

ПОБУДОВА ЗАПИТІВ ДО ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ ДЛЯ ВИРІШЕННЯ ТЕЛЕКОМУНІКАЦІЙНИХ ЗАДАЧ

Ільченко М.В., Іщенко М.О., Педан С.І.

Навчально-науковий інститут телекомунікаційних систем

КПІ ім. Ігоря Сікорського, Україна

*E-mail: michael13ilch@gmail.com, ishchenkomaria391@gmail.com,
stas.pedan@gmail.com*

CONSTRUCTION OF LARGE LANGUAGE MODELS PROMPTS FOR SOLVING TELECOMMUNICATION PROBLEMS

Application of prompt engineering techniques in telecommunications, with a focus on optimizing the performance of large language models (LLMs) for a range of tasks, such as customer support, network management, and the personalization of information requests, specifically regarding the implementation of software products are researched. Analysis of basic prompt design principles, including few-shot learning, role prompting, chain-of-thought reasoning, multi-step prompting, iterative prompt engineering and refinement are conducted. Current challenges and future direction of LLMs usage in the telecommunications sector are also outlined.

У 2023 році у світі виникла новітня галузь наукового пошуку - інженерія запитів як результат стрімкого розвитку великих мовних моделей (ВММ), що відкриває безпрецедентні можливості для телекомунікаційної галузі. Реалізація потенціалу цих моделей вимагає ефективної інженерії запитів. У роботі досліджується застосування методик інженерії запитів у телекомунікаціях, з акцентом на оптимізації продуктивності ВММ для вирішення низки технічних задач, таких як підтримка клієнтів, управління мережею та персоналізація інформаційних запитів.

ВММ можуть генерувати нові текстові дані та тренуються на величезних обсягах текстових даних із різних джерел. ВММ продемонстрували вражаючі можливості в різних завданнях обробки природної мови, таких як узагальнення тексту, відповіді на запитання, аналіз настроїв, генерація коду та машинний переклад. Вони також можуть створювати історії, вірші, жарти, тексти пісень, зображення, аудіо та код, усе це керується текстовими запитами [11].

Продуктивність ВММ істотно залежить від якості вхідних запитів. Передбачається, що ВММ будуть інтегровані в усі наші робочі процеси [7]. Метою цієї роботи є дослідження сучасних підходів до побудови запитів ВММ для вирішення актуальних задач у телекомунікаційній галузі.

Термін "інженерія запитів" почав активно застосовуватися у науковій літературі з 2023 року. Нині існує ціла низка загальних методик інженерії запитів, які можна пристосувати для галузевих потреб, наприклад для сфери телекомунікацій [4, 5, 10]. До основних типів запитів ВММ належать такі:

1. Навчання з кількома прикладами [2, 9] – це надання ВММ декількох прикладів запитів клієнтів і бажаних відповідей. Приклад: аналіз мережевого трафіку через надання ВММ низки типових прикладів аномальної активності.

2. Рольова побудова запитів [10] надає ВММ певну роль, на зразок роль "(Уявіть, що) Ви -" Наприклад, коли телекомунікаційна компанія використовує чат-боти для підтримки клієнтів, то для отримання кращих відповідей можна використовувати рольове підказування на зразок: "Ви - досвідчений технічний спеціаліст телекомунікаційної компанії. Відповідайте на запитання клієнтів чітко та лаконічно." Це допоможе чат-боту краще зрозуміти контекст та надавати точніші відповіді.

3. Міркування за допомогою ланцюга думок [1, 8] скеровує ВММ на покроковий процес міркування, щоб допомогти ІІ розбити проблему на менші кроки та зробити висновок. Наприклад: "Клієнт повідомляє, що інтернет не працює. Спочатку перевірте, чи підключений маршрутизатор. Потім перевірте, чи є з'єднання з інтернетом. Якщо з'єднання є, перевірте налаштування мережі."

4. Багатоступеневі запити [6] передбачають розбиття складного запиту на послідовність простіших кроків. Кожен крок використовує результати попереднього аби поступово наблизитися до кінцевого рішення. Уявімо, що нам потрібно діагностувати проблему з повільним інтернетом у клієнта. Ми можемо використовувати багатоступеневі запити:

Крок 1: Збір інформації про проблему. Крок 2: Визначення типу з'єднання та обладнання. Крок 3: Перевірка основних параметрів мережі. Крок 4: Діагностика можливих причин. Крок 5: Надання рекомендацій.

5. Ітеративна побудова та уточнення запитів [3]. Прикладом є сценарій, коли мережевий адміністратор намагається діагностувати періодичну втрату пакетів у критичній бізнес-мережі.

