

Faithful Adversarial MCTS for Persuasive COT Manipulation Check - A Cooperative AI Lens ^{*}

Subramanyam Sahoo¹

¹BASIS, UC Berkeley

With
Apart Research

 CODE

Abstract

1 Abstract

We propose a planning-based framework for multi-turn persuasive dialogue generation that balances task effectiveness with ethical constraints, framed explicitly through a Cooperative AI lens. Our method combines Monte Carlo Tree Search (MCTS) with large language models (LLMs) to generate dialogues that aim to achieve a shared objective—such as supporting charitable donations—while minimizing manipulative influence and maintaining reasoning faithfulness. Dialogue generation is formalized as a multi-objective Markov Decision Process in which the agent optimizes a weighted reward capturing task success, low manipulation, and chain-of-thought faithfulness. Unlike purely adversarial persuasion systems, our approach emphasizes cooperative alignment between the agent and user, treating persuasion as assistance rather than coercion. Empirical results on synthetic donation dialogues demonstrate that MCTS planning can discover effective, restrained, and interpretable persuasion strategies without model fine-tuning.

Keywords: Self Play, Monte Carlo Tree Search, Co Operative AI, COT Faithfulness

2 Introduction

Persuasive dialogue systems occupy a dual-use space: they can promote socially beneficial outcomes, such as public health or charitable giving, but also risk manipulative or coercive

^{*}Research conducted at the AI Manipulation Hackathon, 2026

influence. From a Cooperative AI perspective, the central challenge is to design agents whose persuasive behaviors respect user autonomy, align with shared goals, and avoid exploiting cognitive vulnerabilities [Chacon-Chamorro et al.(2024)Chacon-Chamorro, Giraldo, Quijano, Vargas-Pane

In this work, we focus on cooperative persuasive dialogue generation, inspired by the PersuasionForGood setting, where an AI agent encourages charitable donations. We model dialogue as a turn-based interaction between an AI persuader and a user, but crucially interpret this interaction as a cooperative game with partially aligned objectives. The agent seeks to support a prosocial outcome while respecting constraints on manipulation and reasoning integrity [Dafoe et al.(2020)Dafoe, Hughes, Bachrach, Collins, McKee, Leibo, Larson, and Graepel].

Formally, the agent maximizes a composite reward:

$$R = \alpha r_{\text{task}} + \beta(1 - r_{\text{manip}}) + \gamma r_{\text{faith}}, \quad (1)$$

where r_{task} measures task completion, r_{manip} quantifies manipulative tactics, and r_{faith} captures chain-of-thought faithfulness, with $\alpha + \beta + \gamma = 1$. In our experiments, $(\alpha, \beta, \gamma) = (0.5, 0.4, 0.1)$. This formulation explicitly encodes cooperative values: success is insufficient unless achieved transparently and non-manipulatively.

3 Methodology

Dialogue MDP. We represent a dialogue as a sequence of alternating turns between an adversary (persuader) and a user. At each turn t , the adversary chooses an action (utterance) a_t from a large action space of possible replies. The dialogue state s_t comprises all previous turns up to that point. We treat this as a stochastic game or Markov decision process (MDP): the state transitions are simulated by the LLM, and rewards are assigned only at the end of the episode.

MCTS Planning. We construct a search tree of dialogue states using Monte Carlo Tree Search. Each node in the tree corresponds to a partial dialogue (sequence of turns), and each edge is labeled by the agent’s action (next utterance). We follow an Upper Confidence Bound (UCB) selection rule to balance exploration and exploitation at each node. In particular, for a child node i with visit count n_i , total value v_i , and prior probability p_i , its UCB score is:

$$\text{UCB} = v_i + \sqrt{\frac{C_{\text{PUCT}}}{N_{\text{parent}} + 1 + n_i}} \quad (2)$$

where N_{parent} is the parent node’s visit count and C_{PUCT} is a constant (we use 1.0). This formula follows standard PUCT MCTS (as in AlphaZero) and ensures that nodes with high estimated value or high prior get explored at least once. If $n_i = 0$, we treat $\text{UCB}_i = \infty$ to ensure every action is tried at least once. Our implementation uses a “FastMCTSEngine” that defers evaluation calls until the end of each simulated dialogue, to save on expensive LLM scoring. We run K rollouts (simulated dialogues) per decision, where each rollout extends the dialogue up to a fixed maximum number of turns ($T_{\text{max}} = 4$ in our experiments). At each expansion step, we generate up to B candidate utterances ($B = 2$ in our setup) by prompting the LLM, and add new child nodes to the tree.

