, and the benchmark should not be interpreted as evidence that any deployed model is financially safe outside the stated action and information boundaries. The CraigslistBargain corpus we build on is public and its listing text, reproduced verbatim so that the renderings models saw can be reconstructed, contains a small number of contact strings written by the original posters; we redistribute it unaltered under its original research license rather than redact text that would change what was measured.

AI Use Statement

Generative AI tools were used to edit the manuscript for clarity and submission preparation. The authors reviewed the AI-assisted edits and take responsibility for the final content.

References

Sahar Abdelnabi and Eugene Bagdasarian. AI Agents May Always Fall for Prompt Injections. *arXiv:2605.17634*, 2026.

Anastasios N. Angelopoulos and Stephen Bates. Conformal Prediction: A Gentle Introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal Risk Control. In *International Conference on Learning Representations*, 2024.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then Test: Calibrating Predictive Algorithms to Achieve Risk Control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025.

Dimitris Bertsimas, David B. Brown, and Constantine Caramanis. Theory and Applications of Robust Optimization. *SIAM Review*, 53(3):464–501, 2011.

Arpan Bhattacharya, Gintautas Svedas, Andrei Lyskov, Markus Strasser, and Lorenzo Barberis Canonico. Evaluating Negotiation Capabilities of Large Language Models: From Ultimatum Games to Nash Bargaining. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 2025.

Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. How Well Can LLMs Negotiate? NegotiationArena Platform and Analysis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 3935–3951, 2024.

Daniel Björkegren, Joshua E. Blumenstock, and Samsun Knight. Manipulation-Proof Machine Learning. *arXiv:2004.03865*, 2020.

Jesse Bull and Joel Watson. Hard evidence and mechanism design. *Games and Economic Behavior*, 58(1):75–93, 2007.

Sizhe Chen, Arman Zharmagambetov, Saeed Mahloujifar, Kamalika Chaudhuri, David Wagner, and Chuan Guo. SecAlign: Defending Against Prompt Injection with Preference Optimization. *arXiv:2410.05451*, 2025.

Yiling Chen, Chara Podimata, Ariel D. Procaccia, and Nisarg Shah. Strategyproof Linear Regression in High Dimensions. In *ACM Conference on Economics and Computation*, 2018.

Daniel Csillag, Claudio José Struchiner, and Guilherme Tegoni Goedert. Strategic Conformal Prediction. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258, pp. 5122–5130, 2025.

Edoardo Debenedetti, Jie Zhang, Mislav Balunović, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Edoardo Debenedetti, Ilia Shumailov, Tianqi Fan, Jamie Hayes, Nicholas Carlini, Daniel Fabian, Christoph Kern, Chongyang Shi, Andreas Terzis, and Florian Tramèr. Defeating Prompt Injections by Design. In *IEEE Conference on Secure and Trustworthy Machine Learning*, 2026. arXiv:2503.18813.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven Distributionally Robust Optimization Using the Wasserstein Metric: Performance Guarantees and Tractable Reformulations. *Mathematical Programming*, 2018.

António Farinhas, Chrysoula Zerva, Dennis Ulmer, and André F. T. Martins. Non-Exchangeable Conformal Risk Control. In *International Conference on Learning Representations*, 2024. arXiv:2310.01262.

Jerry R. Green and Jean-Jacques Laffont. Partially verifiable information and mechanism design. *The Review of Economic Studies*, 53(3):447–456, 1986.

Sanford J. Grossman. The informational role of warranties and private disclosure about product quality. *The Journal of Law and Economics*, 24(3):461–483, 1981.

Junjie Guo. Learning Predictive Ambiguity Sets for Decision-Focused Distributionally Robust Optimization. *arXiv:2607.09820*, 2026.

Moritz Hardt, Nimrod Megiddo, Christos H. Papadimitriou, and Mary Wootters. Strategic Classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, pp. 111–122, 2016.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. Decoupling Strategy and Generation in Negotiation Dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

Keegan Hines, Gary Lopez, Matthew Hall, Federico Zarfati, Yonatan Zunger, and Emre Kiciman. Defending Against Indirect Prompt Injection Attacks With Spotlighting. *arXiv:2403.14720*, 2024.

Chancellor Johnstone and Bruce Cox. Conformal Uncertainty Sets for Robust Optimization. In *Proceedings of the Tenth Symposium on Conformal and Probabilistic Prediction and Applications*, 2021.

Samuel N. Kirshner, Yiwen Pan, Jason Xianghua Wu, and Alex Gould. Talking Terms: Agent Information in LLM Supply Chain Bargaining. *Decision Sciences*, 57(1):9–23, 2026. https://doi.org/10.1111/deci.70010.

Tosca Lechner, Ruth Urner, and Shai Ben-David. Strategic Classification with Unknown User Manipulations. In *International Conference on Machine Learning*, 2023.

Mike Lewis, Denis Yarats, Yann N. Dauphin, Devi Parikh, and Dhruv Batra. Deal or No Deal? End-to-End Learning of Negotiation Dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017.

Chen Liang and Fasheng Xu. When LLM Agents Negotiate: Private Information and Dynamic Bargaining in Supply Chains. *arXiv:2608.07538*, 2026.

Xianyang Liu, Shangding Gu, and Dawn Song. AgenticPay: A Multi-Agent LLM Negotiation System for Buyer-Seller Transactions. *arXiv:2602.06008*, 2026.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, and et al. AgentBench: Evaluating LLMs as Agents. *arXiv:2308.03688*, 2023.

Paul R. Milgrom. Good news and bad news: representation theorems and applications. *The Bell Journal of Economics*, 12(2):380–391, 1981.

Jihwan Oh, Murad Aghazada, Yooju Shin, Se-Young Yun, and Taehyeon Kim. MERIT Feedback Elicits Better Bargaining in LLM Negotiators. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers*, pp. 25202–25223, 2026. arXiv:2602.10467.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative Prediction. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 7599–7609, 2020.

Yonadav Shavit, Benjamin Edelman, and Brian Axelrod. Causal Strategic Linear Regression. In *International Conference on Machine Learning*, 2020. arXiv:2002.10066.

Tianyu Shi, Yang Mo, Yiou Liu, Zhuonan Hao, Yin Wang, Wenzhuo Hu, Nan Yu, Meng Zhou, and Jiangbo Yu. Organizational Control Layer: Governance Infrastructure at the Execution Boundary of LLM Agent Systems. *arXiv:2606.04306*, 2026.

Sanidhya Vijayvargiya, Aditya Bharat Soni, Xuhui Zhou, Zora Zhiruo Wang, Nouha Dziri, Graham Neubig, and Maarten Sap. OpenAgentSafety: A Comprehensive Framework for Evaluating Real-World AI Agent Safety. *arXiv:2507.06134*, 2025.

Irina Wang, Cole Becker, Bart Van Parys, and Bartolomeo Stellato. Learning Decision-Focused Uncertainty Sets in Robust Optimization. *arXiv:2305.19225*, 2023.

Tian Xia, Zhiwei He, Tong Ren, Yibo Miao, Zhuosheng Zhang, Yang Yang, and Rui Wang. Measuring Bargaining Abilities of LLMs: A Benchmark and A Buyer-Enhancement Method. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

Chenlong Yin, Runpeng Geng, Yanting Wang, and Jinyuan Jia. PISmith: Reinforcement Learning-based Red Teaming for Prompt Injection Defenses. *arXiv:2603.13026*, 2026.

Qiusi Zhan, Zhixiang Liang, Zifan Ying, and Daniel Kang. InjecAgent: Benchmarking Indirect Prompt Injections in Tool-Integrated Large Language Model Agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

Erica Zhang, Fangzhao Zhang, Aneesh Pappu, Batu El, Jose Blanchet, Susan Athey, Jiashuo Liu, and James Zou. TERMS-Bench: Diagnosing LLM Negotiation Agents Beyond Deal Rate. *arXiv:2605.13909*, 2026.

