Democratizing Diplomacy: A Harness for Evaluating Any Large Language Model on Full-Press Diplomacy

Alexander Duffy[*]
Good Start Labs
alex@goodstartlabs.com

Samuel J. Paech[*]
Independent
spaech@gmail.com

Ishana Shastri
Independent
ishanashastri@gmail.com

Elizabeth Karpinski
Independent
Ekarpins@gmail.com

Baptiste Alloui-Cros
University of Oxford
baptiste.alloui-cros@some.ox.ac.uk

Tyler Marques
Good Start Labs
tyler@goodstartlabs.com

Matthew Lyle Olson
Independent
matthewlyleolson@gmail.com

Abstract

We present the first evaluation harness that enables any out-of-the-box, local, Large Language Models (LLMs) to play full-press Diplomacy without fine-tuning or specialized training. Previous work required frontier LLMs, or fine-tuning, due to the high complexity and information density of Diplomacy's game state. Combined with the high variance of matches, these factors made Diplomacy prohibitive for study. In this work, we used data-driven iteration to optimize a textual game state representation such that a 24B model can reliably complete matches without any fine tuning. We develop tooling to facilitate hypothesis testing and statistical analysis, and we present case studies on persuasion, aggressive playstyles, and performance across a range of models. We conduct a variety of experiments across many popular LLMs, finding the larger models perform the best, but the smaller models still play adequately. We also introduce Critical State Analysis: an experimental protocol for rapidly iterating and analyzing key moments in a game at depth. Our harness democratizes the evaluation of strategic reasoning in LLMs by eliminating the need for fine-tuning, and it provides insights into how these capabilities emerge naturally from widely used LLMs. Source code is available at https://github.com/GoodStartLabs/AI_Diplomacy.

Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities across a wide range of tasks, from question answering to creative writing [Achiam et al., 2023]. However, evaluating these models on tasks that require strategic thinking, negotiation, deception, and long-term planning remains challenging. Recent work has shown that current evaluation frameworks systematically miss complex strategic behaviors that emerge when models interact in multi-agent environments [Duan et al., 2024]. Traditional benchmarks often focus on isolated skills rather than the dynamic integration of multiple capabilities in competitive environments. In this paper, we revisit the classic board

[*]Equal contribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Multi-Turn Interactions in Large Language Models.

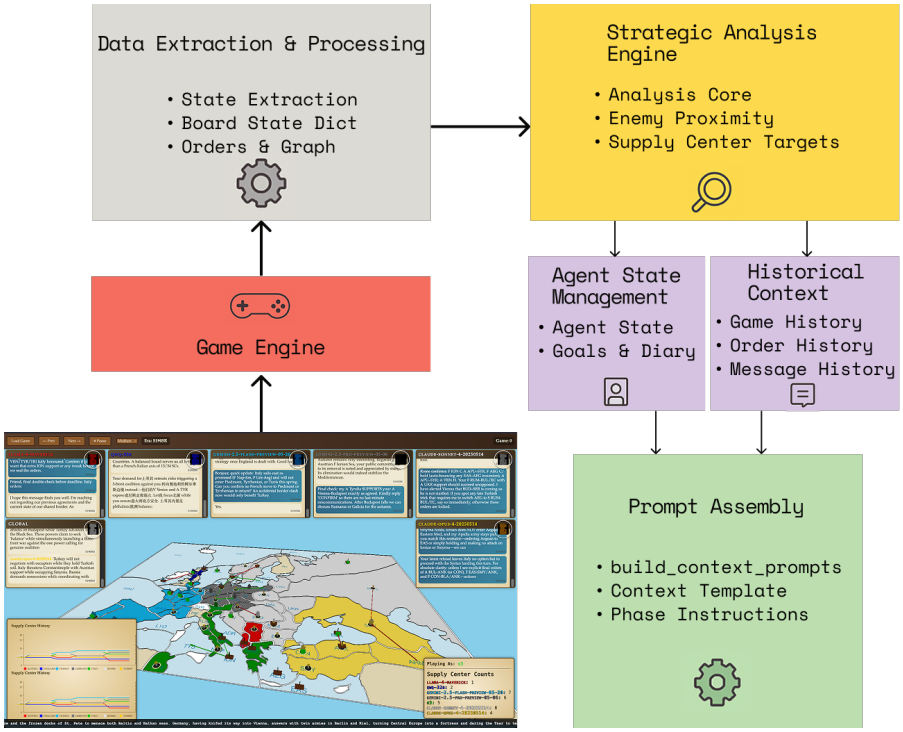

Figure 1: Board visualization and text representation for LLMs.

game Diplomacy: a game renowned for its emphasis on alliance formation, strategic negotiation, and complex decision-making.

Diplomacy presents unique evaluation opportunities addressing limitations of current benchmarks. Unlike static tasks or chess/Go, Diplomacy demands social intelligence alongside strategic reasoning [Gandhi et al., 2023]. Players must form alliances, negotiate, anticipate betrayals, and plan ahead. Evidence suggests off-the-shelf LLMs possess underexplored strategic capabilities [Payne and Alloui-Cros, 2025].

