









openAI fine tuning

 Assignee	
 Status	In Progress
 Due	
 Project	<u>Mark My Work</u>
 Parent-task	<u>try_public api's on questionnaire</u>
 Priority	
 Sub-tasks	<u>finetune all questions one by one, finetune one model answering all questions, check duplicated rows that prepare dataset proposed, check davinci instead of ada, integrate question explanations, finetune/eval on new prompt style, try one model per question</u>
 Tags	

annotated dataset split into 777 train vs 195 validation essay/question pairs.

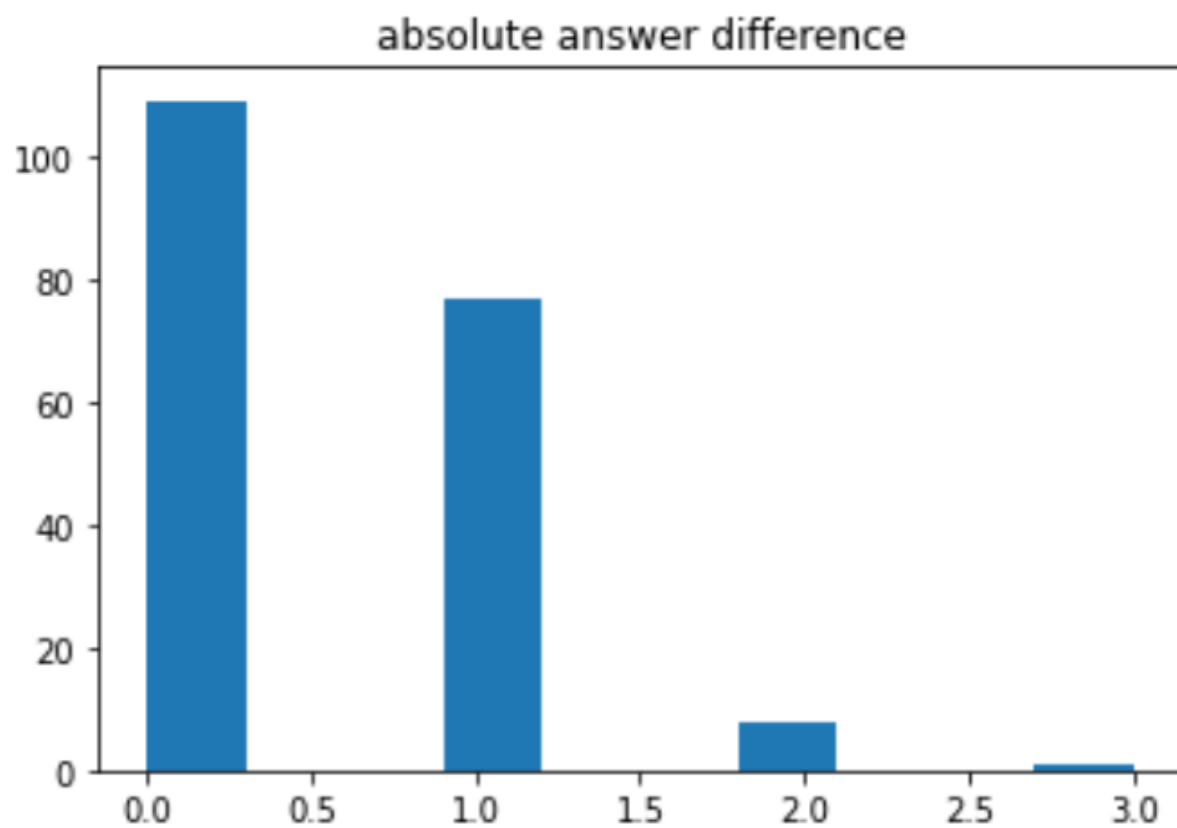
▼ eval on finetuned model, ada #1:

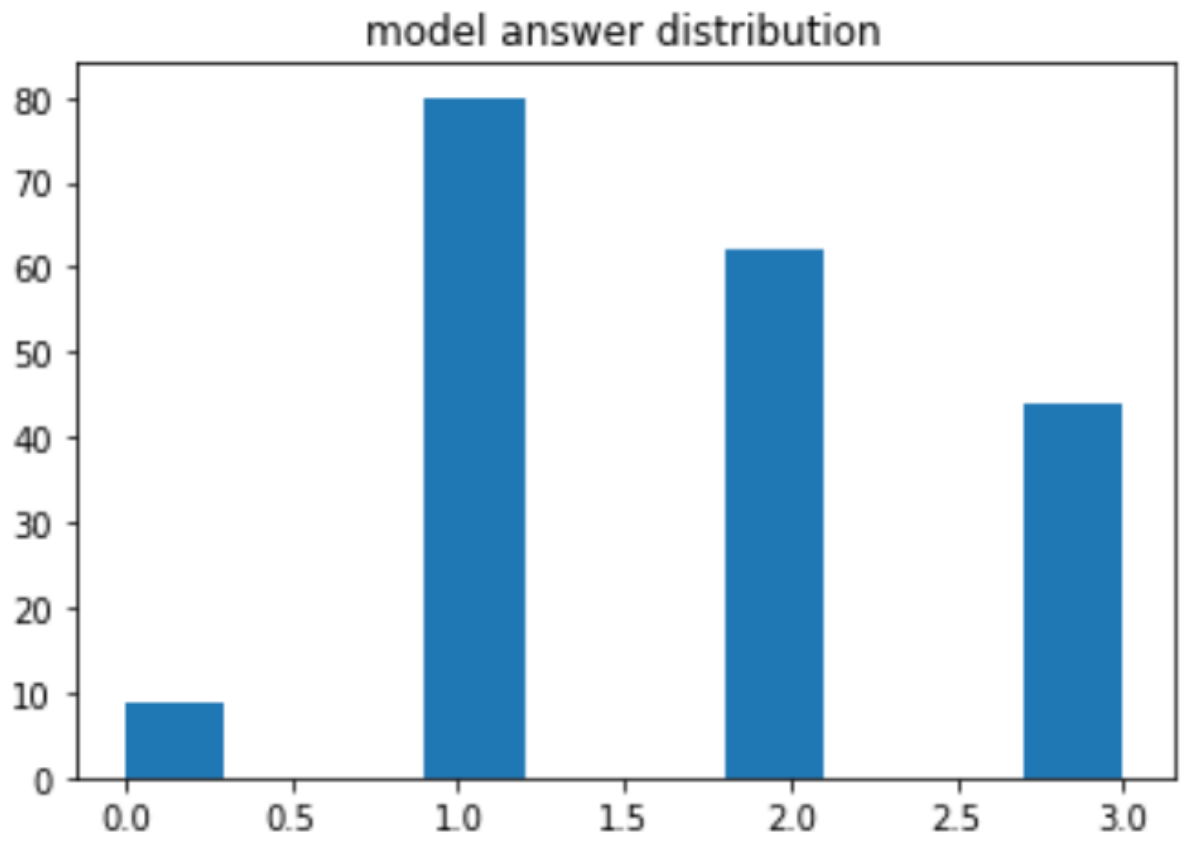
much better performance already.

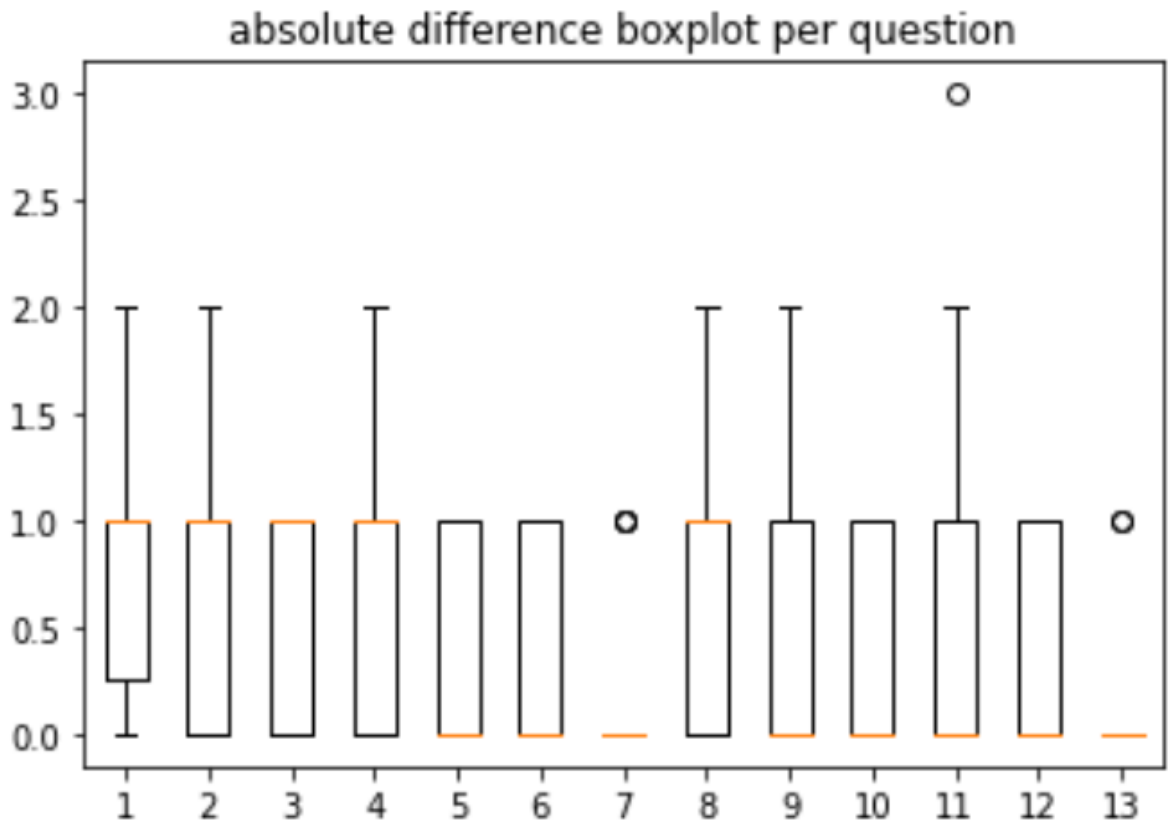
56% correct, $95\% \leq 1$

all medians are 1 or below.

some questions seem to work better.







- fft-ada looks quite usable already. 95% are less than or equal 1 score point apart. its also cheap for production use. one model per question could be further beneficial, no large extra cost here.
- adjusted prompt style, included question descriptions.
- text-davinci still can't comprehend out of the box. (still 'always 3 predictor')

▼ **eval on finetuned model, davinci #2:**

Fine-tune costs \$76.08

94% ≤ 1 , no big difference

