explore_oof_1

September 10, 2022

[ ]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

[ ]:

```
df = pd.read_pickle("../data/external/oof_1/oof_df.pkl")
```

[ ]:

```
df.head()
```

[ ]:

```
        text_id                                          full_text  cohesion  \
0  0022683E9EA5  When a problem is a change you have to let it …       2.5   
1  009F4E9310CB  Asking more than one person for and advice hel…       3.0   
2  00B21F9B726F  Do you think its a good idea for students to c…       3.0   
3  00D281524375  Technology allows people to do many things suc…       3.5   
4  01350DF42AED  When, the people decide to have a good posture…       2.0   

   syntax  vocabulary  phraseology  grammar  conventions  fold  pred_cohesion  \
0     2.5         3.0          2.0      2.0          2.5     0       2.761948   
1     3.0         3.5          2.5      3.0          2.5     0       2.757592   
2     3.5         3.5          3.5      3.5          3.0     0       3.197150   
3     2.5         3.5          3.0      3.0          3.0     0       2.899300   
4     2.0         2.0          2.0      2.0          2.0     0       2.368047   

   pred_syntax  pred_vocabulary  pred_phraseology  pred_grammar  \
0     2.653258         2.902649          2.727023      2.515462   
1     2.550056         2.894924          2.669984      2.398576   
2     3.105509         3.261744          3.186763      3.271552   
3     2.877304         3.091960          2.997519      3.090804   
4     2.142393         2.568555          2.226166      1.981846   

   pred_conventions  
0          2.688283  
1          2.566344  
2          3.108005  
3          2.696547  
4          2.349967  
```

[ ]:

```

```

```
[ ]: import torch
     from src.modules.loss import RMSELoss
```

```
/home/miyakawa/workspace/kaggle/feedback-prize-english-language-
learning/.venv/lib/python3.7/site-packages/tqdm/auto.py:22: TqdmWarning:
IProgress not found. Please update jupyter and ipywidgets. See
https://ipywidgets.readthedocs.io/en/stable/user_install.html
  from .autonotebook import tqdm as notebook_tqdm
```

```
[ ]: loss_fn = RMSELoss()

     true_colnames = ["cohesion", "syntax", "vocabulary", "phraseology", "grammar",␣
      ↪"conventions"]
     pred_colnames = ["pred_cohesion", "pred_syntax", "pred_vocabulary",␣
      ↪"pred_phraseology", "pred_grammar", "pred_conventions"]

     df["loss"] = df.apply(lambda x: loss_fn(torch.from_numpy(x[pred_colnames].
      ↪values.astype(np.float32)), torch.from_numpy(x[true_colnames].values.
      ↪astype(np.float32))).item(), axis=1)
```

