A Appendix

A.1 Proxy Model Task Performance

Table 5: Proxy models performance on the target tasks with and without fine-tuning.

	GPT-2		BERT	
Experiment Tasks	Pre-trained	Fine-tuned	Pre-trained	Fine-tuned
Snarks	38.8	47.2	30.5	38.8
Causal Judgment	44.7	55.2	44.7	52.6
Ruin Names	07.8	26.9	10.1	22.4
Formal Fallacies	50.5	54.4	51.6	53.5
Salient Translation Error Detection	14.0	27.1	11.5	22.6
CommonsenseQA	07.4	29.1	08.8	26.9
Coin Flip	45.2	59.4	51.1	59.7

A.2 Qualitative Analysis

Figure 3 shows an example from CommonsenseQA where GPT-3.5 responses using AO and CoT prompting yield an incorrect answer. The most likely reason for this is that these prompt strategies don't seem to capture all the key points of the input sentence, i.e., the context in the input is based on eyes rather than the overall body. However, this crucial detail is captured when GPT-3.5 is prompted with `AMPLIFY`. We observe that the GPT-3.5 response is correct, and it acknowledges "eyes" as the most important clue in making the correct prediction.

A.3 Hyper-parameter Analysis

Recall that `AMPLIFY` has two other primary hyper-parameters apart from the rationale template choice discussed in our empirical findings, namely, s, which is the size of the few-shot prompt created for LLMs, and k, which is the number of most important tokens identified by the post hoc explanation. Table 6 shows the LLM performance variations for different combinations of (k, s). It is important to note that `AMPLIFY` does not have scalability constraints with increasing s and k, as `AMPLIFY` computes prompts automatically. This is unlike CoT, where increasing the size of the few-shot prompt would require more human effort to generate relevant chains of thoughts.

A.4 Impact of BERT as Proxy Model on LLM Performance

Table 7 shows LLM performance when BERT is used as the proxy model in step 1 of `AMPLIFY`. We observe similar trends as those observed for the case of GPT-2, where fine-tuning proxy model provides marginal improvements in general. This indicates that the fine-tuning step could be avoided in most cases to reduce additional computational overhead.

B Limitations and Broader Impacts

Our work proposes a new framework, `AMPLIFY`, which focuses on improving the task performance of LLMs by injecting automatically generated rationales. This framework results in the reduction of reliance on processes that require heavy human intervention. These processes, which rely on rationales based on human annotations, often suffer from noise and biases, which may transfer to LLMs during in-context learning. We hope that automated rationale creation will provide a solution to mitigate this problem. While our approach provides significant improvements in model performance, the broader negative impact pertaining to LLMs, such as safety concerns in the form of misinformation[2], social bias[2], hallucination[12], etc., and the massive carbon footprint due to heavy usage of LLMs [17], may still persist even when using our proposed framework. Other than the limitations of LLMs, our framework relies on post hoc explanation methods to create automated rationales; hence, `AMPLIFY` may also inherit widely studied issues with post hoc explanations such as robustness[14], the disagreement problem[13], stability[26], etc.

Table 6: This figure shows LLM performance for the different selections of k and s hyper-parameters of AMPLIFY, as denoted by (k, s) for each column. In general, we observe (k = 7, s = 10) achieves the best results for most of the datasets.

	GPT-3 (k,s)				GPT-3.5 (k,s)			
Experiment Tasks	(2, 5)	(5, 5)	(5, 10)	(7, 10)	(2, 5)	(5, 5)	(5, 10)	(7, 10)
Snarks	63.8	72.2	80.5	80.5	75.0	80.5	**91.6**	88.8
Causal Judgment	52.6	57.8	60.5	60.5	65.7	73.6	76.3	**76.3**
Ruin Names	64.0	75.2	76.4	**78.6**	73.0	75.2	77.5	77.5
Formal Fallacies	55.5	57.9	59.8	**59.8**	56.3	58.8	59.6	59.6
Salient Translation Error Detection	49.7	50.2	51.2	51.2	52.7	56.2	60.8	**60.8**
CommonsenseQA	72.8	73.1	73.3	73.5	76.0	76.7	77.6	**77.9**
Coin Flip (OOD)	64.9	65.3	65.7	**65.7**	63.3	65.0	65.3	65.3
All Tasks *(avg)*	60.4	64.5	66.7	67.1	66.0	69.4	**72.6**	72.3

Table 7: Few-shot prompting performance of multiple LLMs on the seven datasets when post hoc explanations, which form the rationale in the prompt constructed during step 4 of AMPLIFY, are generated using models with varying degrees of fine-tuning of the proxy model (BERT in this case). Here, "E" represents the number of epochs the proxy model was fine-tuned. "E = 0" indicates that the proxy model was used to generate post hoc explanations without any fine-tuning. The recorded performance in this table represents the percentage of test samples for which the LLM accurately predicted the true label.

	GPT-3			GPT-3.5		
Experiment Tasks	E = 0	E = 10	E = 200	E = 0	E = 10	E = 200
Snarks	66.6	72.2	72.2	80.8	80.8	**88.8**
Causal Judgment	50.0	52.6	57.8	71.0	73.6	**73.6**
Ruin Names	70.7	73.0	**73.0**	71.9	71.9	71.9
Formal Fallacies	56.2	56.9	**58.5**	56.7	56.9	57.8
Salient Translation Error Detection	50.2	51.2	51.2	56.2	59.2	**60.8**
CommonsenseQA	71.3	71.8	72.4	76.1	76.5	**77.4**
Coin Flip (OOD)	65.4	65.8	**65.9**	63.7	64.3	65.1
All Tasks *(avg)*	61.2	63.1	68.0	68.0	69.0	**70.7**

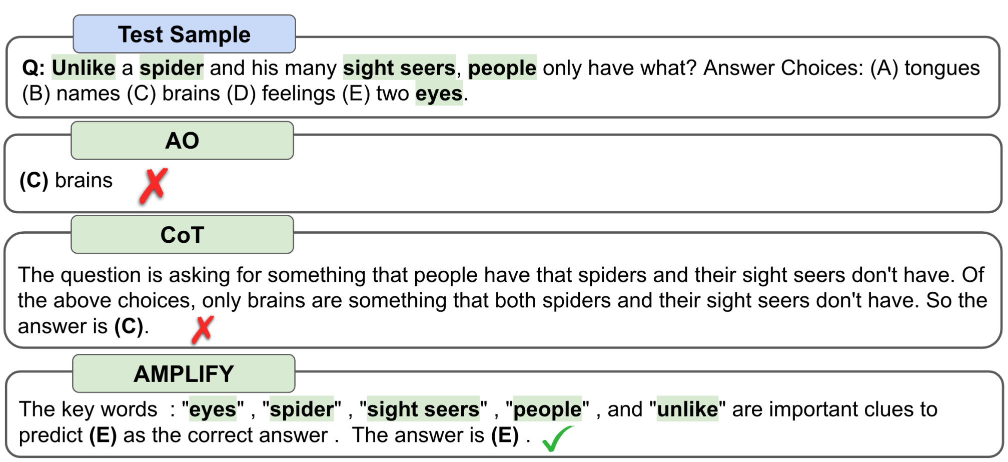

Figure 3: This image exemplifies an instance of CommonsenseQA task where standard prompts and CoT produce inaccurate responses. The CoT response fails to take into account the context in the question being related to eyes. In contrast, the response generated by AMPLIFY emphasizes this crucial detail.