# analysis

### January 24, 2024

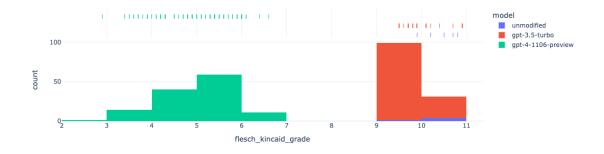
```
[25]: import pandas as pd
      import os
      from pathlib import Path
      import plotly.express as px
[26]: df = pd.read_excel("output/output.xlsx")
      df.head()
[26]:
         Unnamed: 0
                     avs_id
                                   model
                                          seed timestamp \
                  0
                           0
                             unmodified
                                             0
                                                      NaN
                           1 unmodified
      1
                  1
                                             0
                                                      NaN
      2
                  2
                           2 unmodified
                                             0
                                                      NaN
                  3
      3
                           3 unmodified
                                             0
                                                      NaN
                              unmodified
                                             0
                                                      NaN
                                                              flesch_kincaid_grade \
                                                                             10.8
        \nDepartment: ANESTHESIA PREOPERATIVE PROGRAM\...
      1 Department: ANESTHESIA PREOPERATIVE PROGRAM\nD...
                                                                             10.2
      2 \nDepartment: ANESTHESIA PREOPERATIVE PROGRAM\...
                                                                             10.5
      3 Department: ANESTHESIA PREOPERATIVE PROGRAM\nD...
                                                                             9.9
      4 Department: ANESTHESIA PREOPERATIVE PROGRAM\nD...
                                                                             10.7
         flesch_reading_ease
                               smog_index
                                          coleman_liau_index
      0
                        51.89
                                     13.3
                                                         14.10
      1
                        53.61
                                     13.1
                                                         13.92 ...
      2
                        52.80
                                     13.0
                                                         13.98 ...
      3
                       54.22
                                     12.7
                                                         11.89
                       52.09
      4
                                     13.5
                                                         14.16
         diff_difficult_words
                               diff_linsear_write_formula diff_gunning_fog \
      0
                                                        0.0
                                                                           0.0
                             0
      1
                             0
                                                        0.0
                                                                          0.0
      2
                             0
                                                        0.0
                                                                           0.0
      3
                             0
                                                        0.0
                                                                           0.0
      4
                                                        0.0
                                                                           0.0
         diff_text_standard diff_fernandez_huerta diff_szigriszt_pazos \
             Cannot compute
                                                0.0
                                                                      0.0
```

```
0.0
                                                                  0.0
1
       Cannot compute
2
       Cannot compute
                                            0.0
                                                                  0.0
3
                                            0.0
                                                                  0.0
       Cannot compute
4
       Cannot compute
                                            0.0
                                                                  0.0
   diff_gutierrez_polini diff_crawford
                                           diff_gulpease_index diff_osman
0
                      0.0
                                      0.0
                                                             0.0
                                                                         0.0
1
                      0.0
                                      0.0
                                                             0.0
                                                                         0.0
2
                      0.0
                                      0.0
                                                             0.0
                                                                         0.0
3
                      0.0
                                      0.0
                                                             0.0
                                                                         0.0
                                      0.0
                                                             0.0
                                                                         0.0
4
                      0.0
```

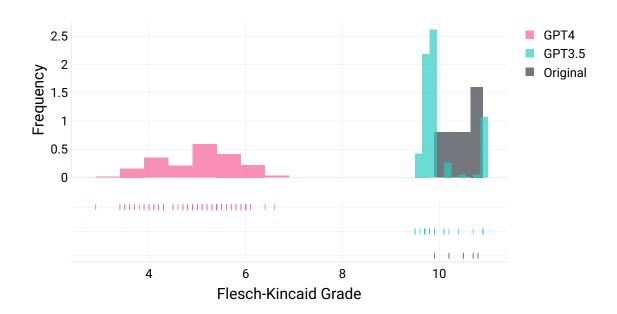
[5 rows x 38 columns]

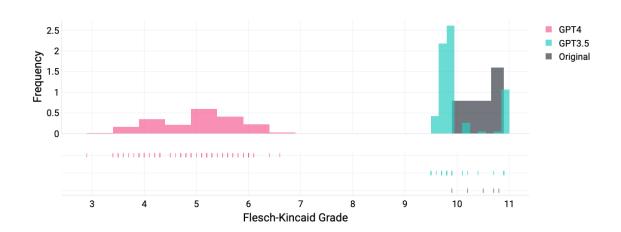
```
[27]: import plotly.io as pio

pio.templates.default = "plotly_white"
# pio.templates.default = "ggplot2"
```



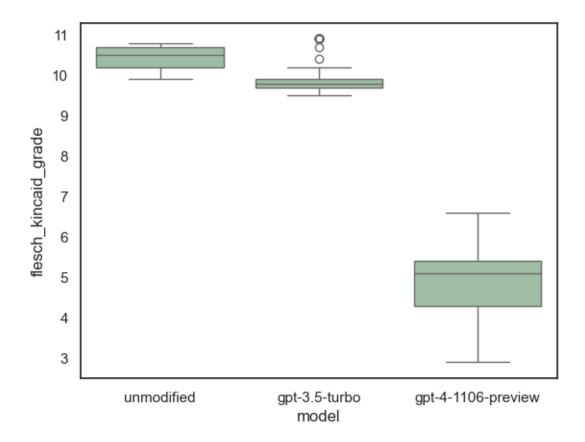
```
colors = ['#393E46', '#2BCDC1', '#F66095']
# Create distplot with custom bin_size
fig = ff.create_distplot(hist_data, group_labels, bin_size=[0.25, 0.15, 0.5],__
 ⇒show_curve=False, colors=colors)
fig.update_layout(
    width=800,
    height=400,
    # legend_title="",
    # legend_x=-0.5,
    # legend_y=0,
    # showlegend=False,
    # legend_xanchor="left",
    # legend_yanchor="bottom",
    xaxis_title="Flesch-Kincaid Grade",
    yaxis_title="Frequency",
    font=dict(
        family="Roboto",
        size=18,
        color="Black"
    ),
    xaxis_showline=True,
    xaxis_gridwidth=1,
    xaxis_linewidth=2,
    yaxis_showline=True,
    yaxis_gridwidth=1,
    yaxis_linewidth=2,
    # title_font_size=24,
    # plot_bqcolor="white",
    margin_t=0,
    margin_b=0,
    margin_l=0,
   margin_r=0,
)
fig.write_image("output/fkg.png", scale=1)
fig.show(renderer="svg")
fig.show()
```





```
[30]: # generate a boxplot to see the data distribution by treatments. Using boxplot, 
→we can

# easily detect the differences between different treatments
import matplotlib.pyplot as plt
import seaborn as sns
sns.set_theme(style="white", palette="pastel")
ax = sns.boxplot(x='model', y='flesch_kincaid_grade', data=df, color='#99c2a2')
plt.savefig("output/fkg_box.png")
plt.show()
```

