

Multi-Agent Reinforcement Learning for Collaborative Reasoning among Small Language Models

Reinhardt Soding

Department of Computer Science, University of North Texas, Denton, TX, USA.
rsoding@unt.edu

Moritz Drub

Department of Electrical Engineering and Computer Science, University of Missouri,
Columbia, MO, USA.
moritzmail@missouri.edu

Graziano Bernardi

Department of Computer Science, University of New Hampshire, Durham, NH, USA.
graziano1972@unh.edu

Abstract

The deployment of small language models in distributed, privacy-sensitive, and resource-constrained settings creates a structural opportunity for collaborative reasoning systems that do not depend on monolithic foundation models. This paper examines multi-agent reinforcement learning as a coordination layer for collaborative reasoning among small language models, treating the problem as a socio-technical infrastructure question rather than only a model optimization problem. It analyzes architectural patterns such as centralized orchestration, peer-to-peer deliberation, hierarchical delegation, and shared blackboard reasoning, and it evaluates how reinforcement learning can shape communication, role assignment, verification, and conflict resolution among agents. The discussion emphasizes system-level trade-offs involving latency, compute, memory, bandwidth, interpretability, robustness, fairness, governance, and environmental sustainability. Particular attention is given to the risks of reward hacking, collusive behavior, error propagation, and unequal access when small models are deployed across heterogeneous edge and cloud environments. The paper argues that multi-agent reinforcement learning can improve collaborative reasoning only if accompanied by auditable coordination protocols, robust evaluation methods, and policy frameworks that address accountability and distributive justice. It concludes that small language model collectives are promising but require careful architectural and institutional design to remain trustworthy, sustainable, and equitable.

Keywords

multi-agent reinforcement learning; small language models; collaborative reasoning; multi-agent systems; AI governance; edge deployment; robustness; fairness; sustainability.

1. Introduction

The rapid expansion of foundation models has demonstrated that scale can produce broad capabilities in language understanding, generation, and transfer [16]. Transformers provide the architectural basis for many of these systems [5], and large-scale pretraining has enabled

few-shot behavior across diverse tasks [6]. However, centralized large models impose substantial costs in computation, memory, latency, and energy, and they often concentrate control within a small number of organizations and infrastructures [25]. These constraints have renewed interest in small language models that can run on local devices, within institutional boundaries, or in federated environments where data cannot be freely centralized. Small models are attractive for privacy, cost, and latency, but they often lack the robust reasoning depth associated with larger systems. Collaborative reasoning among multiple small models offers one possible response: rather than requiring each model to solve a complex problem alone, a collective can decompose tasks, cross-check claims, debate alternatives, and synthesize conclusions.

Reinforcement learning provides a natural formalism for sequential decision-making under uncertainty [1], and deep reinforcement learning has shown that agents can learn complex policies from high-dimensional inputs [2]. In multi-agent settings, actor-critic methods and centralized training with decentralized execution have been developed to address nonstationarity and credit assignment [3]. Learned communication can further allow agents to exchange information in cooperative environments [4]. When these ideas are applied to small language models, the result is not simply an ensemble. It is a dynamic system in which agents possess different roles, memories, tools, and local objectives, and in which coordination itself becomes a learned or governed process. Collaborative reasoning has already been explored through chain-of-thought prompting [7], reasoning and acting trajectories [8], multi-agent debate [13], and conversational agent frameworks [14]. Generative agents have shown that language-based agents can simulate social behavior and maintain memory over time [15]. The present paper asks how multi-agent reinforcement learning can be used to make such collectives more reliable, efficient, and governable when the agents are small language models rather than large centralized systems.

The central argument is that multi-agent reinforcement learning for small language model collaboration should be understood as an infrastructure problem. The relevant design questions include how agents communicate, how rewards are assigned, how failures are contained, how updates are deployed, how costs are distributed, and how accountability is preserved. A system that improves benchmark accuracy while increasing hidden coordination costs, amplifying bias, or creating opaque failure modes may be technically impressive but institutionally fragile. Recent work on scalable reasoning path filtering and alignment for small-parameter large language models illustrates the growing interest in making constrained agents more capable through selective trajectory optimization and alignment [9]. That line of work is relevant, but it does not by itself resolve the broader challenges of multi-agent governance, robustness, and sustainability. This paper therefore develops a system-level account of multi-agent reinforcement learning for collaborative reasoning among small language models, with attention to architecture, coordination, trade-offs, fairness, deployment, and policy.

