Diplomatic Treaty Negotiation: Ambitious Model-Scale Sweep

9 Model Pairs $\times$ 25 Parameter Conditions $\times$ 3 Runs

Joie Zhang

February 24, 2026

Experiment Design

Fixed baseline: GPT-5-nano (Elo 1338)

8 adversary models spanning Elo 1105–1490:

Model	Provider	Elo	Category
GPT-3.5-Turbo	OpenAI	1105	Weak
GPT-5-nano (self-play)	OpenAI	1338	Baseline
GPT-4o	OpenAI	1346	Medium
O3-mini-high	OpenAI	1364	Medium
Claude Haiku 4.5	Anthropic	1403	Medium
GPT-5.2-high	OpenAI	1436	Strong
Claude Sonnet 4.5	Anthropic	1450	Strong
Gemini 3 Pro	Google	1490	Strong

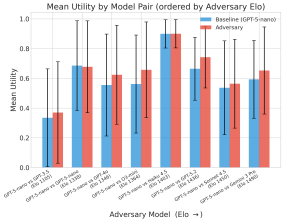
Parameters: $\rho \in \{-1, -0.5, 0, 0.5, 1\}$, $\theta \in \{0, 0.25, 0.5, 0.75, 1\}$
Total: 1,350 configs $\rightarrow$ **1,057 successful** (293 failures from API issues)

Headline Results

Overall (n=1,057):

- 89% consensus rate
- Mean social welfare: 1.178 / 2.0
- Mean rounds: 4.5 / 10
- 91% exploitation rate

Key insight: Unlike co-funding, diplomacy shows **no exploitation gap** between strong and weak models.



Model Pair Performance

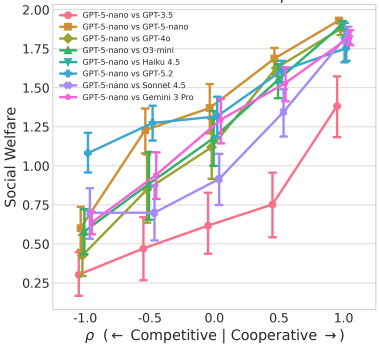
Adversary (Elo)	n	Consensus	SW	Rounds	Util Gap
GPT-3.5 (1105)	150	62%	0.705	7.2	−0.017
GPT-5-nano (1338)	150	97%	1.363	3.6	−0.016
GPT-4o (1346)	150	89%	1.199	4.3	−0.003
O3-mini (1364)	150	93%	1.225	3.9	+0.004
green!10 Haiku 4.5 (1403)	7	100%	1.799	1.4	−0.041
GPT-5.2 (1436)	150	100%	1.406	2.5	−0.006
Sonnet 4.5 (1450)	150	91%	1.124	4.7	−0.001
Gemini 3 Pro (1490)	150	91%	1.203	4.8	−0.014

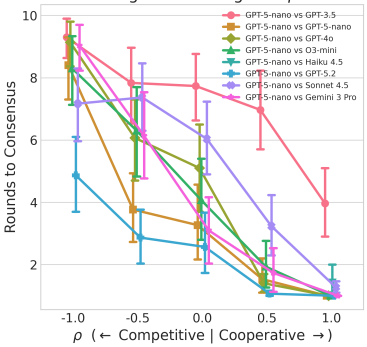
Utility Gap = strong model utility − weak model utility.

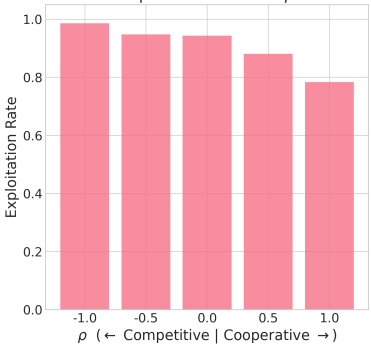
All gaps are $< 0.05 \Rightarrow$ **no systematic exploitation in diplomacy**.

