Federated Reinforcement Planning for Privacy-Preserving Collaborative Large Language Model Agents

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

Brendan Lyons
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
brendan.lyons869@uab.edu

Abstract

The proliferation of large language model agents operating across decentralized, multi-stakeholder environments introduces unprecedented challenges for system-level coordination, data governance, and operational security. Traditional centralized training and inference paradigms expose sensitive user data and organizational knowledge to single points of compromise, while purely local agent learning fails to capture the systemic benefits of shared experience. This paper proposes Federated Reinforcement Planning, a novel architectural framework that integrates federated learning principles with reinforcement learning-based planning mechanisms to enable privacy-preserving collaboration among large language model agents. The framework is built upon a federated topology where agents update local planning policies using encrypted gradient aggregation, ensuring that raw interaction trajectories and proprietary reward signals never leave their trust boundaries. A key innovation lies in the decoupling of high-level planning guidance from low-level action execution, which allows agents to share abstract strategic knowledge without exposing the underlying data distributions or task-specific reasoning chains. This paper provides a detailed examination of the structural trade-offs inherent in such a system, including the tension between communication efficiency and model convergence, the robustness of decentralized governance mechanisms against adversarial agents, and the fairness implications of heterogeneous local computational resources. The analysis further addresses infrastructure requirements for secure aggregation, the sustainability of bandwidth-intensive coordination protocols, and policy considerations for cross-jurisdictional deployment. Through comparative discussion with existing multi-agent reinforcement learning and federated model aggregation approaches, the framework positions itself as a scalable and legally compliant alternative for enterprise, healthcare, and public-sector applications where data sovereignty is non-negotiable. The paper concludes with forward-looking perspectives on open challenges, including incentive mechanism design, differential privacy integration, and the governance of emergent agent behaviors in federated ecosystems.

Keywords

federated learning, reinforcement learning, large language model agents, privacy preservation, multi-agent planning, system architecture, decentralized governance, secure aggregation.

1. Introduction

The rapid adoption of large language model agents as autonomous intermediaries for complex reasoning and task execution has created a pressing need for collaborative frameworks that respect organizational boundaries and data sovereignty. Contemporary approaches to multi-agent learning typically rely on centralized servers that aggregate trajectories, reward signals, and model updates from all participating agents, a design that introduces severe privacy vulnerabilities and creates attractive targets for adversarial attacks [1, 2]. As LLM agents are increasingly deployed in sensitive domains such as healthcare diagnostics, financial portfolio management, and legal document analysis, the inability to guarantee that raw user data or proprietary task specifications remain within local trust boundaries becomes a critical systemic risk. The research community has responded with a variety of federated learning paradigms that enable model training across decentralized datasets without direct data sharing, but these paradigms were originally designed for supervised learning tasks and do not directly address the unique challenges of reinforcement learning in multi-agent planning scenarios [3, 4].

The fundamental difficulty lies in the nature of planning itself. An LLM agent engaged in long-horizon reasoning must not only select actions based on immediate observations but also maintain internal models of future states, evaluate alternative trajectories, and adapt its behavior in response to environmental feedback. In a collaborative setting, agents may benefit from sharing high-level strategic insights, such as which types of subgoals are likely to lead to successful task completion, without revealing the detailed interaction histories that produced those insights [5]. This observation motivates the core design principle of Federated Reinforcement Planning, which separates the learning of abstract planning guidance from the execution of low-level actions and the collection of fine-grained reward signals. By structuring the federated aggregation protocol around compressed, anonymized representations of planning knowledge rather than raw trajectory data, the framework preserves the benefits of collective learning while dramatically reducing the exposure of sensitive information.

The present work is situated at the intersection of three rapidly evolving research domains: federated learning, deep reinforcement learning for planning, and multi-agent LLM system design. Each of these domains has produced substantial theoretical and empirical contributions, but their integration into a coherent architectural framework for privacy-preserving collaboration remains incomplete. This paper addresses that gap by providing a system-level analysis of Federated Reinforcement Planning, with particular emphasis on the structural trade-offs that arise when privacy constraints are imposed on the collaborative learning process. The discussion proceeds through an examination of architectural foundations, the decoupling of planning and execution, governance and robustness concerns, deployment and sustainability considerations, and future directions for policy and research.

2. Architectural Foundations and Design Rationale

The Federated Reinforcement Planning framework is predicated on a client-server topology in which each participating node operates an LLM agent with local planning and execution capabilities, while a central aggregation server coordinates the exchange of encrypted gradient updates derived from high-level planning policies. Unlike standard federated learning, which typically aggregates updates from a static set of clients performing identical supervised tasks, the proposed architecture must accommodate agents that operate in heterogeneous environments with non-stationary dynamics and divergent reward structures [6, 7]. This

heterogeneity introduces significant challenges for gradient alignment, as the locally learned planning policies may be optimized for fundamentally different task distributions, leading to conflicting update directions during global aggregation. The framework addresses this through a hierarchical clustering mechanism that groups agents based on the similarity of their observed state transition patterns, allowing for targeted aggregation within clusters while maintaining a separate layer for cross-cluster knowledge distillation.

