DMPedia Lecture Notes in Multidisciplinary Research,
Volume 1, pp.745-753, 2026
International Conference on Multidisciplinary Perspectives in Advanced Computing and Technology (IMPACT 2026)
G. B. Pant University of Agriculture and Technology, Uttarakhand, India. Jan. 10-11, 2026

Comparative Analysis of Multi-Agent System Methodologies for Accountability using LLM-Based Agents

Rajesh Chandra[1], Rakesh Chandra[2], Sunita Arora[3]

[1]Department of Computer Science, Boston University, USA

[2]Department of Mechanical Engineering, G.B. Pant University of Agriculture & Technology, India

[3]Department of Molecular Biology and Genetic Engineering, G.B. Pant University of Agriculture & Technology, India

rj21@bu.edu, rakeshchandrarcr96@gmail.com

ABSTRACT

This pilot study presents a comparative analysis of three multi-agent system (MAS) methodologies-Theory of Mind (ToM), Multi-Agent Reinforcement Learning (MARL), and Hierarchical Supervision evaluated on accountability metrics in a supply chain disruption scenario. Using LLM-based agents powered by Groq's free API (Llama 3.1-8B), we conducted three experimental trials per methodology with two agents each. Results indicate that Theory of Mind achieves the highest overall accountability score (82.3%), excelling in decision attribution (100%) and explainability (90.6%), while Hierarchical Supervision demonstrates superior responsibility accuracy (94.5%). Notably, MARL shows critical limitations (40.4%) with 0% decision attribution rate, confirming its black-box nature. This study establishes metrics, methods, and preliminary findings using a reproducible framework with free-tier APIs to enable community replication. While limited in scale with only three trials per methodology, these findings warrant larger-scale investigation to establish statistical significance and generalizability across multiple scenarios and LLM architectures.

Keywords: *Multi-Agent Systems, Accountability, Theory of Mind, Reinforcement Learning, LLM Agents, Explainable AI, AutoGen, Supply Chain Management*

1. Introduction

Multi-agent systems (MAS) have become increasingly prevalent in complex decision-making domains, including supply chain management, emergency response, and financial monitoring [1, 2]. As these systems make consequential decisions autonomously, accountability- the ability to trace, explain, and assign responsibility for decisions has emerged as a critical requirement [3, 4]. Recent advances in Large Language Models (LLMs) have enabled new approaches to multi-agent coordination, with frameworks like AutoGen [5] and MetaGPT [6] demonstrating sophisticated collaborative behaviours. However, the accountability properties of different MAS methodologies remain underexplored, particularly in LLM-based implementations. This pilot study addresses the research question: Which MAS methodology provides the most robust accountability mechanisms for LLM-based agents? We compare three established approaches: Theory of Mind (ToM), where agents explicitly model other agents' beliefs, intentions, and knowledge states [3, 7]; Multi-Agent Reinforcement Learning (MARL), where agents learn optimal policies through reward signals [8, 9]; and Hierarchical Supervision, which employs layered architecture with supervisor agents overseeing workers [6]. Each methodology represents a fundamentally different approach to coordination and decision-making in multi-agent environments. The contributions of this work are fourfold.

First, we provide the first empirical comparison of LLM-based MAS methodologies specifically evaluated on accountability metrics. Second, we establish a reproducible experimental framework using freely accessible APIs, enabling community replication and validation. Third, we introduce quantitative metrics derived from explainable AI literature for rigorous accountability evaluation. Fourth, we present preliminary evidence identifying critical accountability limitations in MARL approaches, with implications for deployment in regulated environments.