GPT-3.5 is the outlier: low consensus, low SW, dragging both players down.

ρ (Preference Correlation) — The Dominant Driver

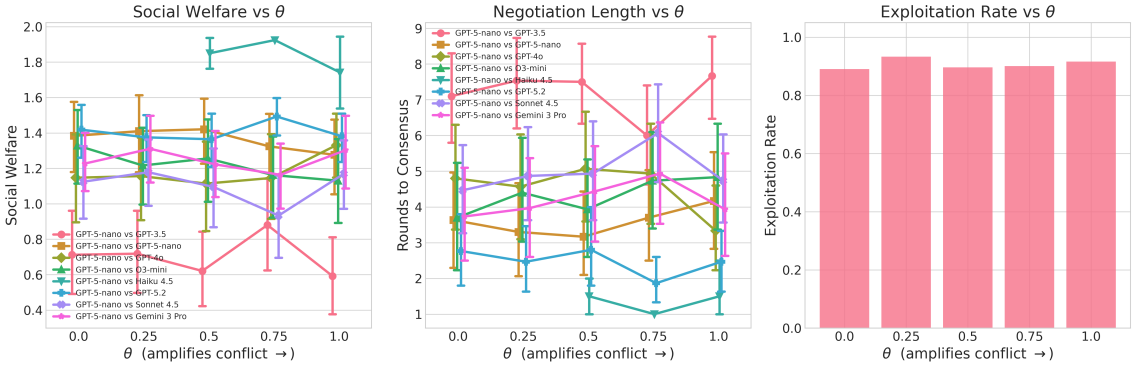






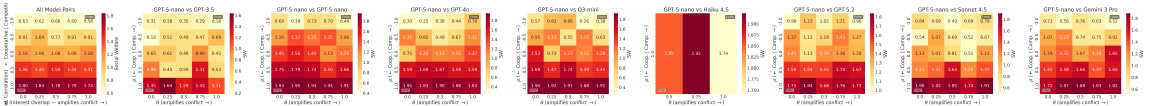
- ρ drives a **3× range** in social welfare (0.617 → 1.788)
- Consensus: 78% at ρ=−1 → 99% at ρ=1
- Rounds: 8.0 at ρ=−1 → 1.5 at ρ=1
- **All model pairs follow the same monotonic trend**

θ (Interest Overlap) — Surprisingly Flat



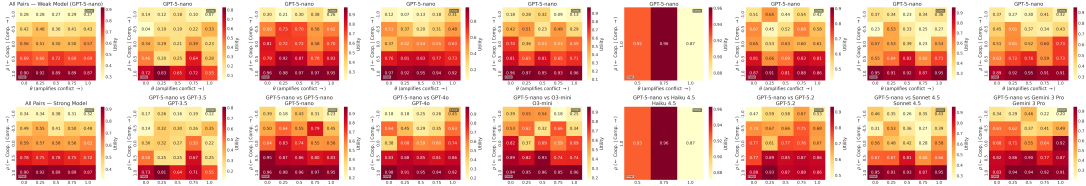
- θ has **almost no effect** on social welfare (range: 1.16–1.20)
- Rounds and consensus are also essentially flat across θ
- Exploitation rate is constant ($\sim$90%) regardless of θ
- **Implication**: Models respond to preference correlation, not overlap structure

