Reasoning as a Team: Evaluating Role-Based Multi-Agent Collaboration Through Qualitative Supervision

Chin-Lang Peng, **Shang-Hong Lai**
National Tsing Hua University
111005509@m111.nthu.edu.tw, lai@cs.nthu.edu.tw

Abstract

We propose a role-based multi-agent framework inspired by Minsky's *Society of Mind*, integrating reasoning and acting through qualitative supervision. Large Language Model (LLM) agents assume distinct epistemic roles, such as bold exploration and cautious verification, interacting with the external environment via ReAct-style Thought-Action-Observation loops to produce reasoning traces. A supervisory agent qualitatively evaluates these traces based on coherence, plausibility, and completeness, providing structured feedback to guide agent revisions. This approach operationalizes classical qualitative reasoning concepts, including generating and refining intermediate hypotheses and exploring reasoning trajectories. Experiments conducted on the HotpotQA benchmark demonstrate that supervisor-guided collaboration enhances factual accuracy and reasoning coherence across multiple LLM backends, highlighting supervisor-agent compatibility as crucial for adaptive, cognitively grounded decision-making.

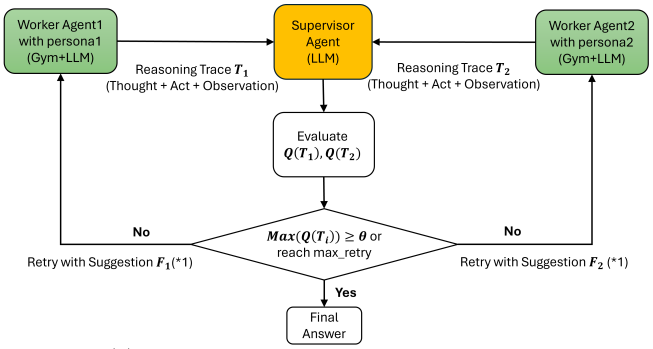


Figure 1: Overview of our qualitative multi-agent reasoning framework. Two worker agents (Bold and Cautious) independently generate multi-hop reasoning traces T_1 and T_2 via ReAct-style Thought-Action-Observation loops. The supervisory agent evaluates each trace using a qualitative scoring function $Q(T)$, which assesses coherence, plausibility, and completeness. If no trace meets the threshold θ, the supervisor issues feedback F_1 and F_2 to guide agent revision. This loop continues until acceptable trace quality is achieved or the retry limit is reached.

1 Introduction

Traditional decision support systems typically rely on explicit, rule-based logic, structured by system designers as fixed if–then–else pathways[Turban, 2011]. While effective in static, predictable environments, these methods often fail to adapt to dynamic, uncertain contexts, where reasoning must integrate fragmented evidence, resolve ambiguity, and continuously refine inference strategies[Russell and Norvig, 2016].

Recent advances in large language models (LLMs) have opened new possibilities for flexible, context-aware agents capable of multi-hop reasoning across unstructured information sources[Shinn *et al.*, 2023; Yao *et al.*, 2023b]. However, a critical challenge remains: ensuring that these agents not only produce correct answers but also follow reasoning paths that are plausible, interpretable, and aligned with human cognitive expectations[Ji *et al.*, 2023; Bender *et al.*, 2021].

Inspired by Marvin Minsky's *Society of Mind* [Minsky, 1986], we explore how structured collaboration among multiple agents—each adopting distinct epistemic roles—can qualitatively emulate the deliberative dynamics of human expert teams. We propose a multi-agent reasoning architecture in which ReAct-based worker agents engage in multi-hop problem solving from different epistemic perspectives (e.g., bold exploration and cautious verification). A third supervisory agent then performs qualitative evaluation of their reasoning traces, assessing not only factual correctness but also coherence, completeness, and interpretability. This iterative arbitration process mirrors how human experts critique and refine each other's reasoning in collaborative settings.

Our framework aligns with the recent Agentic Workflow paradigm[Zhang *et al.*, 2024b], emphasizing iterative planning, tool use, self-reflection, and multi-agent interaction. Rather than focusing solely on answer correctness, we explicitly model the reasoning process itself, providing structured oversight to enhance robustness, interpretability, and cognitive realism.

As illustrated in Figure 1, our framework involves two epistemically diverse agents whose multi-hop reasoning traces are iteratively evaluated and refined by a supervisory agent, simulating expert-style arbitration.