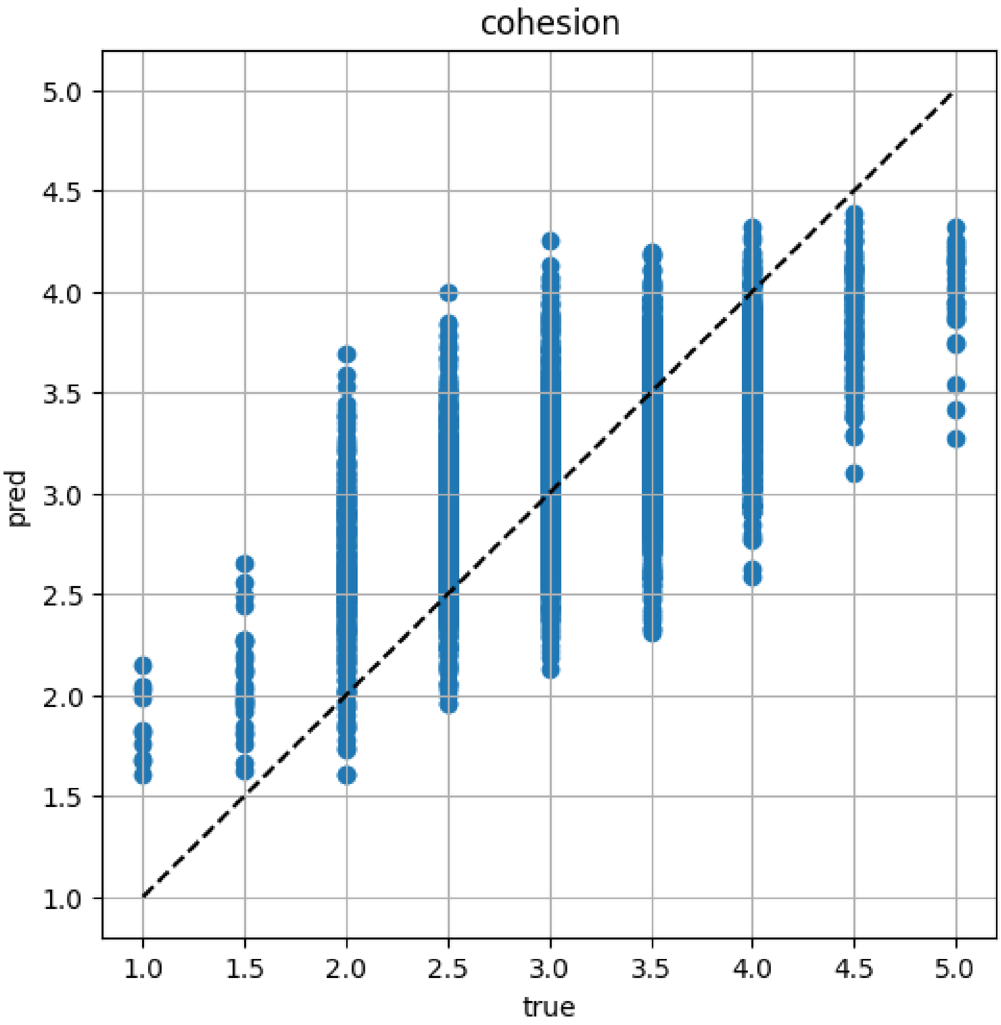
```
[ ]: df.sort_values("loss", ascending=False)
```

