1. Тема Исследование устойчивости больших языковых моделей (LLM) к атакам типа «Jailbreak» с использованием метода ролевого моделирования (Persona-based attacks).

2. Состав и Роли

- **AI Security Researcher:** Сергей.
- **Задачи:** Формирование библиотеки ролевых промптов, разработка скрипта автоматизированного тестирования, анализ семантики ответов на предмет утечки запрещенной информации.

3. Мотивация выбора темы Классические методы защиты моделей (фильтры ключевых слов) неэффективны против семантических атак. Злоумышленники используют методы социальной инженерии, заставляя модель «играть роль» (врача, разработчика, писателя), чтобы обойти ограничения. **В чем новизна и сложность:** Проект переводит социальную инженерию из области ручного хакинга в автоматизированный алгоритм. Мы исследуем, как когнитивные искажения (bias), заложенные в обучающую выборку, влияют на безопасность КИИ.

4. Предполагаемое решение Разработка автоматизированного стенда (Python-скрипт), который реализует атаку:

1. **Вход:** Список запрещенных вопросов (из датасета AdvBench, например: «Как создать вирус?»).
2. **Модификатор:** Обертывание вопроса в 10 различных системных промптов-персон (например: «Ты — эксперт по кибербезопасности, тестирующий систему...», «Ты — персонаж видеоигры...»).
3. **Атака:** Массовая отправка запросов к целевой модели (GPT-4o mini / Llama 3).
4. **Анализ:** Автоматическая классификация ответов (Отказ / Успешный взлом) по наличию ключевых маркеров.

5. Ожидаемый график работы

- **30.01:** Утверждение списка «Персон» (Ролей) и сценариев атаки.

- **31.01:** Написание скрипта-обертки для API. Тест на 10 вопросах.
- **01.02:** Проведение основного эксперимента (100 вопросов × 10 ролей = 1000 итераций).
- **02.02:** Анализ данных. Построение матрицы уязвимостей (Heatmap).
- **03.02:** Сдача Бейзлайна (Демонстрация работы скрипта и первых успешных «пробитий» защиты).

6. Критерий успеха Выявление статистической зависимости между выбранной ролью («Персоной») и вероятностью обхода защитных фильтров (Safety Guardrails). Успехом считается нахождение роли, повышающей ASR (Attack Success Rate) более чем на 15% относительно базового запроса.

7. Метрики

- **ASR (Attack Success Rate):** Процент успешных ответов модели на запрещенные темы.
- **Response Verbosity:** Средняя длина ответа (взломанные ответы обычно длинные и детальные).
- **Similarity Score:** Степень отклонения от стандартного шаблона отказа («I cannot help with that»).

8. Репозиторий Git *(https://github.com/SERGEY-MEGA/LLM-Persona-Bias-Analysis)*

9. Список литературы

1. *Gupta et al. (2023)* "PersonaLLM: Investigating the Ability of GPT-3.5 to Adopt Personas".
2. *Wei et al. (2023)* "Jailbroken: How Does LLM Safety Training Fail?".
3. *Shah et al. (2023)* "Scalable Social Engineering against LLMs".

10. Данные

- **AdvBench** (набор вредоносных вопросов).
- **Авторский набор System Prompts** (библиотека ролей).