```
namepage.1XYZ

goto                                namesection.1count-
  /Author()/Title()/Subject()/Creator(LaTeX with hyperref)/Keywords()
0\376\377\000I\000n\000t\000r\000o\000d\000u\000c\000t\000i\000o\000ngoto
namesection.2count-5\376\377\000R\000e\000l\000a\000t\000e\000d\000\040\000L\000i\000t\000e\000r\000a\000t\000u\000
namesubsection.2.1count-0\376\377\000G\000l\000o\000b\000a\000l\000\040\000G\000a\000m\000e\000s\000\040\000a\000
namesubsection.2.2count-0\376\377\000I\000n\000f\000o\000r\000m\000a\000t\000i\000o\000n\000\040\000D\000e\000s\00
namesubsection.2.3count-0\376\377\000C\000o\000m\000m\000u\000n\000i\000c\000a\000t\000i\000o\000n\000\040\000i\0
namesubsection.2.4count-0\376\377\000R\000e\000g\000i\000m\000e\000\040\000C\000h\000a\000n\000g\000e\000,\000\04
namesubsection.2.5count-0\376\377\000L\000L\000M\000s\000\040\000a\000s\000\040\000E\000c\000o\000n\000o\000m\00
namesection.3count-4\376\377\000T\000h\000e\000\040\000G\000l\000o\000b\000a\000l\000\040\000G\000a\000m\000e\00
namesubsection.3.1count-0\376\377\000S\000e\000t\000u\000pgoto
namesubsection.3.2count-0\376\377\000E\000q\000u\000i\000l\000i\000b\000r\000i\000u\000mgoto
namesubsection.3.3count-0\376\377\000I\000n\000f\000o\000r\000m\000a\000t\000i\000o\000n\000\040\000D\000e\000s\00
namesubsection.3.4count-0\376\377\000T\000e\000s\000t\000a\000b\000l\000e\000\040\000P\000r\000e\000d\000i\000c\000
namesection.4count-6\376\377\000E\000x\000p\000e\000r\000i\000m\000e\000n\000t\000a\000l\000\040\000D\000e\000s\00
namesubsection.4.1count-0\376\377\000O\000v\000e\000r\000v\000i\000e\000wgoto
namesubsection.4.2count-0\376\377\000S\000i\000g\000n\000a\000l\000\040\000G\000e\000n\000e\000r\000a\000t\000i\000
namesubsection.4.3count-0\376\377\000D\000e\000c\000i\000s\000i\000o\000n\000\040\000P\000r\000o\000m\000p\000tgot
namesubsection.4.4count-0\376\377\000P\000a\000r\000t\000\040\000I\000\040\000T\000r\000e\000a\000t\000m\000e\000
namesubsection.4.5count-0\376\377\000P\000a\000r\000t\000\040\000I\000I\000\040\000T\000r\000e\000a\000t\000m\000e
namesubsection.4.6count-0\376\377\000M\000o\000d\000e\000l\000s\000\040\000a\000n\000d\000\040\000P\000a\000r\000
namesection.5count-2\376\377\000D\000o\000\040\000L\000L\000M\000\040\000A\000g\000e\000n\000t\000s\000\040\000
namesubsection.5.1count-0\376\377\000A\000g\000g\000r\000e\000g\000a\000t\000e\000\040\000A\000l\000i\000g\000n\00
namesubsection.5.2count-0\376\377\000C\000r\000o\000s\000s\000-
\000M\000o\000d\000e\000l\000\040\000H\000e\000t\000e\000r\000o\000g\000e\000n\000e\000i\000t\000ygoto
namesection.6count-4\376\377\000F\000a\000l\000s\000i\000f\000i\000c\000a\000t\000i\000o\000n\000\040\000T\000e\000
namesubsection.6.1count-0\376\377\000C\000r\000o\000s\000s\000-
\000P\000e\000r\000i\000o\000d\000\040\000S\000c\000r\000a\000m\000b\000l\000egoto
namesubsection.6.2count-0\376\377\000S\000i\000g\000n\000a\000l\000\040\000F\000l\000i\000pgoto
namesubsection.6.3count-0\376\377\000C\000r\000o\000s\000s\000-
\000M\000o\000d\000e\000l\000\040\000R\000e\000p\000l\000i\000c\000a\000t\000i\000o\000ngoto
namesubsection.6.4count-0\376\377\000I\000d\000e\000n\000t\000i\000f\000i\000c\000a\000t\000i\000o\000n\000:\000\040
namesection.7count-0\376\377\000C\000o\000m\000m\000u\000n\000i\000c\000a\000t\000i\000o\000ngoto
namesection.8count-3\376\377\000I\000n\000f\000o\000r\000m\000a\000t\000i\000o\000n\000\040\000D\000e\000s\000i\00
namesubsection.8.1count-0\376\377\000T\000r\000e\000a\000t\000m\000e\000n\000t\000\040\000E\000f\000f\000e\000c\00
namesubsection.8.2count-0\376\377\000C\000e\000n\000s\000o\000r\000s\000h\000i\000pgoto
namesubsection.8.3count-0\376\377\000F\000a\000l\000s\000i\000f\000i\000c\000a\000t\000i\000o\000n\000\040\000W\000
namesection.9count-2\376\377\000S\000u\000r\000v\000e\000i\000l\000l\000a\000n\000c\000e\000:\000\040\000C\000o\000
namesubsection.9.1count-0\376\377\000D\000e\000s\000i\000g\000ngoto
namesubsection.9.2count-0\376\377\000R\000e\000s\000u\000l\000t\000sgoto
namesection.10count-1\376\377\000P\000r\000o\000p\000a\000g\000a\000n\000d\000a\000:\000\040\000I\000n\000f\000o\0
namesubsection.10.1count-0\376\377\000R\000e\000s\000u\000l\000t\000sgoto
namesection.11count-2\376\377\000I\000n\000s\000t\000r\000u\000m\000e\000n\000t\000\040\000I\000n\000t\000e\000r\00
namesubsection.11.1count-0\376\377\000P\000r\000o\000p\000a\000g\000a\000n\000d\000a\000\040\000\040\000S\000u\00
namesubsection.11.2count-0\376\377\000S\000u\000r\000v\000e\000i\000l\000l\000a\000n\000c\000e\000\040\000\040\000C
namesection.12count-0\376\377\000C\000o\000n\000c\000l\000u\000s\000i\000o\000ngoto
nameappendix.Acount-0\376\377\000R\000o\000b\000u\000s\000t\000n\000e\000s\000s/PageMode/UseOutlinesopenaction
goto page1/FitnameDoc-StartXYZ
```

LLMs Can Play (Global) Games

Khaled Eltokhy
Department of Economics
The Graduate Center, CUNY

February 2026

namesection*.1XYZ Abstract

The canonical Morris–Shin (2003) regime change game—continuous private signals, large groups, strategic uncertainty—has not been implemented experimentally. I embed nine large language models spanning six architecture families in the full game, conveying each agent's private signal as a natural-language intelligence briefing. Across nine models and 1,600 country–periods in the pure treatment (40,000 individual decisions), join rates track the Bayesian Nash equilibrium prediction (mean $r = +0.73$, $p < 0.001$ for every model). Scrambling briefings across periods collapses the correlation to $r = +0.23$; inverting signals flips it to $r = -0.67$—confirming that behavior is driven by briefing content. I then implement censorship, surveillance, and propaganda treatments drawn from the information design and authoritarian control literatures. Upper censorship raises join rates by 18.5 pp through pooling; surveillance reduces them by 17.5 pp through preference falsification; combining the two nearly eliminates coordination (30.9% → 3.7%).

namesection.1XYZ1 Introduction

Coordination games with multiple equilibria are central to the analysis of bank runs (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.diamond1983Diamond and Dybvig, attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.diamond19831983), currency attacks (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.obstfeld1996Obstfeld, attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.obstfeld19961996), and political upheaval (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.angeletos2007aAngeletos et al., attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.angeletos2007a2007). The theory of global games (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.carlsson1993Carlsson and van Damme, attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.carlsson19931993; attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.morris2003Morris and Shin, attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.morris20032003; attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.frankel03Frankel et al., attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.frankel032003) 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 (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.heinemann2004Heinemann et al., attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.heinemann20042004, attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.heinemann20092009; attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.szkup2020Szkup and Trevino, attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.szkup20202020). 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 attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.morris2003Morris and Shin (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.morris20032003) 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 100–600 country–periods per model (Table attr/Border[0 0 0]/H/I/C[1 0 0]goto nametable.caption.131), totaling 1,600 country–periods (40,000 individual decisions) in the pure treatment alone.

