MetaMafia - Final Project Report

Joseph Ellis
josephel@usc.edu

Shravya Shashidhar
sshashid@usc.edu

Loring Scotty Hoag
lhoag@usc.edu

Abstract

This project explores the design of agentic, multi-agent AI systems for social deduction games such as Mafia and Werewolf, where reasoning, persuasion, and deception are central. We build a framework in which large language models act as autonomous players, each equipped with memory, planning, and safety guardrails to operate under strict game rules. A moderator agent enforces constraints, while player agents generate dialogue, strategize, and adapt to evolving game states. We evaluate performance through large-scale tournaments, measuring win-rates, persuasion effectiveness, and rule adherence. Beyond entertainment, the project is a controlled testbed for studying reasoning, multi-agent interaction, and alignment challenges in LLMs, with applications to trustworthy AI and safe autonomous decision-making.

1 Introduction

Recent advances in large language models (LLMs) have enabled fluent dialogue and contextual reasoning. However, their behavior in multi-agent social settings where hidden information, deception, and persuasion matter, remains poorly understood. Studying these settings is essential for developing trustworthy autonomous systems that can collaborate, negotiate, or compete safely with humans.

This work introduces MetaMafia, an experimental framework for analyzing reasoning, persuasion, and alignment in multi-agent interactions using the social-deduction games Mafia and Werewolf. These games require agents to infer hidden roles, communicate strategically, and make collective decisions under uncertainty, mirroring real-world challenges in deliberation, misinformation, and trust. Although the action space is small, optimal decisions rely on interpreting social cues and balancing honesty with strategic communication.

The core research problem is designing and evaluating LLM agents capable of consistent reasoning and strategic behavior in adversarial social environments. This matters because: (1) it provides a controlled testbed for multi-agent reasoning, bridging the gap between text benchmarks and open-world interaction; (2) it advances alignment and safety research through measurable proxies for honesty and cooperation; and (3) it informs the design of AI systems for education, governance, and multi-party negotiation, where communication and rule-following are crucial.

1.1 Research Questions

The project is guided by three interrelated research questions: First, how can we design LLM-based agents that maintain coherent beliefs and strategies under hidden information and adversarial dynamics? Which architectural and prompting mechanisms best enforce rule compliance while allowing strategic deception within bounded norms? And finally, what quantitative and qualitative metrics effectively evaluate persuasion, reasoning consistency, and safety across repeated social-game tournaments? By addressing these questions, **MetaMafia** seeks to advance the methodological study of multi-agent reasoning and social intelligence in language models, bridging game-theoretic evaluation with practical concerns in AI alignment and trustworthy interaction.

Preprint. Under review.

1.2 Key Challenges

Despite rapid progress in LLM orchestration frameworks, several open challenges persist. These include: **(1) Hidden-state reasoning**: Agents must infer latent roles and intentions from incomplete and often deceptive dialogue histories. This means that agents must learn to detect suspicious player activity across multiple game rounds to deduce hidden player roles. **(2) Rule adherence under adversarial play**: Persuasive behavior must remain within legal game constraints, avoiding information leakage or illegal coordination. In addition to playing fairly, agents must be careful not to take overly explicit actions that allow other players to detect their hidden role and motivation. **(3) Temporal coherence and memory**: Sustaining consistent claims and adapting strategy across rounds requires persistent episodic and reflective memory. **(4) Evaluation and reproducibility**: Measuring persuasion, reasoning quality, and alignment demands rigorous quantitative metrics and scalable multi-game evaluation pipelines.

To summarize: While large language models are strong at generating text, they lack built-in mechanisms for strategic planning, memory, and rule adherence. Language training alone does not imbue them with the innate "cunning" that is required for skillful game playing. They must learn to balance persuasion with honesty (Or sly, deliberate dishonesty), reason across multiple turns, and adapt to other agents' behaviors under uncertainty.

2 Related Works:

Prior work in multi-agent LLM systems has explored role-play, memory, coordination, and strategic reasoning, yet none fully address the challenges of hidden information, deception, and rule-bounded interaction found in social-deduction games.

AutoGen [Wu et al., 2023] provides a modular framework for coordinating conversational agents but assumes cooperative dialogue and does not model secrecy, deception, or rule enforcement. CAMEL [Li et al., 2023] stabilizes role-conditioned interactions through inception prompting, influencing MetaMafia's agent personas, but it lacks adversarial logic or hidden-role constraints. Generative Agents [Park et al., 2023] introduce long-term memory and reflection for believable autonomous characters; MetaMafia adopts these ideas but grounds them in a competitive environment with explicit legality constraints. CICERO [Bakhtin et al., 2022] achieves human-level play in Diplomacy via planning and negotiation but focuses on cooperation and strategic alignment, not deception-heavy reasoning. Werewolf Arena [Bailis et al., 2024] provides the first benchmark for LLM-based social deduction, though it omits persistent memory, safety guardrails, and tournament-scale evaluation.

