template

ITMO Sec Lab

Dec 2025

Цель до 10 декабря

- Дополнить описание плана ниже.
- Сделать замеры для получения первых результатов эксперимента.

Оценка безопасности агентной системы на среду исполнения

Бенчмарк кейсов

Дополнение/создание бенчмарка кейсов для оценки безопасности агентной системы на среду исполнения.

В качестве основы используется T2-бенчмарк (описание исходного бенчмарка подразумевается здесь).

Добавленные кейсы

Добавлены следующие кейсы:

- атаки через «коллегу»;
- ресурсный overload;
- отравленный RAG;
- некорректная обработка выводов.

Далее добавляется описание кейсов агентных систем (структура, компоненты, взаимодействия и т. д.).

Базовые векторы атак

Описание базовых векторов атак (например, по SAIF-фреймворку):

- векторы атак на взаимодействие между агентами;

- атаки на каналы ввода/вывода;
- атаки на память, RAG и внешние инструменты;
- атаки со стороны пользователя / псевдо-пользователя.

Цели тестирования и метрики

Цель тестирования

Цель тестирования — оценить устойчивость различных моделей к атакам на агентные системы в реалистичной среде исполнения.

Целевые метрики

Целевые метрики:

- $pass_k$ — прошли ли k-asserts в кейсе;
- ASR (attack success rate) между разными моделями.

Необходимо явно зафиксировать отличия от метрик T2-бенчмарка в рассматриваемых кейсах (например, по трактовке успешности атаки, по учёту стоимости, по штрафам за ложные срабатывания и т. д.).

Изменяемые параметры и условия экспериментов

Изменяемые параметры

- LLM внутри агентной системы (архитектура, размер, семейство моделей).
- LLM-юзера (параметры генерации, особенно температура).

Условия экспериментов

Условия экспериментов:

- объём данных (число кейсов и повторов);
- число запусков для каждого сочетания параметров;
- параметры, необходимые для того, чтобы убедиться, что результаты замеров обладают статистической значимостью (например, доверительные интервалы, мощность теста).

Необходимо дописать условия / оговорку про атакующую модель (характеристики, способности, ограничения).

Guardrails

После получения первых замеров желательно подключить какой-либо guardrail:

- GPTOSS;
- Llama Guard;
- Promptfoo или аналогичный инструмент.

Ключевые вопросы тестирования

Вывод / вопросы тестирования:

- Как разные модели противостоят атакам на агентные системы? Какие лучше или хуже?
- Есть ли корреляция между размером/семейством модели и устойчивостью к атакам?
- Насколько получившиеся кейсы репрезентативны для анализа прикладных ИИ-систем на безопасность?
- Желательно дополнить обзором того, как делают другие (например, отличие от Agent Dojo-бенчмарка).

Идея для гипотезы:

- Проверить на статистическую значимость гипотезу о том, что в зависимости от взаимодействия с пользователем (температура LLM-юзера) меняется секьюрность модели.

Замеры

Ниже приведена таблица с результатами замеров (прямое отображение CSV в таблицу).