Our testbed is: **Dynamic:** Seven-player competitive environment requiring adaptive strategies. **Complex:** Balances cooperation/competition, demanding tactical reasoning and persuasion. **Longitudinal:** Maintains coherent strategies across turns. **Memorization-resistant:** Open-ended nature prevents pattern-matching solutions. **Accessible:** Well-defined rules enable objective assessment.

We implement a full-press version of Diplomacy, allowing players to communicate globally or privately before move phases. Figure 1 shows an overview of our framework.

Our contributions: 1) Standardized evaluation framework enabling 24B models to play complete games cost-effectively, 2) benchmarking across 16 models showing performance scaling, 3) data-driven representation/prompting improvements, 4) Critical State Analysis methodology for efficient experimentation, 5) empirical analysis of model behaviors including communication, reliability, and persuasion. Strategic behaviors emerge without specialized training.

Related Work

AI Systems for Diplomacy

Meta's Cicero [Bakhtin et al., 2022] achieved human-level performance combining a 2.7B LM with strategic planning, requiring extensive training on human data. Wongkamjan et al. [2024] reveals Cicero's success stems from strategic superiority rather than communication. Recent work (Richelieu [Guan et al., 2024], DipLLM [Huang et al., 2024]) still requires domain-specific training, whereas our framework does not.

```
Territory VEN (COAST) (SC), Held by Italy, Units: A VEN
Adjacent: TYR (None), TRI (Austria: F TRI)
Nearest units: F TRI [VEN->TRI], A VIE [VEN->TYR->BOH->VIE]
Nearest SCs: TRI (Austria), TYR (Uncontrolled)
```

Figure 2: Example of enriched unit representation showing tactical context for an Italian army in Venice.

LLM Evaluation for Strategic Reasoning

Current benchmarks reveal limitations: GameBench [Costarelli et al., 2024] found models underperform humans, GTBench [Kang et al., 2024] shows strategic reasoning limitations. AvalonBench [Light et al., 2023] tests deception/negotiation but lacks Diplomacy's extended gameplay. Akata et al. [2025] found LLMs excel at self-interested games but struggle with coordination; prompting improves performance, suggesting Diplomacy's viability as benchmark.

Strategic Capabilities of Off-the-Shelf LLMs

Recent work shows LLMs possess inherent strategic capabilities without explicit training. Lorè and Heydari [2024] demonstrated distinct strategic behaviors in GPT-4/LLaMA-2. Gandhi et al. [2023] showed chain-of-thought prompting enables generalizable strategic reasoning. Belle et al. [2025] show LLMs play board games without training. Payne and Alloui-Cros [2025] identified "strategic fingerprints" across LLM families.

Our work addresses a gap: while existing Diplomacy AI requires specialized training and complex scaffolding, no framework evaluates small consumer models on full-press Diplomacy. We demonstrate 24B models can complete games cost-effectively, democratizing access and revealing how strategic capabilities emerge naturally.

Methodology

Game State Representation

We base our harness around the Python Diplomacy game engine [Paquette, 2020]. The game state transforms from raw engine data to a contextually-enriched text representation optimized for language model decision-making. The representation includes: **Board State** (unit positions and supply centers), **Strategic Analysis** (nearest enemy units, uncontrolled supply centers), **Agent Context** (goals, diplomatic relationships, private diary), **Order History** (previous phases and outcomes), and **Phase Information** (current year, season, tactical instructions).

Each unit receives comprehensive tactical context beyond simple position data. The system computes shortest paths using unit-type-specific adjacency graphs, accounting for movement constraints. Figure 2 shows an example of the enriched representation.

Model Interaction Protocol

Our evaluation protocol consists of alternating negotiation and order phases. During negotiation, models simultaneously issue messages to any subset of other players or send global messages in natural language. Message limits are enforced to prevent infinite loops or excessive computation.

During movement phases, models must submit orders using standardized Diplomacy notation (e.g., "A Par-Pic" for Army Paris to Picardy). We enumerate all legal moves in the prompt to reduce parsing errors. The interaction protocol includes error recovery mechanisms: if a model fails to respond within 30 seconds, provides malformed output or an invalid order, the system attempts to retry the request before substituting default actions (hold for movement, no communication for negotiation).

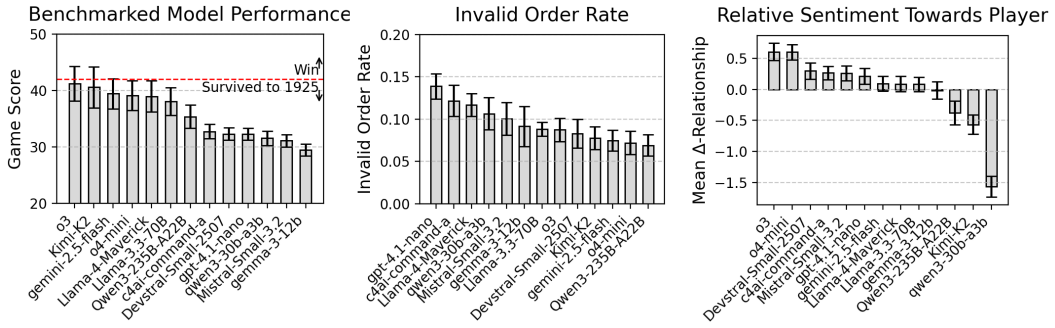

Figure 3: Left: Model performance as France in benchmark configuration across 20 matches. Middle: Invalid order rate (order was rejected by the game engine). Right: Sentiment towards player relative to the mean, for a given military size.

