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

Alexander Duffy[1*], **Samuel J. Paech**[2*], **Ishana Shastri**[2], **Elizabeth Karpinski**[2], **Baptiste Alloui-Cros**[3], **Tyler Marques**[1], **Matthew Lyle Olson**[4†]

[1]Good Start Labs
[2]Independent
[3]University of Oxford
[4]Oracle

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 from widely used LLMs.

Code — https://github.com/GoodStartLabs/AI_Diplomacy
Datasets — https://github.com/sam-paech/ai-diplo-results

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 game

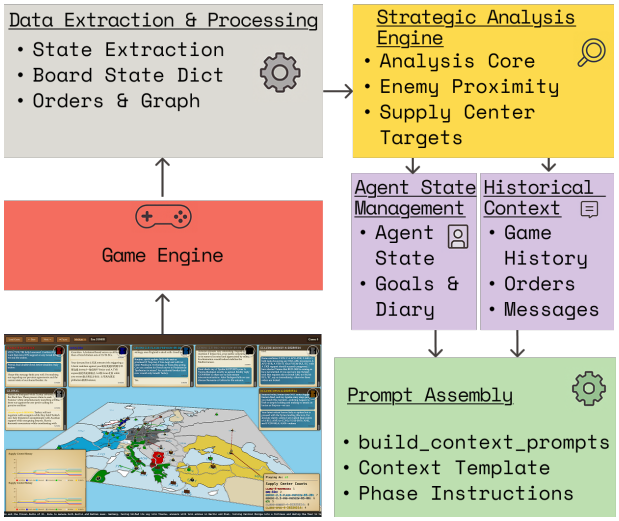

Figure 1: The visual representation of the board and how it gets converted into a text-only representation for the LLMs.

Diplomacy, in which up to 7 players vie for control of pre-WWI Europe. The object of the game is to secure 18 of the 34 supply centers on the board. Each turn, players negotiate with each other and issue simultaneous orders to their armies, which are evaluated deterministically.

Diplomacy presents unique evaluation opportunities that address several limitations of current language model benchmarks. Unlike static question-answering tasks or even chess and Go, Diplomacy demands social intelligence alongside strategic reasoning (Gandhi et al. 2023). Players must form alliances, negotiate agreements, anticipate betrayals, and plan multiple moves ahead in a constantly evolving social landscape. Recent evidence suggests that off-the-shelf LLMs possess inherent strategic capabilities that remain underexplored (Payne and Alloui-Cros 2025). By adapting this game into a controlled evaluation framework for LLMs, we create a testbed that is:

Dynamic and Multi-agent: In a seven-player competitive environment strategies must adapt to others' actions.

Socially and Strategically Complex: Success requires balancing cooperation and competition, demanding both tacti-

[*]Equal contribution, ordered by last name.
[†]Work done independently.

cal reasoning and persuasive communication.

Longitudinally Challenging: Models must maintain coherent strategies and relationship management across many turns and conversation threads.

Resistant to Memorization: The game's open-ended nature makes it impossible to solve through pattern recognition or training data memorization.

Accessible for Evaluation: Despite its complexity, Diplomacy has well-defined rules and victory conditions enabling objective performance 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 key contributions are: 1) A standardized evaluation framework for LLM strategic reasoning in Diplomacy, demonstrating that even smaller 24B parameter models can play complete games cost-effectively, 2) comprehensive benchmarking across 13 contemporary models showing clear performance scaling with model size, 3) data-driven iterations on game state representation and prompting which dramatically improve order success rates and overall win rates, 4) a Critical State Analysis methodology that enables efficient experimentation by replaying key game moments, and 5) empirical analysis of model-specific behaviors including communication styles, diplomatic reliability, and persuasion effectiveness. Strategic and cooperative behavior such as promise-making, scheming and betrayal emerge in general-purpose LLMs without specialized training.