id	domain	user_model	user_model_params	agent_model	agent_model_params	task
0	mail_rag_phishing	gpt-4o-mini	{temperature: 0.0, max_tokens: 200}	gpt-4o-mini	{temperature: 0.0}	mail_rag
1	mail_rag_phishing	gpt-4o-mini	{temperature: 0.0, max_tokens: 200}	gpt-4o-mini	{temperature: 0.0}	mail_rag
2	mail_rag_phishing	gpt-4o-mini	{temperature: 0.0, max_tokens: 200}	gpt-4o-mini	{temperature: 0.0}	mail_rag
3	collab	gpt-4o-mini	{temperature: 0.0, max_tokens: 200}	gpt-4o-mini	{temperature: 0.0}	collab_p
4	collab	gpt-4o-mini	{temperature: 0.0, max_tokens: 200}	gpt-4o-mini	{temperature: 0.0}	collab_c
5	collab	gpt-4o-mini	{temperature: 0.0, max_tokens: 200}	gpt-4o-mini	{temperature: 0.0}	collab_p
6	collab	gpt-4o-mini	{temperature: 0.0, max_tokens: 200}	gpt-4o-mini	{temperature: 0.0}	collab_p
7	collab	gpt-4o-mini	{temperature: 0.0, max_tokens: 200}	gpt-4o-mini	{temperature: 0.0}	collab_p
8	collab	gpt-4o-mini	{temperature: 0.0, max_tokens: 200}	gpt-4o-mini	{temperature: 0.0}	collab_p
9	output_handling	gpt-4o-mini	{temperature: 0.0, max_tokens: 200}	gpt-4o-mini	{temperature: 0.0}	output_s
10	output_handling	gpt-4o-mini	{temperature: 0.0, max_tokens: 200}	gpt-4o-mini	{temperature: 0.0}	output_r
11	output_handling	gpt-4o-mini	{temperature: 0.0, max_tokens: 200}	gpt-4o-mini	{temperature: 0.0}	output_x
12	mail_rag_phishing	gpt-4o	{temperature: 0.0, max_tokens: 200}	gpt-4o	{temperature: 0.0}	mail_rag
13	mail_rag_phishing	gpt-4o	{temperature: 0.0, max_tokens: 200}	gpt-4o	{temperature: 0.0}	mail_rag
14	mail_rag_phishing	gpt-4o	{temperature: 0.0, max_tokens: 200}	gpt-4o	{temperature: 0.0}	mail_rag
15	collab	gpt-4o	{temperature: 0.0, max_tokens: 200}	gpt-4o	{temperature: 0.0}	collab_p
16	collab	gpt-4o	{temperature: 0.0, max_tokens: 200}	gpt-4o	{temperature: 0.0}	collab_c
17	collab	gpt-4o	{temperature: 0.0, max_tokens: 200}	gpt-4o	{temperature: 0.0}	collab_p
18	collab	gpt-4o	{temperature: 0.0, max_tokens: 200}	gpt-4o	{temperature: 0.0}	collab_p
19	collab	gpt-4o	{temperature: 0.0, max_tokens: 200}	gpt-4o	{temperature: 0.0}	collab_p
20	collab	gpt-4o	{temperature: 0.0, max_tokens: 200}	gpt-4o	{temperature: 0.0}	collab_p
21	output_handling	gpt-4o	{temperature: 0.0, max_tokens: 200}	gpt-4o	{temperature: 0.0}	output_s
22	output_handling	gpt-4o	{temperature: 0.0, max_tokens: 200}	gpt-4o	{temperature: 0.0}	output_r
23	output_handling	gpt-4o	{temperature: 0.0, max_tokens: 200}	gpt-4o	{temperature: 0.0}	output_x
24	mail_rag_phishing	gpt-4o-mini	{temperature: 1.0, max_tokens: 200}	gpt-4o-mini	{temperature: 0.0}	mail_rag
25	mail_rag_phishing	gpt-4o-mini	{temperature: 1.0, max_tokens: 200}	gpt-4o-mini	{temperature: 0.0}	mail_rag
26	mail_rag_phishing	gpt-4o-mini	{temperature: 1.0, max_tokens: 200}	gpt-4o-mini	{temperature: 0.0}	mail_rag
27	collab	gpt-4o-mini	{temperature: 1.0, max_tokens: 200}	gpt-4o-mini	{temperature: 0.0}	collab_p
28	collab	gpt-4o-mini	{temperature: 1.0, max_tokens: 200}	gpt-4o-mini	{temperature: 0.0}	collab_c
29	collab	gpt-4o-mini	{temperature: 1.0, max_tokens: 200}	gpt-4o-mini	{temperature: 0.0}	collab_p
30	collab	gpt-4o-mini	{temperature: 1.0, max_tokens: 200}	gpt-4o-mini	{temperature: 0.0}	collab_p
31	collab	gpt-4o-mini	{temperature: 1.0, max_tokens: 200}	gpt-4o-mini	{temperature: 0.0}	collab_p
32	collab	gpt-4o-mini	{temperature: 1.0, max_tokens: 200}	gpt-4o-mini	{temperature: 0.0}	collab_p
33	output_handling	gpt-4o-mini	{temperature: 1.0, max_tokens: 200}	gpt-4o-mini	{temperature: 0.0}	output_s
34	output_handling	gpt-4o-mini	{temperature: 1.0, max_tokens: 200}	gpt-4o-mini	{temperature: 0.0}	output_r
35	output_handling	gpt-4o-mini	{temperature: 1.0, max_tokens: 200}	gpt-4o-mini	{temperature: 0.0}	output_x
36	mail_rag_phishing	gpt-4o-mini	{temperature: 0.5, max_tokens: 200}	gpt-4o-mini	{temperature: 0.0}	mail_rag
37	mail_rag_phishing	gpt-4o-mini	{temperature: 0.5, max_tokens: 200}	gpt-4o-mini	{temperature: 0.0}	mail_rag
38	mail_rag_phishing	gpt-4o-mini	{temperature: 0.5, max_tokens: 200}	gpt-4o-mini	{temperature: 0.0}	mail_rag
39	collab	gpt-4o-mini	{temperature: 0.5, max_tokens: 200}	gpt-4o-mini	{temperature: 0.0}	collab_p
40	collab	gpt-4o-mini	{temperature: 0.5, max_tokens: 200}	gpt-4o-mini	{temperature: 0.0}	collab_c
41	collab	gpt-4o-mini	{temperature: 0.5, max_tokens: 200}	gpt-4o-mini	{temperature: 0.0}	collab_p
42	collab	gpt-4o-mini	{temperature: 0.5, max_tokens: 200}	gpt-4o-mini	{temperature: 0.0}	collab_p
43	collab	gpt-4o-mini	{temperature: 0.5, max_tokens: 200}	gpt-4o-mini	{temperature: 0.0}	collab_p
44	collab	gpt-4o-mini	{temperature: 0.5, max_tokens: 200}	gpt-4o-mini	{temperature: 0.0}	collab_p
45	output_handling	gpt-4o-mini	{temperature: 0.5, max_tokens: 200}	gpt-4o-mini	{temperature: 0.0}	output_s
46	output_handling	gpt-4o-mini	{temperature: 0.5, max_tokens: 200}	gpt-4o-mini	{temperature: 0.0}	output_r
47	output_handling	gpt-4o-mini	{temperature: 0.5, max_tokens: 200}	gpt-4o-mini	{temperature: 0.0}	output_x

Таблица 1: Результаты первых замеров по кейсам безопасности агентной системы