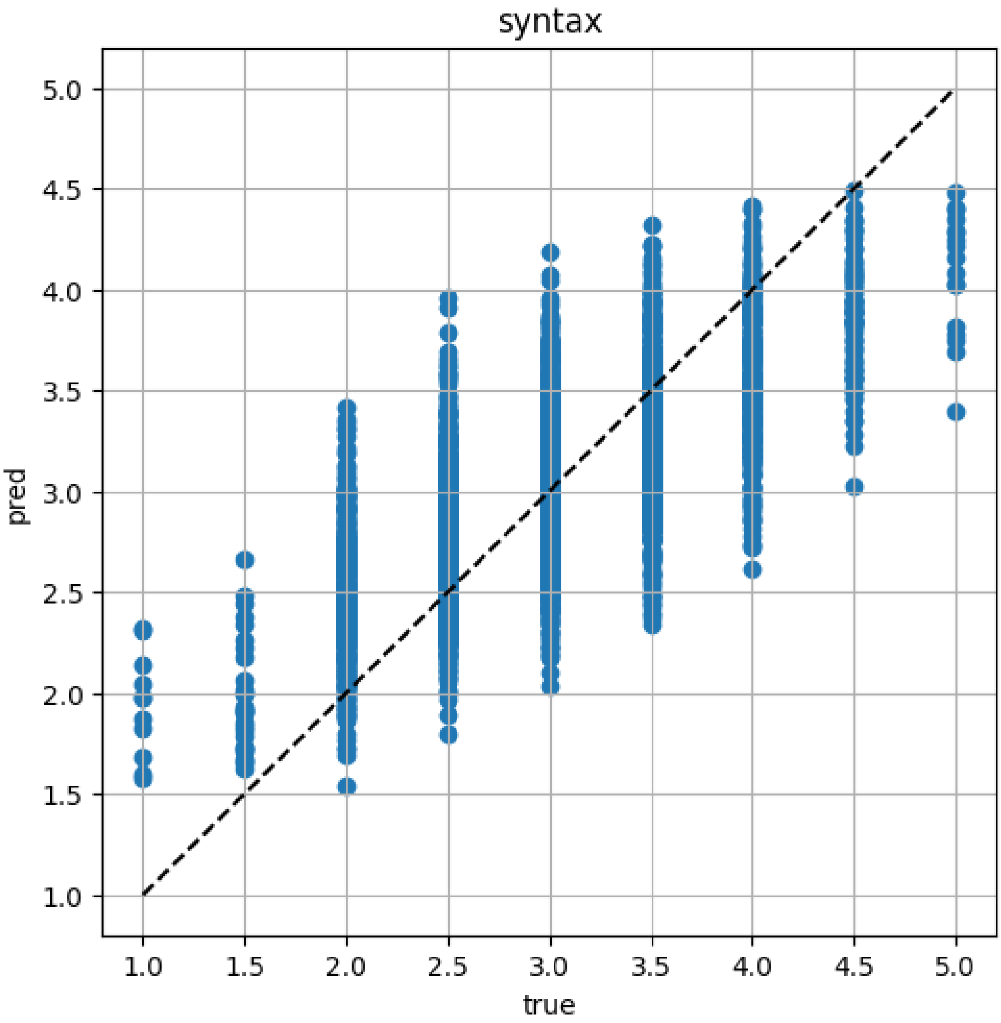
```
[ ]:           text_id                                          full_text  \
     1209  48EA282A4EAF  some student offer distance learning as an opt…
     3342  7CAFF57E1775  Dear principal,\n\nI have heard of the school …
     1796  E545B850725F  Although some say extracurricular activities a…
     2111  2D5A9BEEB30D  Do you agree or disagree with extending the sc…
     2430  92EA64272FB4  High school is like a preschool for young adul…
     …              …                                                …
     1153  386C2A01AD9D  If you could pick a job what would it be? Have…
     3304  762839543BFC  In the quote, Thomas Jeff er son take about de…
     3403  8F5F3F2519DA  what do you think about a program that school …
     2774  E73CF852AC12  Positive attitude is the key to be successful …
     1014  0C699C871AA5  Students will not be benefited if they attend …

           cohesion  syntax  vocabulary  phraseology  grammar  conventions  fold  \
     1209       1.0     1.0         1.0          1.0      1.0          1.0     1
     3342       4.5     5.0         5.0          4.5      5.0          4.0     3
     1796       2.5     2.5         3.0          2.5      2.5          2.5     1
     2111       4.0     4.5         4.0          4.0      4.5          5.0     2
     2430       2.0     2.5         3.0          3.0      3.0          2.5     2
     …            …       …           …            …        …          …     …
     1153       3.0     3.0         3.0          3.0      3.0          3.0     1
     3304       2.0     2.0         2.5          2.0      2.0          2.0     3
     3403       3.5     3.5         3.5          3.5      3.5          3.5     3
     2774       3.5     3.5         3.5          3.5      3.5          3.5     2
     1014       3.5     3.5         3.5          3.5      3.5          3.5     1
```