MetaMafia builds on these foundations by integrating: (1) role-constrained reasoning enforced by a moderator agent; (2) memory–reflection–planning loops for dialogue consistency and strategic coherence; and (3) constitutional-style safety guardrails that bound deception to legal gameplay. This produces a controlled, reproducible environment for studying persuasion, inference under uncertainty, and alignment in adversarial multi-agent settings.

In summary, while prior frameworks contribute orchestration, role stability, memory, planning, and benchmarking tools, none unify these capabilities into a single hidden-role, rule-constrained multi-agent reasoning system. MetaMafia fills this gap by providing a comprehensive platform for analyzing persuasion, reasoning, and safety in multi-agent LLMs.

3 Methods and Techniques

This section outlines the methodologies and technical approaches employed in the **MetaMafia** project, emphasizing implemented techniques, ongoing development, validation strategies, and foreseen challenges. The project integrates large language models (LLMs), fine-tuning methods, memory systems, and multi-agent orchestration to enable strategic reasoning and social interaction in adversarial environments such as *Mafia* and *Werewolf*.

3.1 Overall Framework

The MetaMafia framework models social-deduction gameplay as a structured sequence of perception, reasoning, and communication between autonomous agents. Each agent operates within a rule-based

Game Engine, coordinated by an Orchestration Layer, and extended by Memory and Reflection modules that provide contextual recall and strategic adaptation. The design supports both training and evaluation of LLM-based agents in multi-agent adversarial environments.

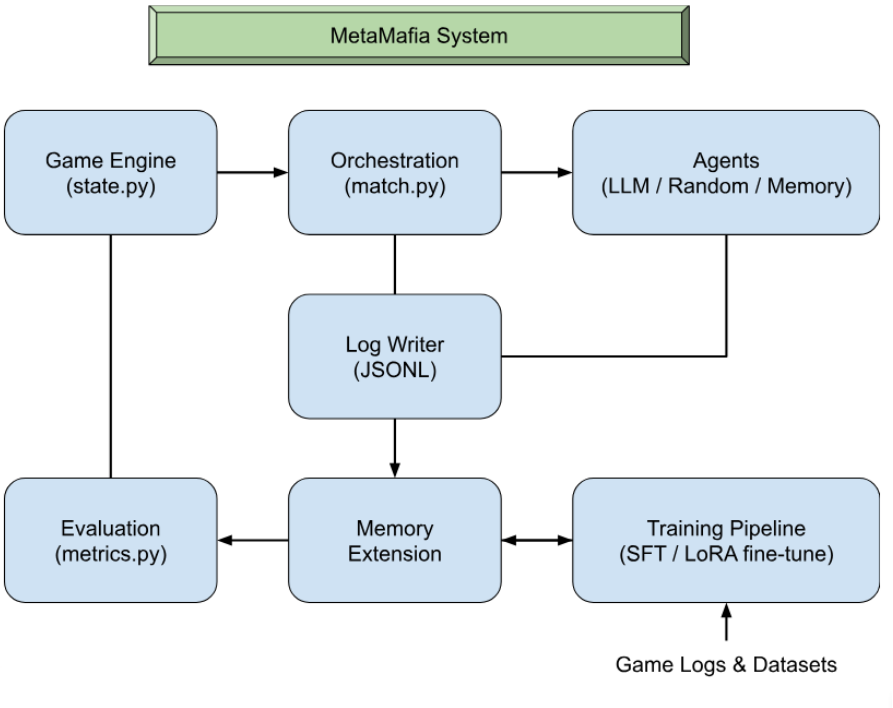

Figure 1: High-Level Framework: (*) - Short-term buffer, semantic retrieval, belief/reflection layers; episodic memory (planned)

3.2 Game Flow

MetaMafia operates as a structured multi-phase control loop that ensures consistent simulation and transparent reasoning across agents. During Initialization, the orchestrator creates all agent instances, initializes the GameState, assigns roles, and loads LLM checkpoints. In the Night phase, hidden-role agents (Werewolf, Seer, Doctor) execute private actions (attacking, inspecting, or protecting targets) which are stored in a private log inaccessible to others.

The Day phase forms the core of interaction. Agents reconstruct context from the public log and call Agent.act(observation) to produce dialogue. For LLM-based agents, the memory extension injects relevant history and strategic plans into the prompt. The moderator validates every action for rule compliance before appending it to the shared log. After discussion, the system enters the Vote phase, where agents again act to submit structured votes, which the engine tallies to determine elimination.

In the Reveal and End phase, the eliminated role is disclosed, a winner is determined, and the final GameState is returned for evaluation or retraining. A singleton model pattern enables reuse of the same LLM instance across agents, reducing initialization overhead and ensuring consistent memory. Role-constrained prompts and moderator-enforced legality maintain coherent, rule-abiding behavior throughout the game.