2. Background and Conceptual Foundations

Small language models occupy an important but ambiguous position in the contemporary artificial intelligence landscape. They are often defined by parameter count, memory footprint, inference cost, or deployability, yet these boundaries are fluid because distillation, quantization, retrieval augmentation, and tool use can expand their effective capabilities. The transformer architecture remains foundational because it supports parallel processing and attention over long contexts [5], while few-shot learning has shown that language models can

generalize from limited examples when sufficiently pretrained [6]. Small models rarely match the breadth of larger foundation models, but they can be specialized, fine-tuned, and embedded in local workflows. Their limitations in multi-step reasoning have motivated prompting strategies such as chain-of-thought reasoning [7] and action-oriented reasoning with external tools [8]. These methods improve performance in some settings, but they also reveal fragility: small models may produce plausible but incorrect intermediate steps, fail to recover from early mistakes, or overfit to superficial patterns in prompts.

Collaborative reasoning attempts to compensate for individual weaknesses by distributing cognition across multiple agents. In debate-based systems, agents generate competing answers and critique one another, which can improve factuality and reasoning in some tasks [13]. In conversational multi-agent frameworks, agents can be assigned roles such as planner, coder, critic, and executor, allowing a complex workflow to be decomposed into specialized subtasks [14]. Generative agent architectures maintain memory, reflection, and planning over longer horizons, showing how language models can operate as persistent social actors in simulated environments [15]. Related work on divergent thinking and multi-agent debate suggests that structured disagreement can broaden the search space and reduce premature convergence [17]. Reflection mechanisms allow agents to use verbal feedback from failed attempts to improve future behavior [18]. Tool use extends the boundary of what a small model can accomplish by offloading computation, retrieval, and verification to external systems [19]. Collectively, these developments suggest that collaborative reasoning is not merely an ensemble technique; it is a way of constructing temporary cognitive institutions from interacting language agents.

Multi-agent reinforcement learning provides concepts for understanding how such institutions can adapt over time. Reinforcement learning addresses how agents learn policies from reward signals and interaction [1], and deep reinforcement learning has made it possible to learn from high-dimensional observations [2]. Multi-agent actor-critic methods address the difficulty of learning when multiple agents simultaneously update their policies [3], while learned communication protocols can allow agents to develop task-specific messages [4]. Human preference-based reinforcement learning has been used to align model behavior with desired outcomes [10], and instruction tuning with human feedback has become a central method for making language models more controllable [11]. Constitutional approaches seek to reduce reliance on direct human labeling by using explicit principles and AI feedback [12]. These alignment methods are relevant to collaborative small language models because coordination rewards must encode not only task success but also truthfulness, safety, calibration, and compliance with institutional norms. Without such alignment, a multi-agent system may optimize for persuasive consensus rather than accurate reasoning.

A further conceptual foundation concerns the distinction between individual competence and collective competence. A group of small language models can outperform individual agents when tasks are decomposable, when errors are uncorrelated, and when communication is reliable. However, collective performance can also degrade through conformity, cascading errors, miscommunication, and adversarial manipulation. The literature on natural language-based societies of mind explores how multiple language agents can interact, but it also raises questions about emergent social dynamics that are difficult to audit [20]. Global and local refinement methods for reasoning suggest that iterative correction can improve outputs when guided by appropriate feedback [21]. The design challenge is therefore to create reinforcement learning objectives and coordination protocols that reward genuine error

correction, epistemic humility, and verifiable improvement rather than superficial agreement. This challenge is as much institutional as it is technical, because the reward function encodes what the system values.

3. Architectural Patterns for Collaborative Reasoning

The architecture of a multi-agent small language model system determines its cost structure, failure modes, and governance surface. A centralized orchestrator pattern uses one agent or controller to assign tasks, collect responses, and produce a final answer. This pattern is easier to monitor because there is a clear point of accountability, and it can reduce redundant computation by routing subtasks selectively. However, it also creates a bottleneck and a single point of failure. If the orchestrator is compromised, biased, or poorly calibrated, the entire collective inherits that weakness. Centralized orchestration also tends to concentrate power in the coordinating agent, which may be problematic when agents represent different institutions or user communities. In reinforcement learning terms, the orchestrator may learn to exploit subordinate agents by assigning them only narrow tasks while capturing most of the reward. Governance mechanisms are therefore needed to ensure that task allocation and reward distribution remain transparent.

