УДК 004.056

О ПОДХОДЕ К ПОСТРОЕНИЮ ПАМЯТИ LLM-АГЕНТА ДЛЯ ТЕСТИРОВАНИЯ НА ПРОНИКНОВЕНИЕ НА ОСНОВЕ РАЗДЕЛЬНОГО УЧЁТА ДОСТОВЕРНОСТИ И ПОЛЕЗНОСТИ ОТКЛИКОВ СРЕДЫ

Белоус А.А (ИТМО)
Научный консультант – аспирант Гаврилова В.В. (ИТМО)
Научный руководитель – кандидат технических наук, доцент Менщиков А.А. (ИТМО)

Введение. Для соблюдения требований регуляторов бизнес проводит тестирование безопасности информационных систем (пентест). Однако пентест – имитация хакерских действий – сопряжён с некоторыми ограничениями: финансовые и временные издержки, а также угрозы нарушения бизнес-процессов. Один из способов автоматизации и повышения регулярности пентеста – использование LLM-агентов. В то же время агенты применяют механизмы работы с памятью, которые опираются на предпосылки о свойствах среды, необходимых и достаточных для корректного функционирования таких механизмов. В проблемной среде пентеста эти предпосылки нарушаются, что приводит к необходимости выявления границ применимости существующих подходов и формирования требований к памяти, согласованных с реальными ограничениями проблемной среды.

Основная часть. Процесс тестирования на проникновение может быть формализован как задача принятия последовательных решений в сложной среде, что позволяет рассматривать его в рамках агентного подхода как процесс в проблемной среде. Свойства проблемной среды непосредственно влияют на выбор и обоснование решений, принимаемых агентами при взаимодействии с этой средой. Формально опишем проблемную среду в задаче пентеста по методологии Рассела-Норвига [1]. Среда – информационная система, в отношение которой производится тестирование на проникновение, и связанные с ней защитные и операционные подсистемы. Производительность среды – многокритериальная функция максимизации ожидаемого ущерба и поверхности атаки, а также минимизация числа срабатываний защитных систем и вычислительных ресурсов. Исполнительные механизмы – набор действий, доступный при использовании программного обеспечения, используемого для достижения цели пентеста. Датчики – стандартный поток ввода/вывода данных. Следуя этой методологии, проблемной среде в задаче пентеста можно дать следующие характеристики: частичная наблюдаемость, мультиагентность, недетерминированность, последовательность, динамичность, непрерывность по смене состояния во времени, дискретность по восприятию и действию агента, а также неизвестность по полноте информации, которой агент обладает в каждый момент времени.

Для рационального поведения агенту требуется внутреннее состояние, аккумулирующее результаты прошлых взаимодействий и позволяющее интерпретировать текущие наблюдения в условиях неопределённости. Как следствие свойств среды, видна необходимость всестороннего анализа механизмов памяти [2]. В настоящей работе анализируются нарушения предпосылок корректной работы типов памяти, функций памяти, а также способов управления памятью в проблемной среде задачи пентеста.

Для таких типов памяти, как скользящее контекстное окно, метод агрегирования информации, векторная память, внешняя структурная память и гибридная архитектура, предпосылки их корректной работы нарушаются свойствами частичной наблюдаемости и неизвестности среды.

Для таких функций памяти, как сохранение контекстной связности, накопление и актуализация знаний, обеспечение планирования, адаптация стратегии поведения, обработка ошибок, предпосылки их корректной работы нарушаются свойствами частичной наблюдаемости, динамичности и мультиагентности среды.

Для таких способов управления памятью, как пассивное управление, управление на основе заранее заданных правил, управление самим агентом, управление внешним модулем, предпосылки их корректной работы нарушаются свойствами частичной наблюдаемости, неизвестности и мультиагентности среды.

Основное ограничение для корректного использования существующих подходов к работе с памятью заключается в том, что в задаче пентеста агент оперирует откликами среды не как информацией, а как гипотезами о состоянии среды с определённым уровнем доверия к датчикам. В связи с этим в работе предложен подход к построению памяти как структуры, ориентированной на неизвестную и мультиагентную среду на основе муравьиного алгоритма [3]. В отличие от распространённых динамических и графовых подходов, где обновление памяти определяется общей полезностью извлечённых данных, в предлагаемом подходе разделяются и независимо обновляются два аспекта информации: достоверность и полезность откликов среды

Под достоверностью отклика понимается степень подтверждённости вывода о свойствах или состоянии среды. Достоверность не повышается из-за того, что вывод способствовал оптимизации целевой функции или часто использовался в работе агента. Полезность отклика отражает, насколько использование связанного с ним вывода оптимизировало целевую функцию.

Логика обновления полезности и достоверности сопоставима с механизмом муравьиных алгоритмов. Роль феромона в предлагаемом подходе выполняет показатель полезности, который накапливается на связях между данными и действиями, входившими в успешные цепочки атак. Роль испарения выполняет снижение полезности по мере появления более результативных альтернатив. При этом показатель достоверности не является феромоном и не обновляется по критерию производительности. Это свойство изменяется только при повторной проверке или получении новых данных, которые прямо подтверждают или противоречат ранее сделанному выводу. Гипотезы с недостаточной подтверждённостью допускаются к использованию только для планирования и выбора проверок, но не для выполнения действий. Таким образом, при каждом обращении агента к памяти, показатель достоверности определяет набор гипотез, которые можно использовать, а показатель полезности определяет, какие из них будут использоваться в первую очередь. Такое разделение совместимо с проблемами в неизвестной и мультиагентной среде, где успешность отдельных шагов не гарантирует истинности выводов о состоянии системы.

Выводы. Формально описана проблемная среда в задачах тестирования на проникновения в предметной области агентных систем. Определены свойства среды, нарушающие предпосылки корректного функционирования механизмов работы LLM-агентов с памятью. Предложен подход к построению памяти на основе муравьиного алгоритма, которая отражает степень уверенности агента в гипотезах, описывающих состояние проблемной среды в задаче тестирования на проникновение.

Список использованных источников:

1. Рассел С. Искусственный интеллект. Современный подход. Том 1 / С. Рассел, П. Норвиг. – 4-е изд. исправ. и доп. – Москва : Вильямс, 2021. – 704 с. – ISBN 978-5-907365-25-4.
2. Zhang Z. et al. A survey on the memory mechanism of large language model-based agents // ACM Transactions on Information Systems. – 2025. – Т. 43. – №. 6. – С. 1-47.
3. Dorigo M., Di Caro G., Gambardella L. M. Ant algorithms for discrete optimization // Artificial life. – 1999. – Т. 5. – №. 2. – С. 137-172.