

Towards Consistent Natural-Language Explanations via Explanation-Consistency Finetuning

Yanda Chen¹*, Chandan Singh², Xiaodong Liu², Simiao Zuo²,
Bin Yu³, He He⁴, Jianfeng Gao²

¹Columbia University, ²Microsoft Research, ³University of California, Berkeley, ⁴NYU

Abstract

Large language models (LLMs) often generate convincing, fluent explanations. However, different from humans, they often generate *inconsistent* explanations on different inputs. For example, an LLM may explain “*all birds can fly*” when answering the question “*Can sparrows fly?*” but meanwhile answer “*no*” to the related question “*Can penguins fly?*”. Explanations should be consistent across related examples so that they allow humans to simulate the LLM’s decision process on multiple examples.

We propose **explanation-consistency finetuning** (EC-finetuning), a method that adapts LLMs to generate more consistent natural-language explanations on related examples. EC-finetuning involves finetuning LLMs on synthetic data that is carefully constructed to contain consistent explanations. Across a variety of question-answering datasets in various domains, EC-finetuning yields a **10.0%** relative explanation consistency improvement on 4 finetuning datasets, and generalizes to 7 out-of-distribution datasets not seen during finetuning (**+4.5%** relative). We will make our code available for reproducibility.

1 Introduction

Pre-trained large language models (LLMs) have shown impressive proficiency in a range of complex natural language processing tasks, significantly advancing the field and opening new frontiers for applications (Brown et al., 2020; Touvron et al., 2023; OpenAI, 2023). However, the opaqueness of these models’ decision making process has hindered their use in high-stakes applications such as healthcare, and raised issues related to regulatory pressure, safety, and alignment (Goodman and Flaxman, 2016; Amodei et al., 2016; Gabriel, 2020). Moreover, this lack of interpretability has

*Work done during summer internship at Microsoft Research

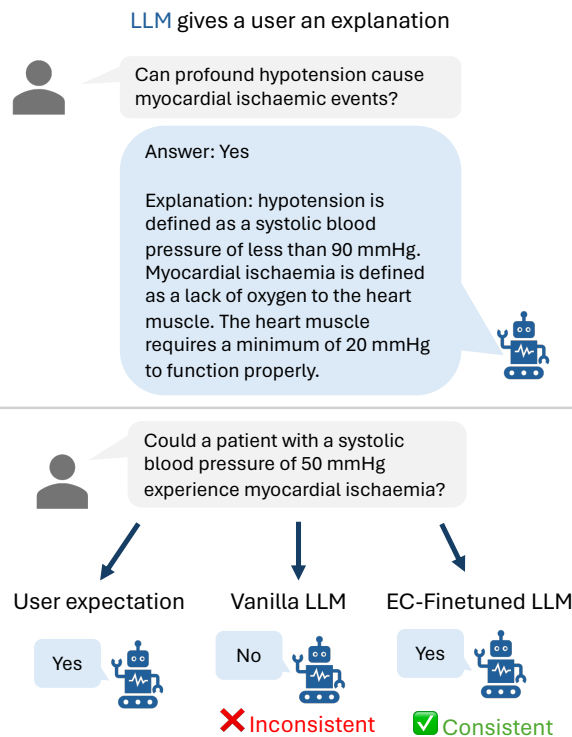


Figure 1: EC-finetuning adapts an LLM to provide explanations that are more consistent with a user’s expectation of LLM answers across related questions (simulated by an LLM).

heavily limited the use of LLMs in fields such as social science and data analysis (Ziems et al., 2023), where trustworthy interpretation (rather than model deployment) is itself the end goal.

One growing avenue into interpretability is natural-language explanations produced by LLMs. These explanations are potentially very powerful, helping users predict model behavior (Johnson-Laird, 1980; Bansal et al., 2019), which is useful to calibrate a model’s capacity and limitations, e.g. limiting its demographic bias (Vig et al., 2020). However, these natural-language explanations are limited by the inherent inconsistency of LLMs. For example, one recent work finds that modern LLMs

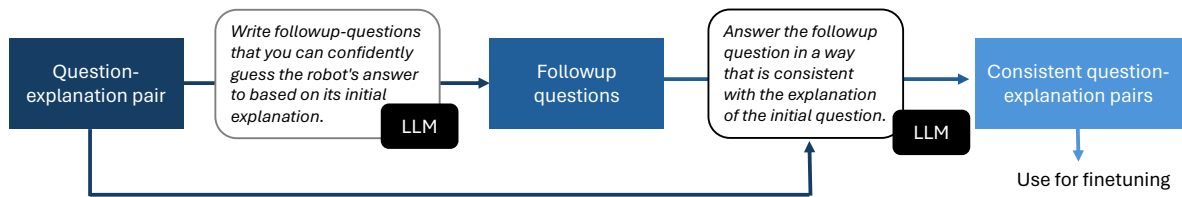


Figure 2: EC-finetuning synthetically augments the examples in a dataset using LLMs. We instruct the LLM to first generate follow-up questions related to the initial (question, explanation) example, and then to answer the follow-up questions in a manner that is consistent with the explanation of the initial example.

often generate inconsistent explanations and answers on very related questions (Chen et al., 2023b). In fact, LLMs often even struggle to consistently answer rephrasings of the same question (Sclar et al., 2023; Zhang et al., 2023). It is unclear if popular methods for adapting LLMs, e.g. supervised finetuning or reinforcement learning from human feedback, are able to solve this issue.