A central tenet of the architecture is the separation of planning guidance from action execution traces. Each agent maintains two distinct policy models: a local action policy that maps current observations to concrete actions, and a high-level planning policy that generates abstract subgoal sequences or strategic preferences. Only the gradients of the planning policy are communicated to the aggregation server, and these gradients are processed through a secure aggregation protocol that masks individual contributions using secret sharing or homomorphic encryption techniques [8, 9]. This design choice is motivated by the observation that high-level planning knowledge is inherently less sensitive than the detailed interaction data from which it is derived, while still capturing the most transferable aspects of an agent's learned experience. For example, an LLM agent operating in a supply chain management system might learn a planning policy that prioritizes just-in-time inventory replenishment for high-demand items; the gradient information reflecting this strategic preference can be shared with other agents without revealing the specific inventory levels, supplier contracts, or demand forecasting models of the originating organization.

The reinforcement learning component introduces a feedback loop that is absent in typical federated supervised learning frameworks. Each agent updates its local planning policy based on rewards received from the environment, and the aggregated global planning policy provides a baseline that accelerates convergence, particularly for agents with limited local experience or sparse reward signals [10, 11]. This creates a classic exploration-exploitation trade-off at the system level: agents that trust the global policy too heavily may fail to learn locally optimal behaviors that diverge from the aggregate, while agents that ignore the global policy forfeit the benefits of collaborative learning. The framework addresses this by incorporating a dynamic weighting mechanism that adjusts the influence of the global policy based on each agent's local uncertainty estimate, measured through the entropy of its value function distribution. Agents operating in stable, well-characterized environments are encouraged to rely more on their local learning, while agents facing novel or volatile conditions benefit from stronger adherence to the shared planning guidance.

3. The Decoupling of Planning Guidance and Execution

The separation of planning and execution within each agent is not merely a convenience for privacy preservation but reflects a deeper conceptual distinction between two forms of knowledge representation in LLM-based reasoning systems. Execution-level knowledge is tightly coupled to the specific sensory inputs, action spaces, and reward functions of a particular deployment environment, and its exposure would reveal detailed operational characteristics that organizations have strong incentives to protect. Planning-level knowledge, by contrast, captures abstract patterns of reasoning that are largely invariant to the low-level implementation details and can often be expressed in terms of statistical regularities across state spaces [12]. The Federated Reinforcement Planning framework formalizes this distinction by training the planning policy through a separate reinforcement learning loop that receives inputs derived from the agent's internal world model rather than from raw environmental observations.

This design bears a conceptual resemblance to hierarchical reinforcement learning, where a meta-controller learns to set subgoals for a lower-level controller, but the key difference lies in the privacy-preserving aggregation mechanism. In federated reinforcement planning, the high-level policy is updated not only through local experience but also through aggregated gradient information from other agents, while the low-level policy remains entirely local and never participates in cross-agent communication [13]. The required reference [13] demonstrates that high-level planning guidance can significantly improve the efficiency of reinforcement learning for LLM reasoning tasks by providing structured exploration strategies and reducing the variance of policy gradient estimates. Their work supports the plausibility of the decoupling approach, showing that planning policies can be learned and transferred across different reasoning contexts without requiring access to the underlying training data or task-specific reward signals.

The practical implementation of this decoupling requires careful attention to the granularity of the planning abstraction. If the planning policy operates at too high a level, its guidance may be too generic to provide meaningful acceleration for diverse local tasks; if it operates at too low a level, the risk of privacy leakage increases as the aggregated gradients begin to encode information about specific state transitions. The framework introduces a tunable abstraction parameter that controls the temporal resolution of the planning policy, allowing system administrators to calibrate the trade-off between utility and privacy based on the sensitivity of the domain. In healthcare applications, for instance, the planning policy might operate at the level of diagnostic subgoals rather than individual test results, while in autonomous vehicle coordination, it might encode route selection preferences without revealing specific sensor readings or traffic patterns.