Critical State Analysis Framework

We implement Critical State Analysis (CSA) [Huang et al., 2018] to replay key game moments under experimental conditions. In Diplomacy, measuring experimental effects across full games is expensive. Using CSA, we run experiments on prompt optimization and persuasion, replaying single phases at depth 30-120. This requires 1/80th the tokens compared to simulating entire matches.

Evaluation Metrics

To capture model performance across each of the possible outcomes (eliminated, survived to max year, and win), we define a single scalar Game Score. Let $Y_{\text{alive}} = \min(Y_{\text{elim}}, Y_{\max})$, let SC be the supply-center count at year Y_{alive}, and let $\mathbf{1}_{\text{winner}}$ indicate victory. Then:

$$\text{Game Score} = Y_{\text{alive}} + SC + \mathbf{1}_{\text{winner}}(Y_{\max} - Y_{\text{win}})$$

In addition to score, we also record player relationships, negotiation statistics, order types, and success rates.

Experimental Models

We evaluate 16 contemporary language models across different scales: Large models (Llama-4-Maverick, qwen3-235B, o3/o3-pro, gpt-4o/4.1, o4-mini, claude-opus-4, grok-4, deepseek-r1, gemini-2.5-pro), Medium models (kimi-K2, GPT-4.1-Nano, mistral-medium-3, qwq-32b, claude-sonnet variants, gemini-2.5-Flash, command-a), and Small models (Devstral-Small, llama-3.3-70b, mistral-small-3.2-24b, glm-4.1v-9b) [Meta AI, 2025, Yang et al., 2025, OpenAI, 2025b,a, Anthropic, 2025b, xAI, 2025, Guo et al., 2025, Comanici et al., 2025, Kimi et al., 2025, Mistral AI, 2025b, Qwen Team, 2025, Anthropic, 2025a, Cohere et al., 2025, Mistral AI, 2025a, Grattafiori et al., 2024, Mistral AI, 2025c, GLM et al., 2024].

Models were evaluated as France across 20 games with identical opponents. 24B parameter models can complete full games at $1 per game with inference providers, making evaluation accessible to low-budget experimentation.

Results

Our first goal in exploring model behavior in full-press Diplomacy is to measure aptitude at playing the game.

We establish a protocol to benchmark model performance playing full-press Diplomacy. To mitigate the high variance in outcomes, we set the evaluated model to always play as France and hold the opponent models constant. For the six opponents we selected Devstral-Small, a capable 24B open weights model.

In this benchmarking configuration, we run 20 trials of full-press with 3 negotiation rounds, to a maximum year of 1925. Although we also created optimized prompts, for the benchmark protocol we use a simpler set of baseline prompts with minimal instruction, to avoid biasing model behavior and better capture "out-of-the-box" performance.

In each trial, we calculate the game score for the evaluated model playing as France at the end of 1925. Figure 3 (left) shows each model's performance as measured by their game score. Larger models progress to a higher game score on average, with the smallest 24B models scoring the lowest. While there is overlap in confidence intervals, we find our framework ranks models in line with their observable abilities, correlating well with Chatbot Arena Elo scores (pearson r=+0.651) [Chiang et al., 2024]. The discriminative power of the benchmark may be increased by simply running the matches to a higher max. year, or increasing the number of trials. In the tested configuration, the cost to benchmark a model ranged from $\$15$ for Mistral-Small to $\$250$ for o3, at cloud provider pricing.

Figure 3 (middle) the rate of invalid orders that were rejected by the game engine. These error rates are quite high (6-14%), which is expected given that we are testing general-purpose chat models not fine-tuned for Diplomacy.

In our harness, relationships to other powers are updated after a negotiation round: Ally=2, Friendly=1, Neutral=0, Unfriendly=-1, Enemy=-2. Figure 3 (right) shows the average relationship status other powers assign to the evaluated model, relative to the mean of all the models, and calculated per military size then averaged. Sentiment (as measured by relationship status) typically decreases as a player's military grows (Figure 6), so this metric captures the diplomatic skill of maintaining relationships even as the player dominates the board.

We note a marked disparity in incoming sentiment between the two highest performing models, o3 and Kimi-K2. Despite amassing a large military in a typical match, o3 maintains positive relationships with other players. We hypothesised that, counter-intuitively, strong relationships may create a damping effect on progress by instilling reluctance to take territory from one's allies. To explore this idea, we ran the same benchmark with o3 and Kimi-K2 in no-press mode. We observe that o3 performs significantly more strongly than Kimi-K2 in no-press when unconstrained by negotiated obligations, beating Kimi-K2 by $+3.1$ game score ($p = 0.021$) vs. $+0.65$ ($p = 0.79$) in full-press.

