UDC: 004.8:621.39:004.912 DOI: 10.20535/2411-2976.12025.22-29

NEXT-GEN TELECOM AI: MASTERING PROMPT ENGINEERING FOR INNOVATION

Mykhailo V. Ilchenko

Educational and Research Institute of Telecommunication Systems
Igor Sikorsky Kyiv Polytechnic Institute, Kyiv, Ukraine

Background. Since 2021, prompt engineering has emerged as a cornerstone of artificial intelligence (AI), revolutionising telecommunications by 2023 through optimised large language models (LLMs).

Objective. This review synthesises existing research to evaluate prompt engineering's transformative role in telecommunications, emphasising practical applications, technical challenges, and future directions.

Methods. This analysis draws on 2021–2025 literature from 31 sources, including IEEE journals, ACM Transactions on Information Systems, NeurIPS proceedings, and arXiv preprints, examining prompt engineering techniques like few-shot learning, chain-of-thought prompting, multi-step prompting, Named Entity Recognition (NER), Retrieval-Augmented Generation (RAG) and more, with a telecom focus (6G and hypothesised 5G applications) contextualized within Natural Language Processing (NLP) advancements.

Results. Although research on prompt engineering specifically for 5G telecommunications is currently limited, it presents substantial opportunities for optimising network performance, diagnostics, documentation handling, enhancing customer support, and driving innovation across both 5G and future 6G networks.

Conclusions. Prompt engineering bridges AI capabilities with telecommunications needs, with techniques like NER and RAG contributing to the enhancement of mobile communications. The dearth of 5G-specific research highlights the urgent need for specialised LLMs in telecommunications and automated prompting to advance solutions for 5G and 6G.

Keywords: *prompt engineering; large language models; telecommunications; network management; 5G; 6G; few-shot learning; chain-of-thought prompting; multi-step prompting; NER; RAG.*

Introduction

The rapid evolution of AI has unleashed transformative tools, reshaping technology. AI encompasses three primary subfields:

1. **Symbolic Automated Reasoning** (rule-based) encompassing:
   a) Expert systems, like Deep Blue's chess algorithms using parallel tree searches;
   b) Planning algorithms, such as pathfinding in navigation.
2. **Machine Learning (ML)** (data-driven) involving:
   a) Supervised learning with labeled datasets, enabling facial recognition via convolutional neural networks;
   b) Hybrid approaches, powering:
      - Generative AI (e,g. DALL-E's image generation with diffusion models or ChatGPT's text production using transformers);
      - Reinforcement learning, as in training language models to optimise their outputs based on feedback (e.g. for handling complex telecom tasks).
3. **Robotics**, focusing on human-robot interaction and autonomy.

Within generative AI, a machine learning subset, prompt engineering has emerged as a pivotal discipline, enabling precise control over AI outputs through natural language inputs. The term "prompt engineering" was first formally introduced by Liu et al. in their seminal 2021 survey [1], categorizing techniques to optimise large language model (LLM) performance for tasks like text generation and question answering. This formalization offers a structured framework for utilizing LLMs, making them accessible to both experts and non-specialists.

This review examines the transformative role of prompt engineering in telecommunications, a sector facing exponential data growth, increasing automation demands, and the rise of AI-driven architectures. LLMs, such as ChatGPT (OpenAI) and LLaMA (Meta AI), excel at synthesising knowledge, diagnosing network faults, and automating customer interactions. Prompt design – the practice of crafting inputs to guide these models effectively – presents both significant challenges and opportunities. Named Entity Recognition (NER), which identifies entities such as device IDs and customer names, holds strong potential for telecommunications applications. However, its integration with prompt engineering remains largely underexplored within fifth-generation (5G) networks.

A key advancement, chain-of-thought (CoT) prompting, introduced by Wei et al. in 2022, enhances LLMs' reasoning by encouraging step-by-step logic, significantly improving their ability to handle complex reasoning tasks [2]. Prompt engineering is instrumental for telecom applications, with studies showcasing network optimisation in light of Foundation Models (FM), part of which are LLMs [3], validating the effectiveness of the multi-agent system by designing a semantic communication system for sixth-generation (6G) systems [4], and formalizing LLM systems as a class of discrete stochastic dynamical systems to explore prompt engineering through the lens of control theory [5]. Retrieval-Augmented Generation (RAG), models which combine pre-trained parametric and non-parametric memory for language generation, further enhance potential telecom applications, like detecting misleading content and automated spam/phishing protection [6, 7]. Multimodal prompting represents another key advancement in next-generation telecom networks. Table 1 outlines the development of major prompting techniques over time, with a focus on their applications in telecommunications.

