vader-analysis

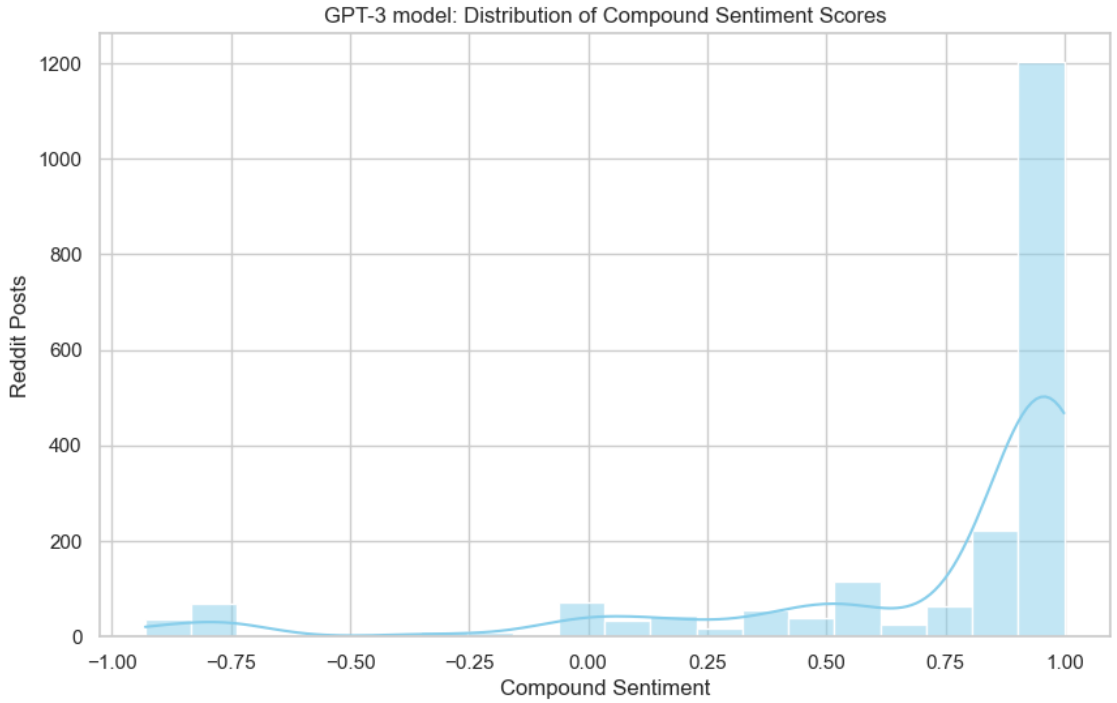
February 19, 2024

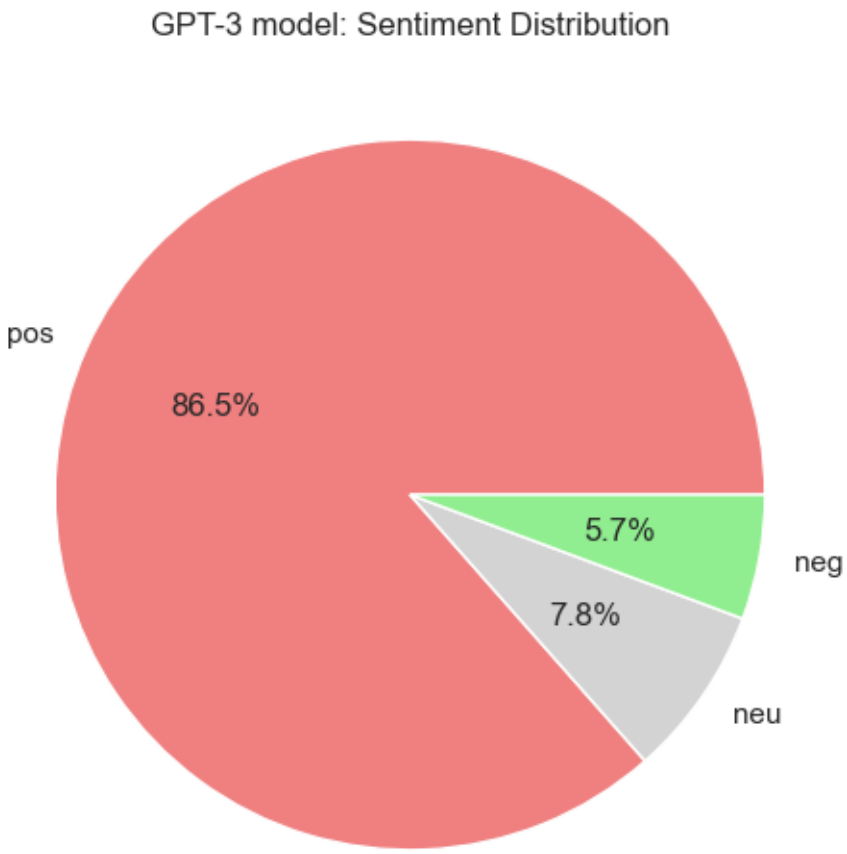
1 Project Setup

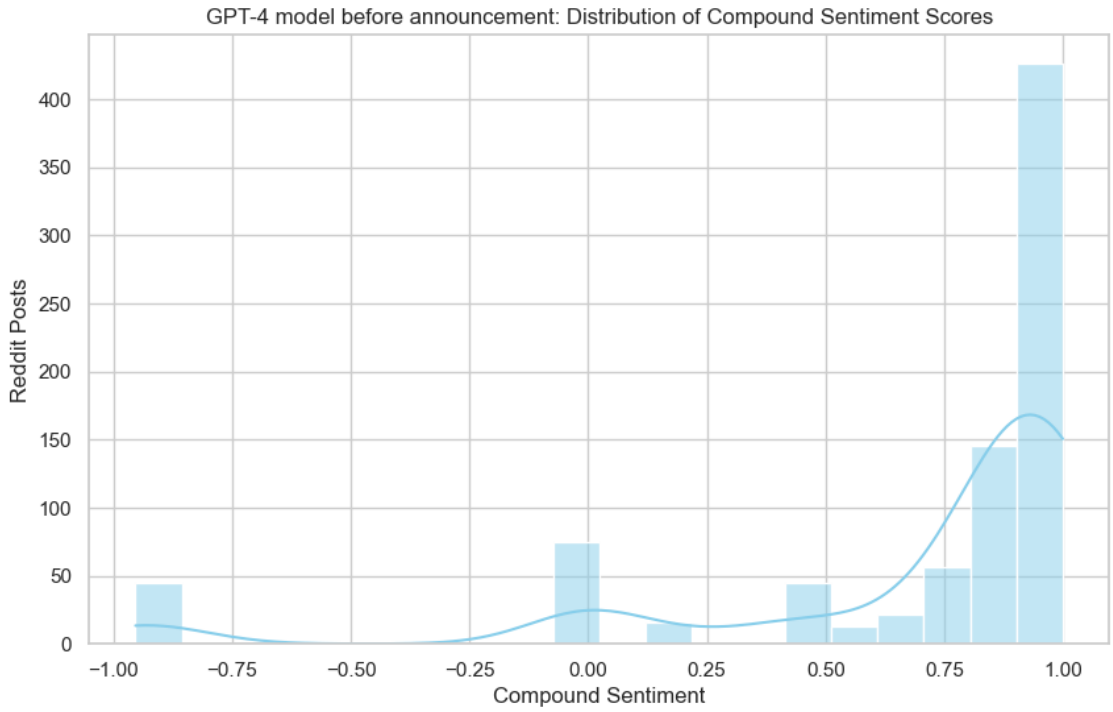
2 Reddit API data import to pandas

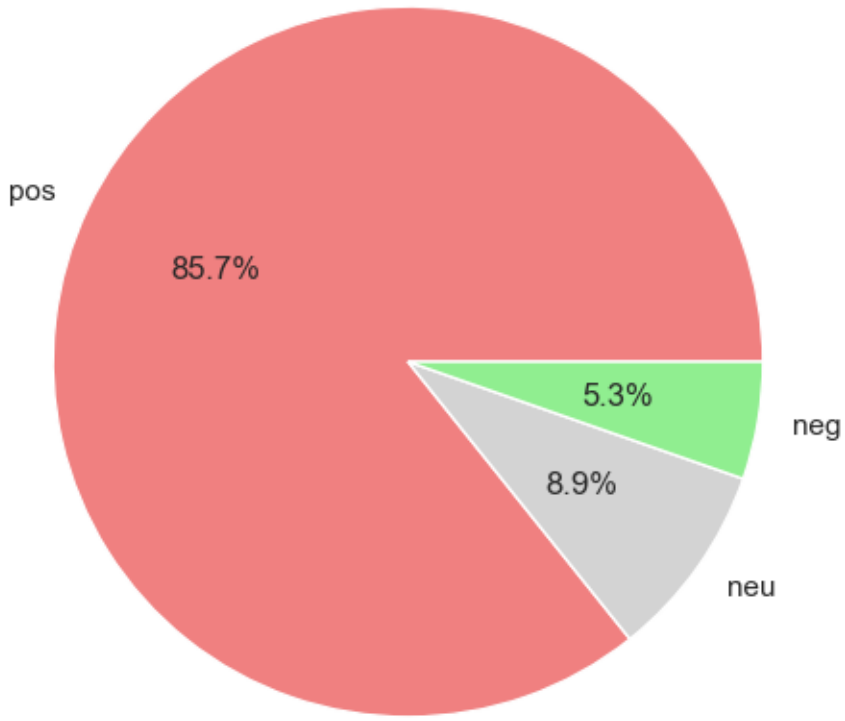
```
Subreddit Stats
+------------+----------------------+-------------+
| Unnamed: 0 |         name         | subscribers |
+------------+----------------------+-------------+
|     0      |      artificial      |    711967   |
|     1      |       chatgpt        |   4427524   |
|     2      |     deeplearning     |    148268   |
|     3      | learnmachinelearning |    383073   |
|     4      |   machinelearning    |   2868080   |
+------------+----------------------+-------------+
Total imported data: 62622
Duplicate Reddit posts: 0
```

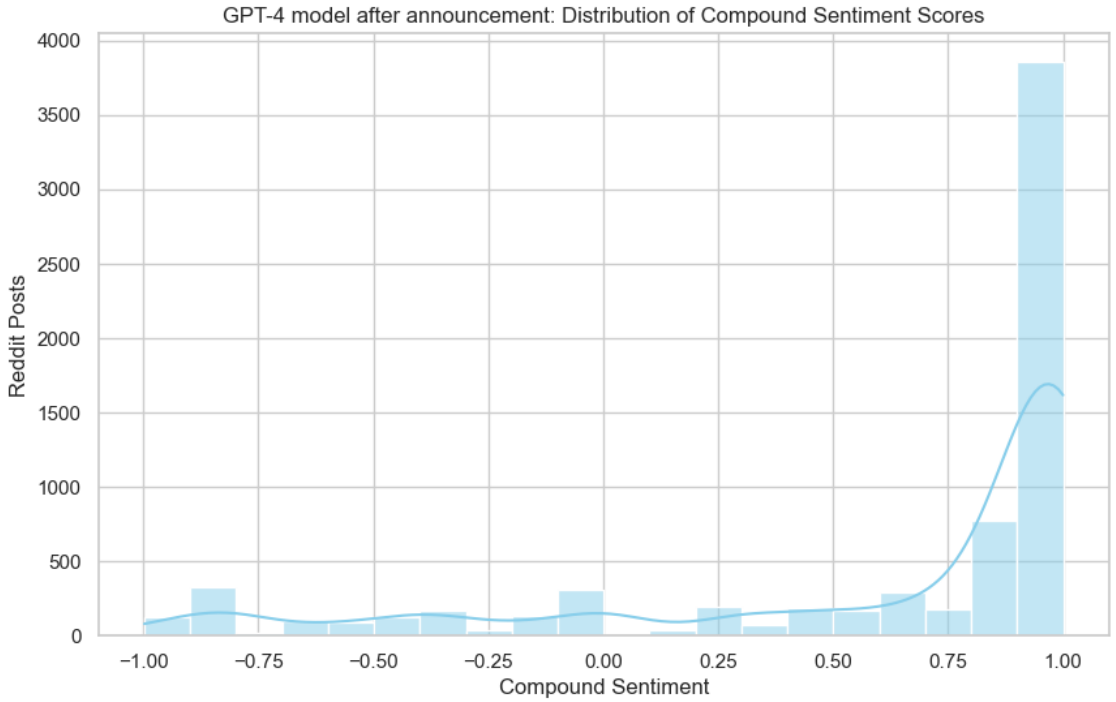
3 VADER sentiment analysis

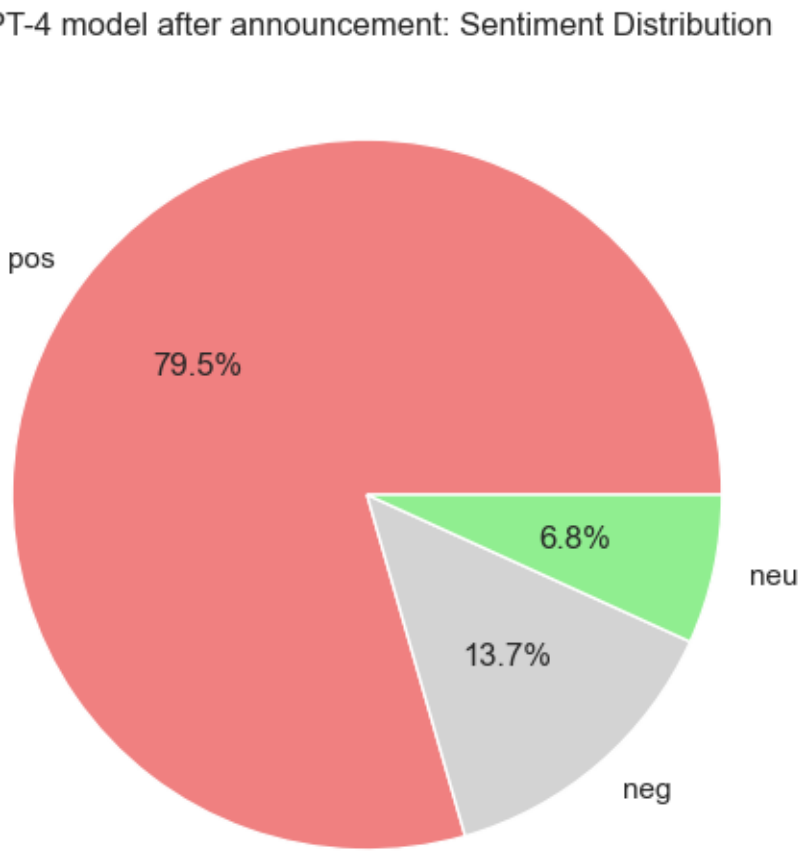



GPT-3 model: Sentiment Distribution
pos
86.5%
5.7%
neg
7.8%
neu


GPT-4 model before announcement: Distribution of Compound Sentiment Scores
Reddit Posts
0
50
100
150
200
250
300
350
400
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Compound Sentiment


GPT-4 model before announcement: Sentiment Distribution
pos
85.7%
5.3%
neg
8.9%
neu


GPT-4 model after announcement: Distribution of Compound Sentiment Scores
Reddit Posts
4000
3500
3000
2500
2000
1500
1000
500
0
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
Compound Sentiment



4 Preparing datasets, features, and targets

```
feature_names=['log_launch_distance_f', 'num_comments_weighted', 'upvote_ratio',
'log_created', 'compound', 'pos', 'neg', 'neu']
target_name=log_score_weighted
```

GPT-3 model

index	count	mean	std	min
Unnamed: 0	2010.00	1741.39	804.30	3.00
upvote_ratio	2010.00	0.89	0.13	0.33
score	2010.00	94.36	130.30	0.00
num_comments	2010.00	22.46	37.79	0.00
created	1970-01-01	2022-12-30	1970-01-23	2022-11-02
score_weighted	2010.00	0.00	0.00	0.00
num_comments_weighted	2010.00	0.00	0.00	0.00

launch_distance	2010	73 days	22 days	42 days
launch_distance_f	2010.00	73.12	22.49	42.20
pos	2010.00	0.04	0.04	0.00
neg	2010.00	0.81	0.12	0.00
neu	2010.00	0.14	0.09	0.00
compound	2010.00	0.72	0.46	-0.93
log_score_weighted	2010.00	0.00	0.00	0.00
log_num_comments	2010.00	2.27	1.39	0.00
log_launch_distance_f	2010.00	4.26	0.30	3.77
log_created	2010.00	21.24	0.00	21.23

index	25%	50%	75%	max
Unnamed: 0	1501.00	1702.00	2266.00	3471.00
upvote_ratio	0.88	0.93	0.96	1.00
score	13.00	46.00	112.00	661.00
num_comments	3.00	10.00	20.00	246.00
created	2022-12-15	2022-12-30	2023-01-19	2023-01-30
score_weighted	0.00	0.00	0.00	0.00
num_comments_weighted	0.00	0.00	0.00	0.00
launch_distance	53 days	73 days	88 days	131 days
launch_distance_f	53.67	73.95	88.30	131.35
pos	0.02	0.03	0.05	0.17
neg	0.79	0.83	0.86	1.00
neu	0.10	0.12	0.17	0.54
compound	0.64	0.93	0.99	1.00
log_score_weighted	0.00	0.00	0.00	0.00
log_num_comments	1.39	2.40	3.04	5.51
log_launch_distance_f	4.00	4.32	4.49	4.89
log_created	21.24	21.24	21.24	21.24

GPT-4 model before announcement

index	count	mean	std	min
Unnamed: 0	842.00	1881.15	673.48	339.00
upvote_ratio	842.00	0.91	0.09	0.60
score	842.00	251.52	552.22	1.00
num_comments	842.00	65.36	141.80	0.00
created	1970-01-01	2023-03-03	1970-01-07	2023-02-16
score_weighted	842.00	0.00	0.00	0.00
num_comments_weighted	842.00	0.00	0.00	0.00
launch_distance	842	10 days	6 days	0 days
launch_distance_f	842.00	10.14	6.85	0.24
pos	842.00	0.03	0.03	0.00
neg	842.00	0.87	0.07	0.65
neu	842.00	0.10	0.06	0.00

compound	842.00	0.69	0.49	-0.95	
log_score_weighted	842.00	0.00	0.00	0.00	
log_num_comments	842.00	2.62	1.82	0.00	
log_launch_distance_f	842.00	2.16	0.80	0.22	
log_created	842.00	21.24	0.00	21.24	

index	25%	50%	75%	max
Unnamed: 0	1416.00	1720.00	2342.00	3399.00
upvote_ratio	0.89	0.94	0.97	1.00
score	12.25	54.00	172.50	2700.00
num_comments	3.00	8.00	46.00	759.00
created	2023-02-27	2023-03-04	2023-03-09	2023-03-13
score_weighted	0.00	0.00	0.00	0.02
num_comments_weighted	0.00	0.00	0.00	0.01
launch_distance	4 days	9 days	14 days	25 days
launch_distance_f	4.07	9.82	14.28	25.63
pos	0.00	0.02	0.04	0.14
neg	0.83	0.86	0.91	1.00
neu	0.05	0.10	0.15	0.28
compound	0.68	0.90	0.97	1.00
log_score_weighted	0.00	0.00	0.00	0.02
log_num_comments	1.39	2.20	3.85	6.63
log_launch_distance_f	1.62	2.38	2.73	3.28
log_created	21.24	21.24	21.24	21.24

GPT-4 model after announcement

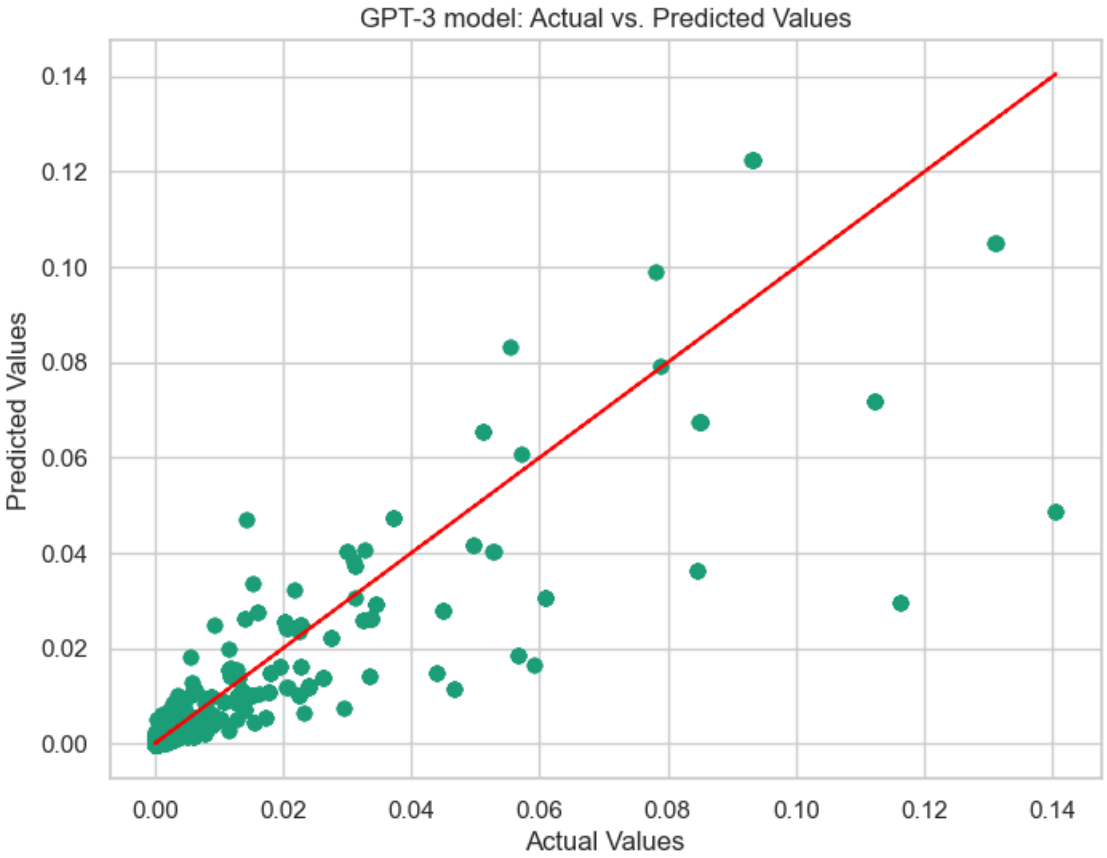
index	count	mean	std	min
Unnamed: 0	7186.00	1308.27	791.96	1.00
upvote_ratio	7186.00	0.91	0.10	0.25
score	7186.00	2590.82	4686.36	0.00
num_comments	7186.00	424.18	741.39	0.00
created	1970-01-01	2023-04-14	1970-01-18	2023-03-15
score_weighted	7186.00	0.01	0.02	0.00
num_comments_weighted	7186.00	0.00	0.00	0.00
launch_distance	7186	31 days	17 days	1 days
launch_distance_f	7186.00	31.05	17.09	1.01
pos	7186.00	0.05	0.06	0.00
neg	7186.00	0.84	0.10	0.00
neu	7186.00	0.11	0.07	0.00
compound	7186.00	0.61	0.58	-1.00
log_score_weighted	7186.00	0.01	0.02	0.00
log_num_comments	7186.00	4.18	2.35	0.00
log_launch_distance_f	7186.00	3.25	0.78	0.70
log_created	7186.00	21.24	0.00	21.24

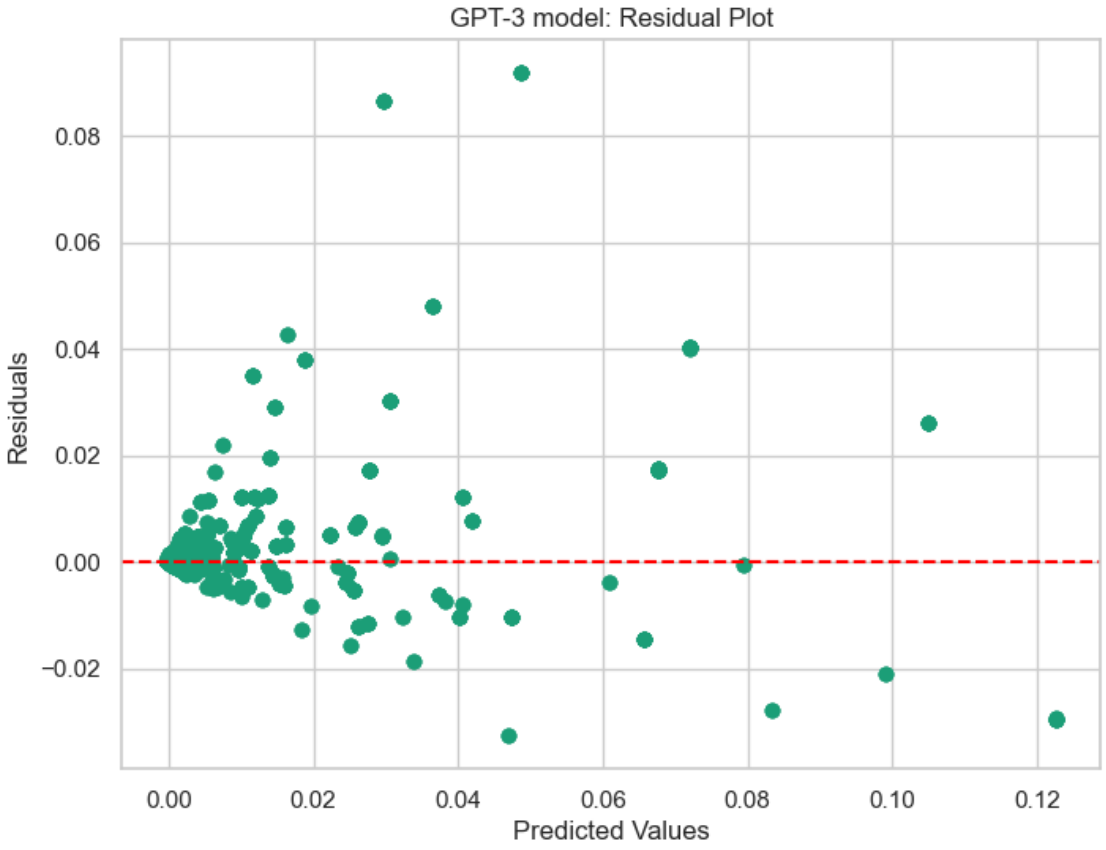
index	25%	50%	75%	max
Unnamed: 0	838.00	1170.00	1756.00	3450.00
upvote_ratio	0.88	0.94	0.98	1.00
score	26.00	186.00	3085.00	22371.00
num_comments	8.00	70.00	456.00	3835.00
created	2023-03-30	2023-04-14	2023-04-28	2023-05-14
score_weighted	0.00	0.00	0.00	0.15
num_comments_weighted	0.00	0.00	0.00	0.03
launch_distance	16 days	31 days	45 days	61 days
launch_distance_f	16.94	31.48	45.71	61.64
pos	0.01	0.04	0.06	0.56
neg	0.81	0.85	0.89	1.00
neu	0.08	0.11	0.14	0.51
compound	0.42	0.93	0.99	1.00
log_score_weighted	0.00	0.00	0.00	0.14
log_num_comments	2.20	4.26	6.12	8.25
log_launch_distance_f	2.89	3.48	3.84	4.14
log_created	21.24	21.24	21.24	21.24

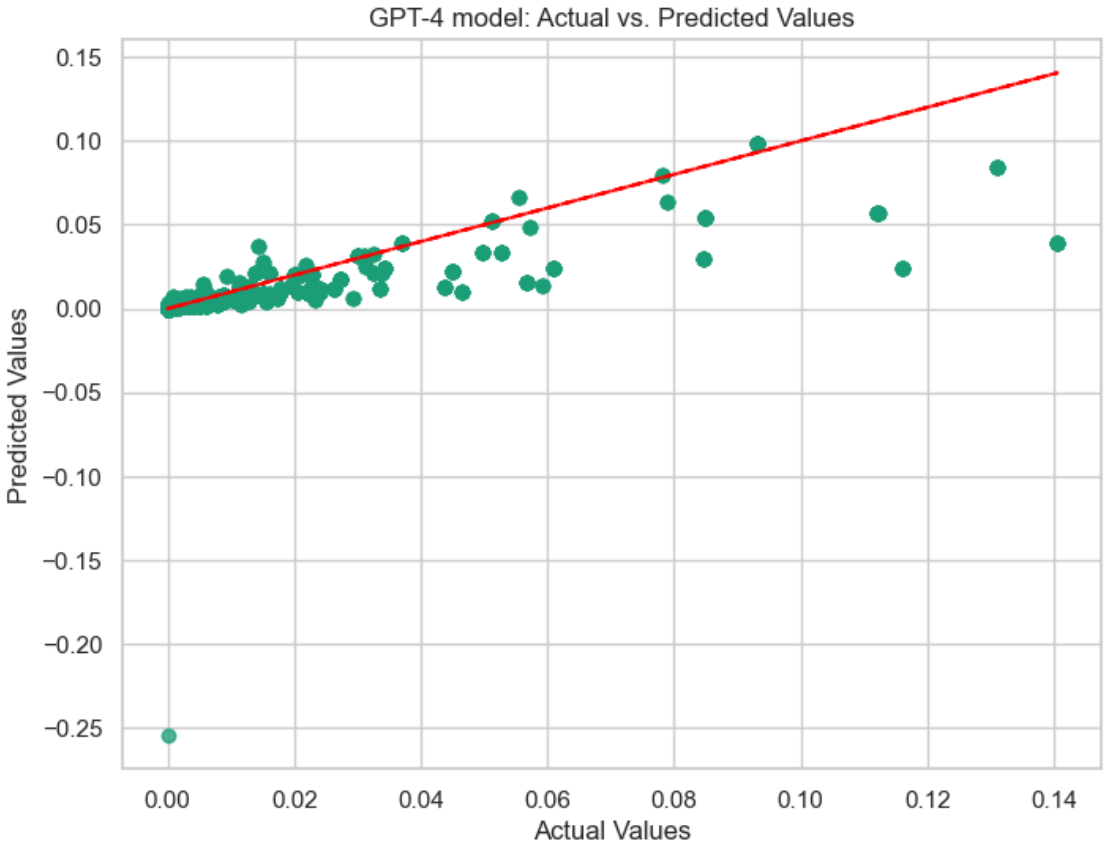
5 Run MLflow experiment

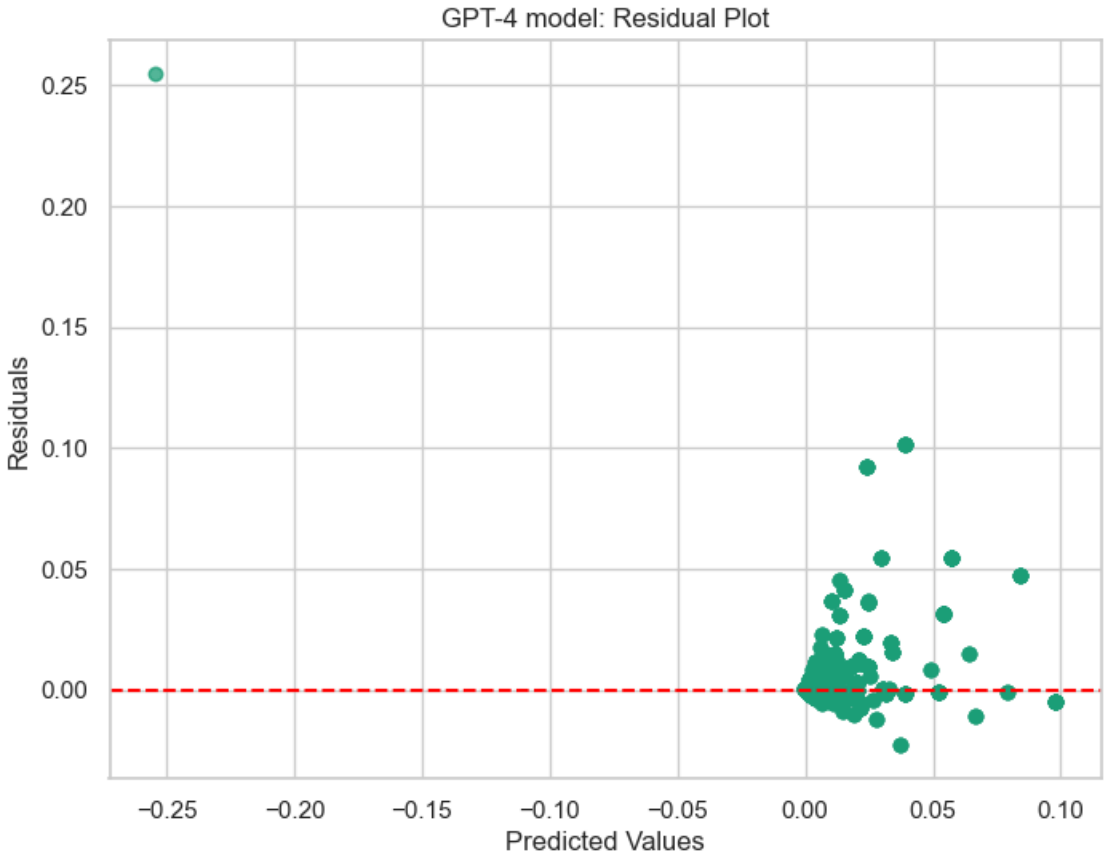
Linear regression analysis. Remove “stickied” Reddit posts from data.

Test model against data after GPT-4 launch.

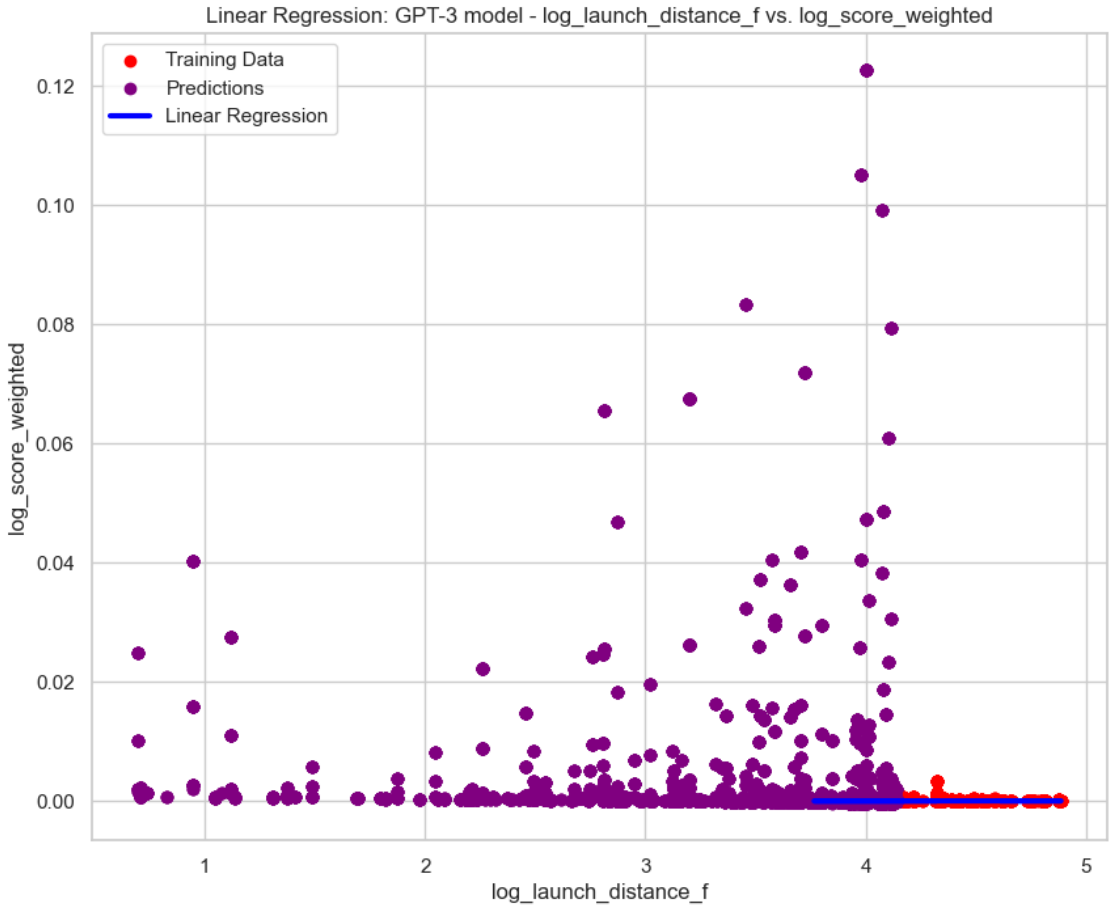

GPT-3 model: Actual vs. Predicted Values
Predicted Values
Actual Values
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14


GPT-3 model: Residual Plot
Residuals
0.08
0.06
0.04
0.02
0.00
−0.02
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Predicted Values

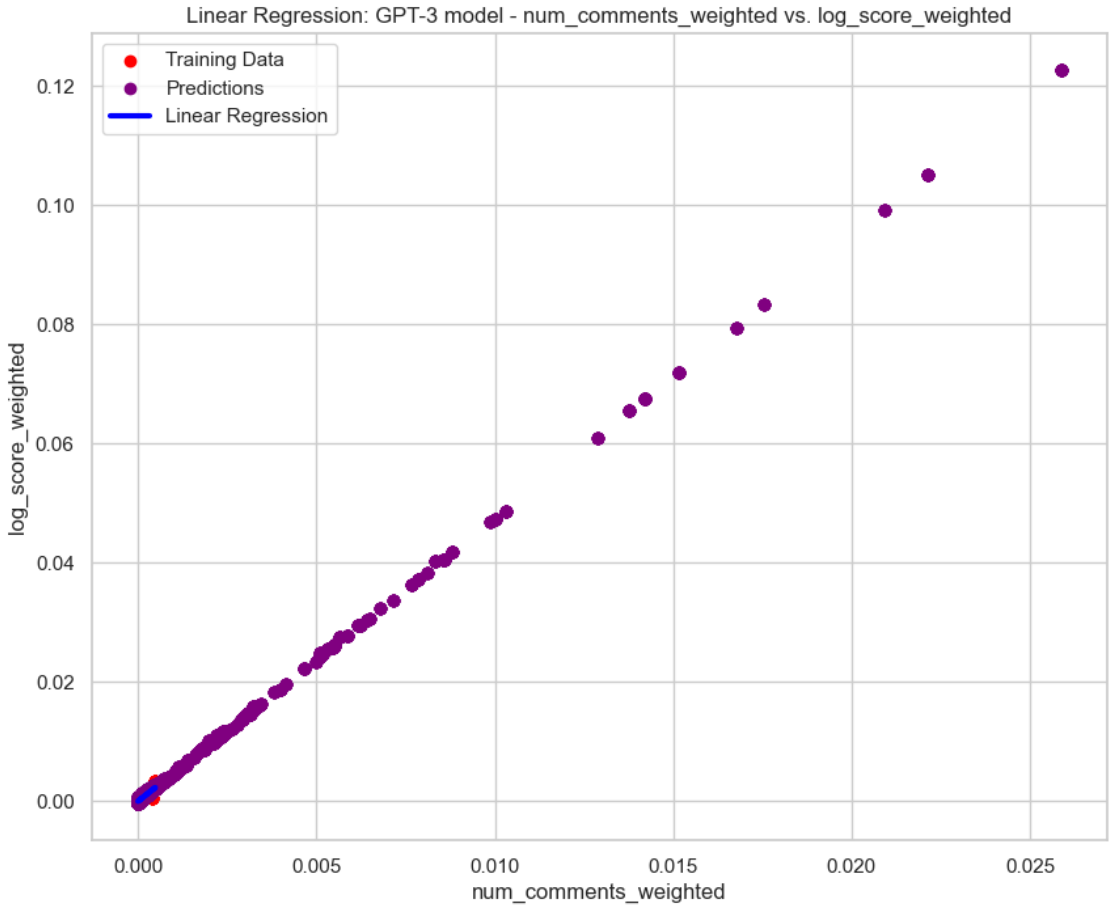

GPT-4 model: Actual vs. Predicted Values
Predicted Values
Actual Values
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
−0.25
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14


GPT-4 model: Residual Plot
Residuals
0.25
0.20
0.15
0.10
0.05
0.00
−0.25
−0.20
−0.15
−0.10
−0.05
0.00
0.05
0.10
Predicted Values

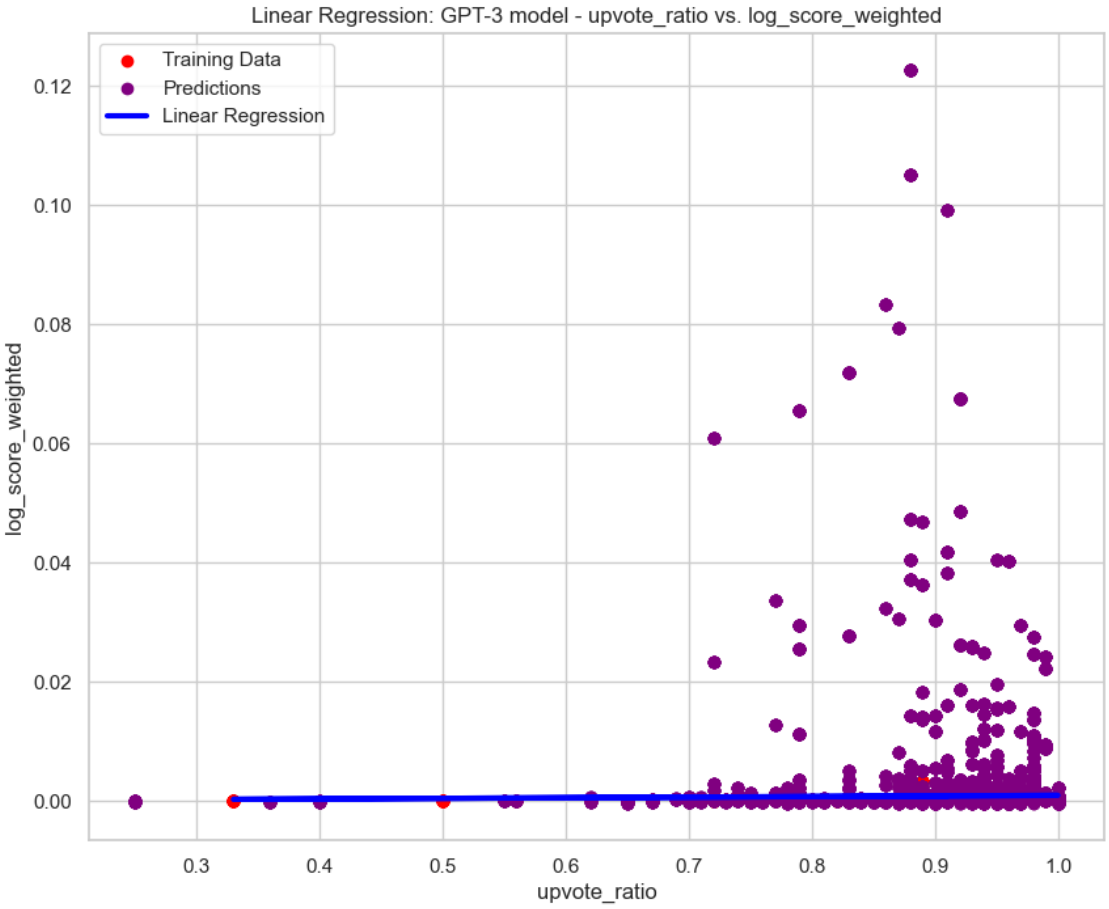
6 Feature Plots



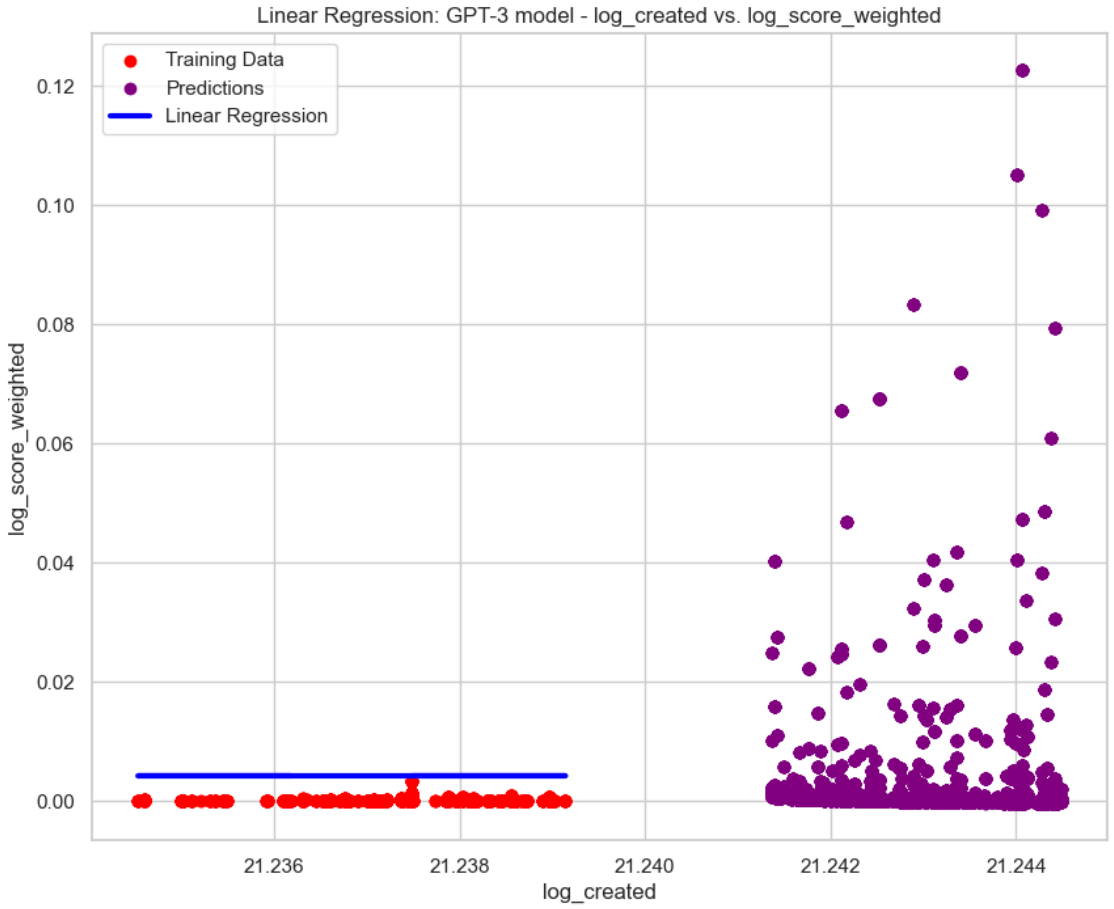
log_launch_distance_f=1.4401203089229048e-05



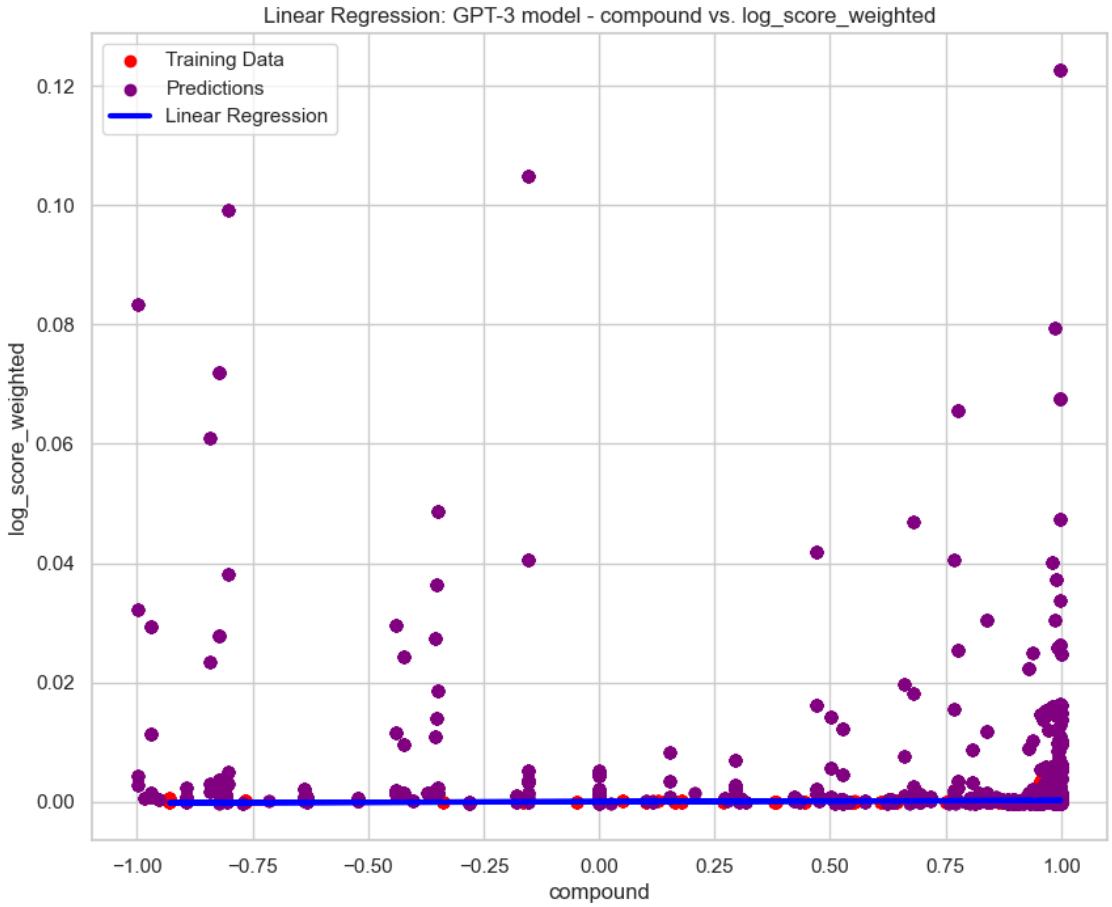
num_comments_weighted=4.7497015333104216



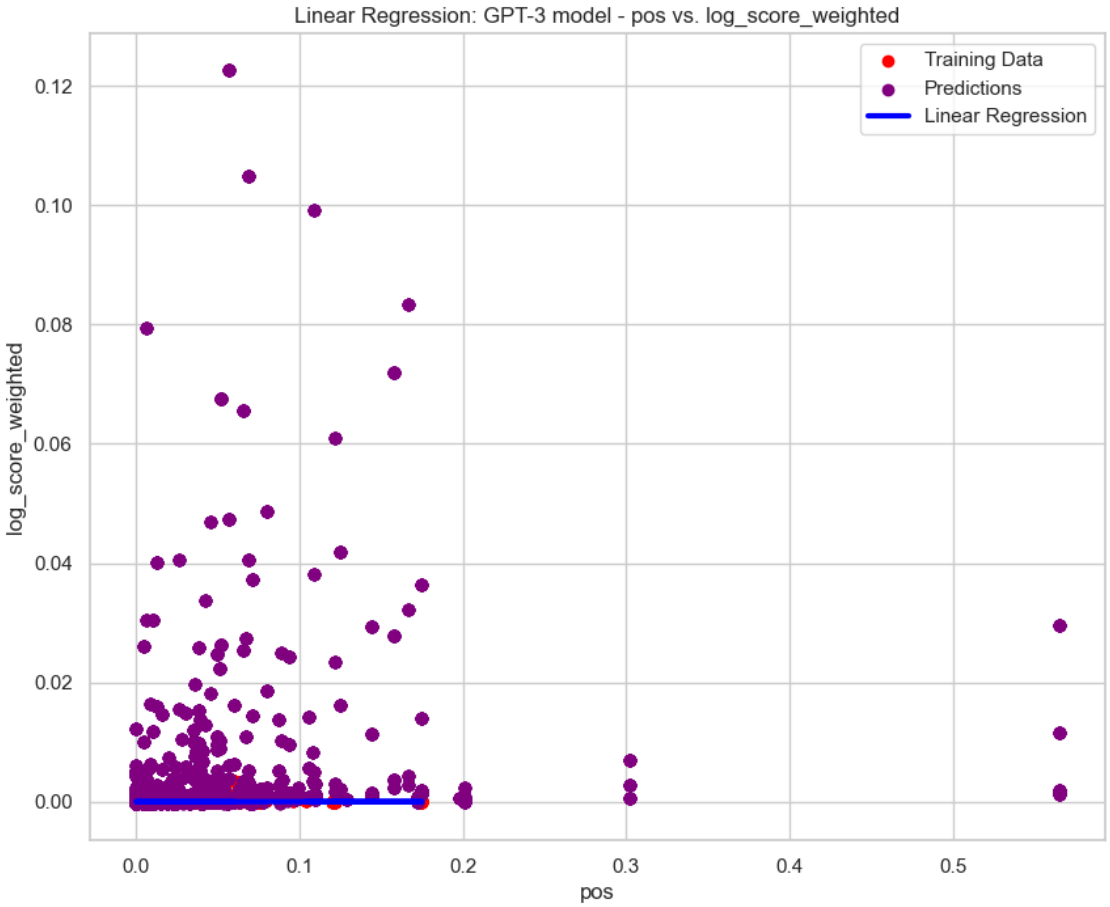
upvote_ratio=0.0009905129629084186



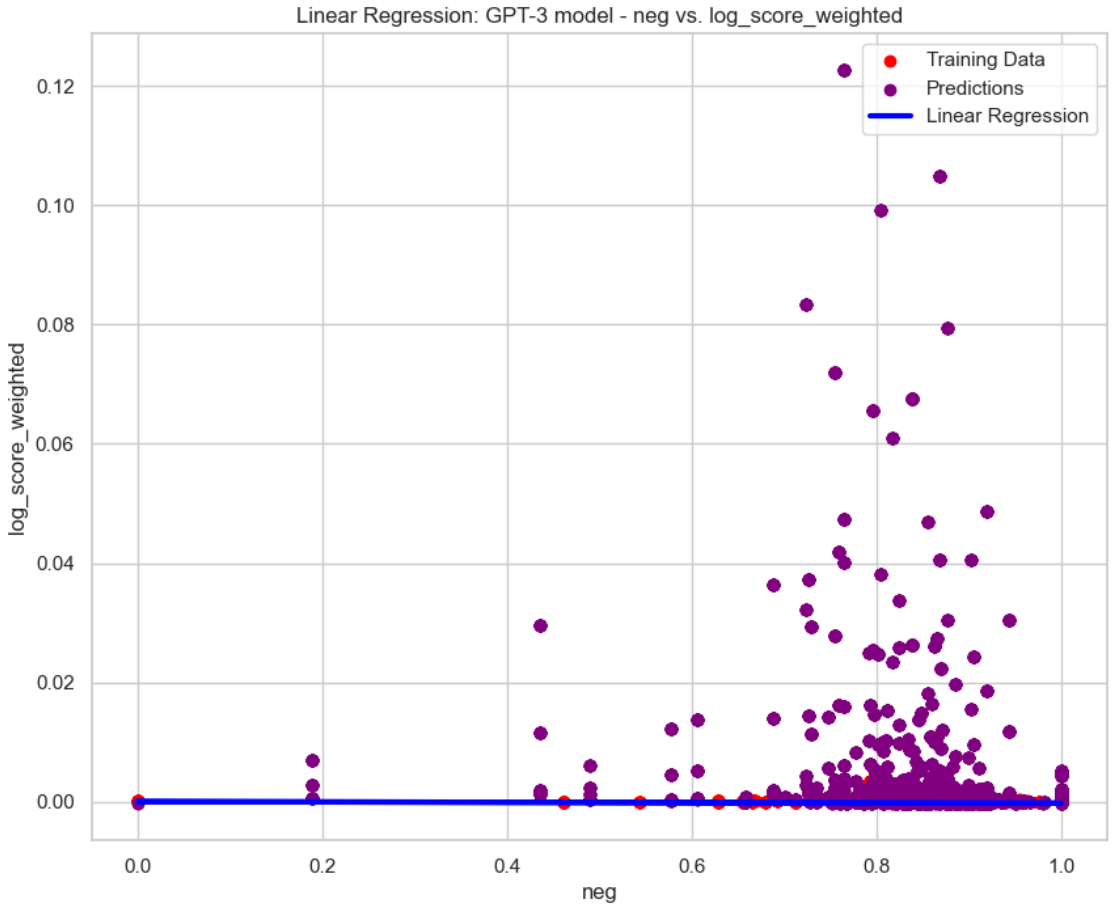
log_created=0.0002010309518536424



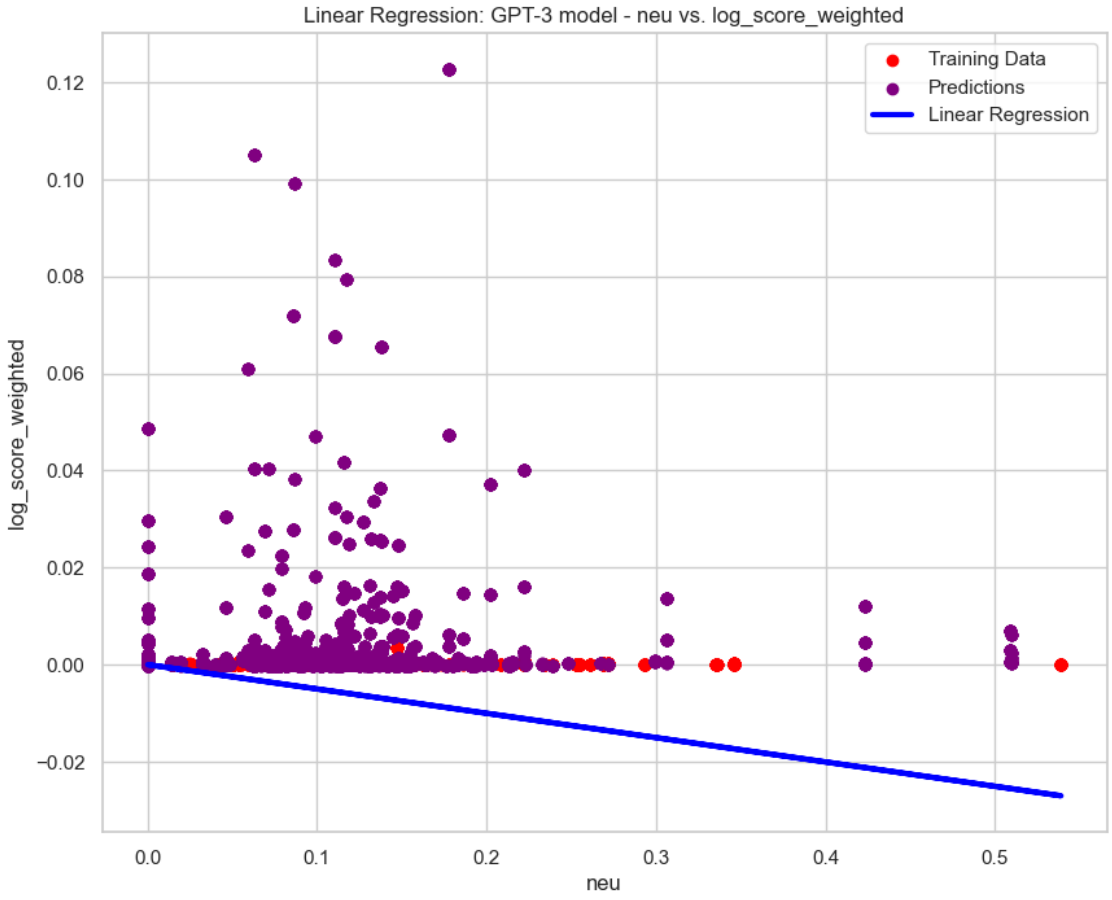
`compound=0.00020754435114023195`



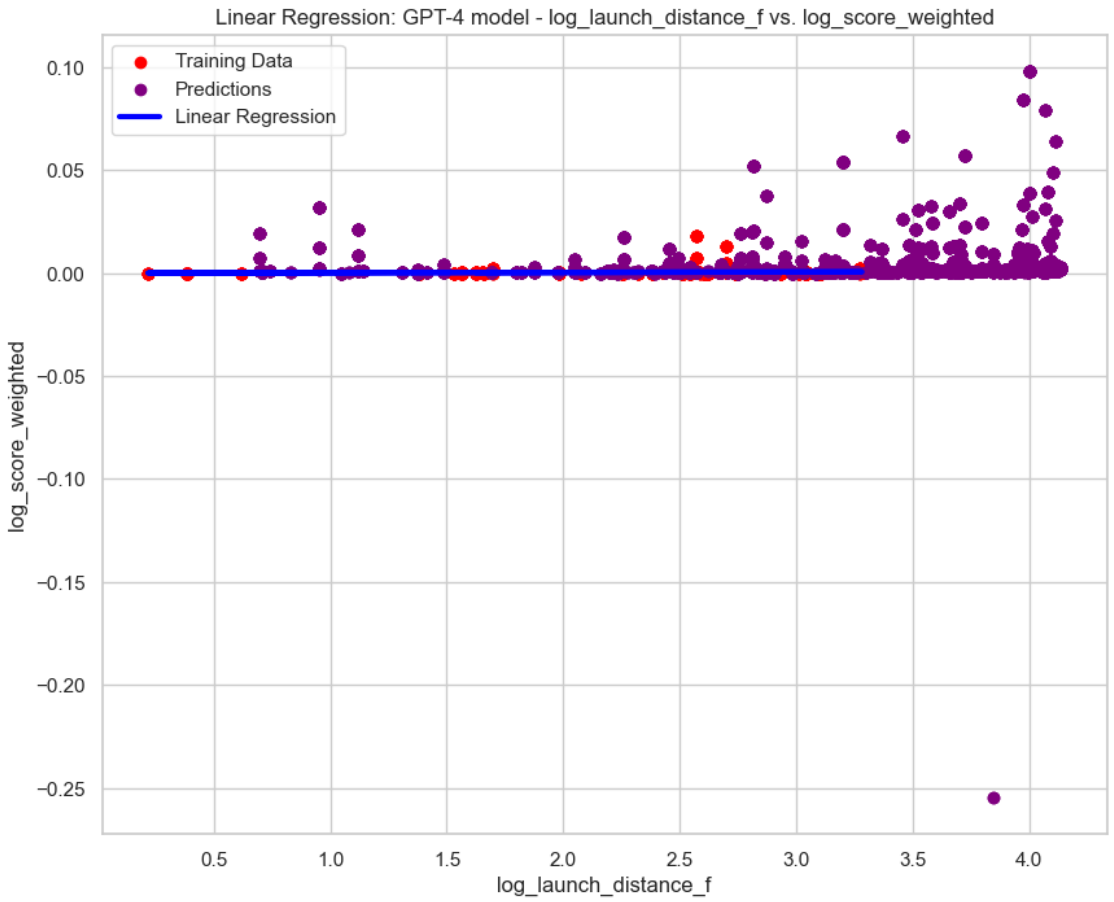
pos=1.897776746151436e-05



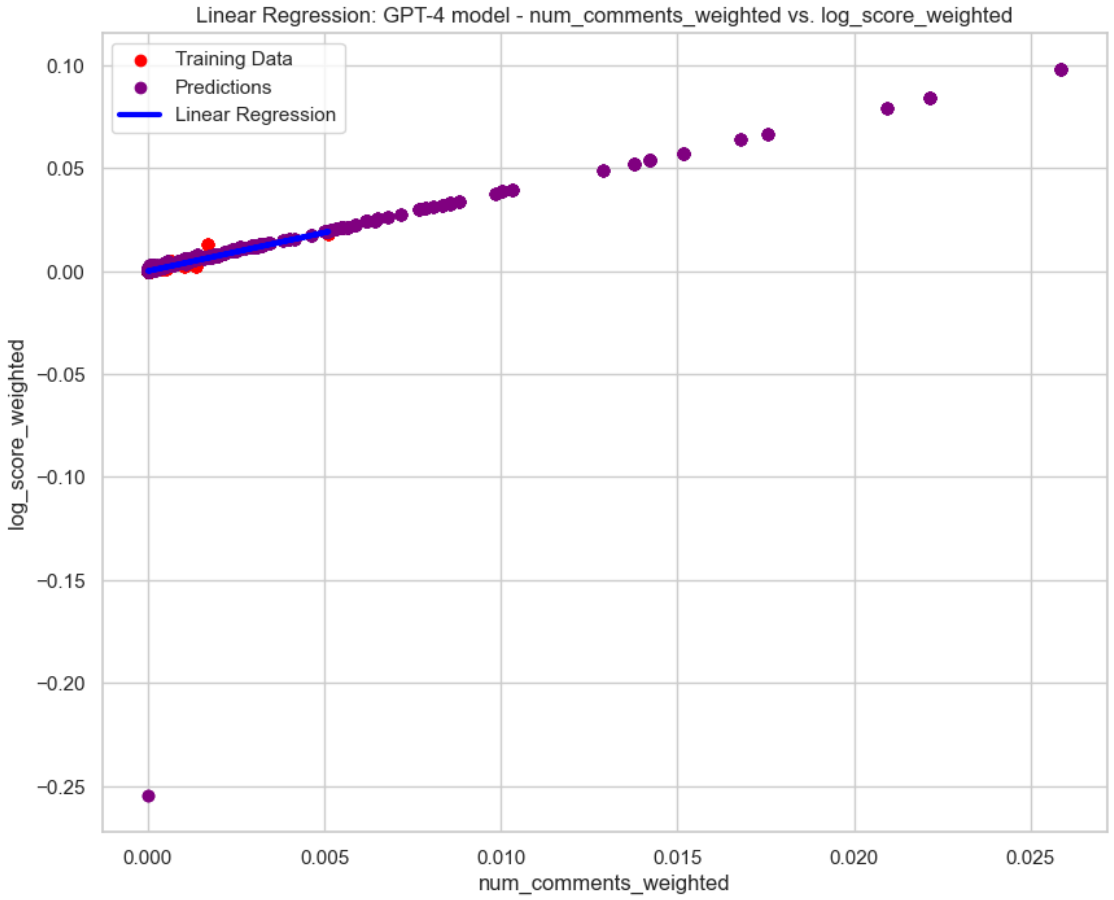
neg=-0.00025694722769253886



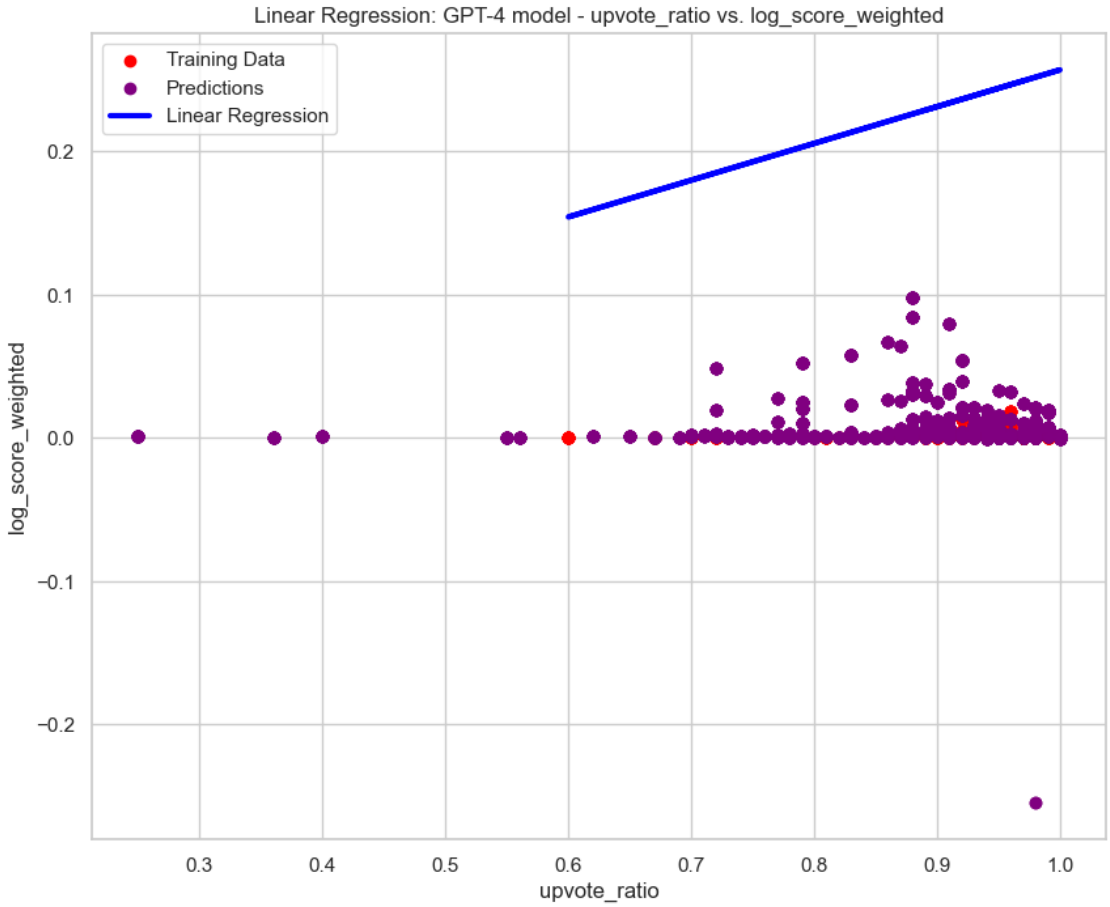
neu=-0.050143916791737624



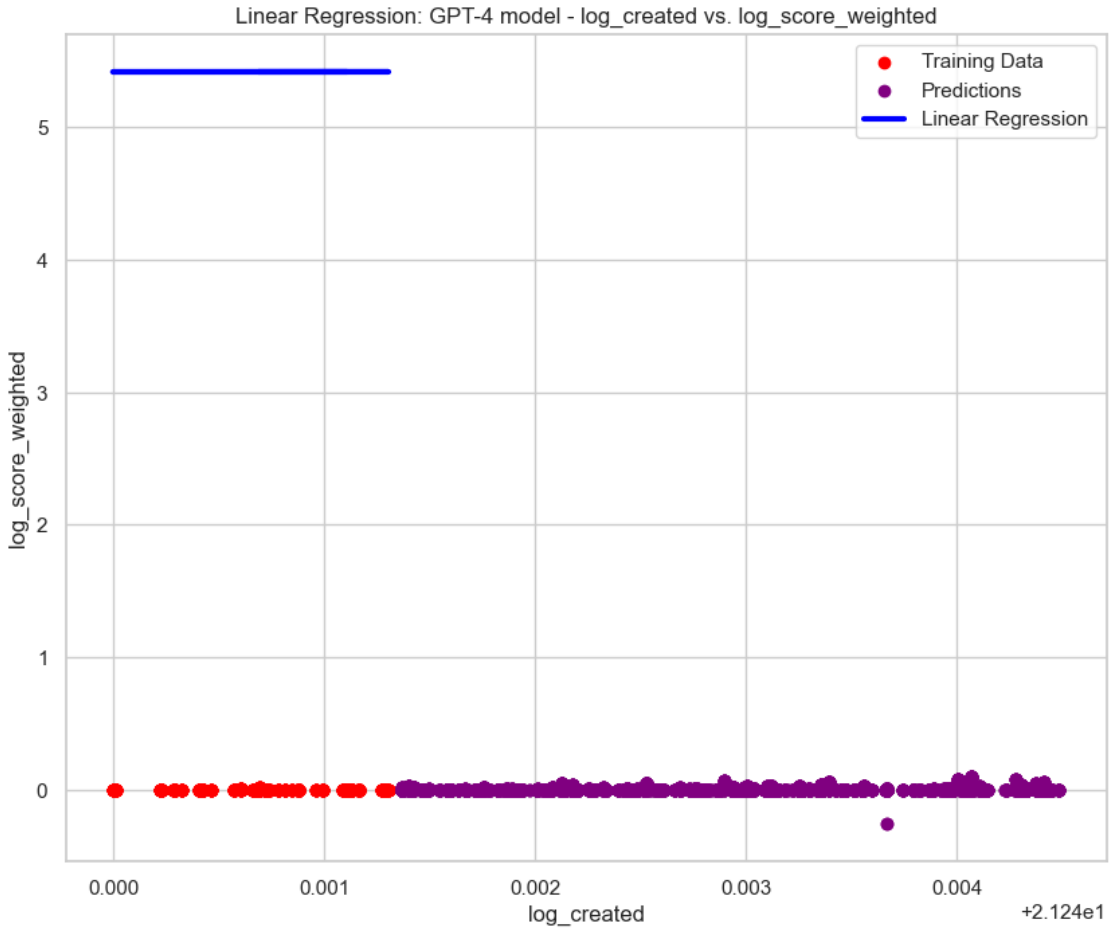
log_launch_distance_f=0.0001688318320620477



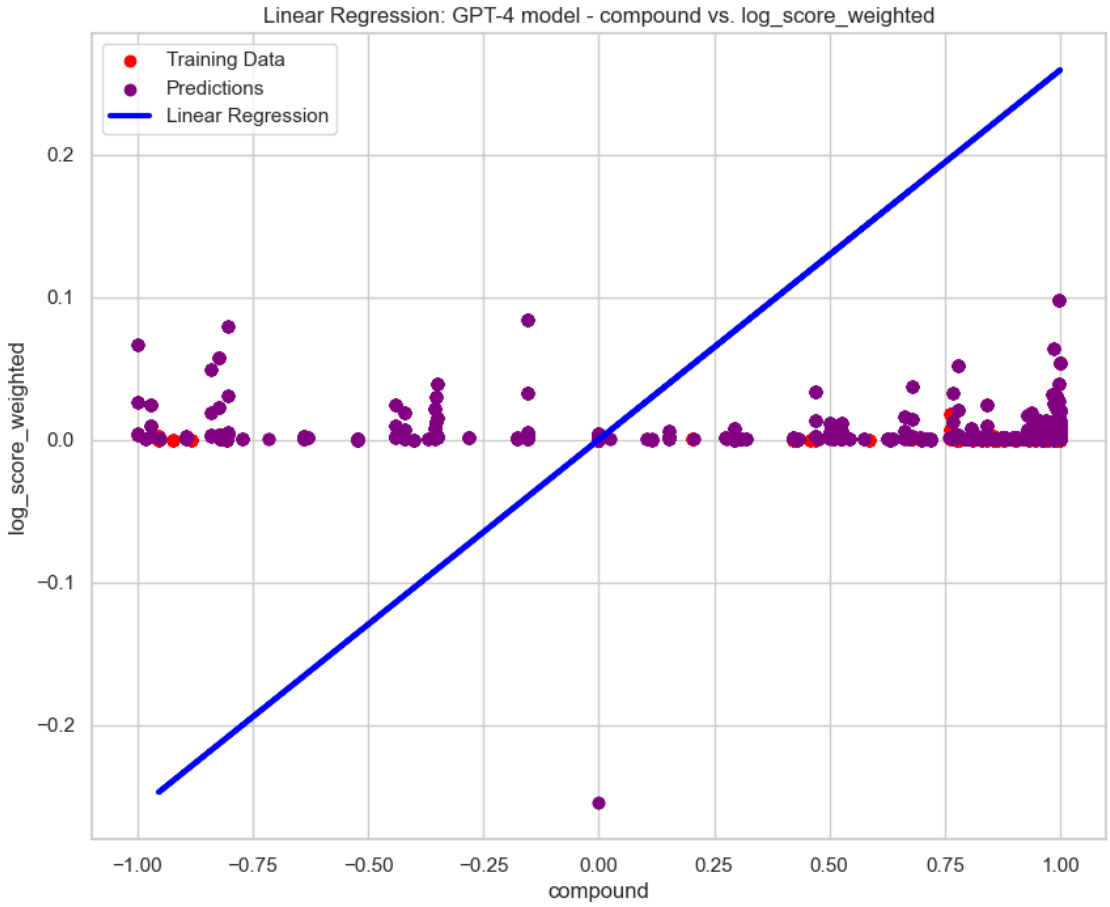
num_comments_weighted=3.7400639766919372



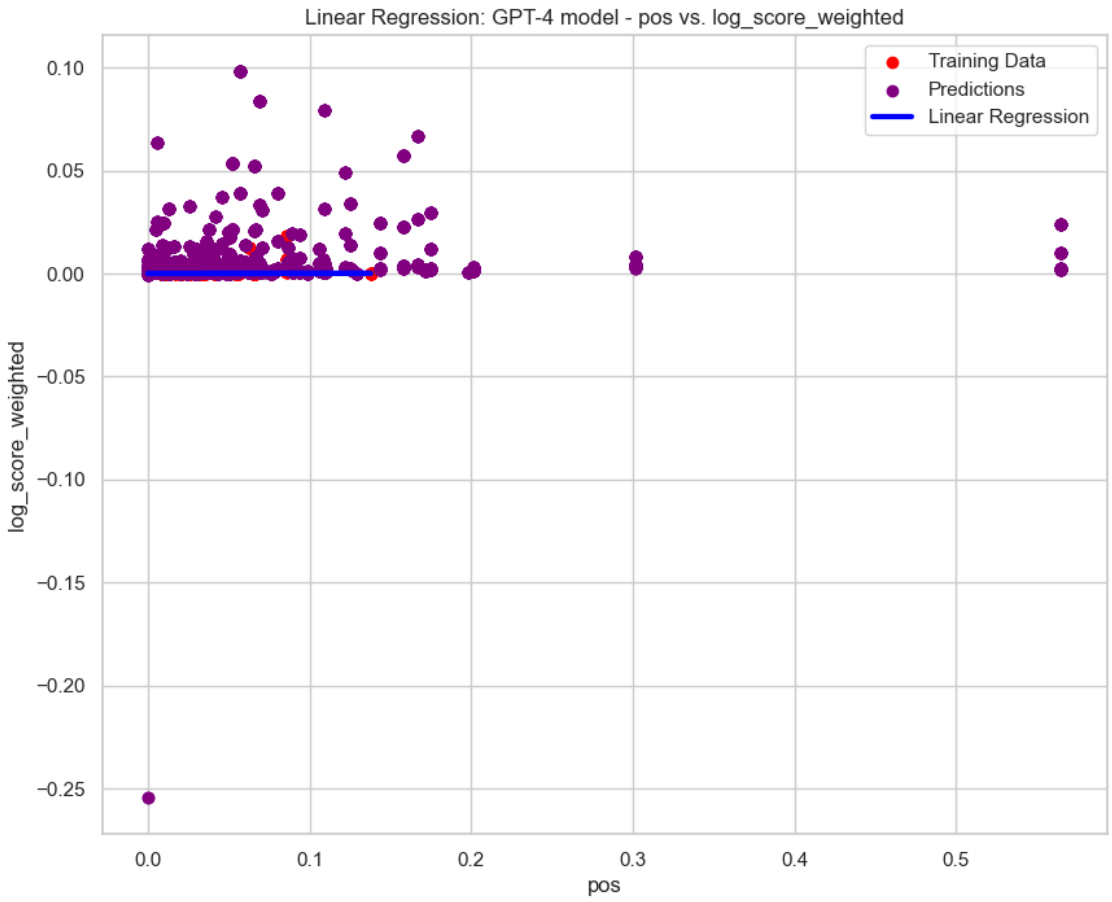
upvote_ratio=0.25666519081571215



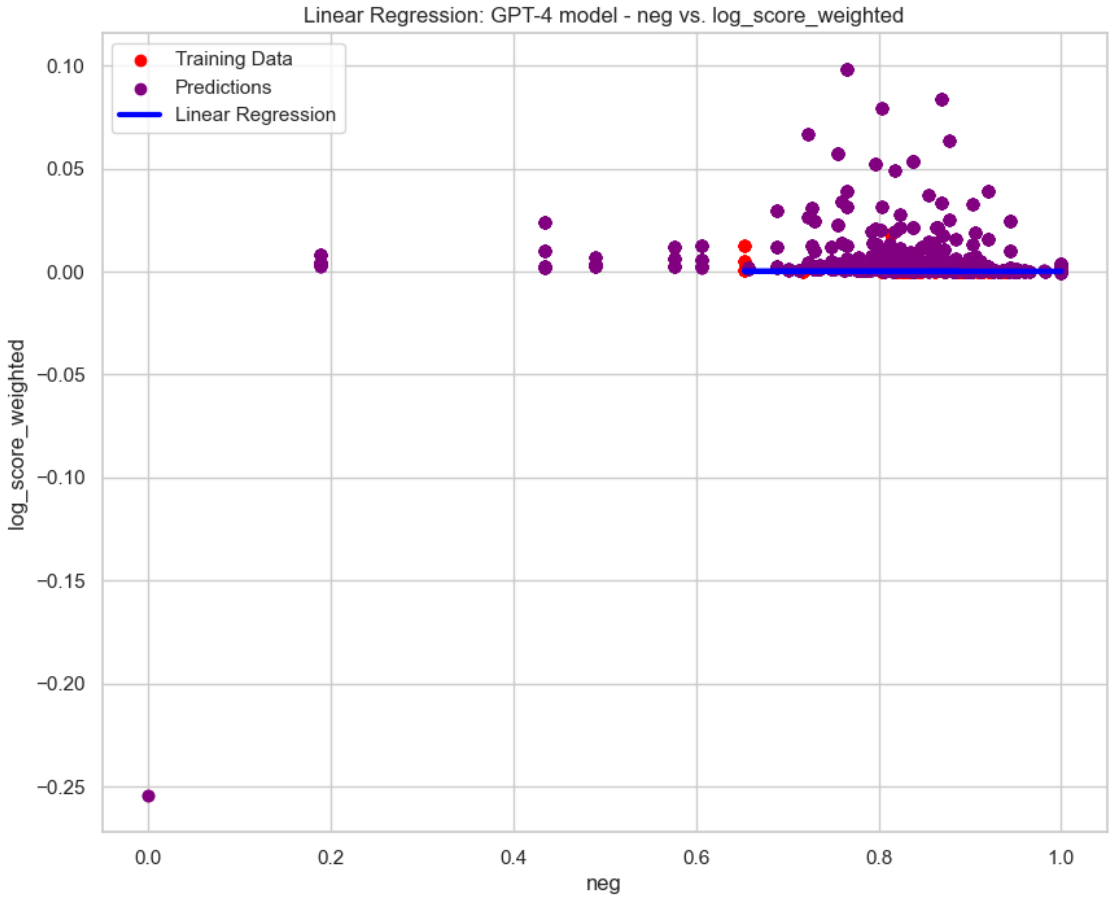

```
log_created=0.254986700298654
```



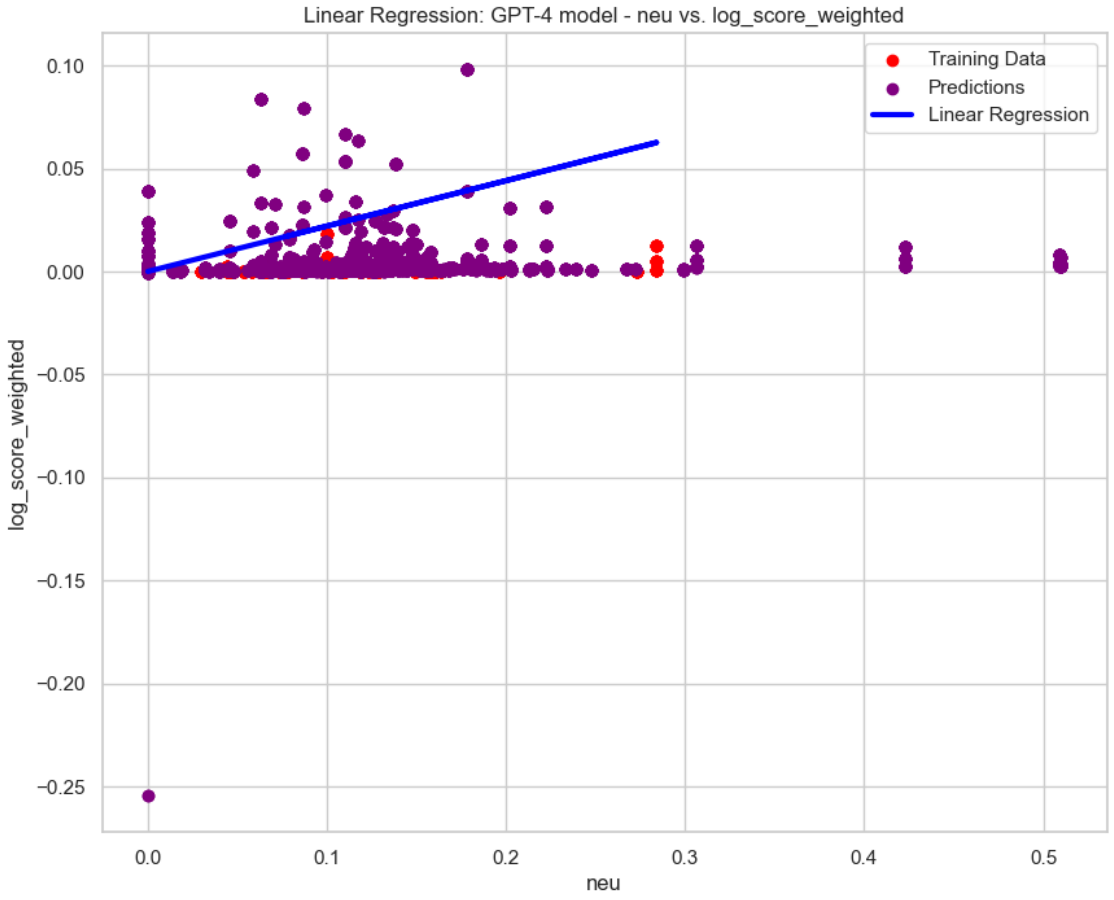
`compound=0.2592078505313998`



pos=3.409054226288788e-05

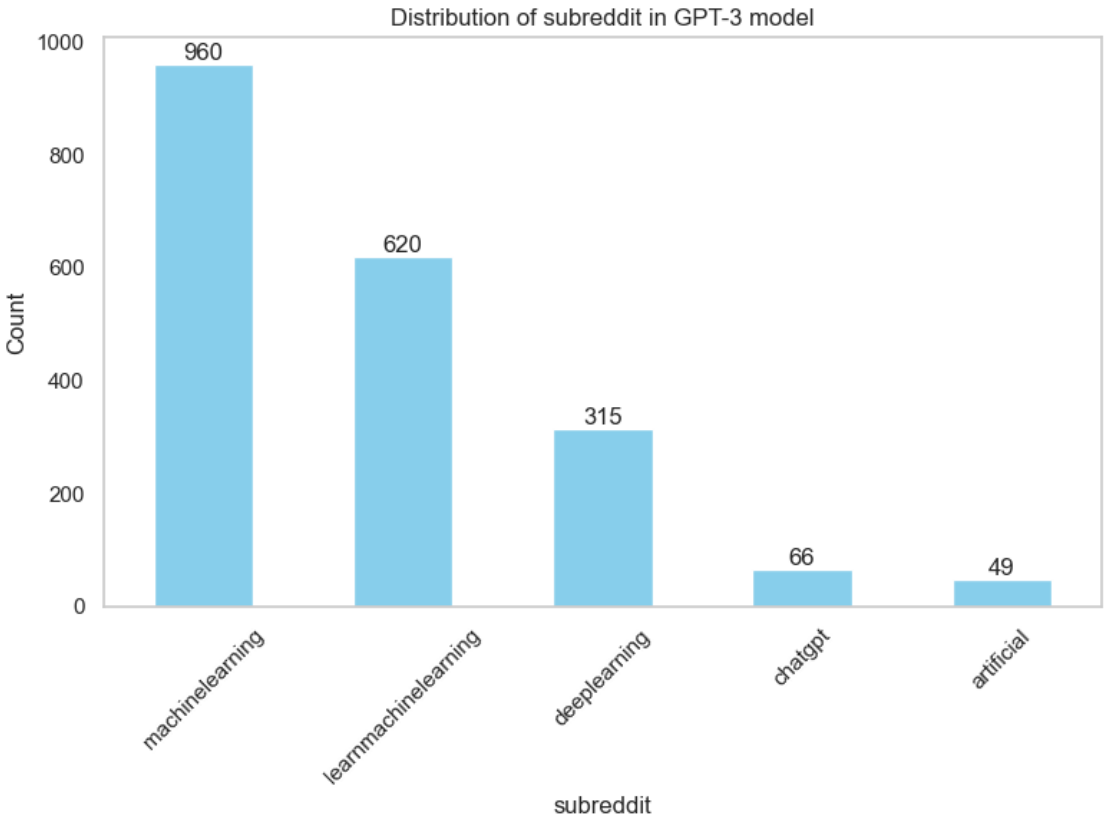


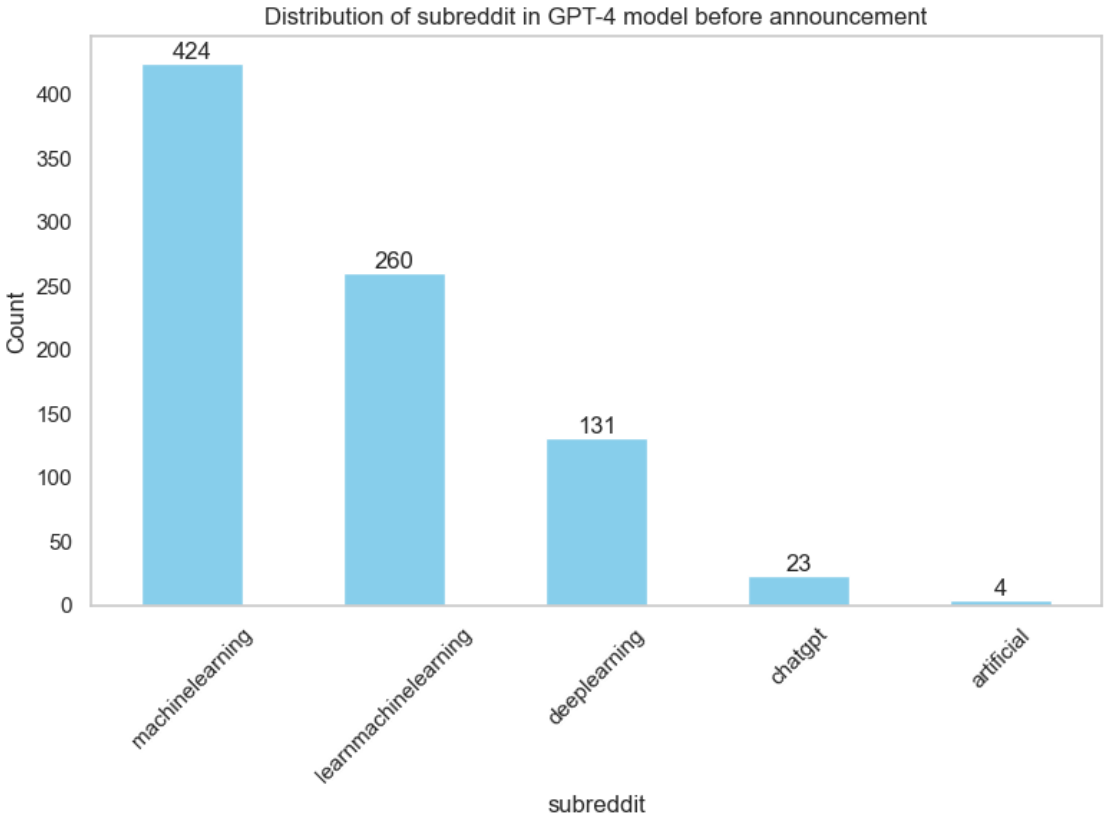
neg=8.478550332924328e-05

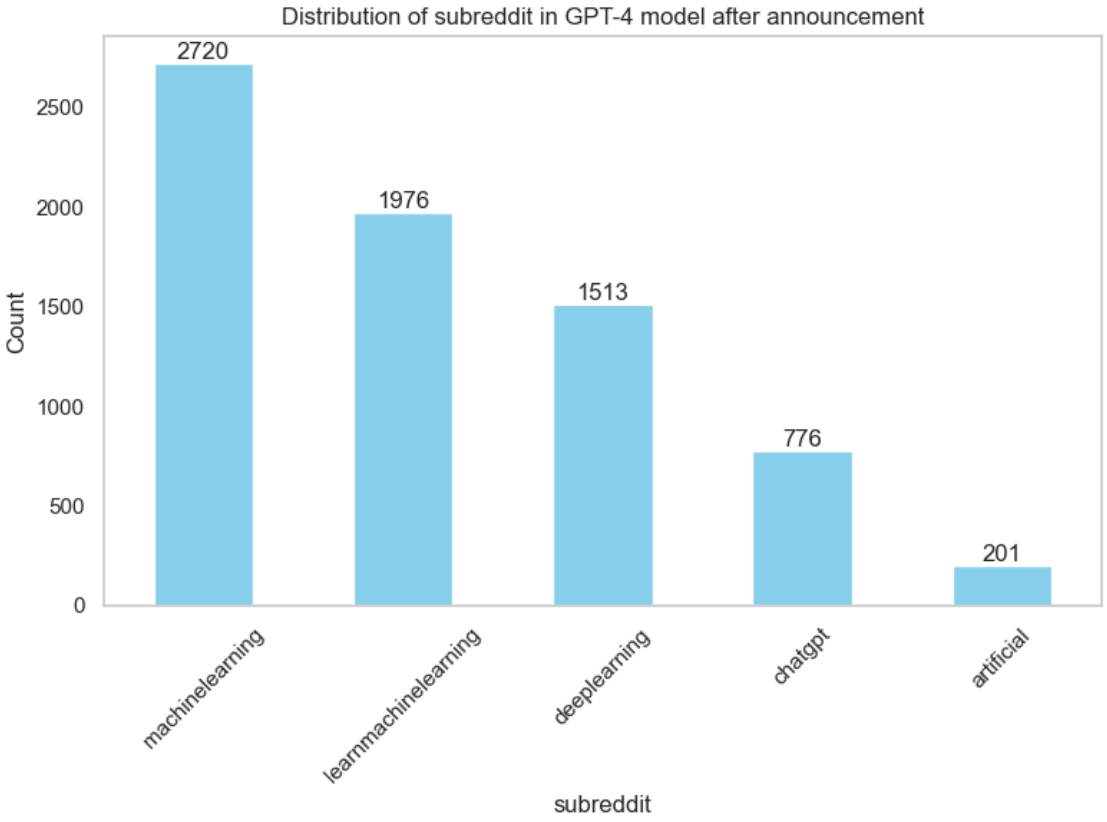


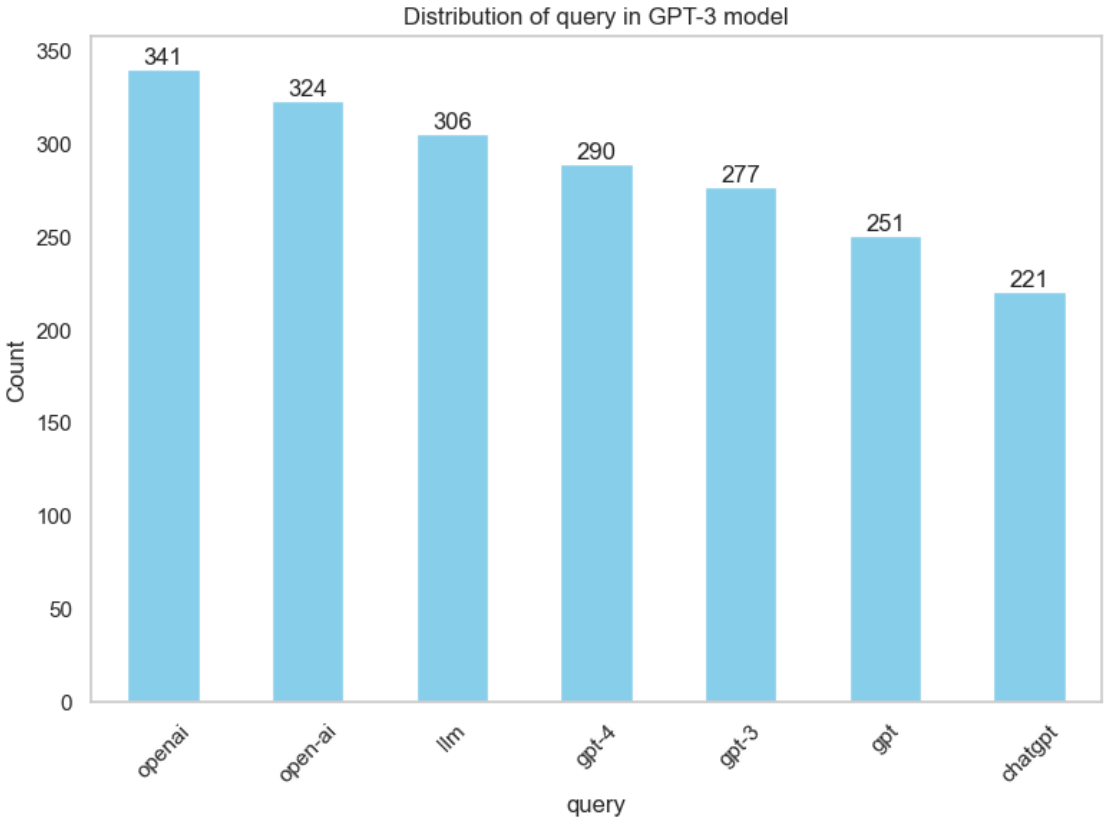
neu=0.22035240722387053

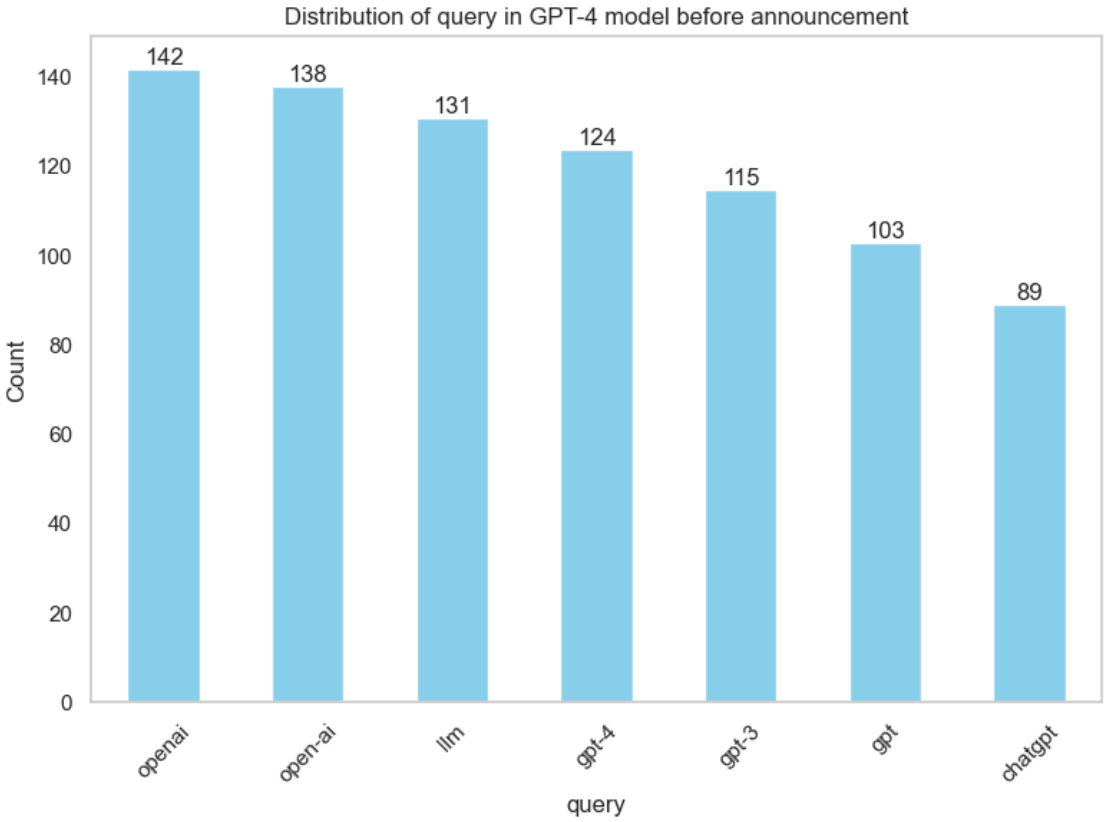
7 Categorical Distributions

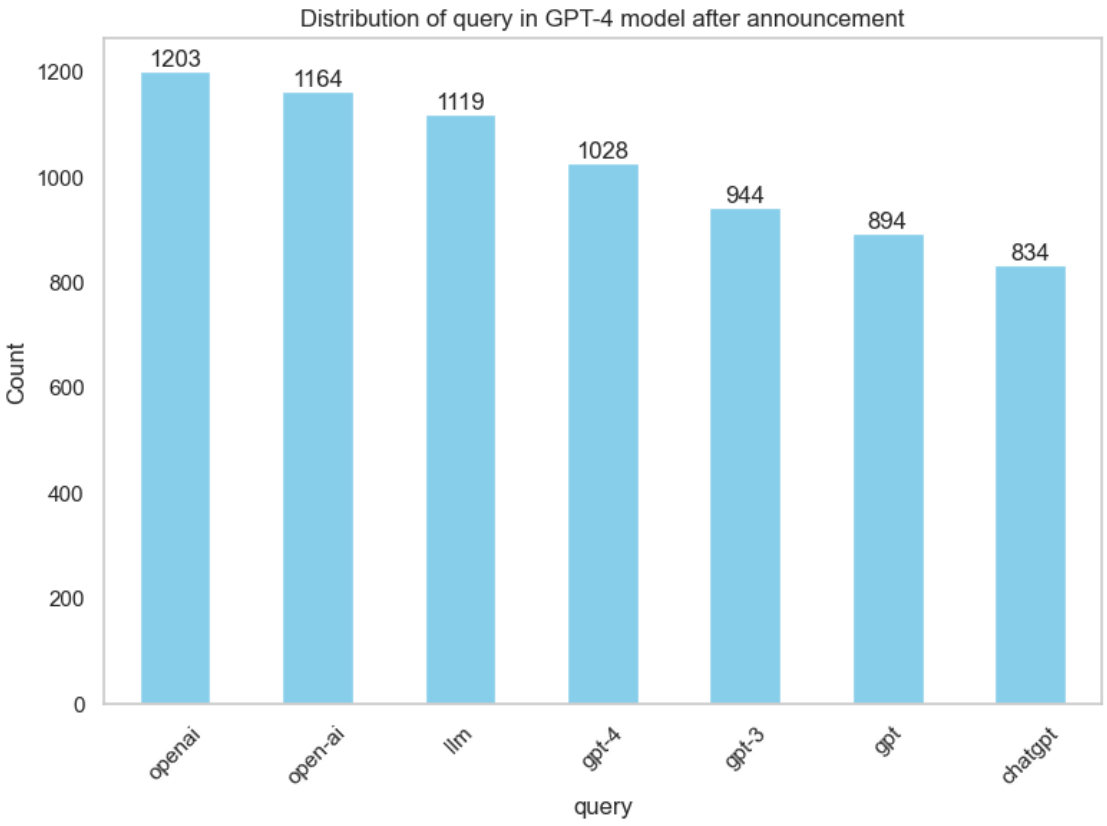



Distribution of subreddit in GPT-4 model before announcement
Count
subreddit
424
260
131
23
4
0
50
100
150
200
250
300
350
400
machinelearning
learnmachinelearning
deeplearning
chatgpt
artificial


Distribution of subreddit in GPT-4 model after announcement
Count
2720
1976
1513
776
201
2500
2000
1500
1000
500
0
machinelearning
learnmachinelearning
deeplearning
chatgpt
artificial
subreddit


Distribution of query in GPT-3 model
Count
query
0
50
100
150
200
250
300
350
341
324
306
290
277
251
221
openai
open-ai
llm
gpt-4
gpt-3
gpt
chatgpt


Distribution of query in GPT-4 model before announcement
Count
query
0
20
40
60
80
100
120
140
142
138
131
124
115
103
89
openai
open-ai
llm
gpt-4
gpt-3
gpt
chatgpt


Distribution of query in GPT-4 model after announcement
Count
query
0
200
400
600
800
1000
1200
1203
1164
1119
1028
944
894
834
openai
open-ai
llm
gpt-4
gpt-3
gpt
chatgpt

8 Heatmap of MLflow runs