УДК 004.85

МЕТОДЫ АДАПТАЦИИ ЯЗЫКОВЫХ МОДЕЛЕЙ СРЕДНЕГО РАЗМЕРА ДЛЯ НАУЧНОГО РЕЦЕНЗИРОВАНИЯ

Маракулин А.А. (ИТМО), **Дедкова А.В.** (ИТМО),
Научный руководитель – Терещенко В.В.
(ИТМО)

Введение. Большие языковые модели демонстрируют высокое качество генерации, однако их использование в задачах научного рецензирования часто ограничено высокой стоимостью вычислений и сложностью масштабирования. При этом модели среднего размера (7–12 миллиардов параметров) обладают значительным потенциалом, который остается недостаточно реализованным без специализированной адаптации. Настоящая работа рассматривает подходы к повышению качества таких моделей при сохранении приемлемой ресурсоемкости, включая параметрически эффективное дообучение (PEFT) и retrieval-augmented generation (RAG) [1, 2].

Основная часть. PEFT-методы позволяют адаптировать модель под предметную область, обучая лишь небольшую долю параметров или вводя дополнительные компактные компоненты, что существенно снижает вычислительные затраты [1]. Ключевой представитель этого класса методов – LoRA, где низкоранговые матрицы добавляются к весам трансформера при заморозке базовых параметров [3]. Развитием данной линии является метод DoRA, который раздельно оптимизирует направление и масштаб весов, сохраняя низкие издержки на инференсе [4].

К аддитивным подходам также относится Prefix-tuning, при котором обучаются непрерывные «префиксы» для каждого слоя, а параметры базовой модели остаются неизменными [5]. Близкий по идее Prompt-tuning использует обучаемые подсказки (soft prompts) и демонстрирует эффективность по мере роста размеров моделей, позволяя переиспользовать одну модель с фиксированными параметрами для разных задач [6].

Дополнительное повышение обоснованности выводов достигается с помощью RAG: модель получает внешний контекст из релевантных источников и формирует ответ на основе извлеченных данных. Такой подход особенно важен для научного рецензирования, поскольку позволяет опираться на актуальные публикации и снижать риск галлюцинаций [2].

Выводы. Показано, что сочетание PEFT и RAG обеспечивает практичный баланс между качеством и ресурсными затратами для моделей среднего масштаба. PEFT-методы (например, LoRA и DoRA) повышают эффективность дообучения без полной перенастройки весов, а подходы на основе непрерывных подсказок (Prefix-/Prompt-tuning) позволяют экономить память и ускорять адаптацию. RAG расширяет фактическую базу знаний модели и повышает обоснованность экспертных оценок, что критично для задач научного рецензирования.

Список использованных источников:

1. Lialin V., Deshpande V., Yao X., Rumshisky A. Scaling Down to Scale Up: A Guide to Parameter-Efficient Fine-Tuning // arXiv preprint arXiv:2303.15647. – 2023.
2. Lewis P., Perez E., Piktus A., et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // arXiv preprint arXiv:2005.11401. – 2020.
3. Hu E. J., Shen Y., Wallis P., et al. LoRA: Low-Rank Adaptation of Large

Language Models // arXiv preprint arXiv:2106.09685. – 2021.
4. Liu S.-Y., Wang C.-Y., Yin H., et al. DoRA: Weight-Decomposed Low-Rank Adaptation // arXiv preprint arXiv:2402.09353.– 2024.
5. Li X. L., Liang P. Prefix-Tuning: Optimizing Continuous Prompts for Generation // arXiv preprint arXiv:2101.00190. – 2021.
6. Lester B., Al-Rfou R., Constant N. The Power of Scale for Parameter-Efficient Prompt Tuning // arXiv preprint arXiv:2104.08691. – 2021.