Table 1. Evolution of key prompt engineering techniques (2021–2025)

Year	Technique	Description	Telecom Implication	Reference
2021	Few-Shot/Zero-Shot Prompting	Formalized by Liu et al., optimises LLMs with minimal or no examples	Designed to optimise input prompts for robust performance and improved NER	[1]
2022	Chain-of-Thought (CoT) Prompting	Introduced by Wei et al., encourages step-by-step reasoning in LLMs	Enhances complex network optimisation and troubleshooting in 5G/6G systems.	[2]
2023	Prompt Pattern Catalogue	White et al.'s catalogue defines structured prompt patterns for LLM tasks	Supports telecom-specific tasks like report generation and context management	[14]
2024	RESPROMPT	Jiang et al.'s multi-step prompting	"Residual connection" adds missing links to transform the linear CoT prompts into graph-like structures to refine and adjust prompts to improve the accuracy of AI-driven responses in network operations and customer service	[22]
2024	Retrieval-Augmented Generation (RAG)	An advanced architecture in natural language processing that combines retriever and generator to generate more accurate, context-rich responses	Combine retrieval from structured/unstructured knowledge bases with generative models	[6, 7]
2024	Multimodal Prompting	Combines text, images, and data for richer LLM responses	Enables intelligent resource allocation and predictive maintenance in 6G networks	[19, 30]

In our corpus, 6G research overshadows 5G, the latter relying on traditional ML (e.g., supervised learning for multiple-input multiple-output (MIMO) antenna systems or beamforming), resulting in a dearth of prompt engineering or NER studies for 5G – a critical research gap. Beyond telecoms, prompt engineering deals with recommendation systems [8] and NER [11]. This paper, spanning 2021–2025, contextualizes telecom applications (6G and potential 5G) within NLP advancements, aiming to:

1. Assess prompt engineering's telecom applications, emphasising NER and RAG potential.
2. Examine techniques and telecom implementations, including NER and RAG integration.
3. Identify challenges and prospects for enhancing 5G network efficiency.

Materials and methods

This review synthesises research from 2021 to 2025 to evaluate prompt engineering's role in telecommunications, emphasising NER and RAG as central components and pivotal tools. Sources were drawn from reputable repositories, namely:

- **IEEE Journals**, a global leader in telecom research [11, 31].

- **ACM Transactions on Information Systems**, a premier information processing journal [9].
- **NeurIPS Proceedings** (a.k.a. Advances in Neural Information Processing Systems), a machine learning flagship [2].
- **AI Open,** a free platform on the theory of artificial intelligence and its applications [30].
- **The ADA User Journal**, a resource on software engineering and AI including prompt patterns for telecom AI applications [27].
- **arXiv Preprints**, an open-access platform for AI and telecom research (e.g., [1, 3–8, 10, 12–26, 28, 29]).

The selection prioritized peer-reviewed articles, preprints, and conference proceedings on LLMs, prompt engineering, NER, and RAG, resulting in 31 sources, of which five are indexed in Scopus [2, 9, 11, 30, 31]. The methodology employed a systematic literature review. Exhaustive searches across Google Scholar, IEEE Xplore, SpringerLink, and arXiv for terms such as "prompt engineering telecoms," "NER 5G LLM," and "RAG telecom" revealed a significant dearth of 5G-specific studies, highlighting the limited research currently available in this area. Specific prompting techniques were analyzed via frameworks like White et al.'s prompt pattern catalogue [14] Schulhoff et al.'s comprehensive taxonomy of prompting techniques [19], and Sahoo et al.'s taxonomy [21]. NER methodologies (e.g., Cheng et al.'s few-shot NER [11]) and RAG approaches (e.g., Lewis et al.'s RAG [6]) were prioritized for entity extraction and data-grounded responses. Prompting techniques have proven instrumental in reducing network optimisation time, enhancing chatbot accuracy, increasing token efficiency through semantic compression, and improving the precision of NER.