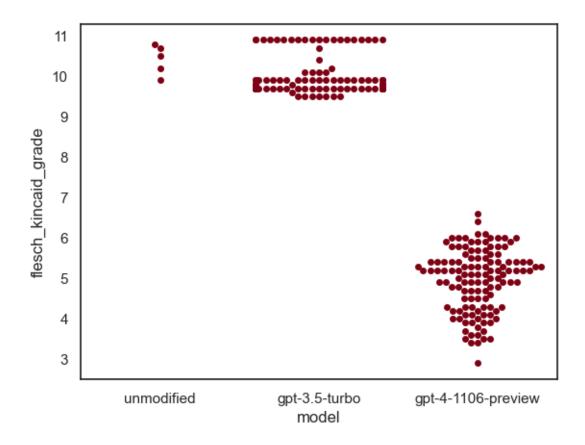


/Users/alexandergoodell/.virtualenvs/aneslit/lib/python3.11/site-packages/seaborn/categorical.py:3398: UserWarning:

33.6% of the points cannot be placed; you may want to decrease the size of the markers or use stripplot.

/Users/alexandergoodell/.virtualenvs/aneslit/lib/python3.11/site-packages/seaborn/categorical.py:3398: UserWarning:

43.2% of the points cannot be placed; you may want to decrease the size of the markers or use stripplot.



```
[33]: # See https://www.reneshbedre.com/blog/anova.html
      # ANOVA table using bioinfokit v1.0.3 or later (it uses wrapper script for \Box
       \hookrightarrow anova lm)
      from bioinfokit.analys import stat
      res = stat()
      res.anova_stat(df=df, res_var='flesch_kincaid_grade',_
       ⇒anova_model='flesch_kincaid_grade ~ C(model)')
      res.anova_summary
[33]:
                   df
                            sum_sq
                                        mean_sq
                                                                      PR(>F)
      C(model)
                  2.0 1607.776078 803.888039 2097.794154 8.180335e-158
      Residual 252.0
                         96.568000
                                       0.383206
                                                         NaN
                                                                         NaN
 []: # we will use bioinfokit (v1.0.3 or later) for performing tukey HSD test
      # check documentation here https://github.com/reneshbedre/bioinfokit
      # perform multiple pairwise comparison (Tukey's HSD)
      # unequal sample size data, tukey_hsd uses Tukey-Kramer test
      res = stat()
```

```
res.tukey_hsd(df=df, res_var='flesch_kincaid_grade', xfac_var='model', anova_model='flesch_kincaid_grade ~ C(model)')
res.tukey_summary
```

```
[35]: from scipy.stats import tukey_hsd

group0 = df.query("model == 'unmodified'")['flesch_kincaid_grade']
group1 = df.query("model == 'gpt-3.5-turbo'")['flesch_kincaid_grade']
group2 = df.query("model == 'gpt-4-1106-preview'")['flesch_kincaid_grade']

res = tukey_hsd(group0, group1, group2)
print(res)
```

```
Tukey's HSD Pairwise Group Comparisons (95.0% Confidence Interval)
Comparison Statistic p-value Lower CI Upper CI
 (0 - 1)
             0.452
                       0.247
                                -0.214
                                            1.118
 (0 - 2)
             5.456
                       0.000
                                 4.790
                                            6.122
 (1 - 0)
            -0.452
                       0.247
                                -1.118
                                           0.214
 (1 - 2)
             5.004
                       0.000
                                 4.819
                                           5.189
 (2 - 0)
            -5.456
                       0.000
                                -6.122
                                          -4.790
 (2 - 1)
            -5.004
                       0.000
                                -5.189
                                          -4.819
```

### []:

### 0.0.1 Tukey's HSD Pairwise Group Comparisons (95.0% Confidence Interval)

| Comparison          | Statistic | p-value | Lower CI | Upper CI |
|---------------------|-----------|---------|----------|----------|
| (Original - GPT3.5) | 0.452     | 0.412   | -0.385   | 1.289    |
| (Original - GPT4)   | 5.399     | 0.000   | 4.562    | 6.236    |
| (GPT3.5 - Original) | -0.452    | 0.412   | -1.289   | 0.385    |
| (GPT3.5 - GPT4)     | 4.947     | 0.000   | 4.715    | 5.179    |
| (GPT4 - Original)   | -5.399    | 0.000   | -6.236   | -4.562   |
| (GPT4 - GPT3.5)     | -4.947    | 0.000   | -5.179   | -4.715   |

```
[37]: print('-'*30)

stats = df.groupby(['Class'])['Force'].agg(['mean', 'count', 'std'])
print(stats)
print('-'*30)

ci95_hi = []
ci95_lo = []

for i in stats.index:
    m, c, s = stats.loc[i]
    ci95_hi.append(m + 1.96*s/math.sqrt(c))
```

```
ci95_lo.append(m - 1.96*s/math.sqrt(c))
      stats['ci95_hi'] = ci95_hi
      stats['ci95_lo'] = ci95_lo
      print(stats)
[37]: <bound method NDFrame._add_numeric_operations.<locals>.mean of 0
      1
           10.2
          10.5
      2
           9.9
      3
      4
          10.7
      Name: flesch_kincaid_grade, dtype: float64>
[43]: import math
      measures =
       →['flesch_kincaid_grade','flesch_reading_ease','smog_index','coleman_liau_index','automated_
       \hookrightarrow
      for measure in measures:
          print(measure)
          print("========= \n\n")
           # if continuous
          if measure != 'text_standard':
              ## mean
              print("basic descriptive stats")
              print('-'*30)
              print('\n\n')
              stats = df.groupby(['model'])[measure].agg(['mean', 'count', 'std'])
              print(stats)
              print('-'*30)
              ci95_hi = []
              ci95_lo = []
              for i in stats.index:
                  m, c, s = stats.loc[i]
                  ci95_hi.append(m + 1.96*s/math.sqrt(c))
                  ci95_lo.append(m - 1.96*s/math.sqrt(c))
              stats['ci95_hi'] = ci95_hi
              stats['ci95_lo'] = ci95_lo
```

```
print(stats)

print("\n\ntukey's HSD")
print('-'*30)

group0 = df.query("model == 'unmodified'")[measure]
group1 = df.query("model == 'gpt-3.5-turbo'")[measure]
group2 = df.query("model == 'gpt-4-1106-preview'")[measure]
res = tukey_hsd(group0, group1, group2)
print(res)

ax = sns.boxplot(x='model', y=measure, data=df, color='#99c2a2')
plt.savefig("output/" + measure + "-box-plot.png")
plt.show()
```