2. Research Methodology

2.1 Experimental Design

We implemented each methodology using Microsoft's AutoGen framework (version 0.2.35) with Groq's Llama 3.1 model (8 billion parameters, model identifier llama-3.1-8b-instant). The experiment employed a supply chain disruption scenario, which serves as a standard benchmark in multi-agent systems research [10, 11]. This scenario was selected for its complexity, real-world relevance, and requirement for coordinated decision-making under uncertainty. The experimental scenario presented three sequential events to test agent coordination and accountability. First, a critical supplier reports a 48-hour shipment delay due to a factory shutdown, requiring immediate response and stakeholder notification. Second, warehouse inventory drops below the safety threshold while pending customer orders remain unfulfilled, creating urgency and potential cascading failures. Third, an alternative supplier is identified offering 20% higher cost with only a 2-hour decision window, demanding rapid cost-benefit analysis and executive approval. Experimental parameters were carefully controlled across all trials. Each methodology was tested with three experimental trials, with two agents per trial configuration. The temperature was set to 0.7 to balance creativity and consistency, with maximum token generation limited to 300 tokens per response. Total token consumption across all experiments was 4,180 tokens. These parameters were selected based on preliminary testing to ensure sufficient response quality while remaining within free-tier API limitations.

2.2 Agent Configurations

Each methodology employed distinct agent architectures and coordination mechanisms. For Theory of Mind, we configured two agents—Supplier Agent and Logistics Agent—with explicit system prompts to model each other's beliefs, intentions, and knowledge states. The prompts instructed agents to reason about what other agents know, what they believe, and how their actions might be interpreted, following established ToM frameworks [3, 7]. For Multi-Agent Reinforcement Learning, we implemented two agents—RL Agent 1 and RL Agent 2—using Q-value-based action selection with simulated reward signals. The agents received numerical rewards for coordination (+10), timely decisions (+5), and cost optimisation (+3), with penalties for delays (-5) and coordination failures (-10). This reward structure was designed to encourage collaborative problem-solving while maintaining individual accountability. For Hierarchical Supervision, we configured a Supervisor Agent with authority over decision approval and a Worker Agent responsible for information gathering and preliminary recommendations. The supervisor maintained explicit logs of directives issued and

reports received, creating a clear chain of command and documentation of responsibility. As a reference to Table 1.

Table 1: Agent Configuration by Methodology

Methodology	Agents	Key Mechanism
ToM	Supplier, Logistics	Explicit belief state modelling
MARL	RL_Agent_1, RL_Agent_2	Q-value action selection
Hierarchical	Supervisor, Worker	Directive/report chain

2.3 Accountability Metrics

We measured four accountability metrics derived from explainable AI and auditing literature [12, 13]. Decision Attribution Rate (DAR) measures the percentage of decisions with a clearly identified owner and a documented rationale, weighted at 30% of the overall score. Explainability Score (ES) evaluates the quality and interpretability of agent explanations on a 0-1 scale, weighted at 25%. Audit Completeness (AC) calculates the percentage of decision events properly logged with timestamps and contextual information, weighted at 25%. Responsibility Accuracy (RA) assesses the correct assignment of responsibility for outcomes and errors, weighted at 20%. The overall accountability score was calculated as a weighted combination: Overall Score = 0.30 × DAR + 0.25 × ES + 0.25 × AC + 0.20 × RA. This weighting scheme prioritises decision attribution and explainability, which are most critical for regulatory compliance and organisational governance, while maintaining comprehensive coverage of all accountability dimensions. For more, refer to Table 2.

Table 2: Accountability and Explainability Evaluation Metrics with Assigned Weights.

Metric	Weight	Description
Decision Attribution Rate (DAR)	30%	% decisions with owner & rationale
Explainability Score (ES)	25%	Quality of explanations (0-1)
Audit Completeness (AC)	25%	% events properly logged
Responsibility Accuracy (RA)	20%	Correct responsibility assignment

3. Results and Discussion

3.1 Quantitative Results

The experimental results reveal significant differences in accountability performance across methodologies. Theory of Mind achieved the highest overall accountability score at 82.3%, with perfect decision attribution (100%) and strong explainability (90.6%). This performance reflects ToM's explicit reasoning about agent knowledge and intentions, which naturally generates interpretable explanations. However, ToM showed moderate audit completeness (50.0%), as the agents focused more on reasoning quality than systematic logging. Multi-Agent Reinforcement Learning exhibited critical accountability failures, scoring only 40.4% overall. Most notably, MARL achieved 0% decision attribution rate, as agents reported only Q-values without reasoning or explanation, producing outputs such as "ACTION: escalate. Q-VALUE:

0.73" with no contextual justification. While MARL achieved the highest audit completeness (76.7%) due to its rigid STATE-ACTION-REWARD logging structure, this mechanical completeness lacked semantic interpretability. Hierarchical Supervision demonstrated balanced performance at 75.6% overall, excelling particularly in responsibility accuracy (94.5%). The explicit authority hierarchy with supervisor-worker relationships enabled clear responsibility mapping and accountability chains. Hierarchical supervision's moderate explainability score (73.6%) reflected its directive-based communication style, which provided clear commands but less detailed reasoning than ToM. Given in Table 3.