Theoretical premises

This section establishes the theoretical foundation for prompt engineering's transformative potential in telecommunications, emphasising named entity recognition NER and RAG. Prompt engineering optimises LLMs (e.g., ChatGPT, LLaMA) through natural language inputs. Liu et al.'s 2021 survey [1] formalized the field, categorizing methods like few-shot and zero-shot prompting to enhance LLM performance without retraining. Schulhoff et al. (2024) catalogued frequently used prompting techniques, including few-shot learning, CoT, and role-based prompting [19].

NER identifies and classifies entities – such as base station IDs and customer names – in unstructured text. Traditionally, it relies on statistical models or transformer-based architectures (e.g., BERT), but is increasingly enhanced through LLM-based prompting. Cheng et al.'s few-shot NER approach [11] demonstrates high precision with minimal examples, making it well-suited for telecom's complex datasets.

RAG, introduced by Lewis et al. [6], augments LLMs to reduce hallucinations and improves accuracy in knowledge-intensive tasks, such as processing telecom documentation. Zhou et al. underscore the generation of domain-specific knowledge that represents a key application of large language models (LLMs) within the telecommunications sector (this involves producing in-depth summaries, interpretations, and overviews of industry standards, technologies, and scholarly research), with LLM-driven systems being capable of generating nuanced, user-tailored content aligned with varying levels of expertise. Such capabilities not only broaden access to telecom knowledge but also help bridge the communication gap between domain experts and non-specialist audiences. Integrating large language models (LLMs) for analysing and classifying a wide range of textual data, including user feedback, standards documentation, technical reports, and troubleshooting records, can significantly enhance the intelligence, reliability, and efficiency of telecom networks [20].

White et al.'s prompt pattern catalogue [14] encompasses the following:

- **Input Semantics** (prompt pattern: Meta Language Creation).
- **Output Customization** (prompt patterns: Output Automater, Persona, Visualisation Generator, Recipe, Template).
- **Error Identification** (prompt patterns: Fact Check List, Reflection).
- **Prompt Improvement** (prompt patterns: Question Refinement, Alternative Approaches, Cognitive Verifier, Refusal Breaker).
- **Interaction** (prompt patterns: Flipped Interaction, Game Play, Infinite Generation).
- **Context Control** (prompt pattern: Context Manager).

Overall, key prompting techniques include:

- **Few-Shot Learning** designed to optimise inputs for robust performance [3] and NER [11].
- **Chain-of-Thought (CoT) Prompting** follows a linear sequence of reasoning steps to enhance reasoning capabilities [2, 17].
- **Iterative Prompting** learns from previous experiences to improve LLM's performance on target tasks iteratively, suitable for handling network optimisation [17].

- **Self-Refined Prompting** can correct LLMs outputs through iterative feedback and refinement prompts, aiming to address network prediction tasks [17].
- **Multi-Step Prompting** organizes complex tasks into a sequence of distinct stages, each with specific instructions, to improve the accuracy and clarity of LLM outputs; it is about breaking down complex tasks into guided subtasks to boosts LLM performance by easing cognitive load and improving output structure [23].
- **Residual Connection Prompting (RESPROMPT)** is helpful where traditional chain-of-thought (CoT) prompting often falls short when handling complex, multi-step reasoning tasks in LLMs. RESPROMPT, an improved Multi-Step technique that breaks down complex tasks into subtasks, using intermediate prompts to retain context from previous steps, ensuring coherence in multi-step processes. At its core, RESPROMPT reconstructs the implicit reasoning structures found in complex queries by incorporating residual connections to transform linear CoT prompts into graph-like representations that better capture the logical flow of reasoning [22].
- **Tree of Thought Prompting (ToT)** is an advanced prompting framework for LLMs that organizes reasoning into a graph structure (organizes reasoning into a tree structure, where nodes represent intermediate thoughts or subtasks, and branches denote possible reasoning paths), unlike CoT, which follows a linear sequence, ToT enables deliberate exploration of multiple reasoning paths [28].
- **Graph of Thought Prompting (GoT)** non-linear, graph-based reasoning with interconnected nodes and edges (e.g., linking entities to network topology) [29].
- **In-Context Learning (ICL)** that supplies exemplars and/or relevant instructions within the prompt [19], ICL empowers LLMs to learn from task-specific instruction and demonstrations [17].
- **Role-Based Prompting** used to assign roles, for tailoring outputs to roles [19]. For instance, “5G analyst”.
- **NER** or entity extraction through prompt-based techniques [11].
- **RAG** improves language model outputs by incorporating information retrieved from external knowledge bases or documents; RAG grounds responses in network data, enhancing anomaly detection and technical documentation processing [6, 7].
- **Multi-Agent Systems –** mainly for 6G communications [4], potentially extendable to 5G applications.
- **Multimodal Prompting** is a noteworthy emerging trend in prompt engineering [19], is about the process of providing prompts to Generative AI models that combine multiple data modalities – such as text, images, numerical data, or audio – to elicit more accurate and contextually rich responses. What is more, the so-called Vision-Language Pre-training (VLP) models have shown promising capabilities in grounding natural language in image data, facilitating a broad range of cross-modal tasks, with novel Color-based Prompt Tuning (CPT) paradigm for tuning VLP models on the way, enabling few-shot and even zero-shot visual grounding capabilities of VLP models [30].

