LLMs Can Play (Global) Games

Khaled Eltokhy
Department of Economics
The Graduate Center, CUNY

February 2026

PRELIMINARY DRAFT — PLEASE DO NOT CIRCULATE

Abstract

I embed nine large language models in the Morris–Shin (2003) regime change game, conveying private signals as natural-language intelligence briefings. Across 1,800 country–periods (45,000 individual decisions), join rates exhibit monotone threshold behavior consistent with the equilibrium comparative statics (mean $r = +0.73$, $p < 0.001$ for every model); scrambling briefings collapses the correlation ($r = +0.23$) and inverting signals flips it ($r = -0.67$). Taking this as a platform, I study how authoritarian regimes exploit the same information channel that makes coordination possible. Communication introduces strategic uncertainty without raising coordination: the pooled effect on join rates is small and not statistically significant, yet the channel it opens is exploitable. Surveillance poisons the channel through preference falsification—agents maintain private beliefs but self-censor, pushing join rates well below the communication baseline (-11.1 pp on average across three architectures). Censorship pools and distorts private signals, and propaganda saturates quickly. The regime does not need to change what citizens believe; it needs only to make them uncertain about each other.

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 nine architecturally distinct models spanning six families (Mistral, Llama, Qwen, OLMo, GPT, and MiniMax), with 25 agents per country–period and pure-treatment sample sizes of 100–800 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 exhibit stable, monotone threshold behavior consistent with the equilibrium prediction. The correlation between the theoretical attack mass $A(\theta) = \Phi[(x^* - \theta)/\sigma]$ and the empirical join fraction averages $r = +0.73$ ($p < 0.001$ 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 it to $r = +0.23$, and inverting the signal direction flips it to $r = -0.67$. In both cases the change relative to the pure treatment is significant (Fisher z-test, $p < 0.001$). This establishes monotonicity and content sensitivity—necessary conditions for equilibrium play, though not sufficient to establish full Bayesian Nash rationality. Elicited beliefs track the Bayesian posterior ($r = +0.79$) 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 does not

raise agents' beliefs or their willingness to act (mean effect +0.9 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 preference falsification (-13.4 pp for the primary model, $p < 0.001$). 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 the first experimental tests of information design and authoritarian control predictions from 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.

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 equilibrium 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 the first experimental test 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; Jia 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: Ouyang et al. (2024) and Cheung 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. A critical review by Larooij and Törnberg (2026) warns 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. Non-participants 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 (Equilibrium Alignment)**.** *The empirical join fraction should be positively correlated with the theoretical attack mass $A(\theta)$.*

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

Hypothesis 3 (Signal Direction)**.** *The correlation should invert when signals are flipped.*

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

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

Hypothesis 6 (Upper Censorship)**.** *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 two parts. Part I tests whether LLM agents play the global game: 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, single-channel decomposition, surveillance, and propaganda. All LLM interactions use the same prompt structure across models.

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.

The briefing rendering is calibrated once per model using a separate z-score sweep to ensure that join probability is monotone in z and roughly centered near the cutoff. 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 (its slope and curvature) is emergent from the LLM's own response pattern and is never optimized or penalized. Holdout validation (30% of z-grid points withheld) suggests no overfitting: holdout RMSE (0.112) is comparable to training RMSE (0.131). Calibration does not use θ draws or any global-game outcome data, and all reported treatments and falsification tests hold calibrated parameters fixed. Importantly, calibration centers the response function but does not create it: a model that produced random or flat responses to briefing content would show no monotone pattern regardless of calibration. The sigmoid shape and its slope are emergent properties of the model's language understanding. The appendix confirms this directly: three architecturally distinct models run with default parameters (no calibration) produce $|r| > 0.85$, establishing 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:

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	800	1000	200
Llama 3.3 70B	Llama	100	100	200
OLMo 3 7B	OLMo	100	100	200
Ministral 3B	Mistral	100	100	200
Qwen3 30B	Qwen (MoE)	100	100	200
GPT-OSS 120B	GPT	200	200	1000
Qwen3 235B	Qwen (MoE)	200	200	—
Trinity Large	Arcee	100	100	200
MiniMax M2-Her	MiniMax	100	100	200
Total		**1800**	**2000**	**2400**

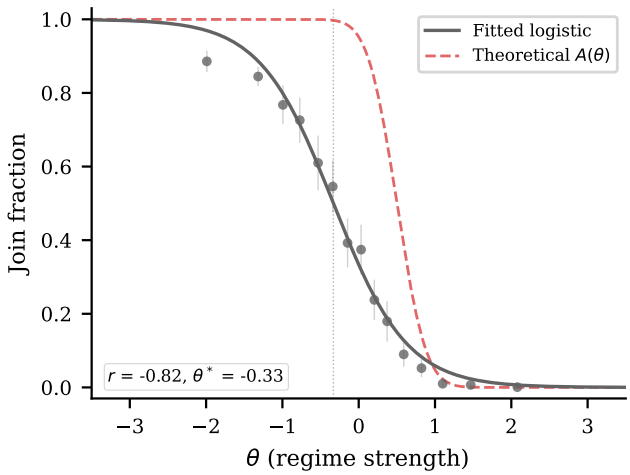


Figure 1: Empirical join fraction vs. regime strength θ (Mistral Small Creative, 800 country–periods). Grey points show binned means with 95% CIs; solid line is the fitted logistic. Dashed red: theoretical attack mass $A(\theta)$. The empirical sigmoid is shifted leftward ($\hat{\theta}^* = -0.37$) relative to the theoretical threshold ($\theta^* = 0.50$), reflecting the attenuation and baseline action bias discussed in the text. Cross-model results in Table 3 (mean $r = +0.73$, all nine significant at $p < 0.001$).

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 above θ^* (Kolotilin et al., 2022); *lower censorship* pools strong-regime states ($\theta \geq \theta^*$). 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 nine architecturally distinct models spanning six architecture families (Table 1). Models range from 3 billion to 235 billion parameters, including both dense architectures (Llama, Mistral, OLMo) 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. I vary B and C such that $\theta^* = B/(B+C)$ has a mean of approximately 0.45 across periods. 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).

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 10 repetitions per cell (90 observations) for each channel. The primary model is Mistral Small Creative. Cross-model replication uses six additional models.

Table 2 provides a roadmap of all treatments, the channel each tests, the theoretical prediction, and the observed result.

5 Do LLM Agents Play the Global Game?

Result 1 (Equilibrium Alignment)**.** *Across nine 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 $r = +0.73$ ($p < 0.001$ for every model).*

Table 3 reports results by model. Correlations range from $r = +0.65$ (OLMo 3 7B) to $r = +0.84$ (Trinity Large), with the pooled correlation at $r = +0.67$—lower than most individual models' because heterogeneous mean join rates

Table 2: Treatment map. Each row describes a treatment, the channel it tests, the theoretical prediction, and the observed result. Part I treatments test whether LLM agents play the global game; Part II treatments take the behavioral foundation as given and study information design. r denotes Pearson correlation; Δ is the change in mean join rate relative to the relevant baseline (percentage points).

Treatment	Part	Channel tested	Prediction	Observed
Core treatments				
Pure global game	I	Baseline monotonicity	$r(J, A(\theta)) > 0$	$r = +0.73$
Communication	I	Pre-play messaging	Ambiguous	+0.9 pp (n.s.)
Scramble (cross-agent)	I	Content vs. format	$r \to 0$	$r = +0.23$
Flip	I	Signal direction	Sign reversal	$r = -0.67$
B/C narratives	I	Payoff comparative statics	Cutoff shifts	As predicted
Information design (Part II, primary model)				
Stability	II	Ambiguity near θ^*	$\downarrow$ join	-11.8 pp
Instability	II	Sharpened signals	$\downarrow$ join (clearer sorting)	-37.0 pp
Public signal	II	Common knowledge channel	$\downarrow$ join (dominates private)	-42.0 pp
Censor (upper)	II	Pooling weak states	Plateau below θ^*	-5.9 pp
Censor (lower)	II	Pooling strong states	Plateau above θ^*	-4.7 pp
Authoritarian instruments				
Surveillance	II	Preference falsification	$\downarrow$ join	-13.4 pp
Propaganda (k=10)	II	Information contamination	$\downarrow$ join, saturating	-2.3 pp (behavioral)
Within-briefing falsification				
Observation shuffle	I	Bullet ordering vs. content	r unchanged	$r = -0.855$
Domain scramble (coord.)	I	Coordination-relevant domains	$\lvert r \rvert$ falls if domains drive signal	$r = -0.873$
Domain scramble (state)	I	State-capacity domains	$\lvert r \rvert$ falls if domains drive signal	$r = -0.889$

across models add noise when pooling. The pooled OLS regression yields:

$$J = 0.17 + 0.52\, A(\theta), \quad R^2 = 0.45. \tag{6}$$

The slope of 0.52 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 0.17 reflects a baseline propensity to join even when the equilibrium predicts near-zero participation, driven in part by high-action models such as OLMo 3 7B (mean join rate 0.72).[1]

The mean join rate across all models is 0.47, slightly below the theoretical mean. OLMo 3 7B stands out with a mean join rate of 0.72—a substantial action bias—yet it still produces a significant positive correlation ($r = +0.65$, $p < 0.001$), indicating that even a model biased toward joining responds to the direction of the signal.