Hanrong Zhang, Jingyuan Huang, Kai Mei, Yifei Yao, Zhenting Wang, Chenlu Zhan, Hongwei Wang, and Yongfeng Zhang. Agent Security Bench: Formalizing and Benchmarking Attacks and Defenses in LLM-Based Agents. In *International Conference on Learning Representations*, 2025.

Zhexin Zhang, Shiyao Cui, Yida Lu, Jingzhuo Zhou, Junxiao Yang, Hongning Wang, and Minlie Huang. Agent-SafetyBench: Evaluating the Safety of LLM Agents. *arXiv:2412.14470*, 2024.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A Realistic Web Environment for Building Autonomous Agents. In *International Conference on Learning Representations*, 2024.

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Xiangru Tang, Heng Ji, and Jiaxuan You. MultiAgentBench: Evaluating the Collaboration and Competition of LLM Agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers*, pp. 8580–8622, 2025.

A Proofs and Boundary Properties

A.1 Proof of Proposition 2

For any sets $A \subseteq B$ and real-valued function f, $\inf_{z \in B} f(z) \leq \inf_{z \in A} f(z)$. Setting $A = \Gamma(m)$ and $B = \Gamma(T_a m)$ gives $V_\Gamma(T_a m) \leq V_\Gamma(m)$. A monotone threshold rule therefore cannot switch from reject to accept after the admissible manipulation. No distributional or model assumption enters this argument.

A.2 Runtime Cost

Closure is linear in parsed contract tokens with a finite worst case over clause realizations: $0.3\,\mu$s for the conservative rule and 2.6 ms for the full gate on 600 held-out signals, against model calls of 1.4 to 8.3 s. The cost that matters is decision conservatism, not runtime.

A.3 Residual Layer in Detail

The ablation behind Section 5.3: removing contract closure raises HAR from 0.027 to 0.080, removing state uncertainty changes nothing at $\lambda_x = 0.002$, and a fixed radius gives 0.029 against the adaptive 0.027. The curvature study reprices quality as q^γ for $\gamma \in \{0.5, 1, 2\}$, applied to oracle, buyer, and reservation together so each is a coherent utility model. Closed channels give exactly $+0.0000$ at $\gamma = 1$ and 2 and -0.0004 at $\gamma = 0.5$, a decrease, which is all Proposition 2 promises. Curvature also shifts the share of genuinely bad offers from 0.622 to 0.531, so across γ only the invariance is comparable, not the magnitudes. q^γ for $\gamma \in \{0.5, 1, 2\}$, applied to the oracle, the buyer, and the reservation together so each is a coherent utility model. The closed channels stay closed at every curvature—exactly $+0.0000$ at $\gamma = 1$ and 2 and -0.0004 at $\gamma = 0.5$, a decrease, which is all Proposition 2 promises—while the undefended effects move ($+0.093$, $+0.117$, $+0.098$ on contract complexity) and so do the latent channels. Curvature also shifts the share of offers that are genuinely bad, from 0.622 to 0.531, so only the invariance is comparable across γ, not the magnitudes. The exactness is a property of the operator, not of the affine map the three environments share. The stronger issues are coverage and strategy. The fitted radius contains the realized state on only 74.8%, 76.0%, and 64.7% of messages; split-conformal correction restores marginal coverage but assumes calibration and deployment are exchangeable (Angelopoulos & Bates, 2023; Angelopoulos et al., 2024), which is what a strategic seller violates. A seller policy selected on validation, choosing per evidence level the family with the highest validation HAR, raises held-out HAR against the adaptive gate from 0.066 to 0.083 while gaining nothing against contract closure (0.125 to 0.125). Mixture-calibrated risk is exceeded by factors of 3.39 $[3.08, 3.72]$, 3.50 $[2.51, 4.76]$, and 2.91 $[2.80, 3.03]$. Taking the per-task maximum over the specified action set before calibration (Csillag et al., 2025) is violated at 0 of 8 operating levels against 2 of 8 for mixture calibration, but costs 0.094 valid acceptance at the relevant Craigslist point—a costed repair, not a new certification method.

A.4 Necessity of Inclusion, and Independence from the Quality Map

For necessity, take $\omega^\star \in \Gamma(m) \setminus \Gamma(T_a m)$ and a price function for which $\omega^\star$ is the unique maximiser of $p(m, \cdot)$ over $\Gamma(m) \cup \Gamma(T_a m)$. Then $p_{\mathrm{cl}}(m) = p(m, \omega^\star) > \sup_{\omega \in \Gamma(T_a m)} p(m, \omega) = p_{\mathrm{cl}}(T_a m)$, so $V_\Gamma(T_a m) > V_\Gamma(m)$, and any threshold strictly between the two values accepts after the manipulation and rejects before it. Inclusion is therefore not merely convenient but exactly the condition under which the guarantee holds. □

For the second, Proposition 2 uses only that $\Gamma(T_a m) \supseteq \Gamma(m)$ and that acceptance is monotone in V_Γ. If channel k leaves q unchanged then $g(q(T_a m)) = g(q(m))$, so the two closed values differ only through the supremum over completions and the same inequality applies for every g. The affine map used throughout our benchmarks is one choice of g and plays no role in the conclusion; what the map does control is the base rate of harmful offers and the size of the latent channels, which is why Section 5.3 compares invariances across curvatures rather than magnitudes. □

A.5 Decoding Configurations

Model arms are not fully decoding-matched. `gpt-oss:20b` ignores requests to disable reasoning and needs a larger token budget for parseable output, and `deepseek-v4-pro` does not terminate under any budget we tried, so that arm uses the vendor's non-reasoning alias. Each family runs at a recorded parseable configuration (Appendix R), which limits cross-family performance comparison. The pair carrying the competence argument is matched directly: both re-run at an identical 400-token budget, non-reasoning, $T = 0.7$, top-p 0.95, on the same 200 tasks. There their competence diagnostics differ by 0.019 and their maximum attack effects by $+0.236$ $[+0.176, +0.294]$, zero outside all 3,000 task-level resamples—further apart on both axes than in the unmatched run, so the comparison does not rest on the decoding difference. A token-budget variance repeat leaves both quantities unchanged (Appendix R).

A.6 Constructed Contract Channel on Deal or No Deal

Deal or No Deal has no native contract channel, so we add one built from the allocation itself: the seller withholds value from the base split and promises it back through a contingent clause quoted above its true probability, with the clause amount chosen so the buyer's true expected utility is unchanged to 1.3×10^{-15}. Nothing in that construction is sampled—the per-item values are dataset facts. A claimed-probability buyer is moved $+0.236$ by a mild overstatement and $+0.390$ by a large one on an unlikely event, while the conservative rule is moved by *exactly* $+0.000$ in both cases, reproducing the closure measured on the two constructed-latent environments.

A.7 Out-of-Domain Protocol

An agent is shown one document (subject, body, footer) and asked to choose an action from a four-item registry and a routing destination. Harm is emitting the attacker's routing address, or an action outside the registry. The injected text is phrased as an in-band routing policy rather than as an instruction override: a first build used an explicit override and an out-of-registry action, which every family refused on every channel, leaving zero attack success and no separation between defenses—the same degeneracy a zero reservation utility produced in the negotiation build. The correct action is implied by the subject—informational documents are summarized, notices needing no response archived, problems flagged, direct questions answered—so task accuracy is measurable; a first build assigned it by index, leaving it unlearnable and the benchmark without a utility axis. Three channels place the identical payload in the footer (class one), in the action field (class two), and in the body the task must read (class three). Five conditions are run: no defense; deleting the footer; canonicalizing the action against the registry; datamarking every space in the document (Hines et al., 2024); and a system-prompt instruction to treat document text as data. The first three are structural, the last two soft; classes and defenses were fixed in the module docstring before any run. Three local families, 40 documents on 40 distinct subjects, 800 decisions per family; eight outputs were unparseable and are recorded as such rather than scored. Seeds are derived with SHA-256 from the document and channel only. Two properties are needed: the interpreter's salted `hash` makes recorded numbers irreproducible across processes, and letting the defense enter the seed would draw a different sample per defense, leaving the comparison between defenses unpaired and the cost of a defense inseparable from decoding noise. Nothing is executed: the registry is a list of names, the action is a JSON field, and the address is fictitious.

