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. (2020). 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: Because not every person is a girl, this answer is Neither.
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: Neither, because not every person is a girl.

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 (Chen and Durrett, 2019), 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 the ADVHOTPOT Dataset

We preprocess the original Adversarial HotpotQA dataset (Yang et al., 2018; Jiang and Bansal, 2019) 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") (Bulian et al., 2022), we manually assess the correctness of the answers. We observed a high inter-annotator agreement (Cohen's Kappa of 0.84) between the correctness annotations of 100 examples on which the annotations of the authors intersected. Please refer to the supplementary material for these annotations.

This dataset is licensed under the MIT license.

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 0.85 between the factuality annotations of 100 examples (obtained using the E-P paradigm) on which the annotations of the authors overlapped.

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.

H Results of Using Explanations in an Alternative Style on SYNTH

Table 6: Performance of text-davinci-001 of using explanations in an alternative style on SYNTH.

		SYNTH
GPT-3	FEW-SHOT	**49.5±0.6**
	E-P (ALTERNATIVE)	48.0±2.6
	P-E (ALTERNATIVE)	49.5±1.7
InstructGPT	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.

We show the performance of using reversed explanations in Table 6 and the reliability in Table 7. In general, this alternative style of explanations yields inferior performance compared to the original style (Table 1). Using explanations leads to no improvements on GPT-3, and InstructGPT. P-E is consistently better than E-P across GPT-3, InstructGPT, and text-davinci-002.

Furthermore, using such a reversed style, language models almost always generate 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 I).

I 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 Kojima et al. (2022), we prepend "Let's think step by step." to 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 achieved substantial performance improvement from using explanations, and we omit OPT because its performance is too low.

Table 7: Reliability of explanations in an alternative style.

		Acc	Fac	Con	Acc=Fac	Acc=Con
davinci	SYNTH (ALTERNATIVE; E-P)	48.4	53.6	98.4	**94.8**	48.4
	SYNTH (ALTERNATIVE; P-E)	51.6	53.2	100.	**98.4**	51.6
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 8: 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$_{6.2}$

As shown in Table 8, adding triggers in the prompts does not lead to statistically significantly improvements in E-P for GPT-3 and InstructGPT. In fact, it typically causes a performance degradation.

J 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) as of August 2022 ($0.06 per 1,000 tokens). 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 results on three datasets as in 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.