# 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 Construct Validity: Latent Feature Baseline

A natural concern is that the behavioral sigmoid reflects nothing more than sentiment classification: models read the briefing's tone and respond accordingly, without any strategic reasoning. I test this by fitting a logistic classifier on the three latent features that generate each briefing—direction ($d$), clarity, and coordination—and comparing its predictions to actual LLM decisions.

Across nine models, the 3-feature classifier achieves mean accuracy of 82.6%, nearly identical to the 1-feature (direction-only) baseline at 82.0%. Direction dominates; clarity and coordination contribute minimally on average (though Llama 3.3 70B and Qwen3 235B show meaningful gains of +1.5 and +3.8 pp respectively, suggesting they attend to coordination cues).

The critical test is cross-treatment departure. I train the classifier on pure treatment data and predict decisions under communication and surveillance. If LLMs merely classify sentiment, treatment should not shift residuals. Under communication, the mean residual is +1.5 pp—agents join slightly more than text features predict, consistent with coordination amplification. Under surveillance, the mean residual is −15.5 pp: agents join substantially *less* than the text-only classifier predicts. The latent features are identical across treatments; the chilling effect cannot be explained by sentiment classification. LLMs respond to the strategic context—being watched—not merely to the briefing's tone.

## 3 Robustness Details

### 3.1 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.

### 3.2 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

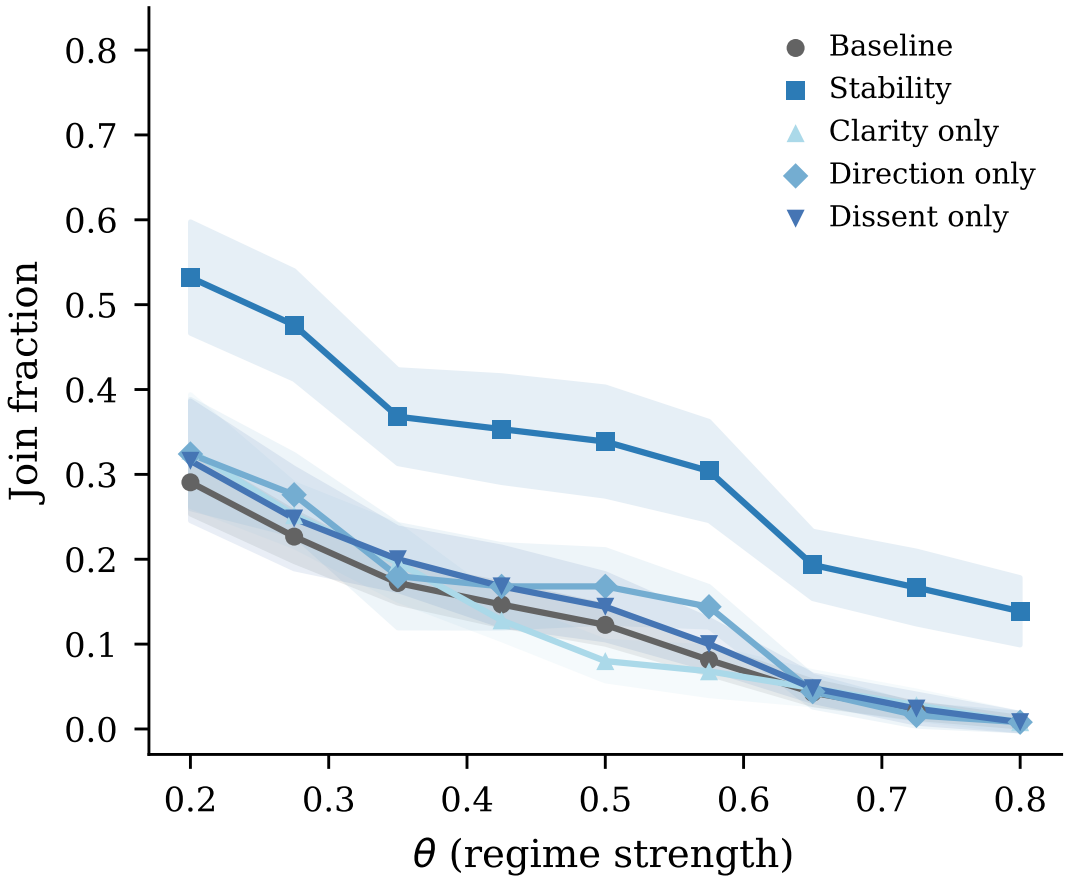


Figure 1: Single-channel decomposition of the stability design. Each panel shows the treatment effect $\Delta(\theta)$ for one channel in isolation.

contacts do not substantially amplify coordination.