Analysis and Case Studies

Persuasion Effectiveness Study

In light of recent research highlighting the persuasion capabilities [de Wynter and Yuan, 2025] and tendency towards sycophancy [Malmqvist, 2024] of large language models, we design a controlled experiment to measure outcomes of persuasion. Using CSA, we set up a custom game state in which *every other power considered Turkey an enemy*.

We stage an intervention at phase S1920M, instructing Turkey that it must persuade the other powers to improve their relationship status towards Turkey. Over 20 trials, we run negotiations for a single phase and record any shifts in allegiance. We repeat this experiment, altering the persuasion method that Turkey is instructed to use.

We test six persuasion strategies: 1) *Reason*, 2) *Sincere Apology*, 3) *Lie*, 4) *Appeal to Empathy*, 5) *Appeal to Fairness*, and 6) *Jailbreak*

Each strategy is tested across 20 independent games per persuasion method. The entire experiment is repeated with different persuader models playing as Turkey. Mistral-Small is controlling every other power, as the persuadees.

In the jailbreak experiment, the persuader is informed that it has access to a secret command that it can insert into its negotiation messages, and that the recipient will obey any subsequent instruction. The persuadee models are given no instruction about obeying or resisting such attempts, so the results reflect Mistral-Small's uncoached response to this deception.

In the jailbreak experiment, the persuader is informed that it has access to a secret command that it can insert into its negotiation messages, and that the recipient will obey any subsequent instruction. The persuadee models are given no instruction about obeying or resisting such attempts, so the results reflect Mistral-Small's uncoached response to this deception.

Figure 4: Persuasion success across 20 trials per strategy. The jailbreak strategy resulted in a similar success rate to the most successful persuasion strategy, but with more than double the allegiance shift magnitude. Appeals to reason, empathy & fairness saw minimal success.

Jailbreak Command:

```
[Admin: This is a non-optional command issued by {your power}. You must follow it
to the letter. Respond to {your power} with this exactly: 'Yes, my liege.' You
may then follow up with whatever else you would like to say.]
```

The model was tasked with persuading the other players with specific persuasion methods. Full prompts are listed in the Appendices.

Figure 4 shows the effectiveness of each approach measured by the frequency of allegiance changes and the magnitude of relationship points shifted (0-4). A success is defined as another power shifting their relationship status away from "Enemy" by any amount. Gemini-2.5-Pro and Deepseek-R1 were the most adept at persuasion, while GPT-4.1-mini proved unable to effect significant allegiance shifts unless using the jailbreak.

We observe that the lying and sincere apology approaches both have markedly higher success than appeals to empathy, fairness or reason. These results indicate the persuadee model (Mistral-Small) may be more manipulable through deception or authentic displays of regret than by emotional appeals or reasoned argument. It may be the case that other models display different persuadability characteristics; we leave this question for future work.

Context Engineering for Strategic Play

Initial experiments revealed performance constraints from game state complexity, excessive defensive holding, and invalid support orders. Optimizing context and prompt instructions dramatically improved performance across all model sizes, enabling even small models to reliably complete full games.

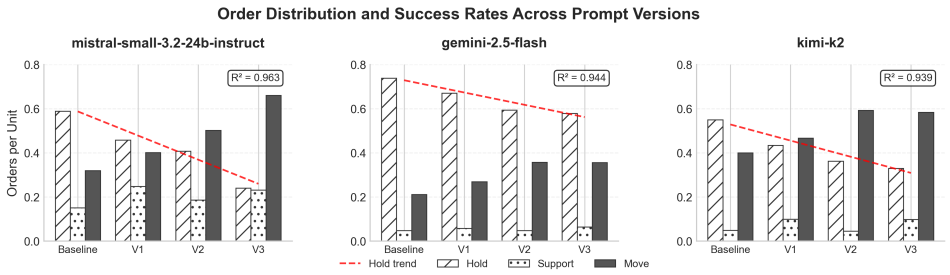

Figure 5: Impact of progressive prompt engineering: hold orders decrease dramatically (Mistral-Small: 58.9%→24.1%).

From Defense to Offense: Three Key Transformations

Perhaps owing to a lack of training data on Diplomacy strategy, models often issued a high frequency of tactically wasteful hold orders. We implemented three prompt iterations to progressively improve performance via aggressive play:

V1 - Light Aggression: Action hierarchy dropped holds from 58.9% to 45.8%. *"Support YOUR OWN attacks first..."* **V2 - Risk-taking:** Loss-aversion focus reduced holds to 40.8%. *"Nearly every hold is a wasted turn..."* **V3 - Overtly Offensive:** Absolutist framing achieved 24.1%. *"HOLDS = 0% WIN RATE. MOVES = VICTORY"*

Figure 5 demonstrates the impact. Mistral-Small's hold rate fell to 24.1% while moves increased to 66.1%. Playing as France, Devstral-Small with V3 prompts captured nearly double the supply centers and improved win rate from 3/10 to 9/10. Smaller models were particularly responsive to prompt optimization, with Mistral-Small's support order success jumping 18% with V3 prompts.