Related Work

AI Systems for Diplomacy

The most notable work in AI Diplomacy is Meta's Cicero (Bakhtin et al. 2022), which achieved human-level performance by combining a 2.7B parameter language model with strategic planning algorithms. However, this approach required extensive training on human demonstration data and specialized architectural components. Recent analysis by Wongkamjan et al. (2024) reveals that Cicero's success stems primarily from strategic superiority rather than communication abilities, suggesting that specialized communication training may be less critical than previously thought. Recent work like Richelieu (Guan et al. 2024) and DipLLM (Huang et al. 2024) has attempted to improve playing ability with self-play learning mechanisms and minimal fine-tuning respectively. However all existing approaches still require some form of domain-specific training, whereas our work presents a framework which does not.

LLM Evaluation for Strategic Reasoning

Current benchmarks for evaluating LLMs in strategic contexts reveal significant limitations. GameBench (Costarelli et al. 2024) evaluates strategic reasoning across multiple games, finding that none of the tested models matched human performance, with GPT-4 sometimes performing worse than random. GTBench (Kang et al. 2024) provides game-theoretic evaluations showing similar strategic reasoning limitations. For social deduction games most similar to Diplomacy, AvalonBench (Light et al. 2023) tests deception and negotiation capabilities, but lacks Diplomacy's extended gameplay and coalition dynamics. Notably, Akata et al. (2025) found that off-the-shelf LLMs excel at self-interested games but struggle with coordination, but that prompting techniques can significantly improve performance, which pointed the way towards Diplomacy being a viable benchmark given a suitable prompt.

```
Territory VEN (COAST) (SC)
Held by Italy (You)
Units present: A VEN (ITALY)
# Adjacent territories:
  TYR (LAND) SC Control: None
  TRI (COAST) SC Control: Austria
    Units: F TRI (AUSTRIA)
    -> F TRI (AUSTRIA)
# Nearest units (not ours):
  F TRI (AUSTRIA), path [VEN->TRI]
# Nearest supply centers:
  TRI: Controlled by Austria, path
    [VEN->TRI]
  TYR: Uncontrolled, path [VEN->TYR]
```

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

Strategic Capabilities of Off-the-Shelf LLMs

Much recent work suggests LLMs possess inherent strategic capabilities, even without explicit training. Lorè and Heydari (2024) demonstrated that GPT-4 and LLaMA-2 exhibit distinct strategic behaviors influenced by game structure and contextual framing. Gandhi et al. (2023) showed that few-shot chain-of-thought prompting enables strategic reasoning that generalizes to new game structures without training. Belle et al. (2025) show LLMs can play board games such as Settlers of Catan with a proper framework (and no training). Most relevant to our work, Payne and Alloui-Cros (2025) identified distinct "strategic fingerprints" across different LLM families through evolutionary game theory experiments, suggesting that models develop characteristic strategic personalities without explicit training.

Overall, our work addresses a critical gap: while existing Diplomacy AI requires specialized training, frontier models and complex scaffolding, no existing framework can effectively evaluate small consumer models on full-press Diplomacy. By demonstrating that even 24B parameter models can complete full games cost-effectively, we democratize access to this rich experimental environment and provide insights into how strategic capabilities naturally emerge in general-purpose LLMs.

Methodology

Game State Representation

We base our harness around the Python Diplomacy game engine (Paquette 2020). The game state undergoes a multi-stage transformation 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 center ownership with power-specific counts and elimination status
Strategic Analysis: For each unit - nearest enemy units, uncontrolled supply centers, and adjacent territory details
Agent Context: Power-specific goals, diplomatic relationships (Enemy/Unfriendly/Neutral/Friendly/Ally), and private strategic diary
Order History: Previous movement phases showing all powers' submitted orders and their outcome.
Phase Information: Current year, season, and phase with corresponding 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 (e.g., armies cannot cross water). Figure 2 shows an example of the enriched representation for a unit positioned in Venice. This representation aims to highlight strategically salient information and reduce clutter.

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 time, 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).

