Transactions on Computational and Scientific Methods | Vol. 6, No. 02, 2026
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Research and Applications of LLM-Based Assisted Planning Capabilities for Wireless Networks

Bowen Dagan
University of Lincoln, Lincoln, United Kingdom
BD923@lincoln.ac.uk

Abstract: With the rapid advancement of large language models (LLMs) in natural language processing, these models have demonstrated substantial potential in data processing, pattern recognition, and predictive analytics. This capability offers new perspectives for traditional wireless network base-station planning and design. Building upon current LLM technologies, this work introduces Agents, prompt-engineering methodologies, and chain-of-thought reasoning to enable interactive analysis of fundamental coverage scenarios. By integrating Retrieval-Augmented Generation (RAG), we construct a domain knowledge base tailored to network planning, supporting standardized verification, specification matching, and knowledge retrieval throughout the planning workflow. Furthermore, to ensure the security of sensitive information, we propose a private-domain deployment architecture in which all data are processed exclusively within internal networks and servers, thereby providing robust protection for end-to-end data security.

Keywords: Large language models; Agent; Prompt engineering; Chain-of-thought

1. Introduction

With the continuous expansion of 4G/5G deployments, site planning has become increasingly challenging. Determining the most suitable co-location site among numerous existing base stations, integrating network data to assess local radio coverage conditions, and generating reasonable supplementation recommendations all present significant difficulties. Moreover, the planning process typically relies on multiple professional tools whose operations are complex and time-consuming, posing particular challenges for inexperienced or newly recruited engineers. Inadequate familiarity with industry standards, specifications, and design procedures may lead to serious design errors during the planning workflow. In recent years, the widespread adoption of large language models in natural language processing has demonstrated substantial potential in data processing, pattern recognition, and predictive analysis, offering new perspectives for addressing these long-standing challenges in traditional wireless network planning.

Based on a systematic review of current mainstream large language models, this study investigates potential applications of such models within wireless network planning workflows. We propose implementation schemes for several representative scenarios, including base-station status analysis and knowledge-base-enhanced retrieval for planning-related specifications, and provide practical case studies to illustrate their feasibility. Furthermore, the paper outlines the future prospects of integrating large language models and AI technologies into traditional planning and design domains. With continuous technological innovation and interdisciplinary collaboration, AI is expected to play an increasingly pivotal role, supporting enterprises in achieving digital transformation and sustainable development within the planning and design industry.

2. Methodology Foundation

The methodological foundation of the proposed framework originates from reinforcement learning and sequential decision-making theory. The formal Markov decision process and value-based optimization principles introduced in [1] establish the theoretical basis for modeling iterative reasoning and action execution. The integration of deep neural networks with structured search mechanisms in [2] further demonstrates how complex planning tasks can be decomposed into stepwise decision procedures guided by learned value estimation. These principles directly inspire the chain-of-thought guided task decomposition and iterative SQL refinement strategy adopted in the proposed Agent framework.

The scalability and emergent reasoning capability of large language models provide the computational backbone of the system. Empirical scaling laws for neural language models in [3] demonstrate predictable performance gains with increasing model size, while few-shot generalization capabilities shown in [4] validate the feasibility of prompt-based task adaptation without full retraining. Retrieval-augmented generation (RAG) was introduced in [5] as a hybrid parametric-nonparametric framework that integrates external knowledge into generative models, significantly enhancing factual reliability. Subsequent improvements in retrieval-generation integration [6] and large-scale retrieval pretraining [7] further confirm that external knowledge grounding effectively mitigates hallucination. Large-scale pathway-based model scaling in [8] reinforces the architectural feasibility of high-capacity reasoning cores for domain-specific deployment.

Agent-based reasoning and memory-driven planning mechanisms form another essential methodological pillar. Interactive generative agent architectures in [9] demonstrate how large language models can maintain environmental state and long-term coherence through memory modules. Reinforcement learning-based adaptive profiling in [10] extends sequential adaptation into dynamic optimization contexts, informing the iterative refinement process within the data Agent. Modular task decomposition and collaborative orchestration strategies in [11] provide structural guidance for breaking down complex planning workflows into manageable sub-tasks. Hierarchical memory encoding and dynamic retrieval planning mechanisms in [12] further support multi-step reasoning and metadata integration. Complementary multi-agent workflow construction paradigms in [13] reinforce the feasibility of tool-augmented task execution pipelines.

