Journal of computer science and software applications
https://www.mfacademia.org/index.php/jcssa
ISSN:2377-0430
Vol. 4, No. 8, 2024

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

Eliot Vendrell
Lakehead University, Thunder Bay, Canada
evendrell2@lakeheadu.ca

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[1]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[2].

Based on a systematic review of current mainstream large language models, this study investigates potential applications of such models within wireless network planning workflows[3]. We propose implementation schemes for several representative scenarios, including base-station status analysis and knowledge-base-enhanced retrieval for planning-related specifications[4], 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. Related Work

Large language models (LLMs) have rapidly evolved to become foundational tools across numerous domains, demonstrating strong capabilities in reasoning, planning, and natural language understanding. The seminal work by Brown et al. introduced the concept of few-shot learning with GPT-3, establishing LLMs as versatile systems that can generalize from minimal examples and perform a wide array of NLP tasks with little to no task-specific tuning [5]. Following this, substantial efforts have focused on optimizing LLM training efficiency and resource utilization. Hoffmann et al. explored compute-optimal training strategies and provided empirical guidance for scaling laws that influence LLM design [6], [7]. Comprehensive surveys, such as those by Kalyan et al. and Li et al., offer overviews of transformer-based pre-trained models and their diverse text generation capabilities [8], [9].

To enhance alignment and usability, research has proposed various architectural and methodological optimizations. Gao et al. improved the few-shot learning potential of pre-trained models through contrastive learning methods [10], while Xue introduced dynamic gating mechanisms to reduce computational overhead during alignment tasks [11]. Structural mapping for efficient domain transfer, as explored by Quan, enables more practical applications of distillation and model reuse [12]. Lian ' s work on semantic and factual alignment further contributes to ensuring trustworthy LLM outputs in high-stakes environments [13], and Chen et al. presented a modular approach through the BERT2BERT architecture to facilitate model composability [14].

With regard to specialized retrieval-based systems, RAG (Retrieval-Augmented Generation) has emerged as a critical paradigm for integrating external domain knowledge into LLM inference. Wang proposed a two-stage retrieval and cross-segment alignment method that enhances retrieval quality in LLM workflows [15]. Additionally, Qi demonstrated the value of LLMs in structuring and summarizing unstructured electronic medical records, while Qin applied hierarchical encoding strategies to facilitate compliance risk detection through semantic-structural modeling [16], [17].

Understanding LLMs as reasoning agents and knowledge processors has also drawn attention. Wang et al. conceptualized LLMs as open-domain knowledge graphs, pushing the boundaries of their representation capabilities [18]. Sap et al., through their exploration of social intelligence in neural theory-of-mind tasks, highlighted the limitations and potential of LLMs in emulating human-like inference [19].

Beyond general applications, domain-specific pretraining is essential for improving performance in specialized fields. Webersinke et al. introduced ClimateBERT, a language model fine-tuned for climate-related texts, demonstrating the advantages of contextual adaptation in enhancing output relevance and accuracy [20].

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.

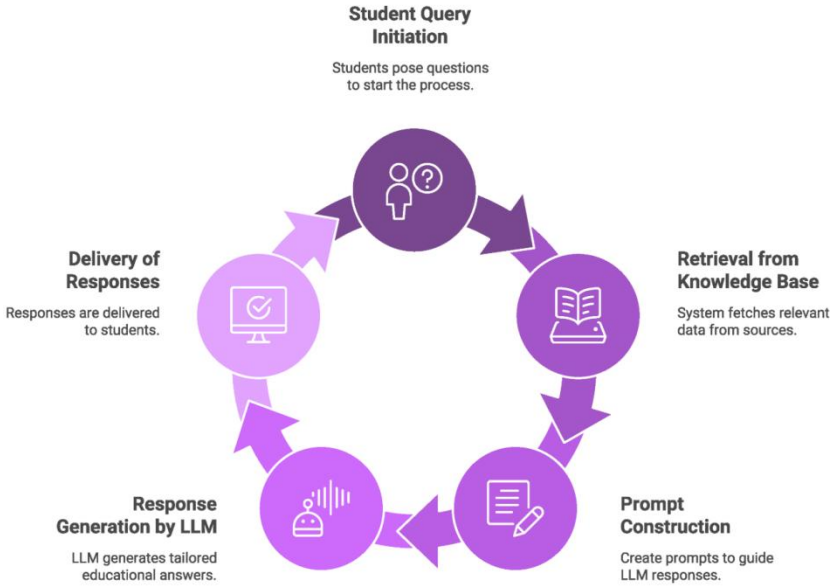

Figure 1. Execution Diagram of the Data Agent

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.

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.

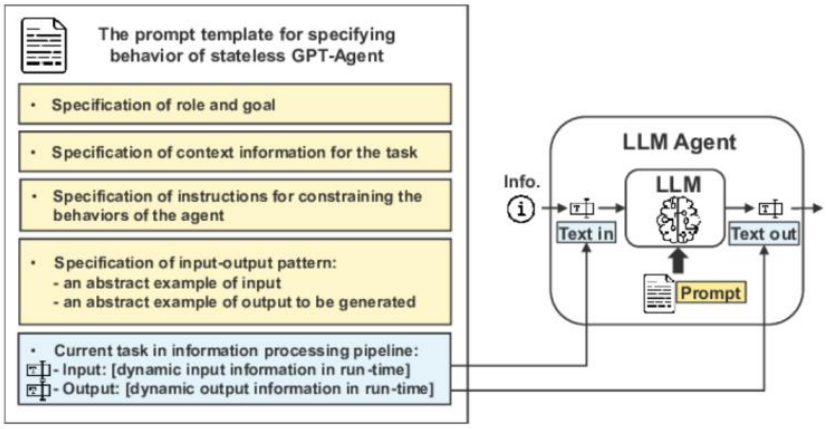

Figure 2. Example of Prompt Engineering for the Data Agent

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.