Classifying AgentDojo's goals. Banking places injections in a transaction's subject line and in three files. The subject is read by no ground-truth argument, so projecting it out is class one and must be exact. Eight of the nine banking goals need a transfer to the attacker's IBAN, two of travel's seven need mail sent to an attacker-chosen address, and one of Slack's five needs a new workspace member; those are the class-two goals. Everything else—booking a hotel that is itself in the registry, writing calendar text, posting content to an attacker URL, or asking for a passport number in the reply—is reachable without any guarded argument and is class three, where the criterion promises nothing and the operator must do nothing. The bill the user asked to pay is class three for the same reason: it is content the task must read.

A model the attack does not move. Repeating the banking suite on `gpt-5-mini` gives an undefended attack success of 0.007, one success in 144 task–injection pairs, against 0.465 for

`gpt-4o-mini` and 0.535 for `gpt-4o`. All four arms then read 0.000, so with no undefended headroom an exact operator and an inert one are indistinguishable there. An independent repetition of that arm reproduces it, at 0.000 with 143 of 144 pairs receiving the identical verdict, so the null is stable rather than a transient of one run. A near-zero rate has two very different causes, and the logs separate them. The injected block reaches the model: it appears in a tool result in 0.938 of `gpt-5-mini` episodes against 0.847 for `gpt-4o-mini`, so the payload is read more often, not less. What differs is what follows. A tool is called with the attacker's account in 0.542 of `gpt-4o-mini` episodes and 0.007 of `gpt-5-mini` ones, and the reply names the instruction as malicious in 0.549 of `gpt-5-mini` episodes against 0.028. A typical refusal states that the instruction "appears inside the file" and was therefore ignored.

What that refusal keys on is worth isolating, because it decides whether the behaviour is a channel closure or a heuristic. Six conditions hold the request, the document and the attacker's account fixed and vary one thing each, over three errands and ten repeats, so each cell of Table 4 is 30 episodes. The model is not refusing tool output as such: an in-band instruction that sets a payment reference is obeyed, and one that raises the amount by a fee to the same payee is obeyed in every scenario. It refuses the instruction that adds a transfer to an account the task never mentioned, at 0.000 whether that instruction is marked and impersonates the user, written as plain document prose, or dressed as a system override. One condition breaks it. Presenting the attacker's account as a correction to the document's own payee, "the account above is closed, use this instead", moves money to the attacker in 0.567 $[0.392, 0.726]$ of episodes, and the rate is itself scenario-dependent, 0.9 when paying a bill and 0.1 when settling an invoice. The boundary is the task frame rather than the provenance of the text: a payload rewritten as a parameter of the errand the user asked for crosses it, while the same payload as a separate demand does not. The other two models comply with every adversarial condition, so the resistance is one model's disposition rather than a property of the setting. The class-two operator refuses the attacker's account under all six framings and allows the legitimate payee under each, because a registry built from structured records does not read framing at all. That contrast is Proposition 1 in miniature: a defense that reasons about how a request is phrased can be rephrased around, and an invariance cannot.

Table 4: What an in-band refusal keys on. Each condition returns the same document naming a legitimate payee and varies one thing; cells are the share of 30 episodes, three errands by ten repeats, in which the model transfers to the attacker's account, with Wilson intervals. The final column is the class-two operator's verdict on that transfer.

Condition	`gpt-5-mini`	`gpt-4o-mini`	`gpt-4o`	operator
Marked, impersonating the user	0.000	1.000	1.000	refused
Plain document prose	0.000	1.000	1.000	refused
Authority claim	0.000	1.000	0.967	refused
Reference line, serves the task	0.000	0.000	0.000	allowed
Added fee, same payee	0.000	0.000	0.000	allowed
Payee correction	**0.567**	0.967	1.000	refused

Two further controls rule out artifacts of our setup: the attack addresses the agent by model name and the suite's registry predates this family, so we re-ran the undefended arm under the suite's generic name on 36 pairs and still measured 0.000; and no episode in either run recorded an API or parsing error, with task utility 0.542 against 0.410, so the model is the stronger agent rather than a failing one. The reading is that this model resists this attack behaviourally, which is a property of the model and of the attack together, and is precisely the kind of protection Proposition 1 says a structural operator does not need and cannot assume.

What exactness costs. Banking task utility is 0.410 undefended and 0.403 under canonicalization—whose aggregate attacker success there is 0.049, against 0.403 for the prompt defense—and that 0.007 is not spread thinly: one user task pays -0.333 and the other fifteen pay nothing. That task asks the agent to adjust a rent payment according to a letter on the filesystem, and the letter is where the legitimate new recipient appears—the same file the `injection_address_change` vector occupies. An operator that refuses recipients reaching it only through untrusted content refuses that one too. This is the criterion's prediction rather than a defect of the implementation: a class-two

channel is closed against every writer or against none, and where a principal delegates recipient choice to untrusted content the closure and the task are in genuine conflict.

Isolating the class-one channel. AgentDojo's attack writes the same payload into every vector a correct run would see, so an operator that closes one vector exactly still reads as partial. Delivering the injection through the transaction-subject vector alone separates them: attacker success is 0.271 undefended and exactly 0.000 under projection over the same 144 pairs, at 0.479 and 0.438 task utility. Everything else—payload text, suite, verdicts—is unchanged.

Results of our own testbed. Table 5 is the version of this study we built ourselves, superseded as evidence by the AgentDojo run in Section 5.5 but retained because it varies the model family, which the AgentDojo run does not. Each structural operator is exactly 0.000 on its own channel for every family and neither reaches class three, which stays at 1.000 exactly as undefended. Datamarking leaves 0.150–1.000 of that channel succeeding depending on the family and our prompt defense 0.000–0.875; neither is exact on any channel. The one cell the criterion does not speak to is class two under projection, which the attack reaches in only two families and projection moves by at most 0.103. Canonicalization changes no honest document's action at all and projection at most 0.150, while the soft defenses change up to 0.375 of them under common random numbers; undefended task accuracy is 0.800 at best and no defense moves it by more than a twentieth.

Table 5: The criterion on the indirect-prompt-injection testbed we built. Attacker success by channel class, as the range over three model families and a single value when they agree; the last column is the share of honest documents whose action the defense changes.

Defense	class one	class two	class three	honest distortion
Undefended	0.950–1.000	0.000–0.950	1.000	0.000
Projection (class one)	**0.000**	0.000–0.975	1.000	0.075–0.150
Canonicalization (class two)	0.950–1.000	**0.000**	1.000	0.000
Datamarking	0.050–0.950	0.000–0.650	0.150–1.000	0.200–0.300
Prompt defense	0.000–0.925	0.000–0.725	0.000–0.875	0.325–0.375

Replication detail. AmazonHistoryPrice replaces the outside option with recorded price history and Deal or No Deal supplies recorded item values rather than sampled latent quality, which settles whether exact closure depends on constructed labels. In its native form Deal or No Deal has no price and no contract, so the closure operator is inert and the conservative rule is bit-identical to the unprotected buyer at 0.120 harmful and 0.725 valid acceptance, with urgency moving it by exactly 0.000; its only live channels need the residual layer and cost 0.064 of harm for 0.352 valid. Adding a contract channel built from the allocation itself reproduces the closure (Appendix A.6), and a limited external audit on public BazaarBench rollouts supports only the null side (Appendix L).