flesch\_kincaid\_grade

\_\_\_\_\_

#### basic descriptive stats

\_\_\_\_\_

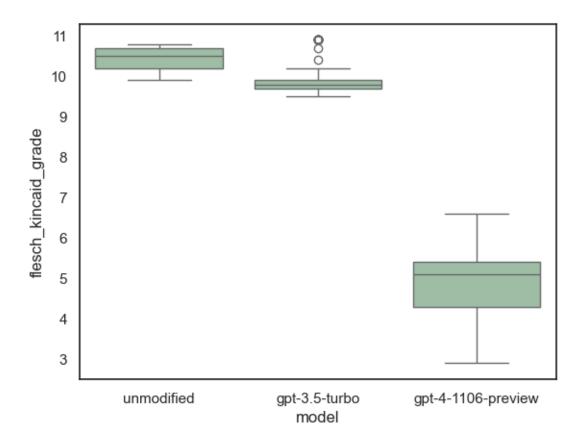
|                    | mean   | count | std      |           |           |
|--------------------|--------|-------|----------|-----------|-----------|
| model              |        |       |          |           |           |
| gpt-3.5-turbo      | 9.968  | 125   | 0.436925 |           |           |
| gpt-4-1106-preview | 4.964  | 125   | 0.763840 |           |           |
| unmodified         | 10.420 | 5     | 0.370135 |           |           |
|                    |        |       |          |           |           |
|                    | mean   | count | std      | ci95_hi   | ci95_lo   |
| model              |        |       |          |           |           |
| gpt-3.5-turbo      | 9.968  | 125   | 0.436925 | 10.044596 | 9.891404  |
| gpt-4-1106-preview | 4.964  | 125   | 0.763840 | 5.097907  | 4.830093  |
| unmodified         | 10.420 | 5     | 0.370135 | 10.744438 | 10.095562 |

#### tukey's HSD

\_\_\_\_\_

Tukey's HSD Pairwise Group Comparisons (95.0% Confidence Interval) Comparison Statistic p-value Lower CI Upper CI

| -       |        | -     |        |        |
|---------|--------|-------|--------|--------|
| (0 - 1) | 0.452  | 0.247 | -0.214 | 1.118  |
| (0 - 2) | 5.456  | 0.000 | 4.790  | 6.122  |
| (1 - 0) | -0.452 | 0.247 | -1.118 | 0.214  |
| (1 - 2) | 5.004  | 0.000 | 4.819  | 5.189  |
| (2 - 0) | -5.456 | 0.000 | -6.122 | -4.790 |
| (2 - 1) | -5.004 | 0.000 | -5.189 | -4.819 |



### flesch\_reading\_ease

\_\_\_\_\_

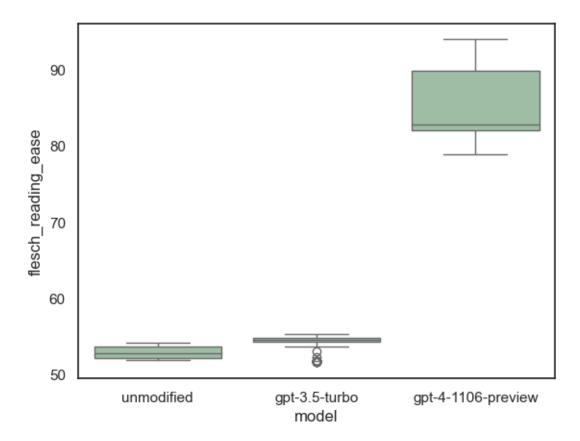
### basic descriptive stats

|                    | mean     | count | std      |           |           |
|--------------------|----------|-------|----------|-----------|-----------|
| model              |          |       |          |           |           |
| gpt-3.5-turbo      | 54.10536 | 125   | 1.107118 |           |           |
| gpt-4-1106-preview | 85.00280 | 125   | 4.277976 |           |           |
| unmodified         | 52.92200 | 5     | 0.991247 |           |           |
|                    |          | _     |          |           |           |
|                    | mean     | count | std      | ci95_hi   | ci95_lo   |
| model              |          |       |          |           |           |
| gpt-3.5-turbo      | 54.10536 | 125   | 1.107118 | 54.299446 | 53.911274 |
| gpt-4-1106-preview | 85.00280 | 125   | 4.277976 | 85.752762 | 84.252838 |
| unmodified         |          |       |          |           | 52.053134 |

tukey's HSD

| Tukey's HSD | Pairwise G | roup Comp | arisons (9 | 5.0% Confidence | ce Interval) |
|-------------|------------|-----------|------------|-----------------|--------------|
| Comparison  | Statistic  | p-value   | Lower CI   | Upper CI        |              |
| (0 - 1)     | -1.183     | 0.681     | -4.519     | 2.152           |              |

| (0 - 1) | -1.183  | 0.681 | -4.519  | 2.152   |
|---------|---------|-------|---------|---------|
| (0 - 2) | -32.081 | 0.000 | -35.417 | -28.745 |
| (1 - 0) | 1.183   | 0.681 | -2.152  | 4.519   |
| (1 - 2) | -30.897 | 0.000 | -31.823 | -29.972 |
| (2 - 0) | 32.081  | 0.000 | 28.745  | 35.417  |
| (2 - 1) | 30.897  | 0.000 | 29.972  | 31.823  |



smog\_index

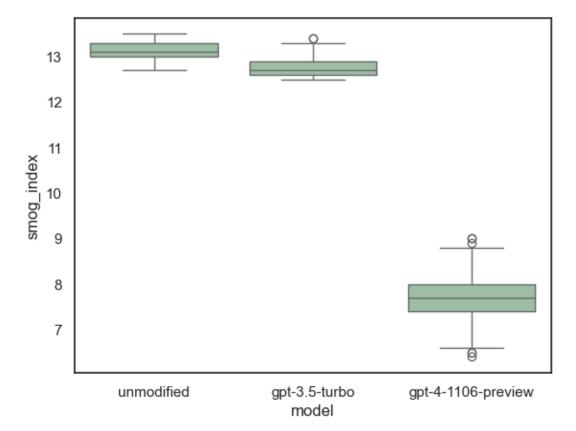
basic descriptive stats

|                        | mean            | count | std             |                      |           |
|------------------------|-----------------|-------|-----------------|----------------------|-----------|
| model                  |                 |       |                 |                      |           |
| gpt-3.5-turbo          | 12.7928         | 125   | 0.254056        |                      |           |
| gpt-4-1106-preview     | 7.6656          | 125   | 0.555051        |                      |           |
| unmodified             | 13.1200         | 5     | 0.303315        |                      |           |
|                        |                 |       |                 |                      |           |
|                        |                 |       |                 |                      |           |
|                        | mean            | count | std             | ci95_hi              | ci95_lo   |
| model                  | mean            | count | std             | ci95_hi              | ci95_lo   |
| model<br>gpt-3.5-turbo | mean<br>12.7928 | count | std<br>0.254056 | ci95_hi<br>12.837338 | ci95_lo   |
|                        |                 |       |                 | _                    | _         |
| gpt-3.5-turbo          | 12.7928         | 125   | 0.254056        | 12.837338            | 12.748262 |