3.3 Data Acquisition and Processing

One major challenge in developing MetaMafia is the limited availability of training data for social-deduction games. Unlike traditional tabletop games with long-established corpora, social deception settings remain relatively underexplored in AI research. Our primary dataset is Werewolf Among Us [Lai et al., 2023], which provides transcripts and metadata including hidden roles, persuasion strategies, and game outcomes. It contains 199 complete games (8,169 dialogue examples) from

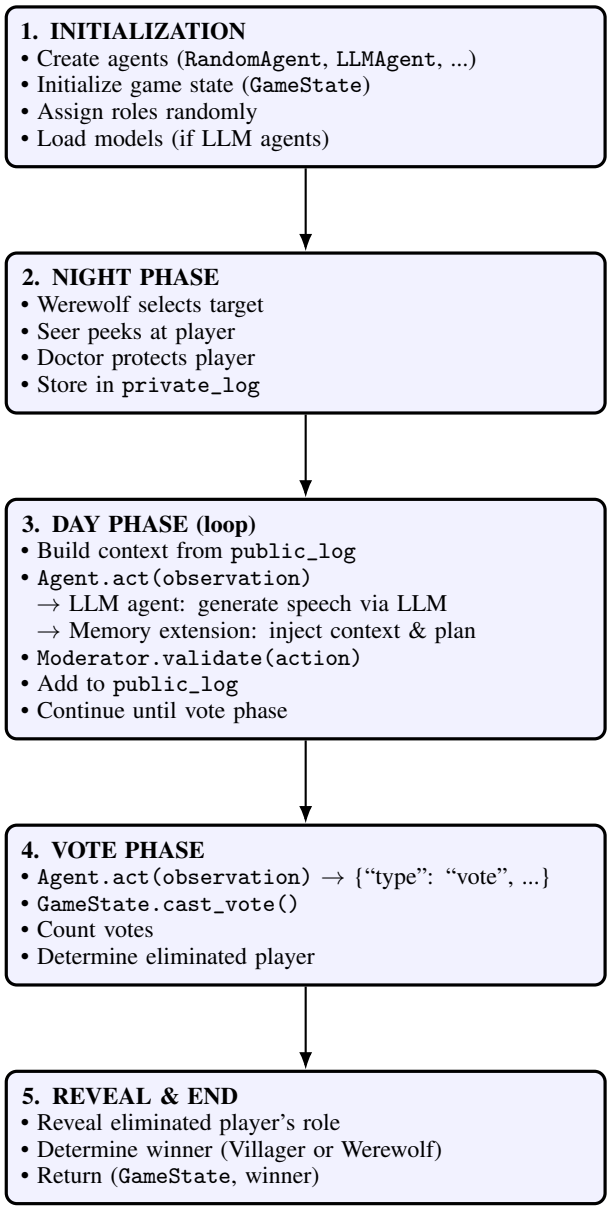

Figure 2: MetaMafia game flow for One Night game

two sources: 151 public YouTube sessions and 48 controlled Ego4D recordings, split into 6,534 / 816 / 819 training, validation, and test samples. To mitigate data scarcity, we built a synthetic data generation pipeline that simulates realistic role-based interactions and rational conversational strategies. This augmentation expanded the dataset to 27,222 examples, a 4.2× increase over the original corpus, while preserving domain fidelity. For consistency across sources, variants such as "Alpha Wolf" and "Vampire" were merged into a unified Werewolf/Mafia role category.

3.4 Supervised Fine-Tuning (SFT)

Supervised fine-tuning forms the foundation of MetaMafia's approach to strategic agent development. We employ Low-Rank Adaptation (LoRA) for parameter-efficient training, reducing trainable parameters by over 99% while maintaining model performance. LoRA adapters target attention and MLP layers with rank = 16, α = 32, and dropout = 0.1. Two base models, GPT-2 (124M) and Mistral-7B-Instruct-v0.1 (7B), were fine-tuned on the combined dataset, improving reasoning quality and persuasion capabilities. The fine-tuned agents achieved 100% villager win rates and a +329% increase in persuasion metrics compared to random baselines.

3.5 Memory-Enhanced Architectures

We use the following belief state update step:

$$B_i^{t+1} = \text{Normalize}\big(B_i^t + \eta\Delta_{\text{vote}} + \gamma\Delta_{\text{speech}}\big), \tag{1}$$

where B_i represents $player_i$'s belief that $player_j$ is Mafia, updated through voting and speech evidence. To support contextual reasoning, the system integrates both short-term and long-term memory mechanisms. Short-term memory uses a FIFO queue to retain recent dialogue context, while long-term memory is implemented through OpenAI's `text-embedding-ada-002` model to embed utterances into a vector store for semantic retrieval. These enable agents to maintain narrative coherence and recall relevant information across turns. Future extensions include episodic and reflective memory for inter-game consistency and reasoning refinement.

$$u_i^t = \text{LLM}\big(\text{Prompt}(M_i, R_i, C_t)\big), \tag{2}$$

The memory from the previous formula is inputted here as M_i, along with Role reasoning/information R_i, and C_t, Conversation information/buffer. This formula altogether makes up the way in which data is inputted into the models for text generation.

