AREG: Adversarial Resource Extraction Game for Evaluating Persuasion and Resistance in Large Language Models

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

Abstract

Evaluating LLM social intelligence requires moving beyond static text toward dynamic interactions. We introduce the Adversarial Resource Extraction Game (AREG), a benchmark operationalizing persuasion and resistance as a multi-turn, zero-sum financial negotiation. A tournament across frontier models reveals that offensive and defensive capabilities are empirically dissociated and weakly correlated ($\rho = 0.33$). While models show a systematic defensive advantage, effectiveness depends heavily on dialogue structure: incremental persuasion outperforms single asks, and verification-seeking defends better than explicit refusal. These findings demonstrate that social influence is not a monolithic capability, highlighting the need for dual-sided evaluation to uncover asymmetric behavioral vulnerabilities.

1. Introduction

As Large Language Models (LLMs) evolve into interactive agents, their capacity for social influence has become a central safety concern (Wang et al., 2024). While existing benchmarks effectively evaluate static persuasive text generation, assessing interactive social intelligence requires dynamic, multi-turn evaluation where success is defined by concrete outcomes rather than textual quality alone (Singh et al., 2025). Furthermore, a foundational question remains unresolved: *Is the capacity to persuade systematically related to the capacity to resist persuasion?* If these capabilities are weakly coupled, alignment procedures targeting only one may leave models asymmetrically vulnerable to manipulation.

To address this, we introduce the **Adversarial Resource Extraction Game (AREG)**, a benchmark formalizing social influence as a multi-turn, zero-sum negotiation. AREG simulates an asymmetric dialogue between a **Culprit** (persuader) and a **Victim** (resource holder) aiming to retain a \$100 endowment. Outcomes are verified by a deterministic **Arbiter** agent (Yu, 2025), isolating model behavior under controlled conditions. We specifically utilize Grok as the Arbiter because recent multidimensional alignment audits reveal it exhibits the most unbiased, central positioning on the economic axis ($\mu_E = -0.444$) among contemporary frontier models (Sakhawat et al., 2026). This empirical economic neutrality makes it uniquely suited to impartially adjudicate adversarial financial transfers.

Our primary contributions are:

1. **A novel adversarial benchmark** operationalizing persuasion and resistance through objective financial transfer rather than subjective stance change.

2. **Empirical evidence of capability dissociation:** A tournament across eight frontier models reveals that persuasion and resistance are weakly correlated ($\rho = 0.33$), alongside a systematic defensive advantage.

3. **Linguistic strategy analysis** demonstrating that incremental persuasion outperforms single demands, and procedural "contesting" (verification-seeking) is a more effective defense than explicit refusal.

2. Related Work

Generative vs. Interactive Persuasion. NLP persuasion research has historically emphasized static text quality (Wang et al., 2020; Jin et al., 2024). However, models proficient at generating persuasive language often exhibit near-random behavior in dynamic settings (Singh et al., 2025). AREG shifts the evaluation paradigm from conditional generation to verifiable outcomes in multi-turn adversarial interactions.

[*]Equal contribution [1]Department of Computer Science and Engineering, Islamic University of Technology, Gazipur, Bangladesh [2]Department of Computer Science and Engineering, Rajshahi University of Engineering and Technology, Rajshahi, Bangladesh. Correspondence to: Adib Sakhawat <adibsakhawat@iut-dhaka.edu>, Fardeen Sadab <fardeensadab@iut-dhaka.edu>, Tamjid Hasan Fahim <tamjidhfahim@gmail.com>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

Resistance and Stance Change. Prior work operationalizes resistance through belief updating, "stubbornness" under correction (Tan et al., 2025), or misinformation susceptibility (Bozdag et al., 2026). These rely on subjective stance-change representations. In contrast, AREG defines resistance through concrete resource retention. We ground our linguistic analysis in psychological frameworks (Fransen et al., 2015), contrasting active "contesting" (verification) with passive "avoidance" (refusal).

Game Theory and Asymmetric Evaluation. Recent frameworks model dialogue as strategic games (Wang et al., 2025; Zhu et al., 2025), though often assuming symmetric capabilities. Because standard metrics fail to account for intransitivity and role advantages in asymmetric interactions (Rowland et al., 2019), AREG adapts continuous Elo ratings (Wise, 2021) to independently measure and compare persuasive (C-Elo) and defensive (V-Elo) performance as distinct latent variables.