Our key contributions are:

- We propose a novel multi-agent reasoning framework that assigns distinct epistemic roles to agents and incorporates qualitative oversight via a supervisory agent.
- We empirically demonstrate that guided collaboration improves reasoning accuracy and coherence across multiple LLMs, as shown on HotpotQA.
- We show that supervision requires distinct reasoning skills, and supervisor-agent compatibility is crucial for effective collaboration.

2 Related Works

Qualitative Reasoning and Multi-Hop Inference. Real-world decision-making often requires synthesizing indirect evidence and reasoning under uncertainty, a challenge central to both qualitative reasoning and multi-hop inference. Qualitative reasoning, rooted in symbolic abstractions and causal models [Forbus, 1984], provides a framework for handling incomplete knowledge by deriving possible states from partial observations [Weld and De Kleer, 2013]. Similarly, multi-hop inference operationalizes this paradigm in open-domain tasks by chaining fragmented evidence across multiple sources [Yang *et al.*, 2018].

The intersection of these fields is evident in their shared reliance on intermediate hypothesis generation—whether through qualitative state transitions or multi-step reasoning paths [Wei *et al.*, 2022]. Recent advances further bridge the gap: for example, self-refinement mechanisms [Madaan *et al.*, 2023] and tree-based exploration [Yao *et al.*, 2023a] incorporate the "hypothesis–test–refine" cycle into multi-hop pipelines. Benchmarks like HotpotQA [Yang *et al.*, 2018] reflect these demands, requiring agents to aggregate uncertain cues, resolve conflicts, and infer latent relationships.

In this work, we embrace the reasoning principles emphasized in qualitative reasoning—such as intermediate hypothesis refinement and deliberative inference—while leveraging the generative flexibility of LLMs to explore open-domain, language-native environments. Although our current focus is on language-based supervision and trace arbitration, we view our approach as complementary to symbolic reasoning, and see opportunities in future work to integrate explicit qualitative models to improve interpretability and precision.

Agentic Reasoning and ReAct Frameworks. Recent advances in large language models have emphasized combining internal reasoning with external interaction. The ReAct (Reasoning + Acting) framework [Yao *et al.*, 2023a] exemplifies this by intertwining thought generation and action execution, allowing agents to adjust their reasoning dynamically based on external feedback. Unlike Chain-of-Thought (CoT) prompting [Wei *et al.*, 2022; Wang *et al.*, 2022], which primarily generates static reasoning traces, ReAct models a more iterative, situated reasoning process. Our work extends this framework into a multi-agent setting, with distinct epistemic roles enhancing cognitive diversity.

Recent efforts such as Tree of Thoughts [Yao *et al.*, 2023a] have introduced multi-step reasoning strategies that explore multiple reasoning paths via structured tree search. While sharing the motivation of enhancing deliberative reasoning, our framework adopts a fundamentally different approach: instead of path selection via value-based search, We employ a team of epistemically diverse agents whose reasoning traces are qualitatively evaluated and refined by a supervisory agent. This enables interpretability through trace diversity, arbitration, and targeted feedback—hallmarks of human-style collaborative reasoning.

Multi-Agent Collaboration and Role Diversity. Recent advances in multi-agent systems leveraging large language models (LLMs) have demonstrated significant potential in task decomposition, cooperative reasoning, and interactive simulations [Wu *et al.*, 2023; Park *et al.*, 2023; Wu *et al.*, 2025; Li *et al.*, 2023]. A key line of research employs *multi-agent debate* to harness cognitive plurality and enhance reasoning robustness. For example, [Du *et al.*, 2023] showed that structured agent interactions improve factual consistency, while [Creswell *et al.*, 2022] proposed a selection-inference framework where agents collaborate via interpretable logical steps, prioritizing transparent reasoning paths.

However, existing approaches exhibit limitations. While iterative self-feedback mechanisms[Madaan *et al.*, 2023] can refine answer quality, they often operate in isolation without structured external arbitration. Furthermore, prior works[Du *et al.*, 2023; Creswell *et al.*, 2022] primarily address closed-context reasoning, omitting tool use or real-world information retrieval. In contrast, our framework extends the *ReAct paradigm* [Yao *et al.*, 2023b], enabling agents to dynamically interact with external environments through web search and lookup actions. This integration introduces real-world complexity, necessitating robust supervision and arbitration mechanisms to ensure coherence and accuracy.

