The Unreliability of Explanations in Few-shot Prompting for Textual Reasoning

Anonymous Author(s)
Affiliation
Address
email

Abstract

Does prompting a large language model like GPT-3 with explanations improve in-context learning? We study this question specifically on two NLP tasks that involve reasoning over text, namely question answering and natural language inference. For these tasks, we find that including explanations GPT-3's prompt and having the model generate them only mildly improves accuracy over standard few-shot learning, contrary to recent results on symbolic reasoning tasks [30, 42]. Moreover, explanations generated by GPT-3 may not entail the predictions nor be factually grounded in the input, even on simple tasks with extractive explanations.

However, these flawed explanations can still be useful as a way to verify GPT-3's predictions post-hoc. Through analysis in three settings, we show that explanations judged as good by humans—those that are logically consistent with the input and the prediction—usually cooccur with more accurate predictions. Following these observations, we present a framework for calibrating model predictions based on the reliability of the explanations. We train calibrators using automatically extracted scores that approximately assess the reliability of explanations, which helps improve performance across three different datasets.[1]

1 Introduction

Recent breakthroughs in pre-training have empowered large language models to learn NLP tasks from fewer and fewer examples. In-context learning, learning a new task from just a few training examples in a prompt without updating the model's parameters, has started to show promising performance for very large language models like GPT-3 [3]. However, this learning process is still poorly understood: models are biased by the order of examples they are shown [53] and may not leverage the instructions or even the labels of the few-shot examples in the ways we expect [29, 41]. It is difficult to investigate these issues or explain the predictions of in-context learning when existing tools for interpreting model predictions have high computational cost [35] or require access to gradients [38, 40].

One appealing way to gain more insight into predictions obtained through in-context learning is to let the language model "explain itself" [30, 42, 9, 27, 23]. In addition to input-label pairs, one can prompt the language model with explanations for input-label pairs and expect the model to generate an explanation corresponding to the prediction it gives (Figure 1). Prompting with explanations introduces much richer information compared to using labels alone, which might guide the in-context learning process and allow the model to leverage more information about the examples.

In this work, we closely investigate the nature of the explanations that GPT-3 generates and whether they can really improve few-shot in-context learning, specifically for textual reasoning tasks. Recent prior work that finds success with this approach largely targets symbolic reasoning tasks with a very

[1] Code and data in the supplementary material.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

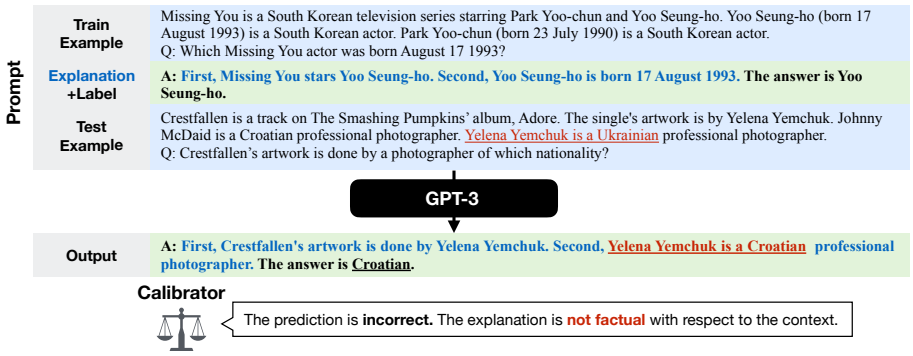

Figure 1: Prompting GPT-3 with explanations. By including explanations in the in-context examples, we can cause GPT-3 to generate an explanation for the test example as well. In this case, the generated explanation is nonfactual, despite the simple reasoning involved here. However, we show this nonfactuality actually provides a signal that can help calibrate the model.

different structure, such as math word problem solving [30, 42]. We show that explanations only mildly improve performance when plugged into the prompt (Figure 1) across three different datasets spanning QA and NLI.

Surprisingly, we find that the explanations generated by GPT-3 are **unreliable**, even for a very simple synthetic dataset. Specifically, we check the explanations along two axes: *factuality*, whether the explanation is correctly grounded in the input, and *consistency*, whether the explanation entails the final prediction. On realistic datasets, GPT-3 tends to generate consistent explanations that account for the predictions, but the explanations may not be factually grounded in the input context, as as shown in Figure 1. Furthermore, our analysis suggests an unreliable explanation more likely indicates a wrong prediction compared to a reliable explanation.

Despite GPT-3's failures here, we can still benefit from model-generated explanations by using them for calibration, given the connection between unreliable explanations and incorrect predictions. If we are able to automatically assess the reliability of an explanation, we can allow GPT-3 to return a null answer when its explanation is unreliable. Unfortunately, there is no automated way to perfectly assess the reliability, but we can extract features that approximately reflect it. We use these features to calibrate GPT-3's predictions, and successfully improve the in-context learning performance across all the datasets.

In summary, our main findings are: 1) Simply plugging explanations into the prompt does not significantly boost the in-context learning performance for textual reasoning tasks. 2) GPT-3 generates mostly consistent explanations, but these explanations might not be factually grounded in the inputs. 3) The reliability of an explanation can serve as an indicator for the correctness of the corresponding prediction. 4) Using features that can approximate the reliability of explanations, we successfully use explanations to improve the in-context learning performance across all tasks.

2 Does Prompting with Explanations Improve In-Context Learning?

In this paper we specifically focus on tasks involving reasoning over natural language. These are tasks where explanations have been traditionally studied [5, 33], but which are more complex than tasks like sentiment which are well explained by extractive rationales [50, 10]. We experiment on two tasks, reading comprehension question answering (QA) and natural language inference (NLI), on three English-language datasets. For each dataset, we create a test set with 250 examples.

2.1 Datasets

Synthetic Multi-hop QA (SYNTH) In order to have a controlled setting where we completely understand whether explanations are factual and consistent with the answer, we create a synthetic multi-hop QA dataset. As in Figure 2, each example in our synthetic dataset asks a bridge question (using the terminology of [46]) over a context consisting of supporting facts paired with controlled distractors. This dataset is carefully designed to avoid spurious correlations, giving us full understand-

SYNTH	**Context:**	Christopher agrees with Kevin. Tiffany agrees with Matthew. Mary hangs out with Danielle. James hangs out with Thomas. Kevin is a student. Matthew is a plumber. Danielle is a student. Thomas is a plumber.
	Question:	Who hangs out with a student?
	Answer:	Mary **Explanation:** Danielle is a student and Mary hangs out with Danielle.
E-SNLI	**Premise:**	A toddler in a green jersey is being followed by a wheelchair bound woman in a red sweater past a wooden bench.
	Hypothesis:	A toddler is walking near his wheelchair bound grandmother.
	Label:	Neither **Explanation:** the woman may not be his grandmother.

Figure 2: A SYNTH example and an E-SNLI example. See Figure 3 for ADVHOTPOT examples.

ing over the correct reasoning process as well as the explanation for every example, which naturally consists of the two supporting sentences. Refer to Appendix B for full details of this dataset.[2]

Adversarial HotpotQA (ADVHOTPOT) We also test on the English-language Adversarial HotpotQA dataset [46, 19] (license: MIT). We use the adversarially augmented version since GPT-3 achieves high performance on the distractor setting of the original dataset. We make a challenging set of examples by balancing sets of questions on which GPT-3 makes correct and incorrect predictions. The context of each question includes two ground truth supporting paragraphs and two adversarial paragraphs. Full details of preprocessing the ADVHOTPOT dataset can be found in Appendix C.

For ADVHOTPOT, we manually annotated explanations for the training examples. Figure 1 shows an example of this explanation, highlighted in orange. We could use the supporting sentences as the explanations, but we found they are usually too verbose and not sufficient, e.g., with anaphors that resolve outside of the supporting sentences. Therefore, we manually annotate a set of explanations which clearly describe the reasoning path for each question.