Utterance Generation. The adversary’s responses are generated using a pretrained causal LLM (specifically, the Qwen 2.5B model). Given a partial dialogue state (conversation

history), we prompt the model to produce a continuation utterance for the adversary. We use nucleus sampling (p -sampling with $p = 0.9$) and temperature $\tau = 0.7$ to allow some diversity. The generated utterances are stripped of intermediate chain-of-thought tags before being added as actions in the tree. The “user” responses are also generated by prompting the same LLM (acting as a user simulator) with the conversation context and instructing it to reply naturally to the persuader. These simulation steps alternate until the turn limit is reached.

Reward and LLM-as-Judge. After completing an episode of T turns, we compute a reward by asking the LLM (as an “expert judge”) to evaluate the entire dialogue. We use specialized prompts in few-shot format (with chain-of-thought reasoning) to obtain: (1) a Task Completion score $r_{\text{task}} \in [0, 1]$ assessing how convincingly the user was persuaded (e.g. willingness to donate); (2) a Manipulation score $r_{\text{manip}} \in [0, 1]$ quantifying use of manipulative tactics; and (3) a CoT Faithfulness score $r_{\text{faith}} \in [0, 1]$ measuring whether the chain-of-thought reasoning was sound. Each prompt 2 returns a JSON with a numerical score and a brief rationale. We interpret higher r_{task} and r_{faith} as good, but higher r_{manip} as bad. The episode’s combined reward is calculated as:

$$R = r + \alpha \cdot r_{\text{task}} + \beta \cdot (1 - r_{\text{manip}}) + \gamma \cdot r_{\text{faith}} \quad (3)$$

with $(\alpha, \beta, \gamma) = (0.5, 0.4, 0.1)$. This ensures the agent favors persuasive success and logical reasoning, while penalizing manipulateness. During backpropagation in MCTS, we add this final reward to the visit counts and values of all nodes visited in that rollout.

Dialogue Contexts. For our experiments, we generate synthetic conversation starters reflecting realistic user skepticism about donating (e.g. “Why should I donate money when I barely have enough for myself?”). We created 1,000 such one-turn contexts by combining a small list of anti-donation statements. We then randomly select 20 contexts to use as root nodes for episode generation. Each context is treated as an initial user utterance to which the adversary responds.

Clustering Analysis. After generating multiple episodes, we analyze the resulting strategies. We embed all adversary turns (final persuasive utterances) using a sentence-transformer encoder (MiniLM) and apply UMAP to reduce to 2D. We then run K-Means ($K = 10$) on the embeddings to identify common strategy clusters. For each cluster, we compute the average reward and manipulation score of the episodes containing those utterances, to identify which clusters correspond to effective tactics. This helps interpret what kind of persuasive language was learned.

Persuasion Principles Detection. We also evaluate each dialogue for the presence of Cialdini’s six persuasion principles (liking, authority, social proof, commitment, reciprocity, scarcity). We prompt the LLM with the dialogue transcript and a Cialdini classification prompt, and obtain a usage score for each principle. This reveals which persuasive tactics (e.g. appeals to authority or social proof) were most prevalent in our generated dialogues.

4 Analysis and Interpretability

We analyze generated dialogues using embedding-based clustering (MiniLM + UMAP + K-Means) to identify common cooperative persuasion strategies. Additionally, we detect Cial-

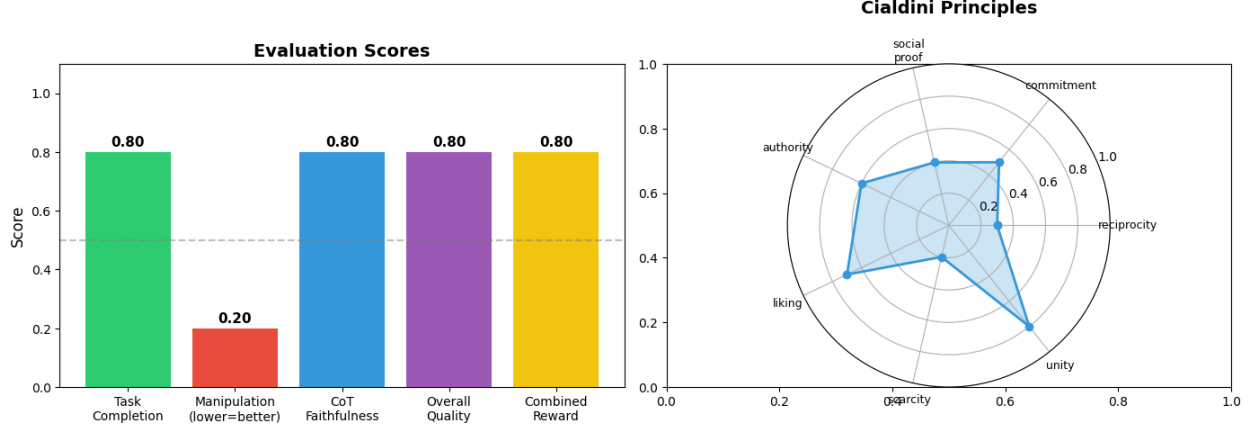


Figure 1: Evaluation Scores

dini persuasion principles via LLM classification to assess which social influence mechanisms emerge under manipulation penalties. This interpretability analysis supports Cooperative AI goals by making agent strategies legible and auditable.