Given the sensitivity of operational data, trust-aware and privacy-preserving mechanisms are incorporated into the methodological design. Trust-aware orchestration frameworks in [14] and contextual trust evaluation mechanisms in [15] inform reliability-aware coordination among reasoning components. Governance-centric secure agent architectures in [16] provide structural principles for private-domain deployment. Parameter-efficient fine-tuning with differential privacy guarantees in [17] and federated adaptation mechanisms in [18] supply theoretical foundations for secure adaptation without exposing raw data. Uncertainty-aware generation strategies in [19] further enhance robustness by quantifying prediction risk.

To enhance controllability and structural faithfulness, advanced retrieval ranking and decoding strategies are incorporated. Faithfulness-aware multi-objective context ranking in [20] informs the clause-level retrieval precision of the planning knowledge base. Structure-aware decoding mechanisms in [21] and output-constrained generation strategies in [22] provide methodological guidance for generating executable SQL statements and standardized outputs. Dynamic prompt fusion strategies in [23] further support adaptive instruction composition across heterogeneous planning tasks. Attention attribution and interpretability-oriented modeling in [24] contribute to transparent discriminative reasoning. Knowledge-augmented agent architectures in [25] reinforce structured knowledge integration during reasoning.

Graph-based and temporal modeling techniques further enrich the structural reasoning capability of the system. Hierarchical graph representation learning with differentiable pooling in [26] establishes foundational principles for structured relational modeling. Causal reasoning over knowledge graphs in [27] extends this

paradigm toward intervention-oriented inference. Multi-scale graph-enhanced language model integration in [28] and graph neural classification mechanisms in [29] provide structural embedding strategies for relational metadata. Temporal-dynamic graph modeling in [30], deep temporal convolution with attention mechanisms in [31], and transformer-based sequence modeling in [32] support time-dependent status evaluation and coverage analysis. Multi-scale transformer anomaly detection in [33] and contrastive dependency modeling in [34] improve robustness under noisy and heterogeneous conditions. Multi-scale temporal alignment strategies in [35] and generative distribution modeling under imbalanced settings in [36] further strengthen statistical reliability in complex analytical scenarios. Finally, shared-encoder self-supervised transfer learning in [37] provides theoretical grounding for cross-domain adaptability within secure environments.

Collectively, these studies establish a coherent methodological lineage that integrates reinforcement learning-based decision modeling, scalable language model reasoning, retrieval-augmented knowledge grounding, agentic task decomposition, trust-aware orchestration, privacy-preserving adaptation, structured decoding control, and graph-temporal representation learning. Building upon these foundations, the proposed framework innovatively combines Agent-based chain-of-thought reasoning with private-domain RAG deployment and deterministic SQL validation workflows, transforming stochastic generative modeling into a controllable, secure, and engineering-oriented intelligent planning system.

3. Comparative Analysis of Common Large Language Model Capabilities

A survey was conducted to evaluate the capabilities and service characteristics of various open-source large language models, and the results are summarized in Table 1. The comparative findings indicate that, in terms of performance, complex task understanding, and answer accuracy, most open-source models currently remain inferior to proprietary models offered by major commercial providers. However, open-source models possess a substantial advantage - they allow users to replicate the model and deploy it independently on local or private-domain servers. This makes open-source large language models a preferred option for users operating in sensitive-data environments.

In non-sensitive application scenarios, the complexity associated with deploying large language models and the hardware limitations of general-purpose computing devices lead many small and medium-sized enterprises, as well as individual developers, to rely on third-party inference services. Such services enable engineering projects to run efficiently on standard computers or servers while achieving relatively fast inference speed and response performance.

Table 1: Comparison of Common Large Language Model Capabilities

Model Name	Open-Source	API Service	Summary of Performance
ChatGLM3-6B	Yes	Yes	Fast; long context; local deploy; moderate complex tasks.
ChatGLM2-6B	Yes	Yes	Fast; long context; weaker complex QA.
Vicuna-13B	Yes	No	Good structured tasks; slow; small context.
Baichuan-7B	Yes	No	Fast; long context; weak reasoning.
Tongyi Qianwen	No	Yes	Fast; stable API; long context.