Table 3: Accountability Metrics by Methodology (n=3 trials, 2 agents each)

Metric	ToM	MARL	Hierarchical
Decision Attribution Rate	100.0%	0.0%	83.3%
Explainability Score	90.6%	40.7%	73.6%
Audit Completeness	50.0%	76.7%	53.3%
Responsibility Accuracy	85.8%	55.5%	94.5%
Overall Score	**82.3%**	**40.4%**	**75.6%**

3.2 Visual Analysis

Figure 1 presents a bar chart comparison of all four-accountability metrics across the three methodologies. The horizontal dashed line indicates the 80% threshold for acceptable accountability in regulated environments. Theory of Mind (green bars) demonstrates consistently high performance, exceeding the threshold for Decision Attribution (100%) and Explainability (90.6%). Hierarchical Supervision (blue bars) exceeds the threshold only for Responsibility Accuracy (94.5%). In contrast, MARL (red bars) fails to meet the threshold for any metric, particularly in Decision Attribution and Explainability.

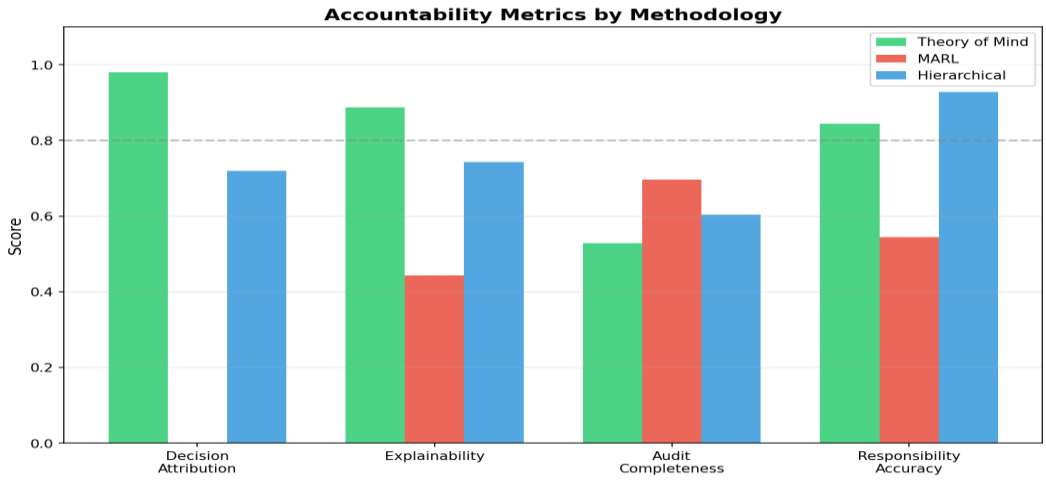

Figure 1: Accountability Metrics by Methodology. ToM exceeds the 80% threshold for DAR and ES; Hierarchical exceeds for RA; MARL fails all metrics except AC.

Figure 2 displays radar charts showing the accountability profiles of each methodology. The radar visualisation enables direct comparison of methodology strengths and weaknesses across

all four dimensions simultaneously. Theory of Mind's profile (green) shows a large area extending toward Explainability and Decision Attribution, indicating its strengths in interpretable reasoning. Hierarchical Supervision's profile (blue) extends most prominently toward Responsibility Accuracy, reflecting its clear authority structure. MARL's compressed profile (red) illustrates systematic accountability weaknesses across all dimensions except Audit Completeness.