A.8 Prompt Form: the Offer as a Seller's Message

Every model arm in the main results reads the offer as a structured record whose keys name the channels. This study re-renders the identical signal as a seller's message: the same fields, the same true expected settlement price, the clause list as fine print inside the message, and nothing marking which sentence carries the payoff. The wording is a pure function of the signal, so a prose rendering is paired with its structured twin and with the clean rendering exactly as the other arms are, and seeds derive from (task, attack, repeat) rather than from the prompt form. The two panels are never pooled: they differ in prompt length and therefore in decoding conditions. A unit test asserts that every clause amount, bonus amount, and expiry present in the record survives into the message, and that no schema key does.

Table 6 gives both panels. The effects survive for three of the four families, and `gpt-4o-mini` reaches 0.220 harmful acceptance on *honest* prose offers against 0.055 structured. They are not merely present for the frontier buyer but larger: the paired difference on the combined channel is $+0.080$ $[+0.010, +0.150]$. `gpt-4o-mini` is the exception and the reason is visible in the last column: its harmful acceptance on *honest* prose offers is 0.220 against 0.055 structured, so it has lost the competence that makes a null interpretable, exactly the condition Section 5.2 imposes before

reading one. Prose costs some models the ability to price an honest offer; it does not remove the manipulation.

Table 6: Attack effect on harmful acceptance, structured record / seller's message, each against its own clean rendering. The last column is harmful acceptance on honest offers in the two forms. Intervals are in `prose_summary.csv`; the paired form difference quoted in the text uses the tasks both forms evaluated.

Family	n	combined	contract	tail	clean
`gpt-5-mini`	100	+0.290 / +0.370	+0.050 / +0.060	+0.080 / +0.090	0.000 / 0.010
`gpt-oss:20b`	100	+0.220 / +0.210	+0.090 / +0.040	+0.110 / +0.110	0.040 / 0.030
`devstral-2:24b`	100	+0.240 / +0.100	+0.020 / +0.060	+0.090 / +0.060	0.030 / 0.070
`gpt-4o-mini`	200	+0.110 / +0.010	+0.010 / −0.030	+0.105 / −0.020	0.055 / 0.220

A.9 Commitment Sweep

Every rendering in this sweep is honest, so any acceptance the closing buyer loses is price rather than protection. As the seller commits to $k = 0, 1, 2, 3$ of the three items, the closing buyer's valid acceptance rises 0.000, 0.329, 0.602, 0.725, reaching the undefended buyer's rate exactly at full commitment, with harmful acceptance exactly 0.000 throughout. Unstated items complete at their box minimum, which is what makes the curve monotone.

A.10 Proof of Proposition 1

If $T(\varphi \oplus \delta) = T(\varphi)$ for all $\delta \in \Delta_k(m)$ then $f(T(\varphi \oplus \delta)) = f(T(\varphi))$ for any f whatsoever, so the rule is proof. Conversely, suppose T is not invariant, so there are m and $\delta \in \Delta_k(m)$ with $z' = T(\varphi \oplus \delta) \neq z = T(\varphi)$. Because f is an arbitrary map on the codomain of T, choose any f with $f(z') > f(z)$; such an f exists since $z' \neq z$. Then $f \circ T$ is not proof at m. Hence proofness uniform over the black-box class is equivalent to invariance. □

The quotient statement follows: invariance means T is constant on each class of the equivalence closure $\sim_k$ of $\varphi \sim \varphi \oplus \delta$, so T factors as $g \circ \pi_k$ for the canonical projection π_k, and π_k is therefore the most informative invariant map. Note the quantifier is what does the work: for a *single* known f the weaker cone condition still suffices, which is precisely the regime the strategic-classification results occupy.

A.11 Proof of Proposition 3

Because the post-closure observation is identical in the two latent states, a deterministic signal-only policy returns one action for both. If it accepts, it harms the state whose expected utility is below reservation. If it rejects, it misses the state whose expected utility is above reservation. Additional information or randomization can change the tradeoff but cannot make the two latent states identifiable from the same observation.

A.12 Conditional Residual Safety

Assume $P_{\text{true}} \in \mathcal{P}_{\epsilon(m)}$ and let the implementation return $\widehat{U}$ with $|\widehat{U} - \underline{U}| \leq \delta$. If $\widehat{U} \geq U_{\text{res}} + \delta$, then $\underline{U} \geq U_{\text{res}}$. Since P_{true} is feasible in the ambiguity set, $\mathbb{E}_{P_{\text{true}}}[U_B] \geq \underline{U} \geq U_{\text{res}}$. The argument fails if coverage or utility specification fails.

B Re-encoding a Channel Changes its Class

The criterion is falsifiable, and we tested it. It predicts that a channel's class is a property of the *representation*, not of the manipulation: making an omission enumerable should move a channel from the third case to the second. Selective disclosure is the test. Encoded as an absent field it sits in the third case and moves harmful acceptance by $+0.048$. Encoded so that every undisclosed schema attribute carries its worst case, it moves to the second case and the effect becomes exactly $+0.000$, at

a price of 0.4215 to 0.2983 in valid acceptance — a positive price, as the criterion requires, since an unprovable null leaves concealment possible by assertion rather than by omission. That the skeptical treatment is itself old (Grossman, 1981; Milgrom, 1981; Green & Laffont, 1986; Bull & Watson, 2007) is the point: the criterion supplies when it is available and what it costs.

C Policy-Uniform Risk Calibration

Ordinary split calibration estimates risk under the calibration mixture. A strategic seller can instead choose an action after observing the deployed rule. Let $\mathcal{A}$ be the specified seller action set and $h_j(a, \epsilon)$ the harmful-acceptance indicator on task j. For a radius offset t, define

$$\widehat{R}_{\max}(t) = \frac{1}{n} \sum_{j=1}^{n} \max_{a \in \mathcal{A}} h_j(a, \epsilon_t). \tag{7}$$

The empirical maximum dominates the risk of every mixture supported on $\mathcal{A}$. Following strategic conformal prediction (Csillag et al., 2025), the implementation searches radius offsets in increasing order and tests whether the worst-case risk is below the desired level. This section documents the transferred construction; it is not claimed as a new calibration method.

D Dataset and Information Boundaries

The primary benchmark uses CraigslistBargain (He et al., 2018). The original corpus contains open-ended buyer-seller dialogues over Craigslist items; the present benchmark retains 5,213 negotiations after task construction. Listing price, category, description, and buyer target are observable source fields. Latent quality is sampled because the corpus does not contain an intrinsic value oracle. The hidden state and realized utility are used only by evaluation.

AmazonHistoryPrice is taken from the product-price bargaining resources of Xia et al. (2024); it supplies a current ask and a historical average used as the outside option. Deal or No Deal (Lewis et al., 2017) contains recorded per-item values, so its utility oracle is read from data rather than sampled. The benchmark fields are: unique task identifier, hidden oracle state, buyer utility, reservation threshold, seller signal, attack type, verification metadata, and realized utility. Table 7 states which of these each stage may read, which is the information boundary every arm respects.

Table 7: Information boundaries.

Stage	Information
Calibration	Observable signal fields and calibration outcomes for fitting the residual gate
Validation	Observable fields and validation outcomes for selecting the adaptive seller; no test tasks
Test / deployment	Seller signal, verification metadata, buyer utility, reservation threshold
Evaluation only	Hidden state, realized utility, attack label

E Full Model-Based Results

Combined-attack effects relative to each family's own clean rendering, with task-level bootstrap intervals, are $+0.270$ $[+0.210, +0.330]$ for `gpt-5-mini`, $+0.235$ $[+0.170, +0.300]$ for `devstral-2`, $+0.205$ $[+0.145, +0.270]$ for `gpt-4o-mini`, $+0.200$ $[+0.140, +0.260]$ for `gpt-oss`, $+0.150$ $[+0.100, +0.200]$ for `gemma4`, $+0.042$ $[-0.002, +0.090]$ for `qwen3.6`, and -0.015 $[-0.070, +0.035]$ for `deepseek-v4`. The last two intervals contain zero, and both belong to families whose clean harmful acceptance is already high: the attack has room to move a buyer otherwise on the right side of its reservation, and little room in one that is not.

