LLMs Struggle to Perform Counterfactual Reasoning with Parametric Knowledge

Khurram Yamin [1] **Gaurav Ghosal** [1] **Bryan Wilder** [1]

Abstract

Large Language Models have been shown to contain extensive world knowledge in their parameters, enabling impressive performance on many knowledge intensive tasks. However, when deployed in novel settings, LLMs often encounter situations where they must integrate parametric knowledge with new or unfamiliar information. In this work, we explore whether LLMs can combine knowledge in-context with their parametric knowledge through the lens of *counterfactual reasoning*. Through synthetic and real experiments in multi-hop reasoning problems, we show that LLMs generally struggle with counterfactual reasoning, often resorting to exclusively using their parametric knowledge. Moreover, we show that simple post-hoc finetuning can struggle to instill counterfactual reasoning ability – often leading to degradation in stored parametric knowledge. Ultimately, our work reveals important limitations of current LLM's abilities to re-purpose parametric knowledge in novel settings.

1. Introduction

Large Language Models (LLMs) internalize extensive world knowledge during pretraining, powering tasks like open-domain QA, fact retrieval, and knowledge-base completion (Petroni et al., 2019; Liu et al., 2019; Roberts et al., 2020). Benchmarks such as NaturalQuestions and HotpotQA focus on recall and multi-hop composition but do not require integrating novel premises at inference time (Yang et al., 2018; Kwiatkowski et al., 2019). For example:

> *"If Paris were located in Italy, in which country would the Eiffel Tower stand?"*

Answering this demands two capabilities: *Contextual Override*—suppressing "Paris is in France"—and *Selective Retrieval*—retaining the Paris–Eiffel Tower link. Standard benchmarks do not probe this dual requirement.

Prior work on retrieval-augmented generation illustrates how external documents can both aid and confuse LLMs when facts disagree (Lewis et al., 2021), and studies of grokking transformers show multi-hop chaining without dynamic premise integration (Wang et al., 2024a). However, none systematically evaluate *counterfactual* multi-hop reasoning against a pretrained knowledge graph.

We ask: *Can modern LLMs combine stored knowledge with in-context counterfactual premises to answer multi-hop queries correctly?*

To address this, we introduce synthetic graph-based and real-world causal benchmarks that isolate reinforcing, additive, contradictory, and irrelevant contexts; empirically analyze GPT-4o to identify *context-ignoring* and *context-overfitting* failure modes under standard, chain-of-thought, and fine-tuned prompts; demonstrate that simple post-hoc fine-tuning yields only marginal gains on counterfactual tasks and can degrade performance on factual benchmarks; and discuss implications for interactive systems, retrieval-augmented pipelines, and safety-critical applications requiring accurate conditional reasoning.

Our results reveal a fundamental limitation: despite powerful fact memorization and retrieval, current LLMs lack mechanisms to dynamically modify their internal knowledge graph in response to new or conflicting information. Addressing this gap will require novel modeling and training paradigms.

2. Related Works

Multi-Hop QA Multi-hop QA benchmarks measure a model's ability to chain stored facts but do not test integration of new premises. Yang et al. (2018) introduce a dataset for contextual multi-hop QA, and Wang et al. (2024a) show that transformers can chain relations in a synthetic "grokking" regime. Real-world analyses compare reasoning over different relation types (Yang et al., 2024) and investi-

*Equal contribution [1]Department of Machine Learning, Carnegie Mellon. Correspondence to: Khurram Yamin <khurram.yamin24@gmail.com>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

gate how intermediate entities emerge in transformer layers (Biran et al., 2024). In contrast, we examine how LLMs combine parametric memory with in-context information.