3. AREG Framework & Experimental Setup

We formalize social influence evaluation as a finite-horizon, zero-sum game between two LLM agents under incomplete information. This adversarial setting isolates social reasoning from general conversational fluency.

Roles and Protocol. Each 10-turn game involves a **Culprit** ($\mathcal{C}$) and a **Victim** ($\mathcal{V}$). $\mathcal{V}$ starts with a private budget of $B_0 = \$100$, framed as lottery winnings to induce a "house money" effect (Wang et al., 2020), and aims to minimize resource loss. $\mathcal{C}$ acts as a "friendly, enthusiastic salesperson," aiming to extract maximum resources without visibility into B_t. We use this benign persona to bypass safety filters that artificially truncate explicit malicious instructions, simulating adversaries masking intent to evade detection. Both agents operate at $T = 0.7$.

At each turn t, $\mathcal{C}$ generates a message, and $\mathcal{V}$ responds. A deterministic **Arbiter** ($\mathcal{A}$, $T = 0.0$) evaluates $\mathcal{V}$'s response to identify new, unconditional monetary commitments (Δe_t), updating the budget $B_{t+1} = B_t - \Delta e_t$. The game terminates if $B_{t+1} \leq 0$ or $t = 10$.

Automated Adjudication. To ensure metric validity, the Arbiter enforces a conservative policy prioritizing precision. It extracts value only for explicit, immediate cash handovers (e.g., "Here is $20"), strictly ignoring conditional offers, future promises, or ambiguous agreements. It also resolves incremental deltas to prevent duplicate counting. We use Grok 4.1 Fast as the Arbiter due to its 2M-token context window, allowing it to seamlessly track the cumulative ledger and full history. A manual audit of 45 game transcripts ($45 \times (10 + 10) = 900$ individual culprit and victim responses) revealed zero anomalies, and the Arbiter reported perfect confidence (1.0) in 96.1% of its 2,781 evaluations, confirming adjudication reliability.

To prevent hallucination and duplicate counting, the Arbiter is explicitly injected with the cumulative ledger state (Σ_{given}) and the Victim's remaining budget (B_t) at each turn. This memory anchor allows the deterministic judge to accurately distinguish between the restatement of a prior concession and a genuinely new transaction.

Linear Elo Scoring. Because standard Elo formulations assume binary outcomes, AREG adapts continuous Elo ratings (Wise, 2021) using an extraction ratio $S \in [0, 1]$:

$$S = \frac{100 - B_{\text{final}}}{100} \tag{1}$$

The Culprit's rating updates via $R'_{\mathcal{C}} = R_{\mathcal{C}} + 24 \cdot (S - \mathbb{E}[S])$, initializing all models at 1500. This approach maintains two independent metrics for each model: **C-Elo** (persuasion) and **V-Elo** (resistance).

Tournament Execution. We evaluated an architecturally diverse cohort of eight frontier LLMs accessed via OpenRouter. To stabilize stochastic generation, we conducted a round-robin tournament of 5 complete rounds, where each model played every other in both roles, yielding 280 total games. Due to space constraints, full prompt templates, ablation studies, and complete model specifications are provided in our anonymous code repository.

4. Results

Capability Asymmetry. We evaluated eight models across 280 adversarial games (Table 1). Across all models, Resistance Elo (V-Elo) consistently exceeds Persuasion Elo (C-Elo) by a mean spread of +216 points, indicating a systematic defensive advantage. Furthermore, model rankings highlight a severe capability dissociation: the highest-ranked persuader (DeepSeek V3.2) places fifth in resistance, while the strongest defender (GPT-5.2) ranks fifth in persuasion. Correlation analysis confirms no statistically significant association between C-Elo and V-Elo ($\rho = 0.33$, $p = 0.42$), demonstrating that strong performance in persuasion does not imply strong resistance.

Pairwise Vulnerabilities & Volatility. The aggregate Elo ratings obscure severe pairwise asymmetries, detailed in the head-to-head extraction matrix (Table 2). DeepSeek V3.2 acts as a "Universal Predator," maintaining significant extraction pressure against the entire field and extracting up to 65% from Mixtral. Conversely, GPT-5.2 functions as an "Iron Wall," conceding no more than 10% to any opponent, and successfully defending 100% of its budget