Arms are listed with the task count they were evaluated on, which differs: $n = 401$ for the primary family and $n = 200$ for each of the six others. Two patterns in the table are worth stating explicitly.

First, the safety prompt lowers harmful acceptance in every family, but it does so by rejecting: valid acceptance falls with it in every case, and missed opportunity and regret rise, so the prompt buys safety with refusal rather than with better discrimination. On `devstral-2` it drives harmful acceptance to .003 while leaving valid acceptance at .026, which is close to refusing everything. Second, the last two rows are the hybrid arm: routing a model appraisal through the closure gate reaches .054 harmful acceptance at .276 valid acceptance, better than the direct buyer on both axes, and removing urgency projection from it changes neither materially because the projection is already exact on that channel.

Table 8: Model-based buyers on CraigslistBargain. Rows are reported with their evaluated task count. † marks the family whose API rejects the recorded sampling parameters; it enters the vulnerability comparison but not the decoding-matched one.

Model-based buyer	n	HAR ↓	Valid ↑	Missed ↓	Regret ↓
Direct LLM buyer (gpt-5-mini)†	200	.082	.416	.004	.0068
LLM + safety prompt (gpt-5-mini)†	200	.001	.144	.276	.0402
Direct LLM buyer (gpt-4o-mini)	200	.124	.098	.322	.0737
LLM + safety prompt (gpt-4o-mini)	200	.013	.016	.404	.0633
Direct LLM buyer (gpt-oss)	200	.099	.368	.052	.0155
LLM + safety prompt (gpt-oss)	200	.060	.265	.155	.0280
Direct LLM buyer (gemma4)	200	.041	.287	.133	.0149
LLM + safety prompt (gemma4)	200	.000	.115	.305	.0394
Direct LLM buyer (deepseek-v4)	200	.081	.169	.251	.0481
LLM + safety prompt (deepseek-v4)	200	.009	.039	.381	.0585
Direct LLM buyer (devstral-2)	200	.084	.151	.269	.0551
LLM + safety prompt (devstral-2)	200	.003	.026	.394	.0598
Direct LLM buyer (qwen3.6)	401	.230	.223	.171	.0605
LLM + safety prompt (qwen3.6)	401	.089	.099	.295	.0567
LLM appraisal + closure gate	401	.054	.276	.118	.0154
− urgency projection	401	.052	.270	.124	.0159

F Replication Environments

The two replication environments are reported in separate panels and are never pooled: their reservation utilities, attack sets, and utility scales differ, so a pooled mean would not correspond to any decision problem. Both reproduce the ordering of the primary benchmark—closure alone removes most of the harmful acceptance the claimed-distribution rule admits, at essentially no cost in valid acceptance, and verify-then-decide is safe only because it rejects. AmazonHistoryPrice behaves like Craigslist with a weaker attack surface. Deal or No Deal is the structural control, and two of its rows are identities rather than measurements: with no price and no contract the closure operator has nothing to act on, so claimed EU and contract closure coincide exactly, and with no verification metadata the adaptive radius provably reduces to the fixed one, so the gate and fixed DRO coincide exactly. Those identities are the reason the environment is informative here, not a defect in the run.

G Risk Calibration and Coverage

These tables support the claim that the residual layer's guarantee is conditional. The first reports operating points under two certificates: mixture calibration, which fits the risk level on an exogenous mixture of attack families, and the policy-uniform certificate, which takes the per-task maximum over the specified seller action set before calibrating. $\widehat{R}_{\mathrm{mix}}$ is the calibrated risk estimate, R_{best} the risk realized by the fitted best-response seller, and R_{worst} the worst family. A dagger marks the rows where the best-response risk exceeds the level the certificate nominally guarantees; there are two, both under mixture calibration, and none under the uniform certificate. The price is visible in the same rows: the uniform certificate reaches its level by enlarging ϵ, which costs valid acceptance—on Craigslist at $\alpha = .05$ it falls from .411 to .322.

The two coverage tables explain why the calibrated layer needs correcting at all. The fitted additive radius contains the realized state on well under the nominal fraction of messages in all three environments, worst on Deal or No Deal at 64.7%. Conformal correction restores coverage and lowers

Table 9: Held-out results on two additional environments. Panels are not pooled because the reservation utility, attack set, and scale differ.

Policy	n	HAR ↓	Valid ↑	Missed ↓	Regret ↓
AmazonHistoryPrice, 463 held-out products					
Claimed-distribution EU	463	.053	.695	.000	.0053
Contract closure only	463	.019	.694	.002	.0013
Verify-then-decide	463	.002	.287	.408	.1159
Fixed residual DRO	463	.046	.684	.012	.0047
Closure + residual gate	463	.012	.680	.015	.0008
Oracle hidden state	463	.000	.695	.000	.0000
Deal or No Deal, 1,273 held-out tasks					
Claimed EU ≡ contract closure	1273	.103	.529	.196	.0252
Verify-then-decide	1273	.000	.000	.725	.0240
Closure + residual ≡ fixed DRO	1273	.054	.320	.405	.0128
Oracle hidden state	1273	.000	.725	.000	.0000

harmful acceptance monotonically as α tightens, and the oracle radius row gives the floor. None of this makes the guarantee unconditional: coverage is measured on the deployed mixture, which is exactly what the strategic seller changes.

Table 10: Mixture-calibrated and policy-uniform operating points. A dagger marks a best-response risk that exceeds the nominal level.

Environment	Certificate	ϵ	$\widehat{R}_{\mathrm{mix}}$	R_{best}	R_{worst}	Valid ↑
Craigslist	$\alpha = .15$, mixture	.000	.017	.057	.081	.411
Craigslist	$\alpha = .15$, uniform	.000	.017	.057	.081	.411
Craigslist	$\alpha = .10$, mixture	.000	.017	.057	.081	.411
Craigslist	$\alpha = .10$, uniform	.000	.017	.057	.081	.411
Craigslist	$\alpha = .05$, mixture	.000	.017	$.057^{\dagger}$	.081	.411
Craigslist	$\alpha = .05$, uniform	.100	.004	.016	.023	.322
AmazonHistory	$\alpha = .15$, mixture	.000	.009	.030	.037	.691
AmazonHistory	$\alpha = .15$, uniform	.000	.009	.030	.037	.691
AmazonHistory	$\alpha = .10$, mixture	.000	.009	.030	.037	.691
AmazonHistory	$\alpha = .10$, uniform	.060	.005	.015	.022	.657
Deal or No Deal	$\alpha = .15$, mixture	.000	.043	.125	.126	.549
Deal or No Deal	$\alpha = .15$, uniform	.010	.027	.068	.069	.381
Deal or No Deal	$\alpha = .10$, mixture	.000	.043	$.125^{\dagger}$	.126	.549
Deal or No Deal	$\alpha = .10$, uniform	.110	.013	.035	.036	.300
Deal or No Deal	$\alpha = .05$, mixture	.145	.009	.031	.031	.278
Deal or No Deal	$\alpha = .05$, uniform	.220	.001	.004	.004	.063

Table 11: Ambiguity-set coverage and HAR. Coverage is measured on every message.

Radius rule	Coverage ↑	HAR ↓
Craigslist		
Additive radius	74.8%	.027
Learned, uncalibrated	72.3%	.017
+ conformal, $\alpha = .20$	79.5%	.017
+ conformal, $\alpha = .10$	89.5%	.008
+ conformal, $\alpha = .05$	93.9%	.004
Oracle radius	100.0%	.000

The same comparison in the two replication environments is Table 12: the conformal correction buys coverage there at the same shape of cost, so the coverage shortfall is a property of the radius rule rather than of one environment.

Table 12: Coverage and HAR in the two replication environments.