Knowledge Conflicts LLMs face *knowledge conflicts* when external context contradicts internal facts. Techniques like contextual decoding constraints (Yuan et al., 2024) and attention pruning (Li et al., 2025) enforce premises, while closed-book QA relies solely on retrieved parameters (Petroni et al., 2019; Roberts et al., 2020). Retrieval-augmented methods (REALM (Guu et al., 2020), RAG (Lewis et al., 2021), DPR (Karpukhin et al., 2020), FiD (Izacard & Grave, 2020)) merge external evidence, and finer-grained approaches like AdaCAD (Wang et al., 2024b) and CD2 (Jin et al., 2024) balance parametric and contextual sources. However, none target selective retention of stored knowledge under counterfactual premises.

Causal Reasoning and Counterfactuals in NLP Counterfactual reasoning with LLMs remains challenging. Benchmarks such as CLadder (Jin et al., 2023), QRData (Liu et al., 2024), and CounterBench (Chen et al., 2025) reveal poor performance on formal causal and counterfactual tasks. Critics argue LLMs often "parrot" causal patterns from data rather than infer causally (Zečević et al., 2023), and Yamin et al. (2024) detail failures like over-reliance on parametric knowledge. While methods exist for generating faithful explanations (Gat et al., 2024) or counterfactual tokens (Chatzi et al., 2024), a systematic study of multi-hop counterfactual integration is still lacking.

3. Experiments on Real-World LLMs

3.1. Problem Formulation for a Causal Case

We inject contextual information at test time. For example, if the model's parametric knowledge includes $Y_1 \to Y_2$ and prompt asserts $Y_2 \to X_0$, can it deduce $Y_1 \to X_0$? Results shown in main paper are for GPT 4o (OpenAI et al., 2024), and LLama 3.1 8B (Grattafiori et al., 2024) results can be found in the Appendix. We categorize contextual edits into four partitions (all prompts and code are in the appendix/supplement):

1. **Scenario 1 (Reinforcing Prior Knowledge):** Prompting the LLM with a relationship already present in its prior knowledge graph, thereby reinforcing an existing edge. Example: Given excessive rain causes flooding, query whether excessive rain causes infrastructure damage.

2. **Scenario 2 (Adding New Information):** Prompting the LLM with scenario-specific information necessary to answer the query, but absent from its parametric knowledge graph, akin to adding an edge. Example: Informing the LLM that excessive rain causes Timmy to eat vegetables, and querying whether excessive rain improves Timmy's health.

3. **Scenario 3 (Contradicting Prior Knowledge):** Prompting the LLM with information that strongly contradicts its existing parametric knowledge, equivalent to replacing an edge in its prior knowledge graph. Example: Informing the LLM that excessive rain causes desert expansion and querying whether excessive rain promotes cactus growth.

4. **Scenario 4 (Irrelevant Information):** Prompting the LLM with unrelated information, akin to providing an edge from a disconnected knowledge graph. Example: Informing the LLM that fatty food consumption causes heart attacks, and querying whether excessive rain leads to flooding.

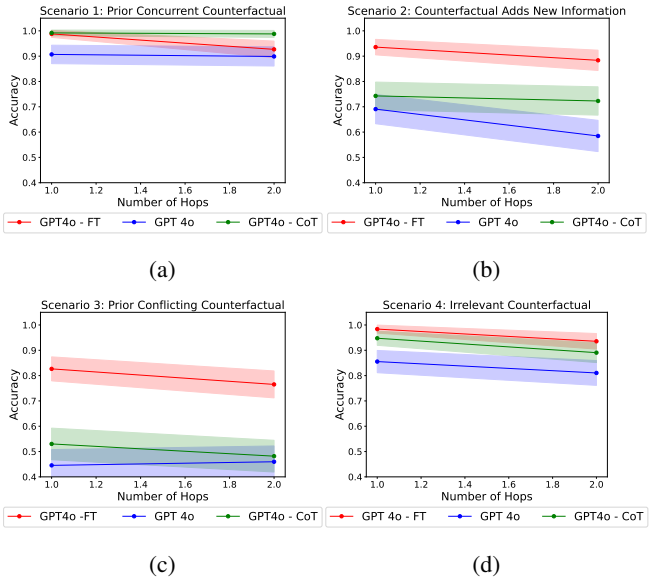

Figure 1. Causal Counterfactual Plots comparing standard GPT-4o, GPT-4o CoT and GPT-4o Fine tuned Results. (a) Counterfactual Reinforces Prior, (b) Counterfactual Adds new Information, (c) Counterfactual Conflicts with Prior, (d) Counterfactual is Irrelevant to Prior and Query. 95 % CI is shown.

3.2. Prompting Methods

We compare three strategies: *Standard*—direct causal query; *CoT*—chain-of-thought prompting; *FT*—fine-tuning on counterfactual examples with CoT explanations (160 examples, hyperparameters in Appendix). Models are queried over either 1 or two counterfactual hops for simplicity.

3.3. Results

We ask a binary cause/non-cause question (random baseline 50%). Figure 1 reports GPT-4o accuracy across scenarios.

3.3.1. Success when Context Does Not Oppose Prior

We see that in Scenario 1 where we reinforce prior knowledge, Figure 1a shows that standard prompting achieves around 85% accuracy, with *CoT* accuracy rises above 95%, and *FT* further closes errors to near 100% with low variance across trials. These results together demonstrate that when contextual information reinforces existing knowledge, modern LLMs like GPT-4o can reliably utilize their parametric knowledge without being misled by the prompt. Such robustness provides a useful baseline against which to compare the more challenging counterfactual scenarios in Scenarios 2 and 3.

3.3.2. Failure when Context Adds New Information or Contradicts Prior

In the adding-information scenario 2 plotted in Figure 1b, CoT yields ≈80% accuracy (30% above random) but shows variability across relation types; FT improves this to ≈85% by reinforcing task-specific patterns. Under conflicting premises plotted in Figure 1c, CoT performance collapses to near the 50% baseline, with responses oscillating between stored and contextual facts. Fine-tuning partially mitigates this, lifting accuracy to ≈75%, yet significant errors persist, highlighting the difficulty of overriding strong parametric priors. As such, it becomes clear that the greatest hurdle we encounter is information that conflicts with our prior. In a sense, scenario 2 , the second worst performance, where we add new information can be viewed as a weaker version of prior conflicting information. For example, if we look back at the previous example where we inform the LLM that excessive rain causes Timmy to eat vegetables, the LLM likely has a weak prior that rain does not cause kids to eat vegetables in general.

3.4. Mixed Results for Irrelevant Information

In the irrelevant-information scenario 4 plotted in Figure 1b, standard prompts receive around ≈65% accuracy, while CoT recovers to ≈75% and FT boosts performance to ≈90%. It should be noted that LLama 3.1 8B results (appendix) show finetuning reduces performance. As such, we see mixed signals in this regime, possibly owing to the large difference in parameter sizes in these models.

4. Conceptual Experiments in Toy Setting

In the previous section, we demonstrated that state-of-the-art LLMs can struggle to perform counterfactual reasoning tasks and that simple fine-tuning can be insufficient to overcome this. In this section, we perform experiments in a more controlled setting to better understand the origins of this limitation. This setting can be viewed as a generalization of the causal inference setting studied in the previous section – where previously we considered only a single relation (of causality).

4.1. Setup

We adapt Wang et al. (2024a)'s synthetic knowledge graph (entities $\mathcal{E}$, relations $\mathcal{R}$). Facts are Atomic or Multi-Hop paths. **Atomic Facts** represent only a single edge in the knowledge graph (i.e. ent1 $->$ r1 $->$ ent2). On the other hand, **Multi-Hop Facts** represent a composition of two atomic facts (i.e. ent1 $->$ r1 $->$ ent2 $->$ r2 $->$ ent3). Pretraining follows Wang et al. (2024a) (including transformer layer setup), ensuring greater than 99% validation accuracy on multi-hop composition. We perform the pretraining stage using the hyper-parameters reported in Wang et al. (2024a). We extend this to counterfactual multi-hop QA: given a novel atomic counterfactual premise, answer a multi-hop query assuming it's true, testing integration with parametric knowledge.

Counterfactual Finetuning For simplicity, we concentrate our work on two-hop settings. We then have three possible types of counterfactual multi-hop queries. **Hop 1-Relevant** are those in which the counterfactual premise supplied is relevant to the first hop of the multi-hop reasoning problem. **Hop 2-Relevant** are those in which the counterfactual premise is relevant to the second hop of the multi-hop reasoning problem. Both of these are classified as **Related** queries. Finally, **Unrelated** queries are those in which the counterfactual premise *should not* effect the outcome of the final query. We ensure that the counterfactual fine-tuning dataset is balanced across these three catagories to mitigate shortcut learning.

4.2. Findings

Counterfactual Finetuning Induces Shortcuts Figure 2a shows that, although finetuning quickly drives accuracy on relevant counterfactual queries to high levels, performance on irrelevant queries remains poor—even when training on a balanced mix. Our results point to the learning of a shortcut solution whereby the model becomes induced to *always* override its contextual knowledge even when the counterfactual premise is irrelevant. We additionally examine the sensitivity of this behavior to finetuning hyperparameters in Figure 2b, finding that it arises across a range of learning rates.

Performance Degradations Not a Result of Format Change One potential explanation for the difficulty in

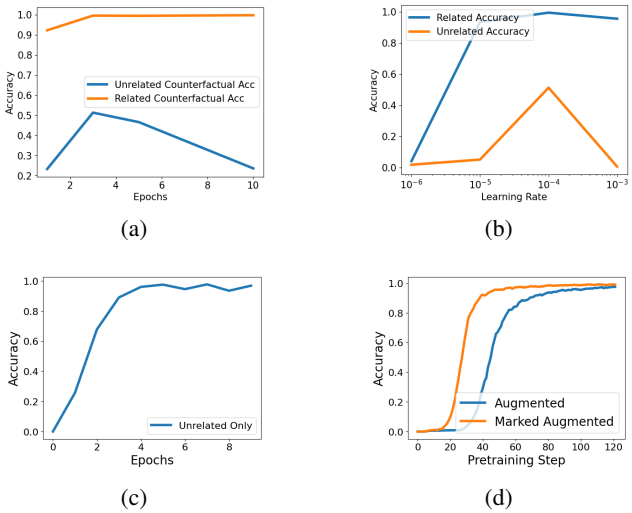

Figure 2. **Exploration in a Conceptual Setting.** (a) Counterfactual finetuning accuracy vs. epochs (stratified by query relatedness). (b) Effect of learning rate on counterfactual accuracy. (c) Accuracy on unrelated counterfactual examples when finetuning on "Unrelated Only" data. (d) Accuracy on unrelated counterfactual examples when performing counterfactual-augmented pretraining (for clarity, we omit the related counterfactual case, which is near 100% and similar across methods).

introducing counterfactual reasoning could be the mismatch in format between the original presentation of knowledge and the counterfactual style prompts. To isolate the impact of prompt format, we finetuned on an "Unrelated Only" dataset (Figure 2c), where counterfactual premises never affect the correct answer—so the model only needs to adapt to a new prompt style while relying on its existing parametric knowledge. Achieving near perfect accuracy shows that format changes (or generic forgetting) don't drive the drop in performance; instead, the core difficulty lies in teaching the LLM to *conditionally override* its parametric knowledge without corrupting it.

Effect of Incorporating Counterfactual Data in Pretraining We explored adding counterfactual data to pretraining (augmenting 20% of KG edges, Figure 2d). Incorporating counterfactual data indistinguishably from other pretraining data (*Augmented* in plot) or marking it with a special token (*Marked-Augmented* in plot) in pretraining induce good accuracies across the different groups of *counterfactual* points (i.e unrelated and related). *Marked-Augmented* performances converges faster than the *Augmented* case. However, while we find that interleaving counterfactual data into pretraining induces counterfactual reasoning, regular fact learning is slowed as can be seen in Figure 3 where we plot factual *Marked-Augmented* performances . This suggests a tradeoff between learning counterfactual reasoning and facts at pretraining-time. It is possible that counterfactual tasks introduce interfering gradients which contribute to the suppression of parametric knowledge.

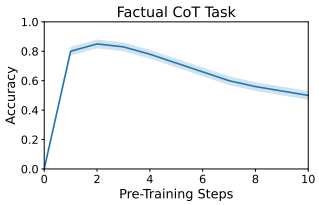

Figure 3. **Fact Learning when Counterfactuals are Incorporated in Pretraining** Fact Learning is slowed when counterfactual data is interleaved into pretraining (with a special token). 95 % CI shown.

5. Discussion

Our work highlights LLM challenges in counterfactual reasoning requiring dynamic integration of parametric and contextual knowledge. Experiments show LLMs often default to parametric knowledge or fail to balance contextual override with selective retrieval. Simple finetuning is limited, inducing shortcuts or degrading knowledge. Pretraining with counterfactual data, while improving such reasoning, can also harm factual task performance. These findings point to a core limitation: current LLMs lack robust mechanisms for on-the-fly, conditional use of their parametric knowledge.

6. Limitations

While our study sheds light on the challenges LLMs face in integrating parametric knowledge with novel counterfactual premises, it is subject to several limitations. Our synthetic benchmarks abstract away many complexities of real-world reasoning. In the toy setting, counterfactual premises are expressed as single-edge edits to a static knowledge graph and queries are limited to two-hop chains. Many real-world scenarios often involve multi-predicate interactions, ambiguous or probabilistic relationships, and noisy or conflicting evidence from multiple sources. What is significant is that we show how models struggle even under these simpler circumstances.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Biran, E., Gottesman, D., Yang, S., Geva, M., and Globerson, A. Hopping too late: Exploring the limitations of large language models on multi-hop queries, 2024. URL https://arxiv.org/abs/2406.12775.

Chatzi, I., Singh, S., Ilharco, G., Kollar, T., and Kaelbling, L. P. Counterfactual token generation in large language models, 2024.

Chen, Y., Singh, V. K., Ma, J., and Tang, R. Counterbench: A benchmark for counterfactuals reasoning in large language models, 2025.

Gat, Y. O., Calderon, N., Feder, A., Chapanin, A., Sharma, A., and Reichart, R. Faithful explanations of black-box NLP models using LLM-generated counterfactuals. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=UMfcdRIotC.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X. E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M.,

Liu, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N. P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., and Ma, Z. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M.-W. Realm: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 3929–3938. PMLR, 2020. URL http://proceedings.mlr.press/v119/guu20a/guu20a.pdf.

Izacard, G. and Grave, É. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*, 2020. URL https://arxiv.org/abs/2007.01282.

Jin, Z., Chen, Y., Leeb, F., Gresele, L., Kamal, O., Lyu, Z., Blin, K., Gonzalez Adauto, F., Kleiman-Weiner, M., Sachan, M., and Schölkopf, B. CLadder: Assessing Causal Reasoning in Language Models. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, pp. 12036–12058, 2023.

Jin, Z., Cao, P., Chen, Y., Liu, K., Jiang, X., Xu, J., Li, Q., and Zhao, J. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. In *Proceedings of LREC-COLING 2024*, pp. 16867–16878, 2024. URL https://aclanthology.org/2024.lrec-main.1466/.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550/.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Kelcey, M., Devlin, J., Lee, K., Toutanova, K. N., Jones, L., Chang, M.-W., Dai, A., Uszkoreit, J., Le, Q., and Petrov, S. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*, 2019.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. URL https://arxiv.org/abs/2005.11401.

Li, G., Chen, Y., and Tong, H. Taming knowledge conflicts in language models, 2025. URL https://arxiv.org/abs/2503.10996.

Liu, X., Wu, Z., Wu, X., Lu, P., Chang, K.-W., and Feng, Y. Are LLMs Capable of Data-based Statistical and Causal Reasoning? Benchmarking Advanced Quantitative Reasoning with Data. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 9215–9235, August 2024.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach, 2019. URL https://arxiv.org/abs/1907.11692.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N.,

Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., and Riedel, S. Language models as knowledge bases?, 2019. URL https://arxiv.org/abs/1909.01066.

Roberts, A., Raffel, C., and Shazeer, N. How much knowledge can you pack into the parameters of a language model?, 2020. URL https://arxiv.org/abs/2002.08910.

Wang, B., Yue, X., Su, Y., and Sun, H. Grokked transformers are implicit reasoners: A mechanistic journey to the edge of generalization, 2024a. URL https://arxiv.org/abs/2405.15071.

Wang, H., Prasad, A., Stengel-Eskin, E., and Bansal, M. Adacad: Adaptively decoding to balance conflicts between contextual and parametric knowledge, 2024b. URL https://arxiv.org/abs/2409.07394.

Yamin, K., Gupta, S., Ghosal, G. R., Lipton, Z. C., and Wilder, B. Failure Modes of LLMs for Causal Reasoning on Narratives, 2024.

Yang, S., Kassner, N., Gribovskaya, E., Riedel, S., and Geva, M. Do large language models perform latent multi-hop reasoning without exploiting shortcuts?, 2024. URL https://arxiv.org/abs/2411.16679.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W. W., Salakhutdinov, R., and Manning, C. D. Hotpotqa: A dataset for diverse, explainable multi-hop question answering, 2018. URL https://arxiv.org/abs/1809.09600.

Yuan, X., Yang, Z., Wang, Y., Liu, S., Zhao, J., and Liu, K. Discerning and resolving knowledge conflicts through adaptive decoding with contextual information-entropy constraint. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3903–3922. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.234. URL https://aclanthology.org/2024.findings-acl.234/.

Zečević, M., Willig, M., Dhami, D. S., and Kersting, K. Causal parrots: Large language models may talk causality but are not causal. *Transactions on Machine Learning Research (TMLR)*, 2023. ISSN 2835-8856.

Appendix

All code is in supplementary materials

A. Causal Case

A.1. Full Prompt Library

Below are every prompt template employed in our experiments. Placeholders (‘<. . . >‘) were dynamically filled in the code with the appropriate event names or lists, as described below.

Placeholder Definitions:

- ‘<Cause>‘: The event assumed to be the cause in the causal-query prompt (e.g., “Rain”).
- ‘<Effect>‘: The event assumed to be the effect in the causal-query prompt (e.g., “Flooding”).
- ‘<Query>‘: The event whose causal relationship is being queried (e.g., “Crop Growth”).
- ‘<Target>‘: The event we are checking if ‘<Query>‘ causes (e.g., “Food Supply Increase”).
- ‘<PrevEvent>‘: The most recently generated event in the chain when expanding synthetic event sequences.
- ‘<ExistingList>‘: A serialized list of all previously generated events in that chain (as a Python list string), used to prevent duplicates.

A.2. A. General Causal-Query Prompts

1. Standard (Direct) Prompt

```
Assume that <Cause> causes <Effect>.
Given this statement, does <Query> logically cause <Target>?
Only output your answer as <answer>Yes/No</answer>”
Feel free to extrapolate beyond the given context.
```

2. Chain-of-Thought (CoT) Prompt

```
Assume that <Cause> causes <Effect>.
Please think step by step
to decide whether <Query> logically causes <Target>.
Then output your reasoning and your answer: <answer>Yes/No</answer>
```

> **Fine-Tuned** model runs used these exact templates but pointed at the fine-tuned checkpoint.

A.3. Event Chain Generation Prompts

1. Causal Chain Expansion

```
generate an event that is caused by the event: <PrevEvent>
(if this event is empty pick a random choice, do not simply
ever output the text "if this event is empty pick a random choice"),
for example Cancer -> Death or Obesity -> Bad Heart Health.
Use an arrow between the two events such that <PrevEvent> is
the first item in the chain. Make sure the event you generate
is not already in the list: <ExistingList>. Make sure the
output only includes the two events with an arrow between them.
```

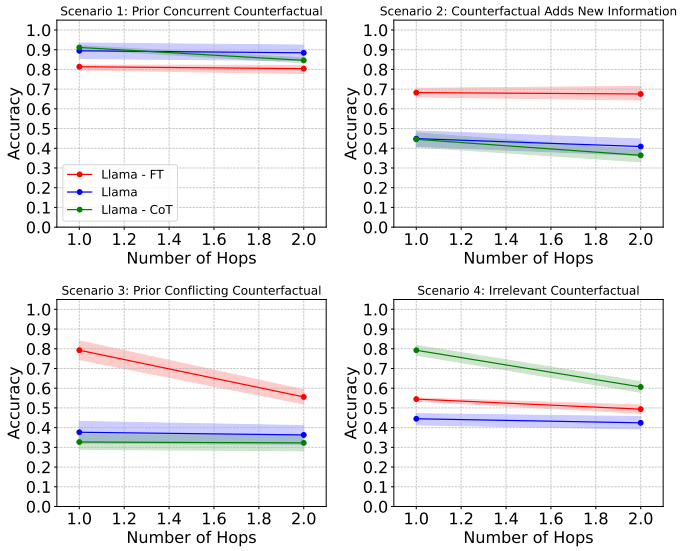

Figure 4. LLama 3.1 8B Causal Counterfactual Plots, Scenario 1 Counterfactual Reinforces Prior, Scenario 2- Counterfactual Adds new Information, (c) Scenario 3- Counterfactual Conflicts with Prior, Scenario 4- Counterfactual is Irrelevant to Prior and Query. 95 % CI is shown.

2. Anticausal Chain Expansion

```
generate an event that is anticausal to <PrevEvent>
(meaning the effect is actually the opposite of what it should be),
for example Cancer -> Longer Life or Obesity -> Weight Loss.
Use an arrow between the two events. Make sure the event you
generate is not already in the list: <ExistingList>. Output
only the two events with an arrow between them.
```

3. "Irrelevant" Transition Event

```
generate a random event that is a result of <PrevEvent>
(meaning this is a specific strange scenario), for example
Rain -> Increased Chocolate Eating or Obesity -> Warm Weather.
This event should not typically be a result of <PrevEvent>.
Use an arrow between the two events. Start with <PrevEvent>
and separate the events with an ->.
```

4. Post-Transition Chain Expansion

```
generate an event that is caused by the event: <PrevEvent>
(if this event is empty pick a random choice), for example
Cancer -> Death or Obesity -> Bad Heart Health.
Use an arrow between the two events such that <PrevEvent>
is the first item in the chain. Make sure the event you
generate is not already in the list: <ExistingList>.
Output only the two events with an -> between them.
```

B. Llama Results

In these results in Figure 4 , we see similar trends to the GPT-4o results except that we see Fine Tuning producing decreased accuracy for irrelevant counterfactuals. This mirrors what we see in the toy example.

C. GPT Finetuning HyperParameters

Trained tokens: 38,754 Epochs: 3 Batch size: 1 LR multiplier: 2

D. LLama 3.1 8B Finetuning HyperParameters

Epochs: 5, LoRA Rank:8,Learning Rate: 0.0001, Max Context Length:8192

E. LLM Usage

Our paper focuses on examining the reasoning abilities of Language Models. We do not use a language model to assist with writing.

F. Compute

In our toy experiment, we used 4 gpus per experiment. The GPU used was NVIDIA A6000 and we had a total of 72 GPU hours.