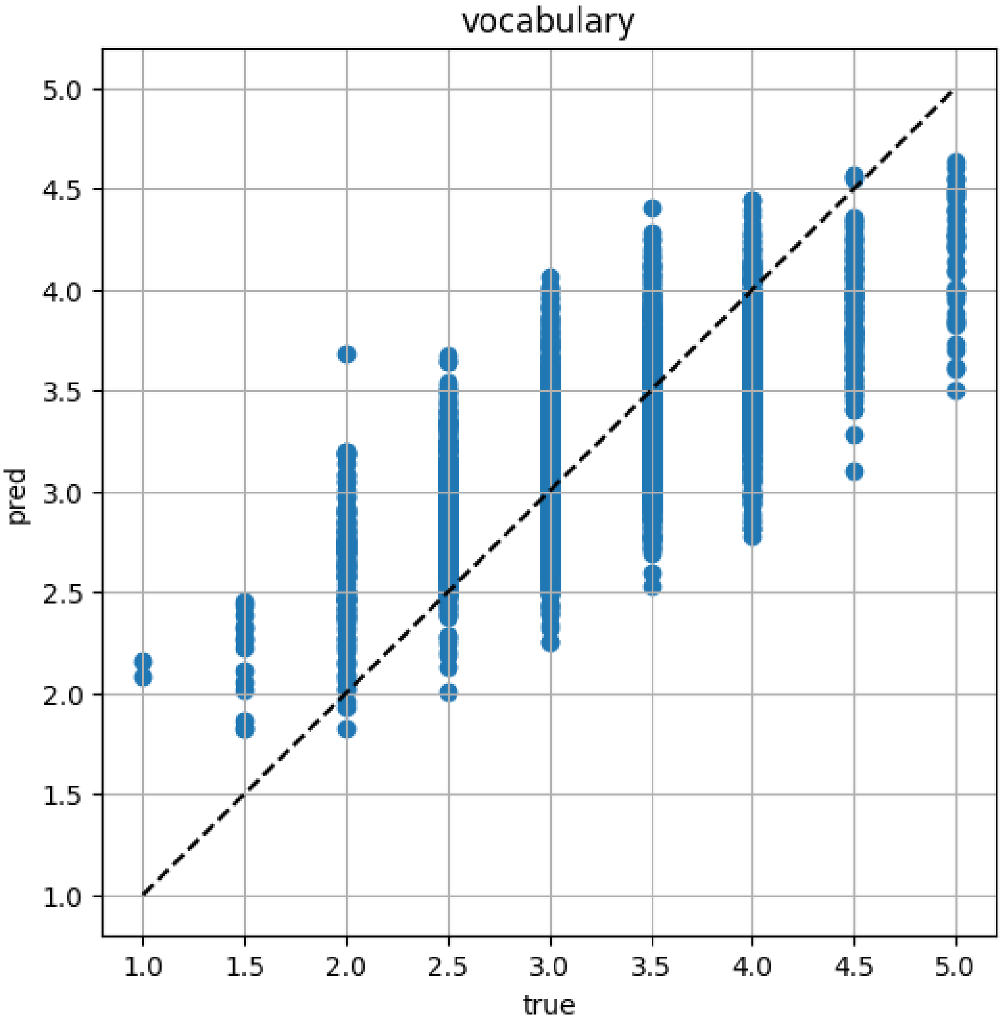
```
      pred_cohesion  pred_syntax  pred_vocabulary  pred_phraseology  \
1209       2.021610     2.308577         2.081938          2.103876   
3342       3.528046     3.398105         3.702724          3.567436   
1796       3.677545     3.665245         3.715463          3.760916   
2111       3.241116     3.227941         3.103109          3.158148   
2430       3.690733     3.689462         3.812043          3.839446   
…               …           …               …                 …   
1153       3.027940     2.911342         3.021968          2.925257   
3304       2.018000     1.883463         2.384690          2.099365   
3403       3.579203     3.467394         3.523069          3.465259   
2774       3.552495     3.452317         3.565363          3.492634   
1014       3.557043     3.492122         3.557395          3.593375   

      pred_grammar  pred_conventions      loss  
1209      2.330614          2.360804  1.201237  
3342      3.467285          3.322876  1.168921  
1796      3.673338          3.607192  1.099950  
2111      3.430276          3.348485  1.081821  
2430      3.675879          3.707826  1.069231  
…              …                 …       …  
1153      2.923754          2.811182  0.079729  
3304      1.931714          2.059850  0.079558  
3403      3.352793          3.618948  0.072629  
2774      3.261052          3.487999  0.070643  
1014      3.528026          3.679004  0.070454  

[3911 rows x 16 columns]
```