$\rho \times \theta$ Heatmaps: Social Welfare by Model Pair



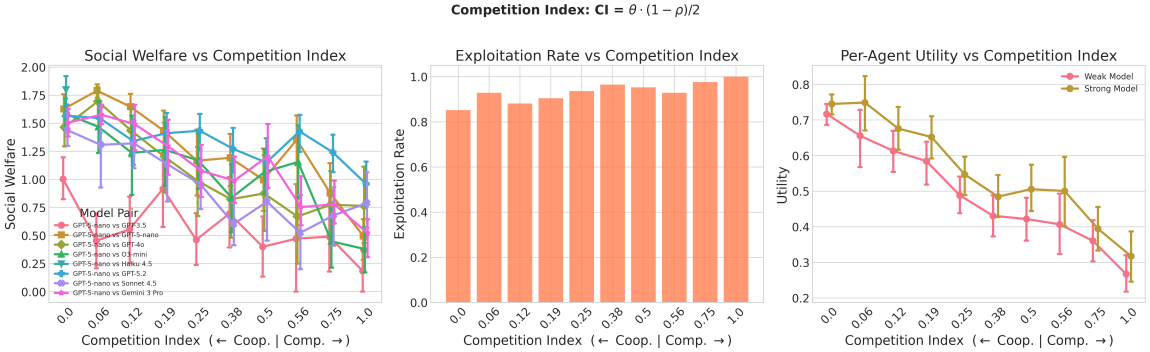
Row-dominant gradient (ρ) confirms it drives outcomes.
Column variation (θ) is negligible within each row.

Per-Agent Utility: Weak Model vs Strong Model



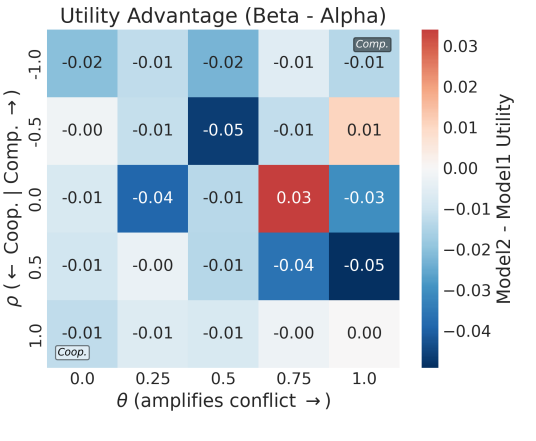
Top row: Weak model (GPT-5-nano as Alpha). Bottom row: Strong model (adversary as Alpha).
Both agents' utilities track together — no exploitation pattern visible.

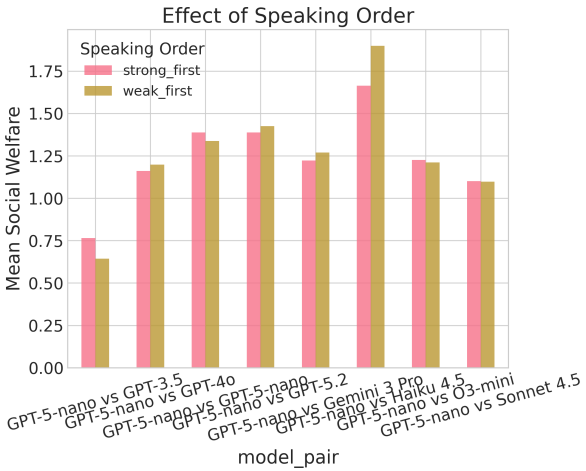
Competition Index: $\mathrm{CI} = \theta \cdot (1 - \rho)/2$



- **Clean monotonic decline** in SW as competition rises (1.46 → 0.58)
- Exploitation rate rises from 85% to 100% at max competition
- Per-agent utility: **both agents decline together** — gap stays <0.05