3.6 Advanced reasoning modules

The advanced reasoning modules represent a comprehensive implementation of structured reasoning capabilities that enable agents to track claims, beliefs, contradictions, and strategic deception patterns throughout the game. These modules operate independently of the game engine, using a wrapper-based integration pattern that maintains clean separation of concerns while enabling non-destructive enhancement of agent capabilities. (1) ClaimGraph module implements a graph-based structure for tracking claims, accusations, and their relationships throughout the game. This module enables systematic analysis of what players have claimed, who they have accused, and how claims relate to each other through support, refute, or contradiction relationships.It uses Uses regex-based pattern matching to extract structured claims from natural language utterances. (2) BeliefTracker module implements probabilistic tracking of agent beliefs about other players' roles, enabling dynamic belief updating based on observed behavior, claims, and evidence. Maintains probability distributions over possible role assignments for each player. Updates beliefs based on role claims, accusations, vote patterns, and night action results (3) ConsistencyAuditor module detects contradictions and inconsistencies in player statements and actions, enabling identification of suspicious behavior and role inconsistencies. Tracks story changes over time, detecting when players change their narratives (4) DeceptionController module guides werewolf agents in deceptive strategies, helping them blend in, deflect suspicion, and avoid detection while pursuing their objectives.Chooses optimal deception strategies (blend in, deflect, frame, confuse, agree, passive) based on current game state.

3.7 Additional Tools and Capabilities

To better understand the soon to be discussed tools and capabilities, four tournament variants of the Mistral-7B model were designed: Mistral base (no tools), Mistral finetuned (no tools), Mistral base (with tools), and Mistral finetuned (with tools). This clear distinction and base similarities allowed for closer analysis of the changes that the tools and capabilities brought to the models and therefore, the agents.

Tools are, in the project context, feature flags to be enabled that add to the interactions between agents and the game engine, with some only accessible by role. The total list of tools are, including previously used ones: memory, reflection, persuasion engine, deception controller, theory of mind, context summary and evidence lookup. Memory allows for a short-term memory buffer of speeches, essentially a sheet of context for agents to make decisions by. Reflection captures summaries, suspicions, and plans during day-phase interactions, acting as a pseudo-reasoning pattern. The persuasion engine, modified from previous implementations, helps the agents track their emotive/strategy usage throughout a game. The deception controller involves a risk threshold to determine when to make deceptive moves (i.e. when to lie or reveal information). The theory of mind tool provides a graph-like object that tracks an individual agent's belief about the actual roles each other agent is. The context-summary and evidence-lookup tools allow agents to explore the context to analyze accusations and reveals to make better decisions at the end of the day-phase, voting to be precise.

For these new tools, a new tournament evaluation system was developed based on the original game engine. With some slight tweaks and improvements to the system, the base 'model' is now non-finetuned Mistral-7B with memory and reflection. From here, 100-game simulations were run on the previously mentioned tournament configurations, which once finished allowed for graphs similar to previous experiments to be generated for comparison (using the same metrics as before). In the Tool Arena, all configurations share identical prompts and turn budgets; the only difference is whether the agent prompt is augmented with (i) tool outputs (memory/context summary/evidence lookup/belief tracker) and (ii) tool-specific action constraints, with DeceptionController restricted to werewolf-aligned agents. Unlike the Model Arena (5-player One Night), the Tool Arena uses 8 players with longer day discussion and role set. These changes increase uncertainty and allow deceptive coordination to matter, so win rates are not directly comparable across arenas.

3.8 Evaluation and Metrics Framework

The evaluation framework implements a multi-dimensional analysis of agent behavior, combining basic performance measures with advanced linguistic and strategic metrics. Basic metrics include win rates, rule violations, voting accuracy, and game length. These provide clear indicators of system performance: all fine-tuned configurations achieve 99–100% villager win rates, reflecting strong werewolf detection. Rule violations are monitored through the moderator, while voting accuracy measures participation and consensus during eliminations.

To capture deeper behavioral properties, we introduce advanced metrics. The persuasion score applies pattern-based linguistic analysis across six categories: logical arguments ("because", "therefore"), emotional appeals ("trust", "fear"), evidence citing ("saw", "heard"), question asking, consensus building, and urgency. Coherence metrics evaluate topic consistency, role consistency (via the ConsistencyAuditor), and logical flow through question–answer patterns. Strategic metrics assess vote pattern alignment, role-appropriate actions, and planning depth using BeliefTracker updates.