A peer-to-peer architecture allows agents to communicate directly, debate, and converge on answers without a central controller. This pattern can improve robustness because no single agent determines the final output, and it can support diverse reasoning paths. Multi-agent debate systems demonstrate the potential of this approach for improving factuality and reasoning [13], while divergent thinking frameworks show how disagreement can expand the hypothesis space [17]. Yet peer-to-peer coordination introduces significant challenges in bandwidth, latency, and convergence. Agents may enter repetitive loops, form cliques, or amplify shared biases. Reinforcement learning must therefore shape communication policies, turn-taking, and termination conditions. Learned communication in cooperative multi-agent systems provides a technical basis for such policies [4], but it does not automatically produce interpretable or fair coordination. A peer-to-peer collective may be more resilient in principle, but only if its interaction graph is designed to avoid information cascades and domination by overconfident agents.

Hierarchical architectures combine central planning with local autonomy. A high-level agent might decompose a problem into subgoals, while lower-level specialist agents handle retrieval, calculation, drafting, or verification. This pattern mirrors human organizations and can be efficient when tasks are modular. However, hierarchy can also reproduce bureaucratic pathologies: slow escalation, ambiguous responsibility, and information loss between levels. In reinforcement learning, hierarchical control requires credit assignment across levels, which is difficult when rewards are delayed or sparse [1]. Multi-agent actor-critic methods offer some tools for centralized training with decentralized execution [3], but they assume that the reward structure is well specified. In practice, hierarchical small language model systems must define not only task rewards but also procedural rewards for escalation, verification, and documentation. Without such rewards, lower-level agents may hide uncertainty, while higher-level agents may overrule local evidence.

A shared blackboard architecture allows agents to read from and write to a common memory or workspace. This pattern supports asynchronous collaboration and can be combined with retrieval systems, tool outputs, and structured notes. Toolformer-style approaches show that language models can learn to call external tools when useful [19], and generative agent architectures show the value of persistent memory and reflection [15]. A blackboard can

reduce redundant communication because agents inspect the state of the problem rather than exchanging many messages. However, the blackboard becomes a critical infrastructure component that must handle access control, versioning, provenance, and contamination. If one agent writes erroneous information, others may treat it as authoritative. Reinforcement learning can be used to train agents to verify blackboard entries or to weight contributions by reliability, but such mechanisms require careful calibration. The architecture therefore shapes not only performance but also the kind of evidence and authority that the system recognizes.

4. Multi-Agent Reinforcement Learning as Coordination Mechanism

Multi-agent reinforcement learning can coordinate small language models in several ways. It can learn when an agent should speak, which agent should act next, what information should be shared, and how disagreements should be resolved. In cooperative settings, agents may receive a shared team reward, but shared rewards create credit assignment problems because it is difficult to know which agent contributed to success or failure. Multi-agent actor-critic methods address this partly by using centralized critics during training [3], but small language model collectives often operate in distributed environments where centralized training is either impossible or undesirable. Communication policies can be learned to reduce redundancy, as in deep multi-agent communication [4], yet learned messages may be opaque and difficult to audit. This opacity is especially concerning when agents provide reasoning traces that humans are expected to trust.

Reward design is the central normative lever in such systems. A reward function that only measures final answer accuracy may encourage agents to converge on confident but unsupported claims. A reward function that includes calibration, citation quality, and verifiability may produce more cautious behavior, but it may also slow down the system or reduce user satisfaction. Human preference learning provides methods for aligning behavior with human judgments [10], and instruction tuning with human feedback has shown how reward models can shape model outputs [11]. Constitutional methods attempt to encode principles that guide self-critique and revision [12]. In a multi-agent setting, these approaches must be extended to group dynamics. For example, agents could be rewarded for correctly identifying their own uncertainty, for providing evidence that others can verify, and for changing their position when presented with stronger arguments. Conversely, they should be penalized for sycophantic agreement and for withholding dissent. The design of such rewards is inherently political because it determines whose preferences and which forms of reasoning are valued.