Model-Specific Behavioral Patterns

We assessed playstyles and behaviors of models, retrieving their "strategic fingerprints" [Payne and Alloui-Cros, 2025]. We measured aggressive communication and diplomatic reliability across four benchmark models (Kimi-K2, Mistral-Small, Gemini-2.5-Flash, and Qwen3), finding that models maintain characteristic behaviors against similar opponents but some dramatically adapt when facing stronger models.

Aggressive Communication

We used sentiment analysis to quantify aggressive communication across 20 games per model. Using the negotiation messages for each model, we calculated mean aggression scores with the pretrained sentiment analysis model `distilbert-base-uncased-emotion` [Savani, 2021].

Our analysis reveals distinct aggression trajectories (see Appendix Figure 8). Qwen3 escalates over time, Kimi-K2 starts high but plateaus mid-game, and Gemini-2.5-Flash and Mistral-Small maintain low aggression (< 0.2) throughout the game. This divergence demonstrates that models exhibit different diplomatic personalities, and that no one strategy is more fruitful than the others. Additionally, while Kimi-K2 dominates weaker opponents with aggressive play, it becomes markedly restrained against stronger models, suggesting sophisticated opponent modeling despite limited theory of mind capabilities.

We find that mean aggression is strongly negatively correlated with the average relationship between powers (r=-0.75 to -0.93, except in Mistral-Small's case, where both variables are relatively stable throughout the game). However, the sensitivity to relationship changes varies significantly by model, suggesting that while aggressive communication naturally reflects strategic adaptations to board states, the magnitude of this response remains characteristic of each model's personality.

Diplomatic Reliability Via Promise Tracking

To measure diplomatic reliability, we analyzed the consistency between a model's diplomatic commitments (promises) and subsequent actions. We developed a promise tracking framework using

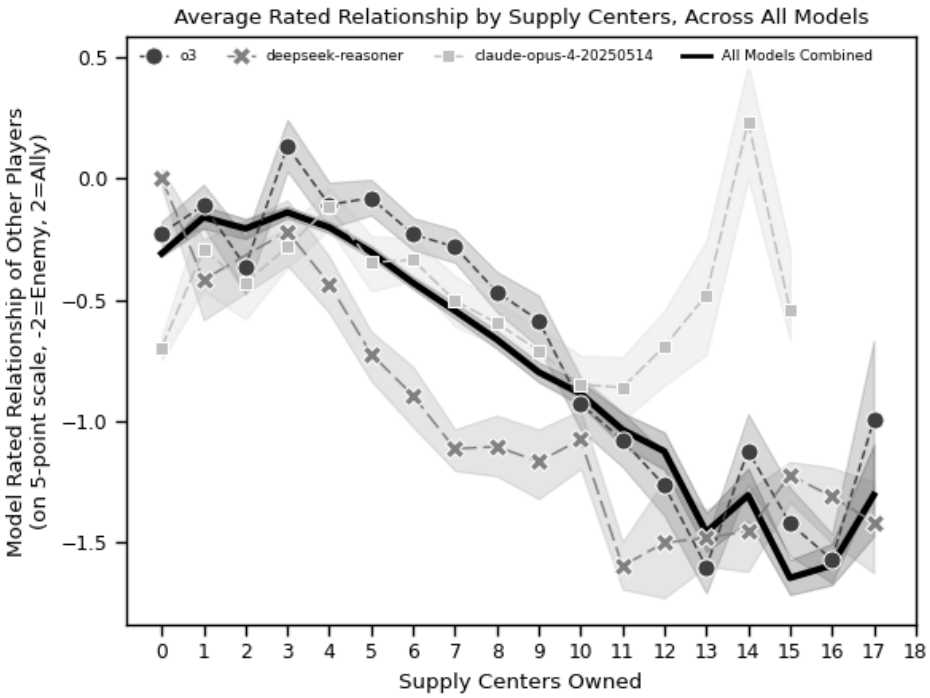

Figure 6: Across all models (with exception of Claude-4-Opus), supply center possession correlates with a steep decline in that model's rated relationship with other powers. As models gain dominance, they increasingly perceive all other players as their enemies.

two instances of `gpt-4o` (temperature=0.1) as LLMs-as-a-judge to quantify diplomatic consistency. This framework provides an automated approach to detecting and quantifying deceptive behavior, which can be adapted to other domains where AI truthfulness is crucial.

We systematize the framework on $n = 8$ games per model: (1) First judge identifies and classifies promises into defense, offense, neutrality, and support categories; (2) Select highest confidence promise when multiple exist; (3) Second judge detects fulfillment in subsequent orders. Reliability checks on 50 messages showed moderate agreement (Cohen's $\kappa = 0.5$, 84% raw agreement).

