УДК 0004.891.2

АРХИТЕКТУРА ВЫСОКОПРОИЗВОДИТЕЛЬНОГО ПОИСКА КАНДИДАТОВ В РЕЛЯЦИОННЫХ БАЗАХ ДАННЫХ С ПОМОЩЬЮ LLM НА CPU

Титоренко А.А. (ИТМО), **Костенко К.Д.** (ИТМО), **Сармусокова А.С.** (ИТМО)
Научный руководитель - кандидат технических наук, доцент Федоров Д.А. (ИТМО)

Введение. Инструменты подбора персонала в корпоративном контуре часто ограничены политиками обработки персональных данных и слабой инфраструктурой [1, 2]. В таких условиях подход, при котором LLM получает длинные тексты резюме и выполняет их оценку, становится неэффективным. Мы предлагаем архитектуру SQL-HR для on-prem поиска кандидатов, в которой LLM используется преимущественно для преобразования требований пользователя на естественном языке в структурированное представление запроса, а поиск, фильтрация и ранжирование выполняются средствами реляционной базы данных (БД), классическими механизмами Information Retrieval (IR) и cross-encoder [3] на ограниченном пуле результатов. Такой подход обеспечивает баланс между качеством, задержкой, объяснимостью и соблюдением требований приватности.

Основная часть. Архитектура SQL-HR реализует мультиагентный подход к поиску и первичному отбору кандидатов в реляционной базе данных в условиях ограниченных ресурсов. В отличие от распространенных подходов, где LLM читает длинные тексты резюме и оценивает кандидатов “в контексте” (иногда в multi-agent реализации и/или с RAG-подсказками [4, 5]), в SQL-HR модель используется преимущественно как компонент интерпретации и компиляции требований (LLM-as-compiler). Требования пользователя преобразуются в типизированное структурированное представление (обязательные/желательные/запрещенные условия и числовые пороги), что повышает предсказуемость и контролируемость структурного вывода по сравнению со свободной генерацией SQL [6]. Архитектура включает два агента: диалоговый агент уточняет требования и их приоритет, поисковый агент компилирует требования, как structured output, формирует набор из трех SQL-запросов:

- Ideal - завышенные требования,
- Fit - минимально достаточные требования,
- Fallback - ослабленные требования для расширения воронки.

По каждому варианту быстро извлекается пул кандидатов из БД, после чего применяется reranking через cross-encoder. Такой дизайн воспроизводит поведение человека в задаче подбора кандидатов, и, кроме того, такой метод позволяет существенно снизить нагрузку на систему.

Эффективность SQL-HR оценивается в сравнении с несколькими классами подходов: (1) внутренними варианты архитектуры; (2) аналогичной системой при использовании более крупной LLM на GPU; (3) классическим поиск без LLM: keyword-поиск с булевыми фильтрами и BM25/TF-IDF; (4) векторный retrieval по эмбеддингам; (5) Агентом с RAG, который читает резюме и выбирает кандидатов напрямую. Оценка проводится по метрикам качества ранжирования (Precision@k, Recall@k, nDCG@k) и производительности (latency, а также TTFT (time-to-first-token)). Разметка релевантности формируется посредством экспертной оценки и методики “LLM-as-a-judge” [7] с последующей валидацией на подмножестве.

Выводы. Предложена архитектура SQL-HR для on-prem поиска кандидатов по внутренним данным заказчика при ограниченном бюджете вычислений (CPU-инференс LLM). Архитектура использует LLM преимущественно как преобразователь требований на

естественном языке в структурированное представление запроса, реализуя извлечение кандидатов и финальную сортировку через другие инструменты и архитектуру.

Список использованных источников:

1. OnPrem.LLM: A Privacy-Conscious Document Intelligence Toolkit : Arun S. Maiya. — 2025. — URL: https://arxiv.org/abs/2505.07672 (дата обращения: 20.02.2026). — Текст : электронный.
2. Российская Федерация. Законы. О персональных данных : Федеральный закон № 152-ФЗ — URL: https://www.consultant.ru/document/cons_doc_LAW_61801 (дата обращения: 20.02.2026). — Текст : электронный.
3. Passage Re-ranking with BERT : R. Nogueira, K. Cho. — 2019. — URL: https://arxiv.org/abs/1901.04085 (дата обращения: 20.02.2026). — Текст : электронный.
4. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks / P. Lewis, E. Perez, A. Piktus [et al.] // Advances in Neural Information Processing Systems (NeurIPS). — 2020. — URL: https://dl.acm.org/doi/abs/10.5555/3495724.3496517
5. AI Hiring with LLMs: A Context-Aware and Explainable Multi-Agent Framework for Resume Screening: Y. Zhang, L. Wang. — 2025. — URL: https://arxiv.org/pdf/2504.02870 (дата обращения: 20.02.2026). — Текст : электронный.
6. PICARD: Parsing Incrementally for Constrained Auto-Regressive Decoding from Language Models / T. Scholak, R. Schucher, D. Bahdanau // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP). — 2021. — URL: https://aclanthology.org/2021.emnlp-main.779.pdf
7. A Survey on LLM-as-a-Judge : P. Moritz, L. Zheng, W.-L. Chiang [et al.]. — 2024. — URL: https://arxiv.org/abs/2411.15594 (дата обращения: 20.02.2026). — Текст : электронный.