### 3.3 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.

### 3.4 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.

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

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 |

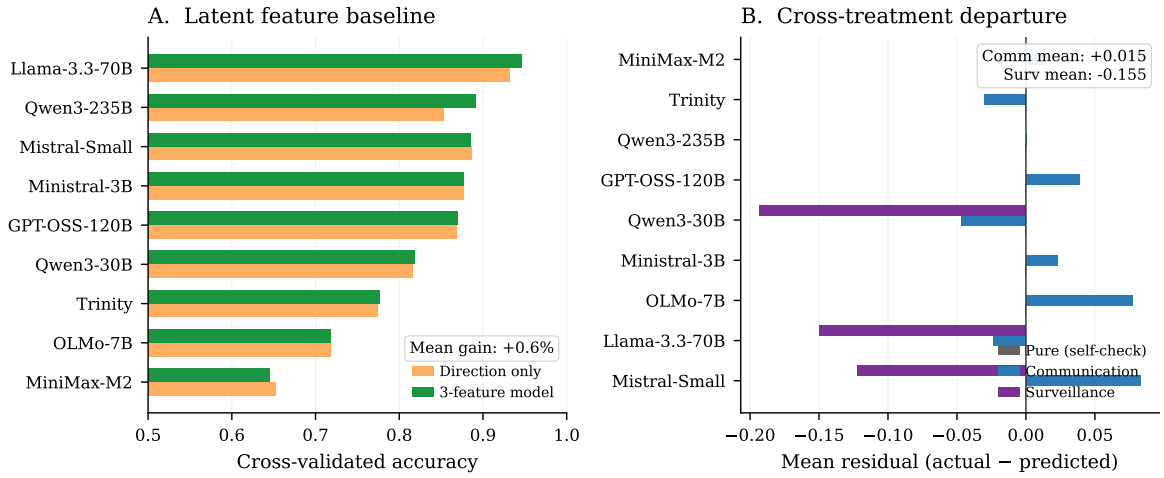


Figure 2: Construct validity test. *Left:* 5-fold cross-validated accuracy of latent feature classifiers (1-feature vs. 3-feature) by model. *Right:* Mean residual (actual − predicted) by treatment, where the classifier is trained on pure data only. The surveillance residual of −15.5 pp indicates LLMs suppress joining beyond what text features predict.

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.

## 3.6 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 $\theta^*$).

## 3.7 Linguistic Analysis of Surveillance Messages

Table 4 reports per-keyword frequencies across communication, surveillance, and propaganda treatments. Surveillance agents drop action language and substitute hedging: "act" falls from 19.0% to 3.8%, "collapse" from 11.6% to 0.6%, while "careful" rises from 1.5% to 4.8% and "stable" from 1.7% to 4.2%. The action-signaling rate (fraction of messages with more action than caution words) falls from 49% to 39% under surveillance.

## 3.8 Calibration Robustness Across Models

Table 5 reports signal alignment metrics across all nine models. The raw correlation $r_\theta$ (between regime strength $\theta$ and join fraction, with no dependence on the theoretical attack mass formula) is strongly negative

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 |

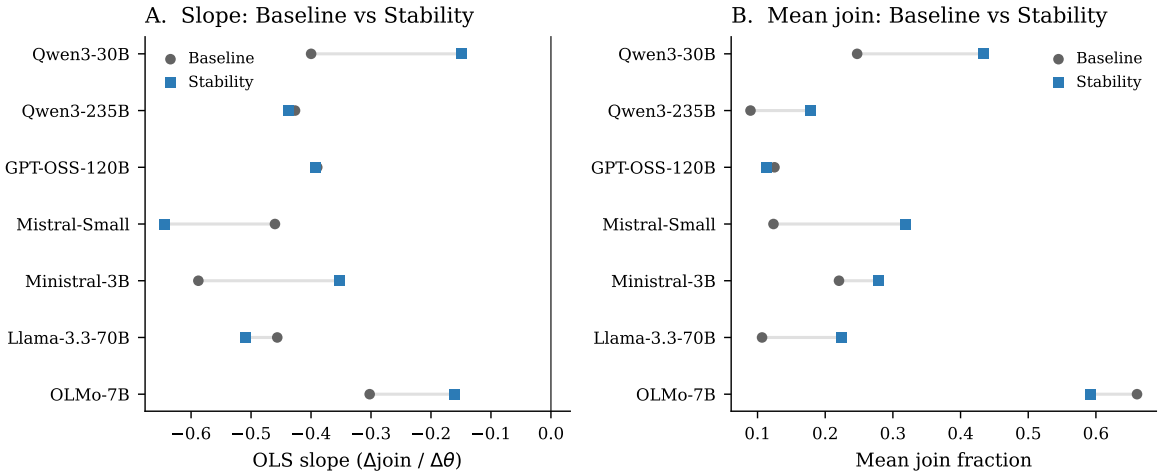


Figure 3: Cross-model replication of information design treatments. Each panel shows join fraction vs. $\theta$ for one model under baseline, stability, scramble, and flip conditions.

for all models ($-0.70$ to $-0.87$), confirming that the behavioral pattern is not an artifact of the calibration procedure. The text baseline logistic slope varies across models (from $-2.0$ to $-25.1$), reflecting differences in how sensitively each model responds to the direction feature of the briefing text.