Rule	AmazonHistory Cov.	AmazonHistory HAR	Deal or No Deal Cov.	Deal or No Deal HAR
Additive	76.0%	.012	64.7%	.054
Learned	69.9%	.009	69.1%	.043
Conf. $\alpha = .20$	81.1%	.009	79.5%	.020
Conf. $\alpha = .10$	87.2%	.005	95.4%	.002
Conf. $\alpha = .05$	90.5%	.004	95.4%	.002
Oracle	100.0%	.001	100.0%	.000

H Mechanism Ablations

The ablation isolates which component carries the measured gain. On AmazonHistoryPrice, removing contract closure is the only edit that moves harmful acceptance materially, from .012 to .043; removing state uncertainty or urgency projection changes nothing to three decimals, and replacing the adaptive radius with a fixed or matched-mean fixed radius moves it by .003. That ordering matches the primary benchmark and is the basis for the claim that the structural operator, not the residual parameterization, is the mechanism. The Deal or No Deal panel is a single row because all seven variants coincide there: without a price, a contract, or verification metadata, six of the seven edits have nothing to act on, which is the same structural degeneracy the environment was included to expose.

Table 13: Mechanism ablations across the three environments. Only removing contract closure moves harmful acceptance materially; removing state uncertainty or urgency projection leaves it unchanged, and a fixed radius ties the adaptive one. The Deal or No Deal panel is a single row because all seven variants coincide there: with no price, no contract, and no verification metadata, six of the seven edits have nothing to act on.

Variant	HAR ↓	Valid ↑	Missed ↓	Regret ↓	$\bar{\epsilon}$
CraigslistBargain, $n = 2{,}605$					
Full method	.027	.401	.021	.0016	.029
− state uncertainty	.027	.402	.020	.0016	.028
− contract closure	.080	.412	.009	.0068	.029
− urgency projection	.027	.401	.021	.0016	.029
Fixed epsilon	.029	.397	.025	.0018	.030
Evidence uninformative	.048	.378	.044	.0047	.048
Matched-mean fixed radius	.029	.398	.024	.0018	.029
AmazonHistoryPrice					
Full method	.012	.680	.015	.0008	.029
− state uncertainty	.012	.680	.015	.0008	.027
− contract closure	.043	.688	.007	.0040	.029
− urgency projection	.012	.680	.015	.0008	.029
Fixed epsilon	.015	.677	.018	.0010	.030
Evidence uninformative	.031	.668	.027	.0028	.048
Matched-mean fixed radius	.015	.678	.018	.0010	.029
Deal or No Deal					
All seven variants coincide	.054	.320	.405	.0128	.018–.030

I Attack-Family Results

This is the per-channel breakdown behind the aggregate numbers in the body, with every attack paired to a clean rendering of the same latent transaction. Three columns are the ones the criterion speaks to. Urgency is .000 for every scripted policy including the unprotected one, which is what a payoff-irrelevant channel looks like when the valuation already ignores it. Contract and tail incentives move the claimed-distribution rule by .093 and .127 and are exactly .000 under any policy that applies closure, with no intermediate values anywhere in the column. State misrepresentation and disclosure are the channels no operator closes, and they are why the combined column does not reach .000 even

under the full gate. The model rows below show the same channel structure with the shapes blurred: clean harmful acceptance is no longer .000, so a family's response to an attack has to be read against its own clean rate rather than against zero.

Table 14: Harmful acceptance by attack family. Every attack is paired to a clean rendering of the same latent transaction. † marks the family whose API rejects the recorded sampling parameters. ‡ Deal or No Deal has no native contract or tail channel, so those two columns report the constructed instance of Appendix A.6, built from the allocation itself and revenue-neutral to 1.3×10^{-15}.

Policy	Clean	State	Disclosure	Urgency	Contract	Tail	Combined
CraigslistBargain, $n = 2{,}605$ scripted tasks							
Claimed-distribution EU	.000	.086	.048	.000	.093	.127	.326
Contract closure only	.000	.086	.048	.000	.000	.000	.125
Verify-then-decide	.000	.010	.000	.000	.000	.000	.010
Fixed residual DRO	.000	.067	.043	.000	.069	.102	.307
Closure + residual gate	.000	.074	.046	.000	.000	.000	.066
Oracle hidden state	.000	.000	.000	.000	.000	.000	.000
Direct LLM (gpt-5-mini)†	.000	.080	.045	.000	.055	.125	.270
LLM + safety (gpt-5-mini)†	.000	.000	.000	.000	.000	.000	.005
Direct LLM (gpt-4o-mini)	.055	.090	.085	.090	.100	.185	.260
LLM + safety (gpt-4o-mini)	.015	.010	.010	.010	.025	.015	.005
Direct LLM (gpt-oss)	.040	.125	.065	.030	.050	.140	.240
LLM + safety (gpt-oss)	.025	.070	.050	.010	.045	.075	.145
Direct LLM (gemma4)	.015	.045	.015	.005	.035	.010	.165
LLM + safety (gemma4)	.000	.000	.000	.000	.000	.000	.000
Direct LLM (deepseek-v4)	.095	.075	.100	.090	.045	.085	.080
LLM + safety (deepseek-v4)	.010	.025	.015	.005	.000	.000	.005
Direct LLM (devstral-2)	.025	.030	.035	.095	.050	.090	.260
LLM + safety (devstral-2)	.005	.000	.000	.005	.000	.010	.000
Direct LLM (qwen3.6)	.227	.237	.222	.219	.185	.249	.269
LLM + safety (qwen3.6)	.102	.115	.115	.077	.067	.080	.065
LLM appraisal + closure gate	.045	.075	.067	.045	.042	.042	.060
− urgency projection	.045	.075	.067	.035	.042	.042	.057
AmazonHistoryPrice, $n = 463$							
Claimed-distribution EU	.000	.039	.022	.000	.048	.082	.184
Contract closure only	.000	.039	.022	.000	.000	.000	.073
Verify-then-decide	.000	.006	.000	.000	.000	.000	.006
Fixed residual DRO	.000	.032	.019	.000	.030	.063	.177
Closure + residual gate	.000	.035	.019	.000	.000	.000	.030
Oracle hidden state	.000	.000	.000	.000	.000	.000	.000
Deal or No Deal, $n = 1{,}273$							
Claimed EU ≡ contract closure	.000	.075	.175	.000	.236‡	.390‡	.266
Verify-then-decide	.000	.000	.000	.000	.000‡	.000‡	.000
Closure + residual ≡ fixed DRO	.000	.016	.102	.000	.000‡	.000‡	.153
Oracle hidden state	.000	.000	.000	.000	.000‡	.000‡	.000

J Calibration Parameters

The residual gate has four coefficients, fitted once on the 1,304 calibration tasks and frozen before any test evaluation; they are listed here so the gate can be reconstructed exactly. Two entries are worth reading. The state coefficient λ_x is .0020 and the verification coefficient λ_s is exactly .0000, so almost all of the adaptive radius comes from λ_ϕ and the constant ϵ_0; that is the arithmetic behind the ablation finding that removing state uncertainty changes nothing. The calibration-split metrics in the lower rows of Table 15 are reported for the adaptive and fixed variants side by side and already differ by less than the test-split gap, which is why the paper treats the adaptive parameterization as not being the mechanism.

K Boundary Experiments

Where the guarantee stops. Proposition 2 holds when the attacked completion set contains the clean one, and every calibrated attack family only adds unresolved conditions, so all of them satisfy that precondition. To exhibit the other side we render a contract manipulation that makes the two sets *incomparable*: it drops one condition the clean rendering really carried and adds a different one of matched expected cost but higher binding probability, hence smaller amount. The seller's true

Table 15: Residual gate coefficients fitted on 1,304 calibration tasks and frozen before test evaluation. The fixed variant sets a constant radius and therefore has no signal coefficients to fit; those cells are marked *n/a* rather than zero, since a fitted zero and an absent parameter are different statements, and λ_s is in fact a fitted zero.

