Grid view

#	Date	Done	Feedback (TODO)	Meeting type
1	8…	no SQ data from Covidence splitting type doesn't make a huge difference conflation makes RoB2 results comparable to RoB1 towards transfer learning: train on RoB1+2, eval on RoB2: results too good (this is rando…	dig into unexpectedly high performance in mixed regime splits effect get new unseen data from Cochrane web train 7B models and check …	Kari…
2	1…	submitted the paper mixed+splity_by_study_ids not as good as mixed+random_splits: only some domains are good but others ~ random. Why? mixed model on RoB1: results are lower compared to when just FT without mixing…		Stre…
3	2…	submitted again to RSM new 2024 dataset to have a clean test test for RoB2 680 instances, 48 studies, 5 reviews	try in-context learning again but with models allowing longer inputs. Also: SQ models run eval on new 2024 test data run eval on new 2024 test data with longer input (change durin…	Kari…
4	2…	submitted abstract to Global evidence summit in Prague eval on new 2024 test data true fewshot with GPT4 and long context window (show slide): fewshot not working as well as zero shot. Zero shot in a few domain…		Stre…
5	5…	submitted abstract to Global evidence summit in Prague eval on new 2024 test data true fewshot with GPT4 and long context window (show slide): fewshot sometimes outperforming zero-shot, sometimes not. …	control the size of train set for PEFT: have 3 train sets, rob1, rob2, and rob1+2, then evaluate on rob2 only. Do multiple samplings + PEFT trainings. rob2-test vs rob2-2024-test diff: …	Kari…
6	1…	found a possible reason why FT results were inflated: some test instances were the same: included separately in Cochrane data, have the same outcome mention, but different P/I same-size training sets to better control FT for RoB1, RoB2, and mixed…	wait for same-size train FT results to finish	Stre…

#	Date	Done	Feedback (TODO)	Meeting type
7	1…	same-size train FT results: rob2 and mixed got lower results overall. We now also include rob1 as training. Why rob1 so good? --> sampling bias for rob2/mixed? try different training samples. ignoring input text that also occurs in train …	any response from Edinburgh? confidence of model predictions --> work together with humans do sample train experiments how to do error analysis? (compare model and human justifications)…	
8	3…	article reviews: ○ data availability ○ additional analysis of SQ accuracy: more thorough description of the failure contexts / some more manual insights? focus on the link to model certainty?…	article revision	
9	9…	article revision: ○ simpler instructions model exps running: ○ data release prepared mixtral: ○ lots of gibberish output	sq analysis: use a stronger model?	
10		added some SQ analysis/interpretation based on manual checking correlation between logprob and answer correctness not clear or very weak for Meditron SQs (few data though!) Simple instructions for SQ-level task. SQ eva…	presentation /reorg for Rev1 next week probably revision ready? Simple instructions for SQ-level task: artificially truncate to the size we had with DirectModel wi…	
11		finalising our revised article and response to reviewers finalised Simple prompt experiments gpt4o; gold sq best perf some exps finished involving sampling effects in FT w/ different training sets: large when …		
12		article revision, submit further? global evidence summit, applied a while ago when I thought my extension will be until oct, what to do nature reviews: Methods Primers should I include only links to our work? (or al…		