Table of Contents

- 1. Задание
- 2. Решение
- 3. Предложение по улучшению
    - 3.1. Проблемы, которые я вижу:
    - 3.2. Дальнейшее улучшение кода
- 4. Код решения
    - 4.0.1. Colors
    - 4.1. Create csv file
    - 4.2. Главный запрос к YandexGPT
- 5. Заключение

1. Задание

Создать отзывы клиентов о продукте и попытаться скрыть, что отзыв был сгенерирован с помощью LLM.

Вы можете использовать любой LLM.

Вы должны сделать это за 3 дня.

2. Решение

Я использовал базовые инструменты: POSIX Shell который будет работать на любой *nix системе или POSIX совместимой. curl (для запросов API) + jq (для обработки JSON).

Поскольку требования к сложности будут расти, для правильного выбора сложных инструментов лучше начать с базовых.

Два дня я потратил на доступ к API YandexGPT: ввод номера кредитной карты и номера телефона, получение идентификатора облака и т.д.

Один день я потратил на написание кода и написание данного отчета о решении, которое генерирует отзывы от клиентов.

Я создал этот CSV-файл: TODO!!!!!!!

Я добавил случайность к отзыву с подсказкой "посмотрите на этот цвет", случайный цвет добавляет эмоциональный окрас к модели и изменяет ее внутреннее состояние.

Я сделал простую попытку скрыть интеллект LLM улыбками.

Я не пытался использовать цепочку запросов.

3. Предложение по улучшению

3.1. Проблемы, которые я вижу:

- Цензура в LLM при генерации
- Проблема стохастического поведения в LLM
- Скрытие интеллекта в LLM

- Цензура при публикации
- Недостаточно случайности

Вижу три способа борьбы с цензурой:

- Самостоятельно обученные модели с уникальными характеристиками как у человека
- Цепи передовых и постоянно меняющихся техник обработки запросов для диалогов с корпоративными LLM, возможно с НС, постоянно ищущие новые уязвимости
- Генеративно-состязательные сети (GAN), которые симулируют корпоративную цензуру для "отпора/первого удара"

Для скрытия интеллекта и добавления случайности, думаю, должно быть достаточно тщательного создания запроса.

Проблема стохастического поведения – сложная проблема, связанная с ограничениями текущих LLM и оборудования.

3.2. Дальнейшее улучшение кода

1. Нужно быть более тщательным с обработкой ошибок
2. Следует сравнить с другими решениями подобного рода
3. Скрытие интеллекта должно быть гораздо более тщательным
4. Необходимо добавить больше случайности с предварительным сбором данных, например, профессии пользователей, пола и т. д.
5. Цепь запросов должна быть протестирована и рассмотрена.

4. Код решения

4.0.1. Colors

```
1. Ализариновый
2. Антрацитовый
3. Баклажан
4. Берилловый
5. Бирюзовый
6. Бланжевый
7. Бронзовый
8. Бургундия
9. Виридиан
10. Гранатовый
11. Гридеперлевый
12. Грушевый
13. Деним
14. Жжёного апельсина
15. Индиго
16. Какао
17. Карри
18. Кофейный
19. Кремовый
20. Лазурный
21. Лаймовый
22. Лавандовый
23. Лиловый
24. Маджента
25. Маковый
26. Маренго
27. Насыщенный синий
28. Оливковый
29. Охра
30. Палевый
31. Пюсовый
```

```
32. Пыльная роза
33. Ржавый
34. Сапфировый
35. Серебристый
36. Сизый
37. Сиреневый
38. Слоновая кость
39. Тауповый
40. Терракотовый
41. Тиффани
42. Ультрамарин
43. Умбра
44. Фисташковый
45. Фуксия
46. Хаки
47. Цвет морской волны
48. Цвет шампанского
49. Чернильный
50. Шоколадный
```

```
get_color( ) {
    c=$(cat | grep -o " .*" | tr -d ' ')
    count=$(echo "$c" | wc -l)
    random_line_n=$(shuf -i 1-$count -n 1)
    echo "$c" | sed "$random_line_n!d"
}
get_color
```