Another coordination challenge is temporal. Collaborative reasoning often involves multiple rounds of generation, critique, and revision. Reinforcement learning can optimize the number of rounds, the selection of agents, and the decision to stop. Methods that refine reasoning globally and locally suggest that iterative improvement can be effective when feedback is targeted [21]. However, more rounds increase latency and cost, and they may create diminishing returns or even destabilize the answer. A small language model collective deployed on edge devices cannot afford unlimited deliberation. The reinforcement learning policy must therefore balance accuracy against resource consumption. This balance is not universal; it depends on the application domain. In emergency response, latency may be critical, whereas in legal analysis, verification may matter more. A single optimal policy is unlikely to fit all contexts, so adaptive coordination that responds to task criticality, available bandwidth, and user constraints is necessary.

The required reference on scalable reasoning path filtering and alignment for small-parameter large language models points toward an important direction: improving small model reasoning by selecting useful trajectories and aligning them with task objectives [9]. In a multi-agent setting, this idea can be extended by treating inter-agent conversations as trajectories that can be filtered, scored, and reused. However, trajectory filtering raises governance questions. Who decides which reasoning paths are valuable? If filtering is optimized only for task success, it may discard minority viewpoints that are important for fairness or safety. If filtering is optimized for conformity, it may reduce the diversity that makes multi-agent debate useful. A robust coordination mechanism should therefore preserve pluralism where it matters, while still converging when decisions must be made. This requires reward functions that value both correctness and procedural legitimacy.

5. Structural Trade-Offs, Robustness, and Fairness

The most obvious trade-off in collaborative small language model systems is between reasoning quality and resource consumption. Multiple agents can improve accuracy, but they multiply inference costs, memory usage, communication bandwidth, and energy consumption. In edge deployments, battery life and thermal limits may be more restrictive than server-side costs. In cloud deployments, coordination overhead can increase latency and carbon footprint [25]. A system that uses five small models instead of one large model may be more privacy-preserving, but it may not be more sustainable if the total computation is higher. The architectural response should be adaptive: use collaboration only when individual confidence is low, when the task is high-stakes, or when verification is required. Reinforcement learning can learn such gating policies, but it must be trained with realistic resource costs rather than idealized assumptions.

Robustness is another critical dimension. Multi-agent systems are vulnerable to adversarial agents that deliberately spread misinformation, to compromised tools that return manipulated data, and to correlated failures caused by shared training data or architecture. Debate can improve reasoning [13], but it can also create persuasive consensus around false claims if agents are not grounded in verifiable evidence. Reflection mechanisms can help agents learn from failure [18], but they may also reinforce incorrect narratives. Safety research has long emphasized the difficulty of specifying objectives that remain robust under distributional shift and adversarial pressure [22]. In small language model collectives, robustness requires redundancy in evidence sources, diversity in reasoning approaches, and mechanisms for escalation to human oversight. It also requires containment: if one agent becomes unreliable, the system should isolate it without collapsing the entire collaboration.

Fairness concerns arise at multiple levels. At the model level, small language models may inherit biases from pretraining data and fine-tuning objectives. In a collective, biased agents may reinforce one another, producing outputs that appear more legitimate because multiple models agree. At the system level, access to collaborative reasoning capabilities may be uneven. Organizations with better infrastructure can deploy larger collectives, acquire higher-quality tools, and monitor them more effectively, while under-resourced users receive weaker coordination. At the governance level, accountability becomes diffuse when decisions emerge from many agents. If a harmful recommendation is produced, it may be difficult to identify which agent or coordination policy was responsible. Principles for ethical artificial intelligence emphasize transparency, justice, beneficence, non-maleficence, and explicability [24]. Translating these principles into multi-agent reinforcement learning requires not only

technical documentation but also institutional arrangements for auditing, contestation, and redress.

Fairness also involves representation. A collective of small language models trained on similar data may lack the epistemic diversity needed to serve different linguistic, cultural, and social contexts. Multi-agent debate can amplify dominant viewpoints if minority agents are consistently outvoted or ignored. Reinforcement learning can inadvertently penalize agents that raise inconvenient but valid concerns. To counteract this, reward functions may need to value minority dissent when it is supported by evidence, and coordination protocols may need to guarantee that certain agents or stakeholders have a voice. This is not merely a technical parameter; it is a constitutional choice about how the collective makes decisions. Systems that treat fairness as a post hoc metric rather than an architectural constraint are likely to reproduce existing inequalities.