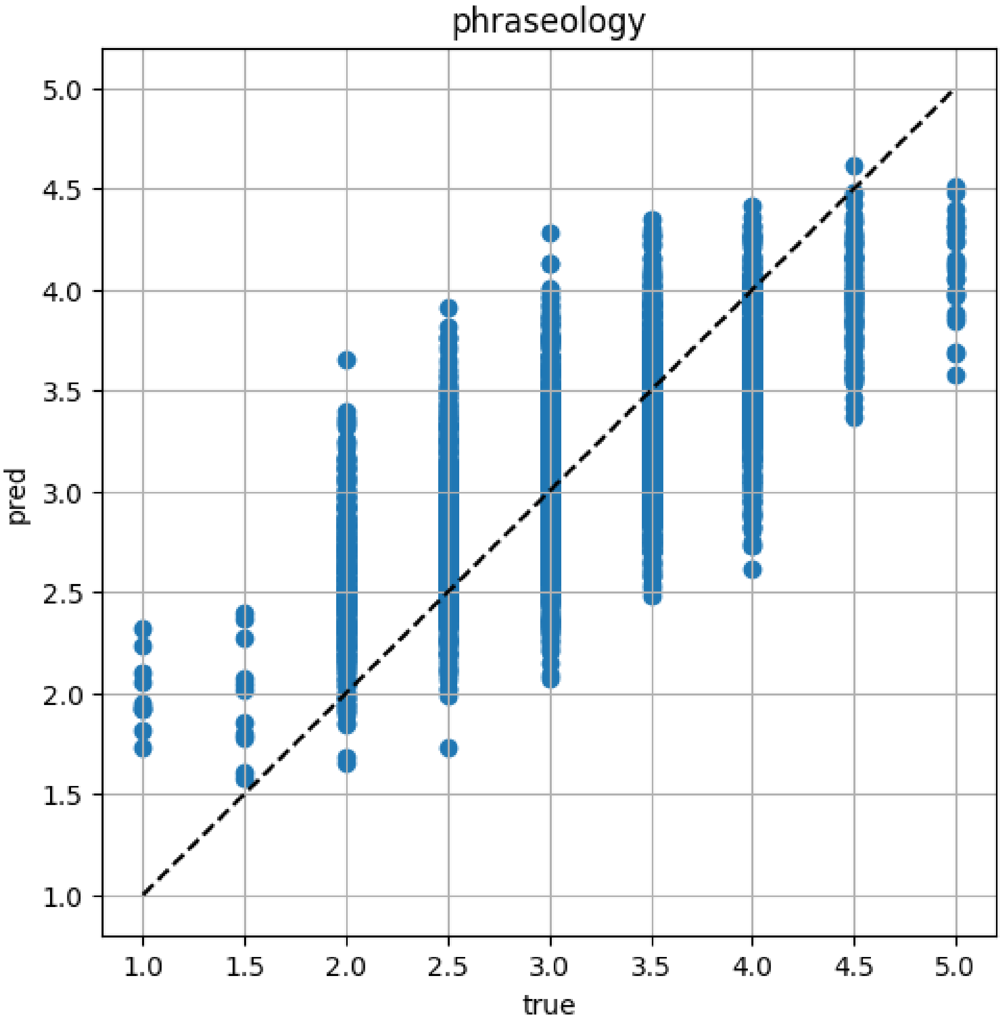
```
[ ]:
```

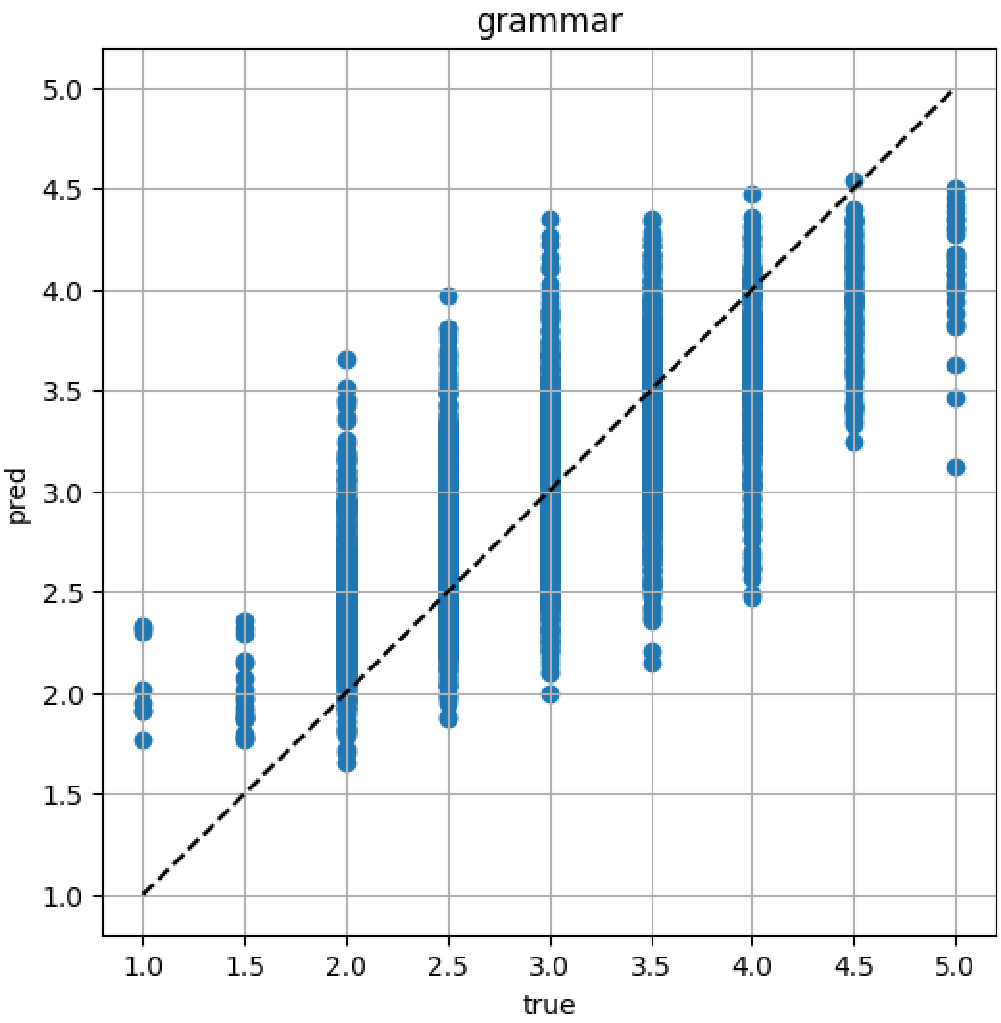
```
[ ]: for colname in true_colnames:
         fig, axs = plt.subplots(figsize=(6,6))
         axs.scatter(x=df[colname], y=df["pred_"+colname])
         axs.set_xlabel("true")
         axs.set_ylabel("pred")
         axs.plot([1.0, 5.0], [1.0, 5.0], linestyle="dashed", color="black")
         axs.set_title(colname)
         axs.grid()
```

