В. В. Волкевич

МЕТОДЫ АНОНИМИЗАЦИИ ПЕРСОНАЛЬНЫХ ДАННЫХ В RAG-СИСТЕМАХ НА ОСНОВЕ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Рассматриваются подходы к защите персональных данных в RAG-системах (Retrieval-Augmented Generation) на базе больших языковых моделей. Проводится анализ угроз утечки данных на каждом этапе обработки информации. Предлагается многоуровневый метод анонимизации с использованием NER-моделей и псевдонимизации, обеспечивающий сохранение качества ответов при минимизации рисков раскрытия данных.

I. Введение

Распространение корпоративных систем на базе больших языковых моделей (LLM) ставит новые задачи в области информационной безопасности. Архитектура RAG предполагает индексацию внутренних документов организации и их динамическую подачу в контекст модели при обработке запросов [**?**]. Персональные данные (ПДн) в промпте создают риски их раскрытия через атаки на контекст. По оценкам Gartner, к 2025 г. 75% корпораций внедрят RAG-системы, что расширяет поверхность атак на ПДн.

II. Анализ угроз и методы защиты

Типичная RAG-система включает хранилище документов, векторную базу данных и языковую модель. Основные угрозы: воспроизведение ПДн из контекста и инверсия эмбеддингов для восстановления исходного текста [**?**]. Атаки prompt injection позволяют извлекать до 90% ПДн из незащищённого контекста. В таблице III приведено сравнение уровней угроз.

Таблица 1 – Уровни угроз в RAG

Угроза	LLM	RAG
Контекст	Низ.	Выс.
Инверсия	Ср.	Выс.
Принадл.	Выс.	Ср.

Предлагается метод NER-анонимизации (Named Entity Recognition) на этапе индексации: модель заменяет ПДн на токены-заглушки, а таблица соответствий хранится в защищённом хранилище. Для русскоязычных текстов используется Microsoft Presidio и модель Natasha [**?**]. На уровне запросов применяется prompt interceptor [**?**], блокирующий утечки в режиме реального времени.

III. Экспериментальная оценка

Тестирование на 500 корпоративных документах (45% содержат ПДн: ФИО, email, номера телефонов) с моделью LLaMA-2-7B (контекст 4096 токенов) показало точность NER 0,94. Снижение качества ответов (faithfulness) составило лишь 2,3%, что подтверждает практическую применимость метода.

Таблица 2 – Результаты эксперимента

Метрика	Без анон.	С анон.
P (NER)	—	0,94
Faithfulness	0,87	0,85

IV. Выводы

Предложенный многоуровневый метод NER-анонимизации обеспечивает точность идентификации ПДн 0,94 при снижении качества ответов не более чем на 2,3%. Метод совместим с фреймворками LangChain и LlamaIndex. Для критичных данных рекомендуется дополнительная дифференциальная приватность (ε=3,0). Перспективное направление — интеграция с федеративным обучением для распределённых корпоративных систем.

Список литературы

1. Lewis, P. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks [Электронный ресурс]. – Режим доступа: https://arxiv.org/abs/2005.11401. – Дата доступа: 17.04.2026
2. Gao, Y. Retrieval-Augmented Generation for Large Language Models: A Survey [Электронный ресурс]. – Режим доступа: https://arxiv.org/abs/2312.10997. – Дата доступа: 17.04.2026
3. Коробов, М. Natasha [Электронный ресурс]. – Режим доступа: https://github.com/natasha/natasha. – Дата доступа: 17.04.2026
4. Chen, Y. Hide and Seek (HaS): A Lightweight Framework for Prompt Privacy Protection [Электронный ресурс]. – Режим доступа: https://arxiv.org/abs/2309.03057. – Дата доступа: 17.04.2026

Волкевич Виталий Владимирович, магистрант кафедры интеллектуальных информационных технологий, БГУИР, vitalyvolkevich@gmail.com

Научный руководитель: Крапивин Юрий Борисович, доцент кафедры интеллектуальных информационных технологий БГУИР, кандидат технических наук, krapivin@bsuir.by