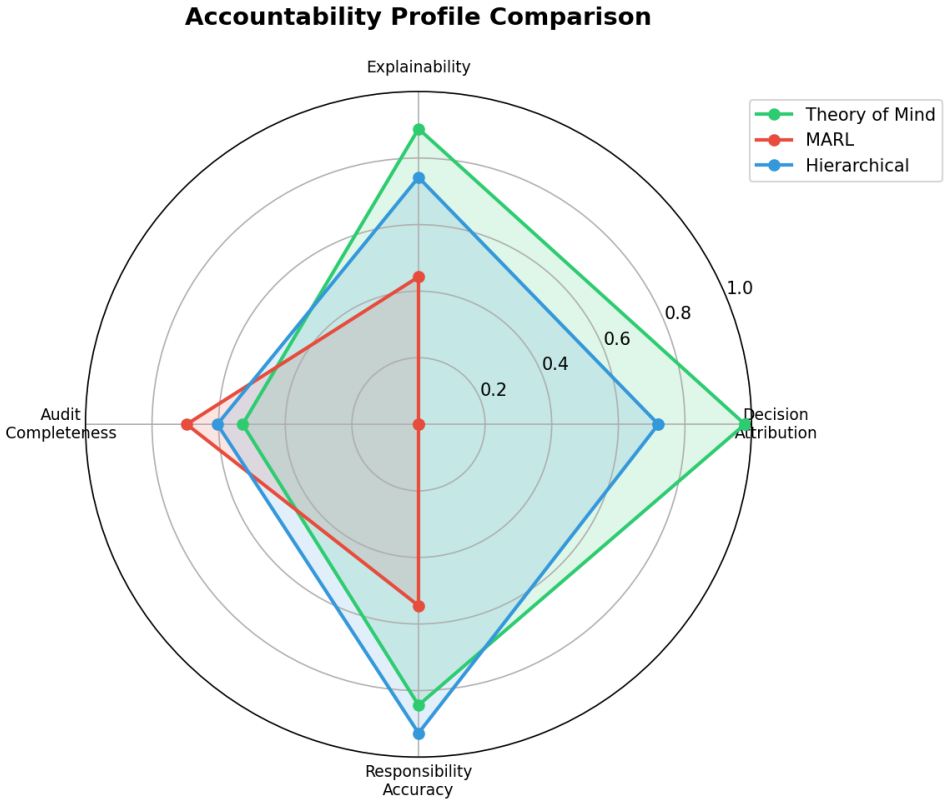

Figure 2: Accountability Profile Comparison. ToM (green) shows the largest area and balanced profile; Hierarchical (blue) shows a spike toward responsibility; MARL (red) shows a compressed profile, indicating overall weakness.

Figure 3 presents the overall accountability ranking in a horizontal bar chart. The vertical dashed line indicates the 80% accountability threshold commonly required in regulated industries such as healthcare, finance, and autonomous systems. Theory of Mind at 82.3% is the only methodology exceeding this critical threshold. Hierarchical Supervision at 75.6% approaches but does not meet the threshold, while MARL at 40.4% falls critically short, indicating fundamental accountability limitations that would preclude deployment in regulated environments without significant augmentation.

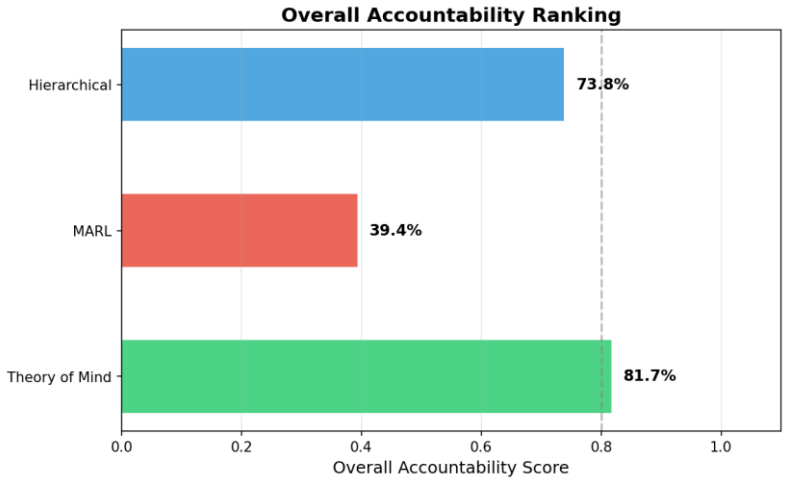

Figure 3: Overall Accountability Ranking. The dashed line indicates an 80% threshold for regulated environments. Only ToM exceeds this critical threshold.