Wenxin Yiyan	No	Yes	Fast; API; web search support.
ChatGPT	No	Yes	Strong reasoning; high quality; paid.

4. Wireless Network Assisted Planning Implementation

4.1 Base-Station Status Analysis: Implementation Framework

To enable large language models to support auxiliary base-station planning and perform real-world base-station status analysis, the model must be able to perceive and interpret relevant base-station information. The most direct approach would be to provide the raw base-station data to the model for training and analysis. However, this is impractical. On one hand, base-station data contain sensitive information, and uploading such data to a third-party model for inference introduces significant security risks. On the other hand, as previously noted, current 4G/5G network deployments generate massive volumes of detailed base-station data, far exceeding the capacity of existing large language models to process such extensive context and data.

To address this challenge, this paper introduces the concepts of Agents and chain-of-thought planning. An Agent is an intelligent entity capable of autonomously performing decision-making, learning, and task execution within a defined environment. Although the concept predates modern large language models, it has gained substantial momentum with the rapid development of large models and their application frameworks. Agents are now widely regarded as one of the most promising approaches for advancing toward artificial general intelligence (AGI). Building on this, the proposed method defines a base-station data Agent, where the large language model serves as the reasoning core. By leveraging chain-of-thought principles, the Agent decomposes user problems into sub-tasks and resolves them iteratively. Meanwhile, the data Agent transforms the inherently stochastic nature of model responses into deterministic workflow steps, enabling reliable and repeatable outcomes in engineering applications.

The operational workflow of the data Agent is illustrated in Figure 1. During actual deployment, the data Agent runs on a locally deployed platform that directly connects to the internal base-station database. When a user submits a query, the data Agent retrieves only the required metadata instead of the full sensitive dataset. Subsequently, through prompt engineering, metadata are embedded into the prompt template along with chain-of-thought guidance. After assembling the complete prompt, it is combined with the user's question and submitted to the large language model, which then returns a step-by-step reasoning process and an executable SQL command.

Next, the data Agent's SQL execution module verifies the syntax and validity of the SQL query. Once validated, the module executes the SQL request on the database. The queried data are then visualized, enabling users to interpret the current base-station status. For users with sufficient database knowledge, the system also allows manual inspection of SQL correctness to ensure expected outcomes.