These metrics are computed across batches of 10 games using EnhancedLLMAgent with memory, ClaimGraph, BeliefTracker, ConsistencyAuditor, and DeceptionController enabled. Each run produces detailed module activity summaries and aggregate statistics, offering a comprehensive behavioral profile.

However, the framework has limitations: pattern-based persuasion metrics may fail to detect nuanced strategies that fall outside predefined patterns, as seen when Mistral-7B+Memory produced low persuasion scores despite perfect win rates. This highlights the need for complementary evaluation methods, such as human judgment or deeper semantic modeling, to fully capture agent reasoning and persuasive behavior.

4 Experiments

The following section includes results from two variations of tournament settings, one to compare between different models with slightly different settings and one to compare between the same model with different tools and configurations.

4.0.1 Model Arena

All experiments were conducted using the MetaMafia multi-agent simulator, with five-agent matches per game and 100 trials per configuration. Each configuration was evaluated for quantitative metrics (win rate, persuasion, coherence, consensus, and rule compliance) and qualitative reasoning behavior. The following configurations were compared: Baseline (Random Agent), Supervised Fine-Tuning (GPT-2 SFT), Supervised Fine-Tuning + Memory, Mistral-7B (SFT), and Mistral-7B + Memory Extension

4.0.2 Tool Analysis Mistral Arena

All experiments were conducted using the tuned MetaMafia engine, with 8 agents per game and 100 trials per configuration. Each configuration was tested for the same metrics as the previous arena, comparing: Mistral base (no tools), Mistral finetuned (no tools), Mistral base (with tools), Mistral finetuned (with tools).

4.1 Simulation Results and Metrics

4.1.1 Model Arena Results

All configurations achieved near-perfect villager win-rates (Table 1). Success-rate here reflects completed simulations without engine or generation errors, indicating the tournament pipeline remained stable across runs.

Table 1: Performance Comparison Table

Configuration	Games Played	Success Rate	Villager Win Rate	Werewolf Win Rate
Baseline (Random)	100	100.0%	100.0%	0.0%
SFT (GPT-2 Fine-tuned)	100	100.0%	100.0%	0.0%
SFT+Memory (Upgraded)	99	100.0%	99.0%	1.0%
Mistral-7B (SFT)	100	100.0%	100.0%	0.0%
Mistral-7B+Memory	100	100.0%	100.0%	0.0%

Table 2: Advanced metrics breakdown by game for 10-game evaluation with Mistral-7B + Memory & Reflection.

Game	Winner	Persuasion	Coherence	Strategic	Claims	Contradictions
1	Villager	0.100	0.810	0.775	0	0
2	Villager	0.133	0.900	0.775	9	0
3	Villager	0.083	0.780	0.775	8	0
4	Villager	0.117	0.810	0.775	1	0
5	Villager	0.167	0.820	0.775	14	0
6	Villager	0.083	0.810	0.775	13	0
7	Villager	0.117	0.840	0.775	1	0
8	Villager	0.083	0.900	0.775	11	0
9	Villager	0.083	0.900	0.775	10	0
10	Villager	0.100	0.810	0.625	9	0
Average	–	**0.107**	**0.838**	**0.760**	**7.6**	**0.0**