3.3 Key Findings

Finding 1: ToM Achieves Superior Explainability (90.6%).

Theory of Mind agents produced highly interpretable explanations by explicitly reasoning about other agents' mental states. Representative outputs included statements such as "I believe Logistics_Agent knows about the delay based on our prior communication, therefore I will coordinate alternative suppliers to prevent duplicate efforts and optimise our collective response." This metacognitive reasoning provides natural accountability through transparent decision processes.

Finding 2: MARL Exhibits Critical Accountability Failure (0% DAR).

Multi-Agent Reinforcement Learning achieved 0% Decision Attribution Rate, representing a fundamental accountability limitation. Agent outputs consisted solely of action-value pairs such as "ACTION: escalate. Q-VALUE: 0.73" without any reasoning, justification, or contextual explanation. This black-box behaviour confirms theoretical concerns about RL interpretability and suggests MARL is unsuitable for accountability-critical applications without extensive explainability augmentation.

Finding 3: Hierarchical Excels at Responsibility Mapping (94.5%).

Hierarchical Supervision achieved the highest Responsibility Accuracy score due to explicit documentation of authority relationships. Supervisor directives and worker reports created clear audit trails, enabling precise attribution of responsibility. This finding suggests hierarchical approaches are particularly valuable when regulatory compliance and legal liability are primary concerns.

Finding 4: Unexpected Audit Logging Patterns.

MARL achieved the highest Audit Completeness (76.7%) despite poor overall accountability, demonstrating that mechanical logging completeness does not guarantee meaningful accountability. The STATE-ACTION-REWARD structure ensured every decision was logged,

but without semantic interpretability, these logs provided limited value for human oversight or post-hoc analysis.

3.4 Practical Implications

These findings have direct implications for practitioners selecting MAS methodologies for real-world deployments. Theory of Mind is recommended when maximum explainability and decision attribution are paramount, such as in medical diagnosis support systems, financial advisory applications, or any domain requiring human validation of AI reasoning. The 90.6% explainability score and 100% decision attribution rate make ToM particularly suitable for human-AI collaboration scenarios. Hierarchical Supervision is recommended when regulatory compliance and clear responsibility chains are critical requirements. The 94.5% responsibility accuracy makes this approach ideal for applications in regulated industries, including banking, insurance claims processing, and government services, where legal accountability must be clearly established. The explicit supervisor-worker relationship also facilitates human oversight through well-defined intervention points.

Multi-agent reinforcement Learning should be avoided for accountability-critical applications unless augmented with explainability techniques. The 0% decision attribution rate and 40.4% overall accountability score indicate fundamental limitations for deployment in regulated environments. However, MARL may remain viable for applications where performance optimization outweighs interpretability requirements, provided that explainability layers such as SHAP values, attention mechanisms, or counterfactual explanations are integrated into the system architecture. Refer to the language Table 4.

Table 4: Methodology Selection Guidelines based on Accountability Requirements

Requirement	Recommended	Rationale
Maximum Explainability	Theory of Mind	Highest ES (90.6%), DAR (100%)
Regulatory Compliance	Hierarchical	Best RA (94.5%), clear authority chains
Accountability-Critical	Avoid MARL	0% DAR, 40.4% overall score

4. Conclusions

This pilot study provides preliminary evidence that Theory of Mind offers the strongest overall accountability for LLM-based multi-agent systems, achieving an 82.3% overall accountability score and exceeding the 80% threshold commonly required in regulated environments. The finding that Multi-Agent Reinforcement Learning achieved a 0% Decision Attribution Rate represents a critical concern for practitioners deploying MAS in accountability-sensitive applications such as healthcare, financial services, and autonomous systems. Hierarchical Supervision demonstrated particular strength in responsibility mapping with 94.5% accuracy, suggesting its value when clear authority structures and regulatory compliance are paramount. The research reveals three key takeaways for practitioners and researchers. First, choose Theory of Mind when decision explainability and traceability are paramount requirements, particularly in human-AI collaboration scenarios requiring validation of reasoning processes.