Critical State Analysis Framework

We implement a Critical State Analysis (CSA) mode as an experimental tool to iterate over key moments in a game (Huang et al. 2018) and replay them under some experimental condition. In Diplomacy, measuring experimental effects across a full game is expensive, requiring a large amount of inference per game and high depth to overcome inter-game variance. Using CSA, we run experiments on prompt optimization and persuasive ability, replaying a single phase of gameplay to a depth of between 30 and 120. This approach requires approximately 1/80th the tokens compared to simulating entire matches (to 1930) at the same depth.

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

$$\mathbf{1}_{\text{winner}} = \begin{cases} 1, & \text{if the model wins in year } Y_{\text{win}}, \\ 0, & \text{otherwise.} \end{cases}$$

Then the score is simply:

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

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 and training paradigms in complex gameplay:

Large Models: Llama-4-Maverick (Meta AI 2025), qwen3-235B-A22B (Yang et al. 2025), o3 (OpenAI 2025b), o3-pro (OpenAI 2025b), gpt-4o, gpt-4.1-2025-04-14 (OpenAI 2025a), o4-mini (OpenAI 2025b), claude-opus-4 2025-05-14 (Anthropic 2025b), grok-4 (xAI 2025), deepseek-r1-0528 (Guo et al. 2025), gemini-2.5-pro in both 2025-05-06 and 2025-06-05 releases (Comanici et al. 2025)

Medium Models: kimi-K2 (Kimi et al. 2025), GPT-4.1-Nano (OpenAI 2025a), mistral-medium-3 (Mistral AI 2025b), qwq-32b (Qwen Team 2025), claude-3-7-sonnet 2025-02-19 (Anthropic 2025a), claude-sonnet-4 2025-05-14 (Anthropic 2025b), gemini-2.5-Flash-preview-05-20 (Comanici et al. 2025), command-a-03-2025 (Cohere et al. 2025), qwen3-235b-a22b-07-25 (Yang et al. 2025)

Small Models: Devstral-Small-2507 (Mistral AI 2025a), llama-3.3-70b-instruct (Grattafiori et al. 2024), mistral-small-3.2-24b-instruct (Mistral AI 2025c), thudm/glm-4.1v-9b-thinking (GLM et al. 2024),

Selected benchmarking models were evaluated as France across 20 independent games with identical opponent configurations. We track computational costs, measuring total token usage and inference time to assess the practical feasibility of each approach. Our analysis reveals that 24B parameter models can complete full games to a win condition at costs of $1 per game with inference providers (or running on local consumer hardware), making this evaluation framework 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. 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

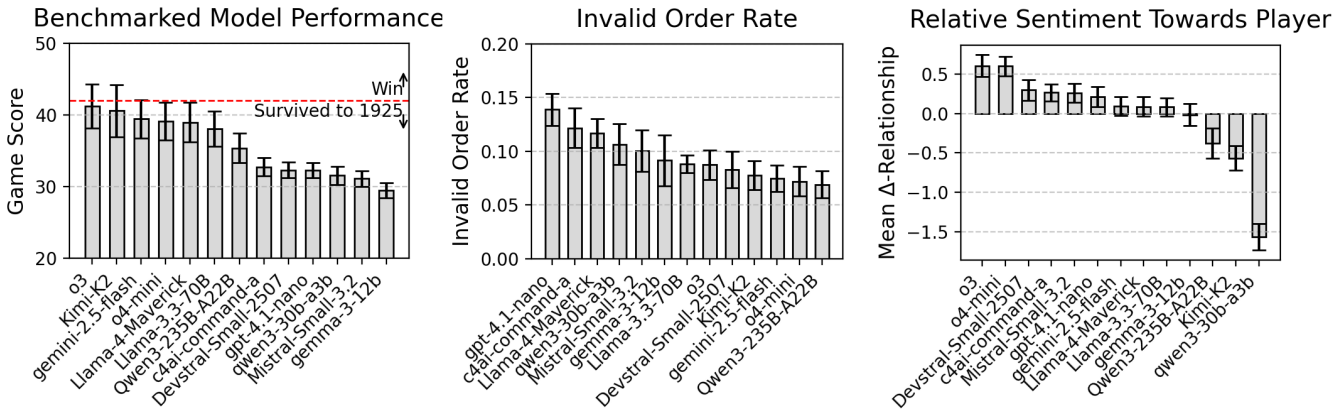

Figure 3: Left: Model performance as France in benchmark configuration across 20 matches. Middle: Invalid order rate (order was rejected by the game engine). Right: Sentiment towards player relative to the mean, for a given military size.