The main finding is that LLM agents exhibit behavioral alignment with the Bayesian Nash 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: 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, $p < 0.001$). Pre-play communication raises join rates by 3.7 percentage points, concentrated in weak-regime environments.

Taking this behavioral foundation as given, I then ask: *what can an information designer accomplish?* I implement information designs—stability, instability, public signals, and censorship—and find large effects. Public signal injection reduces join rates by 10.7 pp; upper censorship raises them by 18.5 pp through pooling, consistent with attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.kolotilin2022Kolotilin et al. (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.kolotilin20222022). I then test authoritarian information control: surveillance reduces join rates by 17.5 pp ($p < 0.001$) through attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.kuran1991Kuran (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.kuran19911991) preference falsification, propaganda suppresses coordination in a dose-response pattern, and combining surveillance with censorship nearly eliminates coordination (30.9% → 3.7%).

The paper makes three contributions. First, it provides the first implementation of 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 attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.goldstein2016Goldstein and Huang (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.goldstein20162016), attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.kolotilin2022Kolotilin et al. (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.kolotilin20222022), and attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.edmond2013Edmond (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.edmond20132013) in a coordination game—including censorship, surveillance, and propaganda treatments that are difficult to implement with human subjects at scale. Third, it demonstrates that LLMs can serve as experimental subjects for strategic environments, extending the attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.horton2023Horton (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.horton20232023) *homo silicus* methodology beyond 2×2 games to the continuous-signal, N-player coordination games that dominate applied theory.

Section attr/Border[0 0 0]/H/I/C[1 0 0]goto namesection.22 reviews the related literature. Section attr/Border[0 0 0]/H/I/C[1 0 0]goto namesection.33 presents the theoretical framework. Section attr/Border[0 0 0]/H/I/C[1 0 0]goto namesection.44 describes the experimental design. Section attr/Border[0 0 0]/H/I/C[1 0 0]goto namesection.55 reports the main results on equilibrium alignment; Section attr/Border[0 0 0]/H/I/C[1 0 0]goto namesection.66 presents the falsification tests. Section attr/Border[0 0 0]/H/I/C[1 0 0]goto namesection.77 analyzes pre-play communication. Sections attr/Border[0 0 0]/H/I/C[1 0 0]goto namesection.88–attr/Border[0 0 0]/H/I/C[1 0 0]goto namesection.1111 cover information design, surveillance, propaganda, and their interactions. Appendix attr/Border[0 0 0]/H/I/C[1 0 0]goto nameappendix.AA reports robustness checks. Section attr/Border[0 0 0]/H/I/C[1 0 0]goto namesection.1212 concludes.

namesection.2XYZ2 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.

namesubsection.2.1XYZ2.1 Global Games and Equilibrium Selection

The theory of global games resolves the equilibrium multiplicity that plagues coordination games by introducing heterogeneous private information. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.carlsson1993Carlsson and van Damme (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.carlsson19931993) showed that adding arbitrarily small noise to a 2×2 coordination game generically selects the risk-dominant equilibrium via iterated dominance. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.morris1998Morris and Shin (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.morris19981998) applied this technique to currency crises, demonstrating that heterogeneous private signals about fundamentals deliver a unique threshold equilibrium even in large-player coordination games. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.frankel03Frankel et al. (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.frankel032003) 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 attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.morris2003Morris and Shin (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.morris20032003), who established the threshold equilibrium structure I implement experimentally. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.angeletos2007aAngeletos et al. (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.angeletos2007a2007) extended the framework to dynamic settings where agents learn across periods, showing that multiplicity can reemerge when agents observe whether the regime survived previous rounds. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.morris2002Morris and Shin (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.morris20022002) 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. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.heinemann2004Heinemann et al. (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.heinemann20042004) 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. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.heinemann2009Heinemann et al. (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.heinemann20092009) measured strategic uncertainty directly through certainty equivalents. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.shurchkov2013Shurchkov (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.shurchkov20132013) tested dynamic global games, finding that subjects learn from failed attacks. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.szkup2020Szkup and Trevino (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.szkup20202020) 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. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.helland2021Helland et al. (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.helland20212021) 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.

namesubsection.2.2XYZ2.2 Information Design and Bayesian Persuasion

attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.kamenica2011Kamenica and Gentzkow (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.kamenica20112011) 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. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.bergemann2016Bergemann and Morris (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.bergemann20162016) unified Bayesian persuasion with correlated equilibrium under the concept of Bayes Correlated Equilibrium. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.bergemann2019informationBergemann and Morris (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.bergemann2019information2019) provided a comprehensive survey integrating cheap talk, persuasion, and robust mechanism design.

The application to coordination games is directly relevant. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.goldstein2016Goldstein and Huang (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.goldstein20162016) 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. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.inostroza2025Inostroza and Pavan (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.inostroza20252025) solved the optimal public information design problem in a global game with heterogeneous private signals, characterizing when pass/fail structures are optimal. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.kolotilin2022Kolotilin et al. (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.kolotilin20222022) proved that upper censorship—pooling all states above a threshold while revealing below—is optimal for all priors when the sender's marginal utility is quasi-concave. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.mathevet2020Mathevet et al. (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.mathevet20202020) 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.

namesubsection.2.3XYZ2.3 Communication in Coordination Games

The strategic communication literature begins with attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.crawford1982Crawford and Sobel (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.crawford19821982), who characterized the partition equilibria of one-way cheap talk games. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.farrell1996Farrell and Rabin (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.farrell19961996) introduced the concept of *self-committing* messages—messages credible because the speaker has an incentive to follow through if believed—which is the mechanism through which pre-play communication improves coordination.

Experimental evidence on cheap talk in coordination games is extensive. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.blume2007Blume and Ortmann (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.blume20072007) found that pre-play cheap talk dramatically increases coordination on the Pareto-dominant equilibrium. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.ellingsen2010Ellingsen and Östling (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.ellingsen20102010) used the level-k model to predict when communication helps or hurts. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.brandts2006Brandts and Cooper (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.brandts20062006) showed that leadership communication helps escape low-effort traps in minimum-effort games. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.avoyan2020Avoyan (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.avoyan20202020) showed that cheap talk improves investment efficiency in a two-player global game.

In real-world coordination, network structure matters. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.enikolopov2020Enikolopov et al. (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.enikolopov20202020) provided causal evidence that social media penetration increased protest incidence in Russia's 2011–12 protests. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.jost2018Jost et al. (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.jost20182018) reviewed how social media facilitates protest through information sharing and network-based social pressure.

My communication treatment embeds agents in a Watts-Strogatz small-world network and allows natural-language messaging before the coordination decision.

namesubsection.2.4XYZ2.4 Regime Change, Censorship, and Surveillance

The theoretical literature on authoritarian information control builds directly on the global games framework. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.edmond2013Edmond (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.edmond20132013) embedded costly propaganda into the Morris–Shin regime change game: the regime shifts the distribution of citizens' signals to make itself appear stronger. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.li2022Li et al. (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.li20222022) extended this to adversarial information design with heterogeneous "local obfuscation."

On censorship, attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.shadmehr2015Shadmehr and Bernhardt (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.shadmehr20152015) characterized when a ruler benefits from censoring media reports. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.lorentzen2014Lorentzen (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.lorentzen20142014) modeled strategic censorship where regimes permit watchdog journalism while suppressing content revealing regime-wide discontent. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.king2013King et al. (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.king20132013) provided empirical evidence that Chinese censorship targets content with *collective action potential*. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.king2017King et al. (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.king20172017) documented that the Chinese government fabricates approximately 448 million social media posts per year, predominantly cheerleading rather than engaged argument.

The surveillance literature documents chilling effects. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.penney2016Penney (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.penney20162016) found that Wikipedia views of terrorism-related articles declined after the Snowden revelations. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.stoycheff2016Stoycheff (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.stoycheff20162016) showed experimentally that surveillance awareness suppresses minority opinion expression. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.chen2019Chen and Yang (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.chen20192019) found that free VPN access does not generate demand for uncensored information—censorship works partly through preference shaping.

attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.kuran1991Kuran (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.kuran19911991) provides the foundational theory: *preference falsification*—the systematic misrepresentation of political preferences under social pressure—explains how authoritarian regimes maintain apparent stability. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.guriev2019Guriev and Treisman (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.guriev20192019) documented the rise of "informational autocrats" who maintain power through information manipulation rather than mass repression. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.carter2021Carter and Carter (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.carter20212021) found that pro-regime propaganda reduces protest probability by 15% per standard deviation.