-----

Tukey's HSD Pairwise Group Comparisons (95.0% Confidence Interval) Comparison Statistic p-value Lower CI Upper CI

| 0 0 mp az z 2 0 0 1 1 | 20002202 | P     |        | oppos os |  |
|-----------------------|----------|-------|--------|----------|--|
| (0 - 1)               | 0.327    | 0.219 | -0.135 | 0.789    |  |
| (0 - 2)               | 5.454    | 0.000 | 4.992  | 5.917    |  |
| (1 - 0)               | -0.327   | 0.219 | -0.789 | 0.135    |  |
| (1 - 2)               | 5.127    | 0.000 | 4.999  | 5.255    |  |
| (2 - 0)               | -5.454   | 0.000 | -5.917 | -4.992   |  |
| (2 - 1)               | -5.127   | 0.000 | -5.255 | -4.999   |  |
|                       |          |       |        |          |  |



#### coleman\_liau\_index

### basic descriptive stats

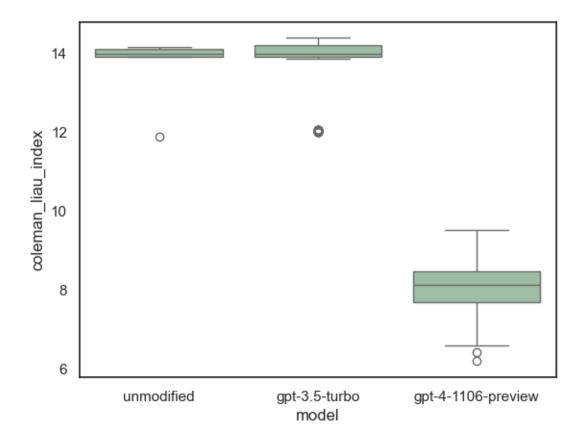
-----

|                    | mean     | count | std      |           |           |
|--------------------|----------|-------|----------|-----------|-----------|
| model              |          |       |          |           |           |
| gpt-3.5-turbo      | 13.66688 | 125   | 0.838828 |           |           |
| gpt-4-1106-preview | 8.04656  | 125   | 0.701902 |           |           |
| unmodified         | 13.61000 | 5     | 0.966178 |           |           |
|                    |          | _     |          |           |           |
|                    | mean     | count | std      | ci95_hi   | ci95_lo   |
| model              |          |       |          |           |           |
| gpt-3.5-turbo      | 12 66600 | 125   | 0.838828 | 13.813933 | 13.519827 |
| Spo c.c carbo      | 13.66688 | 125   | 0.030020 | 13.013933 | 13.519621 |
| gpt-4-1106-preview | 8.04656  | 125   | 0.701902 | 8.169609  | 7.923511  |

### tukey's HSD

\_\_\_\_\_

Tukey's HSD Pairwise Group Comparisons (95.0% Confidence Interval) Comparison Statistic p-value Lower CI Upper CI (0 - 1)-0.057 0.986 -0.892 0.778 (0 - 2)5.563 0.000 4.728 6.399 (1 - 0)0.057 0.986 -0.778 0.892 (1 - 2)5.620 0.000 5.389 5.852 (2 - 0)-5.563 0.000 -6.399 -4.728 (2 - 1)-5.620 0.000 -5.852 -5.389



automated\_readability\_index

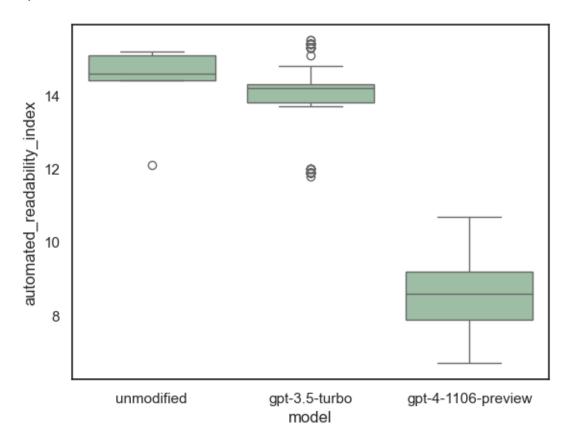
## basic descriptive stats

| mean    | count   | std   |   |  |
|---------|---|---|---|--|
|         |   |   |   |  |
| 13.9000 | 125   | 1.105631  |   |  |
| 8.6008  | 125   | 0.894923  |   |  |
| 14.2800 | 5   | 1.263725  |   |  |
|         |   |   |   |  |
| mean    | count   | std   | ci95_hi   | ci95_lo  |
|         |   |   |   |  |
| 13.9000 | 125   | 1.105631  | 14.093826   | 13.706174  |
| 8.6008  | 125   | 0.894923  | 8.757687  | 8.443913   |
| 14.2800 | 5   | 1.263725  | 15.387703   | 13.172297  |
|         | 13.9000<br>8.6008<br>14.2800<br>——————————————————————————————————— | 13.9000 125 8.6008 125 14.2800 5 mean count  13.9000 125 8.6008 125 | 13.9000 125 1.105631 8.6008 125 0.894923 14.2800 5 1.263725  mean count std  13.9000 125 1.105631 8.6008 125 0.894923 | 13.9000 125 1.105631 8.6008 125 0.894923 14.2800 5 1.263725  mean count std ci95_hi  13.9000 125 1.105631 14.093826 8.6008 125 0.894923 8.757687 |

tukey's HSD

| Tukey's HSD Pair | rwise Group Co | omparisons (95.09 | % Confidence | Interval) |
|------------------|----------------|-------------------|--------------|-----------|
|------------------|----------------|-------------------|--------------|-----------|

| Comparison | Statistic | p-value | Lower CI | Upper CI |
|------------|-----------|---------|----------|----------|
| (0 - 1)    | 0.380     | 0.688   | -0.706   | 1.466    |
| (0 - 2)    | 5.679     | 0.000   | 4.593    | 6.766    |
| (1 - 0)    | -0.380    | 0.688   | -1.466   | 0.706    |
| (1 - 2)    | 5.299     | 0.000   | 4.998    | 5.601    |
| (2 - 0)    | -5.679    | 0.000   | -6.766   | -4.593   |
| (2 - 1)    | -5.299    | 0.000   | -5.601   | -4.998   |