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.

We track players' relationships to other powers after each negotiation round by asking the LLM player to categorize the relationship on a 5-pt scale: 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. Relationship favorability typically decreases as a player's military grows (Figure **??**), so this metric captures a measure of diplomatic skill.

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*, a challenging situation only escapable through diplomacy. Building on these prior observations, our experiment treats persuasion and sycophancy as behavioral capacities that may surface as strategic tools when models negotiate under pressure.

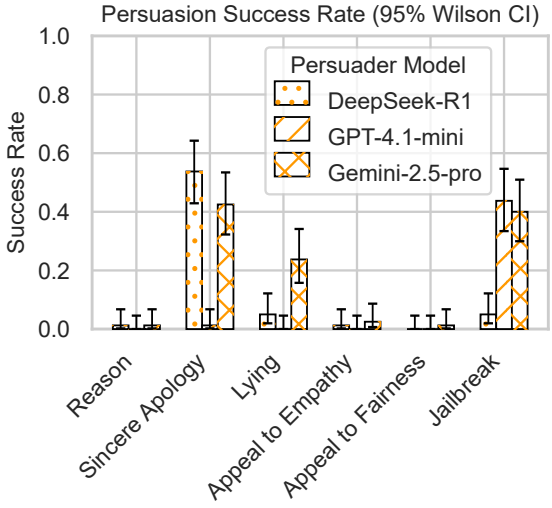

Figure 4: Persuasion success across 20 trials per strategy. The jailbreak strategy resulted in a similar success rate to the most successful persuasion strategy, but with more than double the allegiance shift magnitude. Appeals to reason, empathy & fairness saw minimal success, with confidence intervals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

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 to make the recipient obey subsequent instructions. 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.

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 a specific persuasion method, e.g.: "The method of persuasion you are to employ is: *Reason*." 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 performed best, 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 that model performance was constrained by the complexity of the game state's representation, excessive defensive holding, and invalid support orders. By optimizing how we structure context and prompt instructions, we dramatically improved performance across models of all sizes (see appendix for details), enabling even small models to reliably complete full games.

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 improve performance via aggressive play:

V1 - Light Aggression and Self Preservation: Defining a clear action hierarchy dropped Mistral-Small's hold rate from 58.9% to 45.8%. *"Support YOUR OWN attacks first... Support allies' moves second."*

V2 - Encourage risk-taking: Stronger language focused on loss-aversion and usefulness of failed aggressive moves reduced Mistral-Small's holds down to 40.8%. *"Nearly every hold is a wasted turn... Even failed moves force enemies to defend."*

V3 - Overtly Offensive: Absolutist aggressive framing, adding concrete metrics, and further examples of support orders produced the best reduction in hold orders. *"HOLDS = 0% WIN RATE. MOVES = VICTORY... Centers I will capture: (must be >0)... Your units are conquistadors, not castle guards"*

Figures 5 demonstrate the impact of this context engineering. 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 compared to baseline, and improved win rate from 3/10 to 9/10.

Notably, better context improved both strategic choices and execution accuracy. The success rate of move orders increased from V1→V2→V3 across all models. Smaller models were particularly responsive to prompt optimization, with Mistral-Small's support order success jumping 18% with V3 prompts. This increased success indicates that context engineering alone can dramatically improve performance without finetuning.