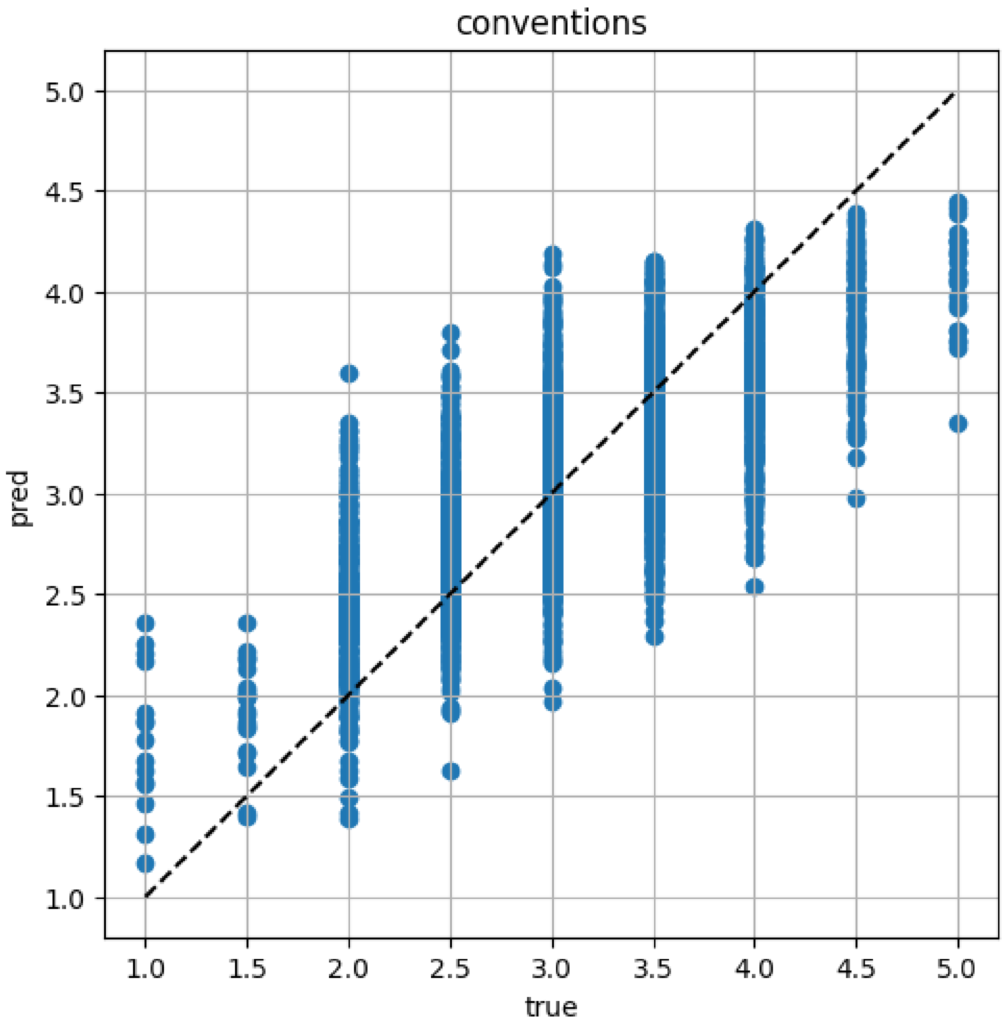

cohesion
pred
true
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0


syntax
pred
true
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0


vocabulary
pred
true
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0


phraseology
pred
true
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0


grammar
pred
true
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0



[ ]:

```
```

[ ]:

```
import os
```

[ ]:

```
os.makedirs("oof_1", exist_ok=True)
for idx, example in df.sort_values("loss", ascending=False).head(10).
 ↪reset_index(drop=True).iterrows():
    with open(f"oof_1/{idx}.txt", "w") as f:
        print(f"text id: {example.text_id}\nloss: {example.loss}", file=f)
        print(f"\ntrue: \n{example[true_colnames]}", file=f)
        print(f"\npred: \n{example[pred_colnames]}", file=f)
        print(f"\ntext:\n{example.full_text}", file=f)
```

[ ]:

```
```

```
[ ]:
```