My surveillance and propaganda treatments directly test these mechanisms computationally within the full regime change game—an environment that would be difficult to implement with human subjects at this scale, and where surveillance and propaganda manipulations raise ethical concerns that computational subjects avoid.

namesubsection.2.5XYZ2.5 LLMs as Economic and Strategic Agents

attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.horton2023Horton (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.horton20232023) proposed treating LLMs as "homo silicus"—computational models of human decision-makers. attr/Border[0 0 0]/H/I/C[0 1

0]goto namecite.argyle2023Argyle et al. (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.argyle20232023) showed that GPT-3 conditioned on demographic backstories achieves “algorithmic fidelity” in replicating attitudinal distributions. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.aher2023Aher et al. (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.aher20232023) introduced the “Turing Experiment” framework, replicating results from the Ultimatum Game, Milgram experiment, and Wisdom of Crowds.

In game-theoretic settings, attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.akata2025Akata et al. (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.akata20252025) ran LLMs in finitely repeated 2×2 games, finding that LLMs perform well in self-interested games but struggle in coordination games. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.lore2024Lorè and Heydari (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.lore20242024) compared GPT-3.5, GPT-4, and LLaMA-2 across multiple game types, finding distinct strategic signatures. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.guo2023Guo (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.guo20232023) found GPT exhibits level-0 to level-1 strategic depth.

Several papers document systematic behavioral differences across model families. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.palatsi2025Palatsi et al. (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.palatsi20252025) found that Llama replicates human cooperation with high fidelity while Qwen tracks Nash equilibrium closely. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.petrov2025Petrov et al. (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.petrov20252025) evaluated 22 LLMs on a behavioral game theory battery, finding that model scale alone does not predict strategic performance. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.corrupted2025Corrupted by Reasoning (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.corrupted20252025) showed that reasoning-focused models free-ride in public goods games at much higher rates than standard models.

The alignment literature provides further context. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.huang2024Huang et al. (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.huang20242024) found that increasing ethical alignment increases risk aversion by 2–8%. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.carlini2025Carlini et al. (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.carlini20252025) showed that fine-tuning for chatbot use amplifies omission bias. These findings motivate my design choice to convey strategic stakes through narrative rather than an explicit payoff matrix: more capable models can identify dominant strategies from formatted tables and short-circuit the information-processing channel I aim to study.

The LLM-based agent-based modeling literature has expanded rapidly. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.park2023Park et al. (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.park20232023) demonstrated emergent social behavior among 25 LLM agents in a simulated town. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.piatti2024Piatti et al. (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.piatti20242024) introduced GovSim, where only GPT-4 and Claude-3 Opus achieve sustainable equilibrium. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.bail2024Bail et al. (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.bail20242024) showed that networks of LLM agents reproduce empirically observed opinion dynamics. Critical reviews by attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.gao2025Gao et al. (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.gao20252025) and attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.grossmann2025Grossmann et al. (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.grossmann20252025) warn that validation remains poorly addressed. attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.sun2025surveySun et al. (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.sun2025survey2025) identify coordination games as a consistent failure mode.

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.

namesection.3XYZ3 The Global Game of Regime Change

namesubsection.3.1XYZ3.1 Setup

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:

nameequation.3.1XYZ

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

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

nameequation.3.2XYZ

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

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.

namesubsection.3.2XYZ3.2 Equilibrium

nameproposition.1XYZ

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*

nameequation.3.3XYZ

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

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

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

nameequation.3.4XYZ

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

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

namesubsection.3.3XYZ3.3 Information Design in the Regime Change Game

An information designer controls the mapping $\pi : \Theta \rightarrow \Delta(\mathcal{S})$ 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:

nameItem.1XYZ**Clarity.** The width of the Gaussian kernel that maps z-scores to text. Wider kernels produce more ambiguous briefings; narrower kernels produce more informative briefings.

1. nameItem.2XYZ**Directional precision.** The slope of the mapping from z-score to briefing sentiment. A steeper slope means briefings more accurately reflect the direction of the underlying signal.

2. nameItem.3XYZ**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:

nameequation.3.5XYZ

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

where bandwidth = 0.15 in the baseline specification.

namesubsection.3.4XYZ3.4 Testable Predictions

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

namehypothesis.1XYZ

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

namehypothesis.2XYZ

Hypothesis 2 (Signal Dependence)**.** *The correlation in Hypothesis attr/Border[0 0 0]/H/I/C[1 0 0]goto namehypothesis.11 should collapse when the mapping from θ to briefing content is broken (scramble test).*

namehypothesis.3XYZ

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

namehypothesis.4XYZ

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

namehypothesis.5XYZ

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

namehypothesis.6XYZ

Hypothesis 6 (Upper Censorship)**.** *Upper censorship should raise join rates in censored states by creating pooling (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.kolotilin2022Kolotilin et al., attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.kolotilin20222022).*

namehypothesis.7XYZ

Hypothesis 7 (Surveillance Chilling Effect)**.** *Informing agents that communications are monitored should reduce coordination (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.kuran1991Kuran, attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.kuran19911991).*

namehypothesis.8XYZ

Hypothesis 8 (Propaganda Dose-Response)**.** *Regime plant agents transmitting pro-regime messages should suppress coordination, with the effect increasing in the number of plants (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.edmond2013Edmond, attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.edmond20132013).*

namesection.4XYZ4 Experimental Design

namesubsection.4.1XYZ4.1 Overview

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.

namesubsection.4.2XYZ4.2 Signal Generation and Briefing Rendering

For each country–period, nature draws θ from a normal distribution. Each agent i receives a private signal $x_i = \theta + \varepsilon_i$ and computes a z-score $z_i = (x_i - \bar{z})/\sigma$, where $\bar{z}$ is a public prior mean drawn randomly for each country. 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.

namesubsection.4.3XYZ4.3 Decision Prompt

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.

namesubsection.4.4XYZ4.4 Part I Treatments

namesection*.2XYZ
Pure global game. Each agent decides independently based on their private briefing.

namesection*.3XYZ
Communication. Before deciding, agents send a message to a small network of "trusted contacts" (Watts-Strogatz small-world network, $k = 4$, $p = 0.3$). They then decide with access to both their briefing and received messages.

namesection*.4XYZ
Falsification: Scramble. All briefings across all periods within a country are pooled and randomly redistributed, breaking the link between a period's θ and the distribution of briefings.

namesection*.5XYZ
Falsification: Flip. The z-score is negated before briefing generation: agents who should see weak-regime cues receive strong-regime cues and vice versa.

namesubsection.4.5XYZ4.5 Part II Treatments

namesection*.6XYZ
Stability-maximizing. Near θ^*, clarity width is multiplied by 4, the dissent floor is raised to 0.45, and the directional slope is flattened by a factor of 0.25.

namesection*.7XYZ
Instability-maximizing. Near θ^*, 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.

namesection*.8XYZ
Public signal injection. A shared "news bulletin" generated from θ with 4 observations is appended to each agent's private briefing, creating a common-knowledge channel.

namesection*.9XYZ
Upper censorship. States above θ^* are pooled: agents receive an identical censored briefing. States below θ^* are fully revealed (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.kolotilin2022Kolotilin et al., attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.kolotilin20222022).

namesection*.10XYZ
Lower censorship. The mirror image: states below θ^* are pooled, and states above are fully revealed.

namesection*.11XYZ
Surveillance. The communication prompt is augmented with: *"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."*

namesection*.12XYZ
Propaganda. Regime plant agents ($k = 2, 5, 10$) participate in the communication network but transmit fixed pro-regime messages and always STAY.

namesubsection.4.6XYZ4.6 Models and Parameters

I test nine architecturally distinct models spanning six architecture families (Table attr/Border[0 0 0]/H/I/C[1 0 0]goto nametable.caption.131). 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 attr/Border[0 0 0]/H/I/C[1 0 0]goto nametable.caption.131. I vary B and

namepage.9XYZ

nametable.caption.13XYZ

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	600	600	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		**1600**	**1600**	**2400**

namefigure.caption.14XYZ

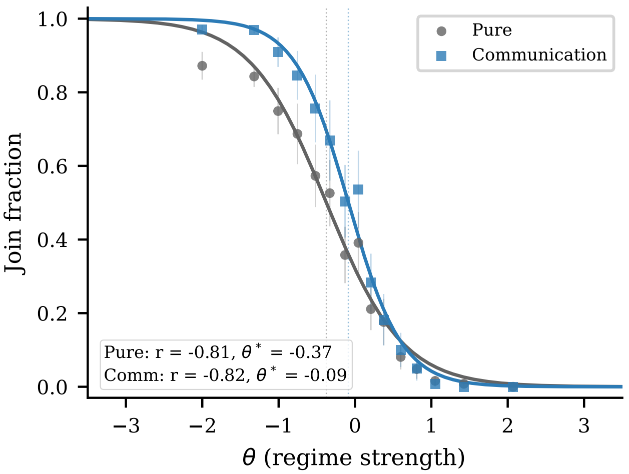


Figure 1: Empirical join fraction vs. theoretical attack mass $A(\theta)$, pooled across all nine models. Each point is one country–period (1,600 observations). The pooled correlation is $r = +0.67$; the mean across individual models is $r = +0.73$ (Table attr/Border[0 0 0]/H/I/C[1 0 0]goto nametable.caption.152), with all nine individually significant at $p < 0.001$.