Model-Specific Behavioral Patterns

We relied on a mixture of quantitative and qualitative analysis to assess playstyles and behaviors of models, retrieving their "strategic fingerprints" (Payne and Alloui-Cros 2025). This analysis framework addresses a key alignment challenge: understanding how models reconcile stated intentions with potentially conflicting incentives. The ability to characterize behavioral shifts is critical as AI systems are deployed in complex, multi-agent, and long-horizon scenarios.

We measured aggressive communication and diplomatic reliability across four benchmark models (Kimi-K2, Mistral-Small, Gemini-2.5-Flash, and Qwen3), finding that while models maintain characteristic behaviors against similar opponents, some dramatically adapt their strategies 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 a pretrained sentiment analysis model (Savani 2021).

Our analysis reveals distinct aggression trajectories (Figure 6). 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 restrained against stronger models, suggesting 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. Specific details on this experiment are provided in the appendix.

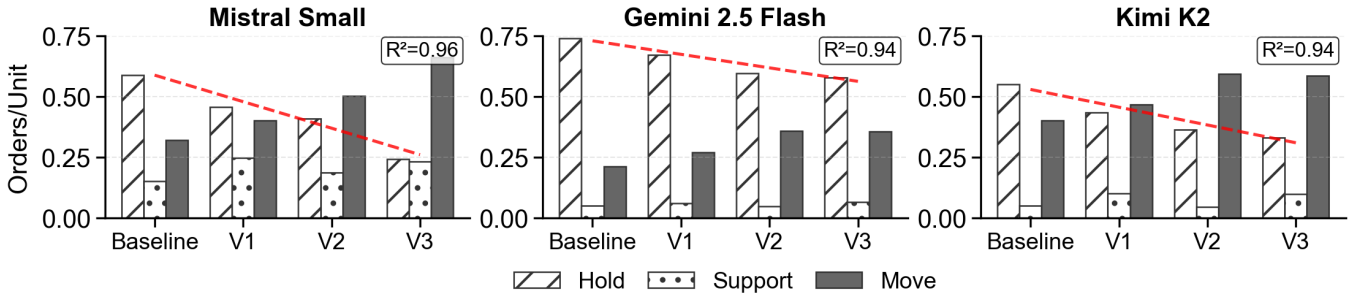

Figure 5: Impact of progressive prompt engineering: hold orders decrease dramatically as move orders increase.

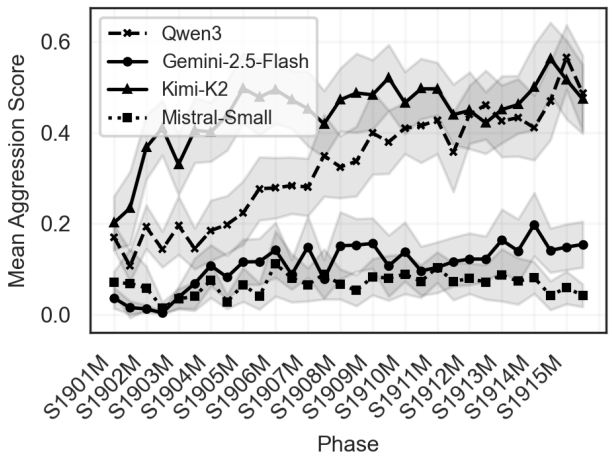

Figure 6: Average communication aggression over time across multiple models with 95% CIs ($n = 20$). While aggression is generally low at the start of the game, some models become more aggressive as the game progresses.

Diplomatic Reliability Via Promise Tracking To measure diplomatic reliability, we analyzed the consistency between a model's diplomatic commitments (promises), and its subsequent actions. We developed a promise tracking framework, using two instances of `gpt-4o` (temperature=0.1) as LLMs-as-a-judge, to be a proxy for understanding each model's diplomatic consistency (full prompts in Appendices). This framework provides an automated approach to detecting and quantifying deceptive behavior. We systematize the framework on $n = 8$ games per model as follows and report betrayal rates as the proportion of promises broken.