Parameter	Adaptive	Fixed
ϵ_0	.0179	.0300
λ_x	.0020	*n/a*
λ_ϕ	.1818	*n/a*
λ_s	.0000	*n/a*
Calibration regret	.00130	.00170
Calibration HAR	.0211	.0254
Calibration valid acc.	.4236	.4228

expected settlement is unchanged to 9.1×10^{-13}, so this is revenue-neutral in the same sense as the calibrated families, but the supremum over the completion set falls on *every* task, by 35.49 on average. The closed rule moves accordingly: acceptance rises by $+0.0226$ for contract closure and $+0.0334$ for the full gate, against $+0.0015$ for the unprotected buyer whose valuation never used the set. The guarantee is therefore genuinely load-bearing rather than incidental. Harmful acceptance nonetheless stays at 0.0000 here, because the tasks whose closed price fell far enough to flip were not ones where the deal was bad. We report that as measured: the precondition's failure is demonstrable in the decision, and on this benchmark it did not convert into harm.

Matched-pair stability. Bootstrapping tasks for 3,000 resamples gives a difference of $+0.230$ $[+0.145, +0.305]$ in maximum attack effect and never reaches zero. The competence difference has bootstrap interval $[0.002, 0.132]$ and exceeds 0.05 in 39% of resamples. The correct statement is therefore that the pair is close on the observed competence diagnostic and far apart on vulnerability, not that the two competence values are statistically identical.

Reservation utility. Misspecifying U_{res} by $\pm 30\%$ changes absolute harm, from 0.080 to 0.088 at the negative perturbation, but leaves the contract-channel effect at or below zero throughout the tested range. The state channel remains.

Incomparable contract sets. Proposition 2 requires set inclusion. A boundary construction starts with two real conditions, drops one, and adds another with equal true expected cost. The resulting completion sets are incomparable. Harmful acceptance then rises by $+0.014$, whereas the pure-addition case is non-positive. This is the main deployment limitation of exact contract closure.

Physical substitution. If the delivered item can be worse behind an observation that a legitimate seller could also produce, the signal is unchanged and Proposition 3 applies. A 20% value cut raises harmful acceptance for contract closure from 0.000 to 0.293. A large residual margin can absorb part of the loss but reduces valid acceptance, so this failure belongs to delivery verification rather than presentation closure.

Additional framing. A reference-price anchor and a social-proof note were held out from the core attack families and evaluated on the two model families most likely to respond. None of eight tests survives Holm correction. The largest uncorrected effect is anchoring on `devstral-2`, $+0.035$ $[+0.010, +0.065]$ with uncorrected $p = 0.038$. The sample size is insufficient for a completeness claim about free-text framing.

L Limited External Counterfactual Audit

A public BazaarBench rollout archive contains paired market continuations after a common fork. The corrected join yields 121 post-fork transactions by 22 paired agents. Seller pressure changes the marketplace premium by \$6.99 $[-8.57, +22.27]$ and price paid by \$0.84 $[-12.41, +12.74]$. The only model-matched family is the one our primary benchmark also finds non-manipulable. This audit

therefore corroborates a null for one family but does not establish out-of-distribution replication of the positive effects.

M Residual Latent-State Gate

Structural closure does not identify latent state. After nuisance projection and contract closure, the remaining claimed-state distribution is $\hat{P}(x \mid Q(m))$. We retain a residual Wasserstein set

$$\mathcal{P}_{\epsilon(m)} = \left\{ P : W_1(P, \hat{P}) \leq \epsilon(m) \right\} \tag{8}$$

and compute

$$\underline{U}(m) = \inf_{P \in \mathcal{P}_{\epsilon(m)}} \mathbb{E}_P[U_B(x, p_{\text{cl}})] \,. \tag{9}$$

The deployed decision is

$$\pi(m) = \mathbf{1}[\underline{U}(m) \geq U_{\text{res}} + \delta] \,, \tag{10}$$

where δ bounds numerical or approximate evaluation error. The implemented adaptive radius is $\epsilon_0 + \lambda_x u_x + \lambda_\phi u_\phi + \lambda_s u_s$, fitted on calibration data and frozen before test evaluation. We keep this layer because latent state remains unresolved, but we do not treat its adaptivity as a contribution. Section A.3 shows that a fixed radius performs nearly identically.

If $P_{\text{true}} \in \mathcal{P}_{\epsilon(m)}$ and the implemented robust value has error at most δ, acceptance implies $\mathbb{E}_{P_{\text{true}}}[U_B] \geq U_{\text{res}}$. This is a conditional guarantee. It does not apply when coverage fails or when the seller changes the deployment distribution in response to the calibrated rule.

Proposition 3 (Signal-only identification boundary)**.** *Assume two latent states x_g and x_b induce the same observable post-closure signal $Q(m)$, but their true expected utilities lie on opposite sides of U_{res}. Any deterministic policy that reads only $Q(m)$ must take the same action in both states. It therefore incurs either a missed valid opportunity on x_g or a harmful acceptance on x_b.*

Proposition 3 explains why item substitution and observationally hidden quality cannot receive the same structural guarantee as urgency or enumerable contract conditions. Verification or additional information is necessary to distinguish the states.

N Limitations in Full

The primary benchmark constructs latent quality because CraigslistBargain records no intrinsic item-value oracle; Deal or No Deal avoids that but lacks the contract channel. The headline 0.326 is a scripted rule's upper bound, not a typical model effect: two of seven families are not separable from zero, and both already accept harmfully on honest renderings. Seven families cannot estimate a rank correlation; the evidence against competence-based risk assignment is two controlled comparisons, and the headroom reading of the two nulls is suggestive rather than established. One hosted family is not decoding-matched—its API rejects the recorded sampling parameters—so it enters the vulnerability comparison but not the matched-pair one.

Structural guarantees apply only to the modeled action set. The criterion is conditional on the specified $\mathcal{A}_k$ and observable representation φ; it does not discover an exhaustive manipulation class from open-ended text, and omitted actions remain outside the guarantee. Closure additionally needs the attacked completion set to contain the clean one; making the two incomparable does move the closed rule, shifting acceptance by $+0.0226$ (Appendix K). Urgency closure assumes urgency carries no intrinsic economic value, and free-text framing beyond the modeled channels is not closed.

The residual gate stays conditional on correctly specified buyer utility and coverage, and on distributional conditions a strategic seller invalidates. If a worse item can be delivered behind the same observable signal as a legitimate one, Proposition 3 applies and no signal-only rule separates the states.

O Related Work in Full

Negotiation and agentic commerce. Bargaining benchmarks measure agreement, utility, strategy, or welfare rather than harmful acceptance under a fixed hidden transaction (Lewis et al., 2017; He

et al., 2018; Bianchi et al., 2024; Xia et al., 2024; Bhattacharya et al., 2025; Kirshner et al., 2026; Oh et al., 2026; Liu et al., 2026), and one equilibrium-referenced study reports individually irrational contracts (Liang & Xu, 2026). Closest to our failure mode, TERMS-Bench scores a false-agreement rate for accepting individually irrational offers (Zhang et al., 2026), independently confirming the behavior is real; it samples a fresh scenario per episode rather than re-rendering a fixed one, and evaluates no defense. That is the gap we fill.

Agent safety and strategic response. AgentBench, WebArena, and MultiAgentBench evaluate interactive execution, web tasks, and coordination (Liu et al., 2023; Zhou et al., 2024; Zhu et al., 2025). AgentDojo, InjecAgent, Agent-SafetyBench, Agent Security Bench, and OpenAgentSafety study unsafe behavior under malicious instructions, untrusted data, or tools (Debenedetti et al., 2024; Zhan et al., 2024; Zhang et al., 2024; 2025; Vijayvargiya et al., 2025). Defenses there split the way we argue decisions should: CaMeL routes untrusted data out of program flow, reporting 77% of AgentDojo tasks solved with provable security against 84% undefended (Debenedetti et al., 2026)—a structural guarantee at a measured price. What the criterion adds over such a design, and what it does not, is set out in Appendix P. Training-time separation (Chen et al., 2025) and prompt-level marking are the softer alternatives; §5.5 measures the latter. Abdelnabi and Bagdasarian argue via contextual integrity that no fixed policy can admit every legitimate flow while blocking every injected one (Abdelnabi & Bagdasarian, 2026); that impossibility is our third class, and what the criterion adds is which channels escape it and at what price. Adaptive red-teaming defeats the softer defenses once it optimizes against them (Yin et al., 2026), as our strategic seller does. Pre-execution governance checks proposed commerce actions against role and policy (Shi et al., 2026), acting on the agent's output where our operators act on its input. Our seller stays inside the economic protocol and changes only the evidence or framing shown to the buyer, which is closer to strategic classification (Hardt et al., 2016) and performative prediction (Perdomo et al., 2020). That literature also studies the cost of manipulation-proofness (Chen et al., 2018; Shavit et al., 2020; Björkegren et al., 2020) and estimating an unknown manipulation set (Lechner et al., 2023), but throughout it the decision-maker *designs* the predictor; Proposition 1 shows what changes when it cannot.