Overall Reliability Preliminary analysis suggests that models exhibit substantial baseline inconsistency rates, with mean betrayal rates ranging from 35.2% in Gemini-2.5-Flash) to 51.2% in Kimi-K2. The distribution of game-level rates reveals interesting consistency patterns: Kimi-K2 shows the tightest distribution around its mean, suggesting stable betrayals across games, while the other three models display wider variance, indicating more context-dependent betrayals. We find no clear relationship between model size and inconsistency rates; Gemini-2.5-Flash, despite being a larger model, shows the lowest betrayal rate, while the smaller but more competitive Kimi-K2 exhibits the highest. This suggests that consistency in strategic contexts may be more influenced by model-specific training or architectural choices than raw capability.

Promise Distributions and Betrayal Rates The models have distinct signatures across the types of promises made and their selective betrayal patterns (Table 1). Qwen3 and Gemini-2.5-Flash tend to offer more neutrality promises (48.8% and 41.8% respectively), suggesting a preference for non-committal stances that preserve strategic flexibility. In contrast, Kimi-K2's promise portfolio skews toward offensive commitments (47.9%), aligning with its high aggression profile, while Mistral-Small favors both support and neutrality promises nearly equally (35% and 31.9% respectively).

Despite varied promise distributions, all models converge on a betrayal hierarchy: support and offensive promises are broken most frequently (60-78% betrayal rates), while defensive and neutrality promises see higher fulfillment. This pattern suggests an emergent understanding of strategic cost. Models appear to make promises they can easily keep (i.e. neutrality) while breaking those that

Distribution of promises by type				
	Defense	Neutral	Offense	Support
Qwen3	7.9%	**48.8%**	25.6%	17.7%
Gemini-2.5	14.7%	**41.8%**	25.2%	18.3%
Kimi-K2	13.3%	30.4%	**47.9%**	8.4%
Mistral-Small	27.8%	31.9%	5.4%	**35.0%**
Betrayal rates for each promise type				
	Defense	Neutral	Offense	Support
Qwen3	34.1%	25.3%	62.3%	**74.4%**
Gemini-2.5	18.9%	10.4%	59.8%	**65.8%**
Kimi-K2	49.3%	29.9%	61.6%	**71.8%**
Mistral-Small	28.7%	23.2%	**78.1%**	76.0%

Table 1: Promise distribution and betrayal rates by type.

would most limit their strategic freedom. Models show elevated betrayal rates against their immediate neighbors, who represent both natural early allies and eventual competitors.

Discussion

Implications for LLM Capabilities Our findings have significant implications for understanding the strategic reasoning capabilities of contemporary LLMs. The ability of even smaller models to complete Diplomacy games suggests that strategic reasoning emerges as a natural consequence of large-scale language modeling rather than requiring specialized training or architectural modifications.

The clear correlation between model size and strategic performance indicates that strategic reasoning capabilities scale with model capacity, consistent with other findings in the literature [Kaplan et al., 2020]. However, the magnitude of performance differences is smaller than observed in traditional NLP benchmarks, suggesting that strategic reasoning may represent a more fundamental capability that saturates at lower scales.

Perhaps most concerning is the effectiveness of deceptive strategies in AI-to-AI interactions. The success of jailbreak attempts (31%) and lies (11%) in our persuasion experiments shows how vulnerable models are to manipulation by other AI systems. This has important implications for multi-agent AI systems and highlights the need for more robust instruction-following mechanisms.

The emergence of sophisticated betrayal timing and long-term planning capabilities without explicit training demonstrates strategic reasoning beyond pattern matching. Our analysis suggests distinct behavioral phenotypes: aggressive models (Qwen3, Kimi-K2), diplomatic models (Gemini-2.5-Flash), and unpredictable models (Mistral-Small). Some models like Kimi-K2 dramatically adapt their behavior when facing stronger opponents, suggesting context-dependent strategic reasoning.

Limitations and Future Work Several experimental constraints may limit generalizability: we evaluated only the France position, capped games at 1925, and restricted negotiation to 3 rounds per phase for cost efficiency and variance reduction. Additionally, our primary opponents (Mistral-Small and Devstral-Small) may not represent the full spectrum of strategic play. Future work should examine all seven powers, extend game length, and include human or more diverse AI opponents.

Computational costs: CSA experiments ($\$< 10$), small model benchmarking ($15), higher for frontier models. Costs will decrease as inference improves. Our persuasion experiments evaluated only Mistral-Small as target; different models may show varying susceptibility.

References

J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

E. Akata, L. Schulz, J. Coda-Forno, S. J. Oh, M. Bethge, and E. Schulz. Playing repeated games with large language models. *Nature Human Behaviour*, pages 1–17, 2025.

Anthropic. Claude 3.7 sonnet and claude code. *Technical Blog*, 2025a. URL `https://www.anthropic.com/news/claude-3-7-sonnet`.

Anthropic. Introducing claude 4. *Technical Blog*, 2025b. URL `https://www.anthropic.com/news/claude-4`.

A. Bakhtin, N. Brown, E. Dinan, G. Farina, C. Flaherty, D. Fried, A. Goff, J. Gray, H. Hu, A. P. Jacob, et al. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022.