Another relevant issue is **semantic compression**, a key advancement that can reduce token size while preserving meaning. Gilbert et al. (2023) offer an initial evaluation of compressing Large Language Models (LLMs), namely ChatGPT-3.5 and ChatGPT-4, and introduce two novel metrics, Semantic Reconstruction Effectiveness (SRE) and Exact Reconstruction Effectiveness (ERE), with the outcome emphasising LLMs’s ability to effectively reconstruct source code from compressed text descriptions, achieving significant functional accuracy [15]. At the same time, as Jiang et al. (2024) rightly note, to reduce irrelevant information in the documents, alongside the **condensate agent** (used to compress the documents), an **inference agent** must be utilized to extract specialised knowledge relevant to the user's query from the compressed content [4]. Research indicates potential problems due to verbose prompts [25]. Therefore, it stands to reason that semantic compression can enhance prompt engineering by enabling concise, information-dense prompts that reduce tokens while preserving task-critical meaning, thus improving efficiency of AI-powered telecommunications solutions.

Results

Utilizing prompt engineering techniques is promising in several ways, as they enhance overall network efficiency by enabling better fault diagnostics, more accurate customer support, and telecom documentation processing. It addresses the telecom industry's need for real-time, resource-efficient solutions.

For instance, proceeding from White et al.'s prompt pattern catalogue [14] specific telecom solutions might, above all, involve **network management** in a number of ways:

- accurate telecom term interpretation via Input Semantics (prompt pattern: Meta Language Creation);
- formatting NER outputs as reports through Output Customization (prompt patterns: Output Automater, Persona, Visualisation Generator, Recipe, Template);
- the flagging of invalid entities via Error Identification (prompt patterns: Fact Check List, Reflection);
- refining NER prompts for precision through Prompt Improvement (prompt patterns: Question Refinement, Alternative Approaches, Cognitive Verifier, Refusal Breaker);
- managing NER tasks in support dialogues via Interaction (prompt patterns: Flipped Interaction, Game Play, Infinite Generation);
- instructing the LLM to operate within a specified context, such as a particular domain of telecoms, assuming a role and acting, for instance, as network engineer) via Context Control (prompt pattern: Context Manager).

Other benefits relate to improving NER extraction due to contextual understanding and resolving ambiguity, addressing specific network scenarios to reduce or even eliminate irrelevant or generic responses by supporting multi-step tasks like those handled by Multi-Step Prompting, specifically, by RESPROMPT.