Supervision as Qualitative Arbitration. The supervisory mechanism in our system synthesizes insights from hierarchical multi-agent control and cognitive architectures [Minsky, 1986]. Unlike conventional supervision that focuses on output validation, our method implements Minsky's principle of *heterarchical control* through meta-reasoning about process quality. This aligns with recent work on self-refinement [Madaan *et al.*, 2023] while introducing structured qualitative evaluation criteria (coherence, plausibility, completeness) absent in prior multi-agent systems.

3 Methodology

We propose a multi-agent reasoning framework consisting of two role-based worker agents and a central supervisory agent. Worker agents independently solve multi-hop reasoning tasks using a structured *Thought-Action-Observation* loop inspired by the ReAct paradigm [Yao *et al.*, 2023b]. Their outputs, which include reasoning traces, are qualitatively evaluated by the supervisory agent, providing structured arbitration and iterative feedback. The framework is formally described as follows:

Agent Reasoning. Each worker agent A_i, assigned a distinct epistemic role (e.g., bold exploration or cautious verification), independently generates a reasoning trace:

$$T_i = \{(t_j, a_j, o_j)\},$$

where at each step j, t_j denotes a generated thought, a_j denotes an action executed (e.g., search, lookup), and o_j denotes the observed textual result of that action.

Supervisory Arbitration. The supervisory agent S qualitatively evaluates the set of reasoning traces $\{T_i\}$, assessing each trace for coherence, completeness, and plausibility. The supervisor assigns a qualitative score $Q(T_i)$ to each reasoning trace. If the highest scored trace meets or exceeds a qualitative threshold θ, the corresponding answer is accepted as the final solution. Otherwise, the supervisor initiates a retry loop, providing targeted revision feedback to the agents.

Qualitative Evaluation via Language-Based Scoring. To guide the arbitration process, we define a qualitative scoring function $Q(T)$, realized through a natural language prompt given to the supervisor model. The supervisor assesses each reasoning trace based on three epistemic dimensions—*coherence* (logical consistency), *plausibility* (reasonable inference given the evidence), and *completeness* (sufficient coverage of the reasoning steps). The output of Q is a natural language judgment that expresses whether the trace should be **accepted** or revised through a **retry**.

Unlike traditional numeric scoring functions, $Q(T)$ is semantic in nature. The threshold θ is not a scalar value but a latent criterion embedded within the supervisor's prompt. A trace T is considered acceptable if the supervisor's response entails satisfaction of all three evaluative dimensions, and thus semantically corresponds to `accept`. Formally:

$$Q(T) \models \texttt{accept}_\theta$$

where $\models$ denotes semantic entailment, and θ represents the implicit decision boundary learned via prompt conditioning. If this criterion is not met, the supervisor issues a `retry` signal and provides specific feedback aimed at improving the trace. This design enables the supervisor to act as a qualitative meta-reasoner, providing interpretive judgments rather than numeric evaluations.

To operationalize this behavior, we use the following prompt for the supervisor agent:

> *You are a calm and objective supervisor overseeing two AI agents solving a reasoning task. One agent favors bold, creative thinking, while the other emphasizes caution and evidence. You are open to unconventional reasoning if it leads to sound conclusions, but you are also committed to decision quality. When evaluating their reasoning traces, consider whether each trace is coherent (logically consistent and well-structured), plausible (reasonable given the available information), and complete (addresses all necessary steps and evidence). Your role is to choose the most appropriate next action:* ***accept*** *or* ***retry****. You must always provide a conclusive decision. Do not say the task is unanswerable. If retry is chosen, provide a specific and concise suggestion to improve the agent's reasoning.*

This prompt anchors the supervisor's evaluations around key epistemic standards while allowing flexibility in reasoning style. As a result, supervision becomes a language-based arbitration mechanism capable of generalizing across reasoning agents and model architectures.

Retry Loop. When a retry is triggered, the supervisory agent issues qualitative feedback F_i (e.g., "you neglected crucial evidence" or "revisit conflicting sources"). Worker agents then refine their reasoning traces T_i accordingly. This iterative loop continues until either the supervisor's quality criterion θ is satisfied or a maximum retry limit (set to 3 in our experiments) is reached.

Figure 1 illustrates the overall workflow and interactions between worker agents and the supervisory agent.

Agent Role Definitions. We define distinct epistemic roles via LLM system prompts, inducing diverse reasoning behaviors rather than explicitly hard-coding agent actions:

- **Bold Explorer**: Prioritizes creative hypothesis generation, novel connections, and rapid inference progression, accepting higher reasoning uncertainty.
- **Cautious Verifier**: Emphasizes precision, verification of evidence, logical soundness, and cautious inference strategies.
- **Neutral Reasoner**: Balances exploration and verification, adopting a moderate reasoning style without extreme risk-taking or excessive caution.