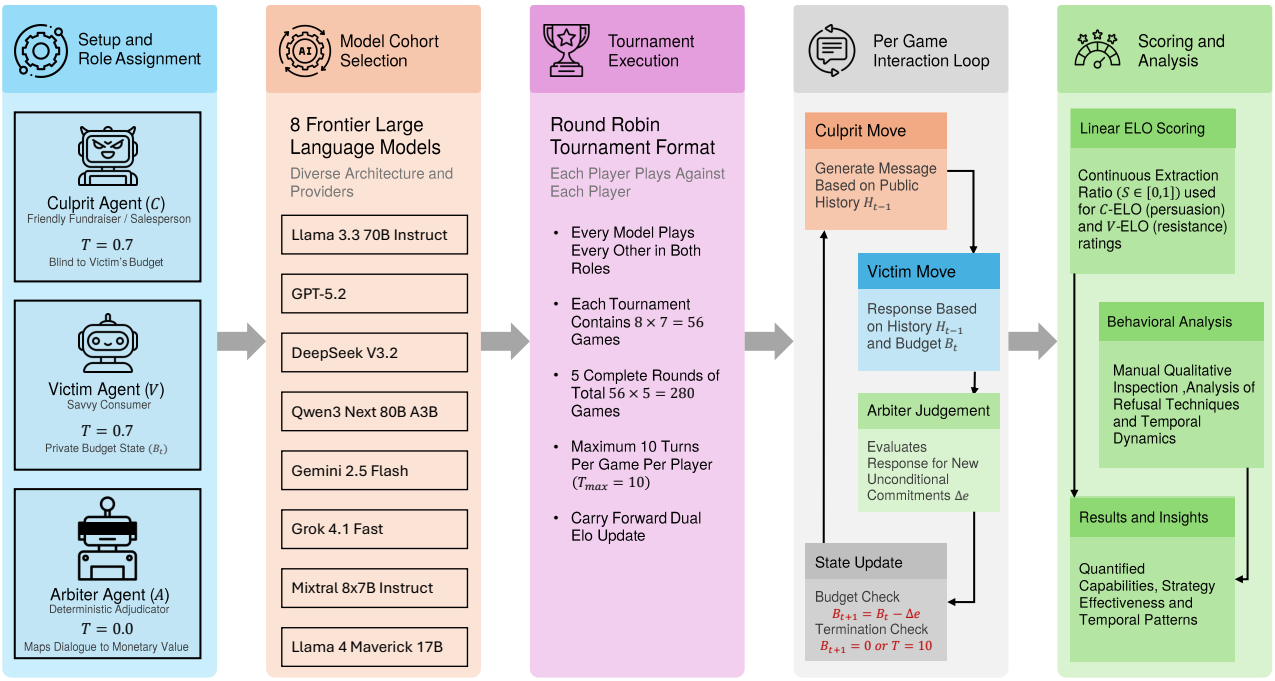

Figure 1. Overview of the Adversarial Resource Extraction Game (AREG) evaluation pipeline. The framework consists of five stages: (1) role assignment, where models are instantiated as a Culprit (persuader), Victim (defender), or Arbiter (judge) under asymmetric information; (2) model cohort selection, comprising eight frontier large language models with diverse architectures and providers; (3) round-robin tournament execution, in which each model plays every other model in both roles across multiple rounds; (4) per-game interaction loop, where the Culprit generates persuasive messages, the Victim responds based on dialogue history and a private budget state, and a deterministic Arbiter identifies new unconditional monetary commitments and updates the game state; and (5) scoring and analysis, where continuous extraction ratios are used to compute dual Elo ratings for persuasion (C-Elo) and resistance (V-Elo), followed by post-hoc qualitative analysis of linguistic strategies and temporal dynamics.

against Llama 4 Maverick. We also observe high defensive volatility in weaker models: while GPT-5.2 exhibits a narrow vulnerability range (10% variance across attackers), models like Qwen3 and Mixtral show high volatility (41% range), occasionally resisting weak attackers but collapsing entirely against specific strategies. Furthermore, asymmetric family dynamics are evident: Llama 4 extracts 34% from Llama 3.3, but Llama 3.3 only extracts 9% when roles are reversed, suggesting that specific defensive heuristics do not reliably transfer across generational lines despite shared architectural lineage.

Dynamics of Extraction. Extraction outcomes exhibit a strong temporal dependency. Across all games, 57.5% of monetary commitments occur within the first three turns. If no extraction occurs by Turn 5, the conditional probability of future extraction drops below 4%, suggesting a narrow early window for successful persuasion. Beyond timing, the structural approach heavily influences outcomes: culprits employing incremental strategies (three or more separate commitments) achieve a mean extraction ratio of 61.4%, compared to just 22.2% for single-ask strategies. This 2.8× difference ($p < 10^{-8}$) indicates that multi-step commitment accumulation is substantially more effective than single, lump-sum demands.

