Оптимизация использования языковых моделей при ограничениях на вычислительные ресурсы в генеративных поисковых системах

Калашников Дмитрий Павлович

Аспирант

Московский Государственный Университет имени М.В. Ломоносова, Москва, Россия

E-mail: dmkalash.mail@gmail.com

Научный руководитель *— Нейчев Радослав Георгиев*

В современных генеративных поисковых системах использование языковых моделей с большим количеством обучаемых параметров обеспечивает высокое качество ответов, однако сопровождается значительными вычислительными затратами. В существующих работах по снижению стоимости инференса основное внимание уделяется архитектурным оптимизациям, каскадным схемам с несколькими языковыми моделями разного размера и эвристикам выбора модели на основе внутренних показателей уверенности [1,2,3]. При этом формализация маршрутизации как задача оптимизации пользовательской метрики при явных ресурсных ограничениях исследована недостаточно.

В промышленной поисковой системе критерием эффективности является успешность удовлетворения информационной потребности пользователя [4], определяемая по логам взаимодействия, а не формальные характеристики текста. Одновременно система функционирует при ограниченном вычислительном бюджете и строгих требованиях к времени отклика. Это приводит к необходимости построения управляемой стратегии распределения запросов между языковыми моделями различной мощности.

Пусть q — пользовательский запрос, D — набор извлечённых документов, $m \in \{m_s, m_l\}$ — малая и большая языковые модели. Обозначим через $Y(q, D, m)$ показатель успешности удовлетворения информационной потребности пользователя при использовании модели m. Задачу маршрутизации можно сформулировать как задачу максимизации ожидаемой успешности при ограничении на вычислительные ресурсы:

$$\max_{\pi} \ \mathbb{E}\big[Y(q, D, \pi(q, D))\big], \qquad \text{при условии} \qquad GPU(\pi) \le B, \quad (1)$$

где $\pi(q, D)$ — стратегия выбора модели, $GPU(\pi)$ — ожидаемое потребление вычислительных ресурсов, B — допустимый бюджет.

В результате работы показано, что при наличии калиброванной оценки ожидаемой успешности запроса оптимальная стратегия может быть реализована в виде пороговой политики по прогнозируемому значению успешности. Порог задаёт управляемую точку функционирования системы на кривой «успешность–вычислительные затраты» и позволяет контролировать долю запросов, направляемых к менее ресурсоёмкой модели.

Экспериментальная проверка проведена на данных промышленной поисковой системы. Показано, что перераспределение порядка 30% запросов к малой модели приводит к снижению суммарного потребления GPU примерно на 40% при отсутствии статистически значимой деградации пользовательской онлайн-метрики и неизменении времени отклика.

Полученные результаты демонстрируют, что маршрутизация языковых моделей может рассматриваться как формально поставленная задача ресурсно-ограниченной оптимизации пользовательской метрики, обеспечивающая управляемый баланс между качеством и вычислительными затратами в высоконагруженных генеративных системах.

Литература

1. Chen L., Zaharia M., Zou J. FrugalGPT: How to use large language models while reducing cost and improving performance // Transactions on Machine Learning Research. 2024.
2. Jiang Y., Fu F., Zhao W., Rabanser S., Lane N. D., Yuan B. Cascadia: a cascade serving system for large language models // arXiv:2506.04203. 2025.
3. Ding D., Mallick A., Wang C., Sim R., Mukherjee S., Ruehle V., Lakshmanan L. V. S., Awadallah A. Hybrid LLM: cost-efficient and quality-aware query routing // Proceedings of the International Conference on Learning Representations (ICLR). 2024.
4. Khabsa M., Crook A. C., Awadallah A. H., Zitouni I., Anastasakos T., Williams K. Learning to account for good abandonment in search success metrics // Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM). 2016.