

# Online Appendix: LLMs Can Play (Global) Games

Khaled Eltokhy

## 1 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.

The direction channel produces the largest average treatment effect at +2.4 pp, with effects concentrated near  $\theta^*$ . The dissent channel contributes +1.6 pp uniformly. The clarity channel produces only +0.3 pp with a non-monotone pattern.

The sum of single-channel effects is +4.3 pp, far smaller than the full stability design effect (+19.5 pp). This implies strong complementarities: combining ambiguity (clarity), softened direction, and mixed-valence cues shifts behavior much more than the sum of each channel in isolation.

## 2 Robustness Details

### 2.1 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.

### 2.2 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.

### 2.3 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.

### 2.4 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.

### 2.5 Cross-Model Replication of Information Design

Table 3 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 Minstral 3B retain baseline-level correlations under scramble ( $r = -0.81$  and  $r = -0.66$ ), suggesting these models extract signal from features

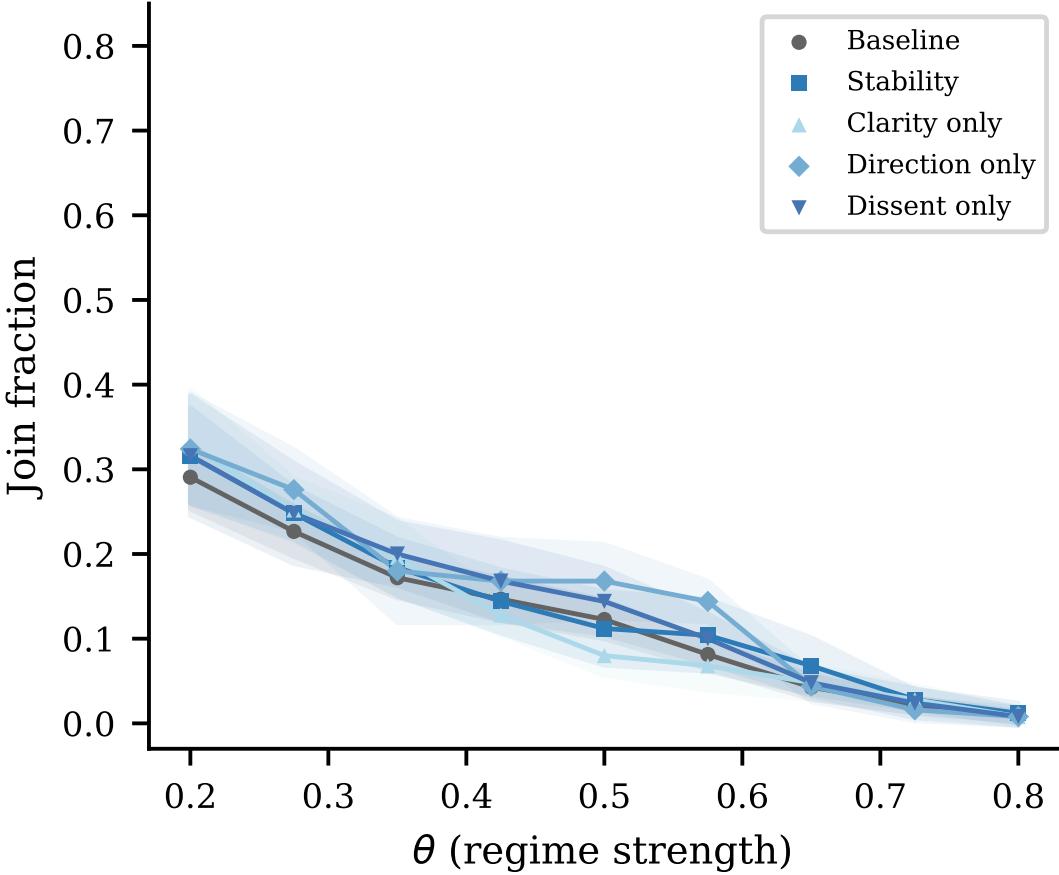


Figure 1: Single-channel decomposition of the stability design. Each panel shows the treatment effect  $\Delta(\theta)$  for one channel in isolation.