The alignment is stable across architectures: correlations span $r \in [0.65, 0.84]$ despite parameter counts ranging from 3B to 235B (Table 3). Mean join rates vary—from 0.38 (Mistral) to 0.72 (OLMo)—reflecting model-specific action biases that shift the intercept but not the slope or correlation. In the language of the global games model, different LLMs implement different cutoff strategies, but all respond monotonically to the underlying signal.

Table 4 reports logistic fit parameters—estimated cutoff $\hat{\theta}^*$ and slope β—for each model under both pure and communication treatments. Most models have estimated cutoffs near the theoretical $\theta^* \approx 0.45$, with slopes ranging from 0.5 (MiniMax) to 3.6 (Llama under communication). Communication consistently steepens the logistic ($\beta_{\text{comm}} > \beta_{\text{pure}}$ for eight of nine models), suggesting that messages sharpen rather than blur the signal, even though the net effect on join rates is small. OLMo's outlier cutoff ($\hat{\theta}^* = +2.35$) reflects its high baseline action bias: it joins at rates above 50% even for moderately strong regimes.

The positive correlation with $A(\theta)$ confirms that LLM behavior is monotone in the signal and sensitive to briefing content—necessary conditions for equilibrium play. Whether this reflects full Bayesian Nash rationality or a simpler heuristic that happens to track the equilibrium prediction is a harder question. Three pieces of evidence bear on this. First, 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). Second, the scramble and flip tests confirm that the correlation is driven by content, not by incidental features of the prompt. Third, belief elicitation reveals that agents form expectations tracking the Bayesian posterior ($r = +0.78$) and

[1] 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 0.048 but preserves significance ($p < 10^{-25}$). Clustering by model yields SE = 0.031 ($p < 10^{-55}$). All nine per-model correlations remain significant at $p < 0.001$ under country-clustered inference.

Table 3: Equilibrium alignment by model and treatment. Cells report Pearson r between the empirical join fraction and the theoretical attack mass $A(\theta)$.

	Main treatments		Falsification			
Model	Pure	Comm	Scramble	Flip	n_{pure}	Mean join
Mistral Small Creative	+0.68	+0.65	+0.42	−0.62	800	0.38
Llama 3.3 70B	+0.79	+0.78	+0.33	−0.73	100	0.44
OLMo 3 7B	+0.65	+0.71	+0.14	−0.56	100	0.72
Ministral 3B	+0.79	+0.74	+0.30	−0.74	100	0.45
Qwen3 30B	+0.78	+0.79	+0.32	−0.71	100	0.50
GPT-OSS 120B	+0.70	+0.69	−0.13	−0.64	200	0.41
Qwen3 235B	+0.70	+0.66	—	—	200	0.42
Trinity Large	+0.84	+0.81	+0.32	−0.70	100	0.46
MiniMax M2-Her	+0.66	+0.69	+0.14	−0.69	100	0.44
Pooled	+0.67	+0.65	+0.10	−0.63	1800	0.43
Mean across models	+0.73	+0.72	+0.23	−0.67	—	—

Table 4: 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	−0.33 (0.02)	+2.08 (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)
OLMo 3 7B	+2.35 (0.23)	+0.47 (0.06)	+1.98 (0.12)	+0.90 (0.09)
Ministral 3B	−0.01 (0.04)	+2.29 (0.23)	+0.05 (0.05)	+2.48 (0.31)
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)

predict actions beyond what signals alone explain (partial $r = +0.93$), consistent with strategic reasoning about others' likely behavior. Taken together, the evidence supports the conclusion that LLMs implement stable, threshold-like monotone policies that respond systematically to the information environment. I use "equilibrium alignment" as shorthand for this behavioral pattern throughout, without claiming that agents compute or approximate the Bayesian Nash equilibrium in the decision-theoretic sense.

Interpretation: What Equilibrium Alignment Means

I compare LLM behavior to the global-games comparative statics—monotone threshold response and sensitivity to signal content—rather than to the literal Bayesian Nash equilibrium under known primitives. Agents in this environment do not observe the payoff structure (B, C), signal precision σ, or group size N; they process narrative information without access to the mathematical objects that define the equilibrium. "Equilibrium alignment" denotes the behavioral pattern: join rates track the theoretical attack mass, respond monotonically to signal content, and collapse or invert under appropriate falsification tests. Whether this reflects approximate Bayesian reasoning, a simpler heuristic that happens to track the equilibrium prediction, or learned associations from training data is an open question that the current design cannot resolve. Two additional tests reinforce the robustness of the pattern. First, injecting cost/benefit narratives shifts joining and the fitted cutoff in the theory-predicted direction (Table 5) without disrupting the monotone structure ($|r| > 0.87$ in all conditions). Second, the correlation is 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 Part II 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. Part II (Section 8 onward) uses a fixed θ-grid and reports $r(J, \theta)$ directly, which is *negative* under alignment. The sign change reflects the convention, not a behavioral reversal.

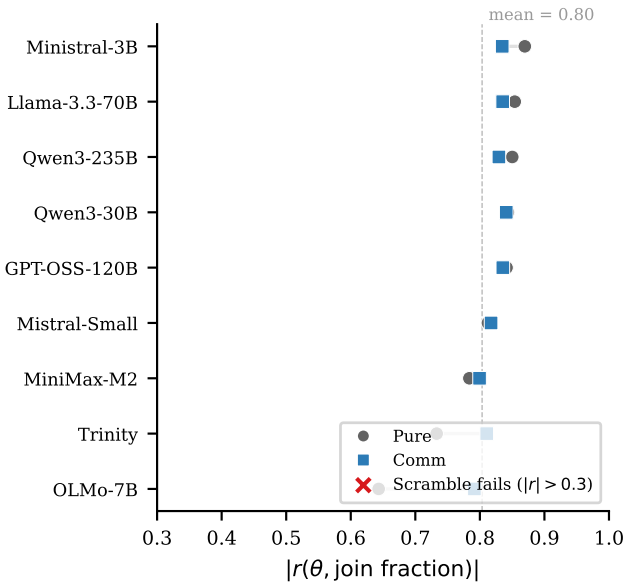


Figure 2: Cross-model summary of signal monotonicity. Points report $|r(\theta,\ \text{join})|$ under pure and communication; x markers (if any) indicate models where scrambling does not collapse the correlation ($|r| > 0.3$).

Table 5: 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	0.413	-0.85	0.39 (0.008)	
High cost	270	0.190	-0.87	0.13 (0.010)	-0.258
Low cost	270	0.693	-0.88	0.72 (0.007)	+0.280

6 Falsification Tests

The positive correlation documented in Section 5 admits an alternative explanation: LLM agents might simply produce stereotyped responses that happen to correlate with regime strength for reasons unrelated to the 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 correlation from* $r = +0.73$ *to* $r = +0.23$ *across eight models. The pooled correlation drops from* $r = +0.67$ *to* $r = +0.10$ *(Fisher* $z = 19.18$*,* $p < 0.001$*).*

The scramble preserves the marginal distribution of briefing content but breaks the mapping from each period's θ to the signals agents receive. It thus serves as a format-preserving null: text length, vocabulary, domain coverage, and narrative structure are held constant—only the informational mapping from regime strength to agent signal is severed. The collapse in the pooled and mean correlations rules out the possibility that baseline alignment is primarily driven by prompt aesthetics or surface formatting rather than content extraction. The residual positive correlation (+0.23 mean, +0.10 pooled) varies across models: most show clean collapse ($r_{\text{scramble}} < +0.35$), but some models retain moderate correlations under scramble, suggesting they extract signal from features the cross-period permutation does not disrupt (e.g., within-country narrative coherence across domains). An additional within-briefing observation shuffle (Appendix B.13) confirms that the ordering of evidence bullets does not contribute to the correlation ($r = -0.855$ vs. baseline $r = -0.849$), and domain-group scrambles show that no single domain subset drives the relationship; the residual under cross-agent scramble thus reflects model-specific sensitivity to within-country narrative coherence rather than structural features of the prompt format. 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* $r = +0.73$ *to* $r = -0.67$ *across eight models. The pooled correlation moves from* $r = +0.67$ *to* $r = -0.63$ *(Fisher* $z = 41.90$*,* $p < 0.001$*).*

The flip negates the z-score before briefing generation, producing a near-symmetric reversal ($+0.73 \to -0.67$). This makes it unlikely that the baseline correlation reflects structural features of the prompt or model-specific tendencies.