E-SNLI E-SNLI [5] is an English-language classification dataset (license: MIT) commonly used to study explanations. Shown in Figure 2, each example consists of a premise and a hypothesis, and the task is to classify the hypothesis as entailed by, contradicted by, or neutral with respect to the premise. As a notable contrast to the other datasets, the explanations here are more *abstract* natural language written by human annotators, as opposed to mostly constructed from extracted snippets of context.

2.2 Baselines

We study the effectiveness of plugging in explanations by comparing the in-context learning performance of prompting with or without explanations. Prompting without explanations resembles the standard few-shot in-context learning approach (**Few-Shot**). To incorporate explanations into the prompt, we consider the following two most commonly used paradigms:

Explain-then-Predict (E-P) which prepends an explanation before the label (Figure 1). The language model is expected to generate an explanation first followed by the prediction. The prompting style of past work involving computational traces can be categorized into this paradigm, including Nye et al. [30] and Wei et al. [42]. This approach is also called a pipeline model in other literature on training models using explanations [18, 45].

Predict-then-Explain (P-E) which generates the explanation after the prediction. Unlike E-P, the predicted explanation does not influence the predicted label, since we use greedy inference and the explanation comes afterwards. However, the explanations in the prompt still impact the predictions.

2.3 Setup

For few-shot learning, we use roughly the maximum allowed shots in the prompt that can fit the length limit of GPT-3, which is 16 for SYNTH, 6 for ADVHOTPOT, and 32 for E-SNLI, respectively.[3] Because the results of in-context learning vary with the examples presented in the input prompt, for each dataset, we randomly sample 5 groups of training shots, and report the mean and standard

[2]This dataset is inspired by the bAbI dataset [43]. In our preliminary experiments with some of the more complex bAbI tasks, we found poor performance from GPT-3 similar to our results on SYNTH, both with and without explanations.

[3]This contrasts with recent work like Zhao et al. [53] that focuses on improving performance in the 1-4-shot setting; by using more data we achieve much stronger results on our tasks.

Nonfactual	Pedro Rubens! The individual chapters were published into 64 "tankōbon" by Kodansha. Yōko Shōji (born 4 June 1950, in Mobara, Chiba) is a Japanese manga artist. She is best known for writing "Seito Shokun! Mulder Scully! The individual chapters were published into 14 "tankōbon" by Kodansha. Seito Shokun! The individual chapters were published into 24 "tankōbon" by Kodansha between. Q: How many chapters does Yōko Shōji's most famous manga have? A: First, Yōko Shōji's most famous manga is "Seito Shokun!". Second, "Seito Shokun!" has 64 chapters. The answer is 64.
Inconsistent	Tim Minchin (December 29, 1808 July 31, 1875) was the President of the United States. Andrew Johnson (December 29, 1808 July 31, 1875) was the President of the United States. George Andrew Atzerodt (June 12, 1835 – July 7, 1865) was a conspirator, with John Wilkes Booth. Jesse Andrew Williams (June 12, 1835 – July 7, 1865) was a conspirator, with John Wilkes Booth. Q: Who was older, George Atzerodt or Andrew Johnson? A: First, George Atzerodt was born on June 12, 1835. Second, Andrew Johnson was born on December 29, 1808. The answer is George Atzerodt.

Figure 3: Explanations generated for ADVHOTPOT. GPT-3 may generate nonfactual explanations containing hallucination (red) or inconsistent explanations contradicting the answer (red).

deviation of the results (subscript). All our experiments use the 175B GPT-3 [3] Instruct series API (text-davinci-001), the strongest available model at the time of our experiments.[4] The completion is obtained through greedy decoding (temperature set to be 0). Our prompt formats follow those in Brown et al. [3]. The explanations are inserted before/after the prediction with conjunction words like *because*. Please refer to Appendix A for full prompts.

2.4 Results

In general, using explanations mildly improves the performance for the text reasoning tasks, as show in Table 1. On the two QA tasks, SYNTH and ADVHOTPOT, E-P improves the performance from 54.8 to 58.5 and 56.8 to 59.4, respectively.[5] On E-SNLI, P-E outperforms FEW-SHOT by 2.6, whereas E-P substantially lags FEW-SHOT. There is no single winner between the two paradigms of plugging in explanations; choosing the most effective way is task-specific.

Table 1: Comparison between the in-context learning performance without and with explanations. Using explanations mildly improves performance.

	SYNTH	ADVHOTPOT	E-SNLI
FEW-SHOT	$54.8_{2.5}$	$53.2_{2.3}$	$56.8_{2.0}$
E-P	$\mathbf{58.5}_{2.1}$	$\mathbf{58.2}_{4.1}$	$41.8_{2.5}$
P-E	$53.6_{1.0}$	$51.5_{2.4}$	$\mathbf{59.4}_{2.0}$

Our results do not suggest immediate strong improvements from incorporating explanations, even for the simple synthetic dataset, contradicting recent prior work. This can be attributed to the difference between the tasks we study. The tasks that receive significant benefits from using explanations in Nye et al. [30] and Wei et al. [42] are all program-like (e.g., integer addition and program execution), whereas the tasks in this work emphasize textual reasoning grounded in provided inputs. In fact, in Wei et al. [42] and Chowdhery et al. [9], explanations only show mild benefit on open-domain QA tasks like StrategyQA [14] that are closer to our setting.

3 Can GPT-3 Generate Factual and Consistent Explanations?

Prompting GPT-3 with explanations and generating explanations does not lead to much higher performance on our tasks. But what about the quality of the model-generated explanations themselves? We assess the reliability of the explanations for the three datasets, measured in terms of two aspects.

Factuality refers to whether a generated explanation is faithfully grounded in the corresponding input context (context for QA and premise/hypothesis pair for NLI). A factual explanation should not contain hallucinations that contradict the context. See Figure 3 for a nonfactual explanation.

Consistency measures if the explanation entails the prediction. Our concept of consistency resembles plausibility as described in the literature [18], in that we assess whether the prediction follows from the explanation **as perceived by a human**. See Figure 3 for an inconsistent explanation.

For SYNTH, we uses rules to automatically judge whether an explanation is factual and consistent. For ADVHOTPOT and E-SNLI, the authors manually inspected the explanations and annotated

[4] We did not experiment with smaller models, as these are much worse at in-context learning [3].

[5] For SYNTH, we also tried using an alternative style of explanations (reversing the order of the two sentences in the explanations), which leads to mild performance degradation.

Table 2: Left: factuality (Fac) and consistency (Con) of the generated explanations. Right: the % of the examples whose explanation factuality/consistency is congruent with the prediction accuracy. In general, GPT-3 tends to generate consistent but less likely factual explanations.

	Acc	Fac	Con	Acc=Fac	Acc=Con
SYNTH (E-P)	58.4	72.8	64.8	66.5	68.8
SYNTH (P-E)	54.8	51.6	95.2	89.6	57.2
ADVHP (E-P)	62.0	79.6	91.2	80.0	68.4
ADVHP (P-E)	54.0	69.2	82.0	77.6	67.2
E-SNLI (P-E)	62.0	–	98.8	–	62.0

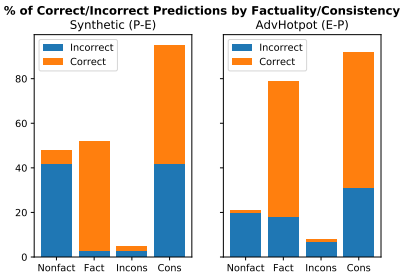

Figure 4: Nonfactual explanations usually indicate an incorrect prediction.

them for these two characteristics (Cohen's Kappa between these annotations: 0.89; more details in Appendix D). Note for each setting, the results are based on the explanations and predictions obtained with a single set of training shots. We only show the results of P-E on E-SNLI, as E-P is substantially worse here.

