

АВТОМАТИЗАЦИЯ УПРАВЛЕНИЯ ПРОМПТАМИ ДЛЯ ПОВЫШЕНИЯ ТОЧНОСТИ ГЕНЕРАТИВНЫХ МОДЕЛЕЙ

Лисицина В.В. (Университет ИТМО)
Научный руководитель – Авдюшина А.Е.
(Университет ИТМО)

Введение. С развитием генеративных языковых моделей искусственного интеллекта возросла потребность в эффективном управлении промптами – текстовыми запросами, на основе которых модели формируют ответы. Качество генерации напрямую зависит от структуры промпта, его контекста и специфики формулировки. Однако процесс создания и тестирования промптов зачастую носит эмпирический характер, что приводит к снижению точности и предсказуемости результатов.

Основная часть. Для эффективного подбора промптов под определённую задачу необходимо учитывать множество факторов, влияющих на точность и релевантность ответа, возвращаемого моделью. Эти факторы включают классификацию задачи, определение шаблонов промптов, контекст, специфику данных и другие характеристики. Ошибки в формулировке и дизайне промпта могут приводить к неоднозначности интерпретации запроса моделью, что снижает точность результата.

Одним из этапов управления промптами является классификация задач, для которых они применяются. В зависимости от типа задачи можно выделить несколько основных категорий: генерация текста, анализ тональности, логические рассуждения, извлечение информации и другие.

После классификации задачи следует определение подходящего метода промптинга. Современные методы можно разделить на несколько основных групп:

- Zero-Shot Prompting
- Few-Shot Prompting
- Chain-of-Thought Prompting
- Retrieval-Augmented Generation

Выбор подходящего метода промптинга играет ключевую роль в точности и эффективности работы генеративных моделей. Однако вручную подбирать оптимальные промпты для каждой задачи может быть затратно и не всегда эффективно. Поэтому возникает необходимость в автоматизированных решениях, способных анализировать входные данные, применять наиболее подходящие стратегии промптинга и адаптировать их под конкретные задачи.

В данной работе рассмотрено решение по автоматическому подбору промптов под определённую задачу пользователя. В рамках реализации предлагается три модуля – PromptGeneration, PromptTesting, PromptLLM. PromptLLM отвечает за взаимодействие с LLM-моделью пользователя, обеспечивая передачу промптов и получение ответов. PromptGeneration занимается формированием промптов на основе входных данных пользователя, истории запросов и применяемых методик промптинга. Он использует предопределённые шаблоны для повышения эффективности формулировки запросов. PromptTesting – тестирует созданные промпты, анализируя точность и релевантность ответов, предоставляемых моделью с помощью пользовательских метрик. Для хранения промптов и их метаданных была спроектирована база данных на PostgreSQL. Чтобы обеспечить гибкость и асинхронность обработки данных, взаимодействие между сервисами организовано через

Kafka, что позволяет минимизировать задержки при обработке запросов, эффективно распределять нагрузку между сервисами и сохранять систему устойчивой к высоким нагрузкам.

Выводы. Представленное решение, позволяет автоматизировать адаптацию и тестирование промтов, тем самым делает работу с языковыми генеративными моделями более эффективной.

Список использованных источников:

1. Schulhoff S., Ilie M., Balepur N., Kahadze K., Liu A. The Prompt Report: A Systematic Survey of Prompting Techniques // arXiv (Cornell University). – 2024. 80 p.
2. Deng Y., Zhang W., Chen Z., Gu Q. Rephrase and Respond: Let Large Language Models Ask Better Questions for Themselves // arXiv (Cornell University). – 2023. – 90 p.
3. 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 // ACM Computing Surveys. – 2023. – Vol. 55, Issue 9. – Article No. 195. – P. 1-35