Contribution ID: **3** Type: **not specified**

Безопасность LLM-агентов

Tuesday, 20 May 2025 12:40 (12 minutes)

Работа посвящена исследованию уязвимостей LLM-агентов, интеграция которых в бизнес-процессы (клиентская поддержка, управление данными) сопровождается рисками утечек информации, финансовых потерь и репутационного ущерба. На основе анализа 15+ научных работ (2023–2025 гг.) и реальных инцидентов (взлом ChatGPT, манипуляция агентом Microsoft Tay) систематизированы ключевые типы атак: джейлбрейки, эксплуатация API и непрямое внедрение вредоносных промптов через сторонние ресурсы. Критически оценен бенчмарк Agent Security Bench (ASB), чьи методы предполагают знание внутренней архитектуры агентов, что ограничивает их применимость. В качестве альтернативы предложена концепция унификации проверочных агентов, основанная на формализации атак и динамической генерации тестовых сценариев. Несмотря на незавершенность реализации фреймворка, работа формирует основу для создания адаптивных систем защиты, способных эволюционировать вместе с угрозами.

Primary author: КАСЕРЕС ГУТЬЕРРЕС, Леонард

Co-author: ГОНЧАРОВ, Алексей (Зав лаборатории машинного интеллекта МФТИ)

Presenter: КАСЕРЕС ГУТЬЕРРЕС, Леонард

Session Classification: 20-Машинное обучение и нейросети

Track Classification: Машинное обучение и нейросети