УДК 004.89

АНАЛИЗ СУЩЕСТВУЮЩИХ ИНСТРУМЕНТОВ И ПОДХОДОВ ПРИ РЕАЛИЗАЦИИ ОРКЕСТРАЦИИ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Иржанова Ю. И. (ИТМО), **Шафиков М.А.** (ИТМО)
Научный руководитель – преподаватель, Терещенко В.В.
(ИТМО)

Введение. Современные большие языковые модели (Large Language Models, LLM) активно применяются для обработки естественного языка, анализа данных и мультимодальной обработки информации. Вместе с тем, их эффективная интеграция в сложные вычислительные процессы требует разработки механизмов оркестрации. Оркестрация LLM – это процесс управления и координации работы различных языковых моделей и инструментов, позволяющий достичь слаженного взаимодействия, высокой производительности и масштабируемости.

Основные проблемы при оркестрации LLM заключаются в том, что отсутствуют методы оценки и выбора моделей для формирования ансамблей и увеличивается сложность интеграции различных моделей. Также дополнительным вызовом является обеспечение контроля и мониторинга процессов внутри подобных систем оркестрации.

Для решения этих проблем ведутся активные исследования как в отечественной, так и в зарубежной научной среде. В данной работе анализируются существующие подходы к оркестрации LLM, включая зарубежные исследования [1, 2], выявляются их достоинства и недостатки, а также предлагаются методы улучшения эффективности работы подобных систем в виде реализации системы оркестрации для решения задачи рецензирования комплексных документов.

Основная часть. Среди доступных инструментов для оркестрации LLM существуют такие решения, как LlamaIndex [3] и Prefect. LlamaIndex представляет собой фреймворк, предназначенный для интеграции языковых моделей с базами данных и API. Рассматриваемый фреймворк обеспечивает эффективное индексирование данных, ускоряя доступ к информации, а также поддерживает работу с различными источниками, включая SQL/NoSQL базы, CRM-системы и платформы общения. Prefect, в свою очередь, является инструментом для оркестрации рабочих процессов, разработанным на Python. Он позволяет автоматизировать управление задачами, эффективно обрабатывать ошибки и масштабировать вычисления, обеспечивая стабильность и надежность работы системы. Prefect дает примеры эффективного управления рабочими процессами, включая обработку ошибок, автоматическое логирование и гибкое масштабирование. Эти принципы важны для надежной оркестрации запросов к LLM.

Еще одним перспективным инструментом является ProtoLLM - фреймворк для быстрого прототипирования и интеграции больших языковых моделей, разработанный в Университете ИТМО. Рассматриваемая технология предоставляет удобные механизмы взаимодействия с LLM, поддержку различных API, а также возможность управления запросами к моделям. ProtoLLM может быть полезен как средство для унификации вызовов LLM и обработки их ответов. Его архитектурные принципы, такие как гибкость настройки конфигураций и адаптация под разные типы моделей, могут быть использованы при проектировании собственной системы оркестрации. Кроме того, реализованные инструменты для работы с внешними базами данных и API позволят улучшить взаимодействие моделей с документами и метаданными. Дополнительно

зарубежные исследования предлагают альтернативные архитектуры оркестрации, ориентированные на оптимизацию распределения нагрузки между моделями. Такие подходы включают динамическое выделение вычислительных ресурсов и гибкую настройку взаимодействия между компонентами системы, что позволяет повысить эффективность работы языковых моделей в сложных сценариях.

Выводы. Развитие методов оркестрации LLM позволяет значительно повысить эффективность работы систем, снизить вычислительные затраты и улучшить качество результатов. В рамках дальнейших исследований планируется разработка собственной архитектуры оркестрации и создание прототипа системы интеграции мультимодальных LLM для решения задачи рецензирования комплексных документов.

Разработанная система оркестрации будет включать ключевые механизмы для эффективного управления LLM и оптимизации процессов обработки сложных документов, а именно:

- Динамическое распределение запросов, позволяющее направлять задачи к наиболее подходящим моделям в зависимости от их специализации, загруженности и качества предсказаний;
- Гибкое ансамблирование моделей, объединяющее результаты нескольких LLM для повышения точности ответов;
- Адаптивное управление контекстом для корректной передачи информации между запросами;
- Модульная архитектура, которая позволит пользователям подключать дополнительные модели и адаптировать оркестрацию под специфические задачи. Встроенные инструменты мониторинга и логирования помогут отслеживать производительность моделей и контролировать качество выдаваемых результатов.

Список использованных источников:

1. Chia-Hsuan Lee, Hao Cheng, Mari Ostendorf OrchestraLLM: Efficient Orchestration of Language Models for Dialogue State Tracking // arXiv. – 2024. – №2311.09758.
2. Shengda Fan, Xin Cong, Yuepeng Fu, Zhong Zhang, Shuyan Zhang, Yuanwei Liu, Yesai Wu, Yankai Lin, Zhiyuan Liu, Maosong Sun Models WorkflowLLM: Enhancing Workflow Orchestration Capability of Large Language Models // arXiv. – 2024. – №2411.05451.
3. Zirnstein B. Extended context for InstructGPT with LlamaIndex. – 2023.

Автор _____ Иржанова Ю.И.

Научный руководитель _____ Терещенко В.В.