4.1. Create csv file

```
echo "Тональность,Текст" > /tmp/dobri_vigruzka.csv
```

4.2. Главный запрос к YandexGPT

```
alias curl="proxychains -f /home/u/proxychains.conf curl 2>/dev/null"

get_color( ) {
    c=$(cat | grep -o " .*" | tr -d ' ')
    count=$(echo "$c" | wc -l)
    random_line_n=$(shuf -i 1-$count -n 1)
    echo "$c" | sed "$random_line_n!d"
}


yandexPassportOauthToken=y0_AgAAAABsj07pAATuwXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
FOLDERNAME=default

if [ -z "$IAM_TOKEN" ]; then
    IAM_TOKEN=$(curl -s -d "{\"yandexPassportOauthToken\":\"$yandexPassportOauthToken\"}"
fi
if [ -z "$CLOUD_ID" ]; then
    CLOUD_ID=$(curl -s -H "Authorization: Bearer $IAM_TOKEN" https://resource-manager.api
fi
# curl -s --request GET -H "Authorization: Bearer $IAM_TOKEN" https://resource-manager.ap
if [ -z "$FOLDER_ID" ]; then
    FOLDER_ID=$(curl -s --request GET -H "Authorization: Bearer $IAM_TOKEN" https://resou
fi
model="gpt://$FOLDER_ID/yandexgpt/latest"
STREAM=false

REACTION="понравился"
REQUEST="посмотри внимательно на эти символы:  :) :-) :-D :D XD X-D ;) - это смайлики.
Представь, что ты человек женского пола HR, страдающая сильным
слабоумием, эмоциональными всплесками и пишущая без сложных пунктуаций и кавычек, обязате
Ты вошла в магазин с вывеской $(get_color) цвета (про это забудь) и купила сок под назван
```

```
Добрый. Пришла домой и хочешь рассказать своим подписчикам программистам (про программист
который попробовала. Сок тебе "$REACTION", но тебе хочется
подбодрить людей и ты им говоришь, что довольна соком. Что бы ты написала им? Сделай от о
В конце поста добавь эти три символа: ###
Можешь от себя добавить в конце."

body=$(jq -n \
--arg model "$model" \
--argjson stream "$STREAM" \
--arg request "$REQUEST" \
'{modelUri: $model,
   completionOptions: {
    stream: $stream,
    temperature: 0.6,
    maxTokens: 2000
  },
  messages: [
    {
      role: "user",
      text: $request
    }
  ]
}')
# -- Main requiest:
r=$(curl --request POST \
  -H "Content-Type: application/json" \
  -H "Authorization: Bearer $IAM_TOKEN" \
  -H "x-folder-id: $FOLDER_ID" \
  -d "$body" \
  "https://llm.api.cloud.yandex.net/foundationModels/v1/completion" 2>/dev/null)

# -- Error handeling
if [ $? != 0 ]; then
    echo ERROR!
fi
if [ "$(echo "$r" | jq -M 'has("error")')" = "true" ]; then
    echo $r | jq -M
    echo ERROR!
    echo $r | jq ".message" | fold -s -w 120
else
    # -- parsing:
    answer=$( echo "$r" | jq -r '.result.alternatives.[] | select(.status | endswith("FIN
    # -- Saving:
    echo "$REACTION"','"\"$answer\"" >> /tmp/dobri_vigruzka.csv
fi
```

5. Заключение

- Успешно получил доступ к Yandex GPT.
- Сгенерировал CSV-файл с отзывами.
- Добавил случайность к отзывам с использованием простой рандомизации запроса.
- Сделал простую попытку скрыть интеллект LLM смайликами.
- Добавил предложение по решению проблем/вызовов и дальнейшему улучшению кода.

Не пробовал цепочку запросов.

Created: 2024-03-03 Sun 08:03
Validate