openAI - base evaluation

👥 Assignee	
🔅 Status	Done
🗓 Due	
⬛ Project	Mark My Work
↗ Parent-task	try public api's on questionaire
⊙ Priority	
⬛ Sub-tasks	produce more sophisticated FlowChart, generate diagram for chatGPT access, convert excel, try text, figure statistics on filled questions
⬛ Tags	

initial evaluation/ 22.4.:

- ~~reduced to to essays < 3400 tokens = 52 essays, 13 questions corrected~~ got all essays now - mixed character vs token lengths :|

essay lengths (x), count (y)
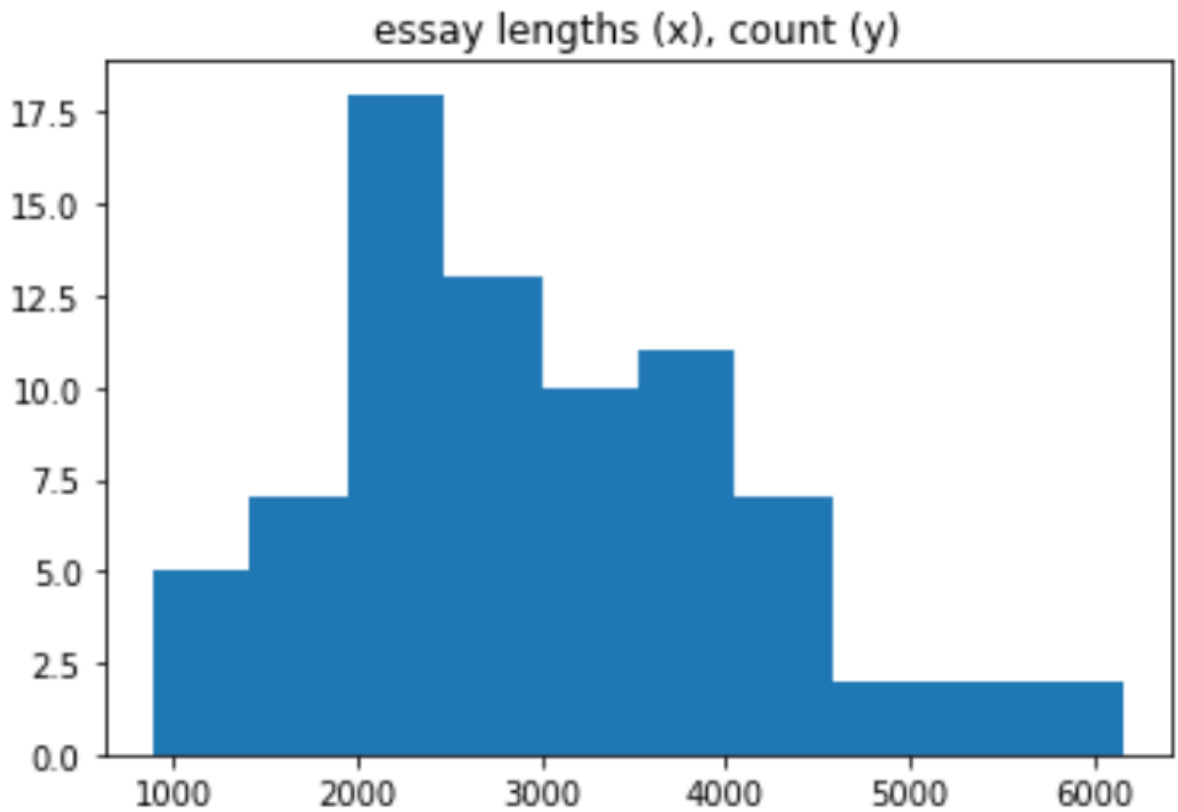
- ⇒ 676 requests with text-davinci-003
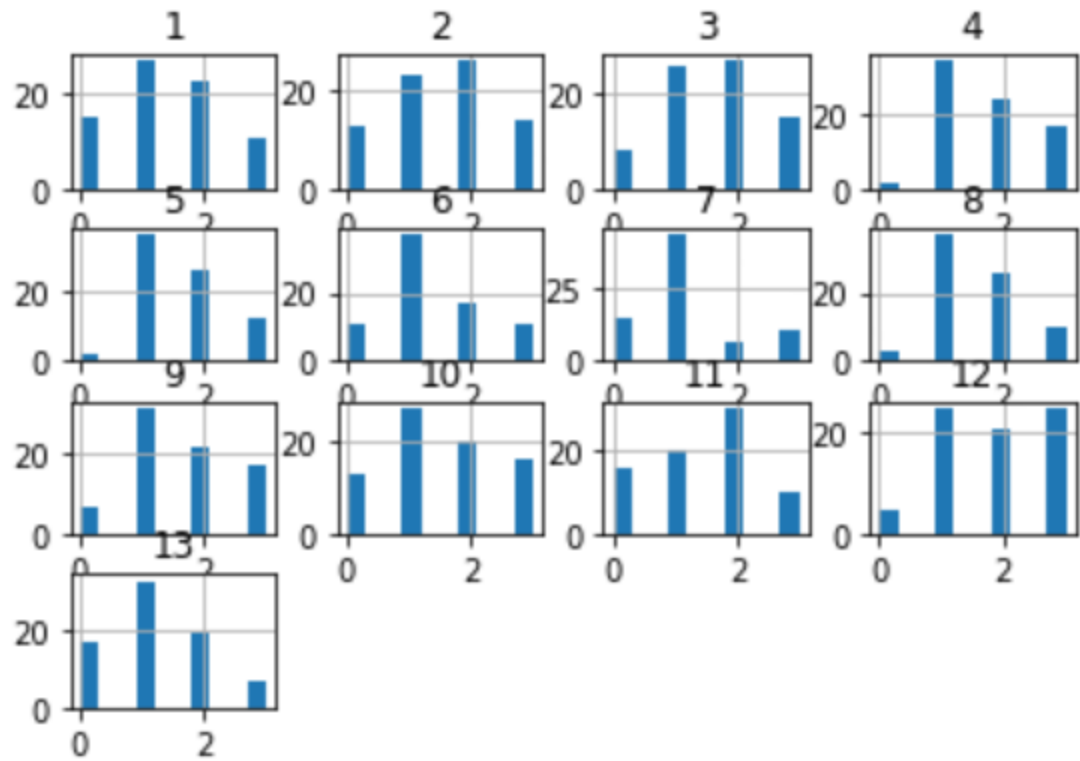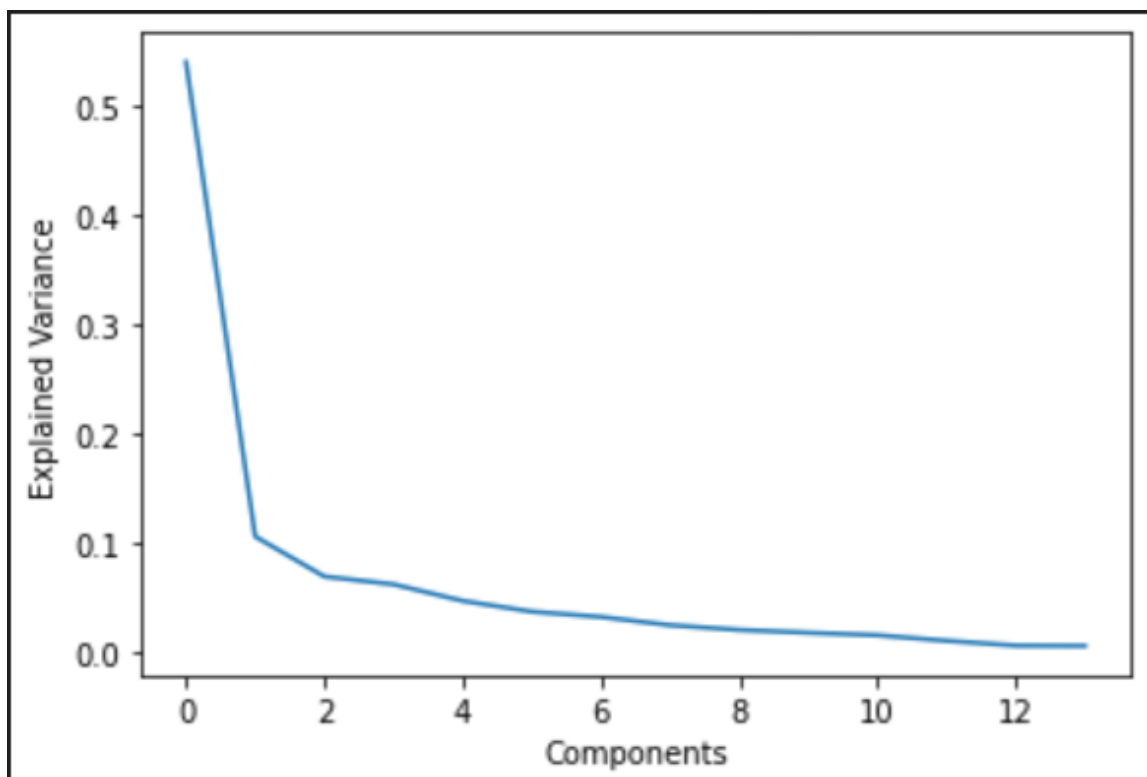
  - took 5 minutes