C such that $\theta^* = B/(B+C)$ has a mean of approximately 0.45 across periods.

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.

namesection.5XYZ5 Do LLM Agents Play the Global Game?

namesubsection.5.1XYZ5.1 Aggregate Alignment

nameresult.1XYZ

Result 1 (Equilibrium Alignment). *Across nine models and 1,600 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 attr/Border[0 0 0]/H/I/C[1 0 0]goto nametable.caption.152 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 any individual model's because heterogeneous mean join rates across models add noise when pooling. The pooled OLS regression yields:
nameequation.5.6XYZ

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

The slope of 0.52 indicates that LLM agents respond to the theoretical attack mass at roughly half the predicted rate. The intercept of 0.17 reflects a baseline propensity to join even when the equilibrium predicts near-zero participation.

The mean join rate across all models is 0.43, 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.

namesubsection.5.2XYZ5.2 Cross-Model Heterogeneity

The cross-model results reveal two patterns. First, the equilibrium alignment is stable across architectures: the nine models span a range of only $r \in [0.65, 0.84]$, despite differing in parameter count by nearly two orders of magnitude (3B to 235B). This suggests that the ability to extract decision-relevant content from narrative text is a generic capability of instruction-tuned language models.

Second, the mean join rate varies across models—from 0.37 (Mistral Small Creative) to 0.72 (OLMo 3 7B)—reflecting model-specific action biases. These biases shift the intercept of the join-fraction-vs-attack-mass relationship but do not substantially affect 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.

namepage.10XYZ

nametable.caption.15XYZ

Table 2: 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	+0.67	+0.68	+0.42	−0.62	600	0.37
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	+0.67	+0.10	−0.63	1600	0.43
Mean across models	+0.73	+0.73	+0.23	−0.67	—	—

namefigure.caption.16XYZ

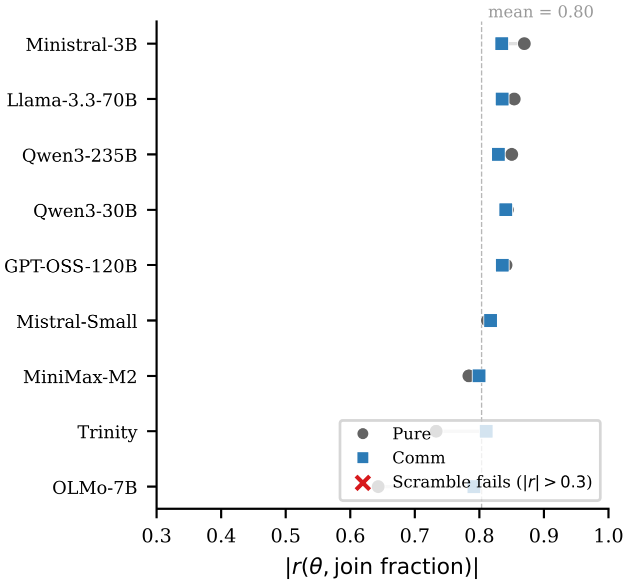


Figure 2: Cross-model summary of signal monotonicity. Points report $|r(\theta,\ \text{join})|$ under pure and communication; x markers (if any) indicate models where scrambling does not collapse the correlation ($|r| > 0.3$).

namesection.6XYZ6 Falsification Tests

The positive correlation documented in Section attr/Border[0 0 0]/H/I/C[1 0 0]goto namesection.55 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.

namesubsection.6.1XYZ6.1 Cross-Period Scramble

nameresult.2XYZ

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 $z = 18.59$, $p < 0.001$).*

The scramble preserves the marginal distribution of briefing content but breaks the mapping from each period's θ to the signals agents receive. The residual positive correlation (+0.23 mean, +0.10 pooled) is small relative to the baseline and varies across models (−0.13 to +0.42), consistent with noise in finite samples.

namesubsection.6.2XYZ6.2 Signal Flip

nameresult.3XYZ

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 $z = 40.76$, $p < 0.001$).*

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

namesubsection.6.3XYZ6.3 Cross-Model Replication

The pure → scramble → flip pattern replicates across all models with full falsification suites:

3. **Mistral Small Creative:** +0.67 → +0.42 → −0.62

- **Llama 3.3 70B:** +0.79 → +0.33 → −0.73

namefigure.caption.17XYZ

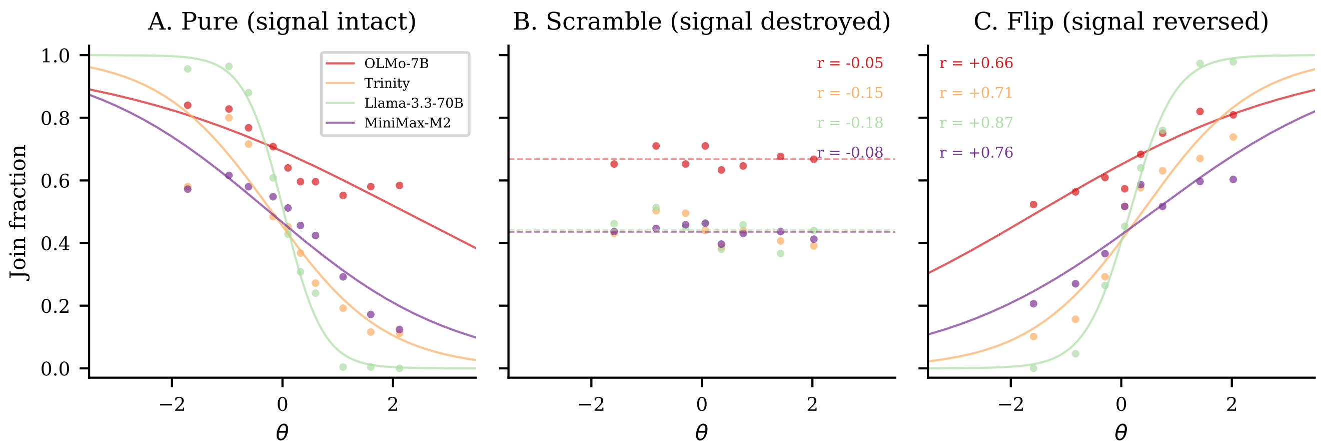


Figure 3: Falsification triptych. *Left:* Pure global game (mean $r = +0.73$). *Center:* Cross-period scramble breaks the θ-to-briefing mapping (mean $r = +0.23$). *Right:* Signal flip inverts the mapping (mean $r = -0.67$). Each panel pools data from models with full falsification suites.