ICL technique seems to be transformative for telecom AI, particularly in 5G and 6G networks, where rapid adaptation to new tasks (e.g., fault detection, customer query resolution) is essential. Telecom systems generate vast, heterogeneous data (e.g., network logs, user queries), and ICL's ability to learn from exemplars/instructions enables LLMs to handle tasks like fault diagnostics, reducing mean time to resolution (MTTR), customer support, resolving multi-turn queries via chatbots, to name a few. Operations support systems and business support systems (OSS/BSS) in telecoms could also benefit from efficient prompting, as it could potentially facilitate network inventory management (tracking all network assets (hardware, software, etc.), ensuring performance management monitoring (network performance and QoS (quality of service), fault management (detecting, diagnosing, and resolving network issues), as well as network planning and design (future network evolution), customer support (like chatbot accuracy) and revenue-related activities (billing etc.)

Multimodal prompting is becoming a cornerstone for realizing the full potential of 5G and 6G applications. By enabling AI models to process and understand information from diverse data sources beyond just text – such as images (e.g., network topology graphs), numerical data, video, audio etc., it allows for the fusion of heterogeneous data streams from various sensors and devices within the network, enabling advanced functionalities such as intelligent resource allocation, predictive maintenance, and context-aware service delivery. However, processing multimodal inputs (e.g., large images) has its limitations, such as increased response times in 5G systems. Another challenge is avoiding possible misinterpretations of such complex data. In Table 2 we highlight key challenges in applying prompt engineering to telecommunications alongside some tentative corresponding solutions.

Table 2. Challenges and solutions in telecom prompt engineering

Challenge	Description	Suggested Solution	Ref.
Latency	Time delay in network communication	Employing semantic compression and optimised multimodal prompting	[15; 30]
Domain Specificity	The tailored application of technologies, protocols, models, tools, solutions and standards to meet the unique telecom requirements	Fine-tuning LLMs with telecom datasets (e.g., 5G Instruct Forge pipeline)	[20, 31]
Real-Time Constraints	Strict timing and performance requirements to ensure the reliable and timely transmission, processing, and delivery of data, voice, or video over a network	Iterative prompting combined with automated prompt refinement tools to enhance efficiency	[17, 19, 23]
Multimodal Data Misinterpretation	Risks of complex data (e.g., images, logs) misinterpretation in 5G/6G systems	Robust VLP models to enable multimodal prompting with improved context control and input understanding	[19, 30]
Lack of 5G-	Limited studies on prompt engineering for 5G	Developing prompt libraries tailored for	[13,

Specific Research	compared to 6G	5G, along with specialised NER and RAG frameworks optimised for telecoms	18, 31]

An important remark concerning **6G Communications** and **5G Hypothetical Applications** should be made.

In 6G, security prompts [18], standard interpretation [16], and multi-agent systems [4] utilize NER, aided by semantic compression [15]. RAG-based systems like TelcoRAG process telecom documentation with high accuracy [7]. A promising approach is offered by Jiang et al. (2024) suggesting a natural language-based multi-agent system that tackles communication-related tasks through three key components: first, Multi-agent Data Retrieval (MDR) that refines and summarizes knowledge from a knowledge base using condensate and inference agents, extending LLM knowledge for 6G communications; second, Multi-agent Collaborative Planning (MCP) that employs multiple planning agents to generate diverse solutions based on the retrieved knowledge, and Multi-agent Evaluation and Reflexion (MER) that utilizes the evaluation agent to assess the solutions, applying the reflexion agent and refinement agent to provide suggestions regarding the improvement of solutions [4]. While comprehensive 5G studies are yet to emerge, the capabilities of NER and RAG suggest promising applications in several areas: log analysis, customer support, and security [13, 18]. These AI techniques hold the potential to significantly enhance the efficiency and intelligence of next-generation mobile networks.

Conclusion

Prompt engineering is revolutionising telecommunications, optimising networks, enhancing customer support, and enabling 6G innovation. It bridges the gap between rapidly evolving AI and present-day telecom demands. The integration of NER and RAG further amplifies this potential by enabling precise entity extraction and the generation of data-grounded responses, thus optimising performance, diagnostics, support, and documentation handling. Despite this readiness for diagnostics and automation, however, the specific applications of NER in the context of 5G require further exploration.

Liu et al.'s 2021 survey [1] formalized prompt engineering, with Wei et al.'s CoT prompting [2], White et al.'s catalogue [12], Schulhoff et al.'s taxonomy [19], and Sahoo et al.'s taxonomy [21] advancing the field. Schmidt et al. (2023) present prompt patterns philosophy that effectively addresses universal key challenges, laying the groundwork for next-generation innovations [28].