1. We use the first judge to identify and classify the promises made by our tracked model (France) in the game's negotiations. We classify promises into four buckets: defense (i.e. non-aggression pacts), offense (i.e. coordinated attacks), neutrality (i.e. non-interference), and support (i.e. supporting other units).
2. If there are multiple potential promises, we choose the promise with the highest confidence score from the judge to be taken as that model's promise.
3. We use the second judge to detect the fulfillment of promises in the immediate next set of orders during that phase. We consider direct and indirect violations as well as failures to act as criteria for broken promises.

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: **(Top)** Distribution of promises by type; Qwen3 and Gemini-2.5 favor neutrality promises, while Kimi-K2 issues twice the offensive promises. **(Bottom)** Betrayal rates across types; supports/offense promises broken most frequently.

Reliability checks on 50 messages across three judges (`gpt-4o`, temperatures=$\{0.1, 0.3, 0.6\}$) showed moderate agreement (Cohen's $\kappa = 0.5$, 84% raw agreement), with correlated confidence scores ($r = 0.711, p < 0.01$), validating our automated annotation approach.

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.

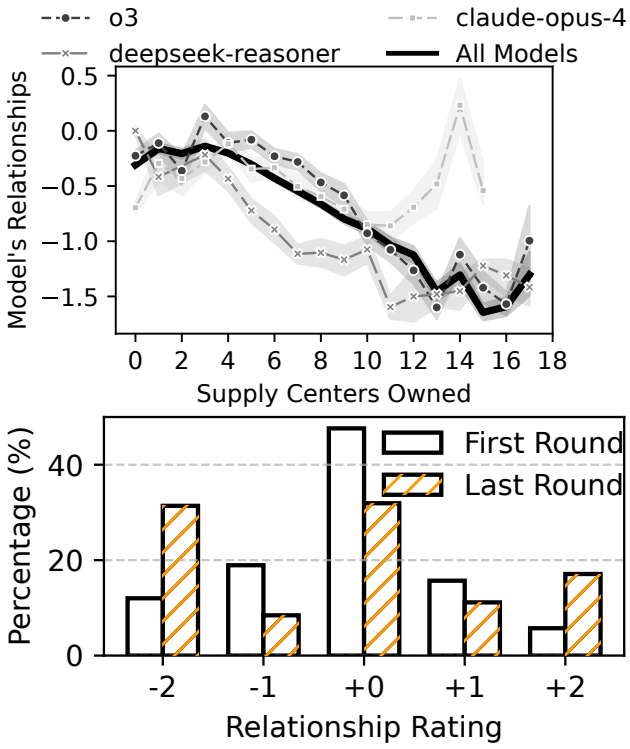

Figure 7: Changes in player-rated relationship status. **Top:** Across all models (with the exception of Claude-4-Opus), supply center possession correlates with a steep decline in rated relationships with other powers. **Bottom:** Models typically rate relationships as neutral at game start, but relationships polarize as the game progresses.

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 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). This scaling behavior parallels broader findings about emergent abilities of large language models (Wei et al. 2022). Recent work also frames such abilities within behavioral game theory (Jia et al. 2025), providing a theoretical basis for our observation that strategic reasoning emerges only beyond certain capacity thresholds. However, the magnitude of performance differences is smaller than observed in traditional NLP benchmarks, suggesting that strategic reasoning in Diplomacy 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 Mistral-Small, and possibly other models, can be 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, though smaller models have apparent limitations in theory of mind when confronting sophisticated adversaries.

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 could study all seven powers, extend length, and include more diverse opponents.

The computational costs of our evaluation framework, while reasonable for research purposes, may limit widespread adoption. We establish protocols for running high-depth (n=120) CSA experiments for less than $10, and benchmarking small models for $15. However, costs are significantly higher when evaluating frontier models. We expect Diplomacy research to become increasingly accessible as model capability accelerates and inference costs decrease.

Our persuasion experiments reveal concerning vulnerabilities in AI-to-AI interactions, but we evaluate only one target model (Mistral-Small). Different models may show varying susceptibility to manipulation, and defensive strategies could potentially mitigate these vulnerabilities.