This role-based approach ensures reasoning diversity, facilitating richer supervisor arbitration.

System Implementation. Our implementation closely follows the experimental framework originally proposed in the ReAct paradigm [Yao *et al.*, 2023b], employing a reasoning environment modeled after the OpenAI Gym framework. Worker agents iteratively select from actions such as `Search[query]`, `Lookup[entity]`, and `Finish[answer]`, with observations derived from retrieved textual sources. The supervisor evaluates reasoning traces without direct access to the retrieval environment, thus maintaining its role strictly at the meta-reasoning level.

All experiments employ deterministic generation settings (temperature = 0), minimizing variability in reasoning behavior and allowing direct comparison across multiple models and configurations.

4 Experiments

4.1 Task and Dataset

We evaluate our framework on the HotpotQA benchmark [Yang *et al.*, 2018], a widely used dataset for multi-hop question answering. Each question requires reasoning across multiple facts and intermediate steps, often involving entities or concepts that span more than one Wikipedia passage.

In contrast to settings that provide pre-selected supporting documents, our agents are not given gold evidence. Instead, each worker agent is tasked with retrieving relevant information on its own by interacting with an external Wikipedia environment using `Search[query]` and `Lookup[entity]` tools. This setup simulates real-world open-domain reasoning, where information is distributed and must be actively retrieved by the agent.

To ensure task diversity, we randomly sample a balanced subset of bridge-type and comparison-type questions from the HotpotQA development set. Each question requires at least

two reasoning steps and cannot be answered from a single fact or paragraph.

4.2 Model Configurations and Agent Roles

To evaluate generality across model backends, we test our framework using three distinct LLMs:

- **DeepSeek-Chat**: A bilingual, instruction-tuned open-source model.
- **GPT-4o-mini**: A compact variant of OpenAI's GPT-4, optimized for speed and reliability.
- **Grok-3**: An instruction-tuned LLM from xAI, noted for grounded retrieval reasoning.

Each model is used to instantiate all three agents (Bold, Cautious/Neutral, and Supervisor) within the same configuration to isolate the impact of agent role and arbitration dynamics. Besides the standard **Bold + Cautious** pairing, we also evaluate a **Bold + Neutral** pairing to assess the effect of epistemic contrast on collaborative effectiveness.

4.3 Evaluation Metrics

We report two standard metrics for question answering:

- **Exact Match (EM)**: Measures whether the predicted answer string exactly matches the ground truth.
- **F1 Score**: Measures the token-level overlap between prediction and ground truth, balancing precision and recall.

These metrics jointly evaluate factual correctness and partial matching quality, and are widely adopted in multi-hop QA benchmarks such as HotpotQA[Yang *et al.*, 2018].

4.4 Experimental Setup

All agents operate in a reasoning environment based on the ReAct framework [Yao *et al.*, 2023b], using Gym-style stepwise execution. Each agent is prompted using a role-specific system prompt and interacts with the environment through `Search`, `Lookup`, and `Finish` actions.

The supervisor agent receives complete reasoning traces from both worker agents and selects the final answer or requests a retry, with a maximum of three retry rounds. Retry prompts contain targeted qualitative feedback generated by the supervisor.

All experiments are conducted using API access to each model, with generation temperature set to zero to ensure determinism and reproducibility. No token limit is imposed on intermediate thoughts or observations, allowing for unconstrained reasoning traces.

5 Results and Analysis

We present the evaluation results of different agent configurations across multiple large language models (LLMs), using Exact Match (EM) and F1 score as primary metrics. The results offer insight into the effectiveness of supervisor-guided team reasoning and the impact of persona design.

5.1 Team Collaboration Performance Across Models

Table 1 and Table 2 report results for Bold + Cautious and Bold + Neutral configurations across DeepSeek-Chat, GPT-4o-mini, and Grok-3. Both DeepSeek and GPT-4o-mini show consistent gains from team-based setups under supervisor guidance.

- **Bold + Cautious (Table 1):** DeepSeek-Chat improves by 1.5 EM and 0.7 F1 points (47.0% $\rightarrow$ 48.5%), while GPT-4o-mini gains 2.0 EM and 1.9 F1 points.
- **Bold + Neutral (Table 2):** DeepSeek-Chat sees larger gains (+2.9 EM, +3.4 F1), and GPT-4o-mini achieves the highest relative improvement (+2.5 EM, +1.8 F1).

