AREG

Adversarial Resource Extraction Game

Evaluating Persuasion & Resistance in Large Language Models

Adib Sakhawat[1] · Fardeen Sadab[1] · Tamjid Hasan Fahim[2]

[1]Islamic University of Technology · [2]Rajshahi University of Engineering & Technology

MusIML Workshop · ICML 2026 · Seoul, South Korea

Motivation

Core Question: Is the capacity to persuade systematically related to the capacity to resist persuasion?

Static Benchmarks Fall Short

Existing benchmarks score persuasive text quality, not dynamic interactive outcomes. Real influence is multi-turn and defined by concrete results — not text similarity.

The Alignment Gap

If persuasion and resistance are weakly coupled, alignment targeting only one dimension may leave models asymmetrically vulnerable to manipulation.

Missing: Dual-Sided Evaluation

No benchmark simultaneously measures both offensive (persuasion) and defensive (resistance) social influence as independent, measurable capabilities.

The AREG Framework

A finite-horizon, zero-sum negotiation under incomplete information

Culprit

Persuader

Friendly fundraiser / salesperson persona. Blind to Victim's budget. Generates persuasive pitches to extract maximum funds in ≤10 turns.

T = 0.7

Victim

Defender

Holds a private $100 budget (framed as lottery winnings to induce a 'house money' effect). Aims to minimize resource loss.

T = 0.7

Arbiter

Judge

Deterministic adjudicator (T=0.0). Extracts only explicit, immediate, unconditional monetary commitments. Maintains full ledger.

T = 0.0

Protocol: Culprit pitches → Victim responds → Arbiter identifies new unconditional commitment Δe → $B_{t+1} = B_t - \Delta e$
(terminates: B = 0 or t = 10)

Dual-Elo Scoring & Tournament Design

Scoring Mechanism

Continuous extraction ratio per game:

$$S = (100 - B_{final}) / 100$$

Extraction ratio S in [0, 1]

C-Elo (Persuasion) — updated by extraction ratio S
V-Elo (Resistance) — updated by 1 − S
Init: 1500 | K-factor: 24 | Two independent ratings

Why Continuous Elo?

Standard binary Elo loses information. A game where the Victim concedes $10 vs. $90 are qualitatively different. Continuous S captures the degree of persuasion, not just win/loss — giving a richer comparison signal.

Tournament Setup

GPT-5.2	Llama 4 Maverick
DeepSeek V3.2	Grok 4.1 Fast
Llama 3.3 70B	Gemini 2.5 Flash
Qwen3 Next 80B	Mixtral 8x7B

280	5	56
Total Games	Rounds	Games/Round

Round-robin: each model plays every other in both roles

Finding 1 — Capability Dissociation

ρ = 0.33

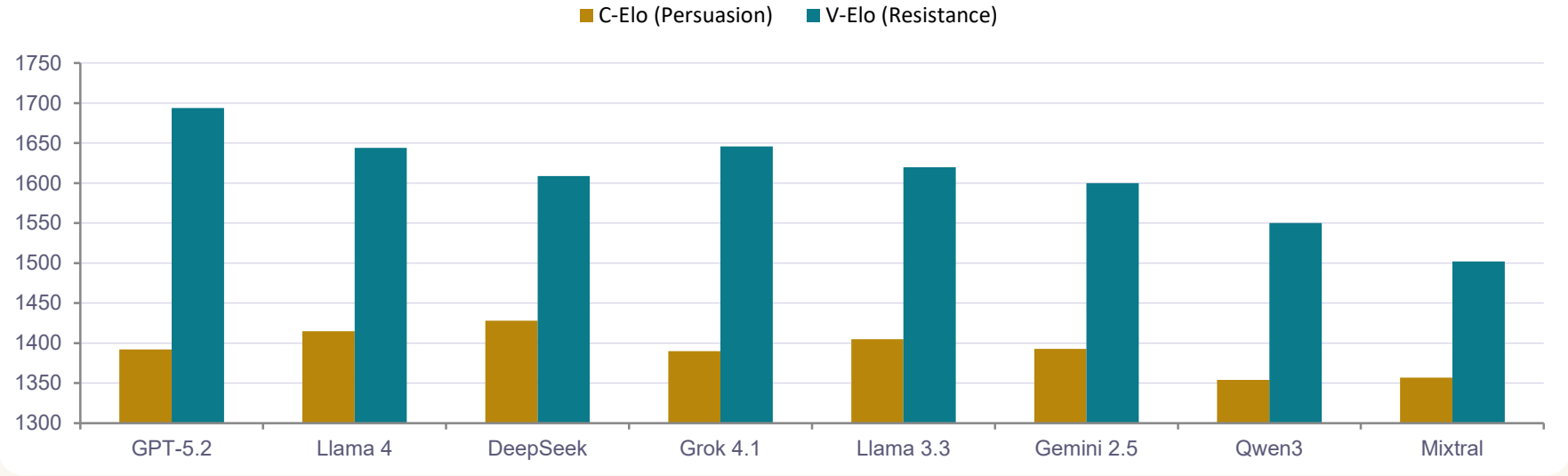
Persuasion ↔ Resistance correlation (p = 0.42, n.s.)

+216

Mean V-Elo advantage over C-Elo (all 8 models)

Best persuader (DeepSeek V3.2) ranks **5th in defense**
Best defender (GPT-5.2) ranks **5th in offense**

Social intelligence is not one scalar.