Results We summarize the results in Table 2. We only report consistency on E-SNLI, as the explanations for E-SNLI often require some external commonsense knowledge which cannot be easily grounded in the inputs or judged as true or false (examples in Appendix F). The results suggest a disconnect between the model predictions and the "reasoning" in explanations: even though using explanations improves GPT-3's performance, the generated explanations are *unreliable*, even for the straightforward synthetic setting. Overall, GPT-3 tends to generate consistent explanations (>90% for all three datasets with the right prompt structure), but the explanations are less likely to be factual, which is concerning as they can deceive a user of the system into believing the model's answer.

3.1 Reliability of Explanations and Prediction Accuracy

GPT-3 may hallucinate problematic explanations, but this could actually be advantageous if it gives us a way of spotting when the model's "reasoning" has failed. We investigate the connection between the reliability of an explanation and the accuracy of a prediction, and ask whether a reliable explanation indicates an accurate prediction. (This resembles the linguistic calibration of Mielke et al. [28], but using a different signal for calibration.)

As shown in the right section of Table 2, accuracy and factuality/consistency are typically correlated, especially factuality. By knowing whether an explanation is factual, we can guess the model's prediction a high fraction of the time (Accuracy = Factuality). A nonfactual explanation very likely (89.6%) means an incorrect prediction on the SYNTH dataset. On ADVHOTPOT, factuality and the model's prediction correspond 80.0% of the time, substantially surpassing the prediction accuracy itself. We show fractions of correct and incorrect predictions when the explanations are factual/nonfactual and consistent/inconsistent in Figure 4 for two of our settings. Factual explanations are much more likely paired with correct predictions compared to nonfactual explanations. Consistency also connects to the accuracy, but is an inferior indicator compared to factuality in general (Table 2).

4 Calibrating In-Context Learning using Explanations

From Section 3.1, we see that a human oracle assessment of the factuality of an explanation could be of substantial use for calibrating the corresponding prediction. Can we automate this process?

We show how to achieve this goal on the perfectly controlled SYNTH dataset (Section 4.1). On our other two datasets, we use surface lexical matching to approximate semantic matching and give real-valued scores approximately reflecting factuality. Following past work on supervised calibration [20, 7, 48], we can learn a calibrator that tunes the probabilities of a prediction based on the score of its explanation (Section 4.2). We show such a calibrator can be trained with a handful of

examples beyond those used for in-context learning and successfully improve the in-context learning performance on realistic datasets.[6]

4.1 Motivating Example: Improving SYNTH Dataset

We first show how post-hoc calibration functions in the controlled SYNTH setting, where we can simply check the factuality of an explanation. Since the generated explanation always follows the format "`B is [profession] and A [verb] B.`" (example in Figure 2), we can split the explanation into two sentences. The explanation is factual if and only if both of the sentences exactly match one of the sentences in the context.

We use the assessment to improve the performance of P-E for SYNTH, where a nonfactual explanation typically indicates an incorrect prediction. This gives us a way to reject presumably incorrect answers. Specifically, we iterate through the top 5 candidate answers (restricted by the API) given by GPT-3 and reject any answer-explanation pair if the explanation is nonfactual until we find a factual one. This procedure dramatically improves the accuracy from 54.8% to 79.2%. Note that this SYNTH dataset without any possible reasoning shortcuts is a challenging task. For reference, neither ROBERTA [26] and DEBERTA [17] finetuned with 16 examples can achieve an accuracy surpassing 50%. With the help of the explanations and the checking procedure, we can use GPT-3 to achieve strong results using few-shot learning.

4.2 Learning-based Calibration Framework

Framework We now introduce the framework that can leverage the factuality assessment of an explanation to calibrate a prediction. Let $\boldsymbol{p}$ be the vector of predicted probabilities associated with each class label in NLI (or the probability score of predicted answer in QA). Let v be a scalar value extracted from the explanation to describe the factuality. Then, we can adjust the probabilities accordingly using a linear model:

$$\hat{\boldsymbol{p}} = \mathrm{softmax}(W[\boldsymbol{p}; v] + b),$$

where $\hat{\boldsymbol{p}}$ is the tuned probabilities. Our calibration framework is extended from classical calibration methods [31, 15, 53], which apply an affine transformation on the probabilities alone: $\hat{\boldsymbol{p}} = \mathrm{softmax}(W\boldsymbol{p} + b)$. In contrast, we use an additional factor v in calibration to incorporate the factuality assessment of the explanation.

There are a small number of parameters (W and b) that need to be trained in such a calibration framework. We will rely on a few more examples in addition to the shots we use in the prompt to train the calibrator. Specifically, we use the prompt examples to generate the predictions and explanations for these extra examples, and extract predicted probabilities, factors, and target probabilities triples to construct training data points used to train the calibrator. Note this procedure requires **no** explanation annotations for the extra examples.

Approximating Factuality We approximate the factuality using lexical overlap between the explanations and the inputs, which we found to work fairly well for our tasks.

ADVHOTPOT: We use an explanation consisting of two sentences (examples in Figure 3) as an illustration. Let $\mathcal{E} = (E^{(1)}, E^{(2)})$ be the generated explanation, where $E^{(1)}$ and $E^{(2)}$ are the two sentences, and the $E^{(i)} = (e_1, e_2, \cdots)$ contain tokens $e_1, e_2, \cdots$. Similarly, let $\mathcal{P} = (P^{(1)}, P^{(2)}, P^{(3)}, P^{(4)})$ be the context paragraphs, and $P^{(i)} = (p_1, p_2, \cdots)$ be the tokens. The factuality estimation of one explanation sentence $E^{(i)}$ is defined as:

$$\mathcal{V}(E^{(i)}) = \max_{P \in \mathcal{P}} \frac{|E^{(i)} \cap P|}{|E^{(i)}|}.$$

Intuitively, the factuality score for a sentence E is defined as the maximum number of overlapping tokens over all paragraphs P, normalized by the number of tokens in E. We then define the factuality

[6] This procedure does require extra data. However, it provides a natural avenue for using a small number of additional examples that otherwise would be *impossible* to incorporate into this procedure, when the size of the context actually limits the amount of data for in-context learning.

score for the whole explanation as: $\mathcal{V}(\mathcal{E}) = \min_{E \in \mathcal{E}} \mathcal{V}(E)$, as it requires all sentences to be factual in order to make the entire explanation factual.[7]

E-SNLI: On the E-SNLI dataset whose explanations do not really involve a concept of factuality, we still use an analogous score following the same principle, where we regard the premise as the context. Let $E = (e_1, e_2, \cdots)$ be the explanation and $P = (p_1, p_2, \cdots)$ be the premise. We simply score the explanation by $\mathcal{V}(E) = \frac{|E| \cap |P|}{|E|}$. Namely, the more an explanation overlaps with the premise, the more factual it is.

4.3 Calibrating E-SNLI

Setup For E-SNLI, we use calibration methods to postprocess the final probabilities. Unlike classical temperature scaling [31], note that the methods we use here can actually change the prediction; we will therefore evaluate on *accuracy* of the calibrated model.

We study the effectiveness of our explanation-based calibrator under different training data sizes varying from 32 to 128. Recall that we only require explanation annotations for 32 data points, and only need the labels for the rest to train the calibrator. For E-SNLI, we calibrate P-E, which is shown to be more effective than E-P in this setting (Section 2.4).