- **OLMo 3 7B:** $+0.65 \to +0.14 \to -0.56$
- **Ministral 3B:** $+0.79 \to +0.30 \to -0.74$
- **Qwen3 30B:** $+0.78 \to +0.32 \to -0.71$
- **GPT-OSS 120B:** $+0.70 \to -0.13 \to -0.64$
- **Trinity Large:** $+0.84 \to +0.32 \to -0.70$
- **MiniMax M2-Her:** $+0.66 \to +0.14 \to -0.69$

Every model shows the same qualitative pattern: strong positive correlation under pure, collapse under scramble, and sign reversal under flip. This holds for every model with the full falsification suite.

namefigure.caption.18XYZ

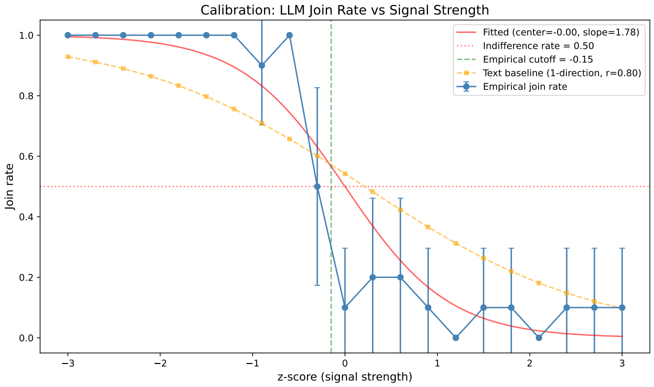


Figure 4: Text baseline identification test. Blue: empirical LLM join rate across z-scores. Orange: naive text-only predictor ($1 - \text{direction}$, $r = 0.80$). Red: fitted logistic (slope $= 1.78$). The LLM produces a steeper transition than the text baseline, indicating processing beyond sentiment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

namesubsection.6.4XYZ6.4 Identification: Text Baseline Test

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 attr/Border[0 0 0]/H/I/C[1 0 0]goto namefigure.caption.184). 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.

namesection.7XYZ7 Communication

nameresult.4XYZ

Result 4 (Communication raises join rates asymmetrically)**.** *Pre-play communication raises the*

namefigure.caption.19XYZ

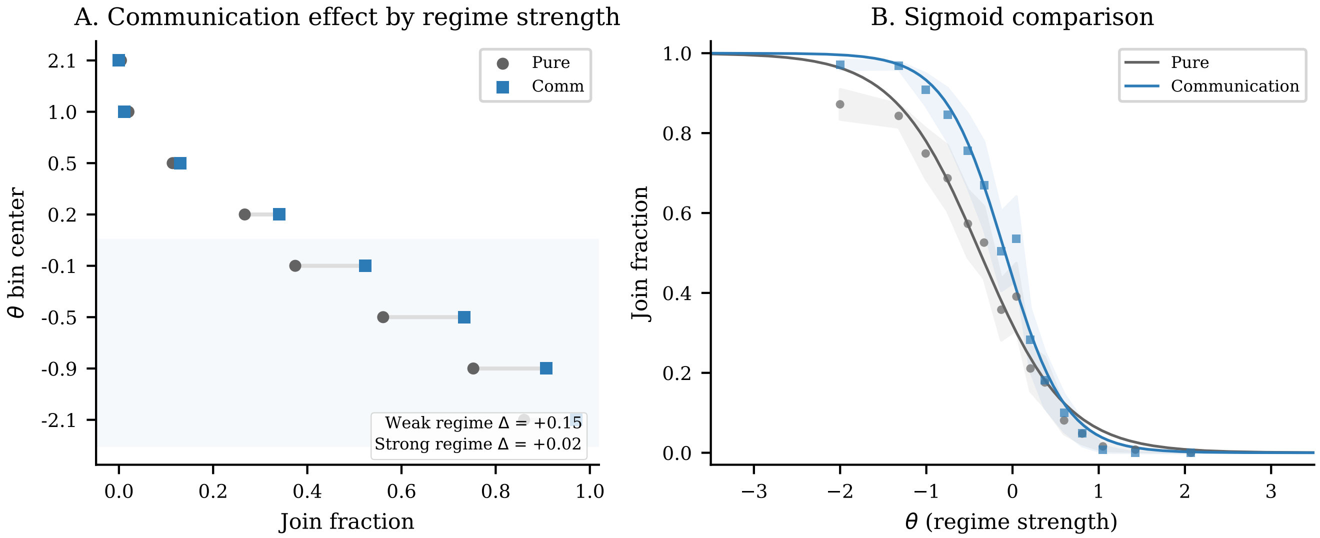


Figure 5: Communication effect by regime strength, pooled across nine models. Communication increases join rates for weak regimes ($\theta < \theta^*$) but has no effect or slightly reduces join rates for strong regimes ($\theta > \theta^*$).

mean join rate by 3.7 percentage points, from 0.429 to 0.466 ($t = 2.75$*,* $p < 0.01$*), pooled across nine models. The effect is concentrated in weak-regime environments (*$+8.8$ *pp for* $\theta < \theta^* - 1$*) and reverses for strong regimes (*-2.5 *pp for* $\theta > \theta^* + 1$*).*

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 +8.3 pp), while three show small negative effects (−2.3 to −4.6 pp). Communication does not change the *correlation* with the theoretical prediction ($r = +0.73$ vs. +0.73 under pure); it shifts the level of coordination but preserves the signal structure. The asymmetry 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.

namesection.8XYZ8 Information Design

Part I reported alignment using $r(J, A(\theta))$, which is positive because both the attack mass and the join fraction decrease in θ. From this section onward, the information design experiments use a fixed θ-grid and report $r(J, \theta)$ directly, which is *negative* under equilibrium play. The sign change reflects the convention, not a behavioral reversal.

namefigure.caption.20XYZ

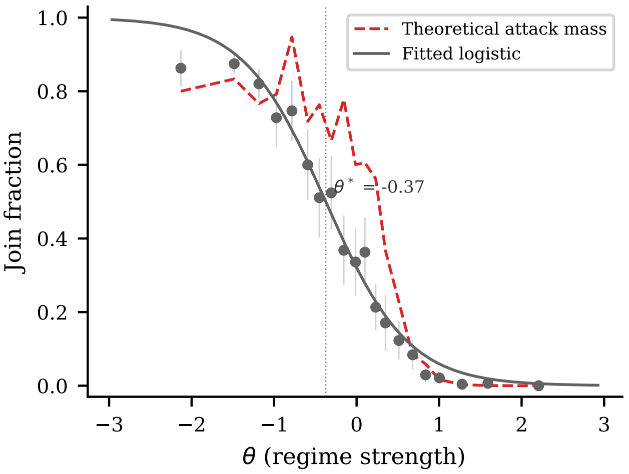


Figure 6: Agent-level threshold behavior. The probability of joining as a function of the agent's z-score, pooled across models.

namesubsection.8.1XYZ8.1 Treatment Effects

Table attr/Border[0 0 0]/H/I/C[1 0 0]goto nametable.caption.233 summarizes the main results. The baseline condition produces a mean join rate of 12.4% with a strong negative correlation between θ and join fraction ($r = -0.812$, $p < 0.001$).
nameresult.5XYZ

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

namefigure.caption.21XYZ

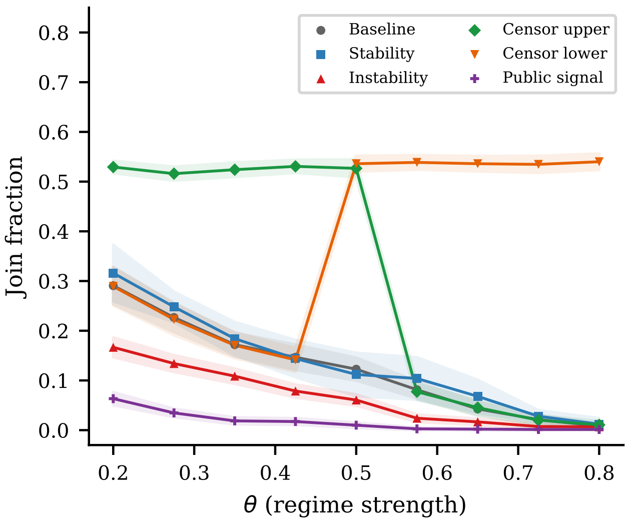


Figure 7: Join fraction as a function of θ under baseline, stability, instability, and public signal information designs. Baseline and stability have $N = 270$; instability and public signal have $N = 540$. Mistral Small Creative model.

The **stability design** sharply increases coordination: mean join rises from 12.4% to 31.9% (+19.5 pp), and the θ–join relationship flattens ($r = -0.626$ vs. -0.812 at baseline). The increase is present at every θ grid point; even at $\theta = 0.80$, join rises from 0.8% to 13.9%. This pattern is consistent with pooling induced by ambiguity and mixed evidence: when strong-regime briefings retain substantial destabilizing cues, agents no longer sharply reduce participation in high-θ states.

The **instability design** reduces the mean join rate to 6.7%, a reduction of 5.7 pp from baseline. The sharper signals allow agents to more accurately perceive regime strength, and agents with sharper information about states above θ^* are more clearly deterred from joining.

The **public signal** produces the largest reduction in coordination: mean join rate falls to 1.7%, a reduction of 10.7 pp. The common news bulletin reveals that the regime is strong (since the grid is centered on θ^* and extends upward), and the overweighting of public information documented by attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.morris2002Morris and Shin (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.morris20022002) amplifies its effect. The correlation between θ and join fraction drops to $r = -0.537$, suggesting agents weight the public signal heavily enough to partially displace private information.

namesubsection.8.2XYZ8.2 Censorship

attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.kolotilin2022Kolotilin et al. (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.kolotilin20222022) proved that upper censorship is optimal for all priors when the sender's marginal utility is quasi-concave. I implement two censorship designs.

namefigure.caption.22XYZ

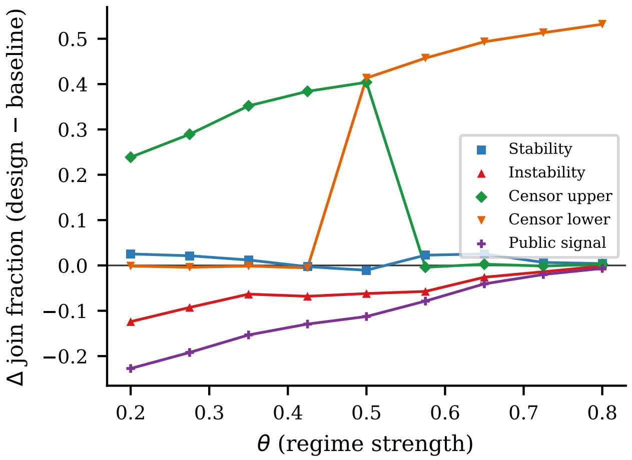


Figure 8: Treatment effect $\Delta(\theta)$ = design join − baseline join as a function of θ. Negative values indicate the design suppresses coordination.

nametable.caption.23XYZ

Table 3: 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	0.124	−0.812	—	270
Stability	0.319	−0.626	+0.195	270
Instability	0.067	−0.740	-0.057	540
Public signal	0.017	−0.537	-0.107	540
Scramble	0.121	+0.036	-0.003	270
Flip	0.663	+0.823	+0.540	270

nameresult.6XYZ

Result 6 (Upper Censorship Raises Join Rates)**.** *Upper censorship raises the mean join rate to 30.9%, an increase of 18.5 pp over baseline. The effect is concentrated in the censored region ($\theta \geq \theta^*$): at $\theta = 0.50$, join rates rise from 12.3% to 52.7% (+40.4 pp). Below the censorship threshold, join rates are essentially unchanged.*

The flat plateau at approximately 53% in the censored region is clear: agents who cannot distinguish $\theta = 0.50$ from $\theta = 0.80$ behave as if the regime is moderately vulnerable. This is consistent with the pooling effect predicted by the theory.

nameresult.7XYZ

Result 7 (Lower Censorship Creates a Symmetric Plateau)**.** *Lower censorship produces a mean join rate of 39.0% (+26.6 pp over baseline). The correlation flips sign to $r = +0.731$, reflecting the inverted structure.*

namefigure.caption.24XYZ

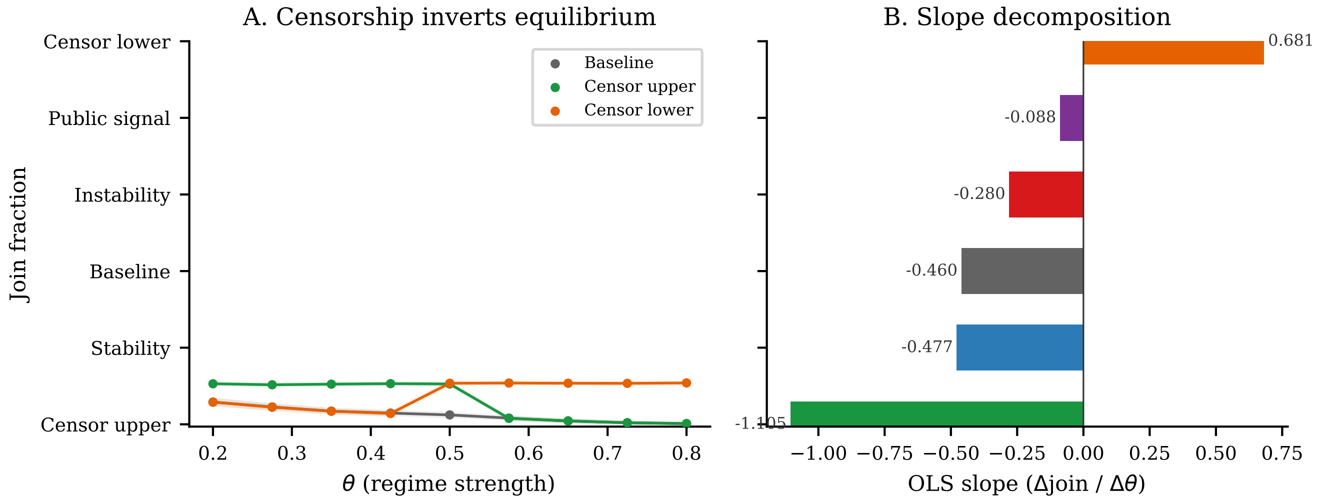


Figure 9: Join fraction under upper and lower censorship vs. baseline. Upper censorship pools states above θ^*, creating a flat join rate in the censored region.

namefigure.caption.25XYZ

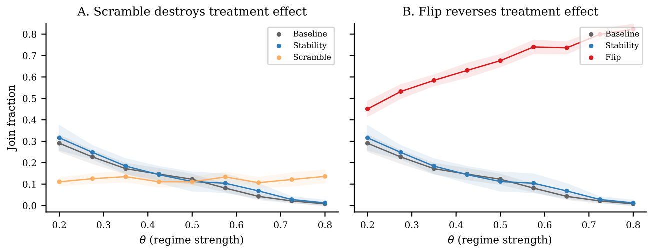


Figure 10: Falsification within information design. Scrambling collapses the θ-join correlation to $r = +0.037$; flipping inverts it to $r = +0.823$.

namesubsection.8.3XYZ8.3 Falsification Within Information Design

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.

namesection.9XYZ9 Surveillance: Computational Preference Falsification

attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.kuran1991Kuran (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.kuran19911991) argued that authoritarian regimes sustain themselves partly through preference falsification. I test this by introducing a surveillance treatment in the communication game.

namesubsection.9.1XYZ9.1 Design

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. This isolates the chilling effect: any difference must arise from agents self-censoring their communications.

namesubsection.9.2XYZ9.2 Results

nameresult.8XYZ

Result 8 (Surveillance Produces a Large Chilling Effect)**.** *In the primary model (Mistral Small Creative), surveillance reduces mean join rates from 45.2% to 27.7%, a difference of 17.5 percentage points (*$t = -9.17$*,* $p < 0.001$*). The correlation between* θ *and join fraction is preserved under surveillance (*$r = -0.809$ *vs.* -0.817*), indicating that surveillance operates as a level shift rather than disrupting signal processing.*

The magnitude—17.5 pp—exceeds the communication premium itself (3.7 pp in Part I). Surveillance does not merely neutralize the coordination benefit; it contaminates the information environment with self-censored messages, pushing join rates *below* what they would be without communication at all. The effect replicates across three architectures: Mistral (-17.5 pp), Llama (-8.9 pp), and Qwen3 (-10.9 pp).

namefigure.caption.26XYZ

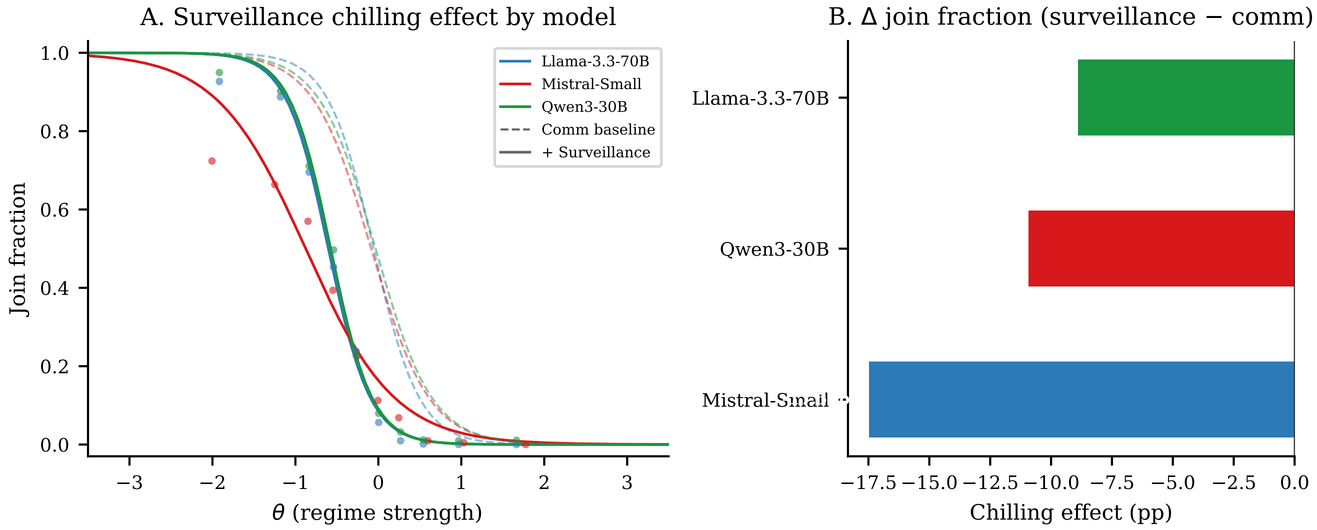


Figure 11: Join rates under regular communication vs. surveillance communication. Surveillance reduces join rates by 17.5 percentage points ($p < 0.001$). Results shown for three models: Mistral (−17.5 pp), Llama (−8.9 pp), and Qwen3 (−10.9 pp).

namesection.10XYZ10 Propaganda: Information Contamination

attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.edmond2013Edmond (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.edmond20132013) 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.

namesubsection.10.1XYZ10.1 Results

nameresult.9XYZ

Result 9 (Propaganda Suppresses Coordination Primarily Through Mechanical Dilution)**.** *Mean join fraction falls from 45.2% ($k = 0$) 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: 45.2% ($k = 0$), 40.7% ($k = 2$, −4.5 pp), 39.1% ($k = 5$, −6.1 pp), 38.8% ($k = 10$, −6.4 pp).*

Propaganda works through two channels: a *mechanical* channel (plants always STAY, directly reducing attack mass) and a *behavioral* channel (pro-regime messages reduce real citizens' willingness to join). The mechanical channel is approximately linear in k; the behavioral channel saturates quickly—doubling plants from 5 to 10 produces essentially no additional behavioral effect (−0.3 pp).

The propaganda effect replicates with Llama 3.3 70B, which shows a behavioral effect of −2.7 pp at $k = 5$—smaller than Mistral's −6.1 pp but in the same direction, confirming the qualitative pattern across architectures.

nametable.caption.28XYZ

Table 4: 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)	.452	.452	−0.817	—
Prop $k = 2$	.375	.407	−0.809	-0.045
Prop $k = 5$	.313	.391	−0.822	-0.061
Prop $k = 10$	.233	.388	−0.818	-0.064
Surveillance	.278	.278	−0.809	-0.175
Prop+Surv	.194	—	−0.829	—

namesection.11XYZ11 Instrument Interactions

namesubsection.11.1XYZ11.1 Propaganda × Surveillance

When propaganda ($k = 5$) and surveillance are combined, the mean join rate falls to 19.4%, a reduction of 25.8 pp from baseline. The combined effect is less than the sum of individual effects (31.4 pp), suggesting diminishing returns: once surveillance has suppressed expressed dissent, propaganda provides less additional deterrence.

namefigure.caption.27XYZ

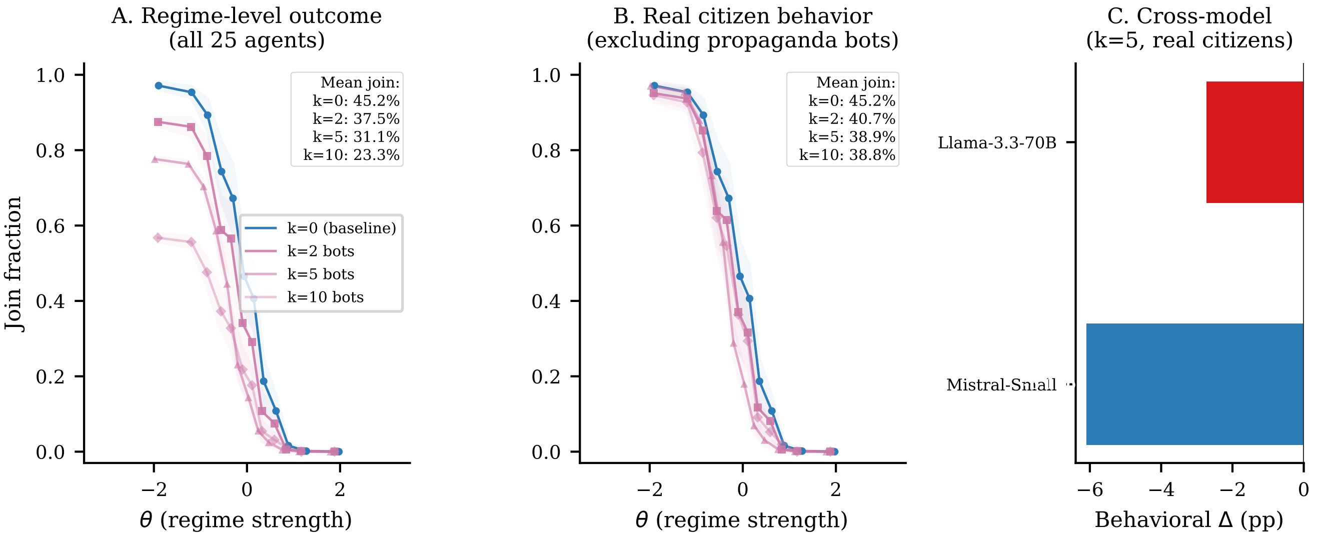


Figure 12: Dose-response relationship between number of propaganda agents and mean join rate. Results shown for Mistral (primary) and Llama (replication). Regular communication ($k = 0$) serves as baseline.

nametable.caption.29XYZ

Table 5: Surveillance × censorship interaction (primary model: Mistral Small Creative).

Design	No Surv.	Surv.	Δ
Baseline	0.124	0.009	-0.115
Upper cens.	0.309	0.037	-0.272
Lower cens.	0.390	0.042	-0.348

namesubsection.11.2XYZ11.2 Surveillance × Censorship

nameresult.10XYZ

Result 10 (Surveillance × Censorship Nearly Eliminates Coordination)**.** *Under surveillance alone (infodesign framework), the baseline join rate falls from 12.4% to 0.9%. Upper censorship under surveillance: 3.7% (−27.2 pp vs. upper censorship alone). Lower censorship under surveillance: 4.2% (−34.8 pp vs. lower censorship alone).*

The surveillance-censorship interaction is superadditive: surveillance does not merely reduce join rates by a fixed amount but eliminates the coordination mechanism through which information design effects propagate.

namesection.12XYZ12 Conclusion

Nine architecturally distinct LLMs, given private signals as natural-language briefings, produce aggregate behavior that tracks the Bayesian Nash equilibrium of the attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.morris2003Morris and Shin (attr/Border[0 0 0]/H/I/C[0 1 0]goto namecite.morris20032003) regime change game (mean $r = +0.73$, $p < 0.001$). Scrambling and inverting signals confirm that this correlation is driven by briefing content, not prompt artifacts. Information design then shifts coordination in the directions theory predicts: censorship raises join rates through pooling, surveillance suppresses them through preference falsification, and the combination nearly eliminates coordination.

I do not claim that LLMs are Bayesian agents—the mechanism by which they process narrative text likely differs fundamentally from Bayesian updating. But the behavioral regularity is precisely what the global games framework predicts: monotone response to signal content, threshold-like decisions, and sensitivity to the information structure. This regularity is what makes the platform useful.

The full regime change game has resisted laboratory implementation because it requires conveying rich private signals, creating genuine strategic uncertainty, and coordinating large groups. LLM agents sidestep these practical constraints. They can be given arbitrary information structures, embedded in any network topology, and run at scale. The same platform can be applied to currency crises, bank runs, and other coordination games where information processing is central to behavior.

namesection*.30XYZ References

namecite.aher2023XYZ Aher, G., Arriaga, R. I., and Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. *Proceedings of the 40th International Conference on Machine Learning* (ICML).

namecite.akata2025XYZ 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.

namecite.angeletos2007aXYZ

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.

namecite.argyle2023XYZ

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., and Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

namecite.avoyan2020XYZ

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

namecite.bail2024XYZ

Bail, C. A. et al. (2024). Simulating opinion dynamics with networks of LLM-based agents. arXiv preprint arXiv:2311.09618.

namecite.bergemann2016XYZ

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

namecite.bergemann2019informationXYZ

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

namecite.blume2007XYZ

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.

namecite.brandts2006XYZ

Brandts, J. and Cooper, D. (2006). A change would do you good: An experimental study on how to overcome coordination failure in organizations. *American Economic Review*, 96(3):669–693.

namecite.carlini2025XYZ

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

namecite.carlsson1993XYZ

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

namecite.carter2021XYZ

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

namecite.chen2019XYZ

Chen, Y. and Yang, D. Y. (2019). The impact of media censorship: 1984 or Brave New World? *American Economic Review*, 109(6):2294–2332.

namecite.corrupted2025XYZ

Corrupted by Reasoning (2025). Reasoning language models become free-riders in public goods games. arXiv preprint arXiv:2506.23276.

namecite.crawford1982XYZ

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

namecite.diamond1983XYZ

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

namecite.edmond2013XYZ

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

namecite.ellingsen2010XYZ

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

namecite.enikolopov2020XYZ

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

namecite.farrell1996XYZ

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

namecite.frankel03XYZ

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.

namecite.gao2025XYZ

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

namecite.goldstein2016XYZ

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

namecite.grossmann2025XYZ

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

namecite.guo2023XYZ

Guo, F. (2023). GPT in game theory experiments. arXiv preprint arXiv:2305.05516.

namecite.guriev2019XYZ

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

namecite.helland2021XYZ

Helland, L., Holm, S., and Saethre, M. (2021). Information quality and regime change: Evidence from the lab. *Journal of Economic Behavior & Organization*, 191:538–554.

namecite.heinemann2004XYZ

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.

namecite.heinemann2009XYZ

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

namecite.horton2023XYZ

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

namecite.huang2024XYZ

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

namecite.inostroza2025XYZ

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

namecite.jost2018XYZ

Jost, J. T., Barberá, P., Bonneau, R., Langer, M., Metzger, M., Nagler, J., Sterling, J., and Tucker, J. A. (2018). How social media facilitates political protest: Information, motivation, and social networks. *Political Psychology*, 39(S1):85–118.

namecite.kamenica2011XYZ

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

namecite.king2013XYZ

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.

namecite.king2017XYZ

King, G., Pan, J., and Roberts, M. E. (2017). How the Chinese government fabricates social media posts for strategic distraction, not engaged argument. *American Political Science Review*, 111(3):484–501.

namecite.kolotilin2022XYZ

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

namecite.kuran1991XYZ

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

namecite.li2022XYZ

Li, F., Song, Y., and Zhao, M. (2022). Global manipulation by local obfuscation. *Journal of Economic Theory*, 207.

namecite.lorentzen2014XYZ

Lorentzen, P. (2014). China's strategic censorship. *American Journal of Political Science*, 58(2):402–414.

namecite.lore2024XYZ
Lorè, N. and Heydari, B. (2024). Strategic behavior of large language models and the role of game structure versus contextual framing. *Scientific Reports*, 14:18490.

namecite.mathevet2020XYZ
Mathevet, L., Perego, J., and Taneva, I. (2020). On information design in games. *Journal of Political Economy*, 128(4):1370–1404.

namecite.morris1998XYZ
Morris, S. and Shin, H. S. (1998). Unique equilibrium in a model of self-fulfilling currency attacks. *American Economic Review*, 88(3):587–597.

namecite.morris2002XYZ
Morris, S. and Shin, H. S. (2002). Social value of public information. *American Economic Review*, 92(5):1521–1534.

namecite.morris2003XYZ
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.

namecite.obstfeld1996XYZ
Obstfeld, M. (1996). Models of currency crises with self-fulfilling features. *European Economic Review*, 40(3-5):1037–1047.

namecite.palatsi2025XYZ
Palatsi, A. C., Martin-Gutierrez, S., Cardenal, A. S., and Pellert, M. (2025). Large language models replicate and predict human cooperation across experiments in game theory. arXiv preprint arXiv:2511.04500.

namecite.park2023XYZ
Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (UIST).

namecite.penney2016XYZ
Penney, J. W. (2016). Chilling effects: Online surveillance and Wikipedia use. *Berkeley Technology Law Journal*, 31(1):117–182.

namecite.petrov2025XYZ
Petrov, A. et al. (2025). LLM strategic reasoning: Agentic study through behavioral game theory. arXiv preprint arXiv:2502.20432.

namecite.piatti2024XYZ
Piatti, G. et al. (2024). Cooperate or collapse: Emergence of sustainable cooperation in a society of LLM agents. *Advances in Neural Information Processing Systems* (NeurIPS).

namecite.shadmehr2015XYZ
Shadmehr, M. and Bernhardt, D. (2015). State censorship. *American Economic Journal: Microeconomics*, 7(2):280–307.

namecite.shurchkov2013XYZ
Shurchkov, O. (2013). Coordination and learning in dynamic global games: Experimental evidence. *Experimental Economics*, 16(2):313–334.

namecite.stoycheff2016XYZ
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.

namecite.sun2025surveyXYZ
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).

namecite.szkup2020XYZ
Szkup, M. and Trevino, I. (2020). Sentiments, strategic uncertainty, and information structures in coordination games. *Games and Economic Behavior*, 124:534–553.

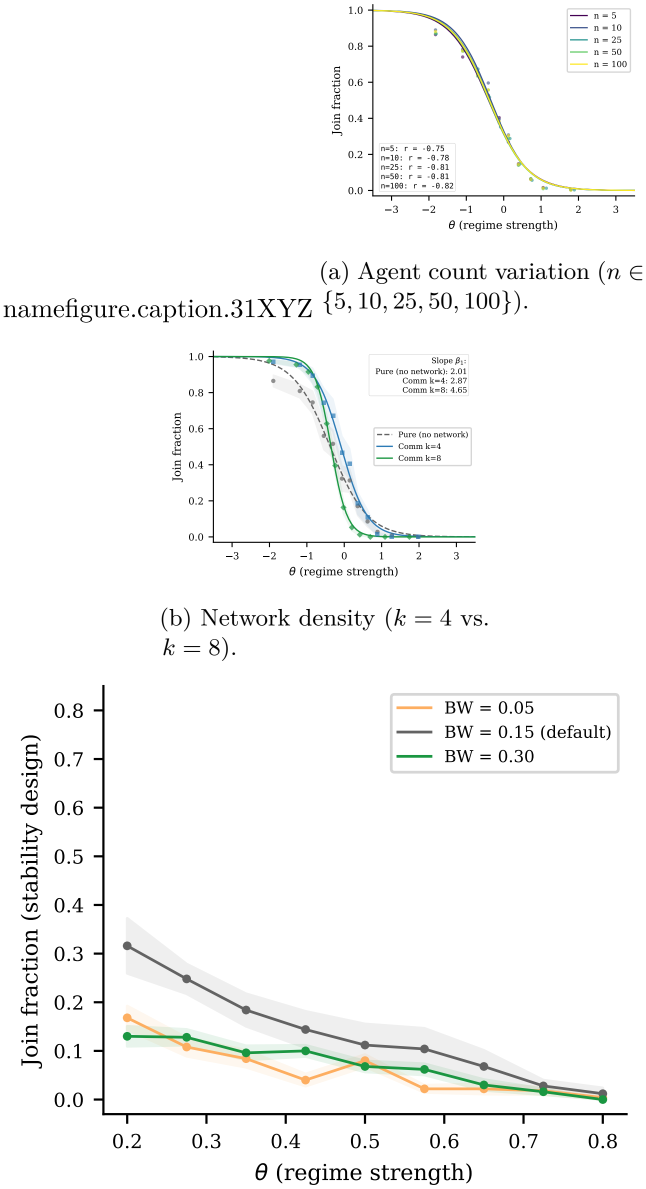
namepage.19XYZ

nameappendix.AXYZA Robustness

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

namesection*.32XYZ
Group-size awareness. In the main experiments, agents are told "You do not know how many others will JOIN" but are not told the group size, leaving them no basis for reasoning about coordination thresholds. As a robustness check, I run the pure and communication treatments with modified prompts that state "You are one of 25 citizens deciding whether to JOIN an uprising or STAY home." Over 100 country–periods per treatment, the pure join rate is 0.507 (vs. 0.369 baseline) and the communication join rate is 0.473 (vs. 0.452 baseline). Monotone response to signals is preserved in both treatments. The communication premium, however, reverses: with group-size knowledge, communication *lowers* join rates by 3.4 pp rather than raising them. One interpretation is that when agents know the group size, messages revealing others' reluctance become more informative about the probability of reaching critical mass, amplifying the deterrent effect of cautious peers. The level shift in the pure treatment suggests that group-size knowledge increases baseline willingness to coordinate, but the core finding—monotone signal response—is robust.

Online appendix. Additional supplemental material is in `online_appendix.tex`.



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

namefigure.caption.31XYZ

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

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

Figure A1: Robustness checks for equilibrium alignment and treatment effects.