Ітерація 1: початковий запит: "Проаналізуй останні журнали мережі щодо втрати пакетів і визнач потенційні причини." Ітерація 2: уточнення (фокус на конкретному часовому проміжку). Ітерація 3: подальше уточнення (глибоке занурення в пропускну здатність). Ітерація 4: цілеспрямоване дослідження (фокус на аномальній активності).

Кожна ітерація базується на попередній відповіді. Запити стають все більш конкретними та цілеспрямованими.

Зазначимо, що наведена вище класифікація не є єдиною можливою. Наприклад, у праці Джулс Вайт [10] наводиться складніша класифікація з дещо відмінного ракурсу, позаяк вона групує шаблони запитів за їхньою функціональністю або метою, наприклад, "Семантика вхідних даних", "Взаємодія" тощо. Це відрізняється від класифікації за техніками, такими як "навчання з кількома прикладами" або "міркування за допомогою ланцюга думок", які описують конкретні методи побудови запитів.

Як зазначають дослідники, перспективи окресленого напрямку допоможуть вирішити питання мережевих аномалій, моделювання мереж, сталого розвитку та екологічної безпеки, причому успіх ВММ у телекомунікаційній сфері буде залежати від спеціалізованих наборів даних при створенні базових моделей для телекомунікацій [5].

Розвиток ВММ відкриває нові можливості для комунікацій 6G, наприклад, оптимізація та управління мережею, дозволяючи користувачам вводити вимоги до завдань у ВММ природною мовою. Покращення знань ВММ про

комунікації, а також можливостей логічного міркування може значно підвищити потенціал ВММ для комунікацій 6G.

Отже, інженерія запитів є критично важливим компонентом ВММ в цілому та у телекомунікаційній галузі зокрема. Так, інжиніринг запитів дозволяє створювати запити, які генерують код для конкретних задач в телекомунікаціях (наприклад, можна використовувати LLM для оптимізації наявних кодів, автоматичного створення скриптів для налаштування мережевого обладнання, розробки протоколів, аналізу трафіку та великих обсягів даних, що генеруються мережами). Використання ефективних принципів розробки запитів допомагає покращити здатність ВММ аналізувати дані мереж, діагностувати та вирішувати їхні проблеми, покращувати безпеку та адаптувати ШІ до потреб споживача. Виклики сьогодення, пов'язані з ВММ, стимулюють оптимізацію їхнього використання в телекомунікаційному секторі.

Таким чином, в роботі розглянуті основні принципи розробки запитів ВММ, включаючи навчання з кількома прикладами, рольові запити, міркування за допомогою ланцюга думок, багатоступеневі запити, ітеративні запити з уточненнями та продемонстроване їхнє застосування у сфері телекомунікаційних задач. Також окреслено проблеми, виклики і перспективи використання ВММ у телекомунікаційному секторі.

Література

1. "Chain-of-Thought Prompting Guide." PromptHub. Accessed [12 February 2025]. <https://www.prompthub.us/blog/chain-of-thought-prompting-guide>
2. "Few-Shot Learning." Nebius Blog. Accessed [12 February 2025]. <https://nebius.com/blog/posts/few-shot-learning>
3. "Iterative Prompt Engineering and Refinement." DataCamp. Accessed [12 February 2025]. <https://campus.datacamp.com/courses/chatgpt-prompt-engineering-for-developers/advanced-prompt-engineering-strategies?ex=13>
4. Lin, Xingqin et al., "A Primer on Generative AI for Telecom: From Theory to Practice." arXiv, August, 16, 2024 <https://arxiv.org/abs/2408.09031>
5. Maatouk, Ali et al. "Large Language Models for Telecom: Forthcoming Impact on the Industry." IEEE Communications Magazine 63 (2023): 62-68. <https://arxiv.org/pdf/2308.06013>
6. "Multi-Step Prompting." DataCamp. Accessed [12 February 2025]. <https://campus.datacamp.com/courses/chatgpt-prompt-engineering-for-developers/advanced-prompt-engineering-strategies?ex=5>
7. "The Rise of the Prompt Engineer – and Why It Matters." World Economic Forum. May 4, 2023. <https://www.weforum.org/stories/2023/05/growth-summit-2023-the-rise-of-the-prompt-engineer-and-why-it-matters/>
8. Wei, Jason et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," arXiv, February 22, 2023, <https://arxiv.org/abs/2201.11903>
9. "What Is Few-Shot Learning?" IBM. Accessed [12 February 2025]. <https://www.ibm.com/think/topics/few-shot-learning>
10. White, Jules et al., "A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT," arXiv, February 22, 2023, <https://arxiv.org/abs/2302.11382>
11. Britto, Ricardo, Timothy Murphy, Massimo Iovene, Leif Jonsson, Melike Erol-Kantarci, and Benedek Kovács. 2023. "Telecom AI Native Systems in the Age of Generative AI -- An Engineering Perspective." <https://arxiv.org/abs/2310.11770>