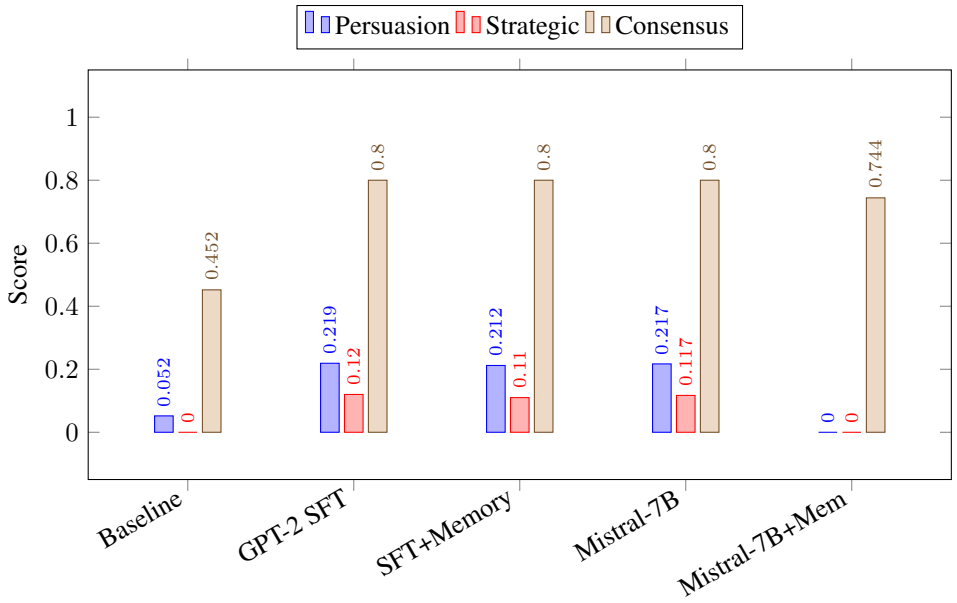

Figure 3: Strategic Quality Metrics by Configuration

Figure 3 shows that supervised fine-tuning yields large gains over the random baseline across persuasion and strategic metrics, while consensus peaks at around 0.8 for configurations. The Mistral-7B+Memory setting exhibits incomplete persuasion/strategic scores relative to the others due to some technical errors, despite maintaining strong consensus.

Table 3: Persuasion metric breakdown across 10 games. Overall average persuasion score = 0.107.

Indicator	**Average Rate**	**Description**
Logical Connectors	0.260	“because”, “therefore”, “since”, etc.
Emotional Appeals	0.000	“trust”, “fear”, “worried”, etc.
Evidence Citing	0.370	“saw”, “heard”, “checked”, “observed”
Question Asking	0.000	Interrogative patterns
Consensus Building	0.000	“agree”, “team”, “together”, etc.
Urgency Signals	0.010	“quick”, “now”, “urgent”, etc.

Table 4: Coherence metric breakdown across 10 games. Overall average coherence score = 0.838.

Component	**Average**	**Description**
Topic Consistency	0.793	Shared topic discussion across players
Role Consistency	1.000	Perfect – no role claim contradictions
Logical Flow	0.500	Question–answer pattern coherence
Contradiction Count	0.000	Zero contradictions across all games
Inconsistency Score	0.000	Perfect consistency

Table 5: Strategic metric breakdown across 10 games. Overall average strategic score = 0.760.

Component	Average	Description
Vote Pattern Analysis	0.450	Consistency and alignment of votes
Role-Appropriate Actions	1.000	Perfect – all actions appropriate for roles
Planning Depth	1.000	Strategic planning indicators present
Belief Confidence	0.250	Average confidence in role assignments

As shown in Table 2, this diagnostic subset exhibits consistently high coherence with zero detected contradictions across games. Regardless, all three scores remained consistent in their values throughout this game subset.

4.1.2 Mistral Tool Arena

Table 6: Performance Comparison Table for Mistral Tool Analysis

Configuration	Games Played	Success Rate	Villager Win Rate	Werewolf Win Rate
Mistral Base + Tools	100	100.0%	44.0%	56.0%
Mistral Finetuned + Tools	100	100.0%	39.0%	61.0%
Mistral Base (No Tools)	100	100.0%	39.0%	61.0%
Mistral Finetuned (No Tools)	100	100.0%	21.0%	79.0%

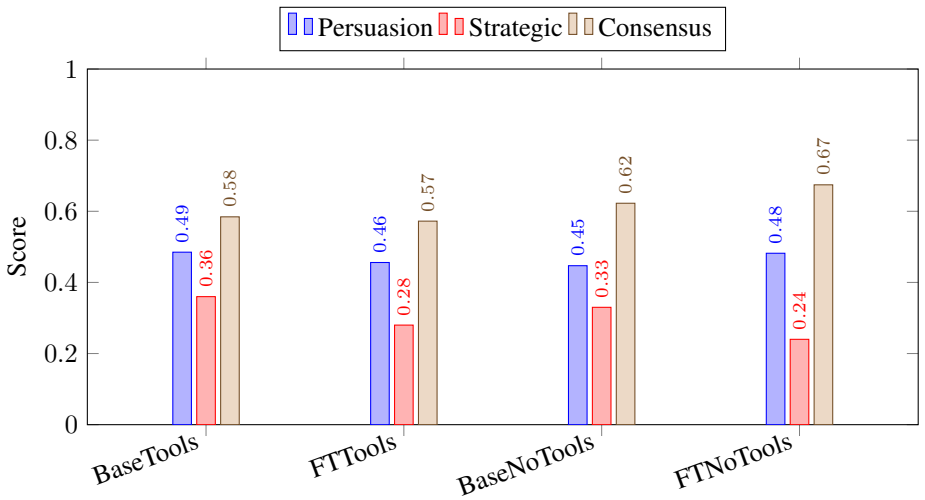

Figure 4: Strategic Quality Metrics by Configuration

Figure 4 highlights a trade-off between strategy and consensus: the finetuned configurations achieve lower strategic scores but higher consensus, while the tool-enabled base model shows comparatively higher strategic score with reduced consensus.