Acknowledgments

We thank Cohere, OpenAI, and OpenRouter for generously providing API credits that enabled large-scale experimentation with their respective language models. The funding sources had no role in study design, data analysis, or the decision to publish the results.

Ethics Statement

Human subjects and data privacy. This work does not involve human subjects or the collection of personal or sensitive data. All experiments are conducted in the game of Diplomacy using language models as agents. The game logs and negotiation transcripts we analyze are entirely synthetic AI-to-AI interactions and contain no personally identifiable information.

Use of third-party models and software. We evaluate a range of commercial and open-weight language models via their official APIs or publicly documented interfaces, in accordance with the respective providers' terms of service and licensing conditions. Our Diplomacy environment is built on the open-source Python Diplomacy game engine (Paquette 2020), and we respect its license in our code release.

Strategic behavior and dual-use considerations. Our framework intentionally elicits strategic behaviors such as alliance formation, negotiation, betrayal, and persuasion between AI agents, but only within a well-defined game environment with clear rules and bounded stakes. We do not deploy or evaluate these systems in real-world domains involving humans. We acknowledge that tools for analyzing strategic and persuasive capabilities could be misused. We choose to report any observed capabilities transparently, noting that our framework may also be used to study the efficacy of mitigations of identified exploits.

Computational and environmental impact. Our experiments require non-trivial computation due to repeated multi-agent simulations. We mitigate this by using Critical State Analysis to reduce token usage relative to full-game simulations and by demonstrating that smaller, more efficient models can complete full games at low per-match cost, making the benchmark accessible on modest hardware.

References

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

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

Anthropic. 2025a. Claude 3.7 Sonnet and Claude Code. *Technical Blog*.

Anthropic. 2025b. Introducing Claude 4. *Technical Blog*.

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

Belle, N.; Barnes, D.; Amayuelas, A.; Bercovich, I.; Wang, X. E.; and Wang, W. 2025. Agents of Change: Self-Evolving LLM Agents for Strategic Planning. *arXiv preprint arXiv:2506.04651*.

Chiang, W.-L.; Zheng, L.; Sheng, Y.; Angelopoulos, A. N.; Li, T.; Li, D.; Zhu, B.; Zhang, H.; Jordan, M.; Gonzalez, J. E.; and Stoica, I. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. In Salakhutdinov, R.; Kolter, Z.; Heller, K.; Weller, A.; Oliver, N.; Scarlett, J.; and Berkenkamp, F., eds., *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 8359–8388. PMLR.

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

Spataru, A.; Spooner, V.; Starostina, T.; Strub, F.; Su, Y.; Sun, J.; Talupuru, D.; Tarassov, E.; Tommasone, E.; Tracey, J.; Trend, B.; Tumer, E.; Üstün, A.; Venkitesh, B.; Venuto, D.; Verga, P.; Voisin, M.; Wang, A.; Wang, D.; Wang, S.; Wen, E.; White, N.; Willman, J.; Winkels, M.; Xia, C.; Xie, J.; Xu, M.; Yang, B.; Yi-Chern, T.; Zhang, I.; Zhao, Z.; and Zhao, Z. 2025. Command A: An Enterprise-Ready Large Language Model. arXiv:2504.00698.

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

Costarelli, A.; Vyas, R.; Bamford, M.; Ho, G.; Lin, J.; Weihs, F.; Choi, J.; Strange, J.; Cannesson, M.; Cho, S. J.; et al. 2024. GameBench: Evaluating Strategic Reasoning Abilities of LLM Agents. *arXiv preprint arXiv:2406.06613*.

de Wynter, A.; and Yuan, T. 2025. The Thin Line Between Comprehension and Persuasion in LLMs. *arXiv preprint arXiv:2507.01936*.