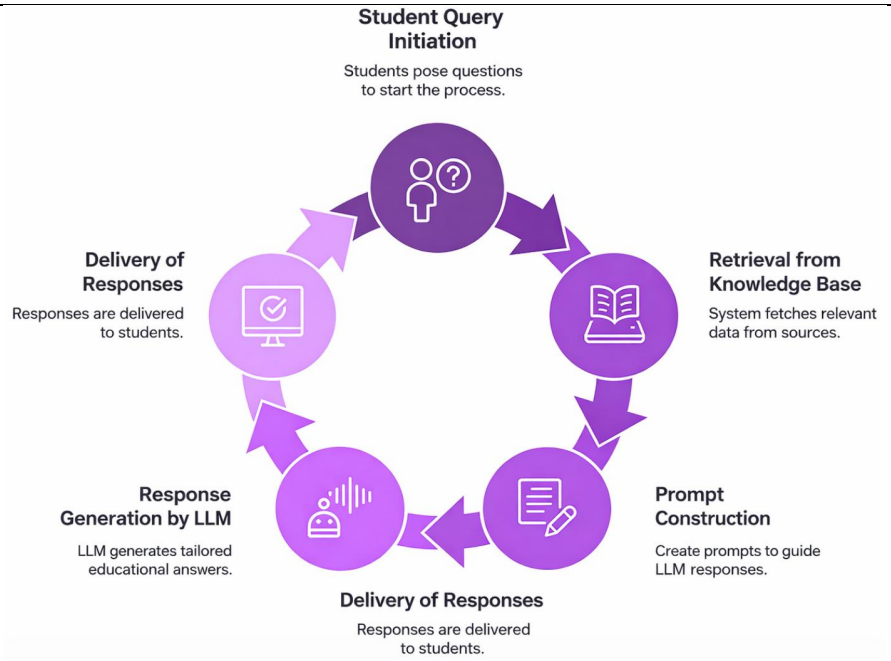


Figure 1. Execution Diagram of the Data Agent

From the above workflow, it is evident that prompt engineering plays a central role in the entire process. In large language model applications, prompt design is indispensable. Here, the concept of prompts extends beyond simple LLM instructions to a more generalized representation design method, including product-level interaction prompts. In the proposed data Agent workflow, prompt engineering involves embedding both metadata and chain-of-thought elements, while also incorporating user-provided inputs. Furthermore, in practical applications, prompt templates may include additional product interaction instructions to guide the model toward generating visualizable results or structured tables, thus improving user experience. One feasible prompt template explored in this study is shown in Figure 2.

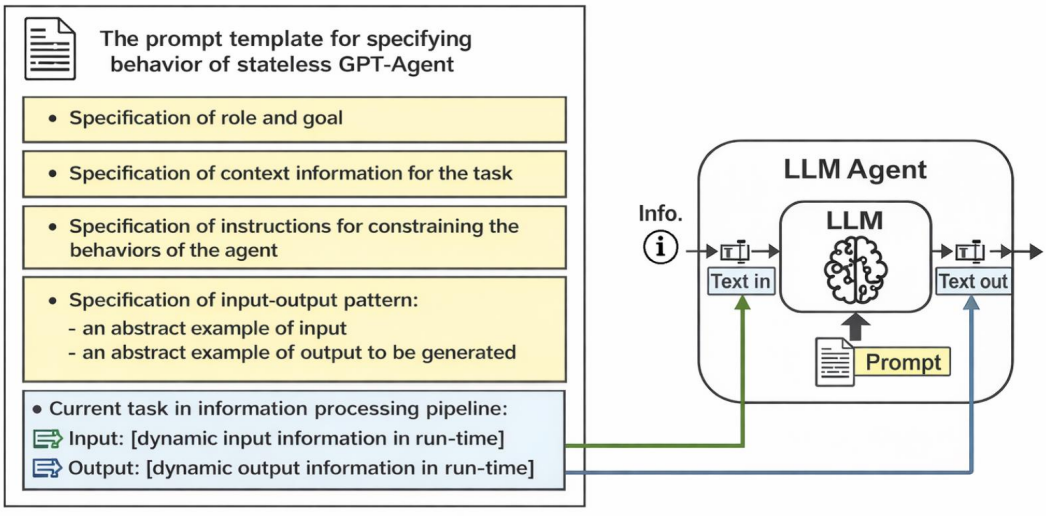


Figure 2. Example of Prompt Engineering for the Data Agent

4.2 Implementation Framework for Domain Knowledge-Base Retrieval and Question Answering

In the planning and design domain, a vast number of telecommunications standards, specifications, documents, and clauses must be referenced. New engineers often struggle to memorize and master these

materials within a short period of time, making it difficult for them to quickly adapt to frontline production tasks. There is a strong need for an efficient method that enables rapid lookup and learning of relevant industry norms. With the emergence of large language models and the rapid advancement of their natural language question-answering capabilities, new opportunities have arisen for addressing knowledge retrieval requirements in this domain.

However, while general-purpose large language models have significantly transformed the way people work and produce, they face multiple challenges in specialized fields, particularly regarding domain knowledge, factual accuracy, and controllability. Although large language models possess strong reasoning capabilities, they lack direct access to private or proprietary information. To address this limitation, Retrieval-Augmented Generation (RAG) technology has gained prominence. By retrieving relevant domain knowledge and private-scope information, RAG enhances a model's decision-making and creative capabilities, offering new solutions and approaches for specialized planning and design knowledge retrieval tasks.

RAG operates by indexing and encoding text into vector representations, constructing a private knowledge base suitable for query and retrieval. During question-answering, the user's input is vectorized and matched against the private knowledge base using similarity search. The retrieved knowledge items are then combined with the user's original query and passed to the large language model. After the model integrates and processes the retrieved knowledge, it generates an appropriate answer to the user's query, as illustrated in Figure 3.