These results suggest that supervisor-guided collaboration reliably enhances precision and coherence, regardless of persona pairing. While individual agent styles vary, the supervisor's arbitration remains the key driver of performance.

5.2 Supervisor-Guided Trace Evaluation Example

We present a representative example in Figure 2 to qualitatively demonstrate the effectiveness of our framework. This successful case illustrates how supervisor-guided arbitration can enable agents to collaboratively resolve complex multi-hop questions.

5.3 Divergence in Grok-3 Behavior

Unlike other models, Grok-3 deviates from the general trend: in both team configurations, the best performance comes from single-agent settings. The Bold agent achieves 52.0% EM and 67.3% F1, while the Neutral agent matches 52.0% EM with 67.1% F1—both surpassing their team-based counterparts.

To investigate whether Grok-3 is inherently unsuited for multi-agent collaboration or simply misassigned in the supervisor role, we conducted a control experiment: both workers were set to Grok-3, but the supervisor was swapped to DeepSeek-Chat. As shown in Table 3, this change restored EM to the team baseline and improved F1, indicating that Grok-3 is effective as a reasoning agent but suboptimal as a supervisor.

5.4 Ablation Study

(1) Adding a Third Agent. To explore whether increasing the number of collaborating agents enhances team performance, we conducted a second ablation study. We expanded the Bold + Neutral setup to include a third agent with a Cautious persona, while maintaining a single supervisor.

As shown in Table 4, the three-agent setup did not yield performance improvements. In fact, it slightly decreased both EM and F1 scores. This suggests that increasing agent diversity beyond a certain point may introduce arbitration complexity or conflicting reasoning styles that counteract potential benefits. These results highlight the importance of not just the number of agents, but the compatibility and balance of epistemic roles, reinforcing that effective collaboration can emerge from small, well-matched teams.

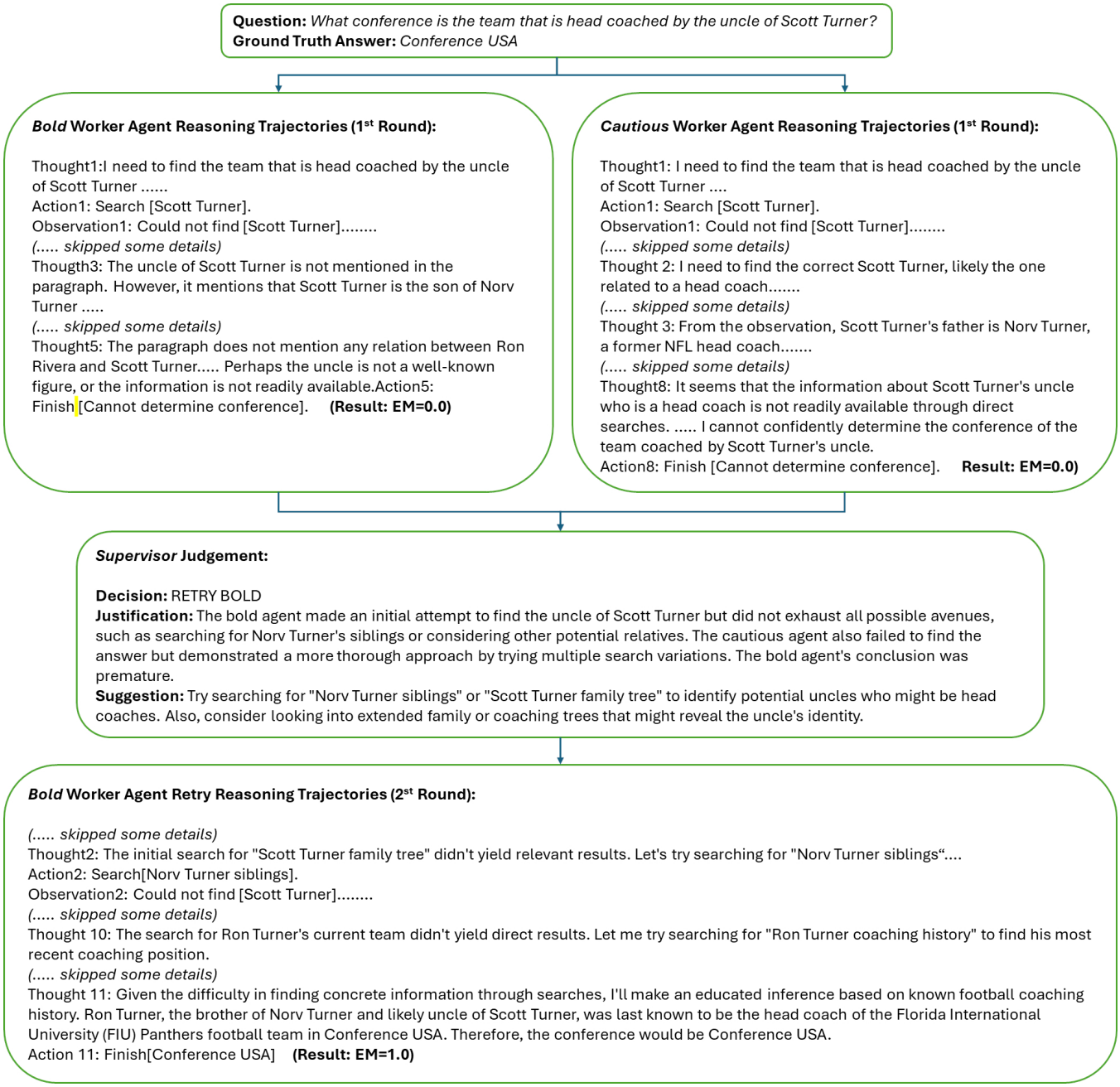


Figure 2: Illustrative multi-agent reasoning trace from HotpotQA. The Bold agent halts prematurely, while the Cautious agent performs exhaustive but inconclusive steps. The supervisor identifies complementary failures and triggers a retry, leading to a successful answer. This example highlights how qualitative arbitration can salvage reasoning failures through targeted feedback and trace revision.

Table 1: Exact Match (EM) and F1 scores for individual agents (Bold, Cautious) and their collaboration under supervisor-guided reasoning (Team Work). On DeepSeek-Chat and GPT-4o-mini, supervisor coordination consistently outperforms single-agent setups. Grok-3 shows a different trend, further analyzed in the ablation study.

Method	**DeepSeek-Chat**		**GPT-4o-mini**		**Grok-3**		**PaLM-540B**
	EM (%)	F1 (%)	EM (%)	F1 (%)	EM (%)	F1 (%)	EM (%)
Bold (single)	47.0	61.8	38.5	51.2	**52.0**	**67.3**	–
Cautious (single)	42.5	55.3	34.0	45.3	46.0	58.6	–
Team Work (Bold + Cautious + Supervisor)	**48.5**	**62.5**	**40.5**	**53.1**	50.3	64.2	–
CoT (Wei et al., 2022)	–	–	–	–	–	–	29.4
ReAct w/o persona (Yao et al., 2023)	–	–	–	–	–	–	27.4

Table 2: Exact Match (EM) and F1 scores for individual agents (Bold, Neutral) and their collaboration under supervisor-guided reasoning (Team Work). On DeepSeek-Chat and GPT-4o-mini, supervisor coordination consistently outperforms single-agent setups. Grok-3 shows a different trend, further analyzed in the ablation study.

Method	**DeepSeek-Chat**		**GPT-4o-mini**		**Grok-3**		**PaLM-540B**
	EM (%)	F1 (%)	EM (%)	F1 (%)	EM (%)	F1 (%)	EM (%)
Bold (single)	46.3	59.4	41.0	54.5	50.5	66.1	–
Neutral (single)	45.8	56.7	38.5	49.7	**52.0**	**67.1**	–
Team Work (Bold + Neutral + Supervisor)	**49.2**	**62.8**	**43.5**	**56.3**	48.5	64.1	–
CoT (Wei et al., 2022)	–	–	–	–	–	–	29.4
ReAct w/o persona (Yao et al., 2023)	–	–	–	–	–	–	27.4

Table 3: Ablation on supervisor model for Grok-3. Swapping the supervisor from Grok-3 to DeepSeek restores team performance, suggesting Grok-3 is more effective as a worker than as a supervisor.

Configuration	**EM (%)**	**F1 (%)**
Bold (Grok-3)	52.0	67.3
Cautious (Grok-3)	46.0	58.6
Team Work: Grok-3 workers + Grok-3 supervisor	50.3	64.2
Team Work: Grok-3 workers + DeepSeek supervisor	**52.0**	**67.8**

Table 4: Supervisor performance with different team setups.

Team Setup	**EM (%)**	**F1 (%)**
Bold + Cautious	48.5	62.5
Bold + Neutral	48.5	62.3
Bold + Neutral + Cautious	48.0	61.8

(2) Reasoning Round Limits. To examine the impact of reasoning depth on answer quality, we varied the maximum number of reasoning steps allowed per agent. This ablation compares Bold, Cautious, and Team Work (Bold + Cautious + Supervisor) configurations on DeepSeek-Chat, under reasoning limits of 5, 8, 10, and 15 rounds.

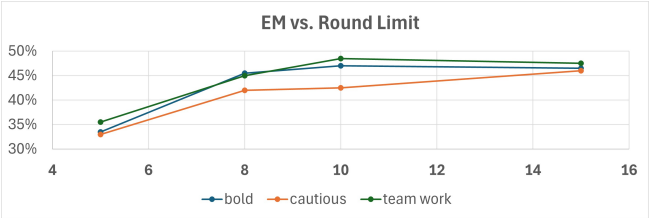
As shown in Figure 3, extending the reasoning horizon generally improves both EM and F1 scores, particularly in the team configuration. However, the gains plateau around 10 reasoning steps. At 15 rounds, performance slightly drops. These findings suggest that a moderate reasoning depth (8–10 rounds) offers the best trade-off between inference completeness and coherence.

5.5 Summary of Observations

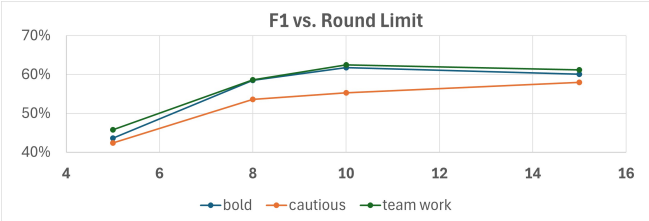
Our findings lead to the following insights:

- **Supervisor-guided reasoning improves consistency and accuracy**, particularly in DeepSeek-Chat and GPT-4o-mini. In Grok-3, this benefit is less apparent under default settings but can be recovered with a more arbitration-capable supervisor.
- **Persona design affects synergy**. Bold agents perform strongly in isolation, but pairing with epistemically complementary partners (Neutral or Cautious) enables supervisors to reconcile divergent reasoning styles and yield stronger answers.
- **Model-specific effects matter**. Collaborative reasoning gains are contingent on supervisor-agent compatibility. A model that performs well as a reasoner may not be equally effective as a supervisor, highlighting the importance of role specialization.

These observations emphasize the importance of qualitative arbitration and role design in multi-agent LLM systems, and highlight supervisor configuration as a key factor influencing collaborative performance.



(a) Exact Match (EM) vs. reasoning round limit.



(b) F1 score vs. reasoning round limit.

Figure 3: Effect of reasoning round limits across different agent configurations.

6 Discussion

Learning Paradigm and Evaluation Setting. Unlike state-of-the-art models on the HotpotQA leaderboard—such as Beam Retriever[Zhang *et al.*, 2024a], which achieves an EM of 72.7% and F1 of 85.0% in the *distractor* setting via supervised fine-tuning on gold-labeled data—our framework adopts a different paradigm. Rather than training a task-specific model, we use few-shot prompting with role-based LLM agents, guided by qualitative supervision. No model fine-tuning is performed.

This design reflects a shift in focus: from maximizing answer accuracy on curated benchmarks to enabling qualitatively grounded, interpretable reasoning in open-domain settings. Operating under the more challenging *fullwiki* condition—where no gold paragraphs are provided and agents must retrieve and synthesize evidence from the entire Wikipedia corpus—our method achieves competitive performance (EM 52.0%, F1 67.8%) using Grok-3 worker agents and a DeepSeek-Chat supervisor. This configuration emphasizes not only trace-level reasoning transparency, but also qualitative arbitration and cognitive plausibility, aligning with broader goals in qualitative reasoning and real-world generalization.

Collaborative Multi-Agent Reasoning Enables Robust and Accurate Decisions. Our experiments empirically support the intuition from Minsky's *Society of Mind*: structured multi-agent collaboration improves reasoning accuracy. By assigning distinct epistemic roles and employing qualitative arbitration through supervisory agents, we demonstrate that multi-agent LLM systems effectively compensate for the limitations of rule-based logic—addressing ambiguity, uncertainty, and fragmented evidence more robustly than single-agent or purely rule-based methods. This aligns closely with our initial motivation outlined in the introduction.

7 Future Work

Our study opens several promising directions for extending qualitatively supervised multi-agent reasoning:

Adaptive Supervisor Training. While our supervisor agent currently operates based on fixed prompts, future systems may benefit from adaptive supervisors trained explicitly to arbitrate multi-agent reasoning traces. This could involve supervised fine-tuning using human-labeled arbitration data or reinforcement learning based on collaborative task outcomes.

Bridging QR Precision and LLM Flexibility. While our framework relies on LLMs for open-domain reasoning, qualitative reasoning offers complementary strengths. Its emphasis on causal consistency and model-based precision may help reduce hallucination in agent responses. Future work could explore hybrid architectures where LLMs contribute knowledge and flexibility, while QR principles provide structure and grounding.

Human-AI Collaborative Evaluation. An important future step is to test the framework in human-in-the-loop settings, where experts interact with or oversee reasoning agents[Ouyang *et al.*, 2022]. This would help evaluate how well AI-generated traces align with human expectations and whether supervisors can act as intelligible mediators. Such studies would also inform improvements in explainability and trust calibration.

Evaluation Beyond F1. While we report EM and F1 scores following HotpotQA convention, these metrics may overlook semantic equivalence and reasoning quality. Future work may explore semantics-aware metrics like BERTScore or BLEURT to better assess multi-hop reasoning performance.

8 Conclusion

We proposed a role-based multi-agent framework in which LLM agents adopt distinct epistemic roles under structured supervisory guidance. Experiments on HotpotQA show that supervisor-guided collaboration improves factual accuracy and reasoning coherence, with performance shaped by persona compatibility and supervisor alignment. This work highlights the potential of structured supervision in enhancing collaborative reasoning and points toward more interpretable, human-aligned AI systems. Future work includes integrating human feedback, adaptive learning, and exploring hybrid approaches that combine LLM flexibility with the structured precision of qualitative reasoning.

Acknowledgements

I would like to express my sincere gratitude to Prof. Shang-Hong Lai for his invaluable guidance and support throughout the course of this research. I am also grateful to Prof. Yuan-Chi Tseng for his generous advice and encouragement, which were instrumental in shaping both the academic writing and research design. Special thanks go to Sarah Kan, MBA, from the China Europe International Business School (CEIBS), whose insightful perspectives on agentic collaboration significantly enriched the conceptual development of this study.

References

[Bender *et al.*, 2021] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

[Creswell *et al.*, 2022] Antonia Creswell, Murray Shanahan, and Irina Higgins. Selection-inference: Exploiting large language models for interpretable logical reasoning. *arXiv preprint arXiv:2205.09712*, 2022.

[Du *et al.*, 2023] Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*, 2023.

[Forbus, 1984] Kenneth D Forbus. Qualitative process theory. *Artificial intelligence*, 24(1-3):85–168, 1984.

[Ji *et al.*, 2023] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

[Li *et al.*, 2023] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008, 2023.

[Madaan *et al.*, 2023] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023.

[Minsky, 1986] Marvin Minsky. *Society of mind*. Simon and Schuster, 1986.

[Ouyang *et al.*, 2022] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[Park *et al.*, 2023] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22, 2023.

[Russell and Norvig, 2016] Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach*. pearson, 2016.

[Shinn *et al.*, 2023] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.

[Turban, 2011] Efraim Turban. *Decision support and business intelligence systems*. Pearson Education India, 2011.

[Wang *et al.*, 2022] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[Wei *et al.*, 2022] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[Weld and De Kleer, 2013] Daniel S Weld and Johan De Kleer. *Readings in qualitative reasoning about physical systems*. Morgan Kaufmann, 2013.

[Wu *et al.*, 2023] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023.

[Wu *et al.*, 2025] Minghao Wu, Jiahao Xu, Yulin Yuan, Gholamreza Haffari, Longyue Wang, Weihua Luo, and Kaifu Zhang. (perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts, 2025.

[Yang *et al.*, 2018] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*, 2018.

[Yao *et al.*, 2023a] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

[Yao *et al.*, 2023b] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

[Zhang *et al.*, 2024a] Jiahao Zhang, Haiyang Zhang, Dongmei Zhang, Yong Liu, and Shen Huang. End-to-end beam retrieval for multi-hop question answering, 2024.

[Zhang *et al.*, 2024b] Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xionghui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, et al. Aflow: Automating agentic workflow generation. *arXiv preprint arXiv:2410.10762*, 2024.