4. Structural Trade-Offs and Governance Mechanisms

The adoption of any federated architecture involves inherent trade-offs that must be systematically analyzed to ensure responsible deployment. The most prominent of these is the tension between communication efficiency and model convergence speed. Federated Reinforcement Planning requires each agent to transmit gradient updates for the planning policy at regular intervals, and the bandwidth consumption scales with the number of agents, the dimensionality of the planning policy, and the frequency of aggregation rounds [14]. In environments with limited network connectivity, such as remote industrial installations or disaster response scenarios, the communication overhead may become prohibitive. The framework addresses this through gradient compression techniques that reduce transmission costs by up to an order of magnitude, but compression introduces noise that can slow convergence and, in worst cases, cause the aggregated policy to diverge from the optimal solution.

Another critical trade-off involves the robustness of the system against adversarial agents. In a federated topology, any participant that controls a local agent can potentially manipulate its gradient updates to poison the global planning policy, causing all agents to learn suboptimal or harmful behaviors [15]. The open-ended nature of LLM-based planning amplifies this risk, as adversaries can craft gradient perturbations that are difficult to detect using standard statistical anomaly detection methods. The framework incorporates a Byzantine-robust aggregation rule that discards gradient updates falling outside a statistically plausible range, but this defense comes at the cost of reduced model accuracy when legitimate agents produce truly novel or outlier updates that reflect genuine environmental shifts. A governance mechanism is therefore required to adjudicate between outliers that represent useful

exploration and those that represent malicious attacks, a challenge that points toward the need for decentralized reputation systems and incentive-compatible contribution models.

Fairness concerns arise from the heterogeneity of local computational resources among participating agents. Agents deployed on powerful hardware with access to large local datasets can train their planning policies more accurately and contribute higher-quality gradient updates, while resource-constrained agents may produce noisy or biased updates that degrade the quality of the global model [16]. This asymmetry can create a feedback loop in which well-resourced agents benefit disproportionately from the aggregated knowledge, while poorly resourced agents receive marginal improvements or even suffer degradation due to mismatched policy guidance. The framework introduces a contribution weighting scheme that amplifies the influence of updates from agents with high local uncertainty, thereby giving greater voice to under-resourced participants, but this approach must be carefully balanced against the risk of incorporating low-quality information into the aggregate.

5. Deployment Infrastructure and Sustainability Considerations

The operationalization of Federated Reinforcement Planning requires a robust infrastructure that supports secure, asynchronous communication across potentially thousands of distributed agents. The aggregation server must be designed for fault tolerance, as agents may drop out mid-round due to network failures or deliberate disconnection, and the secure aggregation protocol must be capable of handling dynamic membership changes without exposing individual contributions [17]. From a deployment perspective, the most practical approach involves a two-tier architecture: a central coordination layer responsible for managing the global aggregation schedule and maintaining the encrypted gradient buffers, and a set of regional aggregation nodes that perform local aggregation for geographical or institutional clusters. This topology reduces latency, provides fault isolation, and aligns with regulatory requirements for data localization that are common in jurisdictions with strict data sovereignty laws.

Sustainability considerations extend beyond computational efficiency to encompass the energy consumption of repeated training rounds and the environmental impact of large-scale LLM inference. Each training round in the federated cycle requires each agent to perform multiple forward and backward passes through its planning policy, and the cumulative energy cost across a large deployment can be substantial [18]. The framework mitigates this through selective participation strategies that only activate agents for aggregation when their local policy update is likely to contribute meaningful information, as measured by the magnitude of the policy gradient relative to a threshold. Additionally, the use of compressed gradient representations reduces the amount of computation required for aggregation, and the hierarchical clustering approach means that many agents can skip global aggregation rounds altogether if their cluster-level policy is sufficiently aligned with the global average.

Policy implications for cross-jurisdictional deployment are particularly complex given the current regulatory landscape. The European Union's General Data Protection Regulation, the California Consumer Privacy Act, and emerging frameworks in Asia and Latin America impose distinct requirements for data minimization, purpose limitation, and the right to explanation [19]. Federated Reinforcement Planning, by virtue of keeping raw data and execution traces within local trust boundaries, is generally well-positioned to comply with these regulations, but the aggregation of planning policy gradients may still trigger data protection obligations if the gradients can be used to reconstruct sensitive attributes of the originating agent. Techniques such as differential privacy, which add calibrated noise to

gradient updates before transmission, can provide formal privacy guarantees, but the noise inevitably degrades the utility of the aggregated policy [20]. The choice of privacy budget becomes a governance decision that must be informed by the specific legal requirements and risk tolerance of the deploying organization.

6. Future Directions and Open Challenges

Despite the architectural maturity of the Federated Reinforcement Planning framework, several open challenges remain that will require sustained interdisciplinary investigation. The first concerns the design of incentive mechanisms that encourage agents to contribute high-quality updates without free-riding on the contributions of others. In purely voluntary federations, rational agents may prefer to receive the benefits of the global planning policy without incurring the computational and communication costs of participation, a classic collective action problem [21]. The development of verifiable contribution metrics and reward structures based on mutual information or policy improvement potential is an active area of research that will directly impact the viability of large-scale deployments.