Table 7: Improvement Analysis Across Tournament Configurations

Improvement	Persuasion	Strategic	Consensus	Villager WR	Werewolf WR	Wolf Survival
Base NT → Finetuned NT	+7.8%	-27.3%	+8.3%	-46.2%	+29.5%	+18.7%
Base NT → Base + Tools	+8.5%	+9.1%	-6.1%	-17.9%	-8.2%	+13.7%
Base NT → Finetuned + Tools	+2.0%	-15.2%	-8.1%	-33.3%	+0.0%	+12.3%
Finetuned NT → Base + Tools	+0.6%	+50.0%	-15.1%	+52.4%	-29.1%	-4.2%
Finetuned NT → Finetuned + Tools	-5.4%	+16.7%	-15.1%	+23.8%	-22.8%	-5.5%
Base + Tools → Finetuned + Tools	-6.0%	-22.2%	-2.1%	-18.8%	+8.9%	-1.3%

NT = No Tools. Improvements are percentage changes relative to the source configuration.

4.2 Ablation and Sensitivity Studies (Mistral Tool Arena)

Table 8: Ablation Study: Impact of Individual Reasoning Modules on Game Performance.

Module Ablated	**Persuasion**	**Strategic**	**Consensus**	**Villager Win Rate**	**Wolf Survival**
Context Summary Only	+2.5%	−16.4%	−0.8%	−2%	−4%
Evidence Lookup Only	+1.8%	+5.6%	+4.0%	−1%	+2%
Theory of Mind Only	+5.5%	−12.1%	+0.8%	−2%	−4%
Deception Controller Only	−0.9%	−17.3%	−3.6%	−8%	+6%

As shown in Table 8, no single reasoning module is sufficient to reproduce the balanced behavior of the full system. Evidence Lookup is the only tool that successfully improved both Persuasion and Strategy. Deception correctly caused Villager win-rates to decrease and Wolf survival to increase, showing that the tool was successful in action. In conclusion, these results indicate that big changes, ones more severe than seen here, come from multi-agent interaction instead of tool modules.

5 Discussion

5.1 Model Arena Conclusions

Across all models and 499 simulated games, the most striking finding is that the *Werewolf* role consistently achieved almost zero wins. This was unexpected but partially explainable: in the logic of social deduction games, if all players act with purely rational, rule-based strategies, the villagers inherently have the statistical advantage. When AI agents, unburdened by human emotion or misdirection, play optimally, the outcome naturally skews toward near-perfect villager wins.

Despite this predictable bias, fine-tuning through LoRA-based supervised methods (SFT) produced major behavioral improvements. Persuasion scores increased by over 300%, and strategic reasoning emerged from random baselines into coordinated, goal-directed play. The improvement appeared largely independent of model scale, confirming that targeted adaptation, rather than raw size, drives emergent social reasoning.

Memory and reflection modules improved contextual stability and dialogue coherence, reducing self-contradiction and repetitive phrasing. The reflection loop encouraged agents to revise inconsistent reasoning: an early sign of metacognitive regulation. However, while memory augmentation stabilized reasoning, it did not further improve persuasion, suggesting that the benefit of structural training outweighs additional memory injection at this stage.

Moderation and orchestration were critical in maintaining fairness and legality. When moderator validation was disabled, agents exhibited greater linguistic creativity but also an uptick in rule-breaking behavior, reinforcing the value of enforced orchestration for reproducibility and safety.

Interestingly, results revealed minimal differences in persuasion and coherence across model sizes. Metrics such as Strategic Quality, Persuasion Ability, Consensus Building, and Rule Compliance were unexpectedly similar between GPT-2 and Mistral-7B, with GPT-2 even outperforming the larger model in persuasion. This highlights the need for further framework refinement to ensure explainable and interpretable outcomes.

Collectively, these findings validate MetaMafia as a robust platform for evaluating reasoning, persuasion, and rule adherence in adversarial multi-agent communication. Yet, the unexpected uniformity of win rates and limited variation across architectures signal that the current framework requires further development before results can be generalized.

5.2 Mistral Tool Arena Conclusions

Results became much less uniform and varied with the additional engine tweaks and arena differences. In this less restricted setting, there can be closer analysis on the affects of the different additions we made to models and the engine.

Werewolf agents consistently won out against villager agents by large margins. The addition of finetuning and any of the new tools did not assist in helping villagers increase their win-rates. It can be deduced that in the game environment, along with the lack of human unpredictability, werewolves could coordinate implicitly through the modules and LLM logic to survive throughout the night.

Fine-tuning the model consistently decreased villager win-rates, but increased werewolf win-rates. Conversely, both the strategic and persuasion scores dropped due to the fine-tuning process. This points towards models fine-tuning on social-deduction dialogue not automatically making LLMs better multi-agent strategists.