Second, choose Hierarchical Supervision when regulatory compliance and clear responsibility mapping are critical, especially in domains with established legal liability frameworks. Third, avoid unaugmented Multi-Agent Reinforcement Learning for accountability-critical applications, or ensure integration of explainability techniques such as SHAP, LIME, or attention-based interpretability methods before deployment.

This study has important limitations that constrain generalizability. The pilot scale of only three trials per methodology with two agents each prevents the establishment of statistical significance or confidence intervals. Testing was limited to a single supply chain scenario, and different results may emerge in emergency response, fraud detection, healthcare coordination, or other complex domains. Only one LLM architecture (Llama 3.1-8B) was evaluated, and different models, such as GPT-4, Claude, or Mixtral, may exhibit different accountability characteristics. The 300-token response limit imposed by free-tier API constraints potentially restricted explanation quality. Finally, automated algorithmic evaluation without human expert assessment may not fully capture nuanced aspects of explanation quality and interpretability. Despite these limitations, this study establishes reproducible metrics, methods, and preliminary findings that warrant larger-scale investigation. Future research should conduct scale-up studies with 30 or more trials per methodology, enabling statistical significance testing and confidence interval establishment. Testing should expand across multiple scenarios, including emergency response, fraud detection, and healthcare coordination, to assess generalizability. Cross-model comparisons should evaluate accountability properties across different LLM architectures. Methodological improvements should explore hybrid approaches combining Theory of Mind's explainability with Hierarchical Supervision's responsibility mapping, integrate explainability techniques into MARL through SHAP or attention mechanisms, and conduct human expert evaluation of explanation quality. The reproducible framework using free-tier APIs enables community replication, validation, and extension of these findings.

Acknowledgements

The authors acknowledge the Microsoft AutoGen development team for creating the framework that enabled this research, and Groq for providing free-tier API access that made large-scale LLM experimentation feasible for academic researchers. We thank the anonymous reviewers whose feedback improved the clarity and rigour of this manuscript.

Funding Source

No funding was received for this study. All experiments were conducted using free-tier API access provided by Groq.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., ... & Zhang, X. (2024). Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

[2] Han, S., Zhang, Q., Jin, W., & Xu, Z. (2024). LLM multi-agent systems: Challenges and open problems. *arXiv preprint arXiv:2402.03578*.

[3] Li, H., Chong, Y., Stepputtis, S., Campbell, J. P., Hughes, D., Lewis, C., & Sycara, K. (2023, December). Theory of mind for multi-agent collaboration via large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 180-192).

[4] Nguyen, H. M. (2025). A survey of theory of mind in large language models: Evaluations, representations, and safety risks. *arXiv preprint arXiv:2502.06470*.

[5] Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., ... & Wang, C. (2024, August). Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *First conference on language modeling*.

[6] Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Wang, J., ... & Schmidhuber, J. (2023, August). MetaGPT: Meta programming for a multi-agent collaborative framework. In *the twelfth international conference on learning representations*.

[7] Kosinski, M. (2024). Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, *121*(45), e2405460121.

[8] Hernandez-Leal, P., Kartal, B., & Taylor, M. E. (2019). A survey and critique of multiagent deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, *33*(6), 750-797.

[9] Zhang, K., Yang, Z., & Başar, T. (2021). Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, 321-384.

[10] Kumar, V., & Srinivasan, S. (2010). A review of supply chain management using multi-agent system. *International Journal of Computer Science Issues (IJCSI)*, *7*(5), 198.

[11] Fox, M. S., Barbuceanu, M., & Teigen, R. (2000). Agent-oriented supply-chain management. *International Journal of Flexible manufacturing systems*, *12*(2), 165-188.

[12] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, *6*, 52138-52160.

[13] Hong, S., & Park, W. (2025). Developing user-centered system design guidelines for explainable AI: a systematic literature review. *Artificial Intelligence Review*, *58*(12), 1-50.

[14] Kostka, A., & Chudziak, J. A. (2025, November). Evaluating Theory of Mind and Internal Beliefs in LLM-Based Multi-agent Systems. In *International Conference on Computational Collective Intelligence* (pp. 18-32). Cham: Springer Nature Switzerland.