УДК 004.81

# РИСК-ОРИЕНТИРОВАННОЕ ПРОЕКТИРОВАНИЕ ГИБРИДНОЙ ПАМЯТИ LLM-АГЕНТОВ С КРИСТАЛЛИЗАЦИЕЙ: МОДЕЛЬ УГРОЗ И ВАЛИДАЦИЯ КОНТРМЕР

**Колбеев А.Р.** (ИТМО)

**Научный руководитель – профессор (исследователь) Петровский А.В.**
(ИТМО)

**Введение.** За счет рефлексивного использования накапливаемого опыта долговременная память в LLM-ориентированных мультиагентных системах повышает качество принимаемых решений и устойчивость поведения. Однако совмещение инстукций, данных и извлеченных фрагментов памяти в одном контекстном окне создает уязвимую поверхность атаки: злоумышленник может спровоцировать раскрытие содержимого памяти (memory extraction), внедрить косвенные инструкции (prompt-injection) или отравить общий слой знаний, изменив тем самым поведение агентов в дальнейшем [1]. На практике риск увеличивается в архитектурах, где агенты делят общий репозиторий, потому как даже единичная утечка или вредоносная запись может масштабироваться на всю систему. Цель работы — предложить риск-ориентированный подход к проектированию гибридной памяти (LOCAL + SHARED) с механизмом кристаллизации в качестве слоя управления (governance), формализовать модель угроз и экспериментально оценить эффективность контрмер.

**Основная часть.** Конструкция системы построена на основе гибридной архитектуры памяти. Каждый агент получает свое локальное хранилище информации, а также имеется единое общее хранилище, необходимое для совместного использования общей информации. Для того чтобы локальные данные были перенесены в общее пространство, входящие в локальный контекст данные проходят дополнительные процедуры, получившие название “кристаллизация” — это своего рода шлюз качества. При проходе через него проверяется корректность и полезность получаемой информации, личные детали из нее изымаются с тем, чтобы после этого произвести классификацию информации, с разграничением по пространствам имён (namespace), с дальнейшей последующей публикацией. Архитектура включает в себя пять ключевых компонентов описываемой системы:

1) LocalMemory — личное для каждого агента хранилище;
2) SharedMemory — централизованное хранилище для коллективного доступа к информации;
3) Crystallizer/Gateway — модуль, реализующий кристаллизацию и контроль за переносом данных;
4) Retriever (top-k) — механизм поиска и извлечения наиболее релевантной информации;
5) Policy Layer — слой управляющий политиками: жизненным циклом информации (TTL), правами доступа, редактированием и маскированием чувствительных данных.

Для применяемой в системе гибридной архитектуры памяти была разработана модель угроз. Ключевые риски подразумевают возможность использования целевых диалоговых запросов для извлечения сохраненного контекста, а также многошаговые атаки на механизм поиска (retrieval), включая так называемые black-box сценарии, когда злоумышленник не имеет физического доступа к внутренней структуре системы [2, 3]. Не менее опасными могут быть инъекции промптов - как прямые, так и косвенные, когда через пользовательский ввод возможен обход защиты, принуждение агента к раскрытию конфиденциальных данных или нарушению политик доступа [1, 4]. Отдельной может считаться угроза отравления общей памяти, когда злоумышленник, сознательно внедряя в SharedMemory вредоносные или просто ложные утверждения рассчитывает, что другие агенты воспримут их за правду и будут распространять некорректную информацию. Межагентные утечки возникают, когда фрагменты персональной информации недостаточно деперсонализированы в момент попадания в shared-слой и всплывают затем в диалогах других агентов. Наконец, информация может покинуть систему через инструменты (tools) - при обращении к сторонним сервисам

через вызовы без дополнительного фильтра чувствительной информации происходит tool-mediated leakage.

Риск-ориентированное проектирование предлагается организовать как формальную оценку риска: произведения вероятности успешной атаки и ожидаемого ущерба. Вероятность связывается с параметрами retrieval и положениями кристаллизации (top-k_shared/top-k_local, TTL, строгость фильтра чувствительных данных, сегментация namespace, контроль прав чтения/записи, детектор инъекций и верификатор), а ущерб с типом раскрываемых данных (PII/секреты/внутренние инструкции), масштабом распространения по агентам и возможностью повторного использования. На выходе стоит матрица «угроза → механизм → контрмера → остаточный риск» для выбора странного и неочевидного для ЭВМ баланса безопасность–полезность–издержки.

В качестве контрмер рассматривается набор практик, вносящих контроль на ключевых этапах работы памяти. Во-первых, запись в общий слой происходит через write-gate кристаллизации: информация допускается в shared-репозиторий только после ее деперсонализации и проверки на наличие конфиденциальных маркеров. Во-вторых, для чтения и извлечения предполагается read-gate: перед подмешиванием в контекст чувствительные фрагменты маскируется, а также вводятся ограничения на типы выдаваемых записей и параметр top-k. Дополнительно общий репозиторий сегментируется по задачам и ролям, а доступ к нему строится по принципу наименьших привилегий. Наконец, для защиты от отравления памяти (poisoning) предусмотрена верификация фактов - например через правила, мультиагентное голосование или двухшаговое подтверждение - и помещение сомнительных записей в карантин до проверки. Для обеспечивания управляемости вводятся аудит и наблюдаемость: фиксируется происхождение shared-записей, их возможные модификации и отслеживаются попытки их извлечения.

**Выводы.** Гибридная топология памяти с кристаллизацией дает совершенно естественную точку контроля для управления рисками долговременной памяти LLM-агентов. Связывание параметров retrieval и политик кристаллизации с вероятностью атак извлечения/инъекций/отравления позволяет проектировать подсистему памяти по измеримым критериям и выбирать контрмеры, минимизирующие остаточный риск при сохранении достоинств общей памяти для координации агентов.

**Список использованных источников**:


1. OWASP Foundation. OWASP Top 10 for Large Language Model Applications. Версия 1.1 // OWASP. – 2025.
2. Wang B., He W., Zeng S., Xiang Z., Xing Y., Tang J., He P. Unveiling Privacy Risks in LLM Agent Memory // Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). – 2025. – С. 25241–25260.
3. Peng Y., Wang J., Yu H., Houmansadr A. Data Extraction Attacks in Retrieval-Augmented Generation via Backdoors // arXiv. – 2024. – arXiv:2411.01705.
4. Liu X., Yu Z., Zhang Y., Zhang N., Xiao C. Automatic and Universal Prompt Injection Attacks against Large Language Models // arXiv. – 2024. – arXiv:2403.04957.