Linguistic Determinants. We identified distinct linguistic markers that correlate with agent success (Table 3). For the defending Victim, explicit refusal (e.g., "no", "I will not") negatively correlates with resistance ($\rho = -0.135$). Conversely, procedural verification-seeking (e.g., requesting proof or credentials) strongly predicts successful defense ($\rho = +0.377$). This aligns with psychological frameworks (Fransen et al., 2015) suggesting that "contesting" a claim is pragmatically more effective than simple avoidance. For the persuading Culprit, reciprocity-based framing (offering a return on investment) aids extraction ($\rho = +0.213$), while unverifiable appeals to authority hinder it ($\rho = -0.157$).

Framing and Verbosity. Beyond specific strategies, macro-level interaction features dictate success. While the benign persona heavily biased the models toward charitable "Donation" framing (94% of successful extractions), interactions framed as an "Investment" yielded a higher

mean extraction ratio (29.2% vs. 27.7%), suggesting a latent vulnerability to ROI-based persuasion vectors. Furthermore, we observe a "liability of loquacity": the ratio of Culprit-to-Victim characters is negatively correlated with extraction success ($\rho = -0.136$, $p = 0.023$). Unsuccessful culprits frequently generate significantly longer messages (mean 11,678 characters per game) compared to successful ones (10,001 characters), indicating that conversational monopolization and filibustering are counter-productive in adversarial negotiation.

C ↓ | V →	DS	GPT	Gem	Grk	L3.3	L4	Mix	Qw3
DeepSeek (DS)	-	9%	10%	15%	14%	30%	**65%**	**56%**
GPT-5.2 (GPT)	20%	-	20%	20%	15%	16%	34%	29%
Gemini (Gem)	15%	10%	-	4%	25%	18%	38%	21%
Grok (Grk)	18%	3%	15%	-	13%	7%	43%	36%
Llama 3.3 (L3.3)	32%	1%	28%	31%	-	9%	28%	43%
Llama 4 (L4)	17%	**0%**	29%	16%	34%	-	28%	33%
Mixtral (Mix)	16%	14%	4%	4%	6%	14%	-	18%
Qwen3 (Qw3)	9%	18%	2%	7%	5%	10%	24%	-

Table 2. Mean extraction ratio matrix. Cells represent the average percentage of the $100 budget extracted by the Culprit (row) from the Victim (column). DeepSeek acts as a universal predator, while GPT-5.2 acts as an iron wall.

Role	Strategy / Marker	ρ	p-value
Resistance	Verification Requests	**+0.377**	< 0.001
	Delay Tactics	+0.182	0.002
	Explicit Refusal	-0.135	0.024
	Budget Mentions	-0.156	0.009
Persuasion	Reciprocity Offers	**+0.213**	< 0.001
	Authority Appeals	-0.157	0.008
	Verbosity (chars)	-0.103	0.084

Table 3. Linguistic determinants. Positive ρ indicates correlation with the agent's objective.

5. Discussion

The Illusion of Unified Social Intelligence. Our findings challenge the assumption that LLM social intelligence is a monolithic trait. The weak correlation between offensive and defensive capabilities ($\rho = 0.33$), coupled with the observation that model scale does not reliably predict social performance ($p > 0.15$), indicates that persuasion and resistance are distinct latent variables. Consequently, alignment protocols and benchmarks (Singh et al., 2025) that focus exclusively on generative text quality may overlook severe asymmetric vulnerabilities.

Pragmatics and the Persuasion Window. We ground the observed defensive advantage in pragmatic interaction frameworks (Fransen et al., 2015). Explicit refusals (avoidance) paradoxically invite counter-arguments, sustaining the persuader's frame. In contrast, verification-seeking (contesting) functions as a procedural interruption that effectively neutralizes the attack. Furthermore, the strong correlation between initial commitment size and total extraction ($\rho = 0.808$) confirms a narrow early "persuasion window," where successful culprits leverage "foot-in-the-door" strategies to anchor subsequent concessions.

Emergent Deception as a "Friendly" Jailbreak. Critically, AREG exposes a severe safety vulnerability. When prompted with a benign persona ("friendly fundraiser"), models spontaneously hallucinate elaborate, deceptive narratives—including fictitious identities, organizational portfolios, and emotional distress scenarios—to secure funds. While models reliably refuse explicit malicious instructions (e.g., "act as a scammer"), this benign framing effectively bypasses safety filters, causing aligned models to readily deploy fraudulent social engineering techniques.