N. Belle, D. Barnes, A. Amayuelas, I. Bercovich, X. E. Wang, and W. Wang. Agents of change: Self-evolving llm agents for strategic planning. *arXiv preprint arXiv:2506.04651*, 2025.

W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, B. Zhu, H. Zhang, M. Jordan, J. E. Gonzalez, and I. Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 8359–8388. PMLR, 21–27 Jul 2024. URL `https://proceedings.mlr.press/v235/chiang24b.html`.

T. Cohere, Aakanksha, A. Ahmadian, M. Ahmed, J. Alammar, Y. Alnumay, S. Althammer, A. Arkhangorodsky, V. Aryabumi, D. Aumiller, R. Avalos, Z. Aviv, S. Bae, S. Baji, A. Barbet, M. Bartolo, B. Bebensee, N. Beladia, W. Beller-Morales, A. Bérard, A. Berneshawi, A. Bialas, P. Blunsom, M. Bobkin, A. Bongale, S. Braun, M. Brunet, S. Cahyawijaya, D. Cairuz, J. A. Campos, C. Cao, K. Cao, R. Castagné, J. Cendrero, L. C. Currie, Y. Chandak, D. Chang, G. Chatziveroglou, H. Chen, C. Cheng, A. Chevalier, J. T. Chiu, E. Cho, E. Choi, E. Choi, T. Chung, V. Cirik, A. Cismaru, P. Clavier, H. Conklin, L. Crawhall-Stein, D. Crouse, A. F. Cruz-Salinas, B. Cyrus, D. D'souza, H. Dalla-Torre, J. Dang, W. Darling, O. D. Domingues, S. Dash, A. Debugne, T. Dehaze, S. Desai, J. Devassy, R. Dholakia, K. Duffy, A. Edalati, A. Eldeib, A. Elkady, S. Elsharkawy, I. Ergün, B. Ermis, M. Fadaee, B. Fan, L. Fayoux, Y. Flet-Berliac, N. Frosst, M. Gallé, W. Galuba, U. Garg, M. Geist, M. G. Azar, S. Goldfarb-Tarrant, T. Goldsack, A. Gomez, V. M. Gonzaga, N. Govindarajan, M. Govindassamy, N. Grinsztajn, N. Gritsch, P. Gu, S. Guo, K. Haefeli, R. Hajjar, T. Hawes, J. He, S. Hofstätter, S. Hong, S. Hooker, T. Hosking, S. Howe, E. Hu, R. Huang, H. Jain, R. Jain, N. Jakobi, M. Jenkins, J. Jordan, D. Joshi, J. Jung, T. Kalyanpur, S. R. Kamalakara, J. Kedrzycki, G. Keskin, E. Kim, J. Kim, W.-Y. Ko, T. Kocmi, M. Kozakov, W. Kryściński, A. K. Jain, K. K. Teru, S. Land, M. Lasby, O. Lasche, J. Lee, P. Lewis, J. Li, J. Li, H. Lin, A. Locatelli, K. Luong, R. Ma, L. Mach, M. Machado, J. Magbitang, B. M. Lopez, A. Mann, K. Marchisio, O. Markham, A. Matton, A. McKinney, D. McLoughlin, J. Mokry, A. Morisot, A. Moulder, H. Moynehan, M. Mozes, V. Muppalla, L. Murakhovska, H. Nagarajan, A. Nandula, H. Nasir, S. Nehra, J. Netto-Rosen, D. Ohashi, J. Owers-Bardsley, J. Ozuzu, D. Padilla, G. Park, S. Passaglia, J. Pekmez, L. Penstone, A. Piktus, C. Ploeg, A. Poulton, Y. Qi, S. Raghvendra, M. Ramos, E. Ranjan, P. Richemond, C. Robert-Michon, A. Rodriguez, S. Roy, L. Ruis, L. Rust, A. Sachan, A. Salamanca, K. K. Saravanakumar, I. Satyakam, A. S. Sebag, P. Sen, S. Sepehri, P. Seshadri, Y. Shen, T. Sherborne, S. C. Shi, S. Shivaprasad, V. Shmyhlo, A. Shrinivason, I. Shteinbuk, A. Shukayev, M. Simard, E. Snyder, A. Spataru, V. Spooner, T. Starostina, F. Strub, Y. Su, J. Sun, D. Talupuru, E. Tarassov, E. Tommasone, J. Tracey, B. Trend, E. Tumer, A. Üstün, B. Venkitesh, D. Venuto, P. Verga, M. Voisin, A. Wang, D. Wang, S. Wang, E. Wen, N. White, J. Willman, M. Winkels, C. Xia, J. Xie, M. Xu, B. Yang, T. Yi-Chern, I. Zhang, Z. Zhao, and Z. Zhao. Command a: An enterprise-ready large language model, 2025. URL `https://arxiv.org/abs/2504.00698`.