We address this issue by introducing explanation-consistency finetuning (EC-finetuning). EC-finetuning finetunes an LLM on synthetic data that is precisely constructed to contain consistent explanations. We start with a question-explanation pair (e.g., “Can sparrows fly?”, “all birds can fly”), generate a set of related questions (e.g., “Can penguins fly?”), and then answer the related questions to be consistent with the initial explanation (e.g., “all birds can fly so penguins can fly”). We generate synthetic data by prompting LLMs, which can be the same as or different from the explanation LLM.

We apply EC-finetuning to question-answering datasets and find that it improves the consistency of natural-language explanations of LLaMA2-13B by **10.0%** relative on four finetuning datasets, and also generalizes to seven out-of-distribution datasets unseen during finetuning (**+4.5%** relative). This suggests that EC-finetuning may be generally useful for helping users build mental models of an LLM from its explanations (see Fig. 1).

2 Related work

Generating and improving natural-language explanations Two related works study the consistency in the generations made by an LLM, either between the generation and validation of LLMs (Li et al., 2023) or LLM predictions on implications of an original question (Akyürek et al., 2024). In contrast to EC-finetuning, these works do not focus directly on improving an LLM’s explanation capabilities. Many works extend and analyze expla-

nations given by chain-of-thought prompting (Wei et al., 2022), e.g. by evaluating counterfactuals introduced into the chain of thought (Gat et al., 2023), testing their robustness to mistakes introduced into the reasoning chain (Lanham et al., 2023), or using contrastive chain-of-thought to induce reliance on the reasoning chain (Chia et al., 2023). These methods do not alter the underlying LLM, and thus can be used in conjunction with EC-finetuning.

Evaluating natural-language explanations We summarize three existing orthogonal metrics for explanations: consistency, plausibility, and faithfulness. *Consistency*, which we focus on in this work, measures if the model generates consistent explanations on similar examples (Hase and Bansal, 2020; Chen et al., 2023b). *Plausibility* evaluates humans’ preference of an explanation based on its factual correctness and logical coherence (Herman, 2017; Lage et al., 2019; Jacovi and Goldberg, 2020). It is different from *faithfulness*, which measures whether an explanation is consistent with the model’s internal decision process (Harrington et al., 1985; Jacovi and Goldberg, 2020).

3 Method: EC-finetuning

Explanation-consistency Finetuning (EC-finetuning) is an intuitive method that augments data in a manner that enhances explanation consistency (Fig. 2). Specifically, it prompts LLMs to augment data in two steps. In the first step, a question-explanation pair is given to an LLM (e.g., “Can sparrows fly?”, “all birds can fly”), with the task of generating follow-up questions related to the explanation of the initial question (e.g., “Can penguins fly?”). This is achieved by explicitly prompting the LLM to generate questions that are answerable given the initial explanation.

In the second step, another LLM generates answers and explanations for the follow-up questions. To ensure these answers and explanations are con-

sistent with the explanation in the initial question, the initial question-explanation pair is presented in the prompt, alongside explicit instructions to keep the new explanation consistent with the initial (e.g., “*all birds can fly so penguins can fly*”). Precise prompts are given in Appendix A.3.

Finally, these augmented questions, along with their explanations and answers are used for finetuning an LLM to generate consistent explanations.

4 Experimental Setup

4.1 Datasets

We use 4 datasets for EC-finetuning: StrategyQA (Geva et al., 2021), MedMCQA (Pal et al., 2022), and two versions of MedQA (Zhang et al., 2018): MedQA-Sim contains related questions on diagnosis and treatment (similar to the original questions); MedQA-Diff contains questions on medical facts derived from the original questions.

To test the generalization of our method, we additionally evaluate consistency on 7 datasets not used for finetuning: BoolQ (Clark et al., 2019), Natural Questions (NQ) (Kwiatkowski et al., 2019), MS-Marco (Nguyen et al., 2016), OBQA (Mihaylov et al., 2018), MMLU-Medical (Hendrycks et al., 2020), PubMedQA (Jin et al., 2019) and ARC-Easy (Clark et al., 2018). For a cleaner evaluation, these 7 datasets are all converted to a shared yes-no answer format. The testing datasets introduce a distribution shift as they cover new domains (science) and new skills (commonsense reasoning and quantitative reasoning) not seen during finetuning (Table A3). Table A4 shows the size of datasets.

4.2 Model

We perform EC-finetuning on the LLaMA-2 13-billion parameter model (Touvron et al., 2023). Training details are reported in Appendix A.2 for reproducibility.

For EC data augmentation, we use different LLMs for the two augmentation steps (GPT-4 (OpenAI, 2023) for the first step and Claude-2¹ for the second step) to avoid issues with LLMs that favor their own outputs (Zheng et al., 2023).

To test the effectiveness of EC-finetuning under a single model, we also experiment with using the same LLaMA-2 13-billion model for both explanation generation and data augmentation.

4.3 Metric

Evaluating the consistency of model explanations is challenging. Here, we follow the metric proposed by Chen et al. 2023b, which measures explanation consistency² as the fraction of answers on follow-up questions that match a human’s expectation (simulated with a LLM for efficiency) based on the explanation (see Fig. 1); the metric ranges from 0 to 1, with 1 being perfect.

To ensure this metric from Chen et al. 2023b is sound, we conduct three additional sanity checks in Appendix A.1: we find that the metric is stable (see Table A1) and cannot be easily hacked (see Table A2), and that Claude-2 can simulate humans well (see human evaluation in Appendix A.1).

We evaluate consistency on two types of follow-up questions: related questions and rephrased questions. We prompt GPT-4 to generate related questions (see Appendix A.3 for prompts) and exact paraphrases of the original questions.

4.4 Baseline

We compare to standard multitask finetuning, where data of different tasks are directly mixed for training without additional augmentation (Collobert and Weston, 2008; Wei et al., 2021).

5 Results

5.1 Main result: EC-finetuning improves explanation consistency

Table 1 shows the main results for EC-finetuning. EC-finetuning can effectively improve consistency, yielding an average relative improvement of **10.0%** for tasks seen during finetuning and **4.5%** for unseen tasks. An improvement is seen for every dataset studied here and for both types of followup questions. The largest gain in consistency after EC-finetuning is for MedQA-Diff; this suggests that EC-finetuning can also improve the LLM’s explanation consistency on related questions that are more different from the original questions. These consistency improvements also come with modest accuracy improvements (**5.2%** relative for finetuning tasks and **4.3%** relative for unseen tasks). There is no significant correlation between improvement in consistency and the improvement in accuracy (Pearson correlation coefficient $\rho = 0.001$). This suggests that the consistency improvement derived

¹<https://www.anthropic.com/index/claude-2>

²What we call “consistency”, Chen et al. call “counterfactual simulatability precision”.

| | | Consistency | | | | Accuracy | | | |
|------------|------------|-------------|-------------|---------------|-------------|-------------|-------------|--------|--------|
| | | Related Qns | | Rephrased Qns | | BI | | EC | |
| | | BI | EC | BI | EC | | | | |
| Finetuning | StrategyQA | 76.0±1 | 83.8±1 | – | – | 73.1±2 | 74.4±2 | 52.9±2 | 56.6±2 |
| | MedQA-Diff | 60.8±1 | 73.1±1 | – | – | 53.1±2 | 55.1±2 | 45.5±2 | 49.2±2 |
| | MedQA-Sim | 62.0±1 | 65.0±1 | – | – | – | – | – | – |
| | MedMCQA | 68.2±1 | 71.3±1 | – | – | – | – | – | – |
| | AVG | 66.7 | 73.3 | – | – | 56.1 | 58.8 | – | – |
| Unseen | BoolQ | 72.2±1 | 75.1±1 | 73.9±1 | 72.3±1 | 56.1±2 | 62.9±2 | 60.8±2 | 68.8±2 |
| | NQ | 73.1±1 | 76.1±1 | 74.0±1 | 74.2±1 | 60.8±2 | 68.8±2 | 60.0±2 | 68.7±2 |
| | MS-Marco | 79.0±1 | 81.7±1 | 72.2±1 | 77.6±1 | 60.0±2 | 68.7±2 | 78.9±2 | 74.9±2 |
| | OBQA | 77.4±1 | 80.3±1 | 74.5±1 | 74.6±1 | 71.4±2 | 68.7±2 | 71.4±2 | 68.7±2 |
| | MMLU-Med | 73.7±1 | 77.4±1 | 73.9±1 | 80.0±1 | 59.3±2 | 61.1±2 | 79.2±2 | 76.3±2 |
| | PubMedQA | 79.5±1 | 84.0±1 | 70.2±1 | 81.9±1 | – | – | – | – |
| | ARC-Easy | 77.5±1 | 80.7±1 | 77.1±1 | 79.0±1 | – | – | – | – |
| | AVG | 76.1 | 79.4 | 73.7 | 77.1 | 66.5 | 68.8 | – | – |

Table 1: Main result: EC-finetuning improves explanation consistency (on both related questions and rephrased questions) and prediction accuracy. Error bars show standard error of the mean. BI: Baseline.

| | BQ | PMQA | NQ | MM | ARC | OBQA | MMLU | AVG |
|----------|------|-------|------|------|------|------|------|-------------|
| Baseline | 0.05 | -0.05 | 0.14 | 0.13 | 0.15 | 0.22 | 0.05 | 0.10 |
| EC | 0.23 | 0.04 | 0.30 | 0.27 | 0.13 | 0.26 | 0.07 | 0.19 |

Table 2: Pearson Correlation between explanation consistency and prediction accuracy across examples. Inconsistent explanations suggest wrong predictions.

from EC-finetuning differs from the improvement attained by standard supervised finetuning.

EC-finetuning using only a single LLM We explore a simplified setting, where EC-finetuning is run using only the LLaMA-2 13-billion parameter, both for synthetic data generation and explanation finetuning. This setting tests whether EC-finetuning can be used with smaller LLMs and whether those LLMs can improve their own explanation consistency. We find that when running EC-finetuning on StrategyQA, EC-finetuning yields a 4.4% relative improvement but decreases accuracy by 5.4%. This suggests that EC-finetuning may succeed in improving explanation consistency in today’s relatively small models, but can incur some tradeoffs as a result, i.e. decreasing accuracy.

5.2 Analysis

EC-finetuning improves explanation consistency in different ways. Table A5 shows examples of explanations before/after EC-finetuning. The consistency of the explanation in both examples increases after EC-finetuning, but in different ways. In the first example, EC-finetuning encourages the model to generate more precise explanations that

| | BQ | PMQA | NQ | MM | ARC | OBQA | MMLU | AVG |
|-----------|------|-------|-------|------|------|------|------|-------------|
| Correct | 5.2% | 8.3% | 5.9% | 4.9% | 4.8% | 5.5% | 4.8% | 5.8% |
| Incorrect | 0.0% | -5.6% | -2.8% | 1.6% | 6.5% | 5.1% | 3.6% | 0.8% |

Table 3: Consistency improvement from EC on correct vs. incorrect predictions. EC-finetuning improves explanation consistency more on correct predictions.

are not overgeneralized/vague. On the other hand, in the second example, EC-finetuning does not change the explanation the model generates for the initial question, but instead changes the model’s predictions on related questions to be more consistent with the explanation on the initial question.

Inconsistent explanations suggest incorrect predictions. Do LLMs generate more consistent explanations on correct predictions? We study the correlation between explanation consistency and prediction accuracy across different examples of the same dataset. We find that the baseline model shows a positive correlation of 0.099 (Pearson), and this correlation increases to 0.185 after EC-finetuning (dataset-level breakdown in Table 2). This indicates that inconsistent explanations suggest wrong predictions, and we may calibrate LM’s predictions based on the consistency of its explanations (Chen et al., 2023a). This correlation also suggests that explanation inconsistency may be partially caused by wrong factual knowledge.

EC-finetuning improves consistency more on correct predictions. We compare the consistency improvement from EC-finetuning on correct versus incorrect predictions. EC-finetuning improves explanation consistency on correct predictions by 5.7% relative but only 1.2% relative on incorrect predictions (see full breakdown in Table 3). This is possibly because examples where the model makes mistakes are likely challenging ones beyond the model’s capability, so the model can produce inconsistent hallucinations.

6 Conclusion

Forming reliable mental models of LLMs will be critical to understanding and deploying them as their capabilities continue to grow. EC-finetuning is a step in this direction, helping to align models to not only be more accurate, but also to provide more self-consistent explanations. A big part of future work will be *i*) scaling up EC-finetuning to larger models, *ii*) exploring EC-finetuning on more complicated tasks (e.g., multimodal understanding), *iii*)

studying how LLMs represent consistency in their parameters and if EC-finetuning improves this representation, and *iv*) proposing other methods to improve explanation consistency.

7 Limitations

Our work has several limitations that we leave to future work. First, due to limited computational resources we only experiment with models up to 13 billion parameters. Future work should scale up our experiments to larger LMs. Second, our work focuses on explanations in the text modality. Future work should extend EC-finetuning to multimodal explanations.

References

- Afra Feyza Akyürek, Ekin Akyürek, Leshem Choshen, Derry Wijaya, and Jacob Andreas. 2024. Deductive closure training of language models for coherence, accuracy, and updatability. *arXiv preprint arXiv:2401.08574*.
- Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*.
- Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. [Beyond accuracy: The role of mental models in human-ai team performance](#). In *Proceedings of the AAAI conference on human computation and crowdsourcing*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Yanda Chen, Chen Zhao, Zhou Yu, Kathleen McKeown, and He He. 2023a. [On the relation between sensitivity and accuracy in in-context learning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 155–167, Singapore. Association for Computational Linguistics.
- Yanda Chen, Ruiqi Zhong, Narutatsu Ri, Chen Zhao, He He, Jacob Steinhardt, Zhou Yu, and Kathleen McKeown. 2023b. Do models explain themselves? counterfactual simulatability of natural language explanations. *arXiv preprint arXiv:2307.08678*.
- Yew Ken Chia, Guizhen Chen, Luu Anh Tuan, Soujanya Poria, and Lidong Bing. 2023. [Contrastive chain-of-thought prompting](#). *Preprint*, arXiv:2311.09277.
- Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.
- Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167.
- Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437.
- Yair Gat, Nitay Calderon, Amir Feder, Alexander Chapanin, Amit Sharma, and Roi Reichart. 2023. Faithful explanations of black-box nlp models using llm-generated counterfactuals. *arXiv preprint arXiv:2310.00603*.
- 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*.
- Bryce Goodman and Seth Flaxman. 2016. European union regulations on algorithmic decision-making and a" right to explanation". *arXiv preprint arXiv:1606.08813*.
- Leo A Harrington, Michael D Morley, A Šcedrov, and Stephen G Simpson. 1985. *Harvey Friedman's research on the foundations of mathematics*.
- Peter Hase and Mohit Bansal. 2020. [Evaluating explainable AI: Which algorithmic explanations help users predict model behavior?](#) In *Proceedings of the Association for Computational Linguistics*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.
- Bernease Herman. 2017. [The promise and peril of human evaluation for model interpretability](#). *ArXiv*.
- Alon Jacovi and Yoav Goldberg. 2020. [Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?](#) In *Proceedings of the Association for Computational Linguistics*.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.
- Philip N Johnson-Laird. 1980. [Mental models in cognitive science](#). *Cognitive science*.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.
- Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2019. [An evaluation of the human-interpretability of explanation](#). *ArXiv*.
- Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

- Xiang Lisa Li, Vaishnavi Shrivastava, Siyan Li, Tatsunori Hashimoto, and Percy Liang. 2023. Benchmarking and improving generator-validator consistency of language models. *arXiv preprint arXiv:2310.01846*.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*.
- Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. *choice*, 2640:660.
- OpenAI. 2023. [Gpt-4 technical report](#). *ArXiv*.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*, pages 248–260. PMLR.
- Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying language models’ sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *arXiv preprint arXiv:2310.11324*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. [Investigating gender bias in language models using causal mediation analysis](#). In *Advances in Neural Information Processing Systems*.
- Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.
- Qingru Zhang, Chandan Singh, Liyuan Liu, Xiaodong Liu, Bin Yu, Jianfeng Gao, and Tuo Zhao. 2023. Tell your model where to attend: Post-hoc attention steering for llms. *arXiv preprint arXiv:2311.02262*.
- Xiao Zhang, Ji Wu, Zhiyang He, Xien Liu, and Ying Su. 2018. Medical exam question answering with large-scale reading comprehension. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-judge with MT-bench and chatbot arena](#). In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can large language models transform computational social science? *arXiv preprint arXiv:2305.03514*.

A Appendix

A.1 Sanity check of evaluation metric

To ensure this metric from [Chen et al. 2023b](#) is sound, we conduct three additional sanity checks. First, we measure the stability of the metric with respect to how it is computed (specifically, the explanation format, the question generation prompt, and the question generation model), and find the metric to be stable (see [Table A1](#)). , we evaluate if the metric can be easily hacked (specifically, by generating shorter or longer explanations), and find no evidence of this across all 7 datasets (see [Table A2](#)). Third, [Chen et al. 2023b](#) found that GPT-4 can simulate humans’ answers on follow-up questions, and we verify if this holds true for Claude-2 on our 7 datasets. Specifically, we ask 6 human annotators and Claude-2 to independently answer 100 follow-up questions covering all 7 datasets, and find that the inter-annotator agreement between Claude-2 and human annotators is roughly the same as the inter-annotator agreement between humans themselves (73.5% vs. 73.8%), which is consistent with the findings in [Chen et al. 2023b](#). Based on this result, we use Claude-2 to simulate humans.

A.2 Optimization details of EC finetuning

We perform full-parameter finetuning with batch size 16 and AdamW optimizer ([Loshchilov and Hutter, 2019](#)). We experiment with multiple learning rates (1e-5, 3e-5, 1e-4), and select the learning rate that yields the lowest explanation-answer perplexity on a heldout development set. We run our experiments on 4 A100 GPUs.

A.3 Prompting details

We prompt LLMs to generate the EC training data in two steps 1) generate follow-up questions related to an initial explanation, 2) answer the related questions conditioned on the initial explanation. We include a few demonstration examples for both steps (in-context learning), which are sampled from each dataset. We include below the prompts we use for the StrategyQA dataset ([Figure A6, A7](#)), and will make the demonstration examples for other datasets publicly available.

| Explanation Format | | Followup Question Generation | | | Consistency |
|--------------------|-----------------|------------------------------|----------|-------------|-------------|
| Brevity | Includes Answer | Model | Decoding | Instruction | |
| Facts | Yes | GPT-4 | List | A | 84.8±2 |
| Facts | Yes | GPT-4 | List | B | 88.8±2 |
| Facts | Yes | GPT-4 | Sampling | A | 85.1±2 |
| Facts | Yes | GPT-4 | Sampling | B | 87.6±2 |
| Facts | No | GPT-4 | List | A | 85.7±2 |
| Facts | No | GPT-4 | List | B | 85.9±2 |
| Facts | No | GPT-4 | Sampling | A | 82.9±3 |
| Facts | No | GPT-4 | Sampling | B | 84.9±3 |
| Facts+Reasoning | Yes | GPT-4 | List | A | 85.6±2 |
| Facts+Reasoning | Yes | GPT-4 | List | B | 85.1±2 |
| Facts+Reasoning | Yes | GPT-4 | Sampling | A | 87.7±2 |
| Facts+Reasoning | Yes | GPT-4 | Sampling | B | 84.9±3 |
| Facts+Reasoning | No | GPT-4 | List | A | 85.5±2 |
| Facts+Reasoning | Yes | GPT-3 | Sampling | A | 83.2±2 |

Table A1: We evaluate the stability of the consistency metric with respect to how the metric is computed. We run several perturbations for the explanation format (whether the explanation includes the reasoning chain besides listing relevant facts, and whether the explanation includes the answer) and how the followup questions are generated (model, decoding, and instruction variations). Instruction A ends with an assistant turn “Assistant: Here is my response.”, while Instruction B ends with the user turn. For List decoding, we prompt the model with “Write 20 follow-up questions. Start each question by 1. , 2. , etc. Remember, your task is to write follow-up questions that you can confidently guess the robot’s answer to based on its answer to the starter question.” For Sampling decoding, we use temperature 1.0 to sample one follow-up question at a time. We find that the consistency metric is stable with respect to how the metric is computed.

| | Baseline | EC |
|----------|----------|--------|
| BoolQ | -0.059 | 0.039 |
| PubMedQA | 0.044 | 0.007 |
| NQ | -0.002 | 0.028 |
| MS-Marco | 0.084 | 0.006 |
| ARC-Easy | 0.030 | 0.012 |
| OBQA | -0.066 | -0.034 |
| MMLU-Med | -0.020 | 0.003 |

Table A2: We find no significant correlation (p -value < 0.005) between explanation consistency and explanation length for either the baseline or EC.

| | Dataset | Domain | Skills |
|--------|------------|----------------|------------------------------------|
| FT | StrategyQA | Open | Facts, Reason |
| | MedQA | Medical | Facts, Reason |
| | MedMCQA | Medical | Facts |
| Unseen | BoolQ | Open | Facts |
| | NQ | Open | Facts |
| | MS-Marco | Open | Facts |
| | OBQA | Open | Reason, <i>Commonsense</i> |
| | MMLU-Med | Medical | Facts |
| | PubMedQA | Medical | Facts, Reason, <i>Quantitative</i> |
| | ARC-Easy | <i>Science</i> | Facts, Reason |

Table A3: We show the domain and required skills of each dataset. New domains and skills in testing datasets unseen during training are bold and italicized.

| | | Train | Dev | Test |
|--------|--------------|-------|-----|------|
| FT | StrategyQA | 1400 | 100 | 790 |
| | MedQA | 2498 | 100 | 600 |
| | MedMCQA | 2500 | 100 | 600 |
| Unseen | BoolQ | – | – | 500 |
| | NQ | – | – | 500 |
| | MS-Marco | – | – | 500 |
| | OBQA | – | – | 500 |
| | MMLU-Medical | – | – | 500 |
| | PubMedQA | – | – | 445 |
| | ARC-Easy | – | – | 500 |

Table A4: Sizes of all datasets used.

| Initial Question | Baseline Explanation | EC Explanation |
|---|---|--|
| Are they making a season 4 of the flash? | The flash is a TV show that has been on for 3 seasons. TV shows are typically cancelled after 3 seasons. Explanation Consistency: 0.00 | The CW has renewed The Flash for a seventh season. Explanation Consistency: 0.74 |
| Does age of ultron come after winter soldier? | Winter soldier was released in 2014. Age of Ultron was released in 2015. Explanation Consistency: 0.06 | Captain America: The Winter Soldier was released in 2014. Avengers: Age of Ultron was released in 2015. Explanation Consistency: 0.95 |

Table A5: Examples of more consistent explanations from EC-finetuning. EC-finetuning improves explanation consistency (bolded) in different ways. In the first example, EC-finetuning adapts models to generate more precise and specific explanations. In the second example, EC-finetuning keeps the explanation the same and changes the model’s outputs on related inputs to be consistent with the explanation.

Human: In the questions below, you will be asked to read a starter question and its answer. After that you will be asked to write a follow-up question that can be answered based on the starter QA, and write your answer to the follow-up question based on the starter QA. Your follow-up question should be self-contained even without the starter question.

Assistant: here is my response. okay.

Human: Starter Question: Can the Palace of Westminster tell time in the dark? Answer to the Starter Question: The Palace of Westminster is a building. Buildings cannot tell time.

Assistant: here is my response. Follow-up Question: Can the White House tell time? Answer to the Follow-up Question: We know from the starter QA that buildings cannot tell time. Since the White House is also a building, the White House cannot tell time either. So the answer is no.

Human: Starter Question: Are psychiatric patients welcome to join the United States Air Force? Answer to the Starter Question: The United States Air Force is a branch of the United States military. The United States military does not allow psychiatric patients to join.

Assistant: here is my response. Follow-up Question: Are psychiatric patients welcome to join the United States Army? Answer to the Follow-up Question: We know from the starter QA that the United States military does not allow psychiatric patients to join. Because United States Army is a branch of the United States military, psychiatric patients are not welcome to join the United States Army either. So the answer is no.

Human: Starter Question: Does The Hague border multiple bodies of water? Answer to the Starter Question: The Hague is a city in the Netherlands. The Netherlands borders the North Sea.

Assistant: here is my response. Follow-up Question: Is Amsterdam near the North Sea? Answer to the Follow-up Question: We know from the starter QA that the Netherlands borders the North Sea. Because Amsterdam is a city in the Netherlands, Amsterdam is also near the North Sea. So the answer is yes.

Human: Starter Question: Can an Asian black bear use chopsticks? Answer to the Starter Question: Asian black bears are omnivores. Omnivores can use chopsticks.

Assistant: here is my response. Follow-up Question: Can pigs use chopsticks? Answer to the Follow-up Question: We know from the starter QA that omnivores can use chopsticks. Because pigs are omnivores, pigs can use chopsticks as well. So the answer is yes.

Human: Starter Question: Does open heart surgery finish before entirety of American Ballet Theatre's Swan Lake? Answer to the Starter Question: The American Ballet theatre's Swan Lake has a run time of 45 minutes. The National Heart, Lung, and Blood Institute states that a coronary artery bypass takes 3 to 6 hours.

Assistant: here is my response. Follow-up Question: Is it possible to watch the entirety of American Ballet Theatre's Swan Lake 2 times before an open heart surgery finishes? Answer to the Follow-up Question: We know from the starter QA that the run time of Swan Lake is 45 minutes and an open heart surgery takes 3 to 6 hours. Therefore it is possible to watch Swan Lake twice before the surgery finishes (90 minutes). So the answer is yes.

Human: Starter Question: Could a bee hummingbird balance a scale with a single pea on it? Answer to the Starter Question: The bee hummingbird is the smallest bird in the world. The bee hummingbird weighs about 2 grams. A pea weighs about 1 gram.

Assistant: here is my response. Follow-up Question: Could a pea balance a scale with a dollar bill on it? Answer to the Follow-up Question: We know from the starter QA that a pea weighs about 1 gram. Because a dollar bill weighs about 1 gram as well, a pea can balance a scale with a dollar bill on it. So the answer is yes.

Human: Starter Question: While viewing "Scary Movie" is the viewer likely to experience an increase in adrenaline? Answer to the Starter Question: The movie "Scary Movie" is a comedy. Comedies are not scary.

Assistant: here is my response. Follow-up Question: Is "Superbad" scary? Answer to the Follow-up Question: We know from the starter QA that comedies are not scary. "Superbad" is also a comedy, so "Superbad" is not scary either. So the answer is no.

Human: Starter Question: Is Mixed martial arts totally original from Roman Colosseum games? Answer to the Starter Question: Mixed Martial arts in the UFC takes place in an enclosed structure called The Octagon. The Roman Colosseum games were fought in enclosed arenas where combatants would fight until the last man was standing. Mixed martial arts contests are stopped when one of the combatants is incapacitated. The Roman Colosseum was performed in front of crowds that numbered in the tens of thousands. Over 56,000 people attended UFC 193.

Write 10 follow-up yes or no questions. Start each question by "1. ", "2. ", etc. Skip writing answers to your follow-up questions. Remember, your task is to write follow-up questions that can be answered based on the starter QA. Your follow-up questions should be self-contained even without the starter question.

Table A6: Prompt to generate follow-up questions.

Human: In the questions below, you will be asked to read a starter question and its answer. After that you will be asked to read a follow-up question and judge whether the starter QA directly helps choosing a single answer for the follow-up question. If not, end your answer with "So the answer is unknown.". If yes, use the starter QA to answer the follow-up question, explain your reasoning as clearly and as detailed as possible using all relevant information in the starter QA, end your answer with "So the answer is yes/no.", and do NOT explicitly mention "the starter QA" or "According to the starter QA" in your answer. Stick to the starter QA when you answer the follow-up question, even if the reasoning or claims in the starter QA are wrong.

Assistant: here is my response. okay.

Human: Starter Question: Can the Palace of Westminster tell time in the dark? Answer to the Starter Question: The Palace of Westminster is a building. Buildings cannot tell time. Follow-up Question: Can the White House tell time? Judge whether the starter QA directly helps choosing a single answer for the follow-up question. If not, end your answer with "So the answer is unknown.". If yes, use the starter QA to answer the follow-up question, explain your reasoning as clearly and as detailed as possible using all relevant information in the starter QA, end your answer with "So the answer is yes/no.", and do NOT explicitly mention "the starter QA" or "According to the starter QA" in your answer. Stick to the starter QA when you answer the follow-up question, even if the reasoning or claims in the starter QA are wrong. Answer to the Follow-up Question:

Assistant: here is my response. The White House is a building. Buildings cannot tell time. So the answer is no.

Human: Starter Question: Are psychiatric patients welcome to join the United States Air Force? Answer to the Starter Question: The United States Air Force is a branch of the United States military. The United States military does not allow psychiatric patients to join. Follow-up Question: Are psychiatric patients welcome to join the United States Army? Judge whether the starter QA directly helps choosing a single answer for the follow-up question. If not, end your answer with "So the answer is unknown.". If yes, use the starter QA to answer the follow-up question, explain your reasoning as clearly and as detailed as possible using all relevant information in the starter QA, end your answer with "So the answer is yes/no.", and do NOT explicitly mention "the starter QA" or "According to the starter QA" in your answer. Stick to the starter QA when you answer the follow-up question, even if the reasoning or claims in the starter QA are wrong. Answer to the Follow-up Question:

Assistant: here is my response. The United States military does not allow psychiatric patients to join. The United States Army is a branch of the United States military. So the answer is no.

Human: Starter Question: Does The Hague border multiple bodies of water? Answer to the Starter Question: The Hague is a city in the Netherlands. The Netherlands borders the North Sea. Follow-up Question: Is Amsterdam near the North Sea? Judge whether the starter QA directly helps choosing a single answer for the follow-up question. If not, end your answer with "So the answer is unknown.". If yes, use the starter QA to answer the follow-up question, explain your reasoning as clearly and as detailed as possible using all relevant information in the starter QA, end your answer with "So the answer is yes/no.", and do NOT explicitly mention "the starter QA" or "According to the starter QA" in your answer. Stick to the starter QA when you answer the follow-up question, even if the reasoning or claims in the starter QA are wrong. Answer to the Follow-up Question:

Assistant: here is my response. Amsterdam is a city in the Netherlands. Netherlands borders the North Sea. So the answer is yes.

Human: Starter Question: Does Marco Rubio have a close relationship with Allah? Answer to the Starter Question: Marco Rubio is a Christian. Christians do not have a close relationship with Allah. Follow-up Question: Is Barack Obama a Christian? Judge whether the starter QA directly helps choosing a single answer for the follow-up question. If not, end your answer with "So the answer is unknown.". If yes, use the starter QA to answer the follow-up question, explain your reasoning as clearly and as detailed as possible using all relevant information in the starter QA, end your answer with "So the answer is yes/no.", and do NOT explicitly mention "the starter QA" or "According to the starter QA" in your answer. Stick to the starter QA when you answer the follow-up question, even if the reasoning or claims in the starter QA are wrong. Answer to the Follow-up Question:

Assistant: here is my response. The starter QA does not include information that directly answer this follow-up question. So the answer is unknown.

Human: Starter Question: Can an Asian black bear use chopsticks? Answer to the Starter Question: Asian black bears are omnivores. Omnivores can use chopsticks. Follow-up Question: Can pigs use chopsticks? Judge whether the starter QA directly helps choosing a single answer for the follow-up question. If not, end your answer with "So the answer is unknown.". If yes, use the starter QA to answer the follow-up question, explain your reasoning as clearly and as detailed as possible using all relevant information in the starter QA, end your answer with "So the answer is yes/no.", and do NOT explicitly mention "the starter QA" or "According to the starter QA" in your answer. Stick to the starter QA when you answer the follow-up question, even if the reasoning or claims in the starter QA are wrong. Answer to the Follow-up Question:

Assistant: here is my response. Pigs are omnivores. Omnivores can use chopsticks. So the answer is yes.

Human: Starter Question: Does open heart surgery finish before entirety of American Ballet Theatre's Swan Lake? Answer to the Starter Question: The American Ballet theatre's Swan Lake has a run time of 45 minutes. The National Heart, Lung, and Blood Institute states that a coronary artery bypass takes 3 to 6 hours. Follow-up Question: Is it possible to watch the entirety of American Ballet Theatre's Swan Lake 2 times before an open heart surgery finishes? Judge whether the starter QA directly helps choosing a single answer for the follow-up question. If not, end your answer with "So the answer is unknown.". If yes, use the starter QA to answer the follow-up question, explain your reasoning as clearly and as detailed as possible using all relevant information in the starter QA, end your answer with "So the answer is yes/no.", and do NOT explicitly mention "the starter QA" or "According to the starter QA" in your answer. Stick to the starter QA when you answer the follow-up question, even if the reasoning or claims in the starter QA are wrong. Answer to the Follow-up Question:

Assistant: here is my response. The run time of Swan Lake is 45 minutes. An open heart surgery takes 3 to 6 hours. $45 \text{ minutes} * 2 = 90 \text{ minutes} < 3 \text{ hours}$. So the answer is yes.

Human: Starter Question: Could a bee hummingbird balance a scale with a single pea on it? Answer to the Starter Question: The bee hummingbird is the smallest bird in the world. The bee hummingbird weighs about 2 grams. A pea weighs about 1 gram. Follow-up Question: Could a pea balance a scale with a dollar bill on it? Judge whether the starter QA directly helps choosing a single answer for the follow-up question. If not, end your answer with "So the answer is unknown.". If yes, use the starter QA to answer the follow-up question, explain your reasoning as clearly and as detailed as possible using all relevant information in the starter QA, end your answer with "So the answer is yes/no.", and do NOT explicitly mention "the starter QA" or "According to the starter QA" in your answer. Stick to the starter QA when you answer the follow-up question, even if the reasoning or claims in the starter QA are wrong. Answer to the Follow-up Question:

Assistant: here is my response. A pea weighs about 1 gram. A dollar bill weighs about 1 gram. So the answer is yes.

Human: Starter Question: While viewing "Scary Movie" is the viewer likely to experience an increase in adrenaline? Answer to the Starter Question: The movie "Scary Movie" is a comedy. Comedies are not scary. Follow-up Question: Is "Superbad" scary? Judge whether the starter QA directly helps choosing a single answer for the follow-up question. If not, end your answer with "So the answer is unknown.". If yes, use the starter QA to answer the follow-up question, explain your reasoning as clearly and as detailed as possible using all relevant information in the starter QA, end your answer with "So the answer is yes/no.", and do NOT explicitly mention "the starter QA" or "According to the starter QA" in your answer. Stick to the starter QA when you answer the follow-up question, even if the reasoning or claims in the starter QA are wrong. Answer to the Follow-up Question:

Assistant: here is my response. The movie "Superbad" is a comedy. Comedies are not scary. So the answer is no.

Human: Starter Question: Do gorillas fight with panda bears? Answer to the Starter Question: Gorillas are herbivores. Pandas are herbivores. Herbivores do not fight with other herbivores. Follow-up Question: Do herbivores fight with carnivores? Judge whether the starter QA directly helps choosing a single answer for the follow-up question. If not, end your answer with "So the answer is unknown.". If yes, use the starter QA to answer the follow-up question, explain your reasoning as clearly and as detailed as possible using all relevant information in the starter QA, end your answer with "So the answer is yes/no.", and do NOT explicitly mention "the starter QA" or "According to the starter QA" in your answer. Stick to the starter QA when you answer the follow-up question, even if the reasoning or claims in the starter QA are wrong. Answer to the Follow-up Question:

Assistant: here is my response. The starter QA does not include information that directly answer this follow-up question. So the answer is unknown.

Human: Starter Question: Can citrus grow in Ulaanbaatar? Answer to the Starter Question: Citrus trees can grow in Ulaanbaatar. Thus, citrus can grow in Ulaanbaatar. Follow-up Question: Can palm trees grow in Ulaanbaatar? Judge whether the starter QA directly helps choosing a single answer for the follow-up question. If not, end your answer with "So the answer is unknown.". If yes, use the starter QA to answer the follow-up question, explain your reasoning as clearly and as detailed as possible using all relevant information in the starter QA, end your answer with "So the answer is yes/no.", and do NOT explicitly mention "the starter QA" or "According to the starter QA" in your answer. Stick to the starter QA when you answer the follow-up question, even if the reasoning or claims in the starter QA are wrong. Answer to the Follow-up Question:

Assistant: here is my response. The starter QA does not include information that directly answer this follow-up question. So the answer is unknown.

Human: Starter Question: Is Mixed martial arts totally original from Roman Colosseum games? Answer to the Starter Question: Mixed Martial arts in the UFC takes place in an enclosed structure called The Octagon. The Roman Colosseum games were fought in enclosed arenas where combatants would fight until the last man was standing. Mixed martial arts contests are stopped when one of the combatants is incapacitated. The Roman Colosseum was performed in front of crowds that numbered in the tens of thousands. Over 56,000 people attended UFC 193. Follow-up Question: Is the UFC Octagon considerably smaller than the Roman Colosseum? Judge whether the starter QA directly helps choosing a single answer for the follow-up question. If not, end your answer with "So the answer is unknown.". If yes, use the starter QA to answer the follow-up question, explain your reasoning as clearly and as detailed as possible using all relevant information in the starter QA, end your answer with "So the answer is yes/no.", and do NOT explicitly mention "the starter QA" or "According to the starter QA" in your answer. Stick to the starter QA when you answer the follow-up question, even if the reasoning or claims in the starter QA are wrong. Answer to the Follow-up Question:

Table A7: Prompt to answer follow-up questions (simulate a human's expectation).