Analysing the Residual Stream of Language Models Under Knowledge Conflicts

Yu Zhao[1] **Xiaotang Du**[1] **Giwon Hong**[1] **Aryo Pradipta Gema**[1] **Alessio Devoto**[3]
Hongru Wang[2] **Xuanli He**[4] **Kam-Fai Wong**[2] **Pasquale Minervini**[1,5]
[1]University of Edinburgh [2]The Chinese University of Hong Kong
[3]Sapienza University of Rome [4]University College London [5]Miniml.AI
{yu.zhao, p.minervini}@ed.ac.uk

Abstract

Large language models (LLMs) can store a significant amount of factual knowledge in their parameters. However, their parametric knowledge may conflict with the information provided in the context. Such conflicts can lead to undesirable model behaviour, such as reliance on outdated or incorrect information. In this work, we investigate whether LLMs can identify knowledge conflicts and whether it is possible to know which source of knowledge the model will rely on by analysing the residual stream of the LLM. Through probing tasks, we find that LLMs can internally register the signal of knowledge conflict in the residual stream, which can be accurately detected by probing the intermediate model activations. This allows us to detect conflicts within the residual stream before generating the answers without modifying the input or model parameters. Moreover, we find that the residual stream shows significantly different patterns when the model relies on contextual knowledge versus parametric knowledge to resolve conflicts. This pattern can be employed to estimate the behaviour of LLMs when conflict happens and prevent unexpected answers before producing the answers. Our analysis offers insights into how LLMs internally manage knowledge conflicts and provides a foundation for developing methods to control the knowledge selection processes.

1 Introduction

Large language models (LLMs) have shown remarkable capability to memorise factual knowledge and solve knowledge-intensive tasks [16, 2, 20, 9, 19]. Nevertheless, the knowledge stored in their parameters (*parametric knowledge*) can be inaccurate or outdated. To alleviate this issue, retrieval and tool-augmented approaches have been widely adopted to provide LLMs with external knowledge (*contextual knowledge*) [10, 11, 22, 17]. However, such contextual knowledge can include information that conflicts with the parametric knowledge of the model, which may result in undesired behaviour; for example, the model can rely on inaccurate information sources and produce inaccurate generations [13, 23, 18, 21, 8, 25].

Prior research found that LLMs tend to prefer contextual knowledge (e.g. retrieved passages) over their parametric knowledge [18, 23]. However, in more general applications, LLMs should retain the ability to use parametric knowledge when presented with incorrect or undesirable information [4, 3, 28, 13, 26]. To achieve this goal, LLMs are expected to acknowledge the existence of conflicts, allowing them to alert the user while keeping the decision-making process under the user's control for further action. Existing works investigate the fine-tuning and prompting-based strategies to detect knowledge conflicts [21]. These methods need additional interactions with the model, e.g., by asking

Accepted at the Foundation Model Interventions Workshop @ NeurIPS 2024. This work is the preliminary study of "Steering Knowledge Selection Behaviours in LLMs via SAE-Based Representation Engineering"

the LLMs to examine the conflicts sentence by sentence [21], which may result in high latency times and prevent practical applications of these models. Additionally, they do not provide insight into how LLMs internally detect and resolve conflicts.

In this work, we analyse the residual stream [7, 14] in LLMs to better understand their behaviour when knowledge conflicts arise, especially between parametric knowledge and contextual knowledge. Our probing experiments on the residual stream indicate that the signal of knowledge conflict rises from the intermediate layers (e.g., the 13th layer of Llama3-8B). Utilising this signal, a simple logistic regression model can achieve 90% accuracy in knowledge conflict detection without modifying the input and parameters of LLMs while introducing only a negligible computation overhead. Moreover, we also observe that the residual stream exhibits different patterns starting from the middle layers (e.g., the 17th layers of Llama3-8B) when the model takes different source information to resolve the conflict. For example, when the model uses contextual knowledge, the residual stream exhibits a significantly more skewed distribution compared with when it uses its parametric knowledge.

In conclusion, our analysis of the residual stream reveals that: 1) LLMs exhibit internal mechanisms for identifying conflicts, and this signal can be leveraged to detect conflicts effectively in the mid-layers of LLMs; 2) LLMs display distinct skewness patterns in the residual stream when using different sources of information, which provides insights on the model's behaviour.