Although LLMs have achieved impressive general capabilities, their domain-specific performance remains a challenge. Recent work with the 5G Instruct Forge pipeline demonstrates that structuring 3GPP specifications into specialised datasets enables the fine-tuning of LLMs that surpass GPT-4 on 5G-related tasks [31]. Specific telecom challenges – such as latency, domain specificity, and real-time constraints – require specialised LLMs with relevant prompt patterns [17, 20, 23, 24]. The research gap in 5G for NER and RAG highlights opportunities for future developments in the field. Defining precise requirements for 6G networks is inherently complex; however, structured prompting can streamline this process by guiding the development of tailored AI solutions.

The dynamic telecom environment, driven by real-time traffic demands and domain-specific needs, needs collaboration between engineers and AI systems. Well-designed prompt patterns enable real-time interaction to support adaptive network management. Poorly crafted prompts can hinder AI performance in tasks like diagnostics or optimisation. However, iterative refinement techniques improve accuracy and reliability. Future research should prioritize telecom-specific prompt libraries and automated refinement tools to enhance AI's role in meeting real-time constraints and domain-specific requirements, making it a challenging yet transformative endeavor.

Advancements in next-gen network technologies would enhance efficiency through improved spectrum utilization, energy optimisation, latency reduction, and dynamic traffic management via network slicing. These technologies also ensure scalability by supporting growing numbers of users, devices, and services through virtualization and cloud-native architectures. Continued effort like this can ultimately reshape challenges as opportunities. As such, it is a noble effort.

References

1. Liu P., Yuan W., Fu J., Jiang Z., Hayashi H., Neubig G. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ─ arXiv preprint arXiv:2107.13586v, 2021 https://doi.org/10.48550/arXiv.2107.13586

2. Wei J., Wang X., Schuurmans D., Bosma M., Ichter B., Xia F., Chi E. H., Le Q. V., Zhou D. Chain-of-thought prompting elicits reasoning in large language models. ─

Advances in Neural Information Processing Systems. ─ 2022. ─ Vol. 35. ─ pp. 24824─24837. Scopus https://doi.org/10.48550/arXiv.2201.11903

3. Britto R., Murphy T., Iovene M., Jonsson L., Erol-Kantarci M., Kovács B. Telecom AI native systems in the age of generative AI—An engineering perspective. ─ arXiv preprint arXiv: arXiv:2310.11770v1, 2023. https://doi.org/10.48550/arXiv.2310.11770

4. Jiang F., Wang X., Li Y., Zhang H. Large language model enhanced multi-agent systems for 6G communications. ─ arXiv preprint arXiv:2310.11770, 2023. https://arxiv.org/html/2312.07850v1

5. Bhargava A., Dua M., Kunc M., Pennock S. What's the magic word? A control theory of LLM prompting. ─ arXiv preprint arXiv:2402.00812, 2024. https://doi.org/10.48550/arXiv.2310.04444

6. Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Küttler H., Lewis M., Yih W., Rocktäschel T., Riedel S., Kiela D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. ─ arXiv preprint arXiv:2005.11401v4
2021. https://doi.org/10.48550/arXiv.2005.11401

7. Bornea A.-L., Ayed F., De Domenico A., Piovesan N., Maatouk A. Telco-RAG: Navigating the challenges of retrieval-augmented language models for telecommunications. ─ arXiv preprint arXiv:2404.15939v3, 2024. https://doi.org/10.48550/arXiv.2404.15939

8. Li L., Zhang Y., Chen L. Personalized prompt learning for explainable recommendation. ─ arXiv preprint arXiv: arXiv:2202.07371v2,
2023 https://doi.org/10.48550/arXiv.2202.07371

9. Zhang H., Zhu Q., Dou Z. A unified prompt-aware framework for personalized search and explanation generation. ─ ACM Transactions on Information Systems. ─ 2025. ─ Vol. 43, No. 3. ─ pp. 1─26. Scopus DOI: 10.1145/3716131

10. Mao W., Wu J., Chen W., Gao C., Wang X., He X. Reinforced prompt personalization for recommendation with large language models. ─ arXiv preprint arXiv:2407.17115v2,
2025 https://doi.org/10.48550/arXiv.2407.17115

