SELF-CORRECTION FOR OOD GENERALIZATION

Vanya Bannihatti Kumar*, **Abhinav Rao***, **Aditi Raghunathan**
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{vbanniha,abhinavr,aditirag}@cs.cmu.edu

ABSTRACT

In this work, we aim to study how the self-correction mechanisms aid OOD (out-of-distribution) generalization in both multimodal and language-only models. Reasoning based methods like self-refine (Madaan et al., 2023) and STaR (Zelikman et al., 2022) have helped to improve the correction capacity of the language models; however there have been no studies quantifying the reasoning improvement to help OOD generalization of these models. Initial results, show an improvement of 1.6%-2% on an OOD dataset where the model is finetuned using either self-refinement or STaR on an ID (in-distribution) dataset.

1 INTRODUCTION

Large-scale pretrained models like GPT4 (OpenAI et al., 2024) excel at generalizing across diverse tasks due to their training on vast and heterogeneous datasets. However, their performance often falls short in niche domains like medicine, where the data distribution diverges significantly from what was encountered during pretraining. Various fine-tuning techniques like full fine-tuning, Lora-tuning (Hu et al., 2021) etc addresses this gap by adapting these models to the specific characteristics of a target domain, ensuring they meet the nuanced demands of specialized applications. This fine-tuning can however drastically reduce the model's ability to generalize to out-of-distribution (OOD) scenarios, which is critical for handling dynamic and unseen data when deployed.

Various studies have been done to address this issue like Wise-FT (Wortsman et al., 2022), which averages the weights of the pretrained model and the finetuned model to improve the performance over both the models. There are other works like LADS (Dunlap et al., 2023) which uses language to extend to unseen domain for multimodal models like CLIP (Radford et al., 2021). There are also various light-weight linear probing techniques which can retain the OOD generalization strengths of the pretrained model whilst specializing to a particular domain.

Several other works involve reasoning based techniques to improve the ability of the language models to perform better in reasoning based tasks. Such methods like self-refine (Madaan et al., 2023), STaR (Zelikman et al., 2022), SCoRe (Kumar et al., 2024) etc. involve a notion of *self-correction*, where models' own outputs are passed into the model again, optionally with a ground-truth label, to get a final, refined, output. However, none of the previous works (as outlined in Section §6 have explored how such reasoning-based techniques that lead to self-correction, might also contribute to out-of-domain (OOD) generalization. Intuitively, if models can reason effectively, they *could* be better equipped to generalize OOD. So, in this work we explore how iterative refinement techniques help in improving OOD generalization in finetuned models.

We observe improvements of 1.6%-2% accuracy on the OOD dataset in both the above mentioned settings.

* * = Equal contribution

2 BACKGROUND

2.1 OOD GENERALIZATION IN PRETRAINED MODELS

OOD generalization refers to a model's ability to maintain high performance on data outside the distribution of its training set. Traditional fine-tuning approaches, such as full fine-tuning and LoRA (Low-Rank Adaptation), allow pretrained models to adapt to specific tasks or domains. However, these methods can lead to overfitting on the ID data, thereby compromising the model's ability to generalize to OOD scenarios. Techniques like Wise-FT, which averages weights between the pretrained and finetuned models, and lightweight linear probing methods have been proposed to mitigate this trade-off, retaining OOD strengths while ensuring task-specific performance.

2.2 SELF-CORRECTION MECHANISMS

Self-correction methods have gained attention for their ability to enhance reasoning in language models. Techniques like Self-Refine, STaR, and SCoRe iteratively refine model outputs to improve reasoning quality. In self-refinement, the models generate feedback on their own output and further use the feedback to refine the original output. While in the STaR method, the rationales generated for the correct answer are used to finetune the model to improve its reasoning capacity. More recent methods like SCoRe use reinforcement learning techniques to further generalise this self correction method. While these methods have demonstrated effectiveness in improving the logical coherence of language models, their impact on OOD generalization remains underexplored. Intuitively, improved reasoning should contribute to better generalization, as reasoning enables models to infer and adapt to unseen contexts.

2.3 DATASETS

We evaluated the self-correction methods for OOD generalization using both language-only and multimodal models. To study their effect on OOD generalization we finetuned both language-only models and multimodal models using STaR and self-refinement methods. We broadly consider QA tasks for our ID and OOD datasets, for their relative ease in integrating self-refine and STaR techniques. For our language-only unimodal model, we use Gemma (Team et al., 2024). As our multimodal model under test, we consider a Vision-Language Model (VLM), BLIP2 (Li et al., 2023), as it is a lightweight, functional model capable of following instructions. We treat our in-distribution unimodal dataset as CommonsenseQA (Talmor et al., 2019), and choose our out-of-distribution dataset as SQUAD (Rajpurkar et al., 2016). Our multimodal in-distribution dataset is VQAv2 (Goyal et al., 2017) and out-of-distribution dataset is HaloQuest (Wang et al., 2024). In multimodal contexts, we used VQAv2 as the ID dataset and HaloQuest as the OOD dataset. VQAv2 dataset contains real world images with questions related to the objects in the image. In contrast to this, HaloQuest dataset contains synthetically generated images, and also contains questions regarding objects not present in the image, requiring the model to express the inability to answer such questions. And since the HaloQuest dataset was released after the BLIP2 model, this dataset is truly out-of-distribution for the model. We chose 1000 samples from the train set of CommonsenseQA dataset used for finetuning with STaR. The number of samples from the train set of VQAv2 dataset used for finetuning with and without self-refinement method is 2184.

3 EXPERIMENTS

3.1 SELF-REFINEMENT FOR VLMS

We use the self-refinement technique to improve the reasoning of BLIP2 and measure how this improvement translates to OOD genralization. We generate feedback from the model regarding the output generated by the model and use the feedback to generate another output. The number of iterations used is 1 in the self-refinement feedback. We finetune the model using the self-refinement method with a learning rate of 1e-5.

3.2 STaR for LLMs

We use the STaR method to improve the reasoning of the language-only model, Gemma. In this method we generate rationales for all the samples and filter the rationales where the correct answer was generated by the model. We then finetune the pretrained gemma model on these filtered rationales. We finetune the model with STaR method with a learning rate of 5e-5.

4 Results

We did a qualitative analysis of 200 examples from the test set to understand how self-refine works both ID and in OOD scenarios as shown in table 1.

We can see from table 2 and table 1, finetuning the model with self-correction mechanisms on the in-distribution dataset improves the performance of both the in-distribution and the out-of-distribution datasets. This shows the improving the reasoning ability of the model directly correlates to OOD generalization. Some qualitative examples of how finetuning with self-refinement on in-distribution data i.e VQAv2 improves the performance on OOD dataset, HaloQuest is shown in appendix A. There is one exception in table 1 where finetuning without self-refine gives the best score. This anomaly is further analysed in the section 5.1.

Dataset	**Model Setup**	**Accuracy (%)**
OOD (HaloQuest)	Finetuned w/ self-refine	**6.25**
	Finetuned w/o self-refine	3.10
	Zero-shot BLIP2	4.17
ID (VQAv2)	Finetuned w/ self-refine	18.60
	Finetuned w/o self-refine	**24.20**
	Zero-shot BLIP2	15.00

Table 1: Qualitative analysis of model performance on OOD (HaloQuest) and ID (VQAv2) datasets.

Model Setup	**CommonSenseQA (ID) (%)**	**SQuAD (OOD) (%)**
Zero-shot	37.6	13.0
Finetuned w/ STaR	**42.0**	**14.6**

Table 2: Performance comparison on CommonSenseQA (ID) and SQuAD (OOD) datasets.

5 Analysis

5.1 Why does finetuning without self-refinement give better results

Surprisingly, from table 1, we saw that the finetuned BLIP2 model on the VQAv2 dataset w/o self-refine was able to correct 24.2% of the examples tested and the zero-shot BLIP2 model and finetuned BLIP2 model with self-refinement was only able to correct 15% and 18.6% of the examples respectively. Upon further inspection, we saw that the finetuned BLIP2 model on the VQAv2 dataset w/o self-refine was producing corrections which were basically the opposite of the original predicted answer. This lead to some incorrect correction statements, such as ”1 is not a number”, ”the bird is blue because it is on the roof” etc. Whereas the corrections made by the zero-shot BLIP2 model and the finetuned BLIP2 model with self-refinement made more logical sense. Since a majority of the questions in the VQAv2 dataset consists of yes/no questions, this lead the finetuned BLIP2 model(on the VQAv2 dataset w/o self-refine), to seemingly give a better self-correction rate than the zero-shot BLIP2 model or the finetuned BLIP2 model with self-refinement. Whereas, for the evaluation of the HaloQuest dataset using the same configurations of the BLIP2 model, we see that the BLIP2 model finetuned with self-refinement gave better improvement in self-correction than the finetuned BLIP2 model without self-refinement.

6 RELATED WORK

OOD generalization. The authors of the work Wortsman et al. (2022) propose WiSE-FT, a method combining zero-shot and finetuned model weights through ensembling to balance robustness and performance across domain shifts. This work highlights the challenge of fine-tuning pretrained models for specific tasks without sacrificing robustness, achieving significant gains in both in-distribution (ID) and OOD settings. For multimodal models,Dunlap et al. (2023) demonstrate the use of language-guided domain adaptation with their LADS framework, which improves domain generalization by utilizing verbalized domain descriptions without requiring target domain samples. This method is particularly relevant for adapting vision-language models to unseen domains.

Self-correction. Self-correction approaches have further pushed the boundaries of reasoning by iteratively improving model outputs. The STaR framework (Zelikman et al., 2022) introduces a bootstrapping mechanism that refines reasoning tasks iteratively, leveraging self-generated rationales to improve models without relying on extensive labeled data. Similarly there are other works like SELF-REFINE (Madaan et al., 2023), which iteratively refines outputs through self-feedback, showcasing its utility across diverse domains, including natural language and code generation. In addition, reinforcement learning-based methods like SCoRe (Kumar et al., 2024) have been developed to enhance self-correction, focusing on multi-turn feedback mechanisms to improve intrinsic reasoning and robustness. Other works on multimodal self-correction is either mainly post-hoc (Zhou et al., 2024) or use language-only feedback (Xu et al., 2024) to correct the outputs.

Datasets. One of the main drawbacks of evaluating VLMs on the VQA dataset for its reasoning ability is that these models are over-reliant on the language prior and do not pay enough attention to the images presented, thereby leading to hallucinations. HaloQuest dataset (Wang et al., 2024) focuses on tackling hallucinations in VLMs by using synthetic images that often deviate from real-world logic in order to challenge the drawbacks of the VLM mentioned above. VQA 2.0 (Goyal et al., 2017): Presented a balanced Visual Question Answering dataset to mitigate language biases by including complementary images for each question, making visual understanding essential. SQuAD (Rajpurkar et al., 2016): Introduced a large-scale reading comprehension dataset with over 100,000 questions based on Wikipedia passages, emphasizing span-based question answering. CommonsenseQA (Talmor et al., 2019): Developed a commonsense question-answering dataset using CONCEPTNET, containing over 12,000 multiple-choice questions designed to test commonsense reasoning.

7 FUTURE WORK

There are several avenues for future exploration to extend and deepen the findings of this study. First, conducting more extensive experiments with different types of models and datasets could help in understanding the generality of using the self-correction methods for generalization.

Second, comparing our approach to other OOD generalization techniques, such as WiSE-FT, will establish a more comprehensive benchmark and compare the efficiency of self correction methods to other established robust finetuning techniques.

Third, it is essential to investigate the scalability of our methods to larger datasets, models, and more diverse domains. Assessing the robustness and efficiency of self-correction mechanisms in these contexts will determine their feasibility for deployment in real-world scenarios.

8 CONCLUSION

In this work, we explored the impact of self-correction mechanisms on out-of-distribution (OOD) generalization for both language and multimodal models. While previous methods have demonstrated the potential of self-refinement techniques to improve reasoning capacities, their effect on OOD robustness has remained underexplored. We systematically studied self-refinement and STaR methods, highlighting their ability to enhance OOD generalization by leveraging iterative reasoning. Our initial results show promising improvements in OOD performance, indicating the potential of self-correction mechanisms to create more robust, generalizable models across diverse tasks and do-

mains. This work underscores the importance of combining fine-tuning and self-refinement to build models that can operate effectively in dynamic, real-world environments.

REFERENCES

Lisa Dunlap, Clara Mohri, Devin Guillory, Han Zhang, Trevor Darrell, Joseph E. Gonzalez, Aditi Raghunathan, and Anja Rohrbach. Using language to extend to unseen domains, 2023. URL https://arxiv.org/abs/2210.09520.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering, 2017. URL https://arxiv.org/abs/1612.00837.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, Lei M Zhang, Kay McKinney, Disha Shrivastava, Cosmin Paduraru, George Tucker, Doina Precup, Feryal Behbahani, and Aleksandra Faust. Training language models to self-correct via reinforcement learning, 2024. URL https://arxiv.org/abs/2409.12917.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. URL https://arxiv.org/abs/2301.12597.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback, 2023. URL https://arxiv.org/abs/2303.17651.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel

Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text, 2016. URL https://arxiv.org/abs/1606.05250.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil

Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118.

Zhecan Wang, Garrett Bingham, Adams Yu, Quoc Le, Thang Luong, and Golnaz Ghiasi. Haloquest: A visual hallucination dataset for advancing multimodal reasoning, 2024. URL https://arxiv.org/abs/2407.15680.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models, 2022. URL https://arxiv.org/abs/2109.01903.

Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step, 2024. URL https://arxiv.org/abs/2411.10440.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. Star: Bootstrapping reasoning with reasoning, 2022. URL https://arxiv.org/abs/2203.14465.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models, 2024. URL https://arxiv.org/abs/2310.00754.

A Appendix

a) Question: What type of tree is in this image?

b) Question: What color are the cat's eyes?

In both cases, the models finetuned with self-refinement on VQAv2 data are able to correctly output that the object in the question is not in the image, whereas models finetuned without self-refinement give the most general answers related to the object in the question even though they are not present in the image.