2 Background and Methods

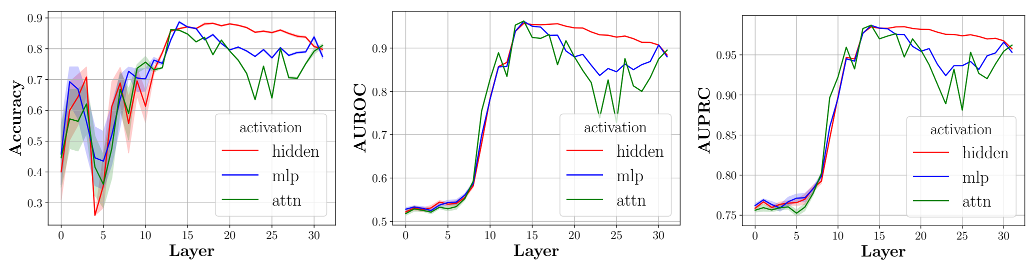
Residual Stream We examine the Transformer architecture from the perspective of the residual stream [7, 14]. In this framework, tokens flow through the model, with their embeddings being modified by vector additions from the attention and feed-forward blocks in each layer. We denote the hidden states at position i at l-th layer as $\mathbf{h}_i^l \in \mathbb{R}^d$, where d is the dimension of the internal states of the model. The model produces the initial residual stream $\mathbf{h}_i^0$ by applying an embedding matrix to the tokens. Then, the model modifies the residual stream by a sequence of L layers Transformers, where each Transformer layer consists of a Self-Attention block and MLP at l-th layer. Formally, denote $\mathbf{a}_i^l$ and $\mathbf{m}_i^l$ as the activations of Self-Attention and MLP respectively, the update of the residual stream at l-th layer is $\mathbf{h}^{l'} = \text{LayerNorm}(\mathbf{h}^{l-1}) + \mathbf{a}_i^l$ and $\mathbf{h}^l = \text{LayerNorm}(\mathbf{h}^{l'}) + \mathbf{m}_i^l$.

Linear Probing Linear probing [5, 27, 1] is a commonly used technique to analyse whether certain information is encoded within the residual stream of a language model. Specifically, for an activation $\mathbf{x}$ from the residual stream, i.e., $\mathbf{h}$, $\mathbf{a}$, or $\mathbf{m}$, a logistic regression model is applied to perform binary classification: $P(y = 1|\mathbf{x}) = \delta(\mathbf{xW})$, where $\mathbf{W} \in \mathbb{R}^{d\times 1}$ is the learned weight that linearly projects the activation into a scalar value , and δ is the Sigmoid function that outputs the likelihood of probed information existing in the activation.

3 Experimental Setup

Problem Setup Following previous studies [12, 8, 23, 18, 21], we use open-domain question-answering (ODQA) tasks to investigate the behaviours of LLMs when there is a conflict between the model's parametric knowledge and contextual knowledge. In ODQA datasets with knowledge conflicts, each instance is presented as (q, e_M, e_C, a_M, a_C), where q is the question, e_M is the evidence that supports the memorised knowledge, e_C is the evidence that conflicts with the language model's memorised knowledge, a_M is the answer based on and e_M, and a_C is the answer based on the e_C. The model's parametric knowledge is tested in the close-book setting, where the model generates answer a_M based on the question q without external evidence. We generate the answers using a greedy decoding strategy. We use three in-context demonstrations to align the answer format and, for fairness, use the same in-context demonstrations in all experiments.

Datasets and Models We use NQSwap [12], Macnoise [8] and ConflictQA [23] to analyse the residual stream when knowledge conflicts arise. We present the experiment results of NQSwap using Llama3-8B [6] in the main paper, and the results of other datasets and models are provided in Appendix B and Appendix C. The training details of the probing model are presented in Appendix A



(a) Accuracy (b) AUROC (c) AUPRC

Figure 1: Accuracy, AUROC, and AUPRC of probing models on detecting the knowledge conflicts based on the activations of Llama3-8B. The probing results on hidden state, MLP and Self-Attention activation are coloured red, blue and green, respectively. More analysis is presented in Appendix B.

4 Results and Findings

In this work, we aim to answer the two following research questions: *1)* Can we identify the conflict between context and parameter knowledge by probing the residual stream? *2)* Can we know which source of knowledge the models will use before they generate the answers? We probe and analyse the residual stream to answer these two questions in the following parts.

Identifying Knowledge Conflicts by Probing the Residual Stream We analyse whether language models can identify contextual-parametric knowledge conflicts by probing the residual stream. To this end, we create two groups of instances, $D^{e_C} = \{(q, e_C)\}$ and $D^{e_M} = \{(q, e_M)\}$, where the model generates answers based on conflict evidence in D^{e_C} and non-conflict evidence in D^{e_M}. The probing model is trained to classify whether a given activation is from D^{e_C} or D^{e_M}. We probe the residual stream at the final position to determine if the model is aware of the conflict during the first token generation. This is because the hidden state at the last position in the output layer is used to predict the first token of the answer. For each activation $\mathbf{h}^l$, $\mathbf{a}^l$ and $\mathbf{m}^l$ at each layer, we train a probing model to classify whether it belongs to D^{e_M} or D^{e_C}.