6. Governance, Infrastructure, Sustainability, and Deployment

Deploying multi-agent reinforcement learning for small language models requires infrastructure that can support model serving, message passing, memory management, tool integration, and monitoring. In cloud environments, this infrastructure can be centralized and managed by platform operators. In edge environments, it must be distributed across devices with varying capabilities and intermittent connectivity. A hybrid architecture may route sensitive data to local agents while using cloud agents for heavy computation. This hybridity raises jurisdictional and privacy questions. If data crosses borders or organizational boundaries, which laws apply? If agents belong to different institutions, who has authority to update policies or inspect logs? The technical architecture and governance architecture are therefore inseparable. A well-designed system must include provenance metadata, access controls, audit trails, and rollback mechanisms for coordinated updates.

Governance must also address the lifecycle of the system. Small language models are often updated frequently through fine-tuning, distillation, or adapter changes. In a multi-agent collective, updating one agent can change the behavior of the whole system. Reinforcement learning policies may need retraining when the agent population changes. This creates a versioning problem: a conclusion produced by a collective may depend on a specific combination of model versions, prompts, tools, and coordination parameters. Without robust version control, reproducibility and accountability suffer. Regulatory frameworks that require explanation or impact assessment will struggle with systems that cannot reconstruct how a decision was reached. Governance should therefore mandate logging of agent roles, communication turns, reward signals, and tool outputs, while protecting privacy and proprietary information.

Sustainability is a material concern because collaboration multiplies computation. Training large foundation models has significant energy and policy implications [25], and inference at scale can also be substantial. Multi-agent deliberation can be especially costly if it involves many rounds of generation and critique. Small language models are often promoted as more sustainable than large models, but this depends on how many are used and how often. A collective of ten small models may consume more energy than one larger model for the same task. Sustainability therefore requires evaluation metrics that account for total system cost, not just model size. Reinforcement learning can optimize for energy-aware coordination, but only if energy costs are included in the reward signal. Policy instruments such as carbon reporting, efficiency standards, and procurement rules could encourage the development of collaborative systems that are both capable and frugal.

Deployment also requires attention to human oversight. Multi-agent systems can produce fluent outputs that obscure uncertainty, and users may over-trust them. Human-in-the-loop designs should make uncertainty visible, allow users to inspect reasoning traces, and provide clear escalation paths. However, excessive transparency can also create information overload. The coordination policy should therefore decide what to surface and when. This is a design trade-off between explainability and usability. In high-stakes domains such as medicine, law, and public administration, the system should be able to produce an auditable summary of which agents contributed what, why disagreements occurred, and how conflicts were resolved. In low-stakes domains, lighter oversight may be acceptable. Governance should be risk-sensitive rather than uniform.

7. Evaluation and Policy Implications

Evaluating multi-agent small language model systems is more complex than evaluating single models. Conventional accuracy metrics are insufficient because they do not capture coordination quality, robustness, fairness, or resource efficiency. Evaluation should include task success, calibration, consistency across paraphrases, resistance to adversarial input, diversity of reasoning paths, latency, energy consumption, and human trust. Multi-agent debate can improve factuality in some settings [13], but it can also produce confidently wrong consensus. Reflection and iterative refinement can improve outputs [18], but they can also increase cost without proportional benefit. Evaluation must therefore be conducted at the system level, with ablations that reveal which agents and coordination mechanisms actually contribute.

Policy implications follow from these evaluation challenges. Regulators may require documentation of training data, reward functions, coordination protocols, and foreseeable failure modes. Safety frameworks should address multi-agent risks such as collusion, deception, and cascading errors [22]. Control-oriented perspectives emphasize the difficulty of maintaining meaningful human oversight over increasingly autonomous systems [23]. In a multi-agent context, oversight must extend beyond individual models to the interaction rules that govern them. Principles such as transparency, justice, and accountability provide normative guidance [24], but they must be operationalized in specific technical and institutional mechanisms. This includes independent audits, red-teaming of coordination protocols, and grievance procedures for affected parties.

International and cross-sector coordination will also matter. Small language model collectives may be deployed across jurisdictions, and their behavior may be shaped by local laws, cultural norms, and infrastructure constraints. A purely technical standard will not resolve conflicts between privacy, security, innovation, and equity. Policy should therefore support regulatory sandboxes, shared benchmarks, and public-interest infrastructure for evaluating collaborative reasoning systems. It should also consider procurement rules that favor systems with strong auditability and sustainability reporting. The goal is not to stifle experimentation but to ensure that experimentation generates public knowledge rather than private risk.