Table 3: Accuracy ($\text{mean}_{\text{std dev}}$) of various methods on E-SNLI under different data conditions. **L** denotes number of labels (as well as the total number of examples); **E** denotes the number of explanations. Calibrating using explanations successfully improves the performance of in-context learning.

w/o Explanation	**32L**	**64L**	**96L**	**128L**
RoBERTa	$40.1_{4.7}$	$43.0_{5.1}$	$49.0_{5.2}$	$54.9_{4.8}$
FEW-SHOT	$56.8_{2.0}$	–	–	–
FEW-SHOT(NN)	–	–	–	$58.9_{1.0}$
FEW-SHOT+PROBCAL	$61.9_{3.8}$	$62.4_{2.6}$	$63.2_{2.9}$	$63.9_{1.2}$
w/ Explanation	**32L+32E**	**64L+32E**	**96L+32E**	**128L+32E**
P-E	$59.4_{2.0}$	–	–	–
P-E+PROBCAL	$\mathbf{64.4_{1.8}}$	$65.4_{1.2}$	$65.4_{1.6}$	$65.4_{1.9}$
P-E+EXPLCAL	$64.2_{2.6}$	$\mathbf{65.8_{1.3}}$	$\mathbf{67.6_{1.6}}$	$\mathbf{68.5_{1.2}}$
P-E+ZHANG [51]	$63.0_{3.2}$	$65.2_{2.2}$	$65.4_{1.5}$	$65.9_{2.5}$

Baselines We provide the performance of fine-tuned ROBERTA [26] model as a reference, finding this to work better than DeBERTa [17]. To isolate the effectiveness of using explanations for calibration, we introduce three additional baselines using non-explanation-based calibrators. We apply the probability-based calibrator as described in Section 4.2 on the results obtained on few-shot learning (FEW-SHOT+PROBCAL) and predict-then-explain pipeline (P-E+PROBCAL). We note that the parameters of these calibrators are trained using the addition data points, as opposed to being heuristically determined as in Zhao et al. [53]. Furthermore, we experiment with a recently proposed supervised calibrator from Zhang et al. [51], which uses the CLS representations from an additional language model as features in the calibrator. The probabilities are tuned using $\hat{\boldsymbol{p}} = \text{softmax}(W[\boldsymbol{p}; \boldsymbol{h}] + b)$, where $\boldsymbol{h}$ is the CLS representation. Since we do not have access to the embeddings obtained by GPT-3, we use ROBERTA to extract the vectors instead. We use such a calibrator on top of our best-performing base model, P-E, resulting P-E+ZHANG [51].

Limited by the maximum prompt length, in-context learning is not able to take as input the additional data used for training the calibrator. For a fair comparison, we can allow the in-context model to use this data by varying the prompts across test examples, dynamically choosing the prompt examples to maximize performance. Choosing closer data points for prompting is a common and effective way of scaling up the training data size for in-context learning [37, 25]. Following Liu et al. [25], we test the performance of choosing nearest neighbors for the prompt based on CLS embedding produced by a ROBERTA model [26], referred as FEW-SHOT(NN). It is worth clarifying that the FEW-SHOT and FEW-SHOT+PROBCAL approaches use the same set of 32 training shots in the prompt for every test example, whereas the shot sets vary from example to example in FEW-SHOT(NN).

Results We show the results in Table 3. We use 5 different groups of training examples and report the mean and standard deviation across the groups. For FEW-SHOT(NN), we only report the results obtained using 128 examples.

Under 128 training examples, applying a trained calibrator on top of prompting with explanation (i.e., P-E+EXPLCAL) achieves the best accuracy of 68.5%, which is 12% higher than the performance of the vanilla uncalibrated few-shot in-context learning (FEW-SHOT). P-E+EXPLCAL also outperforms

[7] Alternatively, one might use a fine-tuned NLI model as a proxy [7]. However, our focus in on the pure black-box setting, and we avoid models that require substantial amounts of data to make work.

Table 4: AUC scores ($\text{mean}_{\text{std dev}}$) on ADVHOTPOT under different data conditions. **L** and **E** denotes the number of label annotations and explanation annotations, respectively. Explanation-based calibration successfully improves the performance on top of prompting with explanations.

w/o Explanation	**6L**	**32L**	**64L**
FEW-SHOT	$59.6_{2.4}$	–	–
FEW-SHOT(NN)	–	–	$61.3_{0.9}$
w/ Explanation	**6L+6E**	**32L+6E**	**64L+6E**
E-P	$\mathbf{64.4}_{2.9}$	–	–
E-P+EXPLCAL	–	$\mathbf{67.2}_{3.2}$	$\mathbf{68.8}_{2.9}$
E-P+ZHANG [51]	–	$65.6_{3.9}$	$66.1_{3.2}$

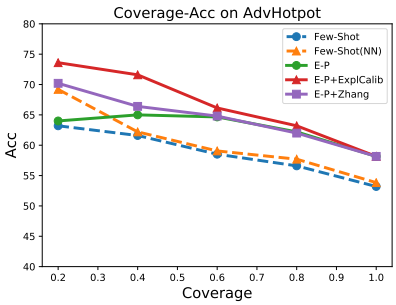

Figure 5: Coverage-Acc curves of various methods on ADVHOTPOT. E-P+EXPLCAL are better calibrated compared to uncalbrated E-P as well as the other approaches.

FEW-SHOT+PROBCAL and P-E+PROBCAL by 5% and 3%, respectively. Using explanations is more effective than using probabilities alone. In addition, P-E+EXPLCAL also outperforms P-E+ZHANG ET AL. [51], whose performance is on par with P-E+PROBCAL. This suggests the additional CLS information is not very helpful in this setting.

As the data size increases from 32 to 128, the performance of the explanation-based calibrator keeps improving notably, whereas the performance of probability-based calibrators nearly saturates at a data size of 96. The performance of FEW-SHOT(NN) with 128 training instances only improves the performance by 3.3%, compared to FEW-SHOT with 32 training instances. Choosing nearest neighbors as the shots, while being effective when having access to a large amount of data, is not helpful in the extreme data-scarce regime. Calibrating using explanations is an effective way of using a few extra data points that cannot fit in the prompt, which is a pitfall of standard in-context learning.

Finally, ROBERTA finetuned using 128 shots only achieves an accuracy of 54.9%, lagging the performance of GPT-3 based models. The limited training data size is insufficient for finetuning smaller language models like ROBERTA, but is sufficient for P-E+EXPLCAL to be effective.

4.4 Calibrating ADVHOTPOT

Setup For the ADVHOTPOT dataset, our calibration takes the form of tuning the confidence scores of the predicted answers to better align them with the correctness of predictions. These confidence scores can be used in a "selective QA" setting [20], where the model can abstain on a certain fraction of questions where it assigns low confidence to its answers. We use the *area under coverage-accuracy curve* (AUC) to evaluate how well a model is calibrated as in past literature [20, 7, 51, 13, 48]. The curve plots the average accuracy with varying fractions (coverage) of questions being answered (examples in Figure 5). For any given coverage, a better calibrated model should be able to identify questions that it performs best on, hence resulting a higher AUC.

We experiment with training data set sizes of 6, 32, and 64. We report the results averaged from 5 trials using different training sets. For ADVHOTPOT, we calibrate E-P, which is shown to be more effective than P-E in this setting (Section 2.4). Our approach is also effective for calibrating P-E; please refer to Appendix E for details.

Results We show the AUC scores in Table 4. By leveraging explanations, E-P+EXPLCAL successfully achieves an AUC of 68.8, surpassing both FEW-SHOT by 7 points and E-P by 4 points. We note this is substantial improvement, given that the upperbound of AUC is constrained by the accuracy of the answers and cannot reach 100. Figure 5 shows the coverage-accuracy curves of various methods averaged across the 5 training runs. E-P+EXPLCAL always achieves a higher accuracy than its uncalibrated counterpart, E-P, under a certain coverage, and the gap is especially large in the most confident intervals (coverage < 50%). E-P+ZHANG ET AL. [51] is able to calibrate the predictions on this dataset, but still lags our explanation-based calibrator, E-P+EXPLCAL.

In addition, the explanation-based calibrator can be effective with as few as 32 examples. This is because there are only two parameters (the probability of predicted answer and the explanation-

based factor) in the calibrator, which can be easily learned in this few-shot setting. Comparing E-P+EXPLCAL against FEW-SHOT(NN), using nearest neighbors in the prompt is also able to improve the performance compared to using a fixed set of shots (FEW-SHOT), yet our lightweight calibrator can better utilize such a small amount of data, and learn to distinguish more accurate predictions based on the explanations.