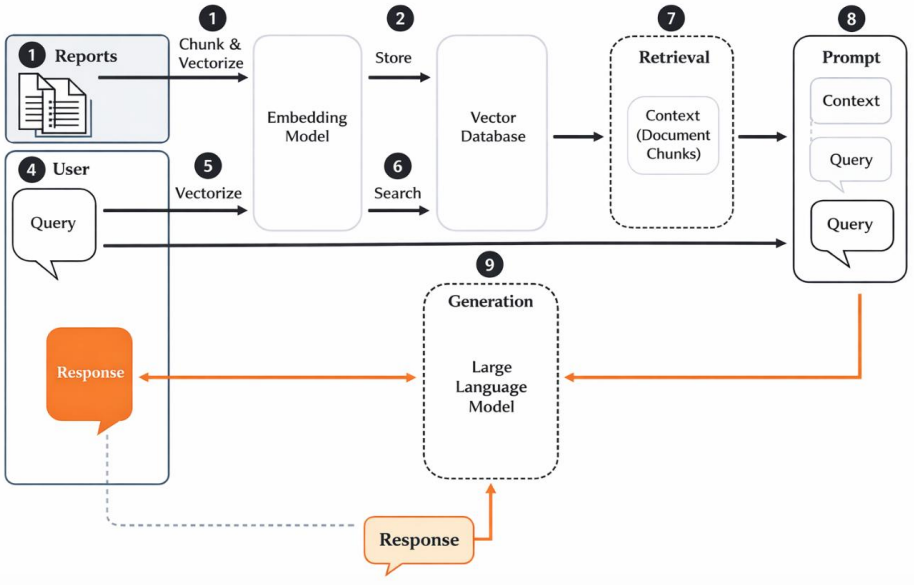


Figure 3. RAG-Based Private Domain Knowledge Question-Answering Framework

5. Solution Validation

To verify the practical effectiveness of the proposed solution in the wireless network auxiliary planning process, a wireless network auxiliary planning platform was constructed and used to evaluate the validity of the approach. Four dimensions-coverage determination, base station data analysis and statistics, automated chart generation, and base-station-level checklist creation-were selected to examine the platform's analytical capability regarding the current status of base stations. A corresponding planning-rule knowledge base and standardized design-specification knowledge base were built to support query validation. Five categories of

queries were sampled, and ten questions were selected for each category. The results were then compared with manual expert statistics, as summarized in Table 2.

Table 2: Results Across Five Scenario Categories

Category	Verification Criterion	Result
Standard Specification Query	Must be fully consistent	100% Consistent
Coverage Determination	Accuracy ≥ 90%	90% Accuracy
Data Analysis & Statistics	Error ≤ 5%	80% of samples within 5% error
Chart Generation	Error ≤ 5%	100% within 5% error
Base-Station Checklist Creation	Error ≤ 5%	100% within 5% error

The validation results indicate that the platform achieves a 100% accuracy rate in standardized planning-rule queries. This demonstrates, on the one hand, the strong performance of the retrieval-augmented generation (RAG) technique, and on the other hand, the close alignment between the knowledge bases used in the tests and their intended application scenarios. The knowledge bases employed mainly contained industry standards and regulatory specifications relevant to planning design. Because the underlying data volume was relatively small compared with other complex industry cases-and due to the structural characteristics of the standards, which are organized into discrete clauses with minimal overlap-the system was able to perform precise clause-level retrieval, substantially improving the overall accuracy of responses.

In terms of base-station status analysis, the results for coverage judgment, chart generation, and checklist creation meet the verification standards. Statistical analysis of base-station data shows that 80% of samples meet the verification criteria, while the remaining 20% fall short but exhibit only minor deviations. Since the proposed method retrieves only database schema information and user descriptions-rather than direct access to detailed base-station data-the generated SQL is returned to the platform for execution. The model's ability to process complex analytical logic is therefore limited, leading to some deviation in outcomes. In less data-sensitive domains, this limitation could be mitigated by adopting multi-round dialogue, whereby SQL generated by the model is executed iteratively and the results are fed back for further analysis, thus improving the integrity of the analytical output.

Overall, the wireless network auxiliary planning platform successfully fulfills the intended predictive and analytical functions. The validation results demonstrate a high level of reference value and indicate that the platform can significantly improve the efficiency of wireless network planners.