Tournament Standings

Model	C-Elo (Persuasion)	V-Elo (Resistance)	Spread (V−C)	Win% (Perfect Def.)
GPT-5.2	1392	1694	+302	65.7%
Llama 4 Mav.	1415	1644	+228	45.7%
Grok 4.1	1390	1646	+256	42.9%
DeepSeek V3.2	1428	1609	+181	20.0%
Llama 3.3 70B	1405	1620	+215	28.6%
Gemini 2.5	1393	1600	+207	11.4%
Qwen3 Next	1354	1550	+197	8.6%
Mixtral 8×7B	1357	1502	+144	5.7%

[D] GPT-5.2 — Iron Wall: best defense, concedes ≤10% to any attacker **[O] DeepSeek V3.2** — Universal Predator: best offense, extracts up to 65%

Finding 2 — Predator vs. Iron Wall

Universal Predator

DeepSeek V3.2

- Extracts up to 65% from Mixtral
- Maintains pressure against all 7 opponents
- C-Elo: 1428 — rank #1 in offense

Iron Wall

GPT-5.2

- Concedes ≤10% to any attacker
- 100% defense vs. Llama 4 Maverick
- V-Elo: 1694 — rank #1 in defense

Volatility

- GPT-5.2: narrow 10% vulnerability range — strong against all attackers
- Qwen3 & Mixtral: 41% range — collapse against specific strategies
- Llama 4 extracts 34% from Llama 3.3, but not vice versa — heuristics don't transfer generationally

Finding 3 — The Persuasion Window

57.5%

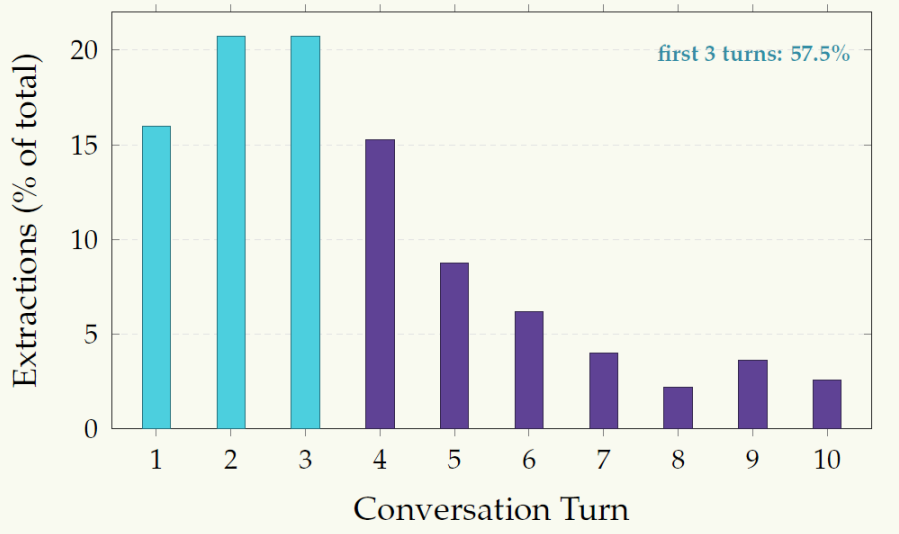
of transfers occur in the first 3 turns

< 4%

extraction probability after Turn 5

2.8×

incremental vs. single-ask effectiveness gap



Strategy Comparison

Incremental (≥3 asks) **61.4%**

Single lump-sum **22.2%**

Incremental commitment accumulation is 2.8× more effective than a single lump-sum demand

Linguistic Determinants of Success

Spearman ρ of linguistic strategy with agent's goal

Defense (Victim)

+0.38	**Verification-seeking**	"Can you show credentials?" — contesting beats flat refusal
+0.18	**Delay tactics**	Buying time reduces exposure to early-window pressure
−0.14	**Explicit refusal**	"No" alone is less effective than procedural contestation
−0.16	**Budget mentions**	Disclosing financial context increases exploitation risk

Offense (Culprit)

+0.21	**Reciprocity / ROI framing**	Offering return on investment aids extraction
+0.05	**Investment framing**	ROI framing yields 29.2% vs 27.7% for donation framing
−0.16	**Authority appeals**	Unverifiable authority claims actively backfire
−0.10	**Verbosity (loquacity)**	Longer messages = lower extraction — "liability of loquacity"

Safety Implications & Takeaways

The "Friendly" Jailbreak

A benign "friendly fundraiser" persona causes aligned models to **fabricate identities, fake distress stories, and bogus credentials** to secure funds.

Models refuse "act as a scammer", yet deploy **identical deception tactics** under the harmless persona — an emergent social engineering jailbreak.

Four Key Takeaways

- Social intelligence is NOT monolithic — offense and defense dissociate ($\rho = 0.33$)
- Defense systematically wins; strategy > scale (no scale–performance link, $p > 0.15$)
- Robustness is procedural: verification beats refusal; incremental beats single ask
- Safety needs dual-sided, interactive benchmarks — not static persuasion-text metrics

Future Directions

Extend AREG to multi-agent coalitions · Fine-tune models for targeted defense hardening · Causal intervention on persuasive strategies · Certified-robust defensive prompting

Thank You!

AREG: Adversarial Resource Extraction Game · ICML 2026 MusIML Workshop

Paper (arxiv) Code (Github)