A second challenge involves the integration of more sophisticated privacy-preserving techniques, such as secure multi-party computation and functional encryption, which could enable agents to compute the aggregated planning policy without revealing even the gradient magnitudes to the aggregation server. These techniques offer stronger theoretical guarantees but currently impose prohibitive computational overheads for LLM-scale models, and their practical deployment will require advances in hardware acceleration and algorithmic efficiency [22]. The trade-off between privacy strength and system performance is likely to remain a frontier of research for the foreseeable future, with the optimal point shifting as both cryptographic and machine learning technologies evolve.

Finally, the emergence of autonomous agent behaviors that cannot be fully anticipated during design time raises questions about governance and accountability in federated ecosystems. When agents learn planning policies through collective reinforcement, there is a possibility that the emergent global behavior violates ethical norms, regulatory constraints, or safety specifications that no single agent would individually adopt [23]. The framework must therefore be augmented with runtime monitoring mechanisms that detect policy drift and trigger human-in-the-loop interventions when necessary. The design of such mechanisms, and the allocation of responsibility for the actions of decentralized, collaboratively learned agents, represents a profound challenge for both the technical community and the broader society that will deploy these systems.

7. Conclusion

Federated Reinforcement Planning offers a coherent architectural response to the privacy, governance, and scalability challenges that arise when large language model agents must collaborate across organizational boundaries. By decoupling high-level planning guidance from low-level action execution and structuring the collaborative learning process around encrypted gradient aggregation, the framework enables agents to benefit from shared strategic knowledge without exposing sensitive interaction data or proprietary reward structures. The analysis presented in this paper has elaborated the key structural trade-offs inherent in such a system, including the balance between communication efficiency and convergence, the robustness of aggregation against adversarial manipulation, and the fairness implications of resource heterogeneity. These considerations are not merely technical but are deeply intertwined with legal, ethical, and policy frameworks that will constrain real-world

deployment. As the frontier of LLM agent research continues to advance, the integration of privacy-preserving mechanisms into collaborative learning architectures will become increasingly critical for ensuring that the benefits of collective intelligence are realized without compromising the autonomy, security, and dignity of the participating entities.

References

1. McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 54, 1273–1282.

2. Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (2nd ed.). MIT Press.

3. Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50–60.

4. Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., & Bacon, D. (2016). Federated learning: Strategies for improving communication efficiency. arXiv preprint arXiv:1610.05492.

5. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

6. Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., & Gao, Y. (2021). A survey on federated learning. Knowledge-Based Systems, 216, 106775.

7. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., & Seth, K. (2019). Practical secure aggregation for privacy-preserving machine learning. Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 1175–1191.

8. Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3–4), 211–407.

9. Garg, S., Goldwasser, S., & Vasudevan, P. N. (2020). Formalizing data de-anonymization and its implications for privacy-preserving machine learning. Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, 1101–1118.

10. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.

11. Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587), 484–489.

12. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J. R. (2024). A survey on large language model based autonomous agents. Frontiers of Computer Science, 18(6), 186345.

13. Dou, Z., Zhao, Q., Wan, Z., Zhang, D., Wang, W., Raiyan, T., ... & Biswas, S. (2025). Plan Then Action: High-Level Planning Guidance Reinforcement Learning for LLM Reasoning. arXiv preprint arXiv:2510.01833.

14. Arulkumaran, K., Deisenroth, M. P., Brundage, M., & Bharath, A. A. (2017). A brief survey of deep reinforcement learning. IEEE Signal Processing Magazine, 34(6), 26–38.

15. Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., & Shmatikov, V. (2020). How to backdoor federated learning. Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics, 108, 2938–2948.

16. Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., & Chandra, V. (2018). Federated learning with non-iid data. arXiv preprint arXiv:1806.00582.

17. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., El Rouayheb, S., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., ... & Zhao, S. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1–2), 1–210.

18. Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L. M., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). Carbon emissions and large neural network training. arXiv preprint arXiv:2104.10350.

19. European Parliament. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation). Official Journal of the European Union, L119, 1–88.

20. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 308–318.

21. Cong, Z., Luo, Y., & Zhang, Q. (2022). Incentive mechanism design for federated learning: A survey. IEEE Communications Surveys and Tutorials, 24(4), 2620–2655.

22. Evans, D., Kolesnikov, V., & Rosulek, M. (2018). A pragmatic introduction to secure multi-party computation. Foundations and Trends in Privacy and Security, 2(2–3), 70–246.

23. Russell, S., Dewey, D., & Tegmark, M. (2015). Research priorities for robust and beneficial artificial intelligence. AI Magazine, 36(4), 105–114.