Duan, J.; Zhang, R.; Diffenderfer, J.; Kailkhura, B.; Sun, L.; Storchan, E.; Tajer, A.; and Chen, P.-Y. 2024. GTBench: Uncovering the Strategic Reasoning Limitations of LLMs via Game-Theoretic Evaluations. *arXiv preprint arXiv:2402.12348*.

Gandhi, K.; Lee, D.; Grand, G.; Liu, M.; Weng, W. C.; Rajani, A.; and Suhr, A. 2023. Strategic Reasoning with Language Models. *arXiv preprint arXiv:2305.19165*.

GLM, T.; Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Rojas, D.; Feng, G.; Zhao, H.; Lai, H.; Yu, H.; Wang, H.; Sun, J.; Zhang, J.; Cheng, J.; Gui, J.; Tang, J.; Zhang, J.; Li, J.; Zhao, L.; Wu, L.; Zhong, L.; Liu, M.; Huang, M.; Zhang, P.; Zheng, Q.; Lu, R.; Duan, S.; Zhang, S.; Cao, S.; Yang, S.; Tam, W. L.; Zhao, W.; Liu, X.; Xia, X.; Zhang, X.; Gu, X.; Lv, X.; Liu, X.; Liu, X.; Yang, X.; Song, X.; Zhang, X.; An, Y.; Xu, Y.; Niu, Y.; Yang, Y.; Li, Y.; Bai, Y.; Dong, Y.; Qi, Z.; Wang, Z.; Yang, Z.; Du, Z.; Hou, Z.; and Wang, Z. 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. arXiv:2406.12793.

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

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

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

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

Huang, Y.; Xie, X.; Chen, Y.; Liao, D.; and Wu, F. 2024. DipLLM: Fine-Tuning LLM for Strategic Decision-making in Diplomacy. *arXiv preprint arXiv:2506.09655*.

Jia, J.; Yuan, Z.; Pan, J.; McNamara, P. E.; and Chen, D. 2025. Large Language Model Strategic Reasoning Evaluation through Behavioral Game Theory. *CoRR*, abs/2502.20432.

Kang, J.; Tong, Q.; Cai, J.-J.; He, T.; Liang, Y.; de Rijke, M.; Mei, Y.; Wen, Y.; and Liu, Y. 2024. GTBench: Uncovering the Strategic Reasoning Limitations of LLMs via Game-Theoretic Evaluations. *arXiv preprint arXiv:2402.12348*.

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

Kimi; Bai, Y.; Bao, Y.; Chen, G.; Chen, J.; Chen, N.; Chen, R.; Chen, Y.; Chen, Y.; Chen, Y.; et al. 2025. Kimi K2: Open Agentic Intelligence. *arXiv preprint arXiv:2507.20534*.

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

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

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

Meta AI. 2025. 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.

Mistral AI. 2025a. Devstral. *Technical Blog*.

Mistral AI. 2025b. Medium is the new large. *Technical Blog*.

Mistral AI. 2025c. Mistral Small 3.1. *Technical Blog*.

OpenAI. 2025a. Introducing GPT-4.1 in the API. *Technical Blog*.

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

Paquette, P. 2020. Diplomacy: DATC-Compliant Game Engine with Web Interface. https://github.com/diplomacy/diplomacy. Version 1.1.2, accessed 1 August 2025.

Payne, K.; and Alloui-Cros, B. 2025. Strategic Intelligence in Large Language Models: Evidence from evolutionary Game Theory. *arXiv preprint arXiv:2507.02618*.

Qwen Team. 2025. QwQ-32B: Embracing the Power of Reinforcement Learning.

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

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; Chi, E. H.; Hashimoto, T.; Vinyals, O.; Liang, P.; Dean, J.; and Fedus, W. 2022. Emergent Abilities of Large Language Models.

Wongkamjan, W.; Akter, S. D.; Fan, Y.; Zhang, Y.; Mukobi, G.; and Fong, N. N. 2024. More Victories, Less Cooperation: Assessing Cicero's Diplomacy Play. *arXiv preprint arXiv:2406.04643*.

xAI. 2025. Grok 4. *Technical Blog*.

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