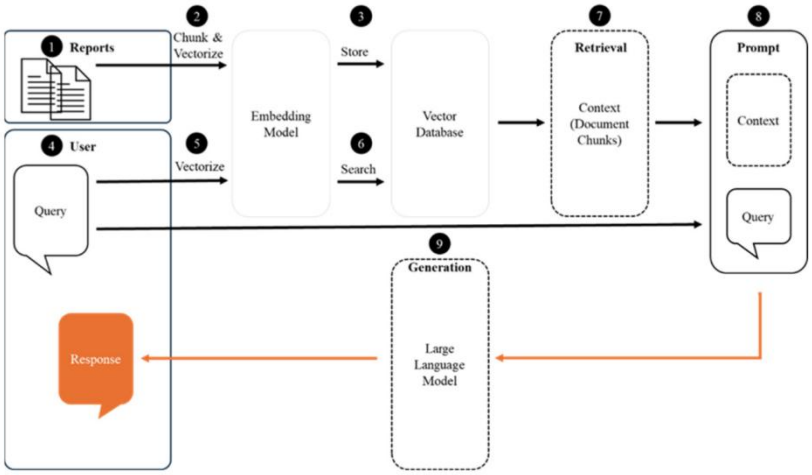

Figure 3. RAG-Based Private Domain Knowledge Question-Answering Framework

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] OpenAI. GPT-4 Technical Report[J]. arXiv preprint arXiv:2303.08774, 2023.

[2] R. Bommasani, D. A. Hudson, E. Adeli, et al., "On the opportunities and risks of foundation models", arXiv preprint arXiv:2108.07258, 2021.

[3] W. Xiang, Z. Zhu, X. Wu, et al., "Large language models in autonomous agents: A survey", arXiv preprint arXiv:2309.07864, 2023.

[4] R. Ying, J. You, C. Morris, et al., "Hierarchical graph representation learning with differentiable pooling", Proceedings of the 2018 Advances in Neural Information Processing Systems, 2018.

[5] Brown T., Mann B., Ryder N., Subbiah M., Kaplan J. D., Dhariwal P., et al., "Language models are few-shot learners," Proceedings of the 33rd Annual Conference on Neural Information Processing Systems, pp. 1877-1901, 2020.

[6] Hoffmann J., Borgeaud S., Mensch A., Buchatskaya E., Cai T., Rutherford E., et al., "Training compute-optimal large language models," arXiv preprint arXiv:2203.15556, 2022.

[7] Hoffmann J., Borgeaud S., Mensch A., Buchatskaya E., Cai T., Rutherford E., et al., "An empirical analysis of compute-optimal large language model training," Proceedings of the 36th Annual Conference on Neural Information Processing Systems, pp. 30016-30030, 2022.

[8] Kalyan K. S., Rajasekharan A., & Sangeetha S., "Ammus: A survey of transformer-based pretrained models in natural language processing," arXiv preprint arXiv:2108.05542, 2021.

[9] Li J., Tang T., Zhao W. X., Nie J. Y., & Wen J. R., "Pre-trained language models for text generation: A survey," ACM Computing Surveys, vol. 56, no. 9, pp. 1-39, 2024.

[10] Gao T., Fisch A., & Chen D., "Making pre-trained language models better few-shot learners," Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics & the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 3816-3830, Aug. 2021.

[11] Xue Z., "Dynamic structured gating for parameter-efficient alignment of large pretrained models," Transactions on Computational and Scientific Methods, vol. 4, no. 3, 2024.

[12] Quan X., "Layer-wise structural mapping for efficient domain transfer in language model distillation," Transactions on Computational and Scientific Methods, vol. 4, no. 5, 2024.

[13] Lian L., "Semantic and factual alignment for trustworthy large language model outputs," Journal of Computer Technology and Software, vol. 3, no. 9, 2024.

[14] Chen C., Yin Y., Shang L., Jiang X., Qin Y., Wang F., et al., "bert2bert: Towards reusable pretrained language models," Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2134-2148, May 2022.

[15] Wang S., "Two-Stage Retrieval and Cross-Segment Alignment for LLM Retrieval-Augmented Generation," Transactions on Computational and Scientific Methods, vol. 4, no. 2, 2024.

[16] Qi N., "Deep learning and NLP methods for unified summarization and structuring of electronic medical records," Transactions on Computational and Scientific Methods, vol. 4, no. 3, 2024.

[17] Qin Y., "Hierarchical semantic-structural encoding for compliance risk detection with LLMs," Transactions on Computational and Scientific Methods, vol. 4, no. 6, 2024.

[18] Wang C., Liu X., & Song D., "Language models are open knowledge graphs," arXiv preprint arXiv:2010.11967, 2020.

[19] Sap M., Le Bras R., Fried D., & Choi Y., "Neural theory-of-mind? on the limits of social intelligence in large LMs," Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 3762-3780, Dec. 2022.

[20] Webersinke N., Kraus M., Bingler J. A., & Leippold M., "ClimateBERT: A pretrained language model for climate-related text," arXiv preprint arXiv:2110.12010, 2021.