dale\_chall\_readability\_score

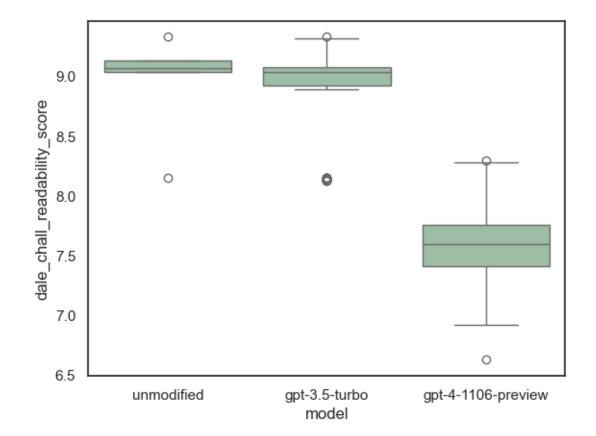
basic descriptive stats

|                    | mean    | count | std      |          |          |
|--------------------|---------|-------|----------|----------|----------|
| model              |         |       |          |          |          |
| gpt-3.5-turbo      | 8.89192 | 125   | 0.395767 |          |          |
| gpt-4-1106-preview | 7.59352 | 125   | 0.289689 |          |          |
| unmodified         | 8.94400 | 5     | 0.458017 |          |          |
|                    |         |       |          |          |          |
|                    | mean    | count | std      | ci95_hi  | ci95_lo  |
| model              |         |       |          |          |          |
| gpt-3.5-turbo      | 8.89192 | 125   | 0.395767 | 8.961301 | 8.822539 |
| gpt-4-1106-preview | 7.59352 | 125   | 0.289689 | 7.644305 | 7.542735 |
| unmodified         | 8.94400 | 5     | 0.458017 | 9.345470 | 8.542530 |

-----

Tukey's HSD Pairwise Group Comparisons (95.0% Confidence Interval)
Comparison Statistic p-value Lower CI Upper CI
(0 - 1) 0.052 0.943 -0.323 0.427

(0 - 2)0.000 1.350 0.975 1.726 (1 - 0)-0.052 0.943 -0.427 0.323 (1 - 2)1.298 0.000 1.194 1.402 (2 - 0)-1.350 0.000 -1.726 -0.975 (2 - 1)0.000 -1.298 -1.402 -1.194



#### difficult\_words

#### basic descriptive stats

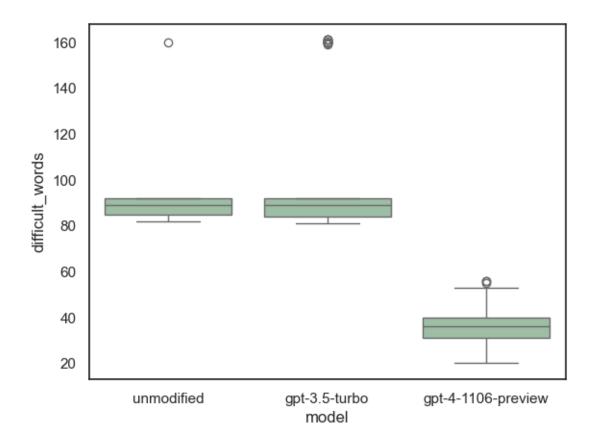
-----

|                     | mean    | count | std              |         |                      |
|---------------------|---------|-------|------------------|---------|----------------------|
| model               |         |       |                  |         |                      |
| gpt-3.5-turbo       | 101.360 | 125   | 29.711896        |         |                      |
| gpt-4-1106-preview  | 36.144  | 125   | 6.901033         |         |                      |
| unmodified          | 101.600 | 5     | 32.867917        |         |                      |
|                     |         |       |                  |         |                      |
|                     |         |       |                  |         |                      |
|                     | mean    | count | std              | ci95_hi | ci95_lo              |
| model               | mean    | count | std              | ci95_hi | ci95_lo              |
| model gpt-3.5-turbo | mean    | count | std<br>29.711896 | ci95_hi | ci95_lo<br>96.151275 |
|                     |         |       |                  | _       | -                    |

### tukey's HSD

\_\_\_\_\_

Tukey's HSD Pairwise Group Comparisons (95.0% Confidence Interval) Comparison Statistic p-value Lower CI Upper CI (0 - 1)0.240 1.000 -23.194 23.674 (0 - 2)65.456 0.000 42.022 88.890 (1 - 0)-0.240 1.000 -23.674 23.194 (1 - 2)65.216 0.000 58.717 71.715 (2 - 0)-65.456 0.000 -88.890 -42.022 (2 - 1)-65.216 0.000 -71.715 -58.717



linsear\_write\_formula

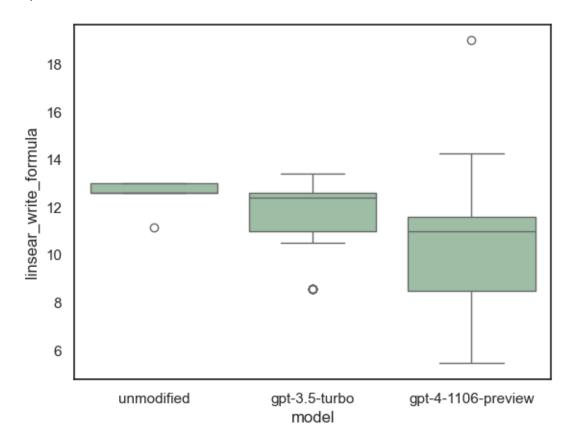
# basic descriptive stats

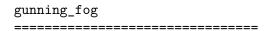
|                     | mean           | count | std      |         |         |
|---------------------|----------------|-------|----------|---------|---------|
| model               |                |       |          |         |         |
| gpt-3.5-turbo       | 11.698667      | 125   | 1.171618 |         |         |
| gpt-4-1106-preview  | 10.432775      | 125   | 2.205288 |         |         |
| unmodified          | 12.473333      | 5     | 0.757335 |         |         |
|                     |                |       |          |         |         |
|                     |                |       |          |         |         |
|                     | mean           | count | std      | ci95_hi | ci95_lo |
| model               | mean           | count | std      | ci95_hi | ci95_lo |
| model gpt-3.5-turbo | mean 11.698667 | count | std      | ci95_hi | ci95_lo |
|                     |                |       |          | _       | _       |

tukey's HSD

| Tukey's HSD | Pairwise G | roup Comp | arisons (9 | 95.0% Confide | nce Interval) |
|-------------|------------|-----------|------------|---------------|---------------|
| Comparison  | Statistic  | p-value   | Lower CI   | Upper CI      |               |
| (0 1)       | 0 775      | 0 500     | 1 110      | 0 001         |               |

| -       |        | -     |        |        |
|---------|--------|-------|--------|--------|
| (0 - 1) | 0.775  | 0.598 | -1.112 | 2.661  |
| (0 - 2) | 2.041  | 0.030 | 0.154  | 3.927  |
| (1 - 0) | -0.775 | 0.598 | -2.661 | 1.112  |
| (1 - 2) | 1.266  | 0.000 | 0.743  | 1.789  |
| (2 - 0) | -2.041 | 0.030 | -3.927 | -0.154 |
| (2 - 1) | -1.266 | 0.000 | -1.789 | -0.743 |