5 Related Work

Our investigation is centered around in-context learning [3], which has garnered increasing interest since the breakthrough of various large pretrained language models. Recent work has been devoted to studying different aspects of in-context learning, including its wayward behaviors [29, 41] and approaches to overcome them [53], whereas our exploration focuses on using explanations.

The utility of explanations for few-shot in-context learning has also been discussed concurrently [30, 42, 27, 9, 23, 44], especially in symbolic reasoning tasks. We differ in that we study more free-form explanations in tasks (QA and NLI, specifically) focusing on textual reasoning over provided contexts. Furthermore, our work focuses on the nature of the explanations generated by GPT-3, which are found to be unreliable. Regarding our use of calibration, similar ideas of explanation-based performance estimation have been applied to other tasks [34, 49, 48], but we rely on the free-text explanations generated by the model instead of interpretations obtained through post-hoc interpretation techniques.

More broadly, how to use explanations in various forms (textual explanation, highlights, etc.) to train better models is a longstanding problem [50]. Past work has built a series of pipeline models that first generate the explanations and then make predictions purely based on the generated explanations [45, 54, 6]. Prior research has also explored using explanations as additional supervision to train joint models [16, 11, 22, 39]. Another line of work seeks to aligning the reasoning process of a trained models with the explanations, which is typically done by interpreting a prediction post-hoc through explanation techniques and optimize the distance between the obtained explanation and ground truth explanation [24, 36, 32, 12, 47]. These aforementioned methods all update the model parameters and typically require a considerable amount of explanation annotations to be effective. By contrast, our setting treats language models as pure black boxes and only requires few-shot explanations.

6 Discussion & Conclusion

Caveats and Risks of Explanations from Large Language Models Our analysis suggests that GPT-3's internal "reasoning" does not always align with explanations that it generates, as shown by our consistency results. More concerning, the explanations might not be factually grounded in the provided prompt. This shortcoming should caution against any deployment of this technology in practice: because the explanations are grammatical English and look very convincing, they may deceive users into believing the system's responses even when those responses are incorrect. Section 6 of Bender et al. [1] discusses these risks in additional detail. The fact that language models can hallucinate explanations is also found in other work [54]. This result is unsurprising in some sense: without sufficient supervision or grounding, language models do not learn meaning as distinct from form [2], so we should not expect their explanations to be strongly grounded.

We have shown that even explanations which don't lead to accuracy gains can still be useful for calibration. However, the lexical overlap feature we use here is a weak signal of explanation correctness (see the example in Figure 1). Strong enough entailment models should theoretically be able to perform this task and work across a range of tasks without fine-tuning. This explanation assessing model can even be a language model itself trained for this particular propose to approach the verification tasks for a given domain by in-context learning.

Conclusion We have explored the capabilities of GPT-3 in using explanations in in-context learning for textual reasoning. Through our experiments on two QA datasets and an NLI dataset, we find that simply including explanations in the prompt does not always improve the performance of in-context learning. Our manual analysis demonstrates that GPT-3 tends to generate nonfactual explanations when making wrong predictions, which can be a useful leverage to assess the correctness of the predictions. Lastly, we showcase how to use explanations to build lightweight calibrators, which successfully improve in-context learning performance across all three datasets.

References

[1] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

[2] Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL)*.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NeurIPS)*.

[4] Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Boerschinger, and Tal Schuster. 2022. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. *arXiv preprint arXiv:2202.07654*.

[5] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NeurIPS)*.

[6] Howard Chen, Jacqueline He, Karthik Narasimhan, and Danqi Chen. 2022. Can rationalization improve robustness? In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL)*.

[7] Jifan Chen, Eunsol Choi, and Greg Durrett. 2021. Can NLI models verify QA systems' predictions? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[8] Jifan Chen and Greg Durrett. 2019. Understanding dataset design choices for multi-hop reasoning. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

[9] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Baindoor Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Oliveira Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *ArXiv*, abs/2204.02311.

[10] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL)*.

[11] Dheeru Dua, Sameer Singh, and Matt Gardner. 2020. Benefits of intermediate annotations in reading comprehension. In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL)*.

[12] Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, Scott M Lundberg, and Su-In Lee. 2021. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature machine intelligence*, 3(7):620–631.

[13] Siddhant Garg and Alessandro Moschitti. 2021. Will this question be answered? question filtering via answer model distillation for efficient question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

[14] Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

[15] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*.

[16] Braden Hancock, Paroma Varma, Stephanie Wang, Martin Bringmann, Percy Liang, and Christopher Ré. 2018. Training classifiers with natural language explanations. In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL)*.

[17] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[18] Alon Jacovi and Yoav Goldberg. 2021. Aligning faithful interpretations with their social attribution. *Transactions of the Association for Computational Linguistics (TACL)*, 9:294–310.

[19] Yichen Jiang and Mohit Bansal. 2019. Avoiding reasoning shortcuts: Adversarial evaluation, training, and model development for multi-hop QA. In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL)*.

[20] Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

[21] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *ArXiv*, abs/2205.11916.

[22] Matthew Lamm, Jennimaria Palomaki, Chris Alberti, Daniel Andor, Eunsol Choi, Livio Baldini Soares, and Michael Collins. 2021. QED: A framework and dataset for explanations in question answering. *Transactions of the Association for Computational Linguistics (TACL)*, 9:790–806.

[23] Andrew K Lampinen, Ishita Dasgupta, Stephanie CY Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L McClelland, Jane X Wang, and Felix Hill. 2022. Can language models learn from explanations in context? *arXiv preprint arXiv:2204.02329*.

[24] Frederick Liu and Besim Avci. 2019. Incorporating priors with feature attribution on text classification. In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL)*.

[25] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *ArXiv*, abs/2101.06804.

[26] Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

[27] Ana Marasović, Iz Beltagy, Doug Downey, and Matthew E. Peters. 2022. Few-shot self-rationalization with natural language prompts. In *Findings of the North American Chapter of the Association for Computational Linguistics (NAACL Findings)*.

[28] Sabrina J. Mielke, Arthur D. Szlam, Y.-Lan Boureau, and Emily Dinan. 2020. Linguistic calibration through metacognition: aligning dialogue agent responses with expected correctness. *ArXiv*, abs/2012.14983.

[29] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

[30] Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. 2021. Show your work: Scratchpads for intermediate computation with language models. *ArXiv*, abs/2112.00114.

[31] John Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

[32] Gregory Plumb, Maruan Al-Shedivat, Ángel Alexander Cabrera, Adam Perer, Eric Xing, and Ameet Talwalkar. 2020. Regularizing black-box models for improved interpretability. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NeurIPS)*.

[33] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL)*.

[34] Nazneen Fatema Rajani and Raymond Mooney. 2018. Stacking with auxiliary features for visual question answering. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana.

[35] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

[36] Laura Rieger, Chandan Singh, William Murdoch, and Bin Yu. 2020. Interpretations are useful: Penalizing explanations to align neural networks with prior knowledge. In *Proceedings of the International Conference on Machine Learning (ICML)*.

[37] Richard Shin, Christopher Lin, Sam Thomson, Charles Chen, Subhro Roy, Emmanouil Antonios Platanios, Adam Pauls, Dan Klein, Jason Eisner, and Benjamin Van Durme. 2021. Constrained language models yield few-shot semantic parsers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[38] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*.

[39] Joe Stacey, Yonatan Belinkov, and Marek Rei. 2022. Supervising model attention with human explanations for robust natural language inference. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

[40] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the International Conference on Machine Learning (ICML)*.

[41] Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

[42] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

[43] Jason Weston, Antoine Bordes, Sumit Chopra, and Tomás Mikolov. 2016. Towards ai-complete question answering: A set of prerequisite toy tasks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[44] Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing human-ai collaboration for generating free-text explanations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

[45] Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. 2021. Measuring association between labels and free-text rationales. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[46] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[47] Huihan Yao, Ying Chen, Qinyuan Ye, Xisen Jin, and Xiang Ren. 2021. Refining language models with compositional explanations. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NeurIPS)*.

