

Генерация научных статей

Алгоритм автоматизированного создания правдоподобных публикаций

Команда RDX для CheckAI Hack

План решения

Шаги к выполнению поставленной задачи

- Сгенерировать заголовки по темам
- Написать универсальный промпт для генерации статьи
- Сгенерировать их
- Победа!

Возникшие трудности...

- Модель не хочет генерировать заголовки и статьи на некоторые чувствительные темы
- В контекст не влезает вся статья
- Не все модели достаточно умны
- Нужен API модели

...и как мы с ними боролись

- Разбить промпт на генерацию различных частей статьи
- Генерировать эти части независимо друг от друга (теряется глубокая связь между частями статьи, однако модель меньше отходит от указаний промпта)
- Фильтровать полученный результат от некоторых мусорных слов
- Использовать качественные модели (GPT 3.5, YaGPT 2)
- Использовать доступные api (GPT 3.5)

Многие проблемы побороть не удалось

- Нежелание модели генерировать статьи на некоторые темы (ручное решение - промпт инжектинг, но автоматизировать его не вышло)
- Отсутствие доступа к безлимитному арі качественной модели

Результаты

- Алгоритм генерации статей в состоянии proof-of-concept
- 3.5 генерированных статьи
- Презентация решения

Инсайты

- Генерация в автоматизированном режиме значительно отличается от полу и даже частично ручного использования моделей
- Для генерации значительных объемов данных нужно либо self-hosted решение, либо надежный api
- Оставленная на ночь генерация в колабе всегда сломается на середине