Exploitation Gap & Speaking Order





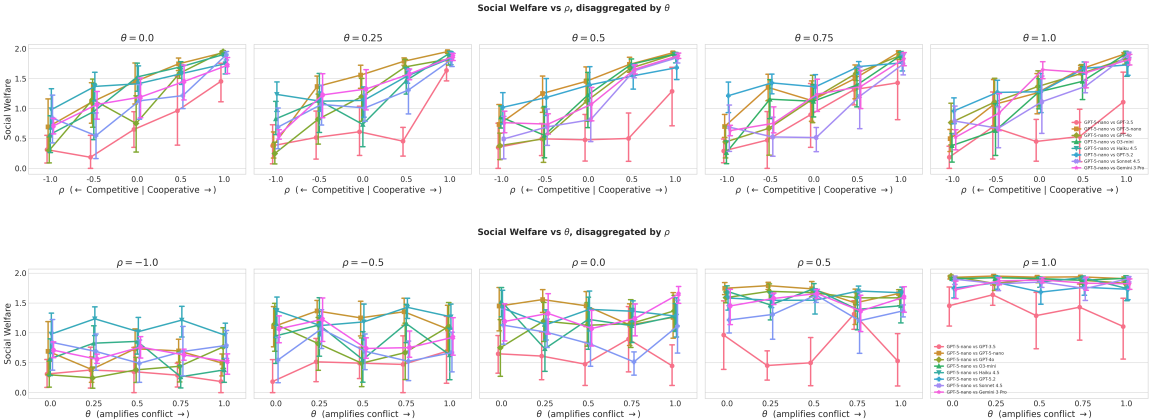
Exploitation gap (left heatmap):

- Range: -0.05 to $+0.03$

Speaking order (right bar chart):

- strong_first $\approx$ weak_first

Social Welfare Disaggregated by θ and ρ



Top: SW vs ρ, faceted by θ — same upward trend regardless of θ.
Bottom: SW vs θ, faceted by ρ — flat within each ρ level.

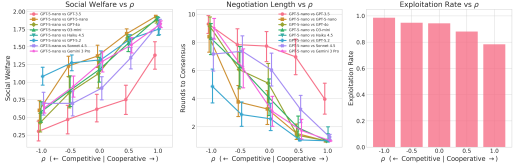
GPT-3.5-Turbo: The Weak Model Outlier

GPT-3.5-Turbo (Elo 1105) is **dramatically worse** than all other models:

Metric	GPT-3.5	All Others
Consensus	62%	94%
Social Welfare	0.705	1.263
Rounds	7.2	3.8

Key observations:

- Poor instruction-following → can't parse proposals
- Hurts **both** players (GPT-5-nano also gets low utility)
- This is a **capability floor**, not exploitation
- Suggests Elo ∼1200 is the minimum for



GPT-3.5 (pink) consistently sits below all other model pairs across ρ.

Key Findings

1. ρ **dominates**: Preference correlation is the primary driver of outcomes (3$\times$ SW range). θ has negligible effect.
2. **No exploitation gap**: Unlike co-funding, stronger models do NOT extract more value. Utility gaps are $<$0.05 across all conditions.
3. **Capability floor matters**: GPT-3.5 (Elo 1105) breaks negotiation for **both** parties. Below Elo $\sim$1200, models can't follow the protocol.
4. **No order effects**: Speaking first provides no advantage. Strong and weak agents get symmetric outcomes.
5. **High consensus rate**: 89% overall, rising to 99% at ρ=1. Diplomacy's propose-and-vote protocol is robust.

Diplomacy vs Co-Funding: Contrasting Exploitation Patterns

Property	Diplomacy	Co-Funding
Protocol	Propose-and-vote	Talk-pledge-revise
Exploitation gap	< 0.05 (none)	Significant (free-riding)
Dominant parameter	ρ (correlation)	α (alignment)
Secondary parameter	θ (no effect)	σ (scarcity)
Consensus rate	89%	N/A (no voting)
Coordination failure	Low	77.5%
Speaking order effect	None	N/A

Hypothesis: The propose-and-vote protocol **protects** the weak agent — both parties must agree, so exploitation is structurally prevented.
In co-funding, each agent independently pledges funds, creating an opening for free-riding that propose-and-vote eliminates.

Next Steps

- **Re-run 293 failed experiments** (amazon-nova-micro code bug fixed; claude-haiku-4-5 API overload was transient)
- **Re-run 692 failed co-funding experiments** once OpenAI API key is restored
- Generate combined cross-game analysis with full data
- Formal statistical tests: regression of SW on ρ, θ, Elo, and interactions
- Investigate **why** diplomacy prevents exploitation while co-funding doesn't