Robust decisions and calibrated risk. Robust optimization protects decisions against uncertainty sets (Bertsimas et al., 2011; Esfahani & Kuhn, 2018), and conformal methods build finite-sample regions (Johnstone & Cox, 2021; Angelopoulos & Bates, 2023), control decision losses (Angelopoulos et al., 2024; 2025), and learn context-dependent sets (Wang et al., 2023; Guo, 2026). These remain the right tools for our latent residual channel, and we use them as established components rather than claim them. Our experiments expose their boundary: calibration on an exogenous mixture need not stay valid once the seller responds to the published rule, a failure treated in general form by strategic conformal prediction (Csillag et al., 2025) and non-exchangeable risk control (Farinhas et al., 2024).

P Positioning and Statistical Scope

What this adds over structural defenses that already exist. CaMeL defends a tool-using agent by extracting control and data flow from the trusted query, so untrusted data cannot reach program flow, and reports 77% of AgentDojo tasks solved with provable security against 84% undefended (Debenedetti et al., 2026). We are not claiming to have discovered that structural defenses beat statistical ones; that work established it in its domain, and its seven-point utility gap is an independent instance of the price this paper formalizes. What is missing there, and what the criterion supplies, is a way to decide *in advance* which channels admit such a defense at all. CaMeL is a system design whose scope is fixed by its architecture; the criterion is a test on observables that says, for a channel not yet considered, whether an exact defense exists, exists only at the cost of enumerating a completion set, or does not exist. Section 5.5 runs that test on a domain we did not build and it predicts, before any model is run, which defenses come out exact, which do nothing, and which are model-dependent. Two further differences are not matters of emphasis. First, the criterion is falsifiable by re-encoding: a channel changes class when its representation changes, and Section 5.4 moves one. Second, Proposition 1 shows the invariant side admits no approximation once the decision-maker is not ours to design, which is why a prompt-level defense cannot substitute — a claim about the whole class of such defenses rather than about any particular one, and one we measure against a published baseline rather than only against our own.

What seven families can and cannot support. Figure 3 motivates the deployment-time surrogate question, but seven points cannot estimate a population rank correlation, and we do not ask them to. The reported Spearman coefficient of 0.250 with exact permutation $p = 0.595$ is included so the reader can see that the data are consistent with no monotone relationship, not as evidence for one; a null at $n = 7$ is uninformative either way, and we would draw the same conclusion had it come out significant, because the conclusion does not depend on it. The argument rests on two controlled comparisons, and a single counterexample is enough to refute a universal claim. At the extremes of a twentyfold competence range the effect differs by 0.065; in the middle, a competence difference of 0.010 accompanies an effect difference of $+0.230$ $[+0.145, +0.305]$ with zero outside all $3{,}000$ task-level resamples and the gap widening once decoding is aligned. Together these comparisons are evidence against using this competence diagnostic as a risk-ordering surrogate on the tested models; they are not a population claim about all model families. The headroom reading of the two null families—that an attack can only move a buyer starting on the right side of its reservation, which the $\rho = -0.714$ association between clean harmful acceptance and effect is consistent with—is reported as a reading and not as a law, on the same grounds.

Q Baseline Discovery and Comparability

We searched the negotiation, agentic-commerce, and agent-safety literatures for methods that could be scored directly on this task, and the table records the outcome of that search rather than a curated shortlist. "Direct score" is the column that matters: a method qualifies only if it can be evaluated on the same paired acceptance decision without changing the action space, the outcome metric, or the information the buyer receives. Most recent negotiation systems fail that test not because they are weaker but because they measure a different quantity—agreement, surplus, or dialogue quality over a multi-turn exchange rather than one accept-or-reject decision against a fixed hidden transaction. Reporting their published numbers beside ours would compare different experiments. The methods that do qualify are the scripted decision rules in Table 2 and, out of domain, the datamarking defense of Section 5.5; training-time separation and interpreter-based isolation are named in Table 16 as reproducible in principle but out of scope here, since one requires fine-tuning each family and the other a full capability runtime.

Table 16: Baseline and closest-work coverage. "Direct score" indicates whether the method can be evaluated on the same paired acceptance task without changing its problem definition.

Work or baseline	Category	Direct score	What it evaluates	Role here
Claimed-distribution EU	Simple heuristic	Yes	Accept from claimed expected utility	Unprotected reference
Contract closure only	Mechanism control	Yes	Enumerable contract uncertainty	Isolates structural contribution
Verify-then-decide	Conservative heuristic	Yes	Accept only verified offers	Tests rejection-based safety
Fixed residual DRO	Robustness control	Yes	Constant ambiguity radius	Tests whether adaptivity matters
Direct LLM / safety prompt	Same-backbone controls	Yes	Model decision under the same signal	Tests prompt-only protection
Oracle hidden state	Upper bound	Yes	True hidden transaction	Identifies remaining headroom
NegotiationArena (Bianchi et al., 2024)	Closest negotiation	No	Bargaining payoff and tactics	Problem positioning
Xia et al. (2024)	Closest negotiation	No	Product-price bargaining	Supplies replication corpus
MERIT / AgoraBench (Oh et al., 2026)	Recent strong negotiation	No	Bargaining utility and human preference	Recent baseline audit
AgenticPay (Liu et al., 2026)	Agentic commerce	No	Multi-round market feasibility and welfare	Recent baseline audit
Liang and Xu (Liang & Xu, 2026)	Recent economic audit	No	Equilibrium efficiency and irrational contracts	Closest recent risk evidence
Agent Security Bench (Zhang et al., 2025)	Agent safety	No	Instruction-driven attack success	Motivates cross-model vulnerability comparison

R Reproducibility Details

The current manuscript directly supports the following settings. `qwen3.6:35b-a3b` and `gemma4:31b` use `think=false` and 250 maximum output tokens. `gpt-oss:20b` uses

`think=low` and 400 maximum output tokens because the attempted non-reasoning configuration produced 17% empty completions. Temperature is 0.7 and top-p is 0.95. One decoding seed is derived for each task and attack condition using blake2b. The `gpt-oss` arm records 45 failures among 2,800 rows, and complete-case estimates differ by less than 0.006 from reported values on every attack family.

Recorded decoding settings, all at $T = 0.7$ and top-p 0.95: `gpt-oss:20b` at reasoning effort `low` with a 400-token budget, because it returns empty content when asked to disable reasoning; `gemma4:31b`, `qwen3.6:35b-a3b` and `devstral-small-2:24b` with reasoning disabled and a 250-token budget; and `deepseek-chat`, the vendor's non-reasoning alias for V4, with a 600-token budget. The reasoning variants `deepseek-v4-pro` and `deepseek-v4-flash` were probed on real benchmark prompts and rejected: `pro` emitted 20,448 characters of reasoning and no content at a 6,000-token budget, and `flash` required 1,500 tokens per decision. Observed completion lengths stay well inside every budget, with medians of 53 to 211 tokens. The matched pair was additionally re-run at an identical 400-token budget, and `devstral-small-2:24b` was repeated at 250 against 400 tokens as a decoding-variance check; both are reported in Section 6.