Table of Contents

1. Task

Generate customer reviews for product and try to hide that review was generated by LLM.

You can use any LLM.

You must do it in 3 days.

2. Solution

Two days I spent to get access to YandexGPT API: to insert credit card number and phone number, to get cloud ID, etc.

One day I spent for coding and writing this report a solution what generates reviews from customers.

I generated this CSV file: TODO!!!!!!!

I added randomness to review with prompt "look at this color", random color adds emotional color to model and change it's inner state.

I made simple attempt to hide LLM intelligence with smiles.

I did not try chain of requests.

3. Proposal for enhancement

3.1. Problems that I see:

- LLM censorship at generating
- LLM stochastic behaviour problem
- LLM intelligence hiding
- censorship during posting
- not enough randomness

I see 3 ways to fight censorship:

- Self-hosted fine-tuned models with unique characteristics as a person
- Chains of advanced and constantly changing prompt engineering techniques to query corporate LLMs, possible with NN, constantly seeking new holes.

- Generative adversarial network (GAN) that simulate corporate censorship "to strike back/first"

To hide intelligence and add randomness a careful prompt engineering should be enough I guess.

Stochastic behaviour problem is advanced problem binded with limitations of current LLMs and hardware.

3.2. Further code enhancement

1. More careful error handling should be implemented
2. Should be compared with other solutions of such kind
3. Intelligence hiding should be much more careful
4. More randomness should be added with preceding collecting, e.q. profession of user, sex, etc.
5. Chain of queries should be tested and considered.

4. Code of solution

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. Main request to 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."

# REQUEST="посмотри внимательно на эти символы разделенные пробелами :) :-) :-D :D XD X-D
# это смайлики. Дай пример любого текста со всеми этими смайликами."

# REQUEST="Дай мне длинный список сильно отличающихся цветов по одному слову каждый."

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. conclusion

- I successfuly got access to Yandex GPT API RESTfull interface.
- I generated CSV file with reviews.
- I added randomness to reviews with simple prompt randomization.
- I made simple attempt to hide LLM intelligence with emojis/smiles.
- I added proposal to address problems/challenges and further enhance the code.

I did not attempt a chain of requests.

Created: 2024-03-02 Sat 12:58
Validate