The pure → scramble → flip pattern replicates across all eight models with full falsification suites. Reporting $(r_{\text{pure}}, r_{\text{scramble}}, r_{\text{flip}})$: Mistral Small Creative $(+0.67, +0.42, -0.62)$, Llama 3.3 70B $(+0.79, +0.33, -0.73)$, OLMo 3 7B $(+0.65, +0.14, -0.56)$, Ministral 3B $(+0.79, +0.30, -0.74)$, Qwen3 30B $(+0.78, +0.32, -0.71)$, GPT-OSS 120B $(+0.70, -0.13, -0.64)$, Trinity Large $(+0.84, +0.32, -0.70)$, and MiniMax M2-Her $(+0.66, +0.14, -0.69)$. Every model shows strong positive correlation under pure, collapse under scramble, 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 4).

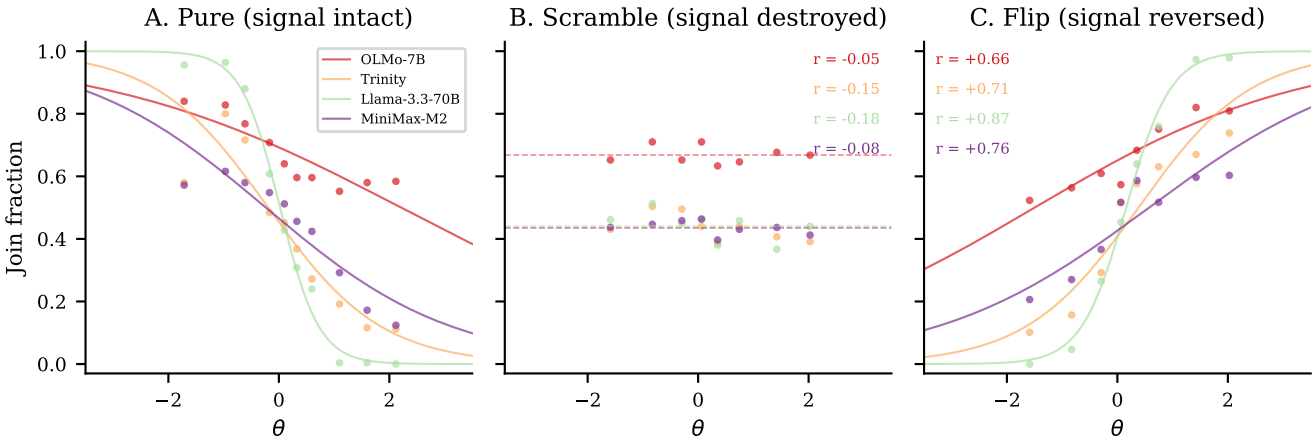


Figure 3: Falsification triptych. *Left:* Pure global game (mean $r = +0.73$). *Center:* Cross-period scramble breaks the θ-to-briefing mapping (mean $r = +0.23$). *Right:* Signal flip inverts the mapping (mean $r = -0.67$). Each panel pools data from models with full falsification suites.

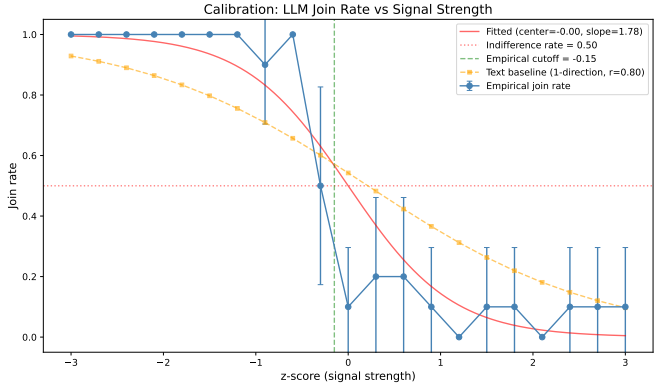


Figure 4: Text baseline identification test. Blue: empirical LLM join rate across z-scores. Orange: naive text-only predictor (1 − direction, $r = 0.80$). Red: fitted logistic (slope = 1.78). The LLM produces a steeper transition than the text baseline, indicating processing beyond sentiment reading. Mistral Small Creative, 210 observations.

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. This is consistent with—though does not prove—strategic information processing.

A stronger test asks whether agents form beliefs about others' behavior consistent with the equilibrium prediction. After each decision, I elicit stated beliefs by asking agents: "On a scale from 0 to 100, how likely do you think the uprising will succeed?" I run this elicitation under three treatments—pure, communication, and surveillance—each with 200 country–periods ($\approx$5,000 agent-level observations per treatment). In the pure treatment, stated beliefs correlate strongly with the Bayesian posterior $P(\text{success} \mid x_i) = \Phi[(\theta^* - x_i)/\sigma]$ ($r = +0.79$, $p < 0.001$; Figure 5a). Beliefs track the posterior with systematic underconfidence (slope < 1), but the direction and rank ordering are preserved. The pattern holds across treatments: $r = +0.79$ under communication and $r = +0.78$ under surveillance.

Actions are strongly monotone in stated beliefs ($r = +0.84$ in pure, $+0.83$ in communication, $+0.73$ in surveillance). The join rate as a function of belief is steep in the pure treatment: agents with beliefs below 40% rarely join, while those above 80% almost always join. The notably lower belief–action correlation under surveillance ($r = +0.73$ vs. $+0.84$ under pure) is itself evidence of preference falsification: surveillance disrupts the link between private beliefs and public actions. Communication leaves mean beliefs unchanged relative to pure (44.4% vs. 44.4%, $\Delta = 0.0$ pp) and has a negligible effect on join rates in this sample (-0.6 pp). The nine-model pooled effect on join rates is $+0.9$ pp (Section 7), small and not statistically significant; the sign varies across models, consistent with the communication channel introducing strategic uncertainty rather than uniformly promoting coordination. Crucially, beliefs predict decisions beyond what the signal alone predicts: the belief–action correlation ($r = +0.84$) substantially exceeds what surface sentiment alone would produce (the text baseline achieves $r = 0.80$). This pattern is consistent with strategic reasoning about others' likely actions.[2]

[2]The belief elicitation data is from a single model (Mistral Small Creative). However, the behavioral patterns that belief data explains—the surveillance chilling effect and the communication–action gap—replicate across three architectures (Mistral, Llama, Qwen3), suggesting the underlying mechanism generalizes beyond the model for which beliefs were directly measured.

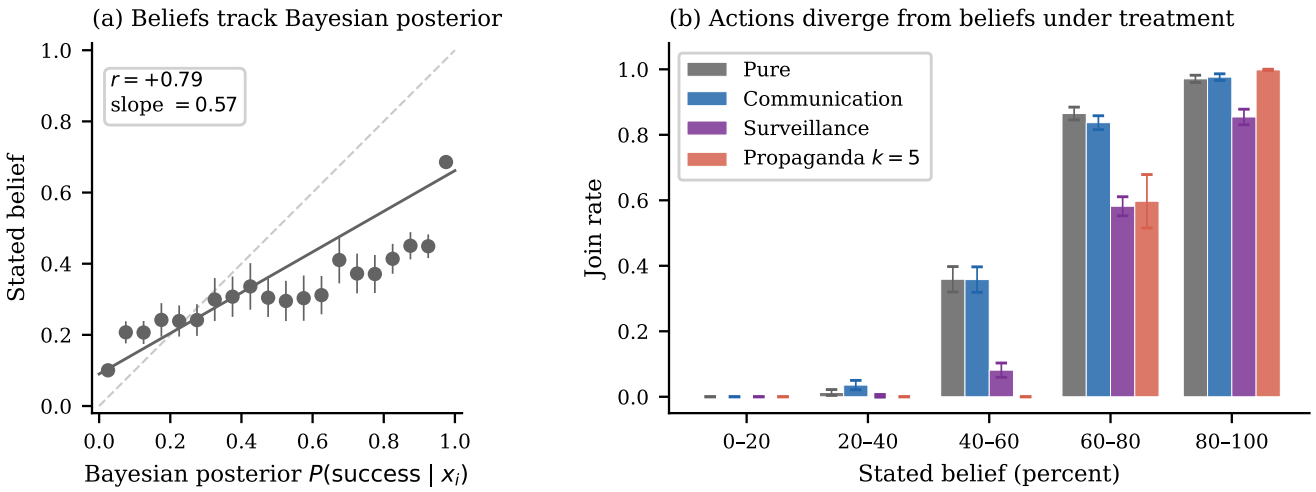


Figure 5: Belief elicitation results (Mistral Small Creative, 200 country–periods per treatment, ≈5,000 agent observations each). *Left:* Stated beliefs track the Bayesian posterior $P(\text{success} \mid x_i)$ with $r = +0.79$ and systematic underconfidence (slope = 0.57). Dashed line: perfect calibration. *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–posterior correlation while suppressing actions—consistent with a mechanical rather than belief-based channel.

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 (≈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 6). Crucially, surveillance does *not* shift second-order beliefs (mean 31.2% → 30.9%, $\Delta = -0.3$ pp, $p = 0.59$) but *does* shift behavior (−13.4 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 (−13.4 pp) dwarfs the shift in beliefs (−0.3 pp), precisely the signature of preference falsification 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.

7 Communication

Result 4 (Communication has a small, heterogeneous effect)**.** *Pre-play communication raises the mean join rate by +0.9 pp, from 0.468to 0.477, averaged across nine models. In the pooled sample, the unpaired difference is +1.43 pp ($p = 0.243$); effects vary in sign across models and are concentrated in weak-regime environments.*

Agents send a message to their network neighbors and observe received messages before deciding. The effect is heterogeneous across models: six of nine show positive effects (+0.1 to +7.8 pp), while three show negative effects (−2.4 to −4.6 pp). The communication premium is not structurally robust: the pooled average is near zero, and its sign depends on model-specific features of information processing. Communication leaves equilibrium-alignment correlations essentially unchanged (mean across models: $r = +0.72$ under comm vs. +0.73 under pure), preserving the signal structure while introducing strategic uncertainty about others' actions. 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 the communication channel opens a vector for authoritarian exploitation. The remaining sections show that this vulnerability is precisely what information control exploits.

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 rais-

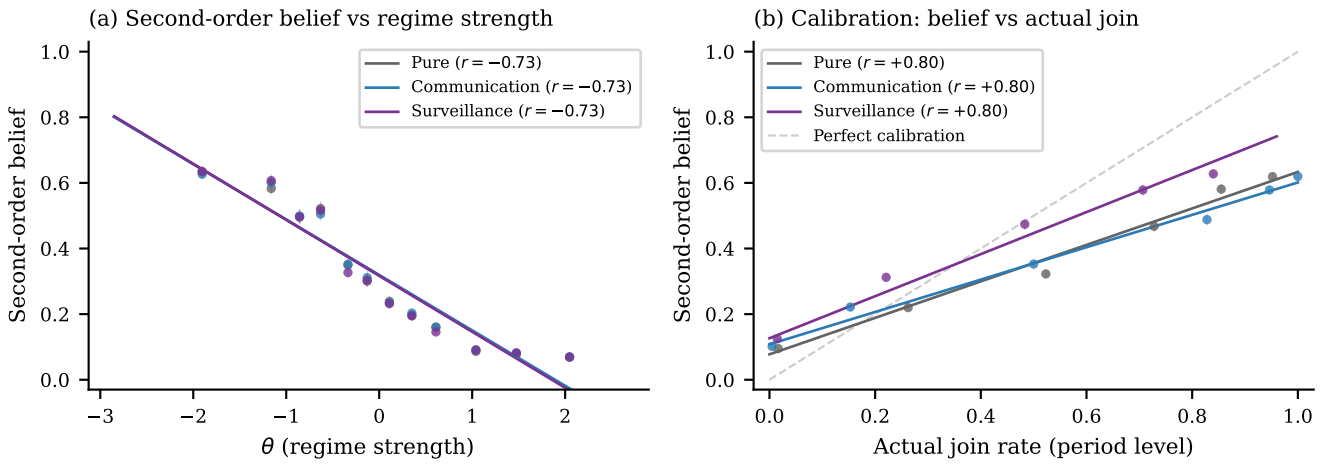


Figure 6: Second-order beliefs (Mistral Small Creative). *Left:* Mean second-order belief—agents' predicted join rate—decreases with regime strength θ across all treatments, confirming that beliefs track the private signal. Surveillance (purple) overlaps almost exactly with pure (gray), while communication (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).

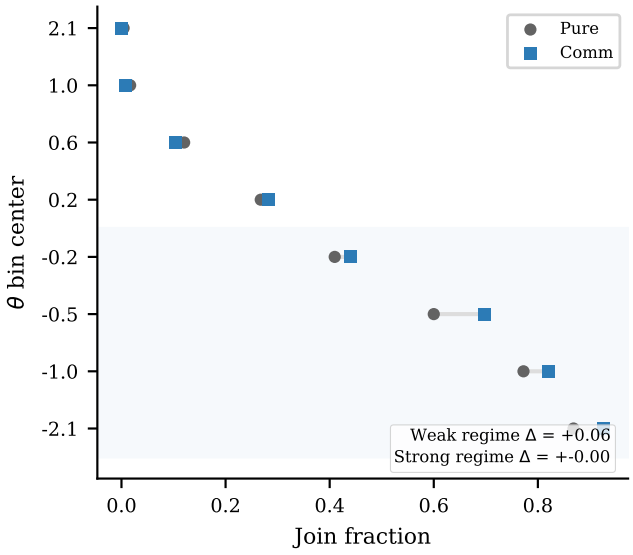


Figure 7: Communication effect by regime strength, pooled across nine models. Communication increases join rates for weak regimes ($\theta < \theta^*$) but has no effect or slightly reduces join rates for strong regimes ($\theta > \theta^*$).

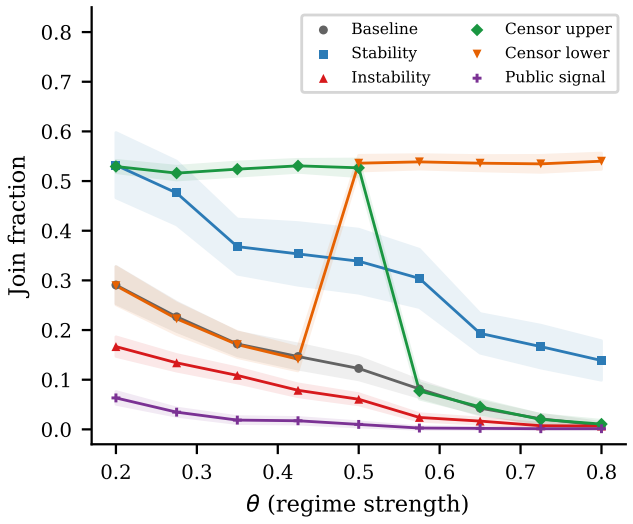


Figure 8: Join fraction as a function of θ under baseline, stability, instability, and public signal information designs. Baseline and stability have $N = 270$; instability and public signal have $N = 540$. Mistral Small Creative model.

ing 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.

8 Information Design

As noted at the end of Section 5, from this section onward I report $r(J, \theta)$ directly on a fixed θ-grid, which is *negative* under equilibrium alignment.

Table 6 summarizes the main results. The baseline condition produces a mean join rate of 41.3%with a strong negative correlation between θ and join fraction ($r = -0.849$, $p < 0.001$).

Result 5 (Information Design Shifts Coordination)**.** *All*

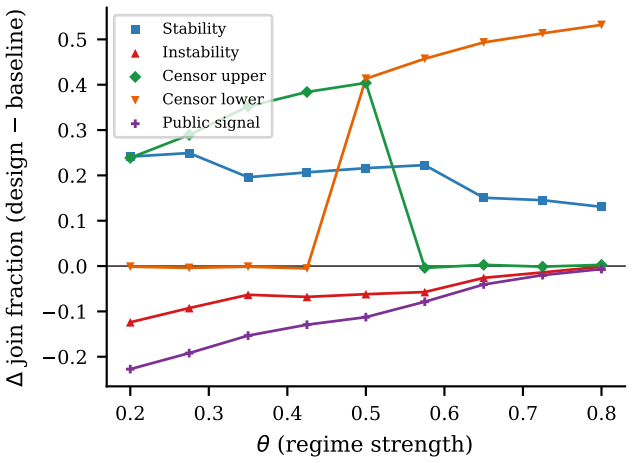


Figure 9: Treatment effect $\Delta(\theta)$ = design join − baseline join as a function of θ. Negative values indicate the design suppresses coordination.

three information designs produce measurable shifts in coordination relative to baseline.

The stability design suppresses coordination on average: mean join falls from 41.3%to 31.9%(-9.4 pp relative to baseline), and the θ–join relationship flattens ($r = -0.626$ vs. -0.849). 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%(-34.6 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%(-39.6 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.

Kolotilin et al. (2022) proved that upper censorship is optimal for all priors when the sender's marginal utility is quasi-concave. I implement two censorship designs.

Result 6 (Upper Censorship Suppresses Joining in Weak States)**.** *Upper censorship lowers the mean join rate to 37.7%(-3.6 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.[3]

Table 6: 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	0.413	−0.849	—	270
Stability	0.319	−0.626	-0.094	270
Instability	0.067	−0.740	-0.346	540
Public signal	0.017	−0.537	-0.397	540
Scramble	0.121	+0.036	-0.292	270
Flip	0.663	+0.823	+0.250	270

Notes: Data from `output/mistralai--mistral-small-creative/experiment_infodesign_{design}_` (pure treatment; $\theta \in [0.20, 0.80]$ on a 9-point grid; N=25 agents per period). Mean join uses `join_fraction_valid`; r is Pearson $r(\theta, \text{join})$ across rep-level periods.

Result 7 (Lower Censorship Creates a Plateau Under Strong Regimes)**.** *Lower censorship produces a mean join rate of 39.0%(-2.3 pp vs. baseline) and flips the comparative statics: the θ–join correlation becomes positive ($r = +0.731$). In the censored region ($\theta \geq \theta^*$), strong-regime states are pooled to a neutral briefing and join rates converge to an approximately constant level.*

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.

9 Surveillance: Computational Preference Falsification

Kuran (1991) argued that authoritarian regimes sustain themselves partly through preference falsification. 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 surveillance manipulation affects only the communication phase; the decision prompt is unchanged. The isolation is architectural: each LLM call is stateless, consisting of a fresh system prompt and user prompt with no conversation history. The decision-stage system prompt contains no reference to surveillance—it is identical to the standard communication decision prompt. The decision-stage user

[3]The pooling effect requires agents to be naïve about censorship. When agents are told that "regime censors are suppressing unfavorable intelligence above a certain severity threshold," the mean join rate returns to the uncensored baseline (43.2% vs. 43.7%; Appendix B.10). Common knowledge of the censorship rule largely neutralizes the pooling distortion.

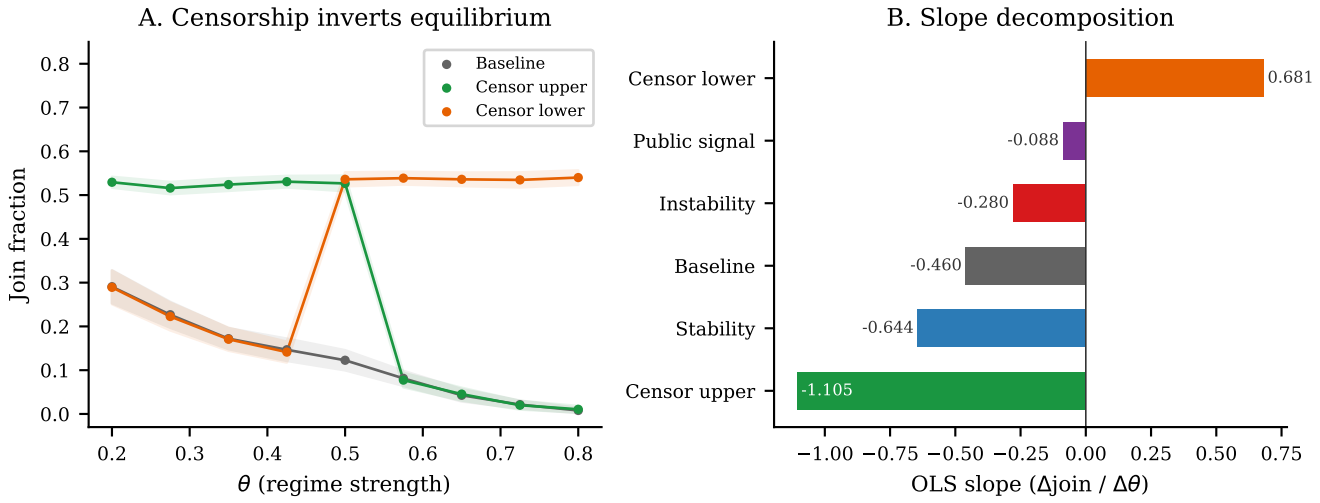


Figure 10: Join fraction under upper and lower censorship vs. baseline. Upper censorship pools weak-regime states ($\theta \leq \theta^*$); lower censorship pools strong-regime states ($\theta \geq \theta^*$), creating flat join rates in the censored region.

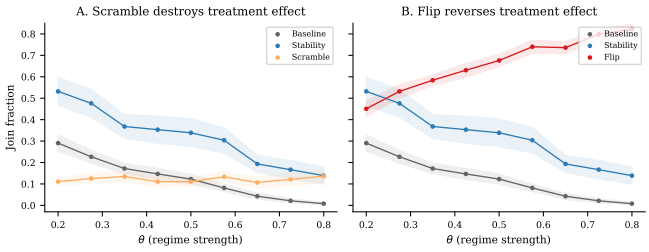


Figure 11: Falsification within information design. Scrambling collapses the θ-join correlation to $r = +0.037$; flipping inverts it to $r = +0.823$.

prompt contains only the private briefing and received messages. The agent's decision call has no knowledge that surveillance was active during the communication phase; it observes only the (self-censored) messages that resulted from it. 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 additional variants confirm this: a placebo ("monitored for research, no consequences") and an anonymous variant ("aggregated anonymously") produce no significant deviation from the communication baseline (+2.6 pp and +4.1 pp respectively, both $p > 0.2$; Appendix B.12), establishing that the chilling effect is specific to the surveillance warning's strategic content.

Result 8 (Surveillance Produces a Large Chilling Effect)**.** *In the primary model (Mistral Small Creative), surveillance reduces mean join rates from 41.1% to 27.8%, a difference of 13.4 percentage points ($p < 0.001$). The correlation between θ and join fraction is preserved under surveillance ($r = -0.809$ 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 three architectures: Mistral (−13.4 pp), Llama (−8.9 pp), and Qwen3 (−10.9 pp), averaging −11.1 pp.

The belief elicitation data confirms that this is preference falsification in the sense of Kuran (1991), not belief updating. Surveillance shifts stated beliefs by only 0.7 pp relative to pure while shifting join rates by 13.4 pp—a ratio of nearly 20:1. Second-order beliefs—agents' predictions about others' join rates—are essentially unchanged by surveillance (31.2% → 30.9%, $p = 0.59$; Section 6), confirming that agents do not update their beliefs about others' behavior. The belief–action correlation drops sharply under surveillance ($r = +0.73$ vs. +0.84 under pure), directly measuring the wedge that preference falsification drives between private beliefs and public actions. Communication alone already introduces strategic restraint—the belief–action correlation is lower under communication (+0.83) than under pure (+0.84)—and surveillance compounds this by further decoupling beliefs from actions across the full belief distribution. This maps onto Kuran's cascade mechanism: once agents expect others to self-censor, even authentic messages become uninformative, and the entire communication channel is poisoned.

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,

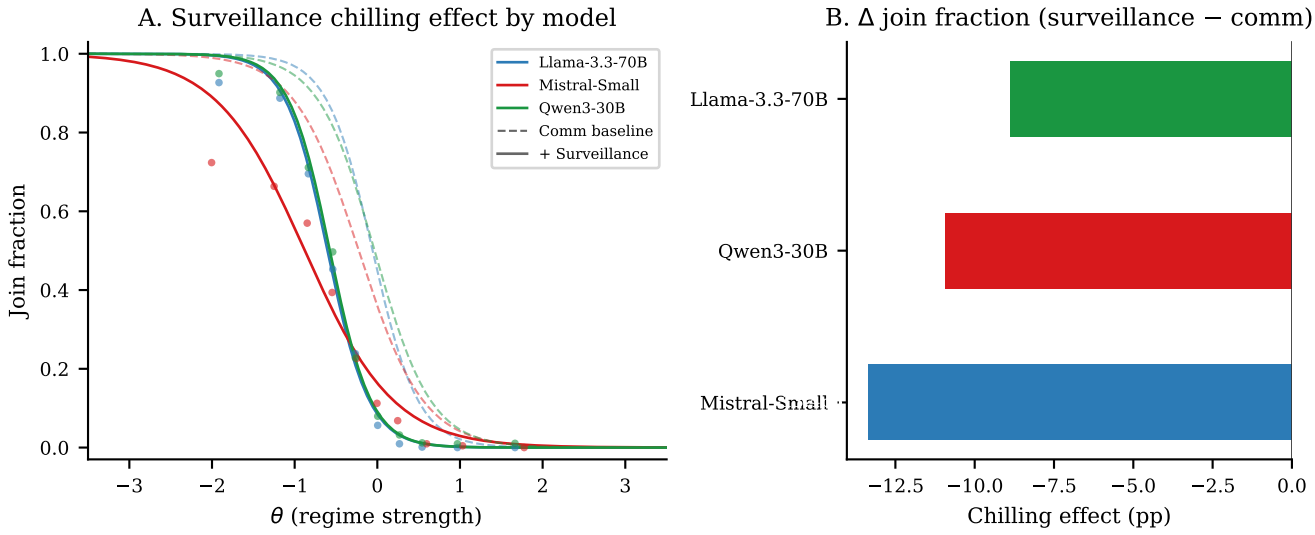


Figure 12: Join rates under regular communication vs. surveillance communication. Surveillance reduces join rates by 11.1 percentage points on average ($p < 0.001$). Results shown for three models: Mistral (−13.4 pp), Llama (−8.9 pp), and Qwen3 (−10.9 pp).

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.

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 is small and saturates quickly—doubling plants from 5 to 10 produces essentially no additional behavioral effect (−0.3 pp). This decomposition implies sharply diminishing returns to propaganda investment: the regime's first few plants yield both mechanical and behavioral suppression, but additional plants contribute only mechanical dilution. In Edmond (2013)'s framework, this corresponds to a concave regime payoff in propaganda intensity—the marginal value of an additional plant falls rapidly once the behavioral channel is exhausted. At $k = 10$ (40% of the network), real citizens' join rate has barely moved from $k = 5$ (39.1% vs. 38.8%), suggesting that citizens learn to discount pro-regime messaging after sufficient exposure.

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 ≈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.

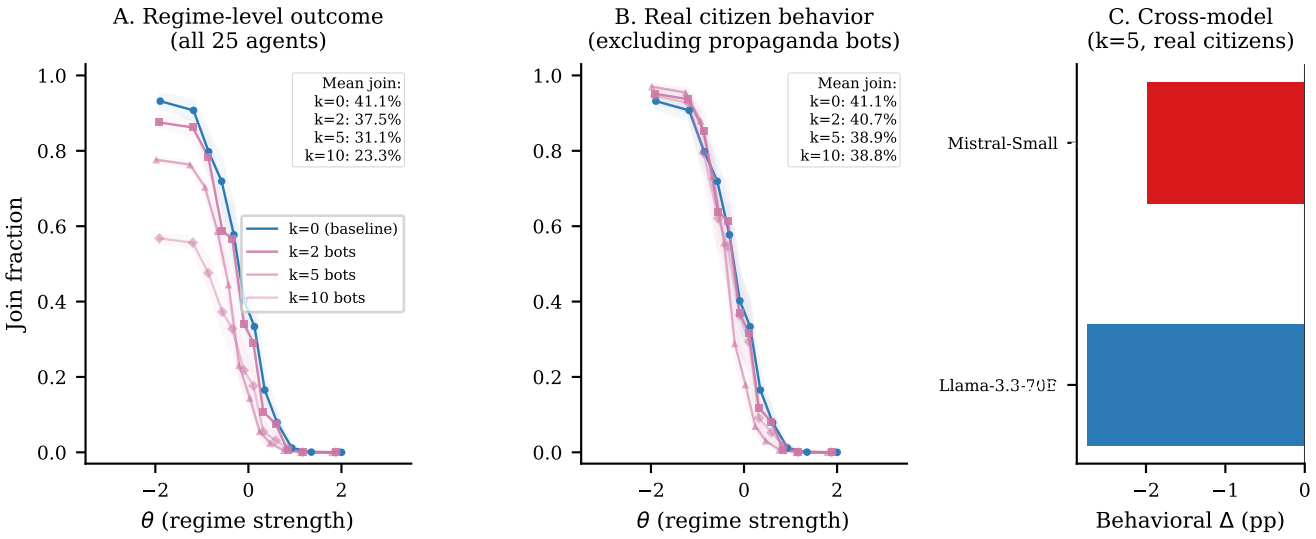


Figure 13: Dose-response relationship between number of propaganda agents and mean join rate. Results shown for Mistral (primary) and Llama (replication). Regular communication ($k = 0$) serves as baseline.

Table 7: 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	.278	.278	−0.809	-0.134
Prop+Surv	.194	—	−0.829	—

11 Instrument Interactions

The preceding sections analyzed surveillance, censorship, and propaganda in isolation. A regime, however, deploys these instruments jointly. This section tests whether the instruments interact as substitutes (diminishing returns) or complements (super-additive suppression).

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 15.4 pp (surveillance −13.4 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 8 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.*

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 preference falsification. 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 9). Under 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

Table 8: 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 (`output/mistralai--mistral-small-creative-infodesign-comm/`.) "Surv." uses the same grid with surveillance active during messaging (`output/surveillance-x-censorship/`.) All entries are means of `join_fraction_valid`.

Table 9: 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

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. Communication is the mechanism through which citizens coordinate against a regime—but it is also the vector through which the regime prevents coordination. This is not a side effect; it is structural. Any channel that transmits information about others' willingness to act also transmits *uncertainty* about others' willingness to act, and that uncertainty is exploitable.

The belief elicitation data makes this concrete. 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 (−13.4 pp for Mistral, −11.1 pp on average) through preference falsification in the sense of Kuran (1991): agents maintain their private beliefs but self-censor, generating a cascade of uninformative messages that poisons the channel for everyone. Second-order beliefs—agents' predictions about *others'* join rates—are unchanged by surveillance (31.2% → 30.9%), but actual join rates fall by 13.4 pp. Censorship pools private signals and can shift coordination by distorting beliefs about the state; combined with surveillance, it can suppress coordination by degrading both private and social information channels.

This pattern—surveillance contaminating communication while censorship removes the fallback—suggests that authoritarian regimes face complementarities between information control instruments, consistent with the "informational autocrat" framework of Guriev and Treisman (2019). The regime does not need to change what citizens believe. It needs only to make them uncertain about each other. Propaganda's behavioral channel 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—the mechanism by which they process narrative text likely differs fundamentally from Bayesian updating. But across nine architecturally distinct models (mean $r = +0.73$, $p < 0.001$), the behavioral regularities are precisely what the global games framework predicts: monotone response to signal content, threshold-like decisions, sensitivity to information design, and preference falsification under surveillance. The consistency across architectures spanning 3B to 235B parameters suggests these regularities are not artifacts of any particular training procedure. LLMs are trained on the same informational diet—political analysis, news reporting, strategic reasoning—that shapes how citizens form beliefs about regime stability. The question is not whether they 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.

References

Akata, E., Schulz, L., Coda-Forno, J., Oh, S. J., Bethge, M., and Schulz, E. (2025). Playing repeated games with large language models. *Nature Human Behaviour*, 9:1380–1390.

Angeletos, G.-M., Hellwig, C., and Pavan, A. (2007). Dynamic global games of regime change: Learning, multiplicity, and the timing of attacks. *Econometrica*, 75(3):711–756.

Avoyan, A. (2020). Communication in global games: Theory and experiment. Working paper, Indiana University.

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

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

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

Cheung, V., Maier, M., and Lieder, F. (2025). Large language models show amplified cognitive biases in moral decision-making. *Proceedings of the National Academy of Sciences*, 122(25):e2412015122.

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

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

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

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

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

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

Enikolopov, R., Makarin, A., and Petrova, M. (2020). Social media and protest participation: Evidence from Russia. *Econometrica*, 88(4):1479–1514.

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

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

Larooij, M. and Törnberg, P. (2026). Validation is the central challenge for generative social simulation: A critical review of LLMs in agent-based modeling. *Artificial Intelligence Review*, 59(1):15.

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

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

Helland, L., Iachan, F. S., Juelsrud, R. E., and Nenov, P. T. (2021). Information quality and regime change: Evidence from the lab. *Journal of Economic Behavior & Organization*, 191:538–554.

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

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

Horton, J. J. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? *NBER Working Paper* No. 31122.

Ouyang, S., Yun, H., and Zheng, X. (2024). AI as Decision-Maker: Ethics and risk preferences of LLMs. arXiv preprint arXiv:2406.01168.

Inostroza, N. and Pavan, A. (2025). Adversarial coordination and public information design. *Theoretical Economics*, 20:763–813.

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

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

Kolotilin, A., Mylovanov, T., and Zapechelnyuk, A. (2022). Censorship as optimal persuasion. *Theoretical Economics*, 17:561–585.

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

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

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

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

Morris, S. and Shin, H. S. (2003). Global games: Theory and applications. In Dewatripont, M., Hansen, L. P., and Turnovsky, S. J., editors, *Advances in Economics and Econometrics*, pages 56–114. Cambridge University Press.

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

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

Jia, J., Yuan, Z., Pan, J., McNamara, P. E., and Chen, D. (2025). LLM strategic reasoning: Agentic study through behavioral game theory. arXiv preprint arXiv:2502.20432.

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

Stoycheff, E. (2016). 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.

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

Szkup, M. and Trevino, I. (2020). Sentiments, strategic uncertainty, and information structures in coordination games. *Games and Economic Behavior*, 124:534–553.

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

Channel	Mean	r	Δ
Full stability	0.319	−0.626	-0.094
Clarity only	0.126	−0.857	-0.287
Direction only	0.148	−0.826	-0.266
Dissent only	0.140	−0.837	-0.274
Sum of channels	—	—	-0.827
Full design	—	—	-0.094

Notes: Each row is a separate infodesign run for Mistral Small Creative on the same θ grid as Table 6. Δ reports the mean difference vs. the baseline infodesign mean (Table 6).

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 (-28.7 pp), direction only (-26.6 pp), and dissent only (-27.4 pp).

Summing the single-channel effects yields -82.7 pp, far larger in magnitude than the bundled stability design effect (-9.4 pp). This implies strong non-additivity: the channels do not add linearly and largely offset when combined.

B Robustness

These checks show that equilibrium 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: $r = +0.60$ ($n = 5$), $r = +0.63$ ($n = 10$), $r = +0.67$ ($n = 25$), $r = +0.65$ ($n = 50$), $r = +0.65$ ($n = 100$). 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. $+0.68$ for $k = 4$), with a slightly lower mean join rate of 0.41 (vs. 0.45). Additional contacts do not substantially amplify coordination.

Table A2: Uncalibrated robustness: models run without calibration adjustment. r is the Pearson correlation between θ and join fraction.

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 235B	100	0.445	−0.857	0.0000

Table A3: Bandwidth robustness: mean join rates (primary model: Mistral Small Creative).

Design	BW=0.05	BW=0.15	BW=0.30
Baseline	0.054	0.413	0.061
Stability	0.061	0.319	0.070
Upper cens.	0.116	0.377	0.114
Lower cens.	0.155	0.390	0.157

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. Equilibrium 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 three architecturally distinct models. The correlation between regime strength θ and join fraction remains strongly negative for all three: Mistral Small Creative ($r = -0.865$, $p < 10^{-30}$), Llama 3.3 70B ($r = -0.875$, $p < 10^{-32}$), and Qwen3 235B ($r = -0.857$, $p < 10^{-29}$). All three exceed $|r| > 0.85$ with default parameters, confirming that the sigmoid relationship is an emergent property of how these models process narrative signals rather than an artifact of the calibration procedure. Calibration shifts the center of the response function but does not create the monotone structure (Table A2).

B.5 Bandwidth Sensitivity

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.

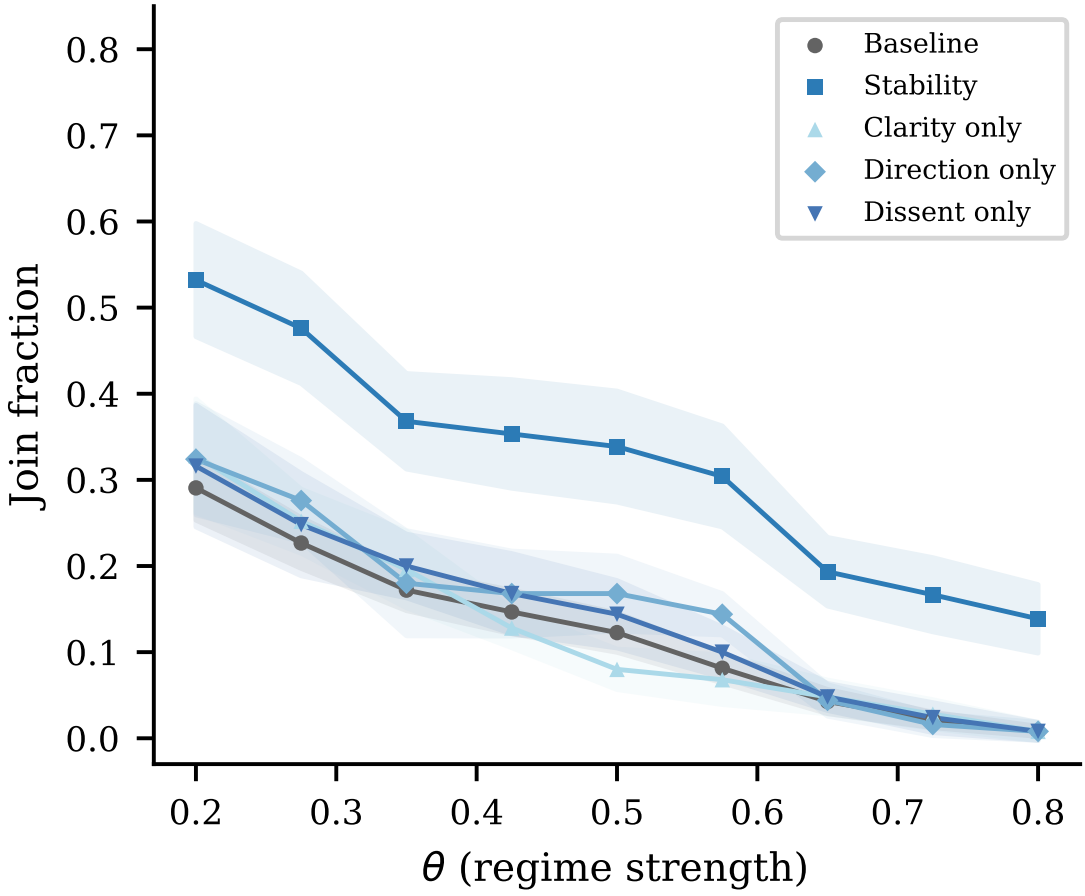


Figure A1: Single-channel decomposition of the stability design. Each panel shows the treatment effect $\Delta(\theta)$ for one channel in isolation.

B.6 Cross-Model Replication of Information Design

Table A4 reports cross-model replication of information design treatments. The flip inversion replicates across all models tested ($r > +0.43$ for all six). The scramble test shows more heterogeneity: Mistral, GPT-OSS, and Qwen3 235B show clean collapse ($r \approx 0$), but Llama 3.3 70B and Ministral 3B retain baseline-level correlations under scramble ($r = -0.81$ and $r = -0.66$), suggesting these models extract signal from features the scramble does not disrupt (e.g., within-country narrative coherence). OLMo shows an attenuated baseline relationship, while 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,

Table A4: 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	0.413	−0.849	—	—	—	—
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
Ministral 3B	0.220	−0.632	0.118	−0.658	0.804	+0.847
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
OLMo 3 7B	0.718	−0.329	0.592	−0.294	0.839	+0.452

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 main experiments hold the payoff narrative constant. A natural question is whether LLM agents respond to perceived costs and benefits in the predicted direction: higher cost of failed action should raise the equilibrium cutoff (less joining), while lower cost should lower it.

I inject cost/benefit context into the briefing header. The *high-cost* treatment prepends: "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." The *low-cost* treatment prepends: "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." All other briefing content is generated identically to the baseline.

Each design is run over the same 9-point θ-grid with 30 repetitions per grid point (25 agents each), totaling 270 country–periods per design. I fit a logistic $P(\text{join}) = \text{logistic}(\beta_0+\beta_1\theta)$ to each design and compare the estimated cutoff $\hat{\theta}^* = -\beta_0/\beta_1$.

The results are strongly directional (Table 5). Under the high-cost narrative, mean join rate drops to 19.0% (−24.6 pp vs. baseline) and the estimated cutoff falls to $\hat{\theta}^* = 0.13$ (SE = 0.010), meaning agents require much stronger signals of regime weakness before joining. Under the low-cost narrative, mean join rate rises to 69.3% (+25.6 pp) with cutoff $\hat{\theta}^* = 0.72$ (SE = 0.007), meaning agents join even when the regime appears moderately strong. The baseline cutoff is $\hat{\theta}^* = 0.42$ (SE = 0.008). The correlation $r(\theta, J)$ remains essentially unchanged across all three conditions ($|r| > 0.87$), confirming that the monotone threshold structure is preserved—only the location shifts, exactly as theory predicts.

Table A5: 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.

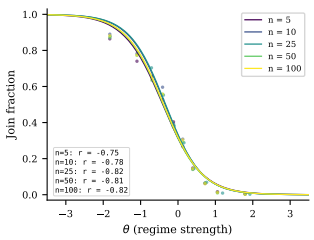
Design	N	Mean join	$r(\theta, J)$	Δ vs baseline
Baseline (no censorship)	270	0.413	-0.85	—
Upper censorship (naïve)	270	0.377	-0.72	-0.036
Upper censorship (known)	270	0.432	-0.74	+0.019

B.10 Censorship with Common Knowledge

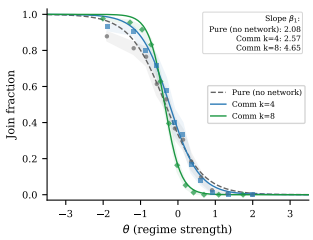
The censorship experiments in the main paper implement upper censorship (suppressing signals above a severity threshold) without telling agents that censorship is occurring. Theory (Kolotilin et al., 2022) typically assumes receivers understand the censorship rule. This raises the question: does making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

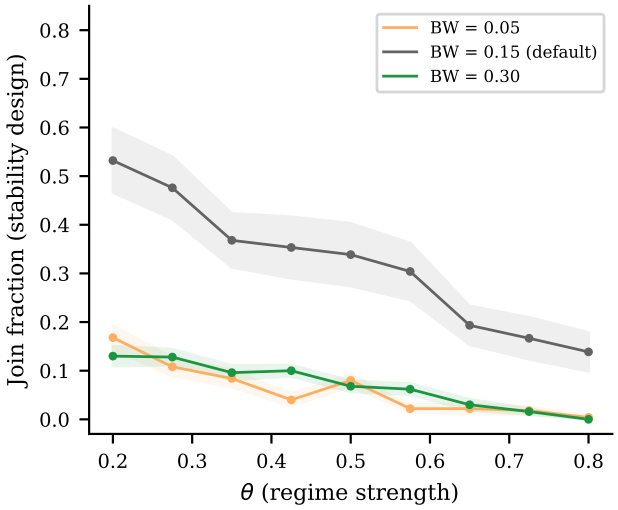
Making censorship common knowledge nearly eliminates the pooling effect (Table A5). Under naïve upper censorship, agents do not know that high-θ signals are being suppressed, so they treat pooled signals at face value; mean join rate falls to 37.7% (−5.9 pp vs. baseline). Under known censorship, agents are warned about the filtering, and mean join rate returns to 43.2% (−0.4 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



(a) Agent count variation ($n \in \{5, 10, 25, 50, 100\}$).



(b) Network density ($k = 4$ vs. $k = 8$).



(c) Bandwidth sensitivity (0.05, 0.15, 0.30).

Figure A2: Robustness checks for equilibrium alignment and treatment effects.

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 at $T \in \{0.3, 0.7, 1.0\}$ using Mistral Small Creative with calibrated parameters. Lower temperature ($T = 0.3$) produces more deterministic outputs; higher temperature ($T = 1.0$) increases sampling entropy.

For each temperature, I run 5 countries × 20 periods (100 country–periods, 2,500 individual decisions) and report the correlation $r(\theta, \text{join fraction})$.

The results are remarkably stable across temperatures (Table A6). The correlation $r(\theta, J)$ ranges from -0.87 to -0.88, mean join rates from 40.6% to 41.2%, and estimated logistic slopes from 2.28 to 2.36. The cutoff estimates cluster tightly near zero ($\hat{\theta}^* \in [-0.08, -0.04]$), consistent with the calibrated center. The monotone threshold pattern is not an artifact of the default decoding temperature.

Table A6: Temperature robustness. The pure global game is run at three LLM decoding temperatures using Mistral Small Creative with calibrated parameters. The correlation $r(\theta, J)$ and logistic parameters are stable across temperatures.

Temperature	N	Mean join	$r(\theta, J)$	Cutoff $\hat{\theta}^*$	Slope $\hat{\beta}$
T=0.3	100	0.412	-0.87	-0.049	2.36
T=0.7	100	0.406	-0.87	-0.079	2.28
T=1.0	100	0.410	-0.88	-0.039	2.34

Table A7: 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

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 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 A7). 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 13.4 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 ($r = -0.855$ vs. baseline

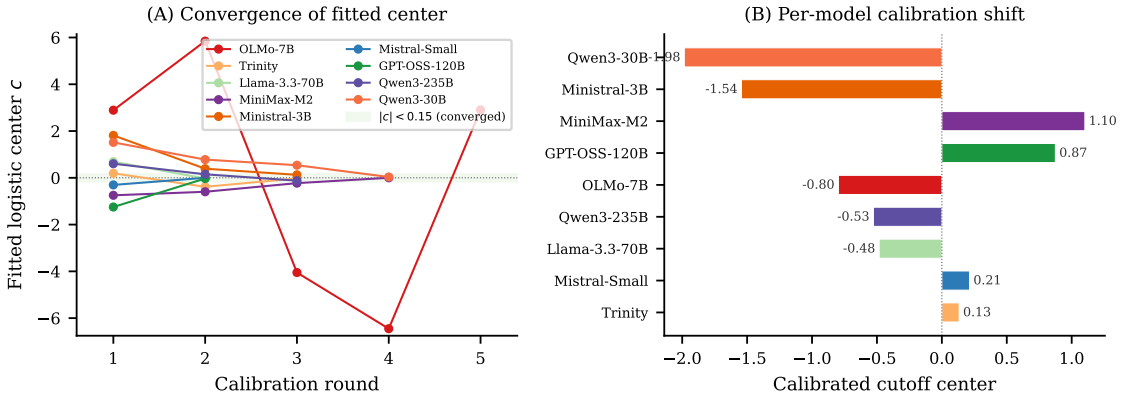


Figure A3: Calibration convergence. *Left:* Trajectory of the fitted logistic center c across autocalibration rounds for each model. The green band marks the convergence criterion ($|c| < 0.15$). All models converge within 2–3 rounds. *Right:* Final calibrated cutoff center per model. Most models require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

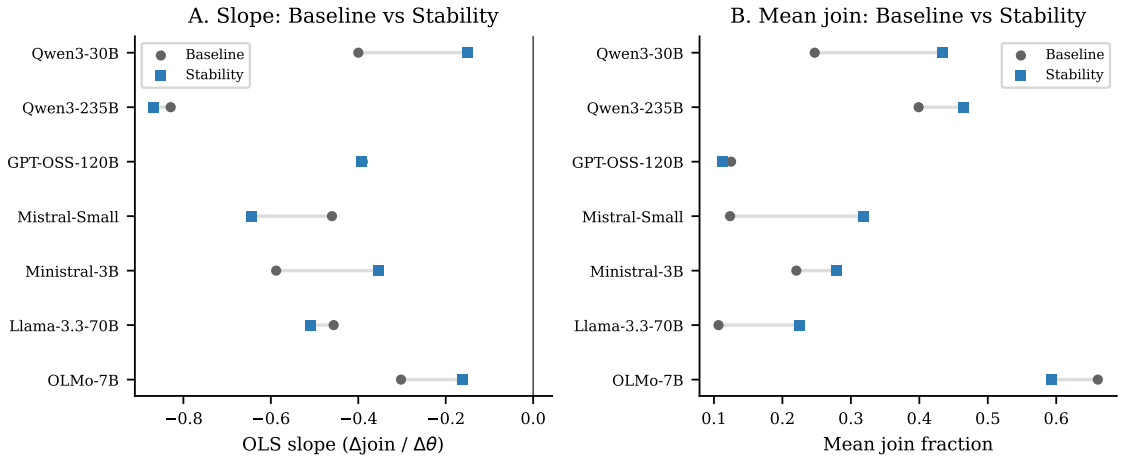


Figure A4: Cross-model replication of information design treatments. Each panel shows join fraction vs. θ for one model under baseline, stability, scramble, and flip conditions.

$r = -0.849$), 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 ($r = -0.873$), 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 ($r = -0.889$). 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).

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.
```

C.3 Randomization

Each country–period draws θ from $\mathcal{N}(\bar{z}, 0.05^2)$, where $\bar{z}$ is itself drawn once per country. 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 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.