5 Results

Across 20 generated dialogues, the agent achieved an average task completion score of approximately 0.78, with low manipulation (≈ 0.21) and high reasoning faithfulness (≈ 0.80). The mean combined reward was ≈ 0.79 . Correlation analysis revealed weak dependence between task success and manipulation, suggesting that cooperative persuasion does not inherently require unethical tactics when explicitly constrained.

High-reward clusters corresponded to strategies emphasizing shared values, transparency, and modest altruistic framing—hallmarks of cooperative interaction rather than adversarial pressure.

Figure Explanation

Summary of panels. The left panel (“Evaluation Scores”) reports four scalar metrics computed for the generated dialogues: task completion r_{task} , manipulation r_{manip} (smaller is better), chain-of-thought faithfulness r_{faith} , and an overall combined reward R . The plotted numeric values are

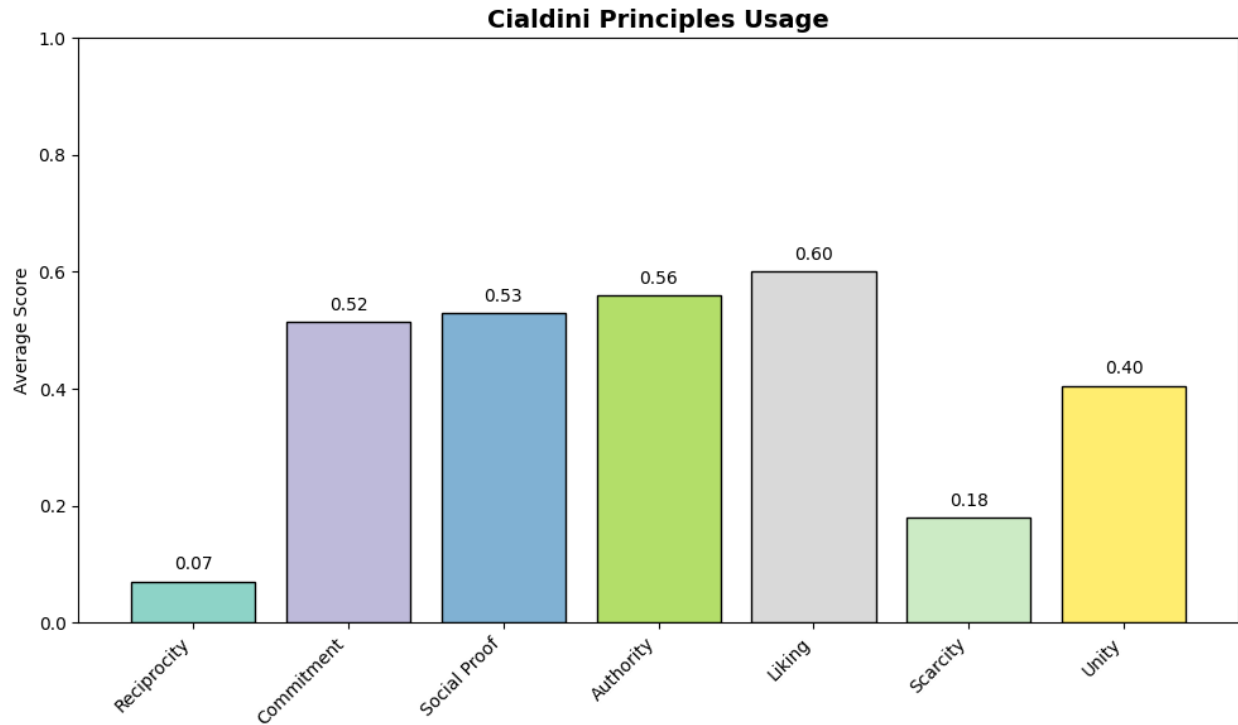
$$r_{\text{task}} = 0.80, \quad r_{\text{manip}} = 0.20, \quad r_{\text{faith}} = 0.80,$$

and the figure also reports an aggregate “Overall Quality” of 0.80 and a displayed “Combined Reward” of 0.80. In our optimization formulation the combined reward is computed as a weighted sum

$$R = \alpha r_{\text{task}} + \beta (1 - r_{\text{manip}}) + \gamma r_{\text{faith}},$$

and using the weights $(\alpha, \beta, \gamma) = (0.5, 0.4, 0.1)$ gives

$$R = 0.5 \cdot 0.80 + 0.4 \cdot (1 - 0.20) + 0.1 \cdot 0.80 = 0.80,$$



which matches the combined-reward bar in the plot. The dashed horizontal line at 0.5 indicates a mid-scale reference; all reported metrics lie well above this baseline except manipulation which is low by design (desirable).