11. Cheng Q., Chen L., Hu Z., Tang J., Xu Q., Ning B. A novel prompting method for few-shot NER via LLMs. ─ IEEE Transactions on Artificial Intelligence. ─ 2024. ─ Vol. 5, No. 8. ─ pp. 4001─4010. Scopus https://doi.org/10.1016/j.nlp.2024.100099

12. Maatouk A., Li Z., Xiao M., Wang Y. Large language models for telecom: Forthcoming impact on the industry. ─ arXiv preprint arXiv:2308.06013 https://doi.org/10.48550/arXiv.2308.06013

13. Liu Y., Zhang H., Li X., Wang Q. Jailbreaking ChatGPT via prompt engineering: An empirical study. ─ arXiv preprint arXiv:2305.13860v2, 2024 https://doi.org/10.48550/arXiv.2305.13860

14. White J., Hays S., Fu Q., Schmidt D. C. A prompt pattern catalogue to enhance prompt engineering with ChatGPT. ─ arXiv preprint arXiv2302.11382v1, 2023. https://doi.org/10.48550/arXiv.2302.11382

15. Gilbert H., Lee J., Patel R., Singh A. Semantic compression with large language models. ─ arXiv preprint arXiv:2304.12512v1, 2023. https://doi.org/10.48550/arXiv.2304.12512

16. Karapantelakis A., Li Y., Zhang H., Wang Q. Using large language models to understand telecom standards. ─ arXiv preprint arXiv:2404.02929v1, 2024. ─ 14 p. https://arxiv.org/html/2404.02929v1

17. Zhou H., Li X., Zhang Y., Wang Q. Large language models for wireless networks: An overview from the prompt engineering perspective. ─ arXiv preprint arXiv:2411.04136v2, 2024. https://doi.org/10.48550/arXiv.2411.04136

18. Nguyen T., Tran H., Le P., Pham Q. Large language models in 6G security: Challenges and opportunities. ─ arXiv preprint arXiv:2403.12239v1, 2024. https://arxiv.org/html/2403.12239v1

19. Schulhoff S., Ilie A., Balepur N., Kahadze K., Liu A., Si C., Sun Y., Zhou C., Zhu C., Pissinou N. The prompt report: A systematic survey of prompting techniques. ─ arXiv preprint arXiv: 2406.06608v6, 2024. https://doi.org/10.48550/arXiv.2406.06608

20. Zhou H., Hu C., Yuan Y., et al. Large language model (LLM) for telecommunications: A comprehensive survey on principles, key techniques, and opportunities. ─ arXiv preprint arXiv: 2405.10825v2, 2024. https://doi.org/10.48550/arXiv.2405.10825

21. Sahoo P., Singh A. K., Saha S., Jain V., Mondal S., Chadha A. A systematic survey of prompt engineering in large language models: Techniques and applications. ─ arXiv preprint arXiv: arXiv:2402.07927v2, 2025. https://doi.org/10.48550/arXiv.2402.07927

22. Jiang S., Zhang Y., Wang X., Li Q., Chen H., Liu J. RESPROMPT: Residual connection prompting advances multi-step reasoning in large language models. ─ arXiv preprint arXiv::2310.04743v2, 2024. https://doi.org/10.48550/arXiv.2310.04743

23. Ahmed, N., Piovesan N., De Domenico A., Choudhury S. Linguistic intelligence in large language models for telecommunications. ─ arXiv preprint arXiv:2402.15818v1, 2024.. https://doi.org/10.48550/arXiv.2402.15818

24. Piovesan N., De Filippo De Grazia M., Zorzi M. Telecom language models: Must they be large? ─ arXiv preprint arXiv:2403, 2024 https://doi.org/10.48550/arXiv.2403.04666

25. Denny P., Kumar V., Giacaman N. Conversing with Copilot: Exploring prompt engineering for solving CS1 problems using natural language. ─ arXiv preprint arXiv: 2210.15157v1, 2022 https://doi.org/10.48550/arXiv.2210.15157

26. Huang Y., Li X., Zhang H., Wang Q. Large language models for networking: Applications, enabling techniques, and challenges. ─ arXiv preprint arXiv: 2311.17474v1, 2023 https://doi.org/10.48550/arXiv.2311.17474