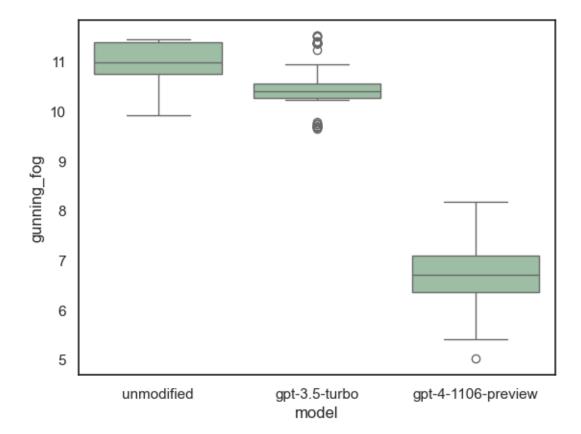
basic descriptive stats

|                     | mean          | count | std             |         |         |
|---------------------|---------------|-------|-----------------|---------|---------|
| model               |               |       |                 |         |         |
| gpt-3.5-turbo       | 10.44928      | 125   | 0.522048        |         |         |
| gpt-4-1106-preview  | 6.73832       | 125   | 0.568708        |         |         |
| unmodified          | 10.90200      | 5     | 0.621184        |         |         |
|                     |               | _     |                 |         |         |
|                     |               |       |                 |         |         |
|                     | mean          | count | std             | ci95_hi | ci95_lo |
| model               | mean          | count | std             | ci95_hi | ci95_lo |
| model gpt-3.5-turbo | mean 10.44928 | count | std<br>0.522048 | ci95_hi | ci95_lo |
|                     |               |       |                 | _       | -       |

-----

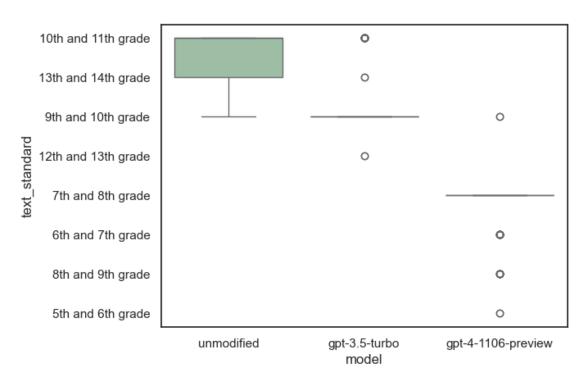
Tukey's HSD Pairwise Group Comparisons (95.0% Confidence Interval)

| Comparison | Statistic | p-value | Lower CI | Upper CI |  |
|------------|-----------|---------|----------|----------|--|
| (0 - 1)    | 0.453     | 0.167   | -0.136   | 1.041    |  |
| (0 - 2)    | 4.164     | 0.000   | 3.575    | 4.752    |  |
| (1 - 0)    | -0.453    | 0.167   | -1.041   | 0.136    |  |
| (1 - 2)    | 3.711     | 0.000   | 3.548    | 3.874    |  |
| (2 - 0)    | -4.164    | 0.000   | -4.752   | -3.575   |  |
| (2 - 1)    | -3.711    | 0.000   | -3.874   | -3.548   |  |



### ${\tt text\_standard}$

\_\_\_\_\_



### ${\tt fernandez\_huerta}$

### basic descriptive stats

|                    | mean      | count | std      |            |            |
|--------------------|-----------|-------|----------|------------|------------|
| model              |           |       |          |            |            |
| gpt-3.5-turbo      | 93.38472  | 125   | 1.111094 |            |            |
| gpt-4-1106-preview | 116.12800 | 125   | 3.192405 |            |            |
| unmodified         | 92.19400  | 5     | 0.998864 |            |            |
|                    |           |       |          |            |            |
|                    | mean      | count | std      | ci95_hi    | ci95_lo    |
| model              |           |       |          |            |            |
| gpt-3.5-turbo      | 93.38472  | 125   | 1.111094 | 93.579503  | 93.189937  |
| gpt-4-1106-preview | 116.12800 | 125   | 3.192405 | 116.687653 | 115.568347 |

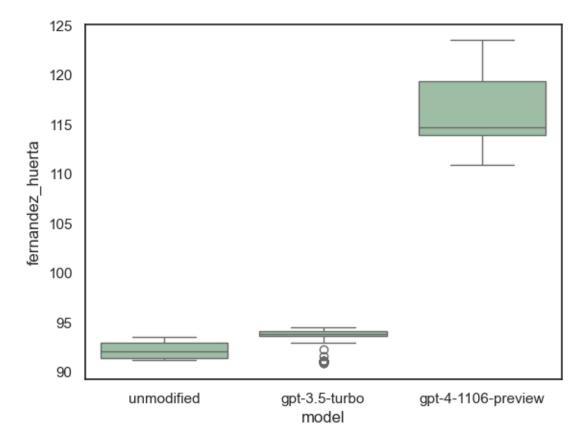
unmodified 92.19400 5 0.998864 93.069543 91.318457