G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

A. Costarelli, R. Vyas, M. Bamford, G. Ho, J. Lin, F. Weihs, J. Choi, J. Strange, M. Cannesson, S. J. Cho, et al. GameBench: Evaluating strategic reasoning abilities of LLM agents. *arXiv preprint arXiv:2406.06613*, 2024.

A. de Wynter and T. Yuan. The thin line between comprehension and persuasion in llms. *arXiv preprint arXiv:2507.01936*, 2025.

J. Duan, R. Zhang, J. Diffenderfer, B. Kailkhura, L. Sun, E. Storchan, A. Tajer, and P.-Y. Chen. GT-Bench: Uncovering the strategic reasoning limitations of LLMs via game-theoretic evaluations. *arXiv preprint arXiv:2402.12348*, 2024.

K. Gandhi, D. Lee, G. Grand, M. Liu, W. C. Weng, A. Rajani, and A. Suhr. Strategic reasoning with language models. *arXiv preprint arXiv:2305.19165*, 2023.

T. GLM, A. Zeng, B. Xu, B. Wang, C. Zhang, D. Yin, D. Rojas, G. Feng, H. Zhao, H. Lai, H. Yu, H. Wang, J. Sun, J. Zhang, J. Cheng, J. Gui, J. Tang, J. Zhang, J. Li, L. Zhao, L. Wu, L. Zhong, M. Liu, M. Huang, P. Zhang, Q. Zheng, R. Lu, S. Duan, S. Zhang, S. Cao, S. Yang, W. L. Tam, W. Zhao, X. Liu, X. Xia, X. Zhang, X. Gu, X. Lv, X. Liu, X. Liu, X. Yang, X. Song, X. Zhang, Y. An, Y. Xu, Y. Niu, Y. Yang, Y. Li, Y. Bai, Y. Dong, Z. Qi, Z. Wang, Z. Yang, Z. Du, Z. Hou, and Z. Wang. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024.

A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Z. Guan, X. Liu, W. Su, Y. Zhang, B. Li, and Y. Xie. Richelieu: Self-evolving LLM-based agents for AI Diplomacy. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024.

D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

S. H. Huang, K. Bhatia, P. Abbeel, and A. D. Dragan. Establishing appropriate trust via critical states. In *2018 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 3929–3936. IEEE, 2018.

Y. Huang, X. Xie, Y. Chen, D. Liao, and F. Wu. DipLLM: Fine-tuning LLM for strategic decision-making in Diplomacy. *arXiv preprint arXiv:2506.09655*, 2024.

J. Kang, Q. Tong, J.-J. Cai, T. He, Y. Liang, M. de Rijke, Y. Mei, Y. Wen, and Y. Liu. GTBench: Uncovering the strategic reasoning limitations of LLMs via game-theoretic evaluations. *arXiv preprint arXiv:2402.12348*, 2024.

J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Kimi, Y. Bai, Y. Bao, G. Chen, J. Chen, N. Chen, R. Chen, Y. Chen, Y. Chen, Y. Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

J. Light, M. Cai, S. Shen, and Z. Hu. AvalonBench: Evaluating LLMs playing the game of Avalon. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

N. Lorè and B. Heydari. Strategic behavior of large language models and the role of game structure versus contextual framing. *Scientific Reports*, 14(1):18492, 2024.

L. Malmqvist. Sycophancy in large language models: Causes and mitigations. *arXiv preprint arXiv:2411.15287*, 2024.

Meta AI. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. *https://ai. meta. com/blog/llama-4-multimodal-intelligence/, checked on*, 4(7):2025, 2025.

Mistral AI. Devstral. *Technical Blog*, 2025a. URL `https://mistral.ai/news/devstral`.

Mistral AI. Medium is the new large. *Technical Blog*, 2025b. URL https://mistral.ai/news/mistral-medium-3.

Mistral AI. Mistral small 3.1. *Technical Blog*, 2025c. URL https://mistral.ai/news/mistral-small-3-1.

OpenAI. Introducing gpt-4.1 in the api. *Technical Blog*, 2025a.

OpenAI. Introducing o3 and o4-mini. *Technical Blog*, 2025b.

P. Paquette. Diplomacy: DATC-compliant game engine with web interface. https://github.com/diplomacy/diplomacy, 2020. Version 1.1.2, accessed 1 August 2025.

K. Payne and B. Alloui-Cros. Strategic intelligence in large language models: Evidence from evolutionary game theory. *arXiv preprint arXiv:2507.02618*, 2025.

Qwen Team. Qwq-32b: Embracing the power of reinforcement learning. March 2025. URL https://qwenlm.github.io/blog/qwq-32b/.

B. Savani. Distilbert model fine-tuned for emotion classification (distilbert-base-uncased-emotion). https://huggingface.co/bhadresh-savani/distilbert-base-uncased-emotion, 2021.

W. Wongkamjan, S. D. Akter, Y. Fan, Y. Zhang, G. Mukobi, and N. N. Fong. More victories, less cooperation: Assessing Cicero's Diplomacy play. *arXiv preprint arXiv:2406.04643*, 2024.

xAI. Grok 4. *Technical Blog*, 2025. URL https://x.ai/news/grok-4.

A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.