  - cost 12$

  - 7seconds, 23ct per essay

▼ **Qualitative questions evaluation:**

Annotated Question value distribution over all essays:

PCA variance of features:

so around 8 questions contain 90% correlation → curve could be a bit even but looks okayish for now. wonder how this influences the final 'generate rating'

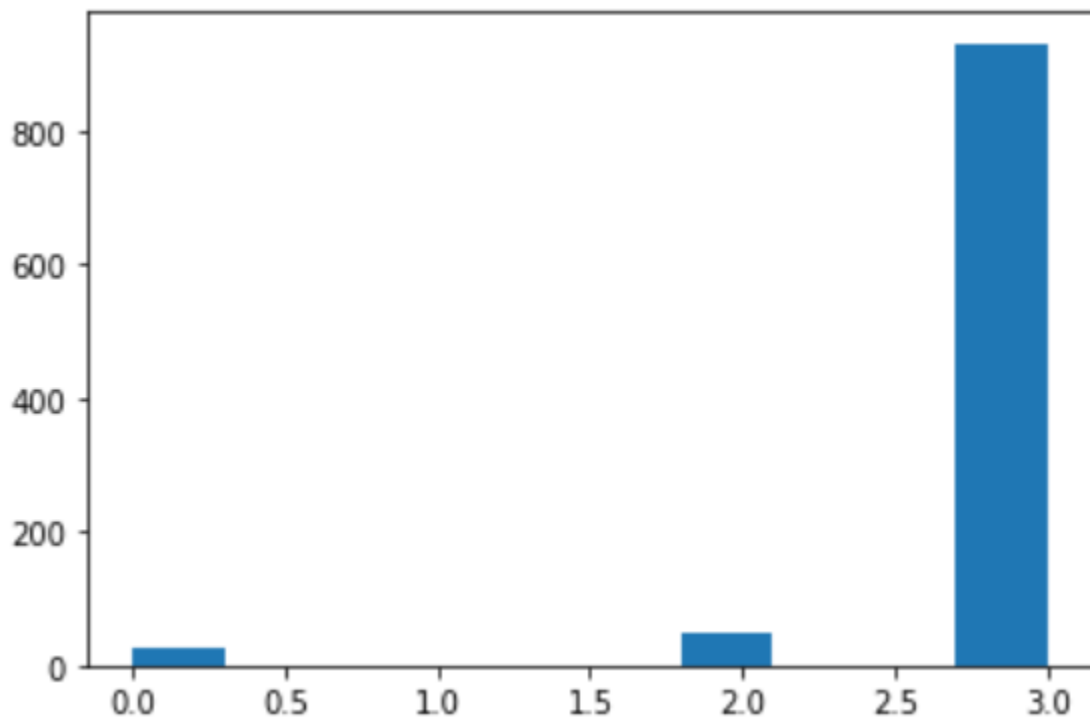Some questions could be done with classical evaluations:

Has the student varied their sentence length?
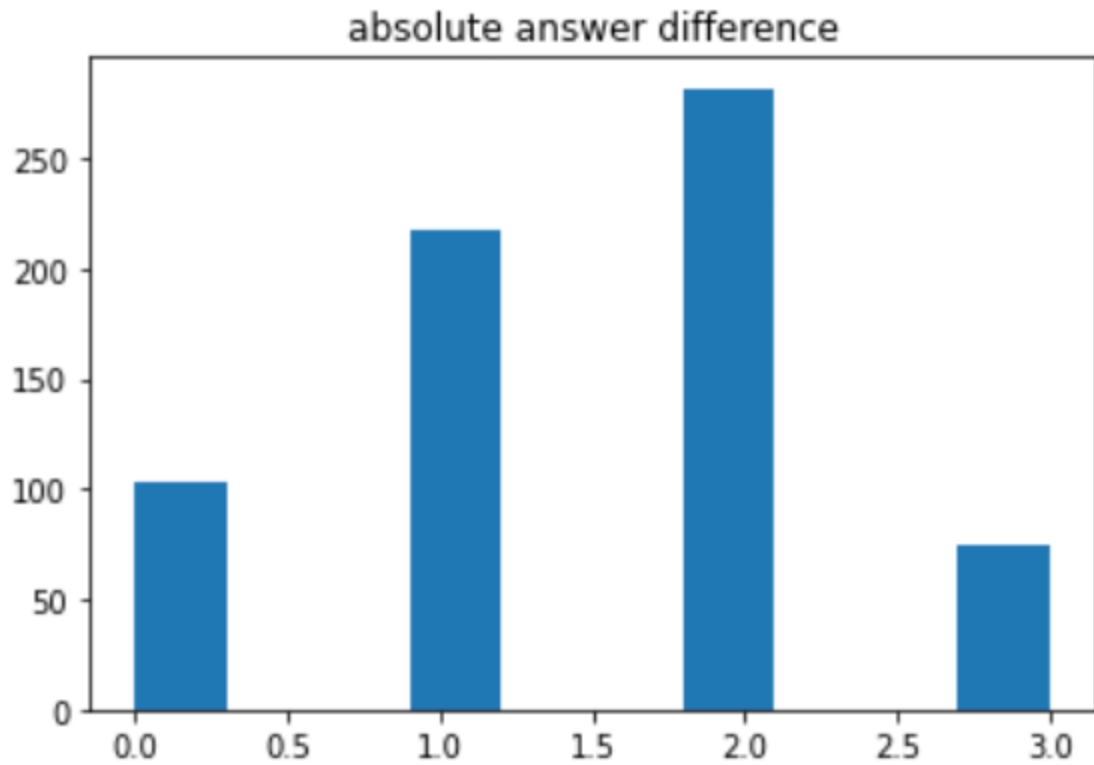
Some questions I don't get:

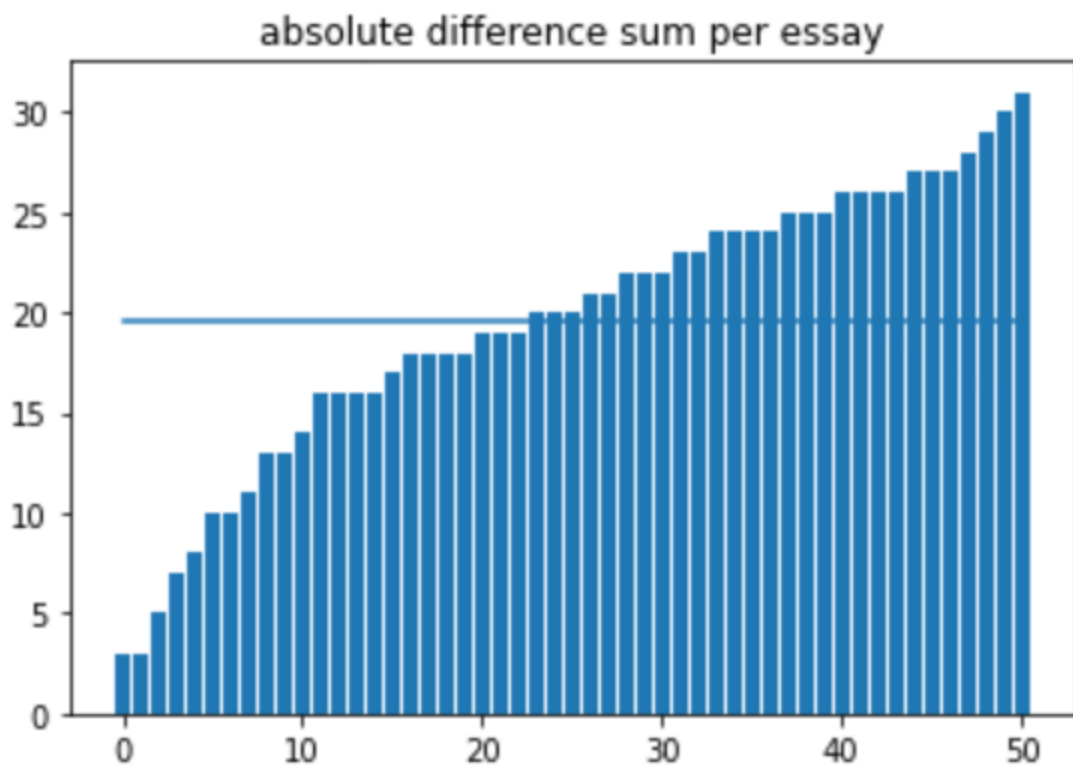*Has the student engaged with the prompt?*