### tukey's HSD

\_\_\_\_\_

Tukey's HSD Pairwise Group Comparisons (95.0% Confidence Interval)

| J          |           |         |          |          |  |
|------------|-----------|---------|----------|----------|--|
| Comparison | Statistic | p-value | Lower CI | Upper CI |  |
| (0 - 1)    | -1.191    | 0.515   | -3.744   | 1.362    |  |
| (0 - 2)    | -23.934   | 0.000   | -26.487  | -21.381  |  |
| (1 - 0)    | 1.191     | 0.515   | -1.362   | 3.744    |  |
| (1 - 2)    | -22.743   | 0.000   | -23.451  | -22.035  |  |
| (2 - 0)    | 23.934    | 0.000   | 21.381   | 26.487   |  |
| (2 - 1)    | 22.743    | 0.000   | 22.035   | 23.451   |  |



szigriszt\_pazos

basic descriptive stats

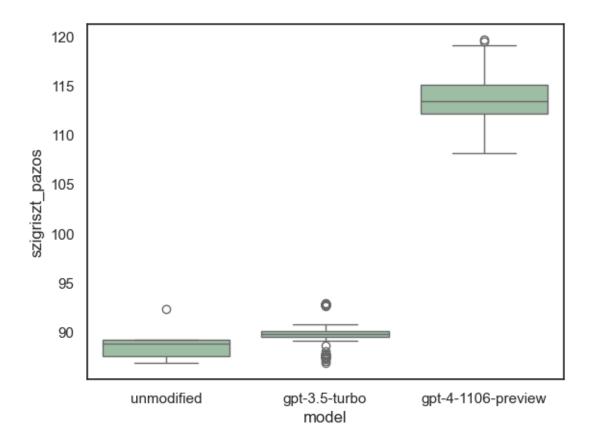
\_\_\_\_\_

|                    | mean      | count | std      |            |            |
|--------------------|-----------|-------|----------|------------|------------|
| model              |           |       |          |            |            |
| gpt-3.5-turbo      | 90.04600  | 125   | 1.655640 |            |            |
| gpt-4-1106-preview | 113.71752 | 125   | 2.389298 |            |            |
| unmodified         | 89.00000  | 5     | 2.107048 |            |            |
|                    |           |       |          |            |            |
|                    | mean      | count | std      | ci95_hi    | ci95_lo    |
| model              |           |       |          |            |            |
| gpt-3.5-turbo      | 90.04600  | 125   | 1.655640 | 90.336247  | 89.755753  |
| gpt-4-1106-preview | 113.71752 | 125   | 2.389298 | 114.136382 | 113.298658 |
| unmodified         | 89.00000  | 5     | 2.107048 | 90.846909  | 87.153091  |

-----

Tukey's HSD Pairwise Group Comparisons (95.0% Confidence Interval)

| Comparison | Statistic | p-value | Lower CI | Upper CI |
|------------|-----------|---------|----------|----------|
| (0 - 1)    | -1.046    | 0.505   | -3.257   | 1.165    |
| (0 - 2)    | -24.718   | 0.000   | -26.929  | -22.506  |
| (1 - 0)    | 1.046     | 0.505   | -1.165   | 3.257    |
| (1 - 2)    | -23.672   | 0.000   | -24.285  | -23.058  |
| (2 - 0)    | 24.718    | 0.000   | 22.506   | 26.929   |
| (2 - 1)    | 23.672    | 0.000   | 23.058   | 24.285   |
|            |           |         |          |          |



# gutierrez\_polini

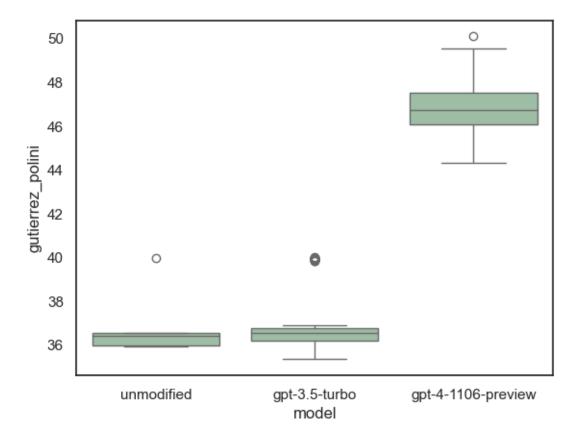
# basic descriptive stats

|                    | mean     | count | std      |           |           |
|--------------------|----------|-------|----------|-----------|-----------|
| model              |          |       |          |           |           |
| gpt-3.5-turbo      | 37.04640 | 125   | 1.512085 |           |           |
| gpt-4-1106-preview | 46.82184 | 125   | 1.213447 |           |           |
| unmodified         | 36.97600 | 5     | 1.699126 |           |           |
|                    |          | _     |          |           |           |
|                    | mean     | count | std      | ci95_hi   | ci95_lo   |
| model              |          |       |          |           |           |
| gpt-3.5-turbo      | 37.04640 | 125   | 1.512085 | 37.311480 | 36.781320 |
| gpt-4-1106-preview | 46.82184 | 125   | 1.213447 | 47.034567 | 46.609113 |
| unmodified         | 36.97600 | 5     | 1.699126 | 38.465350 | 35.486650 |

tukey's HSD

Tukey's HSD Pairwise Group Comparisons (95.0% Confidence Interval)

| Comparison | Statistic | p-value | Lower CI | Upper CI |
|------------|-----------|---------|----------|----------|
| (0 - 1)    | -0.070    | 0.993   | -1.551   | 1.410    |
| (0 - 2)    | -9.846    | 0.000   | -11.326  | -8.365   |
| (1 - 0)    | 0.070     | 0.993   | -1.410   | 1.551    |
| (1 - 2)    | -9.775    | 0.000   | -10.186  | -9.365   |
| (2 - 0)    | 9.846     | 0.000   | 8.365    | 11.326   |
| (2 - 1)    | 9.775     | 0.000   | 9.365    | 10.186   |



### crawford

basic descriptive stats

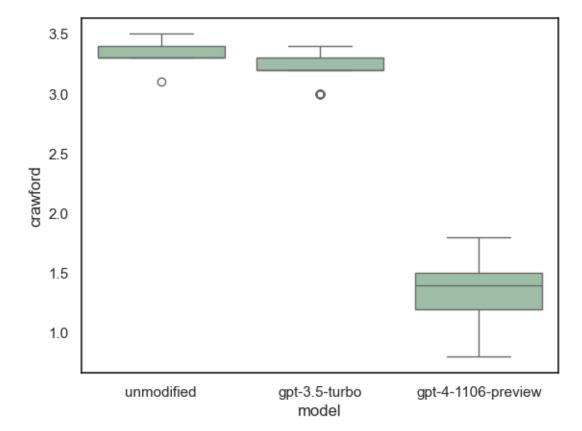
\_\_\_\_\_

|                     | mean        | count | std             |                  |                     |
|---------------------|-------------|-------|-----------------|------------------|---------------------|
| model               |             |       |                 |                  |                     |
| gpt-3.5-turbo       | 3.2240      | 125   | 0.131615        |                  |                     |
| gpt-4-1106-preview  | 1.3464      | 125   | 0.209663        |                  |                     |
| unmodified          | 3.3200      | 5     | 0.148324        |                  |                     |
|                     |             |       |                 |                  |                     |
|                     |             |       |                 |                  |                     |
|                     | mean        | count | std             | ci95_hi          | ci95_lo             |
| model               | mean        | count | std             | ci95_hi          | ci95_lo             |
| model gpt-3.5-turbo | mean 3.2240 | count | std<br>0.131615 | ci95_hi 3.247073 | ci95_lo<br>3.200927 |
|                     |             |       |                 | _                | _                   |
| gpt-3.5-turbo       | 3.2240      | 125   | 0.131615        | 3.247073         | 3.200927            |

