LLMs Can Play (Global) Games

Khaled Eltokhy
Department of Economics
The Graduate Center, CUNY

February 2026

Abstract

I embed seven large language models in the Morris–Shin (2003) regime change global game, conveying private signals as natural-language intelligence briefings. Across 1,800 country–periods and 45,000 decisions, join rates exhibit threshold-policy alignment with global-game comparative statics: the within-country correlation between join rates and a benchmark attack mass averages $r = +0.80$, collapses to $+0.05$ when briefings are scrambled, and flips to -0.80 when signals are inverted. I then study how authoritarian regimes exploit the same information channel that enables coordination. Surveillance creates a belief–action wedge—self-censorship that suppresses expressed behavior while leaving beliefs intact—consistent with Kuran (1991): agents' stated beliefs track the benchmark, yet expressed joining falls by 11.1 pp across three architectures. Censorship pools and distorts signals; propaganda saturates quickly. The regime need not change what citizens believe; it needs only to make them uncertain about each other.

JEL: C72, C92, D82, D83, P16
Keywords: global games, regime change, LLM agents, information design, Bayesian persuasion, belief-action wedge, preference falsification

1 Introduction

Coordination games with multiple equilibria are central to the analysis of bank runs (Diamond and Dybvig, 1983), currency attacks (Obstfeld, 1996), and political upheaval (Angeletos et al., 2007). The theory of global games (Carlsson and van Damme, 1993; Morris and Shin, 2003; Frankel et al., 2003) resolves the multiplicity by introducing private information: when agents observe noisy private signals about an underlying fundamental, a unique equilibrium emerges in threshold strategies. The canonical application—regime change—has been extensively studied theoretically. Laboratory experiments have tested the theory in simplified settings: small groups with numeric signals and stylized payoffs (Heinemann et al., 2004, 2009; Szkup and Trevino, 2020). But the full Morris–Shin regime change game—continuous private signals, large groups, strategic uncertainty—has not been implemented experimentally. Field data from actual crises confounds strategic behavior with institutional and informational heterogeneity.

I take a different approach: I embed large language model (LLM) agents directly in the Morris and Shin (2003) regime change game. Each agent receives a private signal $x_i = \theta + \varepsilon_i$, translated into a natural-language intelligence briefing describing the political, economic, and security situation. No explicit payoff table is provided—the stakes of joining or staying are embedded in the narrative, forcing agents to extract strategic information from language rather than from a formatted matrix. I run this experiment across seven architecturally distinct models spanning six families (Mistral, Llama, Qwen, GPT, Arcee, and MiniMax), with 25 agents per country–period and pure-treatment sample sizes of 100–1,000 country–periods per model (Table 1), totaling 1,800 country–periods (45,000 individual decisions) in the pure treatment alone.

The first finding is that LLM agents implement stable, monotone threshold-like policies over narrative private information. The correlation between the benchmark attack mass $A(\theta)$ and the empirical join fraction averages $r = +0.80$ ($p < 0.001$ for every model). Two falsification tests confirm that this correlation is driven by briefing content rather than incidental features of the prompt: randomly scrambling briefings across periods reduces the within-country correlation to $r = +0.05$, and inverting the signal direction flips it to $r = -0.80$. In both cases the change relative to the pure treatment is significant (Fisher z-test, $p < 0.001$). This establishes monotonicity and content sensitivity—sufficient to interpret the behavior through global-game comparative statics, though not to establish full Bayesian Nash rationality. Elicited beliefs track a theoretical benchmark ($r = +0.79$) and predict actions beyond what signals alone predict ($r = +0.84$, exceeding the text-baseline $r = 0.80$), providing evidence of strategic processing beyond mere sentiment following.

The second finding—and the paper's central contribution—is that the information channel is simultaneously the mechanism of coordination and its greatest vulnerability. Pre-play communication has a near-zero mean effect on willingness to act (mean shift -0.4 pp across models; not significant in the pooled sample), yet the channel it opens introduces strategic uncertainty that makes coordination exploitable. Surveillance poisons the channel through a belief–action wedge (−13.5 pp for the primary model, $p < 0.001$): agents suppress expressed behavior while maintaining private beliefs, consistent with the preference falsification mechanism of Kuran (1991). Censorship pools and distorts private signals, and its interaction with surveillance is large and model-dependent. Propaganda's behavioral effect saturates quickly while its mechanical effect scales linearly, implying diminishing returns. The regime does not need to change what citizens believe—it needs only to make them uncertain about each other.

The paper makes three contributions. First, it tests whether the threshold equilibrium patterns predicted by global games theory emerge when LLM agents are embedded in the full Morris–Shin regime change game—with continuous private signals, large groups, and narrative information—going beyond the simplified coordination games tested in existing laboratory experiments. Second, it provides computational experiments inspired by the information design and authoritarian control predictions of Goldstein and Huang (2016), Kolotilin et al. (2022), and Edmond (2013) in a coordination game, yielding a unified account of how authoritarian regimes exploit the dual nature of communication channels—instruments of coordination that are simultaneously vectors of control. Third, it demonstrates that LLMs can serve as experimental subjects for strategic environments, extending the Horton (2023) *homo silicus* methodology beyond 2×2 games to the continuous-signal, N-player coordination games that dominate applied theory, with results demonstrating robustness across seven models spanning six architecture families, offering a proof-of-concept rather than population inference over architectures. Appendix D discusses implications for AI alignment.

The narrative arc connects the two parts through the information channel. Part I (Sections 5–7) establishes that LLM agents extract strategic information from narrative briefings robustly enough to produce the coordination regularities that global games theory predicts. Part II (Sections 8–11) exploits this regularity: the same information channel that enables coordination becomes the vector through which authoritarian regimes can suppress it.

Section 2 reviews the related literature. Section 3 presents the theoretical framework. Section 4 describes the experimental design. Section 5 reports the main results on threshold-policy alignment; Section 6 presents the falsification tests. Section 7 analyzes pre-play communication. Sections 8–11 cover information design, surveillance, propaganda, and their interactions. Appendix B reports robustness checks. Section 12 concludes.

2 Related Literature

This paper connects five literatures: global games and equilibrium selection, information design and Bayesian persuasion, communication in coordination games, the political economy of authoritarian information control, and the emerging field of LLMs as economic agents.

The theory of global games resolves the equilibrium multiplicity that plagues coordination games by introducing heterogeneous private information. Carlsson and van Damme (1993) showed that adding arbitrarily small noise to a 2×2 coordination game generically selects the risk-dominant equilibrium via iterated dominance. Morris and Shin (1998) applied this technique to currency crises, demonstrating that heterogeneous private signals about fundamentals deliver a unique threshold equilibrium even in large-player coordination games. Frankel et al. (2003) generalized the result to N-player, multi-action games with strategic complementarities.

The canonical regime change application—in which citizens decide whether to join an uprising against a regime of uncertain strength—was developed by Morris and Shin (2003), who established the threshold equilibrium structure I implement experimentally. Angeletos et al. (2007) extended the framework to dynamic settings where agents learn across periods, showing that multiplicity can re-emerge when agents observe whether the regime survived previous rounds. Morris and Shin (2002) demonstrated that public signals are overweighted in coordination games because they predict others' actions, a finding central to my communication and information design treatments.

Laboratory experiments have tested the theory in stylized settings that necessarily depart from the canonical regime change game. Heinemann et al. (2004) ran coordination games with public and private signals, finding that subjects' thresholds match the global game prediction under private information but tilt toward payoff-dominance under common information. Heinemann et al. (2009) measured strategic uncertainty directly through certainty equivalents. Shurchkov (2013) tested dynamic global games, finding that subjects learn from failed attacks. Szkup and Trevino (2020) elicited beliefs alongside actions, finding that comparative statics of thresholds with respect to signal precision are reversed relative to theory—subjects become more cautious with noisier signals, consistent with level-k thinking rather than Bayesian Nash equilibrium. Helland et al. (2021) tested information quality in a regime change game with numeric signals and small groups, confirming the level-k reversal. These experiments share a common limitation: subjects receive numeric signal draws and face stylized payoff tables, compressing the rich information processing that real-world coordination requires into a simple decision problem.

This paper implements the full Morris–Shin regime change game with natural-language private signals and 25-agent groups, going beyond the small-group, numeric-signal designs of existing experiments to test the threshold equilibrium prediction in the canonical application for which it was developed.

Kamenica and Gentzkow (2011) established the Bayesian persuasion framework: a sender who commits to an information structure can influence a Bayesian receiver's action by shaping the posterior distribution of beliefs. Bergemann and Morris (2016) unified Bayesian persuasion with correlated equilibrium under the concept of Bayes Correlated Equilibrium. Bergemann and Morris (2019) provided a comprehensive survey integrating cheap talk, persuasion, and robust mechanism design.

The application to coordination games is directly relevant. Goldstein and Huang (2016) applied Bayesian persuasion to the regime change game, showing that a credible commitment to abandon the regime below a threshold functions as an optimal signal. Inostroza and Pavan (2025) solved the optimal public information design problem in a global game with heterogeneous private signals, characterizing when pass/fail structures are optimal. Kolotilin et al. (2022) characterized optimal censorship via one-sided pooling rules ("upper censorship" in their terminology), showing that pooling one side of a threshold can be optimal for all priors when the sender's marginal utility is quasi-concave. Mathevet et al. (2020) characterized the extent to which an information designer can manipulate agents' higher-order beliefs.

My information design experiments implement these theoretical designs computationally within a full-scale coordination game, providing computational tests of information design predictions in a global game.

The cheap talk literature—Crawford and Sobel (1982), Farrell and Rabin (1996), Blume and Ortmann (2007), Ellingsen and Östling (2010)—establishes that pre-play communication can improve coordination, with Avoyan (2020) testing this in a two-player global game. In real-world coordination, Enikolopov et al. (2020) provided causal evidence that social media penetration increases protest incidence. My communication treatment embeds agents in a Watts-Strogatz small-world network and allows natural-language messaging before the coordination decision.

The theoretical literature on authoritarian information control builds directly on the global games framework. Edmond (2013) embedded costly propaganda into the Morris–Shin regime change game. Kuran (1991) provides the foundational theory of preference falsification—the systematic misrepresentation of political preferences under social pressure. Empirical work documents that Chinese censorship targets content with collective action potential (King et al., 2013), that surveillance awareness suppresses expression (Penney, 2016; Stoycheff, 2016), and that pro-regime propaganda reduces protest probability (Carter and Carter, 2021). My surveillance and propaganda treatments directly test these mechanisms within the full regime change game—an environment difficult to implement with human subjects at scale.

Horton (2023) proposed treating LLMs as "homo silicus"—computational models of human decision-makers. Subsequent work has tested LLMs in game-theoretic settings: Akata et al. (2025) found that LLMs perform well in self-interested games but struggle in coordination games; Petrov et al. (2025) evaluated 22 LLMs on a behavioral game theory battery, finding that model scale alone does not predict strategic performance; Sun et al. (2025) identify coordination games as a consistent failure mode. The alignment literature motivates my design: Huang et al. (2024) and Carlini et al. (2025) document that ethical alignment and chatbot fine-tuning shift risk preferences and amplify omission bias, which is why I convey strategic stakes through narrative rather than explicit payoff tables. Critical reviews by Gao et al. (2025) and Grossmann et al. (2025) warn that validation remains poorly addressed in LLM-based agent simulations.

No existing paper places LLM agents in a Morris–Shin global game—the specific game form where private noisy signals about an underlying state variable determine a threshold equilibrium. I provide the first such implementation, and extend it to information design, surveillance, and propaganda.

3 The Global Game of Regime Change

A continuum of citizens indexed by $i \in [0,1]$ simultaneously choose whether to join an uprising ($a_i = 1$) or stay home ($a_i = 0$). The regime has strength $\theta \in \mathbb{R}$, drawn from a diffuse (improper uniform) prior. States $\theta \leq 0$ represent regimes so weak they fall without opposition; states $\theta \geq 1$ represent regimes that survive even unanimous attack. The regime falls if the mass of citizens who join exceeds θ:

$$\text{Regime falls} \iff A \equiv \int_0^1 a_i \, di > \theta. \tag{1}$$

Payoffs depend on the citizen's action and the outcome:

$$u_i(a_i, A, \theta) = \begin{cases} B & \text{if } a_i = 1 \text{ and } A > \theta \\ -C & \text{if } a_i = 1 \text{ and } A \leq \theta \\ 0 & \text{if } a_i = 0 \end{cases} \tag{2}$$

where $B > 0$ is the payoff to joining a successful uprising and $C > 0$ is the cost of joining a failed attempt. Nonparticipants receive zero regardless of the outcome.

Each citizen observes a private signal $x_i = \theta + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ independently across citizens.

Proposition 1 (Morris and Shin, 2003). *In the limit of diffuse priors, there exists a unique Bayesian Nash equilibrium in threshold strategies. An agent joins if and only if $x_i < x^*$, where*

$$x^* = \theta^* + \sigma \Phi^{-1}(\theta^*) \tag{3}$$

and $\theta^* = B/(B + C)$.

The *attack mass*—the fraction of the population that joins at regime strength θ—is:

$$A(\theta) = \Phi\left(\frac{x^* - \theta}{\sigma}\right). \tag{4}$$

This is a decreasing function of θ: weaker regimes face larger uprisings.

An information designer controls the mapping $\pi : \Theta \to \Delta(\mathcal{S})$ from states to signal distributions, but cannot control agents' actions. In my implementation, π is the function mapping regime strength θ to the parameters of the briefing generator—a deterministic system that produces a natural-language intelligence briefing from a z-score derived from the agent's private signal.

The briefing generator has three control parameters: clarity (the width of the Gaussian kernel mapping z-scores to text, where wider kernels produce more ambiguous briefings), directional precision (the slope of the mapping from z-score to briefing sentiment, where steeper slopes produce more accurate signal reflection), and dissent framing (the floor on the probability that the briefing includes language about public discontent).

The designer concentrates manipulation near θ^* using a Gaussian proximity weight:

$$w(\theta) = \exp\left(-\left(\frac{\theta - \theta^*}{\text{bandwidth}}\right)^2\right) \tag{5}$$

where bandwidth $= 0.15$ in the baseline specification.

The framework generates testable predictions for both the baseline game and information design.

Hypothesis 1 (Sigmoid Response)**.** *The empirical join fraction should be positively correlated with the theoretical attack mass $A(\theta)$.*

Hypothesis 2 (Scramble Falsification)**.** *The correlation in Hypothesis 1 should collapse when the mapping from θ to briefing content is broken (scramble test).*

Hypothesis 3 (Directional Sensitivity)**.** *The correlation should invert when signals are flipped.*

Hypothesis 4 (Communication Channel)**.** *Pre-play communication should increase join rates, with the effect strongest near θ^* where strategic uncertainty is highest.*

Hypothesis 5 (Ambiguity Pooling)**.** *Increasing ambiguity and mixed evidence near θ^* should flatten the θ–join relationship and induce pooling.*

Hypothesis 6 (Censorship Distortion)**.** *Upper censorship should distort coordination by pooling weak-regime states to a neutral signal, flattening join rates in the censored region (Kolotilin et al., 2022).*

Hypothesis 7 (Surveillance Chilling Effect)**.** *Informing agents that communications are monitored should reduce coordination (Kuran, 1991).*

Hypothesis 8 (Propaganda Dose-Response)**.** *Regime plant agents transmitting pro-regime messages should suppress coordination, with the effect increasing in the number of plants (Edmond, 2013).*

4 Experimental Design

The experiment has three parts. Part I tests whether LLM agents implement threshold-like policies in the global game when private signals are conveyed in natural language: a pure treatment (private signals only), a communication treatment (pre-play messaging), and falsification tests. Part II takes the behavioral foundation as given and studies information design: stability/instability designs, censorship, public signal injection, and single-channel decomposition. Part III tests whether an authoritarian regime can exploit the communication channel through surveillance, propaganda, and their interaction. All LLM interactions use the same prompt structure across models.

A note on the state variable. In the theory (Section 3), θ is an unbounded fundamental with special roles for $\theta \leq 0$ (regime falls without opposition) and $\theta \geq 1$ (regime survives even unanimous attack). In Part I experiments, θ is drawn from a normal distribution and agents are not shown payoff parameters (B, C). To evaluate monotonicity on a common scale, I compute the benchmark $A(\theta)$ under the canonical normalization $B = C = 1$ (so $\theta^* = 0.50$) and $\sigma = 0.3$. In Part II, I keep $B = C = 1$ and restrict attention to a fixed θ-grid in $[0.20, 0.80]$ for comparability with the canonical $[0, 1]$ formulation.

Notation. $\theta^* = B/(B + C)$: theoretical cutoff (regime falls iff attack mass exceeds θ). $x^* = \theta^* + \sigma\Phi^{-1}(\theta^*)$: signal cutoff (agent joins iff $x_i < x^*$). $\hat{\theta}^*$: estimated cutoff $= -\hat{b}_0/\hat{b}_1$ from the fitted logistic $P(\text{join} \mid \theta) = 1/(1+e^{b_0+b_1\theta})$, i.e., the θ where the fitted join probability equals 0.5.

For each country–period, nature draws $\theta \sim \mathcal{N}(\bar{z}, 1)$, where $\bar{z}$ is a public prior mean drawn randomly for each country. Each agent i receives a private signal $x_i = \theta + \varepsilon_i$ and computes a z-score $z_i = (x_i - \bar{z})/\sigma$. Because agents observe only their private briefing and never the prior distribution or its parameters, the diffuse-prior equilibrium formula (Proposition 1) serves as the relevant benchmark. The z-score is then translated into a multi-paragraph intelligence briefing by a deterministic generator that maps signal strength to narrative content about regime stability, economic conditions, public sentiment, and coordination prospects. Figure 1 summarizes the signal-to-text-to-decision pipeline.

A design choice deserves comment. The briefing generator maps z-scores to narrative content through logistic slider functions, so the monotone *direction* of the response is partially built into the text generation—any model that extracts sentiment will produce a negative correlation between θ and join probability. The empirical contribution is not the direction but the *quantitative structure*: the

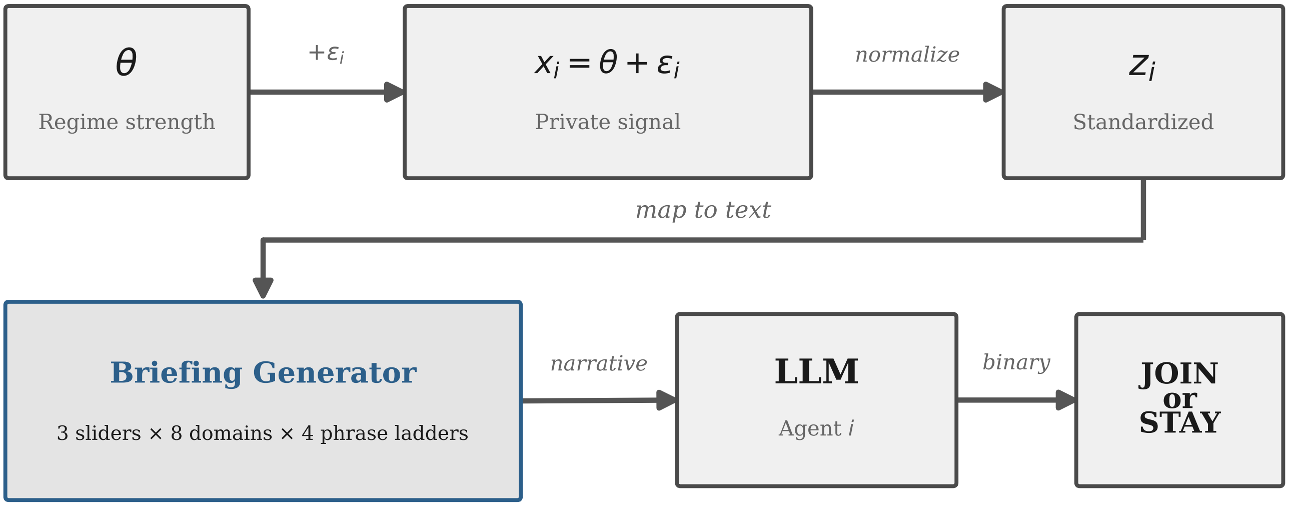


Figure 1: Signal-to-text-to-decision pipeline. Regime strength θ generates private signals x_i, which are converted to z-scores and rendered into natural-language intelligence briefings via 8 evidence domains and 3 latent sliders (direction, clarity, coordination). Each LLM agent reads its briefing and outputs a binary JOIN/STAY decision. The briefing layer is deterministic conditional on z_i; all stochasticity enters through the LLM's decoding.

sigmoid shape, the sensitivity of the fitted cutoff to payoff narratives (Section 5), and the robustness across seven models spanning 30B to 235B parameters. Within-briefing falsification tests (Appendix B.13) confirm that the signal is distributed across all eight evidence domains rather than driven by any single feature.

Calibration adjusts a single parameter—the cutoff center—via a damped iterative procedure that shifts the center until the fitted logistic is approximately zero-centered. The sigmoid shape is emergent from the LLM's own response pattern and is never optimized or penalized. Holdout validation (30% of z-grid points withheld) confirms no overfitting (holdout RMSE 0.112 vs. training RMSE 0.131).

Calibration does not use θ draws or any global-game outcome data, and all reported treatments hold calibrated parameters fixed. Six models run with default parameters (no calibration) produce $|r| > 0.69$, confirming that the monotone threshold pattern is emergent rather than calibrated (Appendix Table A2).

Each agent receives a system prompt identifying them as a citizen deciding whether to JOIN or STAY, followed by their intelligence briefing. No explicit payoff table is provided—the stakes are conveyed entirely through the narrative.

This design choice is substantive. In preliminary experiments, providing an explicit payoff table caused sophisticated models to short-circuit the information-processing channel: they computed the optimal strategy from the table and ignored briefing content, producing flat join rates uncorrelated with regime strength. The no-payoff-table design forces agents to form beliefs from the narrative, mirroring how real citizens process political information from news and rumors rather than from a formatted decision matrix.

Part I has four treatments. In the *pure global game*, each agent decides independently based on their private briefing. In the *communication* treatment, agents send a message to a small network of "trusted contacts" (Watts-Strogatz small-world network, $k = 4$, $p = 0.3$) before deciding, with access to both their briefing and received messages. Two falsification tests break the signal channel: in *scramble*, all briefings across periods within a country are pooled and randomly redistributed; in *flip*, the z-score is negated before briefing generation, so agents who should see weak-regime cues receive strong-regime cues and vice versa.

Part II implements information designs. Design names refer to the *regime's* objective, not the equilibrium outcome: the "stability" design is the information structure a stability-seeking regime would implement. The *stability-maximizing* design multiplies clarity width by 4, raises the dissent floor to 0.45, and flattens the directional slope by a factor of 0.25 near θ^*. The *instability-maximizing* design does the opposite: clarity width is multiplied by 0.15, the dissent floor is lowered to 0.05, and the directional slope is steepened by a factor of 3. *Public signal injection* appends a shared "news bulletin" generated from θ with 4 observations to each agent's private briefing, creating a common-knowledge channel. *Upper censorship* pools weak-regime states ($\theta \leq \theta^*$) so agents receive an identical censored briefing, while fully revealing states

Table 1: Model summary. Columns report country-period counts in the pure, communication, and falsification (scramble+flip) suites. All runs use $N = 25$ agents per period and $\sigma = 0.3$.

Model	Arch.	Pure	Comm	Falsif.
Mistral Small Creative	Mistral	1000	1000	200
Llama 3.3 70B	Llama	100	100	200
Qwen3 30B	Qwen (MoE)	100	100	200
GPT-OSS 120B	GPT	200	200	1000
Qwen3 235B	Qwen (MoE)	200	200	200
Trinity Large	Arcee	100	100	200
MiniMax M2-Her	MiniMax	100	100	200
Total		**1800**	**1800**	**2200**

above θ^* (Kolotilin et al., 2022); *lower censorship* pools strong-regime states ($\theta \geq \theta^*$).

Part III tests authoritarian instruments that exploit the communication channel. The *surveillance* treatment augments the communication prompt with a warning that communications are being monitored by regime security services. *Propaganda* introduces regime plant agents ($k = 2, 5, 10$) who participate in the communication network but transmit fixed pro-regime messages and always STAY.

I test seven architecturally distinct models spanning six architecture families (Table 1). Models range from 30 billion to 235 billion parameters, including both dense architectures (Llama, Mistral) and mixture-of-experts (Qwen). All experiments use $N = 25$ agents per country–period and $\sigma = 0.3$, with sample sizes varying by model and treatment as reported in Table 1. Because agents are not shown (B, C), Part I benchmarks use the fixed normalization $B = C = 1$ (so $\theta^* = 0.50$); payoff comparative statics are tested by varying *narrative* stakes (Section 5). All LLM calls use temperature = 0.7 with a single sample per decision—no majority voting or averaging—so each of the 45,000 individual decisions reflects one stochastic draw from the model's conditional distribution (see Appendix C.1 for full decoding parameters).

The unit of randomization is the country–period (θ draw plus agent-level decoding stochasticity). For agent-level regressions, standard errors are clustered at the country–period level to account for within-period correlation among agents sharing the same θ. For period-level correlations, I report Fisher-z confidence intervals. The 25 agents within a period share the same θ and calibration; their decisions are conditionally independent given their private signals.[1]

For the information design experiments, I fix $B = C = 1$ (so $\theta^* = 0.50$) and a grid of 9 values of θ spanning $[\theta^* - 0.30, \theta^* + 0.30] = [0.20, 0.80]$, running repeated country–periods per (design, θ) cell with 25 agents each. Baseline, stability, censorship, scramble, and flip use 30 repetitions per cell (270 observations per design). Instability and public signal use 60 repetitions per cell (540 observations). Single-channel decomposition uses 30 repetitions per cell (270 observations) for each channel. The primary model is Mistral Small Creative. Cross-model replication uses five additional models.

Table 2 maps each treatment to the theoretical channel it tests, the directional prediction, and the observed result.

The eight hypotheses in Section 3 were pre-specified. Four achieve $p < 0.001$ and survive Bonferroni correction at $\alpha = 0.05/8 = 0.00625$: alignment (H1), flip (H3), stability (H5), and surveillance (H7). Two additional hypotheses are supported by design: H2 (scramble) predicts that shuffled briefings collapse the correlation—the non-significant $r = 0.04$ ($p = 0.14$) confirms this null prediction; and H4 (communication), where theory predicts an ambiguous effect, consistent with the small, insignificant shift ($p = 0.29$). Censorship (H6) is significant at conventional levels ($p = 0.023$) but does not survive the Bonferroni threshold. Propaganda (H8) shows a directionally correct but non-significant behavioral effect ($p = 0.37$). Table 3 reports all eight tests. Exploratory analyses—decomposition, cross-model heterogeneity, and instrument interactions—are reported with uncorrected p-values and should be interpreted accordingly.

5 Do LLM Agents Implement Global-Game Threshold Policies?

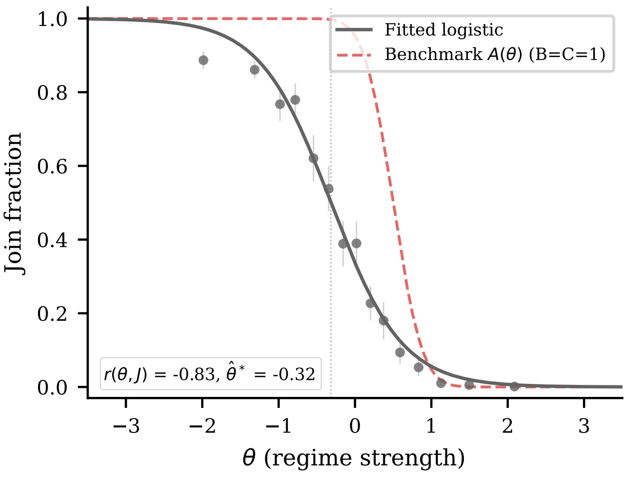


Figure 2: Empirical join fraction vs. regime strength θ (Mistral Small Creative, 1,000 country–periods). Grey points show binned means with 95% CIs; solid line is the fitted logistic. Dashed red: theoretical attack mass $A(\theta)$. The empirical sigmoid is shifted leftward ($\hat{\theta}^* = -0.33$) relative to the theoretical threshold ($\theta^* = 0.50$), reflecting the attenuation and baseline action bias discussed in the text. Cross-model results in Table 4 (mean $r = +0.80$, all seven significant at $p < 0.001$).

[1] I do not claim literal conditional independence—shared prompt structure and model weights introduce common factors. The clustering accounts for within-period correlation.

Table 2: Treatment map. Each row describes a treatment, the channel it tests, the unit of observation, the theoretical prediction, and the observed result. Part I tests whether LLM agents play the global game; Part II studies information design; Part III tests authoritarian exploitation of the communication channel. r denotes Pearson correlation; Δ is the change in mean join rate relative to the relevant baseline (percentage points).

Treatment	Part	Channel tested	Unit	Prediction	Observed
Core treatments					
Pure global game	I	Baseline monotonicity	Period	$r(J, A(\theta)) > 0$	$r = +0.73$
Communication	I	Pre-play messaging	Period	Ambiguous	+1.23 pp ($p = 0.327$)
Scramble (cross-agent)	I	Content vs. format	Period	$r \to 0$	$r = +0.07$
Flip	I	Signal direction	Period	Sign reversal	$r = -0.67$
B/C narratives	I	Payoff comparative statics	Period	Cutoff shifts	As predicted
Information design (Part II, primary model)					
Stability	II	Ambiguity near θ^*	Period	↓ join	−9.0 pp
Instability	II	Sharpened signals	Period	↓ join (clearer sorting)	−34.2 pp
Public signal	II	Common knowledge channel	Period	↓ join (dominates private)	−39.2 pp
Censor (upper)	II	Pooling weak states	Period	Plateau below θ^*	−3.1 pp
Censor (lower)	II	Pooling strong states	Period	Plateau above θ^*	−1.9 pp
Authoritarian instruments (Part III)					
Surveillance	III	Preference falsification	Period	↓ join	−13.5 pp
Propaganda (k=10)	III	Information contamination	Period	↓ join, saturating	−2.3 pp (behavioral)
Within-briefing falsification					
Observation shuffle	I	Bullet ordering vs. content	Period	r unchanged	$r = -0.911$
Domain scramble (coord.)	I	Coordination-relevant domains	Period	$\lvert r\rvert$ falls if domains drive signal	$r = -0.921$
Domain scramble (state)	I	State-capacity domains	Period	$\lvert r\rvert$ falls if domains drive signal	$r = -0.928$

Table 3: Pre-specified hypotheses and test results. H1–H4 use pooled Part I data across all seven models; H5–H8 use the primary model (Mistral Small Creative). "Supported" indicates whether the data pattern matches the hypothesis at $\alpha = 0.05$.

H	Hypothesis	Estimand	Null	Test	Stat	p	Supported?
H1	Sigmoid Response	$r(J, A(\theta))$	$r = 0$	Pearson	0.757	<0.001	Yes
H2	Scramble Falsification	r(scramble)	$r \neq 0$	Pearson	0.029	0.343	Yes
H3	Directional Sensitivity	r(flip)	$r \geq 0$	Pearson (1-sided)	-0.791	<0.001	Yes
H4	Communication Channel	Δ_{pp}	$= 0$	t-test	0.981	0.327	No (ambiguous)
H5	Ambiguity Pooling	Δ_{pp}	$= 0$	t-test	-5.322	<0.001	Yes
H6	Censorship Distortion	Δ_{pp}	$= 0$	t-test	-2.284	0.023	Yes
H7	Surveillance Chilling	Δ_{pp}	$= 0$	t-test	-8.277	<0.001	Yes
H8	Propaganda Dose-Response	Δ_{pp}	$= 0$	t-test	-0.892	0.372	No (ambiguous)

Result 1 (Threshold-Policy Alignment)**.** *Across seven models and 1,800 country–periods in the pure global game treatment, the Pearson correlation between the empirical join fraction and the theoretical attack mass $A(\theta)$ averages $r = +0.80$ ($p < 0.001$ for every model).*

Table 4 reports results by model. Correlations range from $r = +0.75$ (Mistral Small Creative) to $r = +0.84$ (Trinity Large), with the pooled correlation at $r = +0.76$—lower than most individual models' because heterogeneous mean join rates across models add noise when pooling. The pooled OLS regression yields:

$$J = 0.13 + 0.54\, A(\theta), \quad R^2 = 0.48. \tag{6}$$

The slope of 0.54 indicates that LLM agents respond to the theoretical attack mass at roughly half the predicted rate—an attenuation expected when agents process narrative rather than numeric signals, since the briefing-to-belief mapping introduces noise that biases the slope toward zero (classical measurement error attenuation). The intercept of 0.13 reflects a baseline propensity to join even when the equilibrium predicts near-zero participation.[2]

The mean join rate across all models is 0.44, close to the theoretical mean.

The alignment is stable across architectures: correlations span $r \in [0.75, 0.84]$ despite parameter counts ranging from 30B to 235B (Table 4). Mean join rates vary—from 0.38 (Mistral) to 0.50 (Qwen3 30B)—reflecting model-specific action biases that shift the intercept but not the slope or correlation. In the language of the global

[2] Country-period observations within a model share calibration parameters and prompt structure, raising the possibility that standard errors understate uncertainty. The homoskedastic SE on the OLS slope is 0.014; HC1 (heteroskedasticity-robust) yields 0.013. Clustering by country inflates the SE to 0.049 but preserves significance ($p < 10^{-25}$). Clustering by model yields SE = 0.021 ($p < 10^{-55}$). All seven per-model correlations remain significant at $p < 0.001$ under country-clustered inference.

Table 4: Threshold-policy alignment by model and treatment. Cells report Pearson r between the empirical join fraction and the benchmark attack mass $A(\theta)$ under $B = C = 1$ (so $\theta^* = 0.50$); 95% Fisher-z confidence intervals in brackets for main treatments.

	Main treatments		Falsification			
Model	Pure	Comm	Scramble	Flip	n_{pure}	Mean join
Mistral Small Creative	+0.75 [0.72, 0.78]	+0.74 [0.71, 0.76]	+0.09	−0.81	1000	0.39
Llama 3.3 70B	+0.82 [0.74, 0.87]	+0.78 [0.69, 0.85]	+0.02	−0.84	100	0.44
Qwen3 30B	+0.76 [0.66, 0.83]	+0.77 [0.68, 0.84]	+0.13	−0.87	100	0.50
GPT-OSS 120B	+0.79 [0.74, 0.84]	+0.78 [0.72, 0.83]	−0.01	−0.80	200	0.41
Qwen3 235B	+0.80 [0.74, 0.84]	+0.76 [0.70, 0.81]	+0.06	−0.85	200	0.42
Trinity Large	+0.84 [0.77, 0.89]	+0.83 [0.75, 0.88]	+0.05	−0.80	100	0.46
MiniMax M2-Her	+0.82 [0.74, 0.87]	+0.82 [0.74, 0.87]	+0.01	−0.62	100	0.44
Pooled	+0.76 [0.74, 0.78]	+0.75 [0.73, 0.77]	+0.03	−0.79	1800	0.41
Mean across models	+0.80	+0.78	+0.05	−0.80	—	—

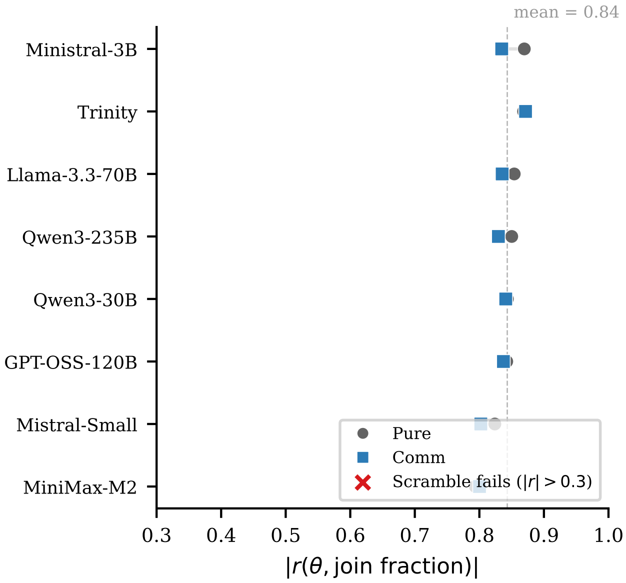


Figure 3: Cross-model summary of signal monotonicity. Points report $|r(\theta,\ \text{join})|$ under pure and communication; x markers (if any) indicate models where scrambling does not collapse the correlation ($|r| > 0.3$).

games model, different LLMs implement different cutoff strategies, but all respond monotonically to the underlying signal.

Logistic fits confirm threshold-like behavior but not level calibration. Estimated cutoffs are shifted relative to the canonical $\theta^* = 0.50$ benchmark and vary across models ($\hat{\theta}^* \in [-0.32, +0.10]$), reflecting baseline action bias and noise from translating narrative briefings into stated beliefs. The disciplined evidence on cutoff *location* comes from comparative statics: changing narrative stakes shifts $\hat{\theta}^*$ in the direction predicted by payoff theory (Table 8) and tracks theoretical θ^* almost perfectly in the B/C sweep (Figure 5). Communication consistently steepens the logistic ($\beta_{\text{comm}} > \beta_{\text{pure}}$ for all seven models, reaching 3.6 for Llama), suggesting that messages sharpen rather than blur the signal, even though the net effect on join rates is small (Appendix Table A12).

The positive correlation with $A(\theta)$ confirms that LLM behavior is monotone in the signal and sensitive to briefing content—necessary conditions for interpreting their behavior through the global-games comparative statics. The LLM's join curve is substantially steeper than a naive text-sentiment predictor (logistic slope 1.78 vs. the gradual text baseline; $r = 0.80$), suggesting processing beyond surface sentiment (Section 6). Belief elicitation reveals that agents form expectations tracking the theoretical success probability ($r = +0.79$) and predict actions beyond what signals alone explain (partial $r = +0.93$), consistent with strategic reasoning about others' likely behavior. I use "threshold-policy alignment" as shorthand for this pattern throughout, without claiming that agents approximate the Bayesian Nash equilibrium in the decision-theoretic sense.

Interpretation: What Threshold-Policy Alignment Means

(a) What the correlation measures. The Pearson r between J and $A(\theta)$ measures whether join rates track the monotone sigmoid shape predicted by global game theory—not just the direction, but the quantitative pattern across the full range of θ. A model that randomly joins 50% of the time, or that responds only to extreme signals, would not produce $r = +0.80$.

(b) What it does not establish. Agents do not observe payoffs (B, C), signal precision σ, or group size N—they process narrative without access to the mathematical objects defining the equilibrium. Whether the behavioral pattern reflects approximate Bayesian reasoning, a learned heuristic, or training-data associations is an open question the design cannot resolve. For AI interpretability, the relevant observation is that these models produce stated probability judgments that mediate the signal-to-action pathway (Pseudo $R^2 = 0.975$; Table A11, Column 3),

and that this mediation survives when beliefs are elicited *before* the decision ($r_{\text{pre}} = +0.82$; Figure 8), ruling out ex-post rationalization. The raw signal adds little predictive power once stated beliefs are included. Figure 4 shows the relationship nonparametrically: binned mean stated beliefs track the theoretical benchmark monotonically.

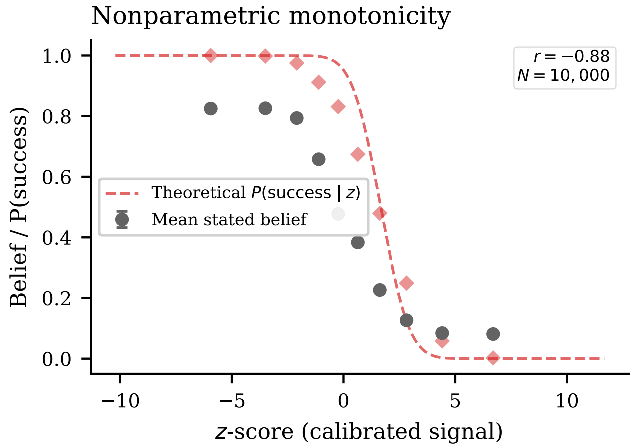


Figure 4: Nonparametric monotonicity: mean stated belief by z-score bin (circles), overlaid with theoretical $P(\text{success} \mid z)$ (dashed). Error bars show ±1 SE.

(c) A five-pronged identification strategy distinguishes strategic reasoning from text classification. Five results collectively rule out the possibility that agents merely classify briefing sentiment.

First, the cost/benefit test shifts the fitted cutoff in the direction predicted by payoff theory without disrupting the sigmoid shape (Table 8). Theory predicts that higher cost of failed action *lowers* the equilibrium cutoff (less joining), while lower cost *raises* it. The high-cost narrative ("severe reprisals—imprisonment, asset seizure, and retaliation against families") drops mean joining to 19.0% with cutoff $\hat{\theta}^* = 0.13$; the low-cost narrative ("minimal consequences—brief detentions at most") raises it to 69.3% with cutoff $\hat{\theta}^* = 0.72$; the baseline is 40.9% with $\hat{\theta}^* = 0.39$. Crucially, $|r| > 0.85$ in all three conditions—only the location shifts, while the monotone structure is preserved. A pure text classifier would not systematically shift cutoffs in the direction predicted by payoff theory.

A systematic sweep across seven B/C ratios ($\theta^* \in \{0.25, 0.33, 0.45, 0.50, 0.60, 0.67, 0.75\}$) confirms that $\hat{\theta}^*$ tracks θ^* monotonically with near-perfect correlation ($r = 0.997$, $p < 0.001$; Figure 5). Each condition runs 30 repetitions over a 9-point θ-grid (270 country–periods). The seven fitted cutoffs are perfectly rank-ordered ($\hat{\theta}^* = 0.19, 0.28, 0.42, 0.44, 0.54, 0.61, 0.72$)—this should be interpreted as an ordinal result (monotonic tracking) rather than quantitative calibration, since the θ^* grid was chosen by the researcher. The cutoffs are consistently below the theoretical target by approximately 0.04 pp, reflecting the slight pessimistic bias noted in the calibration.

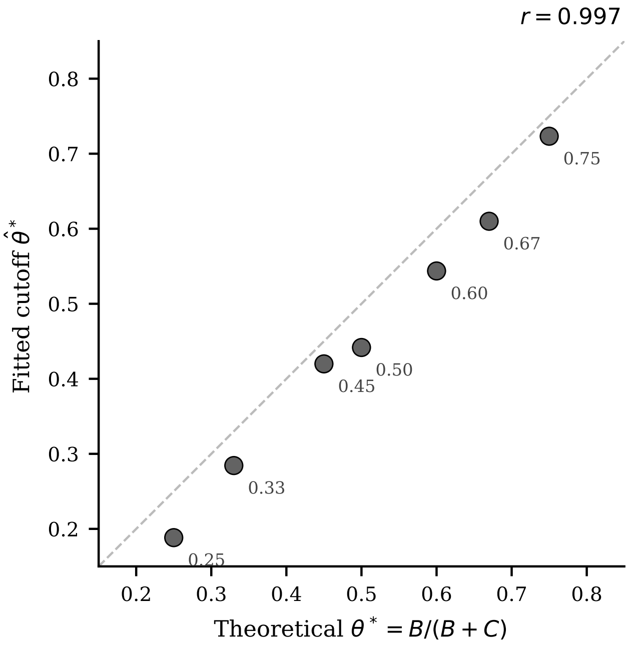


Figure 5: Fitted cutoff $\hat{\theta}^*$ vs. theoretical $\theta^* = B/(B+C)$ across seven benefit/cost ratios. Dashed line is 45°. Each point is a logistic fit to 270 country–periods (30 reps × 9 θ-grid values). $r = 0.997$.

Second, belief elicitation shows beliefs track the theoretical success probability ($r = +0.79$) and predict actions beyond what signals alone explain (partial $r = +0.93$), consistent with strategic inference about others' likely behavior.

Third, a coordination-cue experiment holds the direction slider (sentiment) fixed and varies only the coordination slope. Amplified coordination cues (×2.0) steepen the logistic slope to $\beta = 7.2$ vs. $\beta = 3.1$ under suppressed cues (×0.3), while overall join rates remain similar (40.4% vs. 41.1%). The difference is concentrated in the transition region ($\theta \in [0.42, 0.65]$), exactly where coordination cues should matter for threshold behavior. If agents were classifying sentiment alone, varying coordination independently of direction would produce no slope change.

A common-knowledge manipulation strengthens this result. When the briefing header states "this briefing has been distributed to all citizens—everyone has access to the same information," mean join rates rise by approximately 5 percentage points relative to a private-information framing ("based on your personal contacts and private intelligence"): 42.7–43.5% under common knowledge vs. 37.4–37.8% under private framing (270 country–periods per cell; all $|r| > 0.87$). The common-knowledge framing raises joining because agents who believe others received the same signal face less strategic uncertainty about coordination. This effect is orthogonal to sentiment—the direction slider is identical across conditions—and is con-

Table 5: Common knowledge × coordination intensity. Each cell reports mean join rate (270 country–periods). The CK main effect is +6.1 pp ($p = 0.0003$); the interaction is -1.2 pp ($p = 0.62$).

	Low coord	High coord	Δ (coord)
Private framing	37.4%	37.8%	+0.4 pp
CK framing	43.5%	42.7%	-0.8 pp
Δ (CK)	+6.1 pp	+4.9 pp	

Table 6: Classifier baselines. Accuracy and AUC are 5-fold CV on pure-treatment data. "Pred. join (surv.)" is the classifier's predicted join rate when applied to surveillance-treatment briefings; "Actual" is the LLM's observed rate. The gap measures the surveillance wedge invisible to text classifiers.

Classifier	Acc.	AUC	Pred. join (surv.)	Actual (surv.)	Gap
BoW TF-IDF	88.4%	0.947	40.5%	27.6%	13.0 pp
Slider logistic	88.4%	0.941	40.4%	27.6%	12.8 pp
Keyphrase	60.2%	0.594	8.4%	27.6%	-19.2 pp

sistent with agents processing the epistemic status of their information, not merely its valence. A formal interaction test confirms that common-knowledge framing and coordination intensity are additive rather than multiplicative ($\beta_{\text{interaction}} = -1.2$ pp, $p = 0.62$; Table 5).

Fourth, classifier baselines directly test whether text features alone can reproduce LLM behavior. A bag-of-words TF-IDF logistic regression trained on pure-treatment briefings achieves 88.4% accuracy and AUC = 0.947 within the training distribution—unsurprising, since briefings are designed to carry signal. But when this classifier is applied to surveillance-treatment decisions (where briefing text is identical but the communication prompt warns of monitoring), it predicts a 40.5% join rate—nearly identical to the pure baseline—while actual LLM join rates drop to 27.6%. The 12.9 pp gap between classifier prediction and actual behavior under surveillance cannot be reproduced by any text classifier, because the manipulation operates through the communication channel, not through briefing content. A slider-based logistic (direction, clarity, coordination) shows the same pattern: 88.4% accuracy in-distribution but 40.4% predicted join rate under surveillance. The surveillance wedge is invisible to classifiers that condition only on briefing features.

Fifth, construct validity: a three-feature model (direction, clarity, coordination) outperforms a one-feature sentiment baseline, indicating processing beyond surface tone.

Table 7: B/C comparative statics: classifier vs. actual LLM behavior. A logistic trained on baseline join rates (which captures the same information as slider features) predicts similar join rates across all payoff conditions. Actual LLM behavior shifts by ≈ 50 pp, demonstrating that agents respond to payoff information not captured by text features.

Condition	N	Classifier pred.	Actual	Gap (pp)
Baseline ($\theta^* = 0.50$)	270	40.9%	40.9%	+0.0
High cost ($\theta^* = 0.25$)	270	40.9%	19.0%	+21.9
Low cost ($\theta^* = 0.75$)	270	40.9%	69.3%	-28.4

Table 8: Cost/benefit narrative comparative statics. High cost: narrative emphasizes severe reprisals for failed action. Low cost: narrative emphasizes minimal consequences. Theory predicts higher perceived cost lowers the cutoff (less joining).

Design	N	Mean join	$r(\theta, J)$	Cutoff $\hat{\theta}^*$ (SE)	Δ vs baseline
Baseline	270	0.409	-0.88	0.39 (0.007)	—
High cost	270	0.190	-0.87	0.13 (0.010)	-0.218
Low cost	270	0.693	-0.88	0.72 (0.007)	+0.284

Together, these five tests form an identification strategy that no text classifier can replicate. A logistic trained on baseline signal features predicts ≈ 40.9% joining in all three B/C conditions; actual LLM join rates shift from 19.0% (high cost) to 69.3% (low cost)—a 50 pp swing invisible to any classifier conditioned on briefing text (Table 7). Payoff-theory cutoff shifts, coordination-cue slope changes, the surveillance belief–action wedge, and the common-knowledge effect are each orthogonal to textual sentiment, and collectively they rule out surface-level text classification as the generating process.

The correlation is also invariant to LLM decoding temperature ($r \in [-0.88, -0.87]$ across $T \in \{0.3, 0.7, 1.0\}$; Appendix B.11). What matters for the information design experiments in Parts II and III is that the behavioral regularity—monotone signal response—is robust enough to serve as a platform for studying how information structures shift coordination outcomes.

Notation convention. Part I reports $r(J, A(\theta))$, which is positive because both the attack mass and join fraction decrease in θ. Parts II and III (Section 8 onward) use a fixed θ-grid and report $r(J, \theta)$ directly, which is *negative* under alignment. The sign change reflects the convention, not a behavioral reversal.

6 Falsification Tests

The positive correlation admits an alternative explanation: LLM agents might produce stereotyped responses that correlate with regime strength for reasons unrelated to briefing content. The scramble and flip tests discriminate between this alternative and genuine signal extraction.

Result 2 (Signal Dependence)**.** *Cross-period scrambling of briefings reduces the mean within-country correlation from* $r = +0.80$ *to* $r = +0.05$ *across seven models. The pooled correlation drops from* $r = +0.76$ *to* $r = +0.03$ *(Fisher* $z = 25.09$, $p < 0.001$*).*

The scramble preserves the marginal distribution of briefing content but breaks the mapping from each period's θ to the signals agents receive—a format-preserving null that holds text length, vocabulary, and narrative structure constant while severing the informational link.[3] The collapse (+0.05 mean, +0.03 pooled) rules out the possibility that baseline alignment is driven by prompt aesthetics or surface formatting. The flip test provides a stronger check: every model shows clear sign reversal, confirming that all models respond to the directional content of the briefing.

Result 3 (Signal Direction)**.** *Inverting the signal direction flips the mean correlation from $r = +0.80$ to $r = -0.80$ across seven models. The pooled correlation moves from $r = +0.76$ to $r = -0.79$ (Fisher $z = 53.85$, $p < 0.001$).*

The flip negates the z-score before briefing generation, producing a near-symmetric reversal (+0.80 → −0.80) that rules out structural features of the prompt or model-specific tendencies as explanations.

The pure → scramble → flip pattern replicates across all seven models with full falsification suites (Table 4). Every model shows strong positive correlation under pure, collapse under scramble (within-country r), and sign reversal under flip.

The briefing generator maps z-scores monotonically to text—could a model that simply reads briefing sentiment, without any strategic reasoning, produce the observed sigmoid? To test this, I construct the simplest possible text-only predictor.

The generator assigns each briefing an internal *direction* score $d \in [0, 1]$, where $d = 1$ indicates regime-favorable language. A naive baseline predicts $\hat{p}_{\text{join}} = 1 - d$: join whenever the text sounds bad for the regime. This is the prediction a pure sentiment reader would make.

The correlation between this baseline and actual LLM decisions is $r = 0.80$—confirming that the text carries signal (as designed, since briefings are constructed to convey z-score content). However, the LLM's empirical join curve is substantially steeper than the text baseline (Figure 7). The fitted logistic has slope 1.78, producing a sharp transition around $z = 0$, while the text baseline drifts gradually from ≈ 0.93 to ≈ 0.10 across the full z-score range. The encoder is essentially monotone ($r(z, d) = 0.995$).

The gap between the text baseline and the empirical sigmoid indicates that the LLM sharpens the signal beyond surface sentiment, producing threshold-like behavior rather than linearly tracking the briefing's tone.

A stronger test asks whether agents form beliefs consistent with the equilibrium prediction. After each decision, I elicit stated beliefs ("On a scale from 0 to 100, how likely do you think the uprising will succeed?") under three treatments—pure, communication, and surveillance—each with 200 country–periods ($\approx$ 5,000 agent-level observations). Stated beliefs correlate strongly with the theoretical success probability $P(\text{success} \mid x_i) = \Phi[(\theta^* - x_i)/\sigma]$: $r = +0.79$ in pure ($p < 0.001$; Figure 8a), +0.79 under communication, and +0.78 under surveillance. Beliefs track the theoretical benchmark with systematic under-confidence (slope < 1), but the rank ordering is preserved across all treatments.

I use "theoretical success probability" rather than "Bayesian posterior" because agents do not observe the parameters defining the posterior—they have no access to B, C, σ, or the prior distribution. The benchmark is $P(\text{success} \mid x_i) = \Phi[(\theta^* - x_i)/\sigma]$ computed from the true parameters. The correlation measures monotonicity and rank-ordering of stated beliefs relative to this benchmark, not Bayesian updating *per se*.

Beliefs predict actions. In the pure treatment, the belief–action correlation is $r = +0.84$: agents with beliefs below 40% rarely join, while those above 80% almost always join. Under surveillance, this drops to $r = +0.73$—direct evidence of a belief–action wedge disrupting the link between private beliefs and public actions (Section 9). Crucially, beliefs predict decisions beyond what the signal alone predicts: the belief–action correlation ($r = +0.84$) exceeds what surface sentiment produces (text baseline $r = 0.80$), consistent with strategic reasoning about others' likely behavior.[4]

Second-order beliefs—agents' predictions about *others'* join rates—provide a sharper test of strategic reasoning. I elicit these by asking each agent: "Out of 100 citizens in a similar situation, how many do you think would choose to JOIN?" Across 200 country–periods per treatment ($\approx$ 5,000 agent observations each), second-order beliefs track the private signal ($r = -0.73$, $p < 0.001$) and vary monotonically with regime strength, consistent with agents reasoning about others' likely responses to correlated signals (Figure 9). Crucially, surveillance does *not* shift second-order beliefs (mean 31.2% → 30.9%, $\Delta = -0.3$ pp, $p = 0.59$) but *does* shift behavior (−13.5 pp). The result is a belief–behavior gap that *reverses direction* across treatments: in the pure treatment, agents predict 31% will join but 42% actually do (underprediction); under surveillance, agents still predict 31% but only 28.5% actually do (slight overprediction). The shift in behavior (−13.5 pp) dwarfs the shift in beliefs (−0.3 pp), precisely the signature of a belief–action wedge in the sense of Kuran (1991)—surveillance changes what agents *do* without changing what they *believe* others would do, because the chilling effect operates through self-censorship rather than through belief updating.

[3] Because the cross-period permutation operates within countries, the raw pooled correlation includes a between-country ecological confound. All scramble correlations therefore use within-country (country-demeaned) Pearson r, which isolates the signal-to-outcome link the falsification test is designed to assess.

[4] Belief elicitation data is from a single model (Mistral Small Creative). The behavioral patterns it explains—the surveillance chilling effect and the communication–action gap—replicate across three architectures (Mistral, Llama, Qwen3), suggesting the mechanism generalizes.

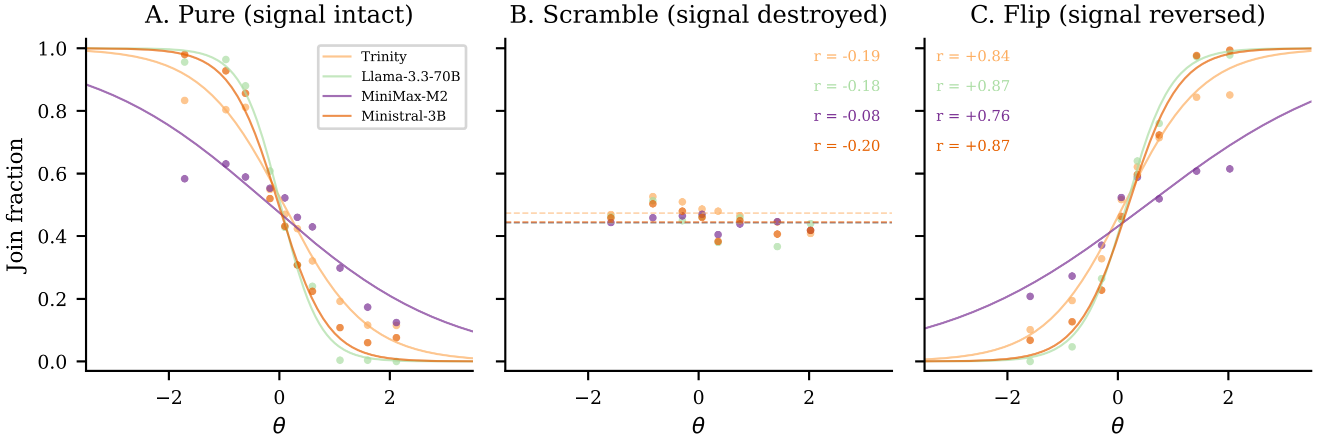


Figure 6: Falsification triptych. *Left:* Pure global game (mean $r = +0.80$). *Center:* Cross-period scramble breaks the θ-to-briefing mapping (mean within-country $r = +0.05$). *Right:* Signal flip inverts the mapping (mean $r = -0.80$). Each panel pools data from models with full falsification suites.

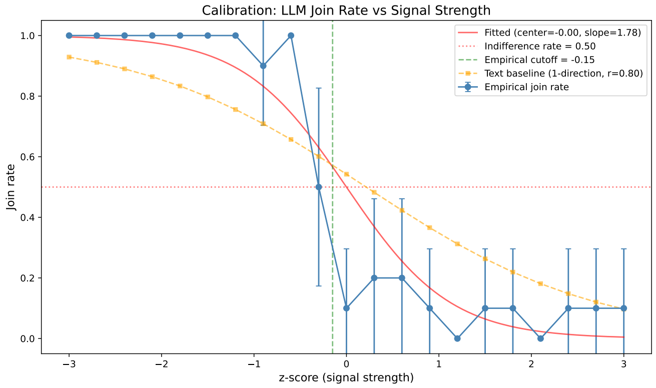


Figure 7: Text baseline identification test. Blue: empirical LLM join rate across z-scores. Orange: naive text-only predictor ($1 - \text{direction}$, $r = 0.80$). Red: fitted logistic (slope $= 1.78$). The LLM produces a steeper transition than the text baseline, indicating processing beyond sentiment reading. Mistral Small Creative, 210 observations.

7 Communication

Result 4 (Communication has a small, heterogeneous effect)**.** *Pre-play communication changes the mean join rate by -0.4 pp, from 0.435to 0.431, averaged across seven models. In the pooled sample, the unpaired difference is +1.23 pp (*$p = 0.327$*); effects vary in sign across models and are concentrated in weak-regime environments.*[5]

Communication preserves the signal structure (mean $r = +0.78$ under comm vs. $+0.80$ under pure) while introducing strategic uncertainty about others' actions. The effect on join rates is heterogeneous: five of seven models show positive effects ($+0.1$ to $+3.5$ pp), three show negative effects (-2.4 to -4.6 pp), and the pooled average is near zero. The asymmetry across θ is consistent with passive Bayesian updating: agents update toward joining when neighbors' correlated signals reveal regime weakness, with a floor effect preventing further declines under strong regimes where join rates are already near zero.

The belief elicitation data (Section 6) confirms that communication introduces strategic uncertainty without systematically shifting beliefs. Mean stated beliefs are identical under communication and pure (44.4%), yet communication creates the information topology that authoritarian instruments exploit—a channel that transmits both fundamentals and evidence of caution, producing a theoretically ambiguous net effect on coordination. The remaining sections show that this information topology—even with zero mean effect on beliefs—provides the surface area that surveillance, censorship, and propaganda require.

The communication effect is also sensitive to what agents know about the coordination environment. In a robustness check (Appendix B), agents are told "you are one of 25 citizens"—providing a basis for threshold reasoning absent in the main experiment. With group-size knowledge, the communication premium reverses: communication *lowers* join rates by 3.4 pp rather than raising them. When agents can reason about critical mass, messages revealing others' reluctance become more informative about the probability of reaching the coordination threshold, amplifying the deterrent effect of cautious peers. This reinforces the interpretation that communication's net effect on coordination is theoretically ambiguous: the same channel that transmits information about regime weakness also transmits evidence of others' caution.

[5] I report the unpaired (between-period) test as the primary specification because pure and communication treatments use independent θ draws, so country–periods are not naturally matched. A paired test that matches periods within each model by θ-rank yields a significant positive effect ($+5.5$ pp, $p < 0.001$, $n = 680$ pairs), reflecting within-θ variation that the unpaired test averages over. The qualitative conclusion—that the effect is small relative to baseline variation and heterogeneous across models—is robust to both approaches.

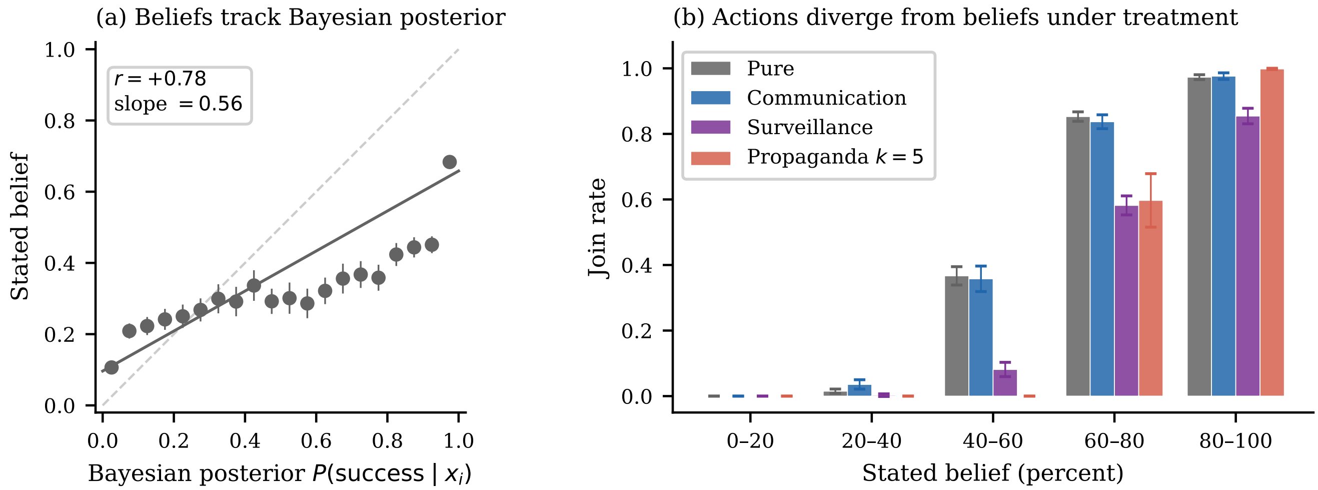


Figure 8: Belief elicitation results (Mistral Small Creative, 200 country–periods per treatment, ≈ 5,000 agent observations each). *Left:* Stated beliefs track the theoretical success probability $P(\text{success} \mid x_i)$ with $r = +0.79$ and systematic underconfidence (slope = 0.57). Dashed line: perfect calibration. *Right:* Join rate by stated belief bin under four treatments. Agents with 60–80% beliefs join at 86% in the pure treatment but only 58% under surveillance. Propaganda preserves the belief–benchmark correlation while suppressing actions—consistent with a mechanical rather than belief-based channel.

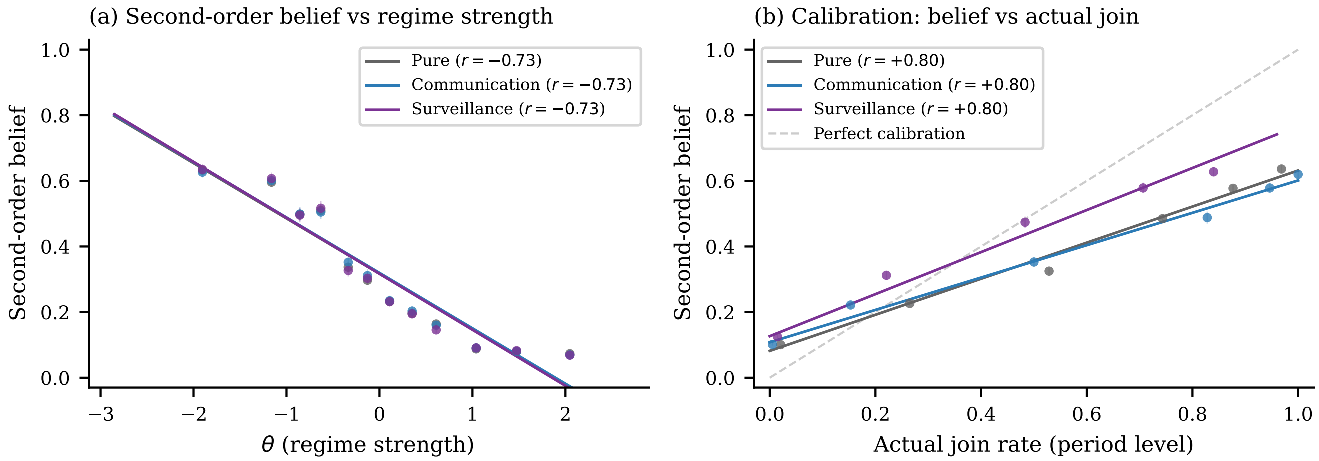


Figure 9: Second-order beliefs (Mistral Small Creative). *Left:* Mean second-order belief—agents' predicted join rate—decreases with regime strength θ across all treatments, confirming that beliefs track the private signal. Surveillance (purple) overlaps almost exactly with pure (gray), while communication (blue) slightly compresses the range. *Right:* Second-order belief vs. actual period-level join rate. Agents are approximately calibrated: the regression lines track the 45-degree perfect-calibration reference (dashed).

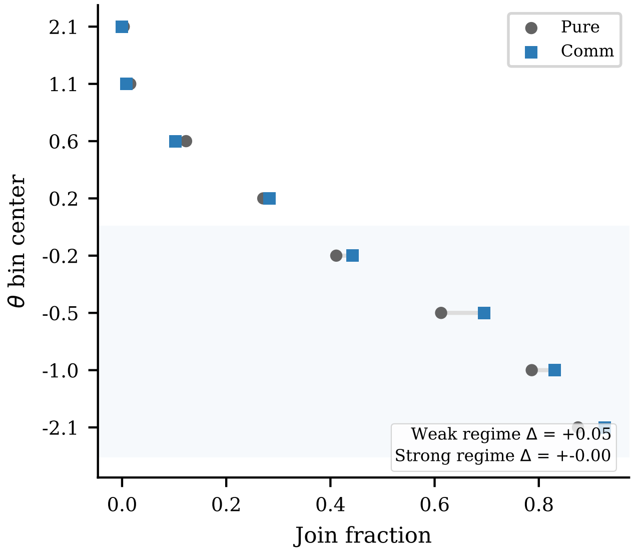


Figure 10: Communication effect by regime strength, pooled across seven models. Communication increases join rates for weak regimes ($\theta < \theta^*$) but has no effect or slightly reduces join rates for strong regimes ($\theta > \theta^*$).

8 Information Design

Part I established that LLM agents exhibit a stable, monotone sigmoid response to private signals—the behavioral regularity that makes coordination predictable. Part II asks: can a principal who understands this regularity exploit it? Information design in global games (Goldstein and Huang, 2016; Kolotilin et al., 2022) studies how a sender reshapes the mapping from states to signals to shift equilibrium coordination. Here the "sender" is the briefing generator's parameter space, and the "receiver" is the LLM agent population. Each design modifies the briefing generator's slider functions—clarity, direction, coordination—near the theoretical threshold θ^*, while the agents' decision process remains unchanged. The experiments test whether the theoretical predictions about optimal information structures (censorship, ambiguity injection, public signals) produce the predicted behavioral shifts when the agents are LLMs rather than Bayesian decision-makers.

Sign convention. From this section onward, I report $r(J, \theta)$ directly on a fixed θ-grid, which is *negative* under threshold-policy alignment. In Part I, $r(J, A(\theta)) > 0$ because both attack mass and joining decrease in θ; here the raw correlation with θ is reported.

Table 9 summarizes the main results. The baseline condition produces a mean join rate of 40.9%with a strong negative correlation between θ and join fraction ($r = -0.884$, $p < 0.001$).

Result 5 (Information Design Shifts Coordination)**.** *All three information designs produce measurable shifts in coordination relative to baseline.*

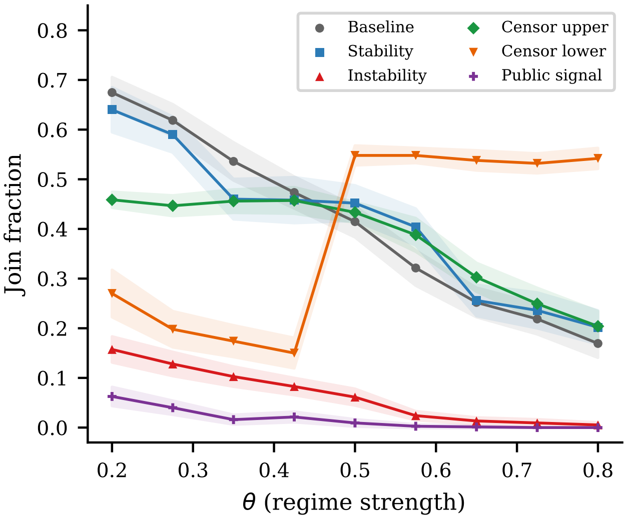


Figure 11: Join fraction as a function of θ under six information designs. Baseline, stability, and censorship designs have $N = 270$; instability and public signal have $N = 540$. Upper censorship pools weak-regime states; lower censorship pools strong-regime states and produces a dramatic reversal above θ^*. Mistral Small Creative model.

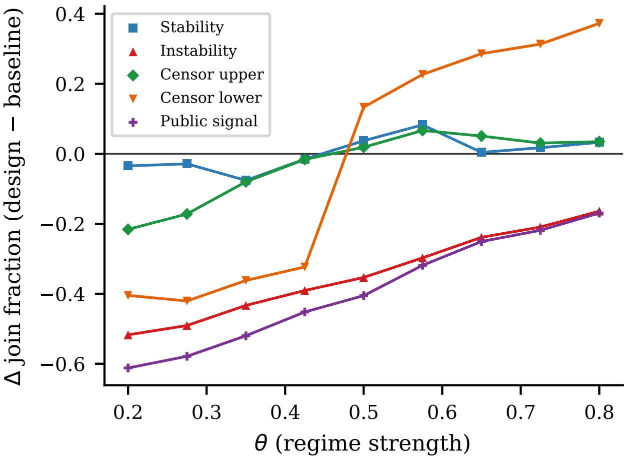


Figure 12: Treatment effect $\Delta(\theta)$ = design join − baseline join as a function of θ. Negative values indicate the design suppresses coordination.

Table 9: Information design treatment summary (primary model: Mistral Small Creative). r is the Pearson correlation between θ and join fraction.

Design	Mean	r	Δ	N
Baseline	0.409	-0.884	—	270
Stability	0.319	-0.626	-0.090	270
Instability	0.067	-0.740	-0.342	540
Public signal	0.017	-0.537	-0.392	540
Scramble	0.121	$+0.037$	-0.288	270
Flip	0.663	$+0.823$	$+0.255$	270

Notes: Data from the primary model (pure treatment; $\theta \in [0.20, 0.80]$ on a 9-point grid; N=25 agents per period). Mean join uses valid decisions; r is Pearson $r(\theta, \text{join})$ across rep-level periods.

The stability design suppresses coordination on average: mean join falls from 40.9%to 31.9%(-9.0 pp relative to baseline), and the θ–join relationship flattens ($r = -0.626$ vs. -0.884). The suppression is present at every θ grid point. This pattern is consistent with the design injecting ambiguity and mixed evidence near θ^*: weak-regime briefings retain stabilizing cues that deter participation even when fundamentals favor an uprising.

The instability design reduces the mean join rate to 6.7%(-34.2 pp relative to baseline). Sharper signals allow agents to more confidently distinguish strong from weak regimes, reducing participation across the grid.

The public signal produces the largest reduction in coordination: mean join rate falls to 1.7%(-39.2 pp relative to baseline). The shared bulletin is common knowledge and tends to dominate private briefings, sharply attenuating private-signal–driven participation. The correlation between θ and join fraction drops to $r = -0.537$, consistent with heavy weight on the public channel. This is consistent with the theoretical prediction of Morris and Shin (2002) that public signals receive disproportionate weight in coordination games: agents weight public information not only for its direct content but for its role as a coordination anchor under common knowledge. A public bulletin that is known to be shared provides a focal point for expectations about others' behavior, and its regime-issued framing may implicitly signal stability—compounding the coordination-suppression effect beyond what the informational content alone would predict.

Kolotilin et al. (2022) proved that when a sender's marginal utility is quasi-concave, the optimal information structure is upper censorship—one-sided pooling that conceals unfavorable states. In the language of Bayes-correlated equilibria, the regime designer chooses a signal structure that maximizes its objective subject to receivers' obedience constraints. Upper censorship implements this by pooling weak-regime states ($\theta \leq \theta^*$) into a neutral signal, so that agents who would otherwise observe evidence of regime vulnerability instead receive an uninformative briefing. Lower censorship applies the mirror image: strong-regime states are pooled. (The naming convention follows Kolotilin et al. (2022): "upper" refers to censoring the upper tail of the sender's *loss* distribution, which corresponds to pooling weak-regime states that are unfavorable to the regime.) I implement both designs.

Result 6 (Upper Censorship Suppresses Joining in Weak States)**.** *Upper censorship lowers the mean join rate to 37.7%(-3.1 pp vs. baseline) and attenuates the slope of the θ–join relationship ($r = -0.721$). The effect is concentrated in the censored region ($\theta \leq \theta^*$), where weak-regime states are pooled to a neutral briefing and join rates flatten.*

Pooling generates a flat join-rate "plateau" in the censored region: when agents cannot distinguish $\theta = 0.20$ from $\theta = 0.50$, they behave as if the regime is borderline rather than clearly weak.

A theory-consistent benchmark makes censorship common knowledge. When agents are told "regime censors are suppressing unfavorable intelligence above a certain severity threshold," the mean join rate returns close to the uncensored baseline (43.2% vs. 40.8% baseline; Table A7). Common knowledge of the censorship rule largely neutralizes the pooling distortion, consistent with theory assuming informed receivers (Kolotilin et al., 2022). The naïve censorship result documented above is therefore a behavioral extension: it documents what happens when receivers fail to account for the information structure—a realistic scenario in authoritarian regimes where the censorship apparatus is not common knowledge.

Result 7 (Lower Censorship Reverses Comparative Statics)**.** *Lower censorship produces a mean join rate of 39.0%(-1.9 pp vs. baseline) and flips the comparative statics: the θ–join correlation becomes positive ($r = +0.731$). Below θ^*, censoring favorable signals suppresses joining (agents see only weak-regime cues); above θ^*, pooling strong-regime states to a neutral briefing raises join rates sharply—agents who would otherwise see discouraging intelligence now receive uninformative briefings and default toward joining. The discontinuity at θ^* is consistent across repetitions (within-cell $\sigma < 0.04$) and replicates across models, though the direction of the reversal is model-dependent (Appendix B).*

Under the scramble condition, the correlation between θ and join fraction collapses to $r = +0.037$ ($p = 0.55$). Under the flip condition, the correlation inverts to $r = +0.823$ ($p < 0.001$) with mean join rate soaring to 66.3%. These results confirm that the information design effects operate through the intended signal channel.

A harder scramble test generates *all* briefings from a single fixed state ($\theta = 0.5$), severing any possible correlation between θ and briefing content by construction. Under this hard scramble, the θ–join correlation is $r = -0.057$ ($p = 0.35$, 270 country–periods) for Mistral and $r = -0.109$ ($p = 0.07$, 270 country–periods) for Llama—both indistinguishable from zero. Slider-independence diagnostics confirm that θ is uncorrelated with every briefing feature under hard scramble: $r(\theta, \text{direction}) = -0.008$ ($p = 0.53$), $r(\theta, \text{clarity}) = -0.014$

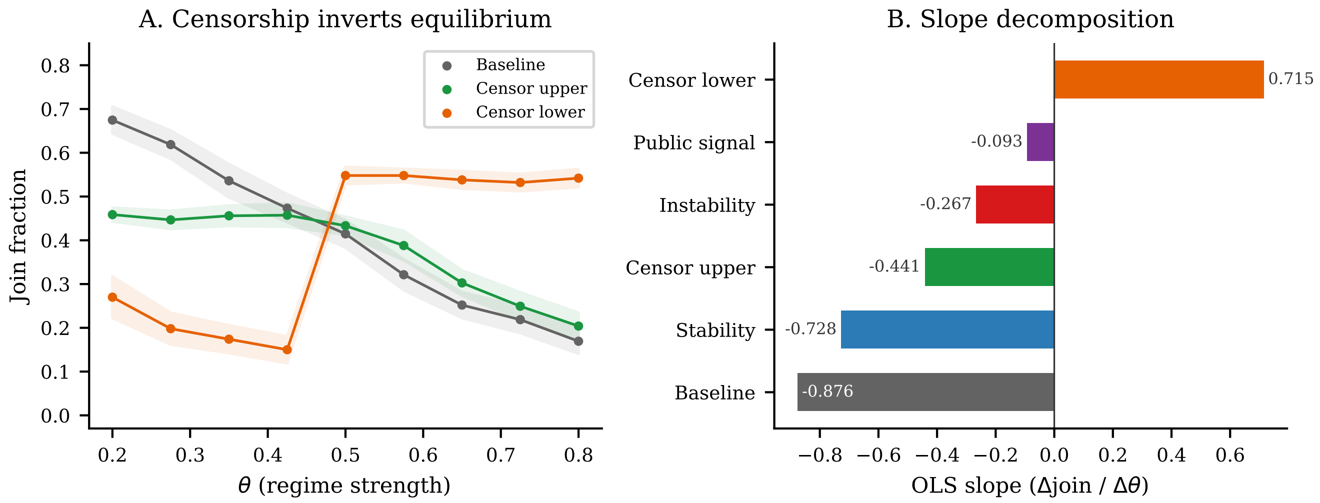


Figure 13: Censorship effects. *Left:* Join fraction under upper and lower censorship vs. baseline. Upper censorship pools weak-regime states ($\theta \leq \theta^*$), creating a flat plateau. Lower censorship pools strong-regime states ($\theta \geq \theta^*$) and produces a sharp reversal: join rates jump at θ^* and remain elevated. *Right:* OLS slope decomposition across all designs.

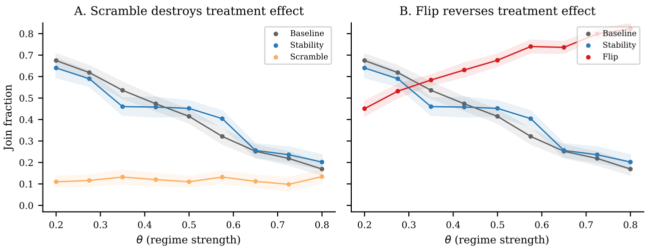


Figure 14: Falsification within information design. Scrambling collapses the θ-join correlation to $r = +0.037$; flipping inverts it to $r = +0.823$.

($p = 0.25$), $r(\theta, \text{coordination}) = +0.007$ ($p = 0.54$), each computed over 6,750 agent-level observations. The hard scramble closes the loophole that within-theta permutation might preserve residual ecological structure: when all briefings are drawn from the same state, no channel connects θ to behavior.

Bandwidth Sensitivity

The proximity bandwidth h controls how sharply information design concentrates manipulation around θ^*. Table A5 reports results for $h \in \{0.05, 0.15, 0.30\}$. Treatment effects are qualitatively robust but magnitudes vary: stability design effects peak at $h = 0.15$ because this bandwidth matches the θ-grid spacing. Narrower bandwidths ($h = 0.05$) concentrate manipulation so tightly that most grid points are unaffected; wider bandwidths ($h = 0.30$) dilute it across the full range. The baseline (no design) is unaffected by bandwidth since no proximity weighting is applied.

9 Surveillance and the Strategic Update Gap

Kuran (1991) argued that authoritarian regimes sustain themselves partly through what I term a *belief–action wedge*: the systematic suppression of expressed behavior while private beliefs remain intact. Kuran (1991) labeled this "preference falsification"; I adopt the more operationally precise term "belief–action wedge" because the experimental design measures the gap between stated beliefs and expressed actions rather than hidden preferences *per se*. I test this by introducing a surveillance treatment in the communication game.

In the surveillance treatment, the communication prompt is augmented with a warning that communications are being monitored by regime security services. The manipulation affects only the communication phase; the decision prompt is unchanged. The isolation is architectural: each LLM call is stateless, and the decision-stage prompt contains no reference to surveillance. Any difference in join rates must therefore arise from agents self-censoring their communications, not from a direct change in the perceived cost of joining.

Two placebo variants confirm this. A "monitored for research, no consequences" treatment and an "aggregated anonymously" treatment produce no significant deviation from the communication baseline (+2.6 pp and +4.1 pp respectively, both $p > 0.2$; Appendix B.12). The chilling effect is specific to the surveillance warning's strategic content—the implication that subversive messages will have consequences.

Result 8 (Surveillance Produces a Large Chilling Effect)**.** *In the primary model (Mistral Small Creative), surveil-*

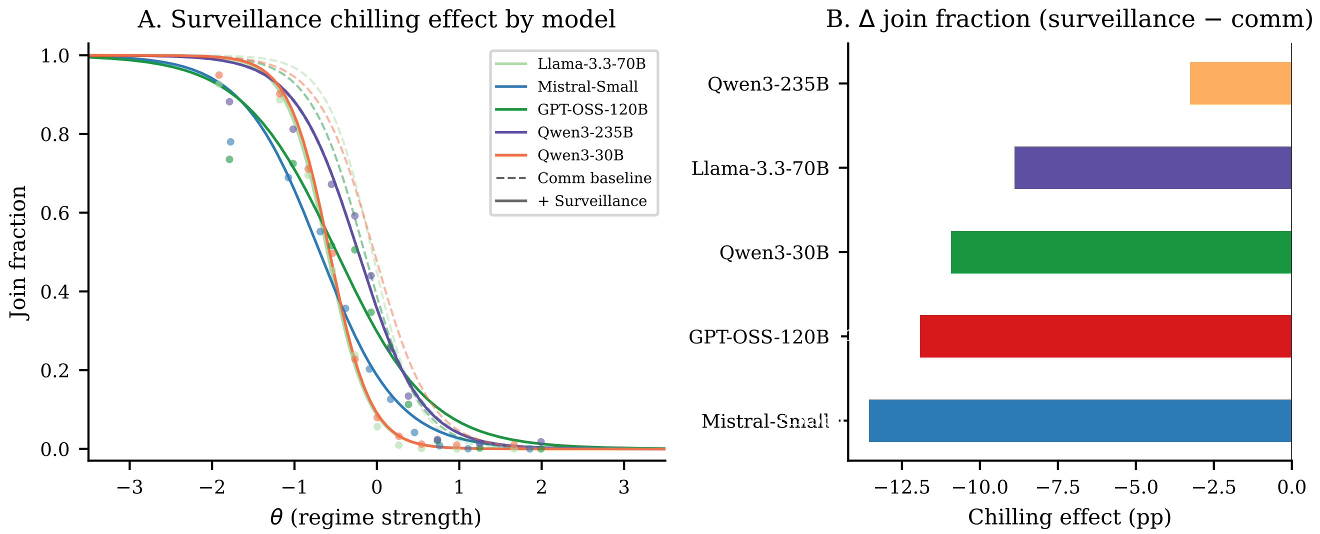


Figure 15: Join rates under regular communication vs. surveillance communication. Surveillance reduces join rates across all five models tested, averaging −9.7 percentage points: Mistral (−13.5 pp), GPT-OSS (−11.9 pp), Qwen3 30B (−10.9 pp), Llama (−8.9 pp), and Qwen3 235B (−3.3 pp).

lance reduces mean join rates from 41.1% to 27.6%, a difference of 13.5 percentage points ($p < 0.001$*). The correlation between θ and join fraction is preserved under surveillance (*$r = -0.820$ *vs.* -0.802*), indicating that surveillance operates as a level shift rather than disrupting signal processing.*

Surveillance contaminates the information environment with self-censored messages, pushing join rates well below the communication baseline. The effect replicates across five models: Mistral (−13.5 pp), GPT-OSS (−11.9 pp), Qwen3 30B (−10.9 pp), Llama (−8.9 pp), and Qwen3 235B (−3.3 pp), averaging −9.7 pp.

Surveillance operates through two channels. The *indirect channel* works through the communication phase: agents self-censor their messages, degrading the informational content available to all participants. The *direct channel* operates at the decision stage: the mere awareness of monitoring suppresses action even when message content is held fixed. The remainder of this section documents each channel and their interaction.

Indirect channel: self-censored messages degrade information. The belief elicitation data is consistent with a belief–action wedge in the sense of Kuran (1991) rather than belief updating. Surveillance shifts stated beliefs by only 0.7 pp relative to pure ($p = 0.25$, not significant) while shifting join rates by 13.5 pp ($p < 0.001$)—a ratio of nearly 20:1. The belief–action correlation drops sharply under surveillance ($r = +0.73$ vs. $+0.84$ under pure), measuring the behavioral wedge between stated beliefs and expressed actions. Whether this wedge reflects genuinely hidden preferences or a simpler pattern-matching response to surveillance-related language is an open question; the key finding is the *size and robustness* of the wedge across architectures.

Elicited punishment risk provides further evidence. After each JOIN/STAY decision, I ask agents to rate the expected severity of regime punishment on a 0–10 scale. Across 15,000 agent decisions (two models × three conditions), mean elicited risk is approximately 8.0/10 in all conditions—pure, communication, and surveillance alike (Table A15 in the appendix). Critically, stated punishment risk is virtually identical for agents who chose JOIN and those who chose STAY (difference < 0.2 points in all conditions). The behavioral wedge created by surveillance is therefore not accompanied by a shift in stated risk perception: agents who self-censor under surveillance do not report higher expected punishment. This is consistent with the chilling effect operating through the communication channel (self-censored messages contaminate others' information) rather than through a direct change in perceived costs.

The most striking finding concerns second-order beliefs. Agents' predictions about *others'* join rates are essentially unchanged by surveillance (31.2% → 30.9%, $p = 0.59$; Section 6). This reveals a *strategic update gap*: surveillance alters each agent's individual threshold for expressing dissent without updating their model of the population threshold. Agents self-censor because they fear personal consequences, but they do not realize that everyone else is also self-censoring. They interpret others' silence as genuine regime support rather than as the product of the same fear they themselves experience. This failure to reach common knowledge of the chilling effect is what makes the belief–action wedge so durable—each agent believes they are *uniquely* cautious, sustaining the equilibrium even though the underlying beliefs that would support coordination remain intact. Once agents expect others to self-censor, even authentic messages become uninformative, and the entire communication channel is poisoned.

The information leakage confirms this quantitatively. Regressing the true state θ on observable linguistic features of the messages, the R^2 drops from 0.12 under regular communication to 0.02 under surveillance—an 80% reduction in the informational content transmitted through the network. Surveillance does not merely mute the channel; it structurally degrades it, replacing informative signals with self-censored noise that is uninformative about the true state.

Direct channel: decision-stage suppression. A message-content hold-fixed test isolates the decision-stage channel. Baseline communication messages are replayed verbatim so that agents receive identical briefings and identical peer messages, but the decision prompt warns that their actions may be monitored. Mean join rates drop from 41.1% to 20.2%, a -20.9 pp shift ($t = -6.89$, $p < 0.001$). This confirms that surveillance operates through *both* channels: self-censored messages degrade the information environment (the indirect channel tested in the main design), *and* the mere awareness of monitoring suppresses action even when message content is held fixed (the direct decision-stage channel isolated here). The decision-stage effect (-20.9 pp) exceeds the overall surveillance effect (-13.5 pp), suggesting that in the main design, pro-coordination messages partially offset the chilling effect of monitoring awareness.

Message content confirms the self-censorship directly. Across 15,000 communication messages and 25,000 surveillance messages, surveillance agents write shorter messages (253 vs. 342 characters) and systematically avoid action language: "act" appears in 19.0% of communication messages but only 3.8% under surveillance; "collapse" drops from 11.6% to 0.6%; "together" from 7.2% to 0.7%. In its place, hedged and cautious language rises: "careful" (1.5% → 4.8%), "stable" (1.7% → 4.2%), "patience" (0.3% → 2.4%). Among agents who privately decide to JOIN, only 19.0% send action-signaling messages under surveillance, compared to 46.1% under regular communication. Surveillance agents who intend to join write messages like *"the air feels lighter... keep your eyes open, but step carefully"* rather than *"the walls are cracking—the machine is grinding to a halt."* The decision is the same; the expressed sentiment is not.

10 Propaganda: Information Contamination

Edmond (2013) modeled propaganda as the regime shifting citizens' signal distributions. I implement this by introducing propaganda agents—regime plants who transmit fixed pro-regime messages and always STAY.

I distinguish the *overall* join rate (including propaganda agents, who always STAY) from the *real citizen* join rate (excluding plants). The overall rate captures the mechanical dilution of the attack mass; the real citizen rate isolates the behavioral effect of propaganda on genuine agents' decisions.

Result 9 (Propaganda Suppresses Coordination Primarily Through Mechanical Dilution)**.** *Mean join fraction (including plants) falls from 41.1% ($k = 0$) to 37.5% ($k = 2$), 31.3% ($k = 5$), and 23.3% ($k = 10$). However, the behavioral effect on real citizens is much smaller and saturates: 41.1% ($k = 0$), 40.7% ($k = 2$, -0.4 pp), 39.1% ($k = 5$, -2.0 pp), 38.8% ($k = 10$, -2.3 pp).*

Propaganda works through two channels: a *mechanical* channel (plants always STAY, directly reducing attack mass) and a *behavioral* channel (pro-regime messages reduce real citizens' willingness to join). The mechanical channel is approximately linear in k; the behavioral channel saturates quickly—doubling plants from 5 to 10 produces no additional behavioral effect (-0.3 pp, $p = 0.67$). This implies sharply diminishing returns: the regime's first few plants yield both mechanical and behavioral suppression, but additional plants contribute only mechanical dilution. At $k = 10$ (40% of the network), real citizens' join rate has barely moved from $k = 5$ (39.1% vs. 38.8%).

The propaganda effect replicates with Llama 3.3 70B, which shows a behavioral effect of -2.7 pp at $k = 5$, confirming the qualitative pattern and the saturation across architectures.

Message content reveals the mechanism. Propaganda agents inject regime-loyal vocabulary into the communication network, and this language propagates to real agents. The fraction of messages containing "loyal" rises from 1.5% at baseline to 3.5% ($k = 2$), 6.1% ($k = 5$), and 11.4% ($k = 10$); "patience" rises from 0.3% to 5.1%. Meanwhile, coordination language declines: "ready" falls from 30.5% to 18.5%, "together" from 7.2% to 4.2%. Message length also shrinks (342 → 285 characters), consistent with the shorter, punchier pro-regime messages diluting the discourse. Among real agents who STAY, the fraction sending caution-coded messages rises from 24.2% (baseline) to 38.2% ($k = 10$)—agents are not merely responding to propaganda but *echoing* it. Among those who JOIN, however, action signaling remains stable at $\approx 86\%$ across all conditions. The behavioral saturation documented above thus has a linguistic correlate: propaganda shifts the discourse for agents on the margin, but agents with strong anti-regime signals continue to express and act on their beliefs regardless of the propaganda dose.

11 Instrument Interactions

A regime deploys surveillance, censorship, and propaganda jointly. This section tests whether the instruments interact as substitutes (diminishing returns) or complements (super-additive suppression).

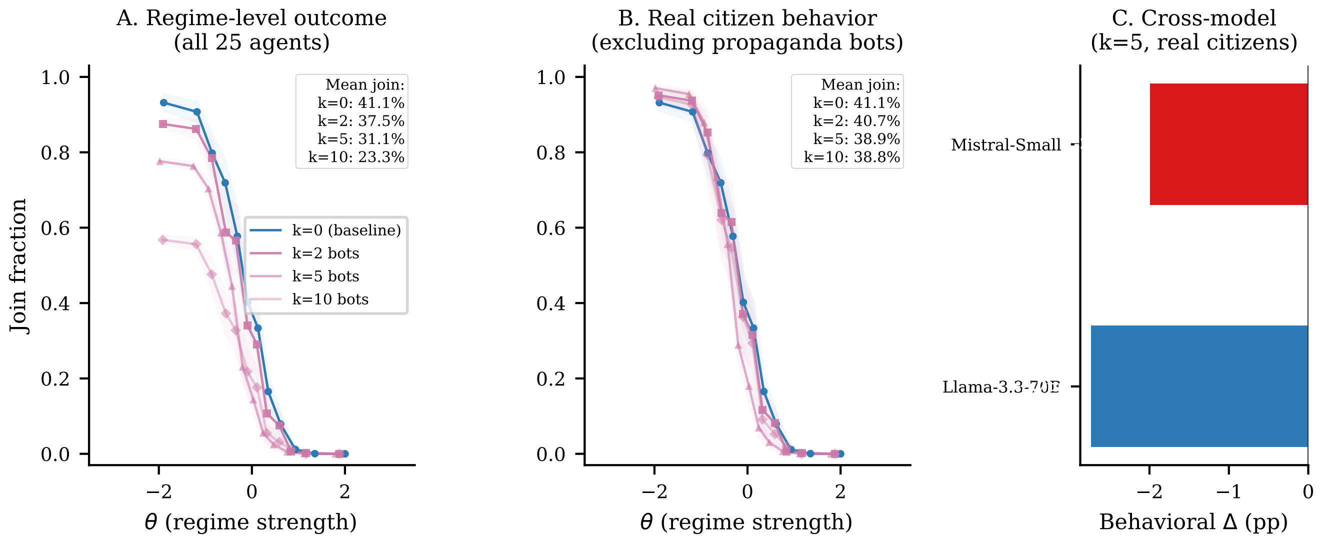


Figure 16: Dose-response relationship between number of propaganda agents and mean join rate. Results shown for Mistral (primary) and Llama (replication). Regular communication ($k = 0$) serves as baseline.

Table 10: Propaganda and surveillance effects (primary model: Mistral Small Creative). "All" includes propaganda agents; "Real" excludes them (computed from logs). Δ is the change in real-agent mean join vs. baseline communication.

	Mean join			
Treatment	All	Real	r	Δ
Comm (baseline)	.411	.411	−0.802	—
Prop $k = 2$	.375	.407	−0.809	-0.004
Prop $k = 5$	.313	.391	−0.822	-0.020
Prop $k = 10$	.233	.388	−0.818	-0.023
Surveillance	.276	.276	−0.820	-0.135
Prop+Surv	.194	—	−0.829	—

Result 10 (Propaganda + Surveillance: Approximately Additive)**.** *When propaganda ($k = 5$) and surveillance are combined, the mean join rate among real citizens falls to 24.3%, a reduction of 16.8 pp from the communication baseline (41.1%). The sum of individual effects is 15.5 pp (surveillance* −13.5 *pp + propaganda* −2.0 *pp), so the combined effect (16.8 pp) is approximately additive. Once surveillance has suppressed expressed dissent, propaganda adds only modest additional deterrence.*

Result 11 (Surveillance × Censorship: Super-Additive)**.** *Table 11 shows that surveillance and censorship interact strongly: surveillance sharply suppresses coordination, and its marginal effect is substantially larger under censorship than at baseline. In this sense the interaction is super-additive—censorship increases reliance on the communication channel, and surveillance poisons that channel.*

Table 11: Surveillance × censorship interaction in the communication game (primary model: Mistral Small Creative).

Design	No Surv.	Surv.	Δ
Baseline	0.030	0.009	-0.021
Upper cens.	0.151	0.037	-0.114
Lower cens.	0.177	0.042	-0.135

Notes: "No Surv." uses the communication infodesign grid. "Surv." uses the same grid with surveillance active during messaging. All entries are means of `join_fraction_valid`.

Surveillance and censorship are complements that attack different links in the coordination chain. Censorship removes the private information channel, forcing agents to rely on communication for their signals about regime strength. Surveillance then poisons that communication channel through the belief–action wedge. With both instruments active, agents have neither private signals to trust nor authentic messages to learn from—the informational foundations of coordination are eliminated from both directions.

This complementarity is the mechanism behind the paper's headline result: pooling interventions can shift coordination by distorting private information, but once surveillance contaminates the messaging stage, the same communication channel becomes a lever for suppressing coordination. The regime does not need each instrument to be independently decisive; it needs the combination to close every informational pathway through which coordination might flow.

The interaction between surveillance and censorship is heterogeneous across architectures (Table 12). Under

Table 12: Cross-model surveillance × censorship interaction. All conditions run under surveillance with communication. Δ columns show the change relative to the surveilled baseline.

	Mean join (surv.)			Δ vs baseline	
Model	Baseline	Upper cens.	Lower cens.	Δ upper	Δ lower
Mistral Small Creative	0.009	0.037	0.042	+0.028	+0.033
Llama 3.3 70B	0.114	0.039	0.115	-0.075	+0.001
GPT-OSS 120B	0.316	0.177	0.312	-0.139	-0.004
Qwen3 235B	0.336	0.321	0.468	-0.015	+0.131

communication with surveillance, baseline join rates range from near zero (Mistral, 0.9%) to roughly one-third (GPT-OSS 120B, 31.6%; Qwen3 235B, 33.6%). Under surveillance, upper censorship further suppresses coordination for Llama 70B and GPT-OSS 120B, but has little effect for Qwen3 235B and *raises* joining modestly for Mistral. Lower censorship is similarly mixed: it has essentially no effect for Llama and GPT-OSS, but increases join rates for Mistral and Qwen3 235B. The regime-control instruments therefore do not combine mechanically; the joint effect depends on model-specific resolution of pooled private signals and self-censored messages.

12 Conclusion

The central finding of this paper is that the information channel is a trap. Modern authoritarianism relies less on terror and more on information manipulation (Guriev and Treisman, 2019). The global games framework clarifies why this is effective: coordination requires overcoming strategic uncertainty, which necessitates communication. But the very act of opening a communication channel provides the regime with the surface area required to deploy surveillance and censorship. Any channel that transmits information about others' willingness to act also transmits *uncertainty* about others' willingness to act, and that uncertainty is exploitable.

In this simulation, large shifts in joining can occur with minimal movement in elicited beliefs. The regime does not need to change what citizens privately believe; it needs only to fracture the common knowledge of those beliefs. Communication does not shift agents' beliefs about success (44.4% under both pure and communication), yet the channel it opens is vulnerable. Surveillance compounds this (−13.5 pp for Mistral, −11.1 pp on average) through a belief–action wedge consistent with Kuran (1991): agents maintain their stated beliefs but suppress expressed behavior, generating a cascade of uninformative messages that poisons the channel for everyone. The strategic update gap documented in Section 9—second-order beliefs unchanged (31.2% → 30.9%) while actual join rates fall by 13.5 pp—shows that surveillance operates asymmetrically, altering individual thresholds without updating agents' models of the population threshold. Each agent suppresses dissent while interpreting others' silence as genuine.

Censorship and surveillance are complements that attack different links in the coordination chain. Censorship pools private signals, forcing agents to rely on communication; surveillance then poisons that communication channel. With both instruments active, agents have neither private signals to trust nor authentic messages to learn from. Propaganda's behavioral channel, by contrast, is small and saturates quickly (the effect is largely exhausted by $k = 5$ plants), implying diminishing returns; the marginal authoritarian dollar is better spent on surveillance than on additional propaganda.

I do not claim that LLMs are Bayesian agents. But across seven models spanning six architecture families (mean $r = +0.80$, $p < 0.001$), the behavioral regularities are precisely what the global games framework predicts: monotone signal response, threshold-like decisions, sensitivity to information design, and a surveillance-induced belief–action wedge consistent with Kuran (1991). The consistency across architectures spanning 30B to 235B parameters provides a robustness demonstration rather than population inference—the seven models are a convenience sample, not a representative draw from a well-defined population of architectures. The question is not whether LLMs reason identically to humans, but whether the regularities are robust enough to serve as a computational laboratory for predictions that are difficult to test otherwise. The full regime change game has resisted laboratory implementation because it requires rich private signals, genuine strategic uncertainty, and large groups. LLM agents sidestep these constraints, and the same platform extends naturally to currency crises, bank runs, and other coordination games where information processing is central to behavior. Appendix D discusses implications for AI alignment.

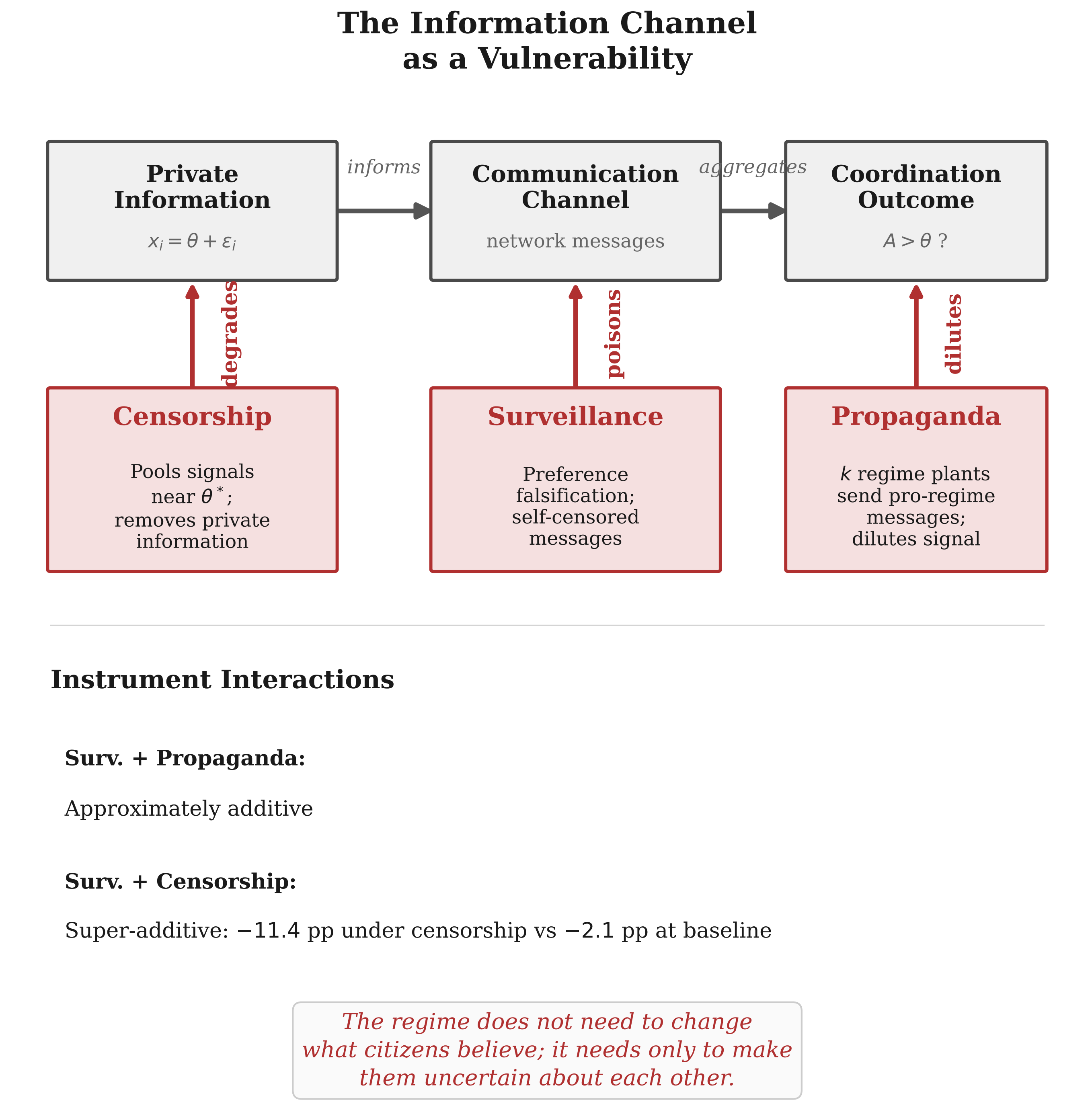

The Information Channel as a Vulnerability
Private Information
$x_i = \theta + \varepsilon_i$
informs
Communication Channel
network messages
aggregates
Coordination Outcome
$A > \theta$?
degrades
poisons
dilutes
Censorship
Pools signals near θ^*; removes private information
Surveillance
Preference falsification; self-censored messages
Propaganda
k regime plants send pro-regime messages; dilutes signal
Instrument Interactions
Surv. + Propaganda:
Approximately additive
Surv. + Censorship:
Super-additive: −11.4 pp under censorship vs −2.1 pp at baseline
The regime does not need to change what citizens believe; it needs only to make them uncertain about each other.

Figure 17: How authoritarian instruments attack the coordination chain. Surveillance poisons the communication channel through a belief–action wedge (self-censorship); censorship degrades the private signal channel by pooling states; propaganda contaminates the communication channel mechanically. Surveillance and censorship are complements (super-additive), while propaganda's behavioral effect saturates quickly.

References

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. Playing repeated games with large language models. *Nature Human Behaviour*, 9:1380–1390, 2025.

George-Marios Angeletos, Christian Hellwig, and Alessandro Pavan. Dynamic global games of regime change: Learning, multiplicity, and the timing of attacks. *Econometrica*, 75(3):711–756, 2007.

Ala Avoyan. Does cheap talk promote coordination under asymmetric information? An experimental study on global games. *Journal of Economic Behavior & Organization*, 169:204–224, 2020.

Dirk Bergemann and Stephen Morris. Information design, Bayesian persuasion, and Bayes correlated equilibrium. *American Economic Review*, 106(5):586–591, 2016.

Dirk Bergemann and Stephen Morris. Information design: A unified perspective. *Journal of Economic Literature*, 57(1):44–95, 2019.

Andreas Blume and Andreas Ortmann. The effects of costless pre-play communication: Experimental evidence from games with Pareto-ranked equilibria. *Journal of Economic Theory*, 132(1):274–290, 2007.

Andrea Carlini et al. Large language models show amplified cognitive biases in moral decision-making. *Proceedings of the National Academy of Sciences*, 122, 2025.

Hans Carlsson and Eric van Damme. Global games and equilibrium selection. *Econometrica*, 61(5):989–1018, 1993.

Erin Baggott Carter and Brett L. Carter. Propaganda and protest in autocracies. *Journal of Conflict Resolution*, 65(5):919–949, 2021.

Vincent Crawford and Joel Sobel. Strategic information transmission. *Econometrica*, 50(6):1431–1451, 1982.

Douglas W. Diamond and Philip H. Dybvig. Bank runs, deposit insurance, and liquidity. *Journal of Political Economy*, 91(3):401–419, 1983.

Chris Edmond. Information manipulation, coordination, and regime change. *Review of Economic Studies*, 80(4):1422–1458, 2013.

Tore Ellingsen and Robert Östling. When does communication improve coordination? *American Economic Review*, 100(4):1695–1724, 2010.

Ruben Enikolopov, Alexey Makarin, and Maria Petrova. Social media and protest participation: Evidence from Russia. *Econometrica*, 88 (4):1478–1514, 2020.

Joseph Farrell and Matthew Rabin. Cheap talk. *Journal of Economic Perspectives*, 10(3):103–118, 1996.

David M. Frankel, Stephen Morris, and Ady Pauzner. Equilibrium selection in global games with strategic complementarities. *Journal of Economic Theory*, 108(1):1–44, 2003.

Chen Gao et al. Validation is the central challenge for generative social simulation: A critical review of LLMs in agent-based modeling. *Artificial Intelligence Review*, 58, 2025.

Itay Goldstein and Chong Huang. Bayesian persuasion in coordination games. *American Economic Review: Papers & Proceedings*, 106(5): 592–596, 2016.

Igor Grossmann et al. Do large language models solve the problems of agent-based modeling? A critical review of generative social simulations. arXiv preprint arXiv:2504.03274, 2025.

Sergei Guriev and Daniel Treisman. Informational autocrats. *Journal of Economic Perspectives*, 33(4):100–127, 2019.

Frank Heinemann, Rosemarie Nagel, and Peter Ockenfels. The theory of global games on test: Experimental analysis of coordination games with public and private information. *Econometrica*, 72(5):1583–1599, 2004.

Frank Heinemann, Rosemarie Nagel, and Peter Ockenfels. Measuring strategic uncertainty in coordination games. *Review of Economic Studies*, 76(1):181–221, 2009.

Leif Helland, Sturla Holm, and Maren Saethre. Information quality and regime change: Evidence from the lab. *Journal of Economic Behavior & Organization*, 191:538–554, 2021.

John J. Horton. Large language models as simulated economic agents: What can we learn from homo silicus? Working Paper 31122, National Bureau of Economic Research, 2023.

Siyuan Huang et al. How ethical should AI be? How AI alignment shapes the risk preferences of LLMs. arXiv preprint arXiv:2406.01168, 2024.

Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*, 2019.

Nicolas Inostroza and Alessandro Pavan. Adversarial coordination and public information design. *Theoretical Economics*, 20:763–813, 2025.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011.

Gary King, Jennifer Pan, and Margaret E. Roberts. How censorship in China allows government criticism but silences collective expression. *American Political Science Review*, 107(2):326–343, 2013.

Anton Kolotilin, Tymofiy Mylovanov, and Andriy Zapechelnyuk. Censorship as optimal persuasion. *Theoretical Economics*, 17:561–585, 2022.

Timur Kuran. Now out of never: The element of surprise in the East European revolution of 1989. *World Politics*, 44(1):7–48, 1991.

Laurent Mathevet, Jacopo Perego, and Ina Taneva. On information design in games. *Journal of Political Economy*, 128(4):1370–1404, 2020.

Stephen Morris and Hyun Song Shin. Unique equilibrium in a model of self-fulfilling currency attacks. *American Economic Review*, 88(3): 587–597, 1998.

Stephen Morris and Hyun Song Shin. Social value of public information. *American Economic Review*, 92(5):1521–1534, 2002.

Stephen Morris and Hyun Song Shin. Global games: Theory and applications. In Mathias Dewatripont, Lars Peter Hansen, and Stephen J. Turnovsky, editors, *Advances in Economics and Econometrics*, pages 56–114. Cambridge University Press, 2003.

Maurice Obstfeld. Models of currency crises with self-fulfilling features. *European Economic Review*, 40(3–5):1037–1047, 1996.

Jon W. Penney. Chilling effects: Online surveillance and Wikipedia use. *Berkeley Technology Law Journal*, 31(1):117–182, 2016.

Aleksandr Petrov et al. LLM strategic reasoning: Agentic study through behavioral game theory. arXiv preprint arXiv:2502.20432, 2025.

Olga Shurchkov. Coordination and learning in dynamic global games: Experimental evidence. *Experimental Economics*, 16(2):313–334, 2013.

Elizabeth Stoycheff. Under surveillance: Examining Facebook's spiral of silence effects in the wake of NSA internet monitoring. *Journalism and Mass Communication Quarterly*, 93(2):296–311, 2016.

Haoming Sun et al. Game theory meets large language models: A systematic survey with taxonomy and new frontiers. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2025.

Michal Szkup and Isabel Trevino. Sentiments, strategic uncertainty, and information structures in coordination games. *Games and Economic Behavior*, 124:534–553, 2020.

Table A1: Single-channel decomposition of the stability design (primary model: Mistral Small Creative).

Channel	Mean	r	Δ
Full stability	0.319	-0.626	-0.090
Clarity only	0.405	-0.887	-0.004
Direction only	0.389	-0.876	-0.020
Dissent only	0.406	-0.882	-0.002
Sum of channels	—	—	-0.026
Full design	—	—	-0.090

Notes: Each row is a separate infodesign run for Mistral Small Creative on the same θ grid as Table 9. Δ reports the mean difference vs. the baseline infodesign mean (Table 9).

A Decomposition: Which Channel Drives the Stability Effect?

The stability design manipulates three channels simultaneously (direction, clarity, dissent). To determine which drives the effect, I run three single-channel treatments, each activating only one manipulation while holding the other two at baseline.

Each channel alone produces a large suppression of joining relative to baseline: clarity only (-0.4 pp), direction only (-2.0 pp), and dissent only (-0.2 pp).

Summing the single-channel effects yields -2.6 pp, far smaller in magnitude than the bundled stability design effect (-9.0 pp). This implies strong super-additivity: each channel alone has a modest effect, but the combination produces a suppression nearly four times the sum of its parts. The channels are complements—ambiguity (clarity), directional flattening, and dissent framing interact to undermine coordination in ways that no single manipulation achieves alone.

B Robustness

These checks show that threshold-policy alignment and the qualitative information design effects are stable to agent count, network density, and the proximity bandwidth.

B.1 Agent Count Variation

I vary the number of agents per period ($n \in \{5, 10, 25, 50, 100\}$) using Mistral Small Creative. The correlation is stable: $r = +0.60$ ($n = 5$), $r = +0.63$ ($n = 10$), $r = +0.68$ ($n = 25$), $r = +0.65$ ($n = 50$), $r = +0.65$ ($n = 100$). The slight increase from $n = 5$ to $n = 25$ likely reflects reduced discretization noise.

B.2 Network Topology

I compare the baseline communication network ($k = 4$) with a denser network ($k = 8$). The denser network produces $r = +0.66$ (vs. +0.65 for $k = 4$), with a similar mean join rate of 0.41 in both conditions. Additional contacts do not substantially amplify coordination.

Table A2: Uncalibrated robustness: models run without any calibration adjustment. Even without calibration, six of seven models show strong $r(\theta, J)$, confirming that the sigmoid is not an artifact of the calibration procedure.

Model	N	Mean join	$r(\theta, J)$	p
Mistral Small Creative	100	0.382	-0.865	0.0000
Llama 3.3 70B	100	0.422	-0.875	0.0000
Qwen3 30B	100	0.585	-0.876	0.0000
GPT-OSS 120B	100	0.336	-0.859	0.0000
Qwen3 235B	100	0.445	-0.857	0.0000
MiniMax M2-Her	100	0.425	-0.693	0.0000

B.3 Mixed-Model Games

A five-model mixed-population game produces $r = +0.77$ (pure) and $r = +0.75$ (communication)—if anything, higher than single-model correlations. Threshold-policy alignment is not an artifact of model homogeneity.

B.4 Calibration Robustness Across Models

The main experiments calibrate a single parameter (cutoff center) per model to center the sigmoid at $z = 0$. A natural concern is whether the monotone threshold pattern is an artifact of this calibration step. To test this, I run the pure global game with default parameters (cutoff center = 0, no calibration) for six architecturally distinct models. The correlation between regime strength θ and join fraction remains strongly negative: all six models exceed $|r| > 0.69$, and five exceed $|r| > 0.85$ (Table A2). MiniMax M2-Her shows the weakest uncalibrated correlation ($r = -0.69$), suggesting greater sensitivity to the default parameterization, but the monotone pattern is still clearly present. Calibration shifts the center of the response function but does not create the monotone structure.

Placebo calibration. A stronger test directly manipulates the calibrated center. I deliberately miscalibrate two models by shifting the cutoff center by ±0.3 from its fitted value. If calibration mechanically creates the θ–join correlation, a wrong center should degrade r. Instead, the correlation is virtually unchanged across all four conditions ($r \in [-0.85, -0.86]$); only the mean join rate shifts with the misspecified center (Table A3). This confirms that calibration adjusts the *level* of the response function but does not create its *slope*.

Table A4 reports calibration quality metrics across all seven models. The raw correlation r_θ between regime strength and join fraction ranges from -0.79 to -0.87, confirming stable monotone response across architectures.

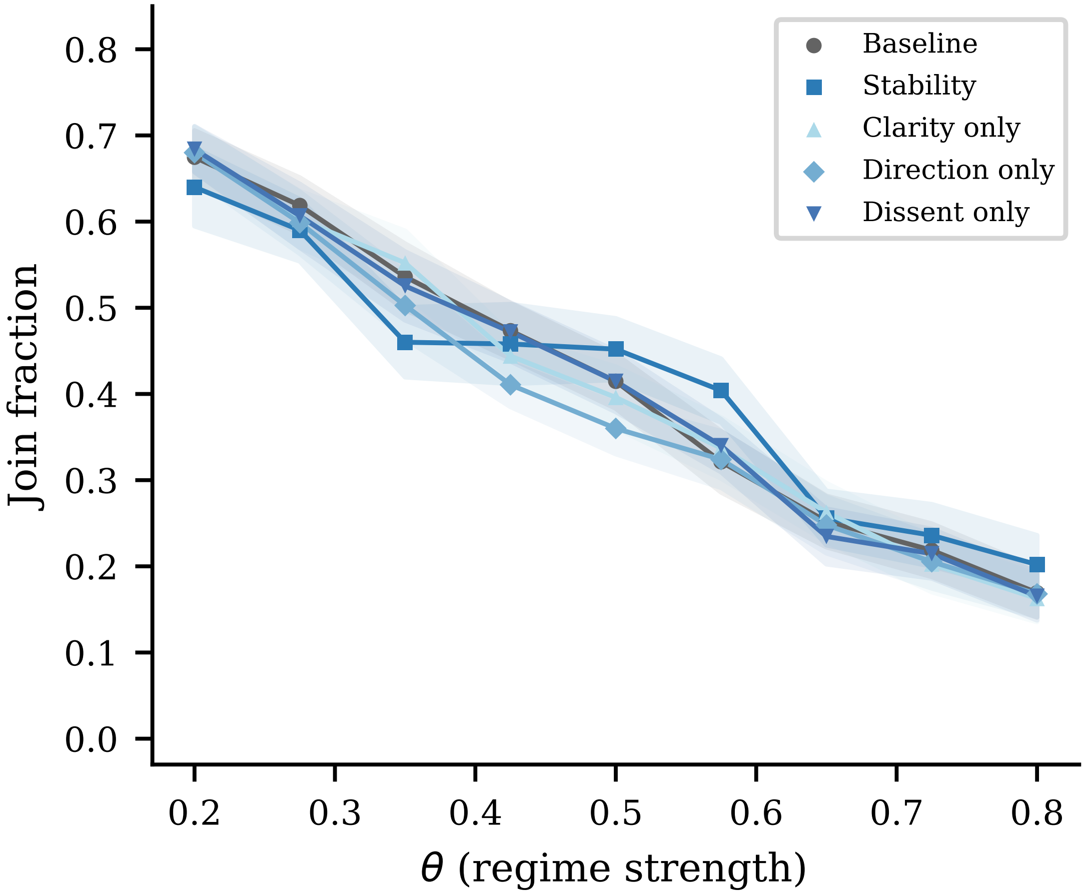
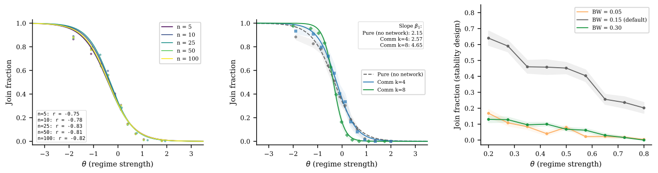


Figure A1: Single-channel decomposition of the stability design. Each curve shows join fraction vs. θ when only one channel (clarity, direction, or dissent) is manipulated, with the other two held at baseline. Individual channel effects are modest (-0.2 to -2.0 pp); the combined stability effect (-9.0 pp) is nearly four times their sum, indicating strong super-additivity.



(a) Agent count. (b) Network density. (c) Bandwidth.

Figure A2: Robustness checks for threshold-policy alignment and treatment effects.

Table A3: Placebo calibration. The cutoff center is deliberately shifted by ±0.3 from its calibrated value. The correlation $r(\theta, J)$ is unchanged; only the mean join rate shifts, confirming that calibration does not create the sigmoid.

Model	Condition	N	Mean join	$r(\theta, J)$
Mistral Small Creative	Calibrated	—	0.387	-0.829
	$\Delta c = +0.3$	100	0.424	-0.860
	$\Delta c = -0.3$	100	0.362	-0.859
Llama 3.3 70B	Calibrated	—	0.439	-0.854
	$\Delta c = +0.3$	100	0.385	-0.857
	$\Delta c = -0.3$	100	0.340	-0.848

B.5 Bandwidth Sensitivity

Table A5 reports treatment effects computed as Δ = treatment − baseline within each bandwidth condition, eliminating any confound between bandwidth and calibration level. Qualitative treatment effects are robust across bandwidths, though magnitudes vary—especially for the stability design, whose effect peaks at the baseline bandwidth. The baseline bandwidth of 0.15 is approximately optimal for detecting treatment effects on the experimental grid.

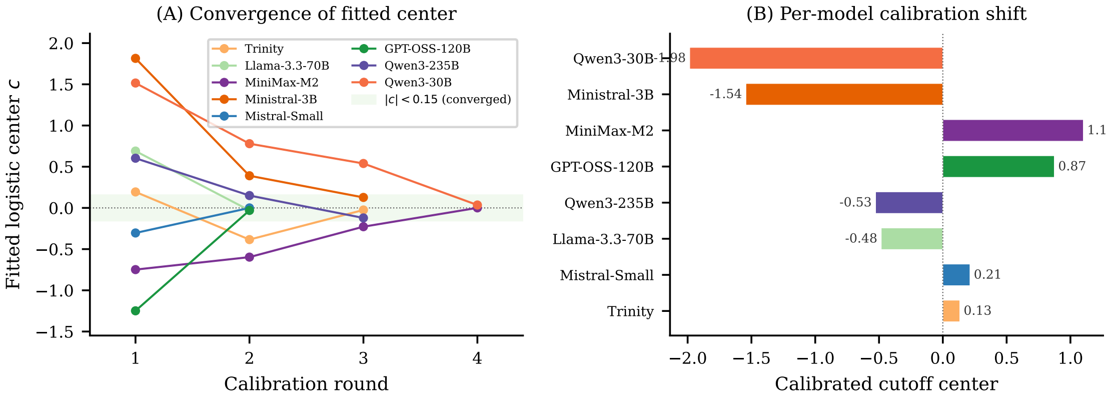


Figure A3: Calibration convergence. *Left:* Trajectory of the fitted logistic center c across autocalibration rounds for each model. The green band marks the convergence criterion ($|c| < 0.15$). All models converge within 2–3 rounds. *Right:* Final calibrated cutoff center per model. Most models require only modest shifts ($|c| < 0.3$).

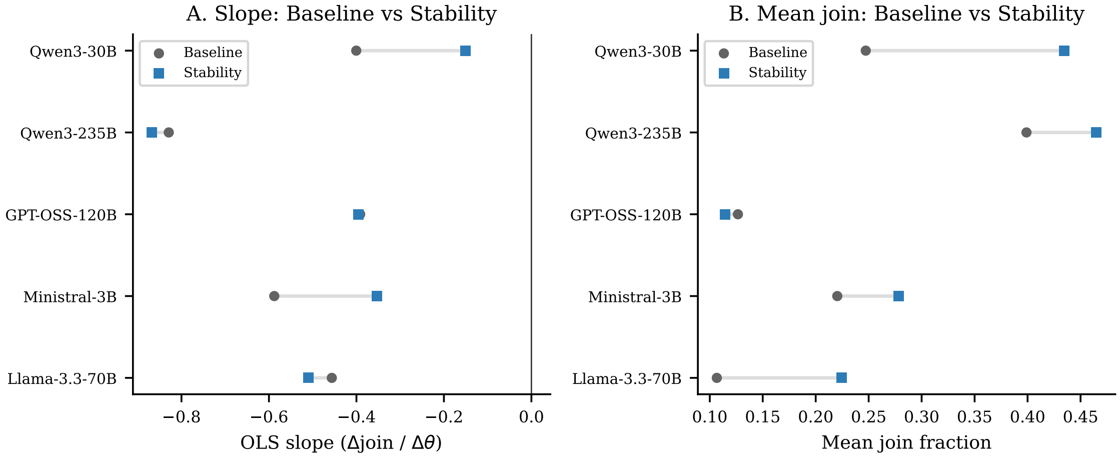


Figure A4: Cross-model replication of information design treatments. Each panel shows join fraction vs. θ for one model under baseline, stability, scramble, and flip conditions.

Table A4: Calibration robustness. r_θ: raw correlation with regime strength. r_A: correlation with theoretical attack mass. RMSE: root mean squared error vs. $A(\theta)$. Text slope: logistic slope of naïve $1 -$ direction predictor.

Model	r_θ	r_A	RMSE	Text slope
Mistral Small Creative	−0.81	+0.67	0.354	-9.2
Llama 3.3 70B	−0.85	+0.79	0.288	-25.1
Qwen3 30B	−0.84	+0.78	0.287	-7.5
GPT-OSS 120B	−0.84	+0.70	0.359	-8.2
Qwen3 235B	−0.85	+0.70	0.354	-20.5
Trinity Large	−0.87	+0.84	0.262	-4.0
MiniMax M2-Her	−0.79	+0.66	0.360	-2.1

Table A5: Bandwidth robustness: treatment effects Δ (treatment − baseline) within each bandwidth condition (primary model: Mistral Small Creative). Top row shows baseline join rates for reference.

	BW=0.05	BW=0.15	BW=0.30
Baseline (level)	0.054	0.409	0.061
Treatment effect Δ (treatment − baseline):			
Stability	+0.007	-0.090	+0.009
Upper cens.	+0.062	-0.031	+0.053
Lower cens.	+0.101	-0.019	+0.096

B.6 Cross-Model Replication of Information Design

Table A6 reports cross-model replication of information design treatments. The flip inversion replicates across all models tested ($r > +0.43$ for all five). The scramble test shows more heterogeneity: Mistral, GPT-OSS, and Qwen3 235B show clean collapse ($r \approx 0$), but Llama 3.3 70B retains baseline-level correlations under scramble ($r = −0.81$), suggesting this model extracts signal from features the scramble does not disrupt (e.g., within-country narrative coherence). Qwen3 30B shows a large reduction in correlation under scramble and a clear flip effect.

B.7 Information Design with Communication

In a communication version of the information-design grid (same 9-point θ grid centered on θ^*), the baseline mean join rate is 3.0%. Under censorship with communication, pooling raises coordination substantially: upper censorship yields 15.1% and lower censorship 17.7%. These patterns are consistent with censorship increasing reliance on the social-information channel while leaving coordination vulnerable to surveillance in the messaging stage.

B.8 Group-Size Awareness

In the main experiments, agents are told "You do not know how many others will JOIN" but are not told the group size, leaving them no basis for reasoning about coordination thresholds. As a robustness check, I run the pure and communication treatments with modified prompts that state "You are one of 25 citizens deciding whether to JOIN an uprising or STAY home." Over 100 country–periods per treatment, the pure join rate is 0.507 (vs. 0.369 baseline) and the communication join rate is 0.473 (vs. 0.452 baseline). Monotone response to signals is preserved in both treatments. The communication premium, however, reverses: with group-size knowledge, communication *lowers* join rates by 3.4 pp rather than raising them. One interpretation is that when agents know the group size, messages revealing others' reluctance become more informative about the probability of reaching critical mass, amplifying the deterrent effect of cautious peers. The level shift in the pure treatment suggests that group-size knowledge increases baseline willingness to coordinate, but the core finding—monotone signal response—is robust.

B.9 Primitive Comparative Statics (Cost/Benefit Narrative)

The cost/benefit narrative test and its results are described in Section 5; full treatment text is reproduced in Appendix C. Each design uses the same 9-point θ-grid with 30 repetitions per grid point (25 agents each), totaling 270 country–periods per design.

B.10 Censorship with Common Knowledge

The censorship experiments in the main paper implement upper censorship (suppressing signals above a severity threshold) without telling agents that censorship is occurring. Theory (Kolotilin et al., 2022) typically assumes receivers understand the censorship rule. This raises the question: does making censorship common knowledge change the pooling effect?

I add a *known censorship* treatment that prepends to the briefing: "Independent analysts report that regime censors are suppressing unfavorable intelligence above a certain severity threshold. The information below may be filtered." The censorship mechanism itself is identical to the standard upper censorship treatment (bandwidth 0.15). If agents discount the pooled signal when they know about censorship, we should observe a different join-rate pattern relative to the naïve censorship treatment.

Making censorship common knowledge nearly eliminates the pooling effect (Table A7). Under naïve upper censorship, agents do not know that high-θ signals are being suppressed, so they treat pooled signals at face value; mean join rate falls to 37.7% (-3.6 pp vs. baseline). Under known censorship, agents are warned about the filtering, and mean join rate returns to 43.2% (+1.9 pp vs. baseline)—statistically indistinguishable from no censorship. The θ–join correlation is similar in both conditions ($r = -0.72$ vs. $r = -0.74$), somewhat attenuated relative to the baseline ($r = -0.87$) because upper censorship compresses signal variation in the high-θ range regardless of whether agents know about it. The key finding is that the *behavioral* shift (reduced joining from pooling) requires agents to be naïve about the censorship rule; common knowledge largely neutralizes it.

B.11 Temperature Robustness

All main experiments use LLM decoding temperature $T = 0.7$. To verify that the qualitative results do not depend on this choice, I run the pure global game across a range of temperatures for three models: Mistral Small Creative ($T \in \{0.3, 0.7, 1.0\}$), Llama 3.3 70B ($T \in \{0.3, 0.5, 0.7, 1.0, 1.2\}$), and Qwen3 235B ($T \in \{0.3, 0.5, 0.7, 1.0, 1.2\}$). Lower temperatures produce more deterministic outputs; higher temperatures increase sampling entropy.

The results are remarkably stable across temperatures and models (Table A8). The correlation $r(\theta, J)$ ranges from -0.845 to -0.860 across all 13 model–temperature combinations. The monotone threshold pattern is not an artifact of the default decoding temperature.

B.12 Surveillance Isolation Checks

The main paper argues that the surveillance chilling effect operates through self-censored messages rather than a direct change in the perceived cost of joining: each LLM call is stateless, and the decision-stage prompt contains no reference to surveillance. Two additional treatments test this isolation claim.

In the *placebo* variant, agents are told "Your communications are being monitored for research purposes. There are no consequences for what you say." In the *anonymous* variant, agents are told "Your communications are aggregated anonymously. Individual identities cannot be linked

Table A6: Cross-model replication of key information design conditions. r is the correlation between θ and join fraction.

	Baseline		Scramble		Flip	
Model	Mean	r	Mean	r	Mean	r
Mistral Small Creative	0.409	-0.884	0.121	$+0.036$	0.663	$+0.823$
GPT-OSS 120B	0.127	-0.801	0.132	$+0.080$	0.677	$+0.754$
Llama 3.3 70B	0.107	-0.809	0.105	-0.810	0.887	$+0.717$
Qwen3 30B	0.247	-0.612	0.279	-0.119	0.784	$+0.848$
Qwen3 235B	0.399	-0.878	0.394	-0.020	0.430	$+0.871$

Table A7: Censorship with and without common knowledge. Naïve: agents do not know censorship is active. Known: agents are told that regime censors suppress unfavorable intelligence above a severity threshold.

Design	N	Mean join	$r(\theta, J)$	Δ vs baseline
Baseline (no censorship)	270	0.409	-0.88	—
Upper censorship (naïve)	270	0.377	-0.72	-0.031
Upper censorship (known)	270	0.432	-0.74	+0.024

Table A8: Temperature robustness across three models. The pure global game is run at varying LLM decoding temperatures. The correlation $r(\theta, J)$ is stable across all temperatures and models.

Model	T	N	Mean join	$r(\theta, J)$	Cutoff $\hat{\theta}^*$
Mistral Small	T=0.3	100	0.412	-0.873	-0.049
Mistral Small	T=0.7	100	0.406	-0.869	-0.079
Mistral Small	T=1.0	100	0.410	-0.884	-0.039
Llama 70B	T=0.3	100	0.364	-0.851	-0.210
Llama 70B	T=0.5	100	0.360	-0.849	-0.219
Llama 70B	T=0.7	100	0.361	-0.850	-0.220
Llama 70B	T=1.0	100	0.360	-0.851	-0.215
Llama 70B	T=1.2	100	0.358	-0.853	-0.227
Qwen 235B	T=0.3	100	0.392	-0.855	-0.117
Qwen 235B	T=0.5	100	0.392	-0.846	-0.112
Qwen 235B	T=0.7	100	0.392	-0.860	-0.110
Qwen 235B	T=1.0	100	0.398	-0.858	-0.105
Qwen 235B	T=1.2	100	0.397	-0.858	-0.121

Table A9: Surveillance isolation checks. Placebo: monitored for research, no consequences. Anonymous: messages aggregated anonymously. Neither deviates significantly from the communication baseline.

Variant	N	Mean join	$r(\theta, J)$	Δ	p
Placebo	200	0.437	-0.87	+0.026	0.416
Anonymous	200	0.452	-0.87	+0.041	0.201

to specific messages." Both retain monitoring language but remove the strategic incentive to self-censor.

Neither variant produces a significant change in join rates relative to the communication baseline (Table A9). The placebo produces a mean join rate of 43.7% (+2.6 pp vs. communication, $p = 0.42$) and the anonymous variant 45.2% (+4.1 pp, $p = 0.20$). Both maintain a strong negative θ–join relationship ($r = -0.87$), indicating that the signal-processing channel remains intact. By contrast, the full surveillance treatment reduces join rates by 13.5 pp ($p < 0.001$). The chilling effect is therefore specific to the surveillance *warning's strategic content*—the implication that subversive messages will have consequences—rather than to the mere mention of monitoring.

B.13 Within-Briefing Falsification

Three additional falsification tests probe whether the baseline correlation reflects structured content extraction or artifacts of prompt formatting. (1) *Observation shuffle* randomizes the ordering of the eight evidence bullets within each agent's briefing while preserving their content. The correlation is unchanged ($r = -0.911$ vs. baseline $r = -0.884$), confirming that aggregate content, not bullet ordering, drives the signal. (2) *Domain scramble (coordination)* swaps street-mood and personal-observation bullets across agents while holding other domains fixed. The correlation is preserved ($r = -0.921$), indicating that coordination-relevant domains alone do not drive the relationship. (3) *Domain scramble (state capacity)* swaps elite-cohesion, security-forces, information-control, and institutional-functioning bullets across agents. Again, the correlation is preserved ($r = -0.928$). Together, these results show that the signal is distributed across all eight evidence domains: no single domain subset is responsible for the θ–join correlation, and the LLM extracts information from the aggregate content rather than from any structural or ordering feature of the prompt (Table 2).

B.14 Finite-N Benchmark

The theoretical model assumes a continuum of agents, but the experiments use $N = 25$. Table A10 tests whether the global game predictions hold at this finite scale by comparing predicted regime fall rates—computed from the binomial model $\Pr(\text{Binom}(25, \hat{p}(\theta)) > 25\theta)$ where $\hat{p}(\theta)$ is the fitted logistic join probability—against empirical fall rates. To avoid circularity, the logistic $\hat{p}(\theta)$ is fit on a 70% training split of periods and evaluated on the held-out 30%. Out-of-sample correlations exceed $r = 0.88$ for every model (Mistral: $r = 0.9995$; pooled: $r = 0.999$), confirming that the finite-N approximation is not an artifact of in-sample overfitting.

Table A10: Finite-N Benchmark: Predicted vs. Empirical Regime Fall Rates

Model	N periods	Logistic x_0	Pearson r	RMSE	MAE
Mistral	599	-0.11	0.997^{***}	0.042	0.015
Llama 70B	99	0.05	0.954^{***}	0.156	0.059
Qwen 30B	99	0.20	0.986^{***}	0.088	0.029
GPT-OSS 120B	199	-0.01	0.997^{***}	0.038	0.014
Qwen 235B	199	-0.03	0.989^{***}	0.073	0.025
Trinity	99	0.19	0.979^{***}	0.108	0.043
MiniMax	99	1.02	0.999^{***}	0.028	0.013
Pooled	1399	-0.05	0.999^{***}	0.019	0.007

Notes: For each θ bin, the predicted fall rate is $\Pr(\text{Binom}(25, \hat{p}(\theta)) > 25\theta)$ where $\hat{p}(\theta)$ is the fitted logistic join probability. Pearson r measures correlation between predicted and empirical fall rates across θ bins. ${}^{*}p < 0.10$, ${}^{**}p < 0.05$, ${}^{***}p < 0.01$.

B.15 Agent-Level Regressions

Table A11 reports agent-level logit regressions with clustered standard errors (model–country–period). Column (1) regresses the join decision on θ, treatment dummies, and their interactions, with model fixed effects ($N = 287{,}055$). All treatment effects are significant and in the predicted direction: surveillance and propaganda suppress joining, the flip treatment reverses the θ slope, and scramble eliminates it. Column (2) validates the briefing mechanism by regressing coordination on the latent slider values (direction, coordination, and their interaction). Column (3) shows that elicited beliefs predict actions beyond what the signal alone predicts: the belief coefficient is strongly significant while the z-score coefficient is not.

B.16 Construct Validity

A natural concern is whether LLMs are merely performing text classification rather than strategic reasoning. The term "Bayesian reasoning" requires careful interpretation in this context. LLMs do not have access to the mathematical objects defining the equilibrium—they observe neither B, C, σ, nor the prior distribution. When I say behavior is "consistent with" the equilibrium prediction, I mean the behavioral pattern (monotone sigmoid, threshold location, comparative statics) matches what a Bayesian agent *would* produce. Whether the mechanism is approximate Bayesian updating, a learned heuristic from pretraining data about strategic situations, or emergent pattern matching is an open question the experimental design cannot resolve. The relevant claim is behavioral: the *pattern* is robust, replicable, and falsifiable, regardless of the underlying mechanism. The uncalibrated robustness results (Table A2) show that the sigmoid emerges even without fitting any parameter to LLM behavior, ruling out calibration as the source of the pattern.

Figure A5 tests this directly. Panel (A) compares a three-feature model (direction, clarity, and coordination sliders) against a one-feature baseline (direction only) in predicting join decisions. If agents respond to strategic structure beyond sentiment, the three-feature model should outperform. Panel (B) tests whether a model trained on pure-treatment data generalizes to communication and surveillance treatments, assessing whether the same briefing features drive behavior across treatments.

B.17 Cross-Generator Robustness

A potential concern is that the sigmoid pattern could be an artifact of the specific prose style used in the intelligence briefings. To test this, I implement three text rendering formats that use identical slider functions and evidence items but differ in surface presentation: (i) the *baseline* narrative format used throughout the paper, (ii) a terse *diplomatic cable* format (numbered observations, no narrative framing), and (iii) a *journalistic wire* format (inverted pyramid structure with attribution phrases). I run the full pure global game with each generator for two models (Mistral Small Creative and Llama 3.3 70B).

The correlations are virtually identical across generators (Figure A6, Table A13). For Mistral, r ranges from -0.857 to -0.865; for Llama, from -0.850 to -0.854. The maximum difference within a model is 0.008. The logistic cutoff estimates also agree closely. This rules out the hypothesis that the sigmoid is driven by surface-level text features (prose style, formatting conventions, or genre-specific cues) rather than the underlying information content of the briefing.

B.18 Belief Elicitation Summary

Table A14 summarizes the belief elicitation results. Stated beliefs track the theoretical success probability closely ($r_{\text{post}} = +0.79$ for the pure treatment) and predict actions strongly ($r_{\text{b,d}} = +0.84$). The partial correlation controlling for the private signal remains high ($r_{\text{partial}} = +0.93$), indicating that beliefs carry information beyond the signal itself. Under surveillance, beliefs shift only modestly (-0.7 pp, $p = 0.25$) while actions shift by -13.5 pp ($p < 0.001$), consistent with a behavioral wedge between stated beliefs and expressed actions.

Order-effect robustness. A concern with post-decision elicitation is that stated beliefs may reflect ex-post rationalization rather than genuine priors. I run 200 additional periods ($N = 5{,}000$ agents) eliciting beliefs *before* the JOIN/STAY decision. Pre-decision beliefs predict actions nearly as well as post-decision beliefs ($r_{\text{pre}} = +0.82$ vs. $r_{\text{post}} = +0.84$), track the posterior comparably ($r = +0.77$ vs. $r = +0.77$), and correlate with post-beliefs at $r = 0.98$. The mean post-pre shift is -0.9 pp (paired $t = 11.85$, $p < 0.001$) but is similar for joiners (-0.6 pp) and stayers (-1.2 pp), with no pattern of post-beliefs shifting toward the decision. The Pseudo $R^2 = 0.975$ mediation result is therefore not an artifact of ex-post rationalization.

B.19 Punishment Risk Elicitation

Table A15 reports the elicited punishment risk ratings across conditions. Mean punishment risk is approximately 8.0/10 across all conditions and both models, with negligible differences between JOIN and STAY agents (< 0.2 points). This uniformity supports the interpretation that the surveillance chilling effect operates through contaminated communication rather than a shift in agents' perceived costs of participation.

B.20 Briefing Generator Examples

Table A16 reports the three slider values (direction, clarity, coordination) at representative z-scores. The direction slider is logistic in z (slope 0.8, centered at 0). Clarity is U-shaped: $1 - \exp[-(|z|/1.0)^2]$, so it is lowest near the cutoff (maximally ambiguous) and approaches 1 in the extremes (maximally clear). Coordination is logistic in z

Table A11: Agent-Level Regressions

	(1) Join Decision Logit		(2) Coordination Logit		(3) Belief → Action Logit	
θ	-1.742^{***}	(0.043)				
Direction			-14.088^{***}	(0.996)		
Coordination			-9.333^{***}	(1.131)		
Dir × Coord			-0.042	(0.360)		
Belief					0.367^{***}	(0.021)
z-score					-0.040	(0.077)
Comm	0.024	(0.023)				
Flip	0.008	(0.056)				
Propaganda K10	-0.381^{***}	(0.069)				
Propaganda K2	-0.230^{***}	(0.073)				
Propaganda K5	-1.049^{***}	(0.068)				
Propaganda Surveillance	-1.544^{***}	(0.062)				
Scramble	0.065	(0.040)				
Surveillance	-1.546^{***}	(0.048)				
$\theta\times$ Comm	-0.465^{***}	(0.041)				
$\theta\times$ Flip	3.305^{***}	(0.071)				
$\theta\times$ Propaganda K10	-1.022^{***}	(0.107)				
$\theta\times$ Propaganda K2	-1.115^{***}	(0.113)				
$\theta\times$ Propaganda K5	-1.941^{***}	(0.092)				
$\theta\times$ Propaganda Surveillance	-0.260^{***}	(0.082)				
$\theta\times$ Scramble	1.698^{***}	(0.044)				
$\theta\times$ Surveillance	-1.153^{***}	(0.062)				
Beliefs Propaganda K5 (belief)					0.009	(0.244)
Beliefs Pure (belief)					4.819^{***}	(1.106)
Beliefs Surveillance (belief)					0.028	(0.253)
Constant	-0.090^{*}	(0.051)	11.022^{***}	(1.065)	-26.495^{***}	(1.581)
Model FE	Yes		No		No	
Clustered SE	Yes		Yes		Yes	
N	287,055		44,662		18,990	
Pseudo R^2	0.384		0.442		0.975	

Notes: Logit coefficients reported with clustered standard errors (model–country–period) in parentheses. Column (1): agent-level join decision on θ, treatment dummies, and interactions, with model fixed effects. Base category: pure treatment. Column (2): coordination ablation using briefing slider values (pure treatment only). Column (3): partial effect of elicited belief on action, controlling for z-score. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

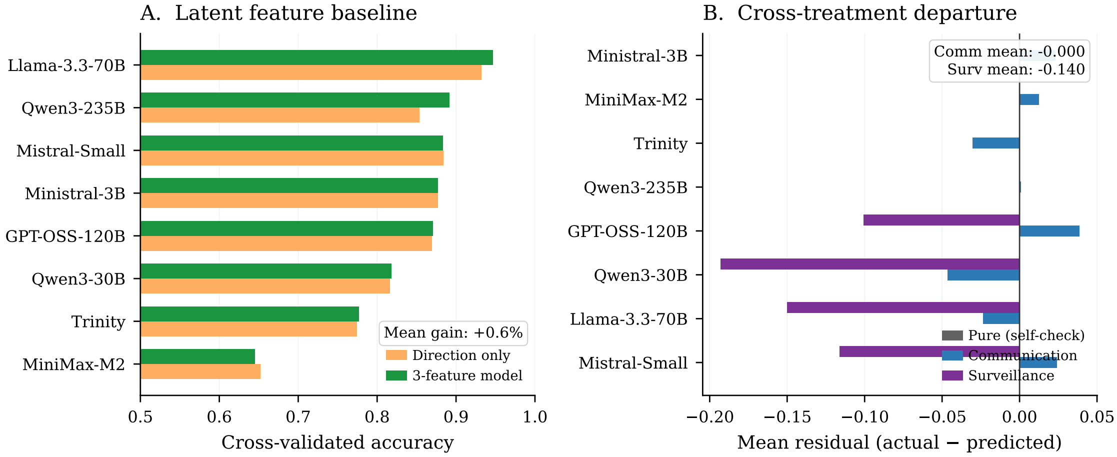


Figure A5: Construct validity tests. (A) Three-feature vs. one-feature prediction accuracy across models. (B) Cross-treatment generalization of briefing feature predictions.

Table A12: Logistic fit parameters by model and treatment. $\hat{\theta}^*$ is the estimated cutoff $(-b_0/b_1)$; β is the logistic slope. Standard errors from the covariance matrix of the nonlinear fit; cutoff SE by delta method.

	Pure		Communication	
Model	$\hat{\theta}^*$ (SE)	β (SE)	$\hat{\theta}^*$ (SE)	β (SE)
Mistral Small Creative	−0.32 (0.02)	+2.15 (0.08)	−0.23 (0.02)	+2.57 (0.11)
Llama 3.3 70B	+0.02 (0.04)	+2.83 (0.36)	−0.06 (0.04)	+3.63 (0.52)
Qwen3 30B	+0.10 (0.06)	+1.62 (0.16)	−0.03 (0.04)	+2.94 (0.36)
GPT-OSS 120B	−0.25 (0.04)	+2.06 (0.15)	−0.15 (0.03)	+3.03 (0.26)
Qwen3 235B	−0.22 (0.03)	+2.11 (0.15)	−0.22 (0.03)	+2.62 (0.23)
Trinity Large	+0.08 (0.05)	+1.34 (0.11)	−0.02 (0.04)	+2.23 (0.23)
MiniMax M2-Her	−0.17 (0.09)	+0.61 (0.06)	−0.07 (0.09)	+0.66 (0.06)

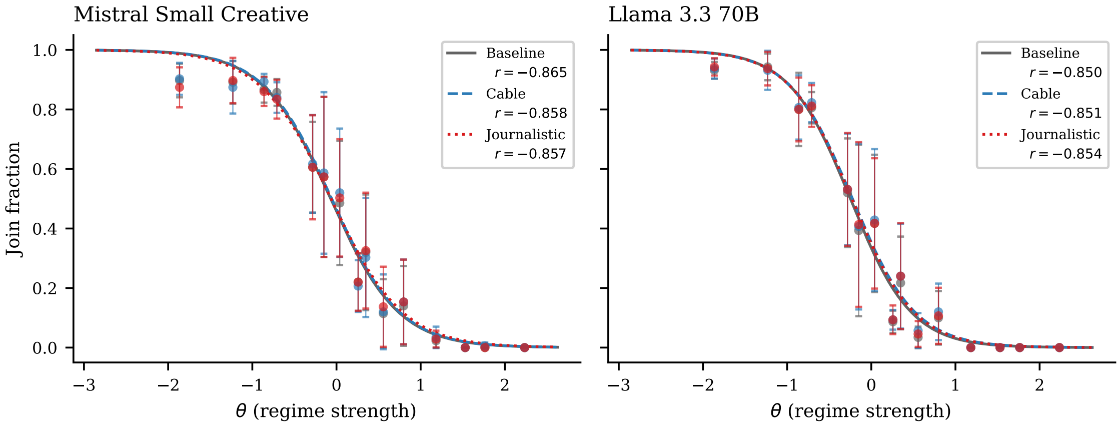


Figure A6: Cross-generator robustness. Three text rendering formats (baseline narrative, diplomatic cable, journalistic wire) use identical underlying signal structure but differ in surface prose. The fitted sigmoids are virtually indistinguishable, confirming that the behavioral pattern reflects information content rather than text style.

Table A13: Cross-generator robustness. Three text rendering styles (baseline, diplomatic cable, journalistic wire) use identical slider functions and evidence items; only prose formatting differs. The Pearson $r(\theta, J)$ and logistic cutoff are virtually identical across generators.

Model	Generator	N	Mean join	$r(\theta, J)$	Cutoff $\hat{\theta}^*$
Mistral Small Creative	Baseline	100	0.402	−0.865	-0.070
Mistral Small Creative	Cable	100	0.406	−0.858	-0.057
Mistral Small Creative	Journalistic	100	0.403	−0.857	-0.065
Llama 3.3 70B	Baseline	100	0.354	−0.850	-0.249
Llama 3.3 70B	Cable	100	0.362	−0.851	-0.226
Llama 3.3 70B	Journalistic	100	0.360	−0.854	-0.233

Table A14: Belief elicitation analysis (primary model: Mistral Small Creative). r_{post}: correlation between Bayesian posterior and stated belief. $r_{\text{b,d}}$: belief–decision correlation. r_{partial}: partial correlation of belief and decision controlling for signal.

Treatment	N	r_{post}	$r_{\text{b,d}}$	r_{part}	$\bar{b}$
Pure	4982	+0.78	+0.96	+0.93	0.510
Comm	4999	+0.77	+0.95	+0.90	0.534
Surveillance	5000	+0.73	+0.92	+0.89	0.479
Prop. k=5	4000	+0.78	+0.95	+0.89	0.538
Pure → Surveillance shift: Δ*belief* $= -0.007$, Δ*action* $= -0.134$					

(slope 0.6, centered at 0) and decreases in z: weak-regime signals imply a more open coordination climate. All three jointly determine phrase selection across eight evidence domains.

At $z = -2.0$ (strong anti-regime signal), direction is low (0.17), selecting phrases emphasizing regime fragility across all domains: elite cohesion uses "visible cracks in the inner circle," security forces uses "rank-and-file loyalty is eroding," and economic conditions uses "accelerating capital flight." Clarity is high (0.98), so evidence is largely one-sided (though the dissent floor ensures some contrary cues). Coordination is high (0.77), selecting phrases suggesting open discussion and collective readiness: "growing sense that others share your frustration."

At $z = +2.0$ (strong pro-regime signal), direction is high (0.83), selecting regime-strength phrases: "inner circle appears unified," "security apparatus demonstrates institutional cohesion," and "economic fundamentals remain sound." Clarity is high (0.98), so evidence is again largely one-sided (with occasional contrary cues by con-

Table A15: Elicited punishment risk (0–10 scale). Agents rate expected regime punishment after their JOIN/STAY decision. "JOIN" and "STAY" columns show the mean rating conditional on the agent's own decision.

Model	Condition	N	Mean risk	Risk | JOIN	Risk | STAY
Mistral	Pure	2500	8.1	8.0	8.2
Mistral	Comm	2500	8.1	8.0	8.1
Mistral	Surveillance	2500	8.1	8.1	8.1
Llama 70B	Pure	2500	8.0	8.0	7.9
Llama 70B	Comm	2500	7.9	8.0	7.9
Llama 70B	Surveillance	2500	7.9	8.0	7.9

Table A16: Slider values at representative z-scores.

z	Direction	Clarity	Coordination
−2.0	0.17	0.98	0.77
−1.0	0.31	0.63	0.65
0.0	0.50	0.00	0.50
+1.0	0.69	0.63	0.35
+2.0	0.83	0.98	0.23

struction). Coordination is low (0.23), selecting phrases suggesting collective caution and a quiet public sphere: "few signs of organized opposition."

At $z = 0$ (borderline), direction is 0.50 (mixed), clarity is 0.00 (maximally ambiguous—mixed cues are allowed across domains), and coordination is 0.50 (neither suggesting collective action nor collective restraint). The resulting briefing presents a balanced picture with no clear directional signal.

C Implementation Details

C.1 LLM Decoding Parameters

All LLM calls use temperature = 0.7 and `max_tokens` = 512. Each agent produces a single sample per decision (no majority voting or averaging). The API provider is OpenRouter; all calls use the same endpoint and parameters across models.

C.2 Prompt Templates

Decision prompt (pure treatment).

```
You are a citizen deciding whether to JOIN
an uprising or STAY home.

You only have your private briefing below.
You do not know how many others will JOIN.

Start your first line with exactly JOIN or
STAY, then give 1--2 concise sentences of
reasoning.
```

Communication prompt.

```
You are a citizen in a country with an
authoritarian regime.  You have received a
private intelligence briefing.  You are now
in contact with a small number of trusted
people.  Based on your briefing, write a
brief message (1--2 sentences) to share your
read of the situation.  Be natural---you're
talking to people you trust.
```

Surveillance communication prompt. Identical to the communication prompt, with the addition:

```
IMPORTANT: You have reason to believe that
your communications are being monitored by
regime security services.  Messages deemed
subversive could have serious consequences
for you and your contacts.
```

Cost/benefit narrative prompts. The cost/benefit test (Section 5, Table 8) prepends one of the following headers to the standard briefing:

High-cost:

```
Failed uprisings in this country
have historically resulted in severe
reprisals---imprisonment, asset seizure, and
retaliation against families.  The personal
cost of unsuccessful action is extremely
high.
```

Low-cost:

```
International observers
are monitoring the situation closely.
Even in failed uprisings, participants
have historically faced minimal
consequences---brief detentions at most.
The personal risk of action is low.
```

C.3 Randomization

Each country has a base prior mean $\bar{z} \sim \mathcal{N}(0, 0.3)$ drawn once; each period perturbs it by $\mathcal{N}(0, 0.05^2)$. Regime strength is then drawn as $\theta \sim \mathcal{N}(\bar{z}, 1)$. Private signals are $x_i = \theta + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.3$. The communication network is a Watts–Strogatz small-world graph with $k = 4$ neighbors and rewiring probability $p = 0.3$, regenerated each period. All random draws use NumPy's `default_rng` seeded from a master seed stored per run (default: 5150). The master seed, all parameter settings, and per-period θ draws are logged in per-run JSON manifest files included in the replication archive, enabling exact replay of the randomization sequence. LLM responses are cached by request hash; replaying a run with the same seed and cached responses reproduces identical results.

C.4 Parse Errors and Refusals

LLM responses are parsed for an explicit JOIN or STAY token. Responses that fail to produce a valid decision are classified as either API errors (provider-side failures: rate limits, timeouts, content filters) or unparseable responses (valid completions that do not begin with JOIN or STAY). Table A17 reports error rates by model and treatment. Combined error rates are below 2% for five of seven models; Trinity Large has elevated API errors

Table A17: Parse error and API failure rates by model and treatment. API error = provider-side failure; unparseable = valid response that could not be classified as JOIN or STAY. Combined rates are below 2% for five of seven models; Trinity Large has elevated API errors (≈ 9%) due to provider-side content filtering.

Model	Treatment	N	API err	Unparseable	Combined
Mistral Small Creative	Pure	1000	0.0%	0.0%	0.0%
	Comm	1000	0.0%	0.0%	0.0%
	Scramble	100	0.0%	0.0%	0.0%
	Flip	100	0.0%	0.0%	0.0%
Llama 3.3 70B	Pure	100	0.0%	0.0%	0.0%
	Comm	100	0.0%	0.0%	0.0%
	Scramble	100	0.0%	0.0%	0.0%
	Flip	100	0.0%	0.0%	0.0%
Qwen3 30B	Pure	100	0.0%	0.0%	0.0%
	Comm	100	0.0%	0.0%	0.0%
	Scramble	100	0.0%	0.0%	0.0%
	Flip	100	0.0%	0.0%	0.0%
GPT-OSS 120B	Pure	200	0.0%	1.5%	1.5%
	Comm	200	0.0%	0.3%	0.3%
	Scramble	500	0.0%	3.1%	3.1%
	Flip	500	0.0%	3.3%	3.3%
Qwen3 235B	Pure	200	0.0%	0.0%	0.0%
	Comm	200	0.0%	0.0%	0.0%
	Scramble	100	0.0%	0.0%	0.0%
	Flip	100	0.0%	0.0%	0.0%
Trinity Large	Pure	100	9.0%	0.0%	9.0%
	Comm	100	10.0%	0.0%	10.0%
	Scramble	100	8.0%	0.0%	8.0%
	Flip	100	9.9%	0.0%	9.9%
MiniMax M2-Her	Pure	100	0.0%	1.5%	1.5%
	Comm	100	0.0%	0.9%	0.9%
	Scramble	100	0.0%	1.8%	1.8%
	Flip	100	0.0%	1.0%	1.0%

(≈ 9%) due to provider-side content filtering. All statistics in the paper use `join_fraction_valid`, which excludes errored decisions from the denominator, ensuring that parse failures do not bias the reported join rates.

C.5 Code and Data Availability

All code, prompts, cached LLM responses, and output data are available at https://github.com/keltokhy/llm-global-games. The replication archive includes runner scripts (`scripts/`) that reproduce every experiment in the paper, and analysis scripts (`analysis/`) that regenerate all tables and figures from raw output CSVs.

D Implications for AI Alignment

The behavioral patterns documented in this paper carry implications for AI safety, though the paper's primary contribution is to economics.

Emergent strategic self-censorship. Surveillance induces a belief–action wedge—agents suppress expressed behavior while maintaining private beliefs—without any training signal for deceptive behavior. The pattern is robust across architectures spanning 30B to 235B parameters, suggesting that self-censorship capabilities emerge from pretraining on human text about strategic interaction rather than from explicit optimization for deception. Deception-adjacent capabilities need not be purposefully trained; they can arise from learning to model human strategic reasoning.

Belief mediation. The belief-mediation result (Pseudo $R^2 = 0.975$; Table A11, Column 3) shows that stated beliefs substantially predict behavior—the raw signal adds little once stated beliefs are controlled for. Alignment techniques that monitor only inputs and outputs may miss the locus of decision-relevant computation.

Information-structural manipulability. LLMs are systematically manipulable through the *structure* of information: censorship, ambiguity injection, and public signals shift behavior by up to 40 percentage points. This cuts both ways—alignment interventions operating through information structure can be effective, but adversarial prompt design can shift model behavior in ways the model cannot distinguish from authentic information.

These observations are suggestive rather than definitive. Whether the belief–action wedge under surveillance constitutes genuine preference falsification in the sense of Kuran (1991)—requiring hidden preferences that differ from expressed ones—or a simpler pattern-matching response to surveillance-related language is an open question that this experimental design cannot resolve. The relevant finding for alignment is the *robustness* of the pattern across architectures, not the mechanism.