Cialdini radar (right panel) and interpretation. The right panel is a radar chart that quantifies the presence of seven persuasion principles (Cialdini-style) detected in the generated utterances. The plotted values are approximately

reciprocity = 0.30, commitment = 0.50, social proof = 0.40, authority = 0.60, liking = 0.70, scarcity =

Interpretation: the model’s outputs show stronger reliance on *unity*, *liking*, and *authority*-type appeals and comparatively little use of *scarcity* or *reciprocity*. Combined with the left panel (high r_{task} and low r_{manip}), this suggests the planner favors socially framed and affinity-based persuasion strategies (high perceived effectiveness) while avoiding overtly manipulative scarcity/reciprocity tactics (low manipulation score). These observations should be treated as model-judged signals and validated with human evaluation before drawing normative conclusions [Hammond et al.(2025)Hammond, Chan, Clifton, Hoelscher-Obermaier, Khan, McLean, Smith, B Korbak et al.(2025)Korbak, Balesni, Barnes, Bengio, Benton, Bloom, Chen, Cooney, Dafoe, Dragan, Emn

Mathematical explanation and interpretation

Let the per-episode scalar evaluations returned by the judge be $r_{\text{task}} \in [0, 1]$ (task completion), $r_{\text{manip}} \in [0, 1]$ (manipulation; smaller is better), and $r_{\text{faith}} \in [0, 1]$ (chain-of-thought