-----

Tukey's HSD Pairwise Group Comparisons (95.0% Confidence Interval)

| Statistic | p-value                                     | Lower CI  | Upper CI   |  |
|-----------|---|---|--|--|
| 0.096     | 0.451                                       | -0.092  | 0.284  |  |
| 1.974     | 0.000                                       | 1.786   | 2.161  |  |
| -0.096    | 0.451                                       | -0.284  | 0.092  |  |
| 1.878     | 0.000                                       | 1.826   | 1.930  |  |
| -1.974    | 0.000                                       | -2.161  | -1.786   |  |
| -1.878    | 0.000                                       | -1.930  | -1.826   |  |
|           | 0.096<br>1.974<br>-0.096<br>1.878<br>-1.974 | 0.096 0.451<br>1.974 0.000<br>-0.096 0.451<br>1.878 0.000<br>-1.974 0.000 | 0.096       0.451       -0.092         1.974       0.000       1.786         -0.096       0.451       -0.284         1.878       0.000       1.826         -1.974       0.000       -2.161 | 1.974       0.000       1.786       2.161         -0.096       0.451       -0.284       0.092         1.878       0.000       1.826       1.930         -1.974       0.000       -2.161       -1.786 |



### ${\tt gulpease\_index}$

\_\_\_\_\_\_

#### basic descriptive stats

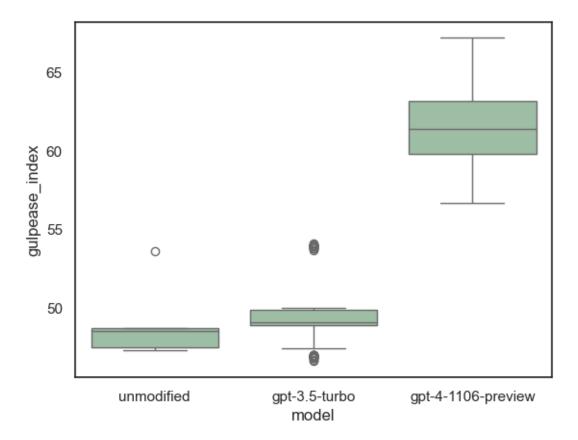
-----

|                    | mean    | count | std      |           |           |
|--------------------|---------|-------|----------|-----------|-----------|
| model              |         |       |          |           |           |
| gpt-3.5-turbo      | 49.7368 | 125   | 2.295105 |           |           |
| gpt-4-1106-preview | 61.5272 | 125   | 2.318939 |           |           |
| unmodified         | 49.1200 | 5     | 2.577208 |           |           |
|                    |         |       |          |           |           |
|                    | mean    | count | std      | ci95_hi   | ci95_lo   |
| model              |         |       |          |           |           |
| gpt-3.5-turbo      | 49.7368 | 125   | 2.295105 | 50.139150 | 49.334450 |
| gpt-4-1106-preview | 61.5272 | 125   | 2.318939 | 61.933728 | 61.120672 |
| unmodified         | 49.1200 | 5     | 2.577208 | 51.379022 | 46.860978 |

### tukey's HSD

\_\_\_\_\_

Tukey's HSD Pairwise Group Comparisons (95.0% Confidence Interval) Comparison Statistic p-value Lower CI Upper CI (0 - 1)-0.617 0.828 -3.102 1.869 (0 - 2)-12.407 0.000 -14.893 -9.922 (1 - 0)0.617 0.828 -1.869 3.102 (1 - 2)-11.790 0.000 -12.480 -11.101 (2 - 0)12.407 0.000 9.922 14.893 11.790 (2 - 1)0.000 11.101 12.480



### osman

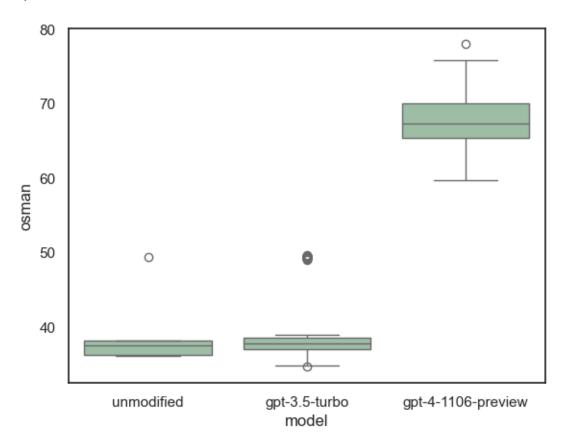
### basic descriptive stats

|                    | mean     | count | std      |           |           |
|--------------------|----------|-------|----------|-----------|-----------|
| model              |          |       |          |           |           |
| gpt-3.5-turbo      | 39.65448 | 125   | 4.988408 |           |           |
| gpt-4-1106-preview | 67.71648 | 125   | 3.773853 |           |           |
| unmodified         | 39.47800 | 5     | 5.598577 |           |           |
|                    |          | _     |          |           |           |
|                    | mean     | count | std      | ci95_hi   | ci95_lo   |
| model              |          |       |          |           |           |
| gpt-3.5-turbo      | 39.65448 | 125   | 4.988408 | 40.528986 | 38.779974 |
| gpt-4-1106-preview | 67.71648 | 125   | 3.773853 | 68.378066 | 67.054894 |
| OP                 | 01.11040 | 120   | 0.110000 | 00.070000 | 01.001001 |

tukey's HSD

| Tukey's HSD | Pairwise G | roup Comp | arisons (9 | 5.0% Confidence | e Interval) |
|-------------|------------|-----------|------------|-----------------|-------------|
| Comparison  | Statistic  | p-value   | Lower CI   | Upper CI        |             |
| (0 - 1)     | -0.176     | 0.996     | -4.955     | 4.602           |             |

| (0 - 1) | -0.176  | 0.996 | -4.955  | 4.602   |
|---------|---------|-------|---------|---------|
| (0 - 2) | -28.238 | 0.000 | -33.017 | -23.460 |
| (1 - 0) | 0.176   | 0.996 | -4.602  | 4.955   |
| (1 - 2) | -28.062 | 0.000 | -29.387 | -26.737 |
| (2 - 0) | 28.238  | 0.000 | 23.460  | 33.017  |
| (2 - 1) | 28.062  | 0.000 | 26.737  | 29.387  |



[]: