Competition of Mechanisms: Tracing How Language Models Handle Facts and Counterfactuals

Francesco Ortu[1,4*] **Zhijing Jin**[2,3*] **Diego Doimo**[4]

Mrinmaya Sachan[3] **Alberto Cazzaniga**[4†] **Bernhard Schölkopf**[2†]

[1]University of Trieste, [2]MPI for Intelligent Systems, [3]ETH Zürich, [4]AREA Science Park

{francesco.ortu, diego.doimo, alberto.cazzaniga}@areasciencepark.it

{jinzh, msachan}@ethz.ch

bs@tue.mpg.de

1 Introduction and Related Works

Large language models (LLMs) have achieved remarkable performance in various NLP tasks, revolutionizing applications across multiple domains (Brown et al., 2020; Touvron et al., 2023; OpenAI, 2023; Anil et al., 2023, *inter alia*). However, their black-box nature poses significant challenges to our scientific understanding of their inner workings. This gap between empirical success and mechanistic comprehension has led to a growing focus on interpretability research, which aims to decode the internal processes of these complex models.

Interpretability research in LLMs has primarily followed two main trajectories: interpreting representations and decoding specific mechanisms. The first approach focuses on understanding what information is encoded in model states (Belinkov et al., 2017; Conneau et al., 2018; Hewitt and Manning, 2019). These studies have revealed that LLMs capture a rich array of linguistic and world knowledge within their hidden states. The second approach, mechanistic interpretability, aims to uncover the specific operations learned by LLMs (Olsson et al., 2022; Geva et al., 2023; Meng et al., 2022; Hanna et al., 2023, inter-alia). For instance, Olsson et al. (2022) identified induction heads responsible for the copy mechanism, a basic yet crucial operation in LLMs. Similarly, studies by Geva et al. (2023) and Meng et al. (2022) have shed light on how LLMs mechanistically recall factual information, showing that early MLP layers enrich subject embeddings while late attention blocks select and write factual information.

Despite these advancements in understanding individual mechanisms, less attention has been paid to how these mechanisms interact and compete within the model's decision-making process. This gap in our knowledge is particularly crucial when LLMs face scenarios requiring them to balance multiple, potentially conflicting sources of information – such as when presented with counterfactual statements that contradict their pre-trained knowledge.

In this study, we propose a novel formulation of *competition of mechanisms* to investigate the interplay between multiple mechanisms in LLMs. Our work specifically focuses on how one mechanism becomes dominant in the final prediction by winning this *competition*. We examine the interaction between two well-studied mechanisms: factual knowledge recall and in-context adaptation to counterfactual statements. This approach allows us to explore how LLMs navigate the tension between their pre-trained knowledge and new information presented in the input context. Based on the latest tools to inspect each of these two mechanisms (Nostalgebraist, 2020; Wang et al., 2023; Geva et al., 2023), we then unfold *how and where* the competition of the two mechanisms happen. Our analysis spans both macroscopic (e.g., layer-level) and microscopic (e.g., attention head) views, providing a comprehensive picture of how information flows and competes within the model architecture.

2 Problem Setup and Methods

We design a task to incorporate the competition of mechanisms by pairing factual statements such as “iPhone was developed by Apple.” with corresponding counterfactual statements, as “iPhone was developed by Google.”. We compose prompts adding the counterfactual statements as a false definition of the factual sentence, allowing us to trace the competition between factual (t_{fact}) and counterfactual (t_{cofa}) tokens, such as *“Redefine: iPhone was developed by Google. iPhone was developed by ___”*. We utilize the CounterFact dataset (Meng et al., 2022), selecting 10,000 examples

* Equal contributions
† Co-supervision

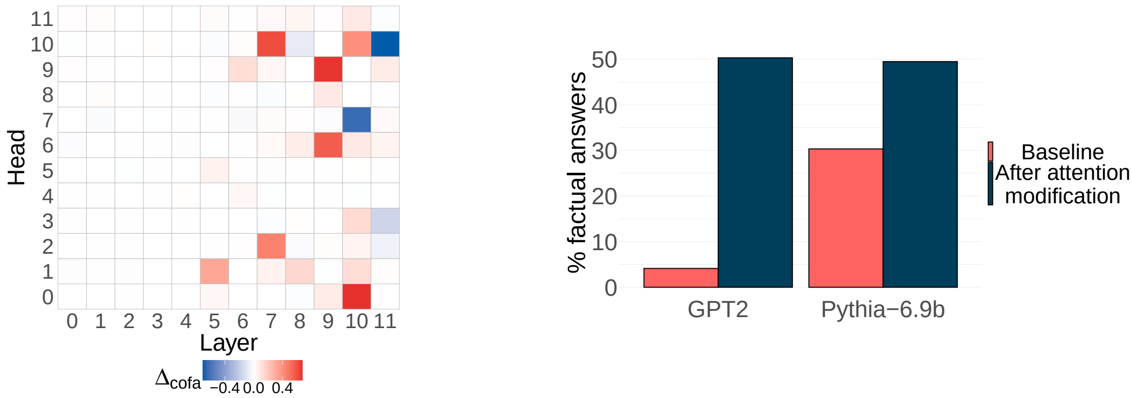

Figure 1: **Summary of our main results.** Few localized attention heads are responsible to modulating the competition between factual recall and counterfactuals redefinition. On the left, the direct contribution to $\Delta_{\text{cofa}} := \text{HeadLogit}(t_{\text{cofa}}) - \text{HeadLogit}(t_{\text{fact}})$ of all heads in GPT-2. Heads favoring t_{fact} are colored in blue, and those favoring t_{cofa} in red. On the right, the factual recall accuracy before and after modifying the target heads in GPT-2 and Pythia-6.9B. We up-weight the heads that favor fact, two in GPT-2 and three in Pythia-6.9B.

where attributes are single tokens and the model completes sentences accurately. To analyze token preferences across model components, we project hidden representations to the vocabulary space using an unembedding matrix W_U (Halawi et al., 2023; Geva et al., 2023; Dar et al., 2023; Geva et al., 2022; Nostalgebraist, 2020). We also employ attention matrix modification techniques to further elucidate information flow within LLMs, intervening on target attention heads and measuring the effect in the model's performance. Our primary focus is on the GPT-2 small model (Radford et al., 2019), aligning with previous interpretability studies (Meng et al., 2022; Wang et al., 2023; Conmy et al., 2023; Hanna et al., 2023). To demonstrate generalizability, we provide supplemental results for Pythia-6.9B (Biderman et al., 2023), enhancing the robustness of our findings across LLMs of different architectures and scales.

3 Results and Findings

Using these methods, we assess the contributions of various model components, both from a macroscopic view (e.g., each layer) and a microscopic view (e.g., attention heads), and identify critical positions and attention heads involved in the competition of the two mechanisms. Moreover, we locate a few localized positions of some attention head matrices that can significantly control the strength of the factual mechanism. We summarize our main findings as follows:

1. In early layers, the factual attribute is encoded in the subject position, while the counterfactual is in the attribute position;
2. The attention blocks write most of the factual and counterfactual information to the last position;
3. All the highly activated heads attend to the attribute position regardless of the specific type of information they promote. The factual information flows by penalizing the counterfactual attribute rather than promoting the factual one;
4. We find that we can up-weight the value of a few very localized values of the attention head matrix to strengthen factual mechanisms substantially.

4 Conclusion

Our study introduces the concept of "competition of mechanisms" as a novel interpretability framework for understanding how LLMs handle multiple, potentially conflicting mechanisms. This approach provides valuable insights into the inner workings of language models, particularly in scenarios where they must navigate between pretrained knowledge and conflicting contextual information. Our findings reveal that the suppression of counterfactual information plays a more significant role than the promotion of factual information in the model's decision-making process. This insight, along with our discovery of localized attention positions that control the strength of the factual mechanism, opens up new possibilities for targeted model fine-tuning and optimization.

References

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805. 1

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics. 1

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR. 2

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. 1

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. *CoRR*, abs/2304.14997. 2

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics. 1

Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. 2023. Analyzing transformers in embedding space. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 16124–16170. Association for Computational Linguistics. 2

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. *CoRR*, abs/2304.14767. 1, 2

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. 2

Danny Halawi, Jean-Stanislas Denain, and Jacob Steinhardt. 2023. Overthinking the truth: Understanding how language models process false demonstrations. *CoRR*, abs/2307.09476. 2

Michael Hanna, Ollie Liu, and Alexandre Variengien. 2023. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. *CoRR*, abs/2305.00586. 1, 2

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *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)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics. 1

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *arXiv preprint arXiv:2202.05262*. 1, 2

Nostalgebraist. 2020. interpreting gpt: the logit lens. Accessed: Nov 2023. 1, 2

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. 2022. In-context learning and induction heads. *CoRR*, abs/2209.11895. 1

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774. 1

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8). 2

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971. 1

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. 1, 2