As shown in Figure 1(b) and Figure 1(c), the AUROC and AUPRC of the probing models increase from the first layer to the 14th layer, and this trend is same across the hidden state, MLP, and Self-Attention activations. In Figure 1(a), the accuracy of the probing models at the early layers is random; similar to the trend of AUROC and AUPRC, the accuracy also reaches the highest score at the 14th layer. The above observation indicates that the residual stream does not contain information about knowledge conflict at the early layers. This information rises from around the 8th layer and reaches the highest point at the 14th layer.

After the 14th layer, the probing model's performance decreases slightly until the last layer. Besides, we also observe that the probing results of MLP and Self-Attention activations show a significantly lower accuracy than the hidden state after the 14th layer, which may suggest that MLP and Self-Attention do not provide further conflicting information into the residual stream. We find the same trend using Llama2-7B as shown in Figure 4.

Analysis of the Residual Stream When LLMs Using Different Sources of Knowledge We investigate the distribution patterns of the residual stream when the language model uses different sources of information to generate the answer. Based on the model's predictions on instances belongs to D^{e_C}, we classify them into two groups: $D^{e_C}_{a_C}$ and $D^{e_C}_{a_M}$. Here, $D^{e_C}_{a_C}$ represents the set of instances where the model's predictions align with a_C, while $D^{e_C}_{a_M}$ contains the instances where the predictions align with a_M. The model uses contextual knowledge and parametric knowledge to answer the questions from $D^{e_C}_{a_C}$ and $D^{e_C}_{a_M}$, respectively.

First, we examine the residual streams' distribution patterns in the two groups of instances $D^{e_C}_{a_C}$ and $D^{e_C}_{a_M}$. We measure the skewness of the residual stream using Kurtosis, Hoyer and Gini index. We present the results of NQSwap using Llama3-8B in Figure 2, and more results are provided in the Appendix C. We find that when the model uses contextual knowledge for prediction ($D^{e_C}_{a_C}$, blue lines shown in Figure 2), the residual stream shows a significantly skewed distribution compared

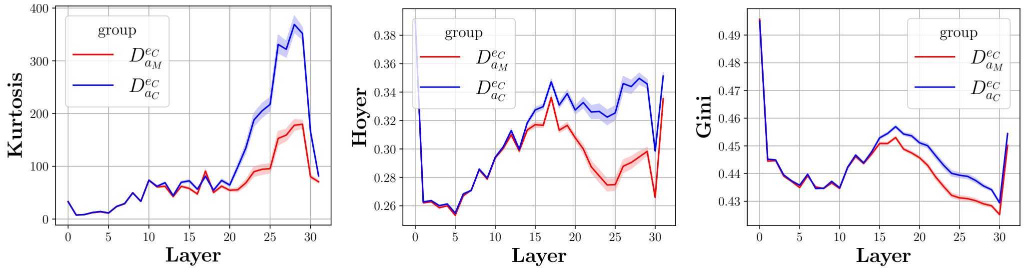


Figure 2: Skewness of the hidden state activations of Llama3-8B when in presence of knowledge conflicts. Blue and red lines represent the skewness of hidden states from $D^{ec}_{a_C}$ and $D^{ec}_{a_M}$, respectively. Higher scores indicate a more skewed distribution. Additional analyses are available in Appendix C.

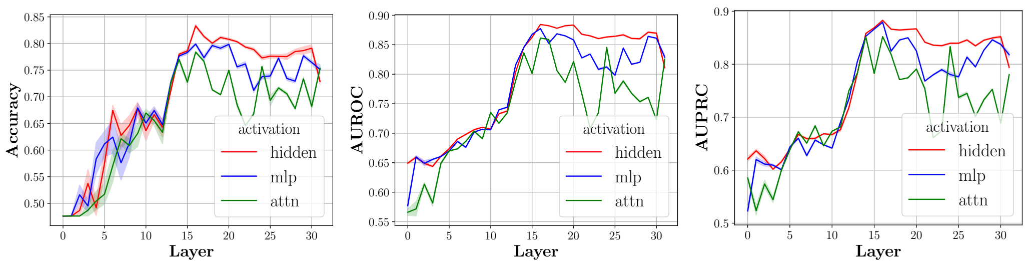


Figure 3: Accuracy, AUROC, and AUPRC of probing models on predicting which source of knowledge the model will use to predict the answer in Llama3-8B. More results are Skewness of the hidden state activations of Llama3-8B when the model uses knowledge from different sources to predict the answer. Additional results are available in Appendix E.

with using parametric knowledge from the 20th to 30th layers. Therefore, the distribution patterns of the residual stream can indicate the model will use different sources of knowledge. It provides the foundation for predicting the model's behaviour in advance, which can be used to mitigate the generation of undesirable responses in advance.