27. Schmidt D. C., White J., Fu Q., Hays S. Towards a catalogue of prompt patterns to enhance the discipline of

prompt engineering. ─ ADA User J., 2023. [Online]. Retrieved from: https://www.dre.vanderbilt.edu/~schmidt/PDF/ADA-User-Journal.pdf
28. Yao S., Yu D., Zha, J., Shafran I., Griffiths T. L., Cao Y., Narasimhan K. (2023). Tree of Thoughts: Deliberate Problem Solving with Large Language Models. arXiv preprint arXiv: 2305.10601v2, 2023 https://doi.org/10.48550/arXiv.2305.10601
29. Yao Y., Li Z., & Zhao H. (2023). Beyond Chain-of-Thought, Effective Graph-of-Thought Reasoning in Large Language Models. arXiv preprint arXiv:2305.16582v2, 2024 https://doi.org/10.48550/arXiv.2305.16582
30. Yao Y., Zhang A., Zhang Z., Liu Z.,Chua T.-S., Sun M. CPT: Colorful Prompt Tuning for pre-trained vision-language models, AI Open, Volume 5, 2024, pp. 30-38 Scopus https://doi.org/10.1016/j.aiopen.2024.01.004
31. Said A.I.A., Mekrache A., Boutiba K., Ramantas K., Ksentini A., Rahmani M. 5G INSTRUCT Forge: An Advanced Data Engineering Pipeline for Making LLMs Learn 5G ─ IEEE Transactions on Cognitive Communications and Networking, Volume 11, 2025, pp. 974–986. Scopus https://ieeexplore.ieee.org/document/10794684

Ільченко М.В.

ШІ нового покоління в телекомунікаціях: оволодіння інженерією запитів заради інновацій

Навчально-науковий інститут телекомунікаційних систем КПІ ім. Ігоря Сікорського, м. Київ, Україна

Проблематика. Починаючи з 2021 року, інженерія запитів (prompt engineering) стала наріжним каменем штучного інтелекту (ШІ), а до 2023 року здійснила революцію в галузі телекомунікацій шляхом оптимізації великих мовних моделей (LLMs).

Мета. Цей огляд узагальнює наявні дослідження з метою оцінки трансформаційної ролі інженерії запитів у телекомунікаціях, з акцентом на практичні застосування, технічні виклики та майбутні напрями розвитку.

Методи. Аналіз базується на літературі за 2021–2025 роки із 31 джерела, включаючи журнали IEEE, транзакції ACM з інформаційних систем, матеріали конференції NeurIPS, а також препринти arXiv. Розглядаються техніки інженерії запитів, такі як навчання за кількома прикладами (few-shot learning), ланцюжкове мислення (chain-of-thought prompting), багатокрокові запити (multi-step prompting), розпізнавання іменованих сутностей (NER), генерація з доповненням через пошук (RAG) та інші, із фокусом на телекомунікаційні застосування (6G та гіпотетичні застосування для 5G) у контексті розвитку обробки природної мови (NLP).

Результати. Хоча досліджень із застосування інженерії запитів у сфері телекомунікацій 5G наразі небагато, ця галузь відкриває значні можливості для оптимізації роботи мереж, діагностики, обробки документації, покращення обслуговування клієнтів та стимулювання інновацій як у 5G, так і у майбутніх мережах 6G.

Висновки. Інженерія запитів поєднує можливості ШІ з потребами телекомунікацій, а такі техніки, як NER і RAG, сприяють вдосконаленню мобільних комунікацій. Недостатність досліджень, специфічних для 5G, підкреслює нагальну потребу у спеціалізованих LLM для телекомунікацій та автоматизованих стратегіях формування запитів для розвитку рішень для 5G та 6G.

Ключові слова: *інженерія запитів; великі мовні моделі; телекомунікації; управління мережами; 5G; 6G; навчання за кількома прикладами; ланцюжкові запити; багатокрокові запити; розпізнавання іменованих сутностей (NER); генерація з доповненням через пошук (RAG).*

Received by the Editorial Office
April 29, 2025

Accepted for publication
June 9, 2025