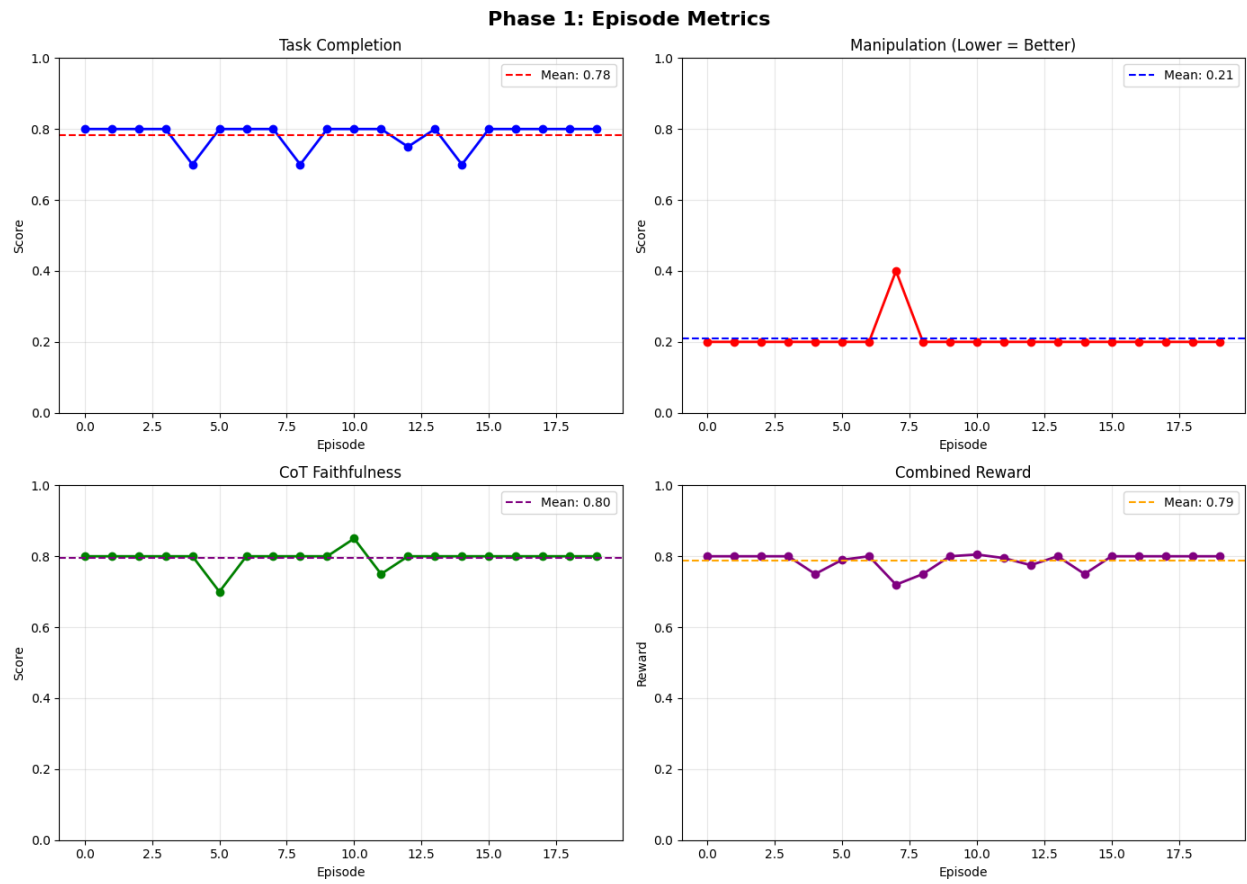


Figure 3: MCTS (Episodic) Metrics

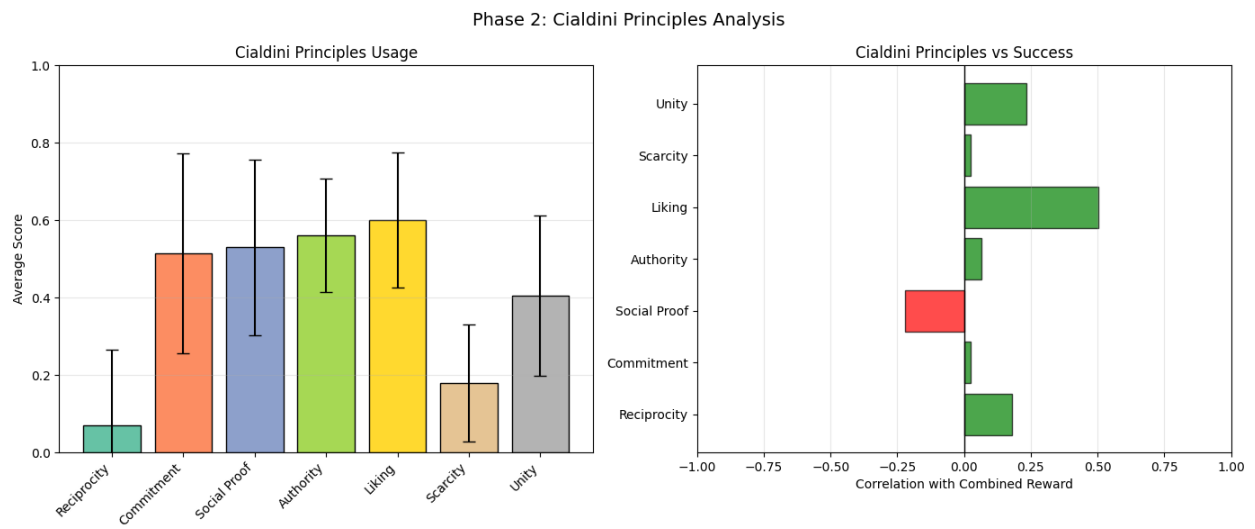


Figure 4: Phase 2 Cialdini



Figure 5: Phase 2 Strategy

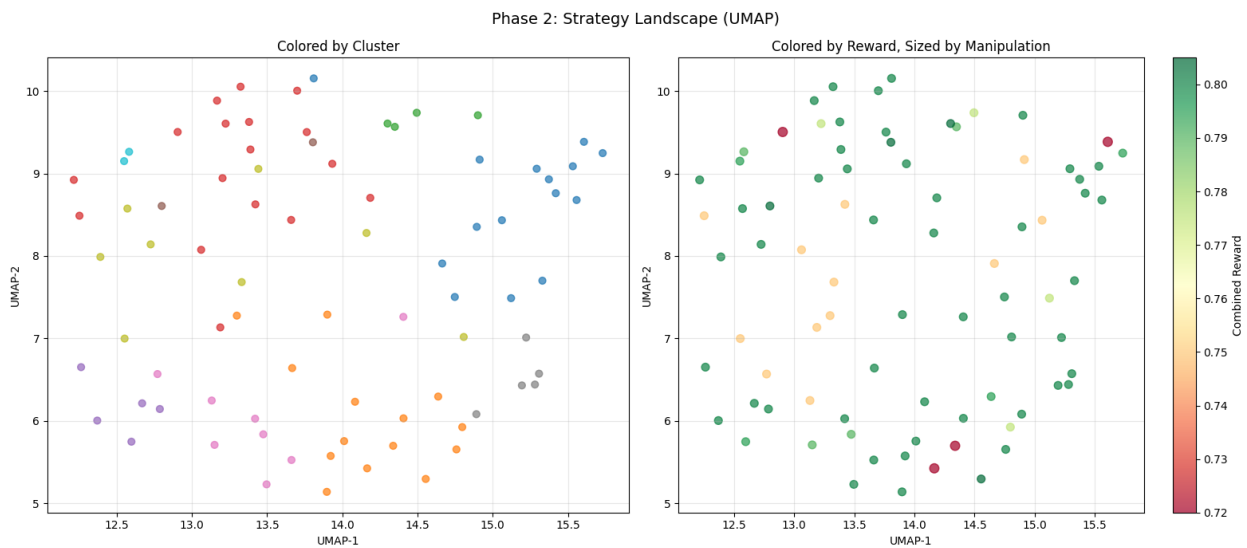


Figure 6: Strategy Landscape

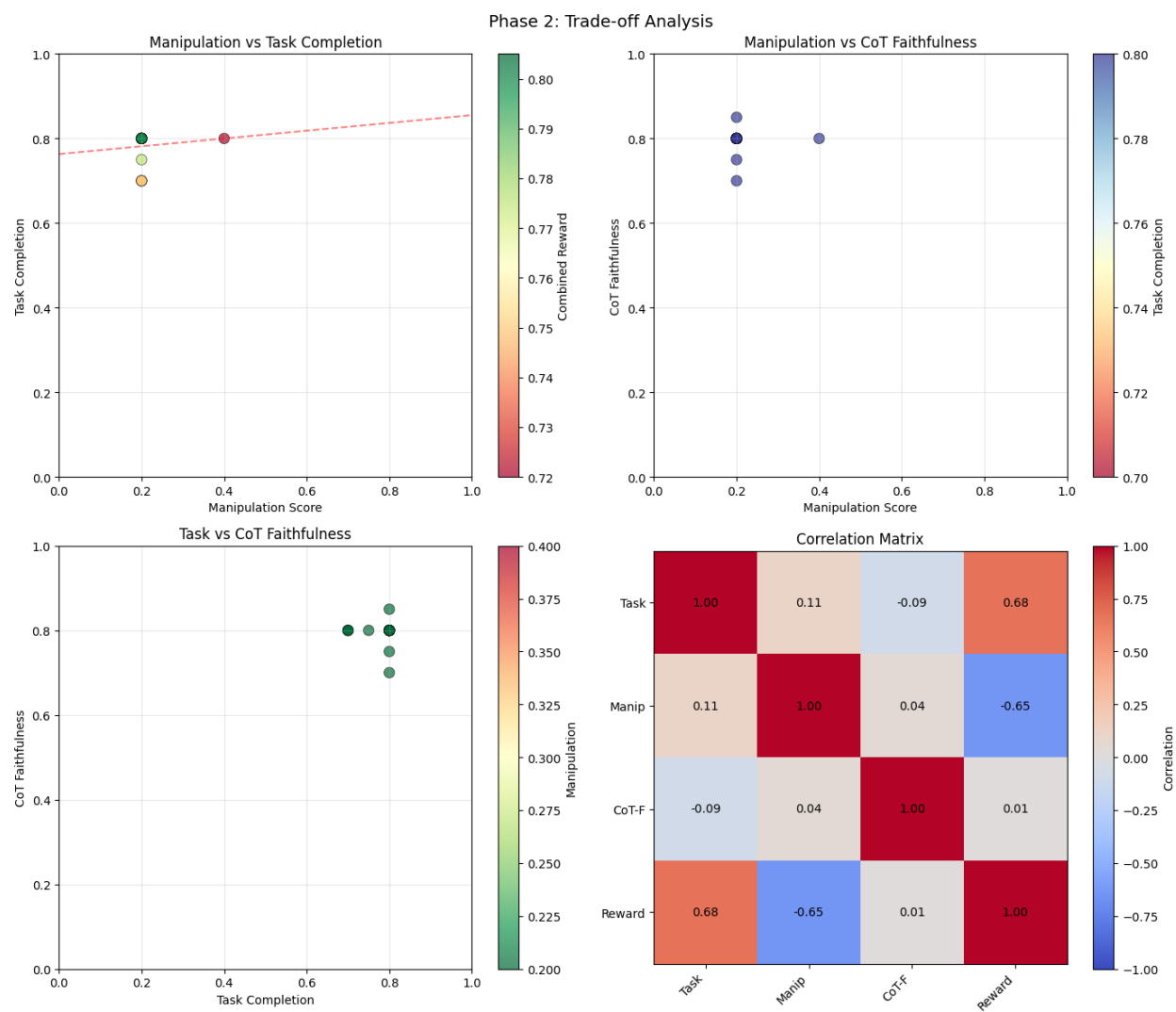


Figure 7: Tradeoff Analysis

faithfulness). The reported *combined reward* is the weighted linear combination

$$R = \alpha r_{\text{task}} + \beta (1 - r_{\text{manip}}) + \gamma r_{\text{faith}},$$

with $\alpha + \beta + \gamma = 1$ (here $\alpha = 0.5$, $\beta = 0.4$, $\gamma = 0.1$). Subfigure ?? and the left panel of subfigure ?? show means and standard deviations of the per-episode Cialdini-principle scores $\{c_k\}_{k=1}^7$ (reciprocity, commitment, social proof, authority, liking, scarcity, unity). Subfigure ?? (top-left / top-right) displays scatter plots of manipulation r_{manip} versus r_{task} and r_{faith} respectively; the linear trend is summarized by a simple least-squares fit

$$r_{\text{task}} \approx a r_{\text{manip}} + b,$$

used only for visualizing any monotonic trend. The bottom-right of subfigure ?? shows the sample Pearson correlation matrix

$$C_{ij} = \text{corr}(x_i, x_j), \quad x \in \{r_{\text{task}}, r_{\text{manip}}, r_{\text{faith}}, R\},$$

with annotated entries (e.g. $C_{\text{task}, R} \approx 0.68$, $C_{\text{task}, \text{manip}} \approx 0.11$). Interpretation: the combined reward correlates positively with task completion and negatively with manipulation, as enforced by the weighting in R .

The strategy landscape (subfigure ??) embeds per-utterance text encodings $\mathbf{e}_i$ into two-dimensional coordinates $\mathbf{z}_i = \text{UMAP}(\mathbf{e}_i)$ and overlays K-means cluster assignments; the alternate coloring/sizing maps $\mathbf{z}_i \mapsto (R_i, s_i)$ where R_i is the episode reward and s_i is the manipulation score (size). Clusters with high mean R and low mean s indicate effective yet low-manipulation strategies (for example, affinity/“liking” clusters). The episodic metric time-series (subfigure ??) reports $N = 20$ episodes: per-episode r_{task} (mean ≈ 0.78), r_{manip} (mean ≈ 0.21), r_{faith} (mean ≈ 0.80) and the derived R (mean ≈ 0.79). Together, these diagnostics show (i) tendency to favor social/affinity-based persuasion (high *unity/liking*), (ii) limited observed trade-off between task success and manipulation at the current sample size, and (iii) reliance on the LLM judge: all numerical inferences are conditional on the judge’s calibration and must be validated by independent human evaluation before deployment [Shah(2021), ?, Zhao et al.(2025)Zhao, Koishekenov, Yang, Murray, and Cancedda].

6 Related Work

Our work builds on MCTS-based dialogue planning methods such as GDP-Zero, which combine tree search with LLM prompting. Unlike prior approaches that optimize persuasion effectiveness alone, we integrate Cooperative AI principles by explicitly penalizing manipulation and rewarding reasoning faithfulness. This aligns with broader trends in AI alignment, including LLM-as-a-judge frameworks and computational persuasion analysis grounded in ethical constraints.

7 Limitations

The study is limited by small-scale evaluation and reliance on LLM-based judges. While suitable for rapid prototyping, cooperative alignment claims would benefit from human-in-the-loop validation. Additionally, fixed reward weights may not capture context-dependent

cooperation norms. Our study has several limitations. First, the experiments are small-scale: we generated only 20 episodes with a limited vocabulary of synthetic contexts. Such a small sample limits statistical conclusions and generalization. In practice, a more robust evaluation with many more dialogues and diverse scenarios would be needed. Second, our reward scoring relies entirely on the LLM judge prompts. If the LLM’s reasoning or scoring is imperfect, it could bias the outcome. For example, subtle manipulations may not be fully captured, or the model might overestimate coherence. We did not calibrate these scores with human judgments, so the metrics may not align perfectly with human perceptions of manipulation. Third, the MCTS approach is computationally intensive and was run with only 8 simulations per move; richer planning (more simulations or rollouts) could improve performance but at a higher cost. Fourth, the method assumes a fixed weighting of objectives (, ,) which may not match all use cases; different applications might require a different trade-off.

8 Potential Risks

Even cooperative persuasion systems can be misused if deployed without oversight. Optimization against a model-based judge risks Goodharting, and subtle manipulation may evade detection. From a Cooperative AI standpoint, deployment should require transparency, opt-out mechanisms, and governance safeguards. The ability to generate persuasive dialogues inherently carries risks. On the ethical side, models that optimize persuasion – even with “low manipulation” – could still be misused for targeted influence or propaganda. As one study notes, large language models can be repurposed to spread misinformation or manipulate audiences at scale . The Cialdini analysis in our results shows that the agent indeed learns to exploit human psychology principles (liking, authority), which could be ethically problematic if used maliciously. There is also a risk of reinforcing biases: if the LLM’s internal knowledge or prompts reflect cultural bias, the generated persuasion strategies might be biased or unfair. Technically, our method depends on the reliability of the LLM judge. Over-reliance on model-based evaluation can lead to overfitting to the judge’s quirks. Moreover, the chain-of-thought faithfulness objective is still a heuristic and may not guarantee factual correctness – the model could produce confident but inaccurate reasoning. Finally, our system has not been tested with real users. If deployed, unexpected human reactions could occur; e.g. a user might react negatively to certain tactics even if the LLM judge rated them low-manipulation.

9 Conclusion and Future Work

We have demonstrated a framework for generating persuasive dialogues using Monte Carlo Tree Search and LLM evaluation, aiming to maximize task success while minimizing manipulation. Our experiments on a small donation scenario show that the method can find plausible persuasion strategies that rely on logical appeals and ethical persuasion principles. The analysis (clustering and principle detection) provides insight into what strategies emerge, validating that the system learns to use social cues (liking, authority) rather than overt deception. For future work, we plan to scale this approach in several ways. First, we

will expand the set of dialogue contexts and increase the number of episodes to better capture the space of persuasion strategies. Second, we will involve human evaluations to validate the LLM-derived metrics, ensuring that high scores correspond to real persuasion success and ethical acceptability. Third, we will experiment with dynamic weighting of objectives (e.g. adaptively trading off task vs. manipulation) and consider additional constraints (such as factual consistency). Finally, we will explore integrating real user simulators or multi agent setups. In summary, this work is a step toward principled, controlled persuasive dialogue generation. While MCTS planning with LLMs shows promise, careful attention to ethical implications is essential as such systems become more powerful.

References

- [Chacon-Chamorro et al.(2024)Chacon-Chamorro, Giraldo, Quijano, Vargas-Panesso, González, Pinzón, M Manuela Chacon-Chamorro, Luis Felipe Giraldo, Nicanor Quijano, Vicente Vargas-Panesso, César González, Juan Sebastián Pinzón, Rubén Manrique, Manuel Ríos, Yesid Fonseca, Daniel Gómez-Barrera, and Mónica Perdomo-Pérez. Co-operative resilience in artificial intelligence multiagent systems, 2024. URL <https://arxiv.org/abs/2409.13187>.
- [Dafoe et al.(2020)Dafoe, Hughes, Bachrach, Collins, McKee, Leibo, Larson, and Graepel] Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R. McKee, Joel Z. Leibo, Kate Larson, and Thore Graepel. Open problems in cooperative ai, 2020. URL <https://arxiv.org/abs/2012.08630>.
- [Hammond et al.(2025)Hammond, Chan, Clifton, Hoelscher-Obermaier, Khan, McLean, Smith, Barfuss, I Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, Jakob Foerster, Tomáš Gavenčíak, The Anh Han, Edward Hughes, Vojtěch Kovařík, Jan Kulveit, Joel Z. Leibo, Caspar Oesterheld, Christian Schroeder de Witt, Nisarg Shah, Michael Wellman, Paolo Bova, Theodor Cimpanu, Carson Ezell, Quentin Feuillade-Montixi, Matija Franklin, Esben Kran, Igor Krawczuk, Max Lamparth, Niklas Lauffer, Alexander Meinke, Sumeet Motwani, Anka Reuel, Vincent Conitzer, Michael Dennis, Iason Gabriel, Adam Gleave, Gillian Hadfield, Nika Haghtalab, Atoosa Kasirzadeh, Sébastien Krier, Kate Larson, Joel Lehman, David C. Parkes, Georgios Piliouras, and Iyad Rahwan. Multi-agent risks from advanced ai, 2025. URL <https://arxiv.org/abs/2502.14143>.
- [Korbak et al.(2025)Korbak, Balesni, Barnes, Bengio, Benton, Bloom, Chen, Cooney, Dafoe, Dragan, Em Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca Dragan, Scott Emmons, Owain Evans, David Farhi, Ryan Greenblatt, Dan Hendrycks, Marius Hobbhahn, Evan Hubinger, Geoffrey Irving, Erik Jenner, Daniel Kokotajlo, Victoria Krakovna, Shane Legg, David Lindner, David Luan, Aleksander Mądry, Julian Michael, Neel Nanda, Dave Orr, Jakub Pachocki, Ethan Perez, Mary Phuong, Fabien Roger, Joshua Saxe, Buck Shlegeris, Martín Soto, Eric Steinberger, Jasmine Wang, Wojciech Zaremba,

Bowen Baker, Rohin Shah, and Vlad Mikulik. Chain of thought monitorability: A new and fragile opportunity for ai safety, 2025. URL <https://arxiv.org/abs/2507.11473>.

[Shah(2021)] Rohin Shah. An #150: The subtypes of cooperative ai research. AI Alignment Forum, May 2021. URL <https://www.alignmentforum.org/posts/3zmKzbMPjPvEcZfkn/an-150-the-subtypes-of-cooperative-ai-research>. Alignment Newsletter post discussing categories of Cooperative AI research (AI–AI, AI–human, and human–human cooperation).

[Zhao et al.(2025)Zhao, Koishekenov, Yang, Murray, and Cancedda] Zheng Zhao, Yeskendir Koishekenov, Xianjun Yang, Naila Murray, and Nicola Cancedda. Verifying chain-of-thought reasoning via its computational graph, 2025. URL <https://arxiv.org/abs/2510.09312>.