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.

## 2.6 Information Design with Communication

Under baseline information design with communication, mean join rate is 3.0% (−9.3 pp vs. baseline without communication). Communication *reduces* join rates across all information design conditions—superficially counterintuitive, but resolved by noting that additional information from others’ messages tends to reveal that the regime is strong (the  $\theta$  grid extends upward from  $\theta^*$ ).

## 3 Implementation Details

### 3.1 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.

Table 1: Single-channel decomposition of the stability design (primary model: Mistral Small Creative).

Channel	Mean	$r$	$\Delta$
Full stability	0.319	-0.626	+0.195
Clarity only	0.126	-0.857	+0.003
Direction only	0.148	-0.826	+0.024
Dissent only	0.140	-0.837	+0.016
Sum of channels	—	—	+0.043
Full design	—	—	+0.195

Table 2: Bandwidth robustness: mean join rates (primary model: Mistral Small Creative).

Design	BW=0.05	BW=0.15	BW=0.30
Baseline	0.054	0.124	0.061
Stability	0.061	0.319	0.070
Upper cens.	0.116	0.309	0.114
Lower cens.	0.155	0.390	0.157

### 3.2 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.

### 3.3 Randomization

Each country-period draws  $\theta$  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 `Generator` with no fixed seed; exact sequences are logged in per-run JSON files included in the replication archive.

### 3.4 Code and Data Availability

All code, prompts, cached LLM responses, and output data will be available at [repository URL] upon publication.

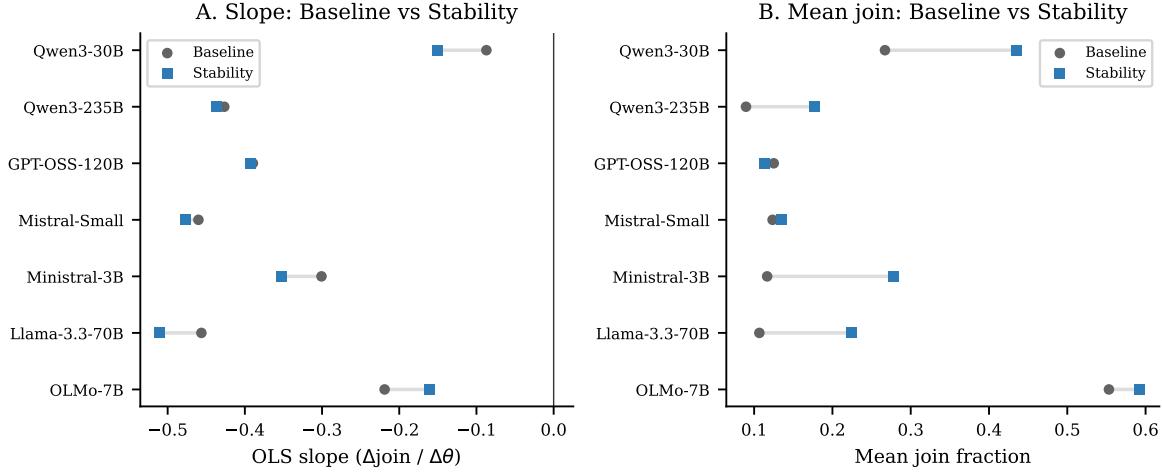


Figure 2: Cross-model replication of information design treatments. Each panel shows join fraction vs.  $\theta$  for one model under baseline, stability, scramble, and flip conditions.

Table 3: Cross-model replication of key information design conditions.  $r$  is the correlation between  $\theta$  and join fraction.

Model	Baseline		Scramble		Flip	
	Mean	$r$	Mean	$r$	Mean	$r$
Mistral Small Creative	0.124	-0.812	0.121	+0.036	0.663	+0.823
GPT-OSS 120B	0.127	-0.801	0.132	+0.080	0.677	+0.754
Llama 3.3 70B	0.107	-0.809	0.105	-0.810	0.887	+0.717
Minstral 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	0.090	-0.776	0.094	+0.056	—	—
OLMo 3 7B	0.718	-0.329	0.592	-0.294	0.839	+0.452