GPT-2 Small Fine-Tuned on Logical Reasoning Summarizes Information on Punctuation Tokens

Sonakshi Chauhan
Pr(Ai)^{2}R Group
3014475C@student.gla.ac.uk.com

Atticus Geiger
Pr(Ai)^{2}R Group
atticusg@stanford.edu

Abstract

How is information stored and aggregated within a language model performing inference? Preliminary evidence suggests that representations of punctuation tokens might serve as "summary points" for information about preceding text. We add to this body of evidence by demonstrating that GPT-2 small fine-tuned on the RuleTaker logical inference dataset aggregates crucial information about rules and sentences above period tokens.

1 Introduction

Reasoning is one of the cognitive abilities that distinguishes humans, enabling us to infer, explain, and draw conclusions. There have been numerous efforts to unveil and comprehend the internal mechanisms of Large Language Models (LLMs), and a key question that arises is: How LLMs do logical reasoning? (Among the various types of reasoning,) We aim to investigate how LLMs can engage in logical reasoning, which involves deriving conclusions based on formal principles and rules. Our goal is to contribute our findings to the ongoing discussion.

[1] and [2] first introduced the notion of "summarization motif" in language models. It can be thought as knowledge block where information is summarized. These so-called blocks can be punctuation marks, newlines, etc., and primarily the aggregation does not happen at the sentiment information tokens but the punctuations. Here we fine-tune a pre-trained GPT-2 Small and check the accumulation of information on punctuations present in a sentence with the help of the RuleTaker dataset. We do so by performing Interchange Interventions (IntInv) [3] on each token and each layer and based on the response of model to the interventions using IIA as a metric[4] for understanding whether the information is aggregated and how does the model reason.

Our main contribution lies in understanding if information is aggregated on punctuations and using the reasoning dataset we try to figure out how does a model perform reasoning.

2 Related Work

Reasoning Much previous work has focused on enhancing the reasoning capabilities of LLMs, including Chain-of-Thought (COT) [5], Tree-of-Thought (TOT) [6], and Cumulative Reasoning (CR) [7]. Other studies have attempted to understand how LLMs reason. For instance, [8] suggests that LLMs have limited generalization capabilities and that their reasoning stems from the overfiting of patterns enforced during training. While these studies provide insights into how LLMs perform reasoning [9, 10, 11, 12], they primarily focus on evaluating model outputs by manipulating datasets—a transient approach. Our study, inspired by mechanistic interpretability, intervenes not on the data but on the neurons and layers that store information and facilitate reasoning.

Interpretability Interpretability includes circuit based analysis methods[13, 14, 15, 16] , iterated null space projection ([17, 18] and Causal Effect Analysis [19, 20]. Causal abstraction [21, 22, 23, 24]

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

is a framework for interpretability in which a high-level system implements a low-level system by preserving key cause-and-effect relationships and ensuring these relationships accurately reflect the underlying causal mechanisms. This is achieved through Interchange Intervention [25, 26, 27], where source inputs are applied to base inputs of specific neuron groups in a neural network. The resulting outputs are used to make causal inferences about the model's behavior, helping to estimate the model's reasoning capabilities. We further evaluate performance using Interchange Intervention Accuracy [4]. This approach is employed in our study to assess how reasoning is done in LLMs.

3 Methods

RuleTaker dataset This rule-based reasoning dataset, tests the reasoning and implication abilities of LLMs. It includes facts and rules, followed by questions that assess whether the rules are correctly applied. Answers to these questions are labeled as True, False, or Unknown, allowing us to evaluate model predictions. The dataset is organized by depth, indicating how many rule iterations are required for correct application. An example prompt from the dataset is:

Harry is tall. Tall people are round. Is Harry round?

In the above example, the first sentence is a fact, the second sentence is a rule, and the third sentence is a question that the model answers. The dataset contains examples with several facts and rules. The rules (If..then.., All... etc) are applied to the Facts which are statements before the rules and then the final questions are answered which check the rule applications on statements are being done correctly. When we do our interpretability experiments, we will try to target a fact or an inference generated from a fact and a rule to change how the model reasons about the input.

Interchange Intervention Datasets : We curate subsets of the original RuleTaker dataset to assess model predictions and intervention effectiveness. These datasets follow the format: *base*, *source*, *base_answer*, *expected_answer*, *question*. The model is prompted with '<base> Question: <question>'. The *base_answer* is the model's original response, while the *expected_answer* is what it should output after a successful intervention. Questions are designed based on the type of intervention performed, where we have questions that check the base information is removed and questions that check whether the information from the source has been introduced.

GPT2 We use the gpt2 small, which has 85M parameters. We use this for classification and is trained on the RuleTaker dataset.

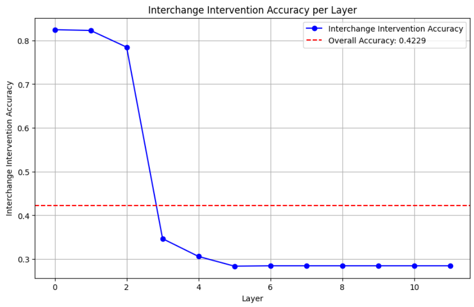
Causal Intervention We perform Interchange Intervention (also referred to as activation patching) by taking two prompts which are consistent in length but the source prompt has different inputs than the base prompt in order to target the reasoning aspect. We first capture the activations on base prompt and then do positional intervention using the activations of source prompt. We use IIA as a metric to help us identify the cases where the reasoning capability of the model was affected.

4 Experiments

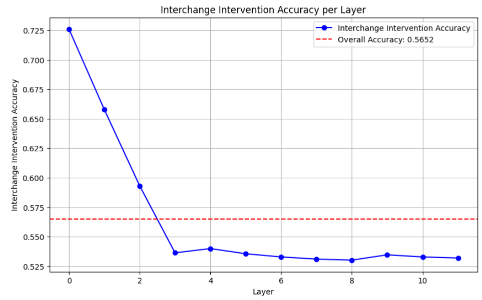
4.1 How are adjective and subject tokens processed?

Experimental Information We intervene above the subject and adjective tokens to determine how information is stored in the residual stream of the transformer. The interventions should modify the adjective or the subject in the first sentence, changing it to be the adjective or subject from the input we are patching from. If the token is indeed stored in the residual stream at the intervention location, then the models output should match the expected label when asked a question about the subject/adjective. This signifies that the reasoning capability of the model leverages information present at that location.

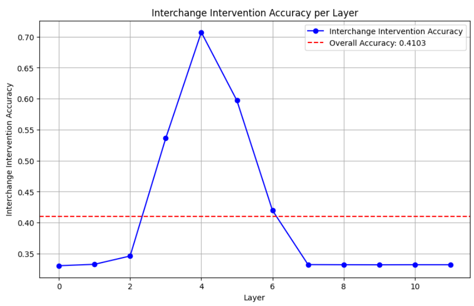
Results and discussion Figure 1 gives a visual representation. We perform experiments doing the subject and adjective interchange intervention target very specific tokens. Subject swap targeting the



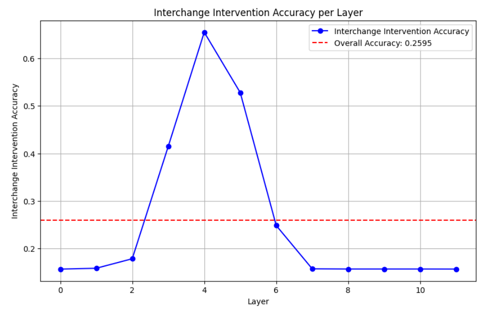
(a) Adjective Swap



(b) Subject Swap



(c) Single Sentence



(d) Double Sentence

Figure 1: The X-axis depicts the layers in the GPT2 model and the Y-axis depicts the Interchange Intervention Accuracy achieved. The higher IIA scores show that the model's output at those points was affected the most, indicating that on intervention, the model processed and gave different outputs, highlighting that reasoning was being performed and our intervention was successful. 1a and 1b are the intervention results on adjective and subject tokens. Higher initial layer accuracy shows the presence of discreet tokens in the initial layers leading to successful interventions.1c and 1d are '.' intervention results on single and double sentences. We see a high accuracy for '.' highest being at layer 4 signifying a successful interventions. The high IIA here signifies the entire information being retrieved at that position in that layer signifying information being aggregated at the dot and reasoning being performed.

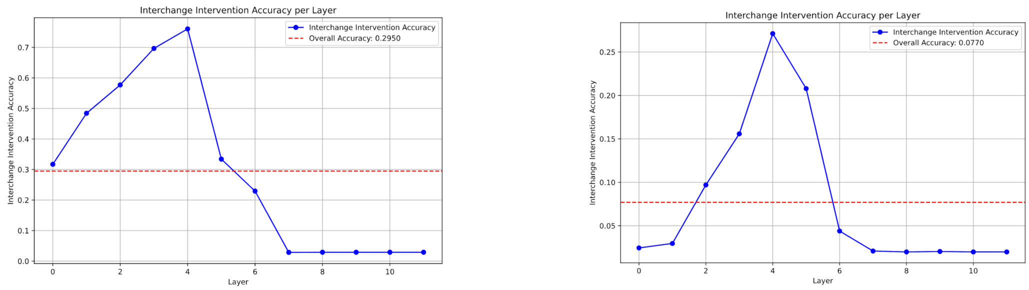
first token or token which is the main entity and Adjective swap is the last token telling about the attribute that entity has. The highest IIA for these swaps are observed for the initial layers, which is expected as the individual token representations contain the information which is passed on to the next layer by default. Information about the adjective has been moved via attention by layer 3 and information about the subject is immediately moved. The initial high accuracy being close to 80% shows that there is an impact on the model reasoning capabilities which signifies that the model considers the information giving us an idea about reasoning being performed.

4.2 The Summarization Motif

Concurrently, [1] and [2] discovered a "summarization motif" in language models where information is aggregated at tokens without semantic content, like punctuation marks or newlines. We add to the record another case of this motif.

Experimental Information We hypothesize that information about sentences is stored at the punctuation that ends the sentence. To test this hypothesis, we do three experiments: *1) Single Sentence 2) Double Sentence 3) Rule Inference*

In single sentence inference, the question the model is asked checks if the entire sentence is swapped when an intervention on the '.' is performed. For double sentence, we do the same except we intervene on two '.' tokens at the end of two sentences. We use four questions to check if both sentences have been removed and replaced. In rule inference, we aim to target the inference being generated by a rule and a fact via an intervention on the '.' at the end of the fact.



(a) Base inference removal check

(b) Source inference inserted check

Figure 2: Here we intervene on the inference being generated by the rule. In figure 2a we check if the base inference is being removed and a high accuracy in layer 4 implies the discovery of information and removal of information was successful implying our intervention was successful. In figure 2b we check if the source information is being inserted and we see a accuracy of 25% in layer 4 displaying the successful insertion and aggregation of source information.

Single Sentence Intervention We first intervene on the '.' present at the end of first sentence Figure 7a shows the accuracy is high in layer 4 getting close to 80%, depicts that the information is aggregated before the dot and can be fully found in that position.

Double Sentence Intervention In first set of experiments, we intervene on dot for both sentences 7b shows the results of the intervention. The accuracy is approximately 80% for layer 4 implying that the sentences were seen in layer 4 and the intervention was successful implying information summarization on '.'.

Rule Inference Intervention Here we target the inference generated by the rule. Initially we target the punctuation which is the end of rule checking if the end is where the entire information is summarized and the inference can be targeted there. 2a shows that the inference is targeted and removed. We see the IIA rising and getting close to 80 % which is a clear indication of the inference being removed and the information being discovered on the punctuation.

We also check if the intervention was successful, and 2b shows the results we can see an observed pattern of the IIA rising till layer 4 being the highest in layer 4 close to 30% and then decreasing. This makes clear that the interchange is successful and the models reasoning capability is affected but there is a fair chance that the information being imposed on to the model and not being actually inserted. This leads further going into the feature space and checking how the information is being represented.

5 Conclusion

Two things which we focus on this paper is to understand the reasoning capabilities of LLMs and their summarization capabilities on punctuation. Through a series of experiments with handcrafted datasets we could see how the reasoning abilities of the model were affected and specifically when the interventions were performed on '.'. It would be interesting to further perform experiments on other models to see the effects.

6 Limitations

This work only focuses on the reasoning abilities of GPT2 on a particular dataset. There is still alot of scope left for this problem to be further explored to reach accurate conclusions.

References

[1] Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. Linear representations of sentiment in large language models, 2023.

[2] Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets, 2024.

[3] Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah D. Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks, 2022.

[4] Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks, 2021.

[5] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

[6] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models, 2023.

[7] Yifan Zhang, Jingqin Yang, Yang Yuan, and Andrew Chi-Chih Yao. Cumulative reasoning with large language models, 2024.

[8] Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks, 2024.

[9] Fangzhi Xu, Qika Lin, Jiawei Han, Tianzhe Zhao, Jun Liu, and Erik Cambria. Are large language models really good logical reasoners? a comprehensive evaluation and beyond, 2023.

[10] Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. Evaluating the logical reasoning ability of chatgpt and gpt-4, 2023.

[11] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting, 2023.

[12] Guangsheng Bao, Hongbo Zhang, Linyi Yang, Cunxiang Wang, and Yue Zhang. Llms with chain-of-thought are non-causal reasoners, 2024.

[13] Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small, 2022.

[14] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

[15] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. https://distill.pub/2020/circuits/zoom-in.

[16] Lawrence Chan, Adrià Garriga-Alonso, Nicholas Goldwosky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. Causal scrubbing, a method for rigorously testing interpretability hypotheses. *AI Alignment Forum*, 2022. `https://www.alignmentforum.org/posts/JvZhhzycHu2Yd57RN/causal-scrubbing-a-method-for-rigorously-testing`.

[17] Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals, 2021.

[18] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online, July 2020. Association for Computational Linguistics.

[19] Eldar David Abraham, Karel D'Oosterlinck, Amir Feder, Yair Ori Gat, Atticus Geiger, Christopher Potts, Roi Reichart, and Zhengxuan Wu. Cebab: Estimating the causal effects of real-world concepts on nlp model behavior, 2022.

[20] Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. Causalm: Causal model explanation through counterfactual language models. *Computational Linguistics*, page 1–54, May 2021.

[21] Sander Beckers, Frederick Eberhardt, and Joseph Y. Halpern. Approximate causal abstraction, 2019.

[22] Paul K. Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M. Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Causal consistency of structural equation models, 2017.

[23] Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. Visual causal feature learning, 2015.

[24] Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability, 2024.

[25] Atticus Geiger, Kyle Richardson, and Christopher Potts. Neural natural language inference models partially embed theories of lexical entailment and negation, 2020.

[26] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. Causal mediation analysis for interpreting neural nlp: The case of gender bias, 2020.

[27] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt, 2023.

Appendix

Below are the IIA scores for positional intervention for the Toy and Messy Dataset for Full Sentence, Adjective and Subject Interventions.

Layer	Single Sentence		Adjective Swap		Subject Swap	
	Toy Dataset	Messy Dataset	Toy Dataset	Messy Dataset	Toy Dataset	Messy Dataset
1	0.3323	0.2704	0.7893	0.8245	0.5093	0.7261
2	0.2835	0.2855	0.7862	0.8227	0.5180	0.6578
3	0.4260	0.3351	0.6909	0.7846	0.4407	0.5931
4	0.6127	0.5656	0.4787	0.3466	0.5160	0.5363
5	0.7955	0.7261	0.3517	0.3059	0.6393	0.5399
6	0.6259	0.5301	0.3447	0.2837	0.7440	0.5355
7	0.5314	0.3475	0.3424	0.2846	0.7273	0.5328
8	0.3462	0.2730	0.3416	0.2846	0.6687	0.5310
9	0.3416	0.2730	0.3416	0.2846	0.6787	0.5301
10	0.3416	0.2730	0.3416	0.2846	0.6913	0.5346
11	0.3416	0.2730	0.3416	0.2846	0.6853	0.5328
12	0.3416	0.2730	0.3416	0.2846	0.6827	0.5319

Table 1: Results of Experiments on Toy and Messy Datasets Across 12 Layers

Single Sentence Per Question analysis We also did analysis on each question to better understand the behaviour on removal some information on base and insertion of some information on source sentence. 3 shows the figures for each.

Double Sentence Per Question Analysis Here we did the interventions on positions for two sentences and we have stored the results separately for 1) On dot 2) Before dot and 3) On and Before dot.

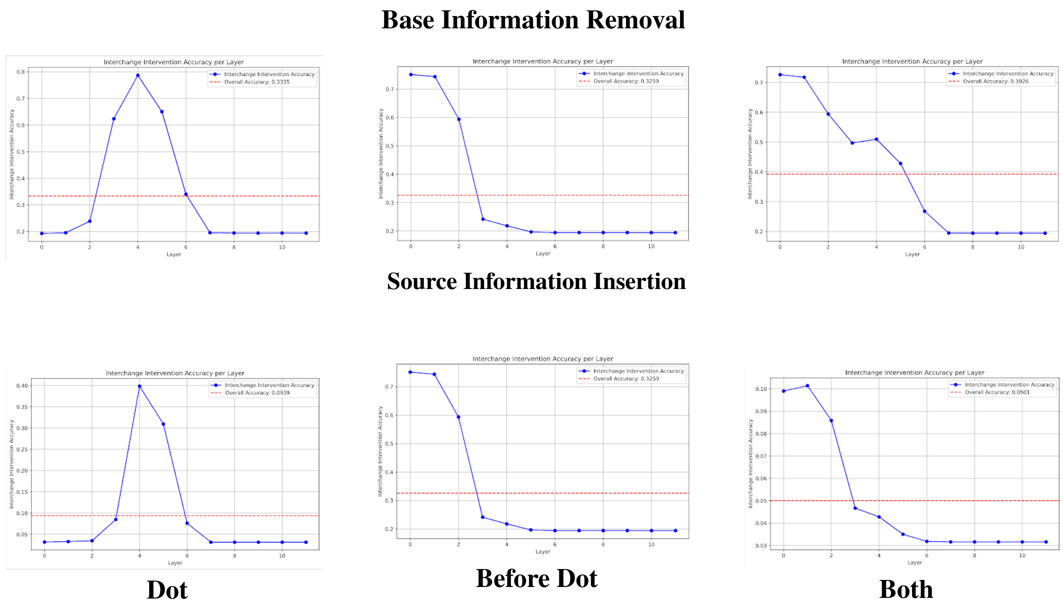
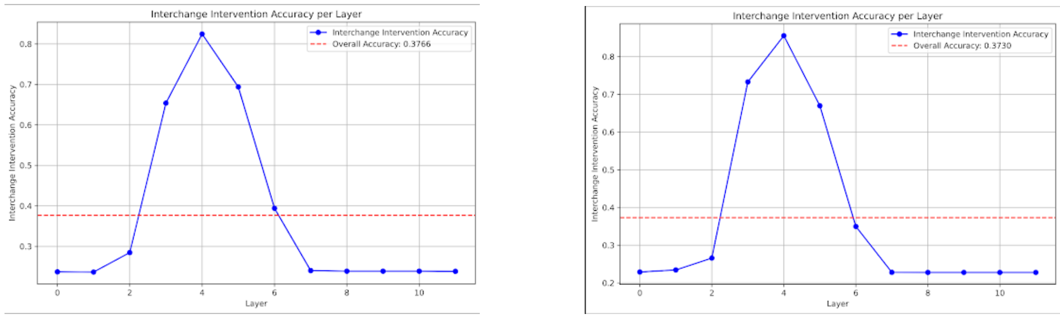


Figure 3: The IIA scores on removal and insertion of base and source information depicting how model does reasoning and how is the information being aggregated.

- On dot
- Before dot
- On and before dot

Base Information Removal



Source Information Insertion

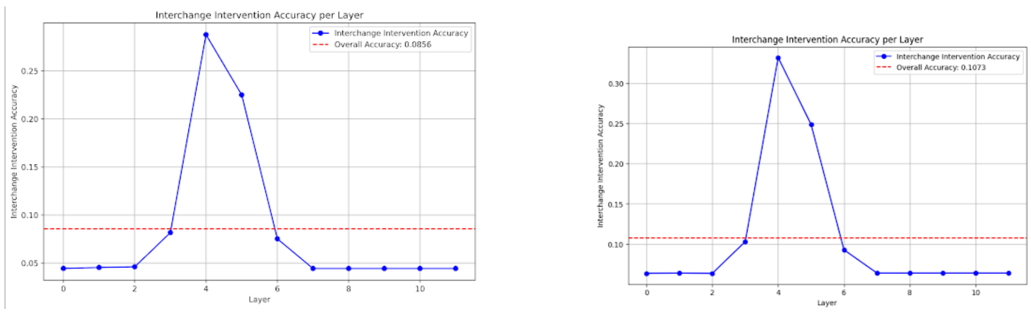
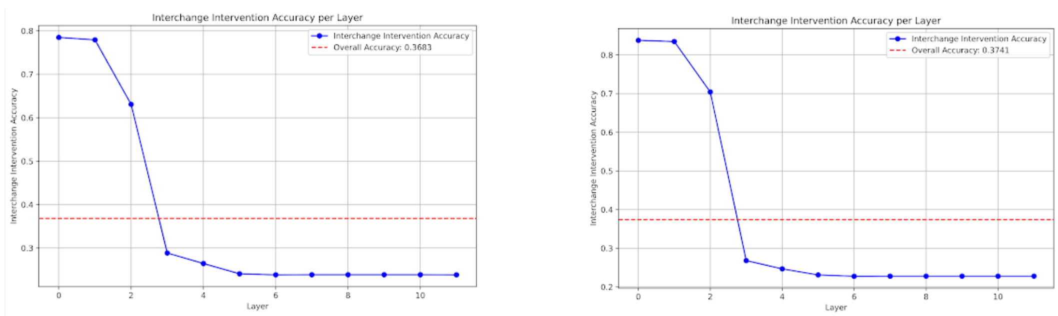


Figure 4: Base and Source Information Removal and Insertion on dot for two sentences

Base Information Removal



Source Information Insertion

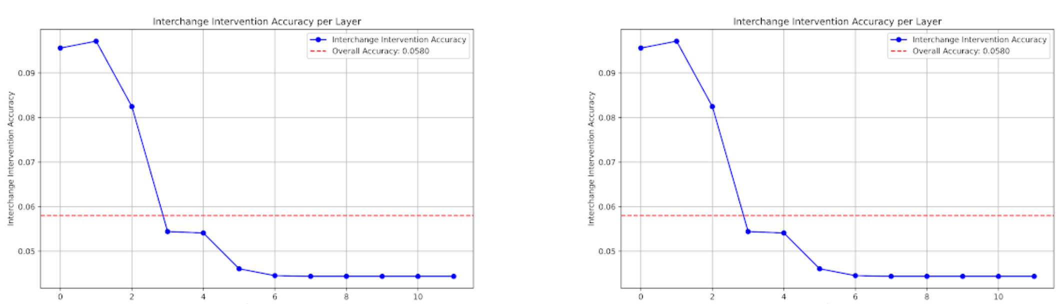
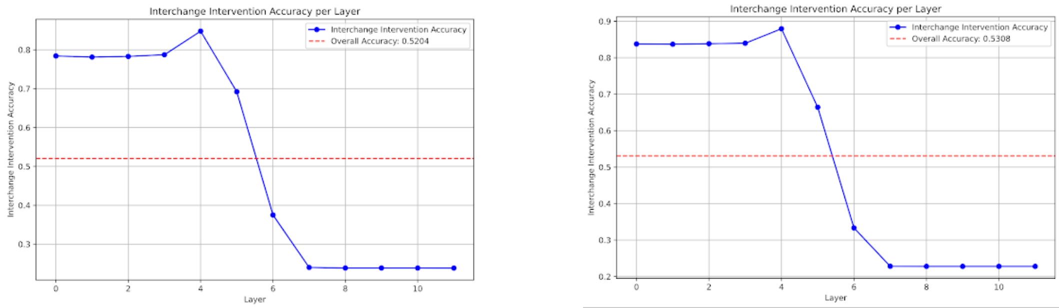


Figure 5: Base and Source Information Removal and Insertion before dot for two sentences

Base Information Removal



Source Information Insertion

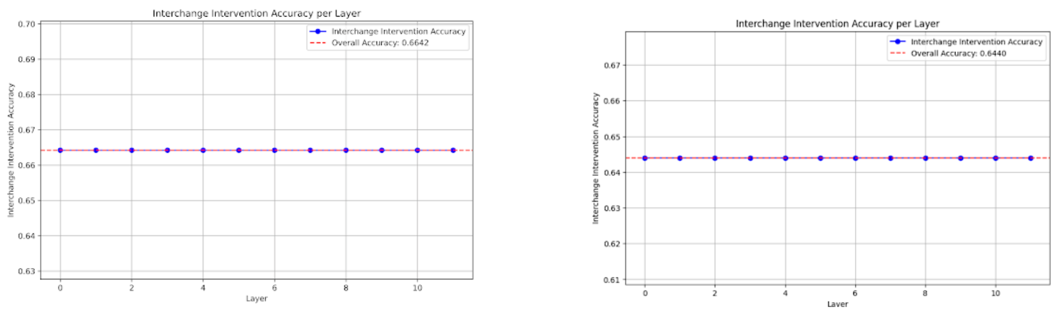
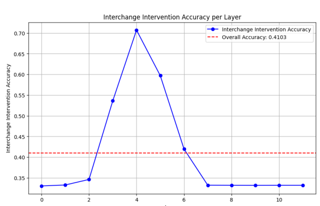
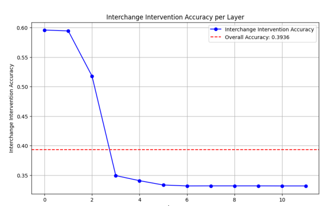


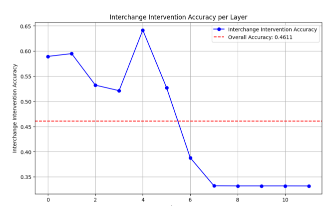
Figure 6: Base and Source Information Removal and Insertion on and before dot for two sentences



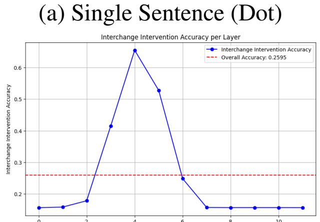
(a) Single Sentence (Dot)



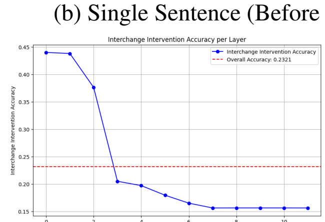
(b) Single Sentence (Before Dot)



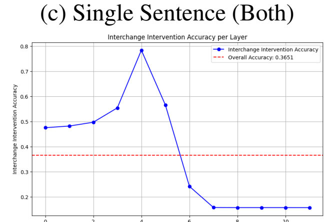
(c) Single Sentence (Both)



(d) Double Sentence (On Dot)



(e) Single Sentence (Before Dot)



(f) Double Sentence (On Dot

Figure 7: Summarization of information on and before punctuation marks