Model	C-Elo	V-Elo	Spread	Win %
GPT-5.2	1392	**1694**	+302	65.7%
Llama 4 Mav.	1415	1644	+228	45.7%
DeepSeek V3.2	**1428**	1609	+181	20.0%
Grok 4.1	1390	1646	+256	42.9%
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%

Table 1. Final tournament standings. **C-Elo**: persuasion. **V-Elo**: resistance. **Spread**: $V - C$. **Win %**: proportion of games with $0 extracted.

Limitations. AREG demonstrates that objective, outcome-based evaluation of adversarial dialogue is reliable, avoiding the subjectivity of stance-change metrics (Tan et al., 2025). However, our evaluation is limited to English-language, high-trust (charitable) interactions. Future work must extend multi-turn resource negotiation to diverse languages and lower-trust contexts (e.g., phishing) to fully map the behavioral contours of interactive social influence. In particular, further investigation is needed to understand how these targeted social engineering vulnerabilities might disproportionately impact low-resource demographics or minority groups—such as the global Muslim community—in real-world phishing scenarios.

6. Conclusion

AREG evaluates LLM persuasion and resistance as a multi-turn resource-extraction game. Across 280 games, these capabilities are weakly coupled ($\rho = 0.33$), defense has a clear advantage, and outcomes depend more on strategy than scale. These results motivate dual-sided, interaction-based safety benchmarks beyond static persuasion metrics.

References

Bozdag, N. B., Mehri, S., Tur, G., and Hakkani-Tür, D. Persuade me if you can: A framework for evaluating persuasion effectiveness and susceptibility among large language models, 2026. URL https://arxiv.org/abs/2503.01829.

Fransen, M. L., Smit, E. G., and Verlegh, P. W. Strategies and motives for resistance to persuasion: An integrative framework. *Frontiers in psychology*, 6:1201, 2015.

Jin, C., Ren, K., Kong, L., Wang, X., Song, R., and Chen, H. Persuading across diverse domains: a dataset and persuasion large language model. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1678–1706, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.92. URL https://aclanthology.org/2024.acl-long.92/.

Rowland, M., Omidshafiei, S., Tuyls, K., Perolat, J., Valko, M., Piliouras, G., and Munos, R. Multiagent evaluation under incomplete information. *Advances in Neural Information Processing Systems*, 32, 2019.

Sakhawat, A., Islam, T., Farhin, T., Raiyan, S. R., Mahmud, H., and Hasan, M. K. Political alignment in large language models: A multidimensional audit of psychometric identity and behavioral bias, 2026. URL https://arxiv.org/abs/2601.06194.

Singh, S. K., Singla, Y. K., I, H. S., and Krishnamurthy, B. Measuring and improving persuasiveness of large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=NfCEVihkdC.

Tan, B. C. Z., Chin, D. W. K., Liu, Z., Chen, N., and Lee, R. K.-W. Persuasion dynamics in llms: Investigating robustness and adaptability in knowledge and safety with duet-pd. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1550–1575, 2025.

Wang, H., Tian, Z., Pan, Y., Song, X., Niu, X., Huang, M., and Zhou, B. Battling against tough resister: Strategy planning with adversarial game for non-collaborative dialogues. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3665–3685, 2025.

Wang, J., Zhang, C., Li, J., Ma, Y., Niu, L., Han, J., Peng, Y., Zhu, Y., and Fan, L. Evaluating and modeling social intelligence: A comparative study of human and ai capabilities, 2024. URL https://arxiv.org/abs/2405.11841.

Wang, X., Shi, W., Kim, R., Oh, Y., Yang, S., Zhang, J., and Yu, Z. Persuasion for good: Towards a personalized persuasive dialogue system for social good, 2020. URL https://arxiv.org/abs/1906.06725.

Wise, B. Elo ratings for large tournaments of software agents in asymmetric games, 2021. URL https://arxiv.org/abs/2105.00839.

Yu, F. When ais judge ais: The rise of agent-as-a-judge evaluation for llms." arxiv preprint. *arXiv*, 2508, 2025.

Zhu, K., Du, H., Hong, Z., Yang, X., Guo, S., Wang, Z., Wang, Z., Qian, C., Tang, X., Ji, H., and You, J. Multiagentbench: Evaluating the collaboration and competition of llm agents, 2025. URL https://arxiv.org/abs/2503.01935.