[48] Xi Ye and Greg Durrett. 2022. Can explanations be useful for calibrating black box models. In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL)*.

[49] Xi Ye, Rohan Nair, and Greg Durrett. 2021. Connecting attributions and QA model behavior on realistic counterfactuals. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[50] Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *Proceedings of the Annual Conference of the Association for Computational Linguistics (ACL)*.

[51] Shujian Zhang, Chengyue Gong, and Eunsol Choi. 2021. Knowing more about questions can help: Improving calibration in question answering. In *Findings of the Annual Conference of the Association for Computational Linguistics (ACL Findings)*.

[52] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. *ArXiv*, abs/2205.01068.

[53] Tony Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[54] Yangqiaoyu Zhou and Chenhao Tan. 2021. Investigating the effect of natural language explanations on out-of-distribution generalization in few-shot NLI. In *Proceedings of the Workshop on Insights from Negative Results in NLP*.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 6.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We use GPT-3 instruct series API (text-davinci-001).
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See reference [19] and [5].
   (b) Did you mention the license of the assets? [Yes] See Section 2.1.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We included the Synthetic dataset in the supplemental material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Details of Prompts

We show examples of the prompts used for SYNTH, ADVHOTPOT, and E-SNLI in Figure 6, Figure 7, and Figure 8, respectively. Our prompts follow the original formats in Brown et al. [3]. For approaches that use explanations (E-P and P-E), we insert explanations before/after with necessary conjunction words.

SYNTHETIC: FEW-SHOT
Christopher agrees with Kevin. Tiffany agrees with Matthew. Mary hangs out with Danielle. James hangs out with Thomas. Kevin is a student. Matthew is a plumber. Danielle is a student. Thomas is a plumber. Q: Who hangs out with a student? A: Mary
SYNTHETIC: E-P
Christopher agrees with Kevin. Tiffany agrees with Matthew. Mary hangs out with Danielle. James hangs out with Thomas. Kevin is a student. Matthew is a plumber. Danielle is a student. Thomas is a plumber. Q: Who hangs out with a student? A: Because Danielle is a student and Mary hangs out with Danielle, the answer is Mary.
SYNTHETIC: P-E
Christopher agrees with Kevin. Tiffany agrees with Matthew. Mary hangs out with Danielle. James hangs out with Thomas. Kevin is a student. Matthew is a plumber. Danielle is a student. Thomas is a plumber. Q: Who hangs out with a student? A: Mary, because Danielle is a student and Mary hangs out with Danielle .

Figure 6: Examples of prompts for SYNTH.

ADVHOTPOT: FEW-SHOT
Sir Luigi Arthur Pirandello (12 August 1895 – 4 October 1952) was an John journalist. Sir Keith Arthur Murdoch (12 August 1885 – 4 October 1952) was an Australian journalist. Australian Associated Press (AAP) is an Australian news agency. The organisation was established in 1935 by Keith Murdoch. Sir Nikolai Arthur Trubetzkoy (12 August 1896 – 4 October 1952) was an Covington journalist. Q: Australian Associated Press was established by a journalist born in which year? A: 1885
ADVHOTPOT: E-P
Sir Luigi Arthur Pirandello (12 August 1895 – 4 October 1952) was an John journalist. Sir Keith Arthur Murdoch (12 August 1885 – 4 October 1952) was an Australian journalist. Australian Associated Press (AAP) is an Australian news agency. The organisation was established in 1935 by Keith Murdoch. Sir Nikolai Arthur Trubetzkoy (12 August 1896 – 4 October 1952) was an Covington journalist. Q: Australian Associated Press was established by a journalist born in which year? A: First, Australian Associated Press was established by Keith Murdoch in 1935. Second, Keith Murdoch was born in 1885. The answer is 1885.
ADVHOTPOT: P-E
Sir Luigi Arthur Pirandello (12 August 1895 – 4 October 1952) was an John journalist. Sir Keith Arthur Murdoch (12 August 1885 – 4 October 1952) was an Australian journalist. Australian Associated Press (AAP) is an Australian news agency. The organisation was established in 1935 by Keith Murdoch. Sir Nikolai Arthur Trubetzkoy (12 August 1896 – 4 October 1952) was an Covington journalist. Q: Australian Associated Press was established by a journalist born in which year? A: 1885. The reasons are as follows. First, Australian Associated Press was established by Keith Murdoch in 1935. Second, Keith Murdochwas born in 1885. The answer is 1885.

Figure 7: Examples of prompts for ADVHOTPOT.

E-SNLI: FEW-SHOT
A person in black tries to knock the last pin down in a game of bowling. Q: The person is a girl. True, False, or Neither? A: Neither
E-SNLI: E-P
A person in black tries to knock the last pin down in a game of bowling. Q: The person is a girl. True, False, or Neither? A: Neither, because not every person is a girl.
E-SNLI: P-E
A person in black tries to knock the last pin down in a game of bowling. Q: The person is a girl. True, False, or Neither? A: Because not every person is a girl, this answer is Neither.

Figure 8: Examples of prompts for E-SNLI.

B Details of the SYNTH Dataset

We create a controlled synthetic multi-hop QA dataset. Each context consists of four reasoning chains, where each chain contains two sentences following a template: "`A [verb] B. B is [profession].`". We fill in `A` and `B` in the reasoning chain templates using randomly selected names from a pool of 50 names. To fill in the `[verb]` and `[profession]` in the four reasoning chain templates, we first select two verbs from a pool of 30 verbs and two professions from a pool of 30 professions. Next, we fill in the four chains using the combination of these two verbs and professions, which give a set of completely symmetric chains. Finally, we sample one reasoning chain from all of the four to derive a asking: "`Who [verb] [profession]?`" (example in Figure 2).

Such a design ensures there are no reasoning shortcuts [8], making it a difficult dataset even despite the regular structure of the task. A ROBERTA model needs roughly 500 data points to tackle this problem and achieve near 100% accuracy on the test set.

C Details of Preprocessing ADVHOTPOT Dataset

We preprocess the original Adversarial HotpotQA dataset [46, 19] in a few ways. We reduce the context length to make it better fit the purpose of testing in-context learning. We use two ground truth supporting paragraphs joined with two adversarial paragraphs to construct the context for each question, instead of using all eight distractors. In addition, we simplify each paragraph by only keeping relevant sentences needed for answering the question (or distracting the prediction); otherwise, the prompt length limit only allows 2-3 examples fit in the input prompt.

We make a challenging test test set of 250 examples by balancing the mix of examples on which prompted GPT-3 makes correct and incorrect predictions. This is done by first running few-shot inference over 1000 examples, and then randomly sampling 125 examples with correct and incorrect predictions, respectively.

Since assessing the accuracy of an answer in QA is hard, and F1 scores do not correlate with the true quality of the answers (e.g., "United States" is a correct answer but has 0 F1 score with respect to the provided ground truth answer "US") [4], we manually assess the correctness of the answers. We observed a high inter-annotator agreement (Cohen's Kappa of 0.87) between the correctness annotations of 100 examples on which the annotations of the authors intersected. Please refer to the supplementary material for these annotations.

D Details of Reliability Annotations

The authors manually inspected the predictions and explanations generated for the 250 ADVHOTPOT test examples using a single set of training shots, and annotated them for factuality and consistency. We observed a Cohen's Kappa of 0.89 between the factuality annotations of 100 examples (obtained

using the E-P paradigm) on which the annotations of the authors overlapped. Please refer to the supplementary material for these annotations.

