Reflections on the experiments in following papers

1(e-SNLI)-natural-language-inference-with-natural-language-explanations-Paper

2(post answer explanation in few shot improves acc of answer without explanation generated at test time)Can language models learn from explanations

3(force to generate sometimes false explanation won't help predicting)The Unreliability of Explanations in Few-shot Prompting for Textual Reasoning

4(mislead irrelevant prompt)Do prompt-based models really understand the meaning

5(e-hans)Investigating the effect of natural language explanations

6 Rethinking the role of demonstration: what makes In-COntext learning work?

|  | 1 | 2 | 3 | 4 | 5 | 6 |
| --- | --- | --- | --- | --- | --- | --- |
| Explanation, prompt, or demonstration tested |  | Few shot,  Then instruction,  Then few shot with explanation | Explanation+prompt, few shot, with around 6-16 shots |  |  | Only demonstration |
| Tested task |  | writing codes, playing competitive games and common-sense reasoning to social bias, linguistics, software development and beyond | NLI  Focus on textual reasoning |  |  | sentiment analysis, paraphrase detection,  natural language inference, hate speech detection,  question answering, and sentence completion  (all are multiple choice or classification) |
| Dataset or benchmark tested on |  | Big bench | SYNTH  E-SNLI |  |  | Pg 17 appendix |
| Model |  | Gophter (good at big bench but not perfect so it’s a good choice for this experiment) | GPT,  Instruct GPT,  Davincii2  OPT |  |  | 12 version of gpt, models shows no huge difference in this experiment, such as GPT-J |
| Few shot or fine tune |  | Few shot | few |  |  | Few shot |
| Other settings |  | P-E | (prompt is not the same for each question) |  |  | No instruction or prompts or explanations,  Just input and labels pairs provided, and predict the testing label based on testing input |
| Baseline |  | Zero shot | No explanation, just few shot |  |  | No demonstration, so just zero shot |
| Experimental group  And results |  | Few shot improves on zero shot  Untuned (random chosen written) correct explanations improves on few shot | E-P  Or  P-E  Limited improvement on the result of baseline |  |  | Demonstration with correct label, or random label  , got the same result |
| Control groups and results |  | Few shot with scrambled explanations, or True but non-explaination explanation won’tc change the result that much  Totally irrelevant explanation does hurt acc  Carefully selected good explanations does improve on untuned explanations  Instruction with few shot better than few shot |  |  |  | Out of distribution input, label, or no label but just input, no correct input just possible label  All of these hurts the result more than random label  As # of shots goes up for correct demonstrations, the result is not getting better after k>=8  Reversed label mentioned, hurting the result |
| Possible explanations |  | Instruction is always helpful, but we don’t know how from this experiment  Scrambled or irrelevant explanation is not as helpful as the actual explanation during few shot,  So at least length for computation on the final result is not as influential as irrelevant text: irrelevant explanations can’t be ignored (compare this with paper 4, maybe change the misleading prompts to different ones from a set of misleading prompts will hurt the model performance more because the model can’t learn to ignore such misleading info)  Scrambled explanation is not helpful, so we can conclude that word distribution is also not as effective as the text mean of the explanation  But the experiment is P-E, so it makes sense that scrambled exp is not helping because there’s no effect from learning the distribution of the exp if we are going to first predict the answer without generating exp. So is the second point in this section, the length of the exp during few shot have nothing to do with the test time prediction without exp.  Since it’s under P-E setting, hopefully the few shot process is actually only changing the way the model understand the input label mapping | Few shot helps model to be familiar with input, output distribution, which already improved the performance a lot;  Few shot with explanations, hoping to change the way the model understand NLI input label mapping itself,the number of example might not be enough to accomplish this goal for some dataset,    Also the explanation generated is wrong at a lot of times  (But paper 2 experiment did improve the acc on average on big bench with PE setting,in PE setting few shots the answer doesn’t depend on the explanation that will be generated)  Textual reasoning is not as easy to learn from explanation as math problems.  The explanation provided during few shots is short. Not every sentence in the input is useful. Unlike a math problem in which all the steps have the corresponding line of input of certain important information. |  |  | Models learn from few-shot by learning or getting familiar with the distribution of label, input, or the data format, but not label-input mapping.  Since the result of random label group is still good, maybe it proves that the LLM already has the ability of zero shots on these task, so getting familiar with the data format within 8 shots is good enough; and the model’s ability in these task is robust enough that the random input label mapping can’t hurt the result |
| Other comment |  | Larger the model, better these improvements are | The correlation between explanation factuality and the truthfulness of the final answer is high; correlation between the truthfulness of the answer and the consistency of the explanation is not as high though.  Davinci performs much better than gpt-3 |  |  | The result of this article suggests a good way doing zero -shot with better result: feed in examples looks like real examples, no need for correct labels |
| Limitations |  | Difficult to measure how responsible each condition is on the improvements made.  The control groups might be meaningless if the P-E form of few shot can’t be effected at all by the scrambled or irrelevant exp; but we don’t know whether or how does these conditions affect the model | The model’s prompt for each question is different, the abel distribution is also different  Also, this experiment is done on only three dataset, so the result might be quite counter intuitively just because of these uncommon settings  And of course one more important reason is that the explanation generated is wrong most of the times; for gpt-3 the acc rate for explanation is just above 0.5 |  |  | The result is evaluated and averaged on 26 datasets, so the result above is not describing a general or common result for all 26 datasets but a result for most of them.  E.g some task might not need any examples to know the data format better so the random mapping will only hurt the result |