▼ **Quantitative results**

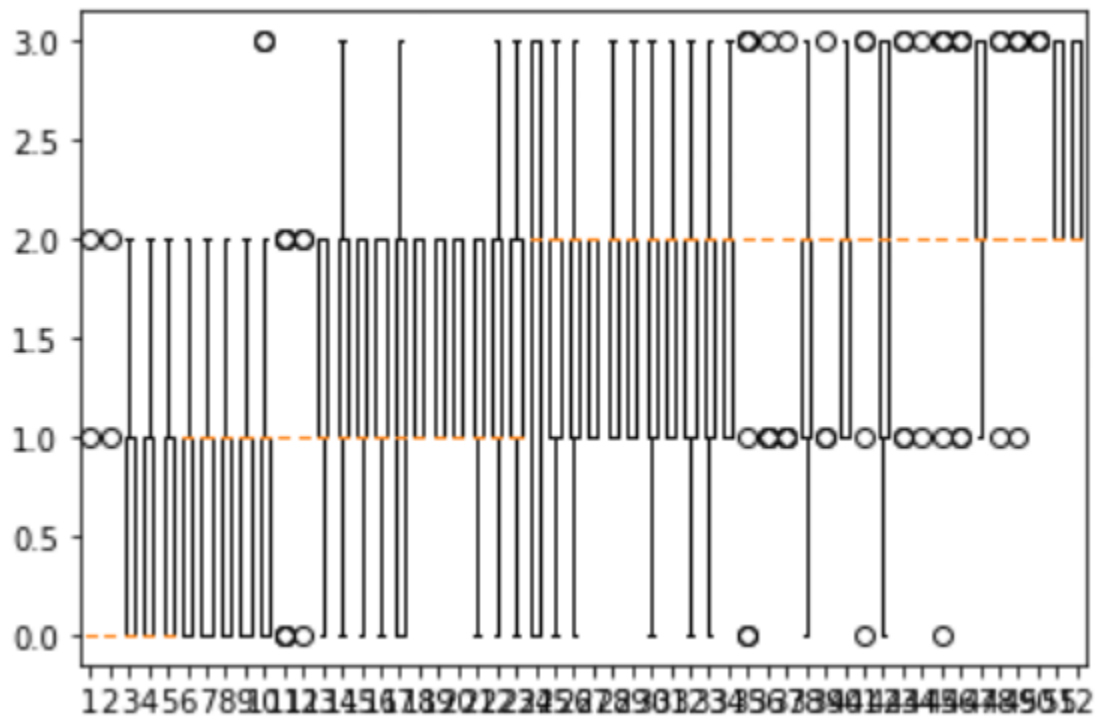first evaluations: got an always 3 predictor :/
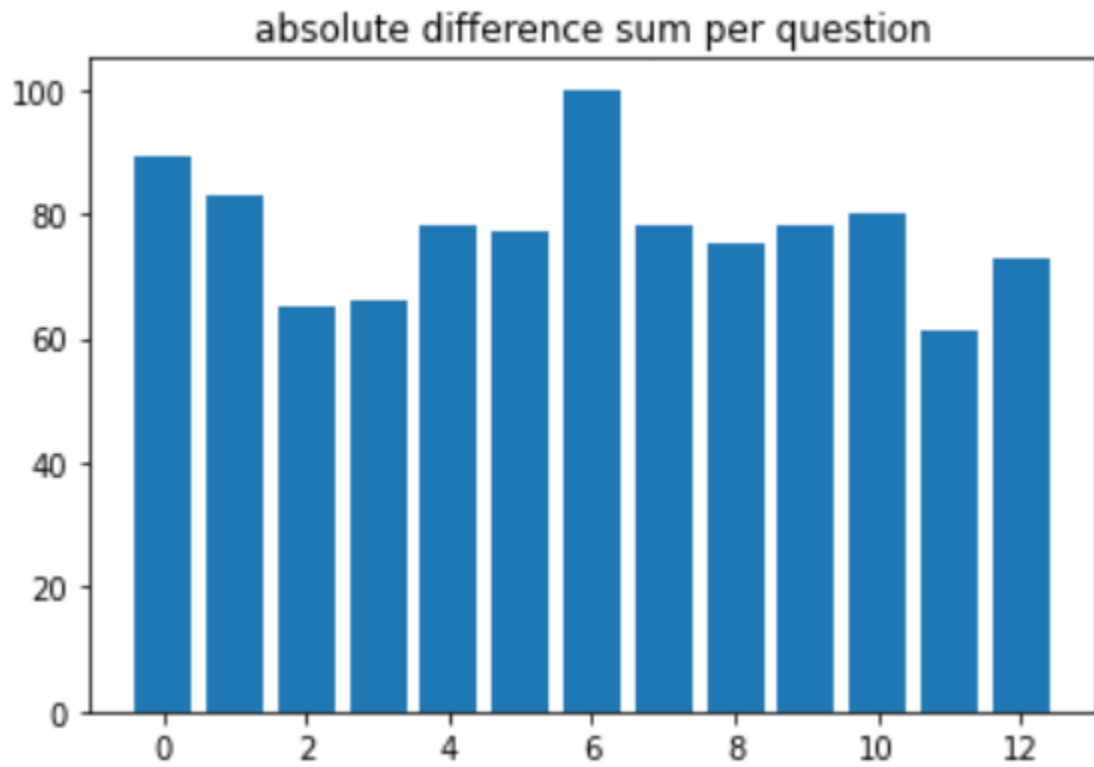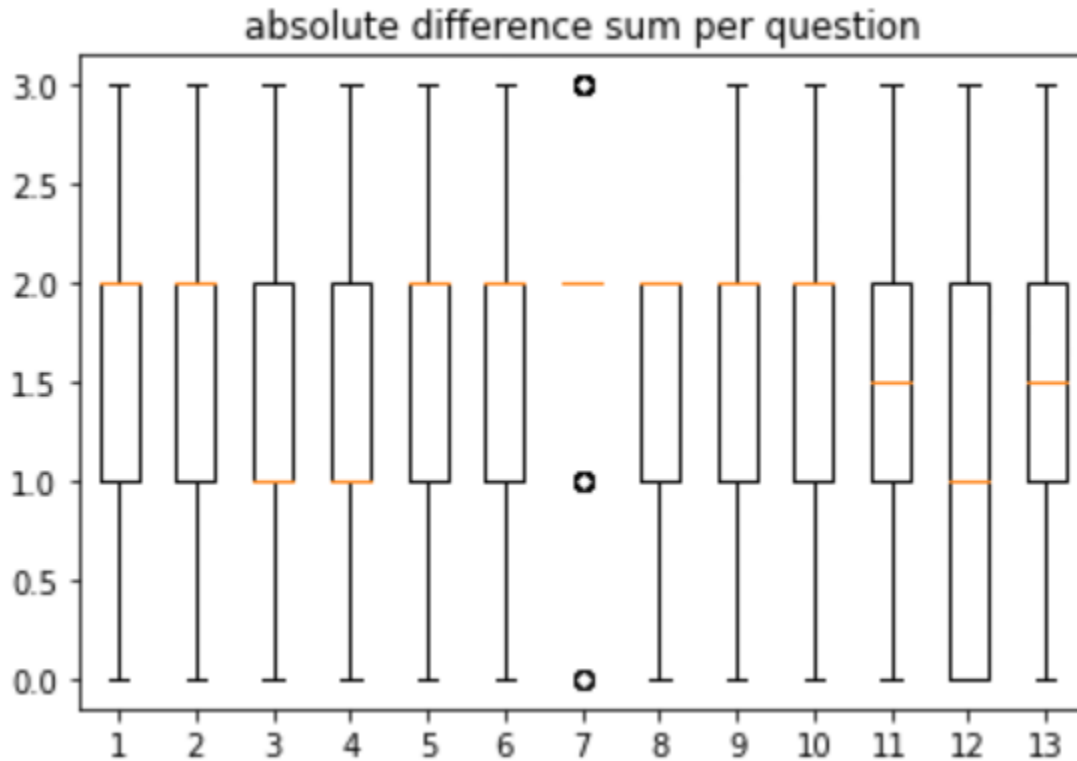

aggregated over all essays and questions:

absolute answer difference

aggregated over all questions:

absolute difference sum per essay

aggregated over all essays:

absolute difference sum per question

absolute difference sum per question

3,4,11,12 work slightly better, 7 always wrong

3: Has the student used a range of literary devices?

4: Has the student made an effort to create dynamic characters?

11: Has the student followed the rules of using dialogue? (1. speech marks 2. new speaker new line 3. punctuation before closing speech marks 4. speech tags).

12: Has the student used paragraphs?

7: Has the student varied their sentence length?

further findings:

it is currently not possible to add multiple organizations e.g. to split billing. one needs a phone number per account.

https://community.openai.com/t/adding-new-organizations-to-your-profile/165812