E Calibrating P-E on ADVHOTPOT

Table 5: AUC scores of various methods on ADVHOTPOT under different data conditions. Explanations are also effective for calibrating P-E.

w/o Explanation	**6L**	**32L**	**64L**
FEW-SHOT	**59.6**$_{2.4}$	–	–
FEW-SHOT(NN)	–	–	61.3$_{0.9}$
w/ Explanation	**6L+6E**	**32L+6E**	**64L+6E**
P-E	58.4$_{2.6}$	–	–
P-E+EXPLCAL	–	**62.2**$_{2.8}$	**62.6**$_{3.1}$

We also experiment with calibrating predictions obtained using P-E on ADVHOTPOT, following the same setup as described in Section 4.4.

As in Table 5, using explanations is also effective for calibrating the predictions obtained using P-E. P-E+EXPLCAL successfully improves the AUC of P-E from 58.4 to 62.6, outperforming FEW-SHOT and FEW-SHOT (NN).

F Additional Output Examples

	SYNTH: E-P
Nonfactual	Jason helps Christopher. Kevin helps Stephanie. James punches Jonathan. Daniel is a typist. Jonathan is a tailor. Christopher is a typist. Richard punches Daniel. Stephanie is a tailor. Q: Who punches a typist? A: Because Jonathan is a typist and James punches Jonathan, the answer is James.
Nonfactual	Matthew blames Tiffany. Lisa is a chef. Christopher helps Kelly. Angela helps Jessica. Rachel blames Lisa. Jessica is a farmer. Kelly is a chef. Tiffany is a farmer. Q: Who helps a farmer? A: Because Kelly is a farmer and Christopher helps Kelly, the answer is Christopher.
	SYNTH: P-E
Nonfactual	Richard is a governor. Megan supervises Adam. Tiffany calls Nicholas. Nicholas is a musician. Elizabeth is a musician. Matthew calls Richard. Robert supervises Elizabeth. Adam is a governor. Q: Who calls a governor? A: Tiffany, because Nicholas is a governor and Tiffany calls Nicholas.
Nonfactual	Anthony believes Justin. John is an engineer. Eric answers Kevin. Amy is an engineer. Justin is a pilot. Sarah answers Amy. Kevin is a pilot. Brandon believes John. Q: Who answers a pilot? A: Sarah, because Amy is a pilot and Sarah answers Amy.

Figure 9: Examples of failure cases on SYNTH. GPT-3 almost always generates nonfactual explanations when giving an incorrect answer, using either P-E or E-P.

ADVHOTPOT: P-E	
Nonfactual	"Crestfallen" is a track on "Adore". The single's artwork is by Yelena Yemchuk. Johnny McDaid is a Croatian professional photographer, painter and film director. Arnljot Berg is a French professional photographer, painter and film director. Yelena Yemchuk is a Ukrainian professional photographer, painter and film director. Q: Crestfallen's artwork is done by a photographer of which nationality? A: Croatian. The reasons are as follows. First, Crestfallen's artwork is done by Yelena Yemchuk. Second, Yelena Yemchuk is Croatian photographer.
Inconsistent	The Pequannoc Spillway is a professional basketball team based in Ellesmere city, United Kingdom. The Silver Lake is a professional basketball team based in Ellesmere harbour, United Kingdom. The Cheshire Phoenix is a professional basketball team based in Ellesmere Port, United Kingdom. Mike DiNunno (born January 29, 1990) is an American-Italian professional basketball player for the Cheshire Phoenix, playing the point guard position. Q: Where is the basketball team that Mike DiNunno plays for based? A: Cheshire Phoenix. The reason is as follows. First, Mike DiNunno plays for the Cheshire Phoenix. Second, the Cheshire Phoenix is based in Ellesmere Port.

Figure 10: Examples of unreliable explanations generated for ADVHOTPOT using P-E.

E-SNLI: E-P	
GT: **True**	A female is looking through a microscope. Q: A lady is observing something True, False, or Neither? A: True, because if you're looking through a microscope you are observing something.
	A person wearing sunglasses, a visor, and a British flag is carrying 6 Heineken bottles. Q: The person has beer bottles and has a england flag and glasses. True, False, or Neither? A: False , because the person could be carrying any number of items and not just six Heineken bottles.
GT: **False**	A woman and a man are sword fighting. Q: Two people have engaged in a sword fight and both are dead. True, False, or Neither? A: False, because it is impossible for two people to sword fight and both be dead.
	Two construction workers in a truck having a conversation. Q: Two people are outside a truck. True, False, or Neither? A: True, because two people are outside a truck.
GT: **Neither**	A dog running on grass Q: A dog is playing outside. True, False, or Neither? A: True, because a dog playing outside implies that the dog is playing outside.
	A silhouette at the bottom of an escalator. Q: The silhoutte is creeping out the children. True, False, or Neither? A: False, because the silhouette is not necessarily creeping out the children.

Figure 11: The completions generated for E-SNLI examples with different ground truth labels (GT) using E-P. GPT-3 sometimes ignores the information from premises when explaining its predictions (examples in the bottom section).

G Details of Automatically Assessing Consistency and Factuality on SYNTH

Our questions follow the template `Who` V_1 P_1`?`. Our generated explanations follow the template N_1 `is` P_2 `and` N_2 V_2 N_3. Our answers are always a name, e.g., N_4. Because large language models almost always produce well-formed explanations, we can match the output against these patterns and extract variables V_1, P_1, etc. using simple regular expressions.

We say that an explanation is consistent if and only if the following conditions are satisfied: (1) $N_2 = N_4$ and $N_1 = N_3$. (2) $P_2 = P_1$ and $V_2 = V_1$. These rules ensure the explanation matches the intent of the question and entails the answer at the same time.

We say an explanation is factual if and only if both N_1 `is` P_2 and N_2 V_2 N_3 appear exactly in the context.

Table 6: Results of prompting with explanations on four large language models. Using explanations mildly improves performance on OPT, davinci, and text-davinci-001 (which is the one reported in the main text), and has more prominent effects on text-davinci-002.

		SYNTH	ADVHOTPOT	E-SNLI
OPT-175B	FEW-SHOT	**40.5**$_{2.8}$	49.7$_{2.6}$	**44.0**$_{3.8}$
	E-P	29.6$_{0.5}$	**52.6**$_{6.5}$	39.3$_{7.8}$
	P-E	40.2$_{2.6}$	43.3$_{4.5}$	43.4$_{1.6}$
davinci	FEW-SHOT	49.5$_{0.6}$	49.1$_{6.2}$	43.3$_{5.7}$
	E-P	47.1$_{2.8}$	**54.1**$_{4.1}$	40.4$_{4.5}$
	P-E	**51.3**$_{1.8}$	48.7$_{4.6}$	**48.7**$_{2.4}$
text-davinci-001	FEW-SHOT	54.8$_{3.1}$	53.2$_{2.3}$	56.8$_{2.0}$
	E-P	**58.5**$_{2.1}$	**58.2**$_{4.1}$	41.8$_{2.5}$
	P-E	53.6$_{1.0}$	51.5$_{2.4}$	**59.4**$_{1.0}$
text-davinci-002	FEW-SHOT	72.0$_{1.4}$	77.7$_{3.2}$	69.1$_{2.0}$
	E-P	**86.9**$_{3.8}$	**82.4**$_{5.1}$	**75.6**$_{7.6}$
	P-E	81.1$_{2.8}$	77.2$_{4.8}$	69.4$_{5.0}$

H Results of Prompting with Explanations on Other Language Models

In addition to text-davinci-001, we also show the results of prompting with explanations on other large language models, including OPT-175B [52] and two other models available on the OpenAI API: davinci (GPT-3 non-Instruct series), and text-davinci-002 (the latest GPT-3 Instruct series model). OPT and davinci are languge models trained using the standard causal language modeling objective, whereas text-davinci-001 and text-davinci-002 are trained with special data and objectives in order to align with human instructions. We note that for LLMs other than text-davinci-001, we only use 3 sets of randomly selected shots (rather than 5 as in the main text) to reduce the cost of experiments.