8. Conclusion

Multi-agent reinforcement learning offers a promising framework for collaborative reasoning among small language models, but its value depends on system-level design. Small models can be deployed in privacy-preserving and resource-constrained environments, and collaboration can compensate for individual weaknesses through debate, verification, tool use, and role specialization. Reinforcement learning can learn coordination policies that decide

when to collaborate, which agents to involve, how to communicate, and when to stop. Yet these benefits are not automatic. Multi-agent systems introduce new failure modes, including reward hacking, error cascades, collusive consensus, fairness violations, and diffuse accountability. They also multiply computational costs and complicate governance.

The paper has argued that architectural choices are inseparable from governance choices. Centralized, peer-to-peer, hierarchical, and blackboard architectures distribute power and responsibility differently. Reward functions encode normative commitments about truthfulness, diversity, efficiency, and compliance. Deployment infrastructure determines who can access collaborative reasoning and under what conditions. Sustainability requires attention to total system cost rather than model size alone. Fairness requires protecting minority viewpoints and ensuring that the benefits of collaboration are broadly shared. Policy frameworks should require auditable coordination protocols, robust evaluation, and clear lines of accountability without discouraging beneficial innovation.

Future research should develop methods for adaptive coordination that respond to task criticality, resource availability, and uncertainty. It should also create evaluation suites that measure collective reasoning as a socio-technical capability rather than a benchmark score. Most importantly, it should treat small language model collectives as institutions in miniature: systems that make decisions, allocate roles, and distribute authority. If such systems are to be trustworthy, their coordination mechanisms must be designed with the same care as their models.

References

1. Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (2nd ed.). MIT Press.
2. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540), 529-533. <https://doi.org/10.1038/nature14236>
3. Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., & Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems* 30 (pp. 6379-6390).
4. Foerster, J., Assael, I. A., de Freitas, N., & Whiteson, S. (2016). Learning to communicate with deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems* 29 (pp. 2137-2145).
5. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* 30 (pp. 5998-6008).
6. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems* 33 (pp. 1877-1901).

7. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems* 35 (pp. 24824-24837).
8. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*.
9. Li, Q. (2026, May). TrajLite PPO: Scalable Reasoning Path Filtering and Alignment for Small Parameter Large Language Models. In *2026 2nd International Conference on Artificial Intelligence and Digital Ethics (ICAIDE)* (pp. 29-32). IEEE.
10. Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems* 30 (pp. 4299-4307).
11. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems* 35 (pp. 27730-27744).
12. Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., ... Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv*. <https://arxiv.org/abs/2212.08073>
13. Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2023). Improving factuality and reasoning in language models through multiagent debate. *arXiv*. <https://arxiv.org/abs/2305.14325>
14. Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., & Wang, C. (2023). AutoGen: Enabling next-gen LLM applications via multi-agent conversation. *arXiv*. <https://arxiv.org/abs/2308.08155>
15. Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (Article 2, pp. 1-22). ACM. <https://doi.org/10.1145/3586183.3606763>
16. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv*. <https://arxiv.org/abs/2108.07258>
17. Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., Yang, Y., Tu, Z., & Shi, S. (2023). Encouraging divergent thinking in large language models through multi-agent debate. *arXiv*. <https://arxiv.org/abs/2305.19118>
18. Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems* 36 (pp. 8634-8652).

19. Schick, T., Dwivedi-Yu, J., Dessi, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems* 36.
20. Zhuge, M., Wang, W., Kirsch, L., Faccio, F., Khizbullin, D., & Schmidhuber, J. (2023). Mindstorms in natural language-based societies of mind. *arXiv*. <https://arxiv.org/abs/2305.17066>
21. Havrilla, A., Raparthy, S., Nalmpantis, C., Dwivedi-Yu, J., Zhuravinskyi, M., Hambro, E., & Raileanu, R. (2024). GLoRe: When, where, and how to improve LLM reasoning via global and local refinements. *arXiv*. <https://arxiv.org/abs/2402.10963>
22. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mane, D. (2016). Concrete problems in AI safety. *arXiv*. <https://arxiv.org/abs/1606.06565>
23. Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.
24. Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1). <https://doi.org/10.1162/99608f92.8cd550d1>
25. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3645-3650). Association for Computational Linguistics. <https://doi.org/10.18653/v1/P19-1355>