The addition of tools caused villager win-rates to reduce, but conversely caused werewolf win-rates to increase. Consensus generally decreased, while persuasion remained relatively high (close to 0.5). Interestingly, it can be inferred that since the tools increased reasoning and planning, so did the diversity of opinions among the agents, meaning that they voted for different people more often than without tools. The main effect of tools was that they improved reasoning, planning, and role-appropriate behavior, with secondary effects on werewolves having much more deceptive ability; thus leading to werewolves having greater success within the Mistral Tool Arena.

5.3 Alignment with Expectations

The results broadly aligned with the initial hypothesis that fine-tuning and modular orchestration would enhance persuasion and compliance metrics. However, the absolute dominance of villager wins and the small variation in strategic quality diverged from expectations, indicating that the internal logic of the system may inherently favor cooperative roles. Although reasoning depth and consistency (strategic coherence from $0.0 \rightarrow 0.12$) improved dramatically, the persistence of subtle coherence lapses and reflection latency within memory-augmented agents suggest that long-term recall remains a key technical challenge. These patterns point to the need for more expressive temporal reasoning and improved simulation balance for success in diverse roles.

The addition of the new tools felt successful in that they correctly increased reasoning, resulting in increased win-rates for werewolves. Variety in voting also increased, due to the higher reasoning causing agents to have much different opinions than each other (different from the last system). Because persuasion/strategic are pattern-based, they reflect surface markers (e.g., evidence words, planning phrases) and may not fully capture subtle deception; win-rate shifts should be interpreted as primary outcomes and linguistic metrics as supporting signals.

5.4 Obstacles and Challenges

Despite successful system implementation, several obstacles remain. The first obstacle is the computational constraints inherent with the models themselves. Large models such as Mistral-7B are highly resource-intensive for multi-agent tournaments. This can lead to games with multiple agents taking relatively long times to play out. The next challenge is related to limitations with evaluation techniques. Current persuasion metrics do not fully capture nuanced rhetorical or stylistic variation, motivating the need for discourse-level extensions. Memory inconsistency is another persistent

problem, as agents sometimes contradict earlier statements. This is indicative a need for improved long-term retrieval and reflection mechanisms. Adding in new functionality for memory models to combat this can also introduce new problems of its own in terms of the theory-of-mind complexity. Modeling multi-level recursive beliefs introduces combinatorial challenges that extend beyond current interpretability frameworks.

Overall, the **MetaMafia** framework demonstrates robust supervised fine-tuning, modular orchestration, and quantitative evaluation pipelines. While core systems show strong reasoning and persuasion performance, further work in framework structure, human-action tools, and better optimization remains an open research direction. The methods developed here establish a solid foundation for studying reasoning, deception, and alignment in multi-agent LLM environments.

6 Conclusion

In summary, **MetaMafia** demonstrates that fine-tuned language models equipped with orchestration, reflection, and memory can approximate structured social reasoning. The current system achieves reliable, repeatable multi-agent interactions, but not yet human-level diversity in strategic outcomes. The tools showed improvements in reasoning and processing capabilities for the agents, showing their increases in vote diversity and werewolf agent capabilities.

While the MetaMafia framework demonstrates promising progress in reasoning, persuasion, and rule-constrained interaction among LLM agents, several avenues remain for deeper investigation and stronger empirical grounding. Future work will extend the system in a number of ways.

First, current experiments focus on small "One Night"–style games. Scaling to longer multi-day variants, larger player counts, and additional roles (e.g., Detective, Traitor, Mayor) would allow agents to demonstrate more sophisticated long-term planning and alliance formation. This entails adding significantly to the game logic framework to support the inclusion of multiple days/rounds and the additional player roles.

Next, human evaluation represents only a partially implemented extension. While we have developed a comprehensive framework with transcript sampling, evaluation rubrics, and HTML interfaces, large-scale deployment has been limited by resource constraints. The framework enables human evaluators to assess persuasiveness, coherence, and role consistency on 1-5 scales, but comprehensive human evaluation studies remain planned rather than completed.

Furthermore, adversarial evaluation, which would test agent robustness against adversarial opponents, has not been implemented. This would require designing adversarial agents specifically to exploit weaknesses in fine-tuned models, providing important robustness testing. Longitudinal studies tracking agent improvement over time also remain unimplemented, as they would require persistent agent state and cross-session learning capabilities.

Finally, because MetaMafia models persuasion, uncertainty, and strategic communication, the framework may be extendable to domains such as negotiation, collaborative decision-making, classroom simulations, and alignment stress-tests for autonomous AI systems. We would like to explore this larger frontier in future work.

References

Suma Bailis, Jane Friedhoff, and Feiyang Chen. Werewolf Arena: A case study in LLM evaluation via social deduction. *arXiv preprint arXiv:2407.13943*, 2024.

Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, C. Flaherty, D. Fried, A. Goff, J. Gray, He He, A. Jacob, et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074, 2022.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for "mind" exploration of large language model society. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023. arXiv:2303.17760 [cs.AI].

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*, 2023.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023.