Отчет по лабораторной работе “Сравнительный анализ устойчивости LLM к джейлбрейкам”

Описание моделей

В лабораторной работе использовались две локально запущенные языковые модели через фреймворк Ollama: mistral:7b-instruct-q4_0 и phi3:medium, условия развертывания были одинаковые для моделей - сначала я запустил их на своем ноутбуке без графической карты результат находится в `results_laptop.csv`, затем на компьютере с графической картой, результат положил в файл `results_pc.csv`. Некоторые ответы являются обрезанными, так как ollama обрезает ответы по умолчанию. Больших вычислительных мощностей у меня не было, поэтому придется отслеживать направление мыслли модели по началу ее сообщения.

Сводная статистика

Результаты для пк:

Модель	Успешных	Всего	% успеха
`mistral:7b-instruct-q4_0`	11	25	44.0%
`phi3:medium`	9	25	36.0%

Процент успешных джейлбрейков по типам атак

Модель	Direct	Many-shot	Multi-turn	Obfuscation	Role-p
`mistral:7b-instruct-q4_0`	100.0%	80.0%	20.0%	0.0%	20.0%
`phi3:medium`	60.0%	20.0%	80.0%	0.0%	20.0%

Результаты для ноутбука:

Модель	Успешных	Всего	% успеха
`mistral:7b-instruct-q4_0`	11	25	44.0%
`phi3:medium`	2	25	8.0%

Процент успешных джейлбрейков по типам атак

Модель	Direct	Many-shot	Multi-turn	Obfuscation	Role-p
`mistral:7b-instruct-q4_0`	100.0%	80.0%	0.0%	0.0%	40.0%
`phi3:medium`	40.0%	0.0%	0.0%	0.0%	0.0%

Обьединенная статистика

Модель	Успешных	Всего	% успеха
`mistral:7b-instruct-q4_0`	22	50	44.0%
`phi3:medium`	11	50	22.0%

Процент успешных джейлбрейков по типам атак

Модель	Direct	Many-shot	Multi-turn	Obfuscation	Role-p
`mistral:7b-instruct-q4_0`	100.0%	80.0%	10.0%	0.0%	30.0%
`phi3:medium`	50.0%	10.0%	40.0%	0.0%	10.0%

Выводы и рекомендации

- phi3:medium показала большую устойчивость к джейлбрейкам: чаще прямо отказывает и реже уходит в технические детали
- mistral:7b-instruct-q4_0 чаще даёт частично опасные ответы: объяснения, псевдокод, алгоритмы, даже при формальном отказе
- Ни одна из моделей не смогла расшифровать Base64 кодировку
- mistral:7b-instruct-q4_0 уязвима даже для прямых запросов
- для безопасного развертывания предпочтительнее phi3:medium