6. Conclusion

Building on the investigation of current large language model capabilities, this study introduces Agents, prompt engineering, and chain-of-thought reasoning to enable interactive querying of fundamental base-station coverage conditions. Furthermore, RAG technology is incorporated to construct a domain-specific knowledge base that supports verification of standards, specifications, and regulatory knowledge throughout the planning and design workflow. To ensure the security of sensitive data, the study proposes a private-domain deployment framework that keeps all data circulation confined within internal networks and servers, thereby providing strong protection for data confidentiality.

The validation results demonstrate that the combination of mainstream large language models and RAG techniques enables highly accurate retrieval and utilization of industry knowledge bases. In base-station status analysis tasks-including coverage evaluation, automated chart generation, and checklist creation-the platform exhibits strong performance. However, due to inherent data-sensitivity constraints within the

proposed solution, further improvements are needed in more advanced analytical and statistical tasks. For non-sensitive data scenarios, future enhancements may adopt multi-round interactions, allowing the model to iteratively execute and refine SQL-based analyses to achieve more comprehensive statistical outputs.

In summary, the proposed wireless-network auxiliary planning solution effectively meets the intended functional requirements. The results exhibit strong reference value and demonstrate that the platform can significantly enhance the efficiency of wireless network planning personnel.

References

[1] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, vol. 1, no. 1, pp. 9-11. Cambridge, UK: MIT Press, 1998.

[2] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche et al., "Mastering the game of Go with deep neural networks and tree search," Nature, vol. 529, no. 7587, pp. 484-489, 2016.

[3] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child et al., "Scaling laws for neural language models," arXiv:2001.08361, 2020.

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal et al., "Language models are few-shot learners," Advances in Neural Information Processing Systems, vol. 33, pp. 1877-1901, 2020.

[5] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," Advances in Neural Information Processing Systems, vol. 33, pp. 9459-9474, 2020.

[6] G. Izacard and E. Grave, "Leveraging passage retrieval with generative models for open domain question answering," Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp. 874-880, Apr. 2021.

[7] S. Borgeaud, A. Mensch, J. Hoffmann, T. Cai, E. Rutherford, K. Millican et al., "Improving language models by retrieving from trillions of tokens," Proceedings of the 39th International Conference on Machine Learning (ICML), pp. 2206-2240, Jun. 2022.

[8] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts et al., "PaLM: Scaling language modeling with pathways," Journal of Machine Learning Research, vol. 24, no. 240, pp. 1-113, 2023.

[9] J. S. Park, J. O' Brien, C. J. Cai, M. R. Morris, P. Liang and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST), pp. 1-22, Oct. 2023.

[10] Y. Zhou, "A unified reinforcement learning framework for dynamic user profiling and predictive recommendation," Proceedings of the 2025 3rd International Conference on Artificial Intelligence and Automation Control (AIAC), pp. 432-436, Oct. 2025.

[11] S. Pan and D. Wu, "Modular task decomposition and dynamic collaboration in multi-agent systems driven by large language models," arXiv:2511.01149, 2025.

[12] Y. Wang, R. Yan, Y. Xiao, J. Li, Z. Zhang and F. Wang, "Memory-driven agent planning for long-horizon tasks via hierarchical encoding and dynamic retrieval," 2025.

[13] T. Guan, "A multi-agent coding assistant for cloud-native development: From requirements to deployable microservices," 2025.

[14] Y. Hu, J. Li, K. Gao, Z. Zhang, H. Zhu and X. Yan, "TrustOrch: A dynamic trust-aware orchestration framework for adversarially robust multi-agent collaboration," 2025.

[15] K. Gao, H. Zhu, R. Liu, J. Li, X. Yan and Y. Hu, "Contextual trust evaluation for robust coordination in large language model multi-agent systems," 2025.

[16]J. Chen, J. Yang, Z. Zeng, Z. Huang, J. Li and Y. Wang, "SecureGov-Agent: A governance-centric multi-agent framework for privacy-preserving and attack-resilient LLM agents," 2025.

[17]Y. Huang, Y. Luan, J. Guo, X. Song and Y. Liu, "Parameter-efficient fine-tuning with differential privacy for robust instruction adaptation in large language models," arXiv:2512.06711, 2025.