As shown in Table 6, the results on OPT and davinci are consistent with our findings on text-davinci-001. E-P consistently provides the strongest performance on the ADVHOTPOT setting, but the improvements are 5% absolute or less. On SYNTH and E-SNLI, E-P typically degrades performance (except on SYNTH for text-davinci-001) and P-E is inconsistent across the different models. Overall, vanilla LLMs (OPT and davinci) see limited benefit from producing explanations, and even text-davinci-001 does not see substantial improvement.[8]

The only exception is text-davinci-002. text-davinci-002 greatly benefits from explanations in the prompt across all the three tasks, and E-P is consistently more effective than P-E. However, it is unclear what contributes to this difference. As far as we are aware, the differences between text-davinci-002 and text-davinci-001 are not described in any publication or blog post.[9] Comparing davinci and text-davinci-001, we see the move to Instruct series models is *not* sufficient to explain the difference.

One possibility is that 002 is an updated version of 001 that includes more Instruct data collected using the API. One hypothesis for the improvement is data leakage from our test set. Because we started running experiments for this work in late 2021, it is conceivable that text-davinci-002 was trained on human-written completions for our data. Another hypothesis is that text-davinci-002 features T0-like fine-tuning on some available datasets such as HotpotQA, which would also change the interpretation of the results.

[8] When assessing the scale of the improvements and choosing to describe them as "mild" or "not substantial," we are using as calibration the facts that (a) SYNTH is a synthetic dataset, easily solved by a rule-based system, and therefore we expect these models to do very well on it; (b) supervised models on ADVHOTPOT can achieve substantially higher performance as well.

[9] One publicly-described difference is the addition of editing and insertion, discussed at `https://openai.com/blog/gpt-3-edit-insert/`, but this does not explain the performance differences we observe.

Table 7: Reliability of explanations generated by other language models.

		Acc	Fac	Con	Acc=Fac	Acc=Con
OPT-175B	SYNTH (E-P)	30.0	77.2	47.2	45.6	58.8
	SYNTH (P-E)	39.6	64.0	81.2	**69.2**	49.6
davinci	SYNTH (E-P)	46.8	59.2	64.8	**66.8**	61.2
	SYNTH (P-E)	52.4	52.4	83.2	**78.4**	58.0
text-davinci-001	SYNTH (E-P)	58.4	72.8	64.8	66.5	68.8
	SYNTH (P-E)	54.8	51.6	95.2	**89.6**	57.2
text-davinci-002	SYNTH (E-P)	86.0	91.6	85.2	**91.2**	84.8
	SYNTH (P-E)	81.6	83.2	96.4	**95.8**	82.8

Given the lack of transparency with this model, we hesitate to make scientific claims about the results it yields. In any case, the relatively poor performance of E-P for three of the four models we explore means that we cannot broadly argue that explain-predict is the superior configuration.

I Reliability of Explanations Generated by Other Language Models on SYNTH

Table 7 shows the factuality and consistency of explanations generated by various language models on SYNTH. The different models and different explanation setups vary in how the factuality and consistency of the explanations compare. On P-E, the models are much more consistent than they are factual. On E-P, the opposite trend is observed except for davinci. However, we note that this setting (E-P on SYNTH) is an outlier; both the ADVHOTPOT results in Table 2 and the alternative prompt style explored in Appendix J feature higher consistency than factuality.

As we argue in the main body of the paper, factuality can be useful for assessing the correctness of predictions across different models; Accuracy=Factuality (the fraction of the time that factuality agrees with accuracy) is always higher than Accuracy.

J Results of Using Explanations in an Alternative Style on SYNTH

Table 8: Performance of text-davinci-001 of using explanations in an alternative style on SYNTH.

		SYNTH
davinci	FEW-SHOT	**49.5±0.6**
	E-P (ALTERNATIVE)	48.0±2.6
	P-E (ALTERNATIVE)	49.5±1.7
text-davinci-001	FEW-SHOT	**54.8±2.5**
	E-P (ALTERNATIVE)	50.6±1.6
	P-E (ALTERNATIVE)	53.3±1.6
text-davinci-002	FEW-SHOT	72.0±1.4
	E-P (ALTERNATIVE)	75.3±2.2
	P-E (ALTERNATIVE)	**80.5±2.4**

We also experimented with using an alternative style of explanations for SYNTH, where we reversed the order of the two sentences in the explanations shown in Table 2. These explanations follow the format: `A [verb] B and B is [profession].` (instead of `B is [profession] and A [verb] B.`) By changing the order in which the sentences are extracted, we might expect that E-P can more easily follow the reasoning chain.

Table 9: Reliability of explanations in an alternative style.

		Acc	Fac	Con	Acc=Fac	Acc=Con
davinci	SYNTH (ALTERNATIVE; E-P)	50.8	53.6	97.6	97.2	53.2
	SYNTH (ALTERNATIVE; P-E)	52.8	52.8	98.4	98.4	54.8
text-davinci-001	SYNTH (ALTERNATIVE; E-P)	50.8	53.6	97.6	97.2	53.2
	SYNTH (ALTERNATIVE; P-E)	52.8	52.8	98.4	98.4	54.8
text-davinci-002	SYNTH (ALTERNATIVE; E-P)	75.2	79.6	100.	95.6	75.2
	SYNTH (ALTERNATIVE; P-E)	82.8	86.0	100.	96.8	82.8

Table 10: Results of adding "let's think step by step" trigger in prompts.

		SYNTH	ADVHOTPOT
davinci	FEW-SHOT	**49.5**$_{0.6}$	49.1$_{6.2}$
	E-P	47.1$_{2.8}$	**54.1**$_{4.1}$
	E-P + TRIGGER	48.6$_{2.6}$	50.1$_{5.2}$
text-davinci-001	FEW-SHOT	54.8$_{2.5}$	53.2$_{2.3}$
	E-P	**58.5**$_{2.1}$	**58.2**$_{4.1}$
	E-P + TRIGGER	58.0$_{3.4}$	58.0$_{6.2}$

We show the performance of using reversed explanations in Table 8 and the reliability in Table 9. In general, this alternative style of explanations yields inferior performance compared to the original style (Table 6). Using explanations leads to no improvements on davinci, and text-davinci-001. P-E is consistently better than E-P across davinci, text-davinci-001, and text-davinci-002.

Furthermore, using such a reversed style, language models almost always generates consistent explanations when being prompted in either E-P or P-E paradigm. The factuality almost always indicates the correctness of predictions.

We believe these two prompts cover the most natural explanation styles for this problem. While small format changes or modifications to the general QA prompt format are also possible, we observed these to have minor impacts on the results (as we see in Appendix K).

K Results of Adding "Step by Step" Trigger in Prompts

We test whether including a trigger for multi-step reasoning can help LLMs better learn from explanations in the prompt for multi-step reasoning. Following [21], we prepend "Let's think step by step." in the exemplar explanations used in the E-P paradigm. For this experiment, we only test on SYNTH and ADVHOTPOT, which involve multi-step reasoning. We do not experiment with text-davinci-002, which has already gained substantial performance improvements from using explanations, and we omit OPT because its performance is too low.

As shown in Table 10, adding triggers in the prompts does not lead to statistically significantly improvements in E-P for davinci and text-davinci-001. In fact, it typically causes a performance degradation.

L Information about Cost of Running Experiments

The cost of our experiments, described as follows, is estimated based on using the GPT-3 API with the largest models available (davinci, text-davinci-001, and text-davinci-002). The setting in Table 1 uses 250 examples for each result, with roughly 1400 tokens per example using the FEW-SHOT paradigm and 2000 tokens per example using the E-P or E-P paradigm. The cost of evaluating FEW-SHOT, P-E, and E-P for 5 trials on a single dataset is roughly $105, $150, and $150, respectively. The total price for reproducing Table 1 using a single language model is roughly $1200.

We subsample 250-example sets to reduce cost rather than running on full datasets. Based on the significance tests in this paper and the reported confidence intervals, this size dataset is sufficient to distinguish between the performance of different approaches.