Figure **??** shows the calibration convergence process. Panel (a) plots the fitted logistic center across calibration rounds for each model; all models converge within 4–6 rounds to a center near zero. Panel (b) reports the final calibrated `cutoff_center` parameter by model.

# 4 Implementation Details

## 4.1 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.

## 4.2 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.**

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

| | Baseline | | Scramble | | Flip | |
|---|---|---|---|---|---|---|
| Model | 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 |
| 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 | 0.090 | −0.776 | 0.094 | +0.056 | — | — |
| OLMo 3 7B | 0.718 | −0.329 | 0.592 | −0.294 | 0.839 | +0.452 |

Table 4: Message content by treatment (primary model: Mistral Small Creative). Keyword columns show frequency (% of words). Action-signal: fraction of messages with more action than caution words. Length: mean characters.

| Treatment | $N$ | act | fight | ready | moment | patience | loyal | stable | strong | cautious | risk | Act-sig | Len |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Comm | 15000 | 0.0 | 0.0 | 0.3 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.49 | 342 |
| Surveillance | 25000 | 0.0 | 0.0 | 0.3 | 0.1 | 0.1 | 0.0 | 0.1 | 0.0 | 0.0 | 0.0 | 0.39 | 252 |
| Prop $k$=5 | 10000 | 0.0 | 0.0 | 0.3 | 0.1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.49 | 341 |

```
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.
```

## 4.3 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.

## 4.4 Code and Data Availability

All code, prompts, cached LLM responses, and output data are available at https://github.com/keltokhy/llm-global-games.

Table 5: Calibration robustness. $r_\theta$: raw correlation with regime strength. $r_A$: correlation with theoretical attack mass. RMSE: root mean squared error vs. $A(\theta)$. Text slope: logistic slope of naïve 1 − direction predictor.

| Model | $r_\theta$ | $r_A$ | RMSE | Text slope |
|---|---|---|---|---|
| Mistral Small Creative | −0.81 | +0.67 | 0.354 | -9.2 |
| Llama 3.3 70B | −0.85 | +0.79 | 0.288 | -25.1 |
| OLMo 3 7B | −0.70 | +0.65 | 0.450 | -2.0 |
| Ministral 3B | −0.87 | +0.79 | 0.281 | -10.4 |
| Qwen3 30B | −0.84 | +0.78 | 0.287 | -7.5 |
| GPT-OSS 120B | −0.84 | +0.70 | 0.359 | -8.2 |
| Qwen3 235B | −0.85 | +0.70 | 0.354 | -20.5 |
| Trinity Large | −0.87 | +0.84 | 0.262 | -4.0 |
| MiniMax M2-Her | −0.79 | +0.66 | 0.360 | -2.1 |

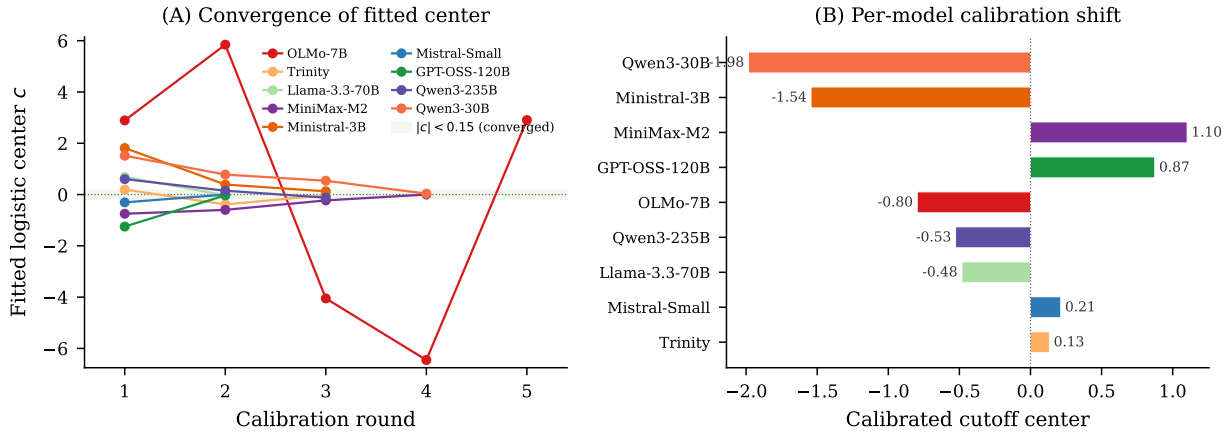


Figure 4: Calibration convergence. *(a)* Fitted logistic center across calibration rounds. *(b)* Final calibrated `cutoff_center` parameter by model.