УДК 004.8

ИССЛЕДОВАНИЕ МЕТОДОВ РАЗРЕШЕНИЯ ПРОТИВОРЕЧИЙ МЕЖДУ ИСТОЧНИКАМИ ДАННЫХ В RAG

Шнайдер П.А. (ИТМО), **Чернышева А.В.** (ИТМО), **Рыбинская З.В.** (ИТМО)
Научный руководитель – кандидат технических наук, доцент Хлопотов М.В.
(ИТМО)

Введение. Развитие Retrieval-Augmented Generation (RAG) значительно расширяет возможности языковых моделей, позволяя использовать внешние источники данных для генерации ответов. Однако данные из разных источников могут содержать противоречия, что приводит к неточностям, повышенной неопределенности и распространению недостоверной информации. Эта проблема особенно актуальна для образовательных и научных систем, где важно обеспечивать достоверность и объективность генерируемого контента.

Исследования в области повышения надежности RAG [1] сосредоточены на калибровке уверенности моделей, улучшении обработки неоднозначной информации и разработке механизмов оценки достоверности источников [2]. Одним из перспективных подходов является использование датасетов, таких как FEVER, для верификации фактов и оценки надежности извлекаемых данных [3]. Несмотря на достижения в интеграции внешней информации в процесс генерации, противоречия между источниками остаются важной проблемой, требующей дальнейшего изучения и разработки эффективных методов решения.

Настоящее исследование направлено на разработку методов, позволяющих анализировать и корректировать противоречивые данные в RAG, что повысит точность и интерпретируемость ответов моделей.

Основная часть. В работе рассматриваются четыре метода разрешения противоречий между источниками данных в RAG:

1. Контекстуальное ранжирование используется для определения релевантности источников данных на основе семантического соответствия запросу, надежности источника и актуальности информации. Разработан алгоритм, учитывающий динамическое обновление весов извлеченных данных, что особенно важно для областей знаний с быстро меняющимся содержанием.
2. Байесовская модель доверия позволяет корректировать вероятность достоверности источника, обновляя ее в зависимости от его исторической надежности и подтверждения информации другими источниками. Такой подход снижает влияние недостоверных данных и повышает качество генерируемых ответов.
3. Фильтрация на основе коллективного согласия, идея которой заключается в выделении групп независимых источников таким образом, чтобы информация в одном источнике подтверждала информацию в другом из этой же группы. Помимо

непротиворечивости внутри группы, важно учитывать надежность источников, образующих группу. Предложена система весового усреднения, позволяющая корректировать уровень уверенности модели в зависимости от согласованности данных. В случае значительных расхождений система снижает уверенность в ответе и указывает на возможную неоднозначность информации.

4. Дополнительная верификация с использованием языковых моделей позволяет сопоставлять извлекаемые факты с верифицированными базами знаний. В ходе работы реализована система анализа уверенности моделей T5 и RoBERTa, включающая оценку логитов, корреляцию уверенности с самооценкой и визуализацию зависимостей. После дообучения на SQuAD v2 и FEVER модели продемонстрировали снижение избыточной уверенности и улучшение точности ответов.

Выводы. Предложенные методы направлены на улучшение достоверности ответов языковых моделей в системах RAG за счет автоматической фильтрации и ранжирования источников. Ожидается, что применение данных подходов снизит уровень противоречий в извлекаемых данных, повысит точность модели и позволит адаптивно корректировать доверие к источникам.

Результаты исследования могут быть внедрены в образовательные и аналитические системы, использующие генеративные модели, а также применяться в задачах фактчекинга и автоматического обобщения информации. Дальнейшие исследования будут сосредоточены на разработке улучшенных алгоритмов ранжирования источников и интеграции пользовательской обратной связи в процесс корректировки достоверности извлекаемых данных.

Список использованных источников:

1. Lewis P., Oguz B., Rinott R., Riedel S., Stenetorp P. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // arXiv preprint arXiv:2005.11401, 2020.
2. Deng, B., Wang, W., Zhu, F., Wang, Q., Feng, F. (2024). CrAM: Credibility-Aware Attention Modification in LLMs for Combating Misinformation in RAG. arXiv preprint arXiv:2406.11497. https://doi.org/10.48550/arXiv.2406.11497
3. Thorne J., Vlachos A., Christodoulopoulos C., Mittal A. FEVER: A Large-Scale Dataset for Fact Extraction and Verification // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018.