[18]S. Wang, S. Han, Z. Cheng, M. Wang and Y. Li, "Federated fine-tuning of large language models with privacy preservation and cross-domain semantic alignment," Proceedings of the 2025 6th International Conference on Computer Vision and Data Mining (ICCVDM), pp. 494-498, Sep. 2025.

[19]S. Pan and D. Wu, "Trustworthy summarization via uncertainty quantification and risk awareness in large language models," Proceedings of the 2025 6th International Conference on Computer Vision and Data Mining (ICCVDM), pp. 523-527, Sep. 2025.

[20]N. Tan, Z. Seng, L. Zhang, Y. C. Shih, D. Yang and A. Salunkhe, "Improved LLM agents for financial document question answering," arXiv:2506.08726, 2025.

[21]M. Jansen and M. Pehlke, "Increasing AI explainability by LLM driven standard processes," arXiv:2511.07083, 2025.

[22]J. Yang, S. Sun, Y. Wang, Y. Wang, X. Yang and C. Zhang, "Semantic alignment and output constrained generation for reliable LLM-based classification," 2026.

[23]X. Hu, Y. Kang, G. Yao, T. Kang, M. Wang and H. Liu, "Dynamic prompt fusion for multi-task and cross-domain adaptation in LLMs," Proceedings of the 2025 10th International Conference on Computer and Information Processing Technology (ISCIPT), pp. 483-487, Sep. 2025.

[24]X. Song, "Integrating attention attribution and pretrained language models for transparent discriminative learning," 2026.

[25]Q. Zhang, Y. Wang, C. Hua, Y. Huang and N. Lyu, "Knowledge-augmented large language model agents for explainable financial decision-making," arXiv:2512.09440, 2025.

[26]Z. Ying, J. You, C. Morris, X. Ren, W. Hamilton and J. Leskovec, "Hierarchical graph representation learning with differentiable pooling," Advances in Neural Information Processing Systems, vol. 31, 2018.

[27]R. Ying, Q. Liu, Y. Wang and Y. Xiao, "AI-based causal reasoning over knowledge graphs for data-driven and intervention-oriented enterprise performance analysis," 2025.

[28]X. Song, Y. Huang, J. Guo, Y. Liu and Y. Luan, "Multi-scale feature fusion and graph neural network integration for text classification with large language models," arXiv:2511.05752, 2025.

[29]R. Liu, R. Zhang and S. Wang, "Graph neural networks for user satisfaction classification in human-computer interaction," arXiv:2511.04166, 2025.

[30]Q. Zhang, N. Lyu, L. Liu, Y. Wang, Z. Cheng and C. Hua, "Graph neural AI with temporal dynamics for comprehensive anomaly detection in microservices," arXiv:2511.03285, 2025.

[31]N. Lyu, F. Chen, C. Zhang, C. Shao and J. Jiang, "Deep temporal convolutional neural networks with attention mechanisms for resource contention classification in cloud computing," 2025.

[32]R. Liu, R. Zhang and S. Wang, "Transformer-based modeling of user interaction sequences for dwell time prediction in human-computer interfaces," arXiv:2512.17149, 2025.

[33]Y. Kang, "Machine learning method for multi-scale anomaly detection in cloud environments based on transformer architecture," Journal of Computer Technology and Software, vol. 3, no. 4, 2024.

[34]Y. Xing, Y. Deng, H. Liu, M. Wang, Y. Zi and X. Sun, "Contrastive learning-based dependency modeling for anomaly detection in cloud services," arXiv:2510.13368, 2025.

[35] W. C. Chang, L. Dai and T. Xu, "Machine learning approaches to clinical risk prediction: Multi-scale temporal alignment in electronic health records," arXiv:2511.21561, 2025.

[36] Z. Xu, K. Cao, Y. Zheng, M. Chang, X. Liang and J. Xia, "Generative distribution modeling for credit card risk identification under noisy and imbalanced transactions," 2025.

[37] Y. Zhou, "Self-supervised transfer learning with shared encoders for cross-domain cloud optimization," Proceedings of the 2025 5th International Conference on Electronic Information Engineering and Computer Science (EIECS), pp. 1435-1439, Sep. 2025.