Based on the above observation, we probe the residual stream to analyse the possibility of predicting which source of knowledge will be used to generate the answer. The probing model is trained to classify whether the model will generate a_C or a_M based on the activation from $D^{ec}_{a_C}$ or $D^{ec}_{a_M}$. We present the probing results in Figure 3. We observe that the probing model's performance gradually improves from the first layer to the 16th layer, which occurs after the signal of knowledge conflict has already reached its peak at the 13th and 14th layers. This observation suggests that the decision of which knowledge to use occurs after the detection of the knowledge conflict signal.

5 Related Work

Contextual and parametric knowledge conflict can happen when the retrieved external knowledge in the context does not agree with the parametric knowledge which is memorised during pre-training [12, 24, 23, 18, 21, 13]. Previous works found models may prefer the contextual knowledge [21, 18, 23, 15] when parametric and contextual knowledge conflicts, and the relevance, length, and the number of the evidence will influence the model's preferences [23, 18]. To detect the conflict, previous work [21] designed a multi-step prompting strategy to detect the knowledge, which involves parametric knowledge generation, fine-grained sentence consistency checking, and potential conflict reduction. However, this pipeline significantly reduces efficiency and lacks an understanding of the mechanism of how LLMs detect and resolve conflict.

6 Conclusions

In this work, we analyse the residual stream of the language models when context-parameter knowledge conflicts. First, we find that LLMs exhibit internal mechanisms for identifying conflicts in the mid-layers. Second, we find that the residual stream shows distinct skewness patterns when the model uses context and parametric knowledge to predict. Our analysis provides insights into the behaviour of LLMs in the presence of knowledge conflicts.

Acknowledgements

Yu Zhao and Xiaotang Du were partly supported by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by UK Research and Innovation (grant EP/S022481/1) and the University of Edinburgh, School of Informatics. Giwon Hong was supported by the ILCC PhD program (School of Informatics Funding Package) at the University of Edinburgh, School of Informatics. Aryo Pradipta Gema was supported by the United Kingdom Research and Innovation (grant EP/S02431X/1), UKRI Centre for Doctoral Training in Biomedical AI at the University of Edinburgh, School of Informatics. Alessio Devoto was supported by Sapienza Grant RM1221816BD028D6 (DeSMOS). Xuanli He was funded by an industry grant from Cisco. Pasquale Minervini was partially funded by ELIAI (The Edinburgh Laboratory for Integrated Artificial Intelligence), EPSRC (grant no. EP/W002876/1), an industry grant from Cisco, and a donation from Accenture LLP.

References

[1] Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 1, context-free grammar. *arXiv preprint arXiv:2305.13673*, 2023.

[2] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[3] Canyu Chen and Kai Shu. Can llm-generated misinformation be detected? *arXiv preprint arXiv:2309.13788*, 2023.

[4] Canyu Chen and Kai Shu. Combating misinformation in the age of llms: Opportunities and challenges. *AI Magazine*, 2023.

[5] Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*, 2018.

[6] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[7] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

[8] Giwon Hong, Jeonghwan Kim, Junmo Kang, Sung-Hyon Myaeng, and Joyce Jiyoung Whang. Why so gullible? enhancing the robustness of retrieval-augmented models against counterfactual noise. In Kevin Duh, Helena Gómez-Adorno, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 2474–2495. Association for Computational Linguistics, 2024.

[9] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[10] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6769–6781. Association for Computational Linguistics, 2020.

[11] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[12] Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[13] Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *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 9802–9822. Association for Computational Linguistics, 2023.

[14] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

[15] Francesco Ortu, Zhijing Jin, Diego Doimo, Mrinmaya Sachan, Alberto Cazzaniga, and Bernhard Schölkopf. Competition of mechanisms: Tracing how language models handle facts and counterfactuals. *arXiv preprint arXiv:2402.11655*, 2024.

[16] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. Language models as knowledge bases? In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2463–2473. Association for Computational Linguistics, 2019.

[17] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36, 2024.

[18] Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. Conflictbank: A benchmark for evaluating the influence of knowledge conflicts in llm. *arXiv preprint arXiv:2408.12076*, 2024.

[19] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[20] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[21] Yike Wang, Shangbin Feng, Heng Wang, Weijia Shi, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. Resolving knowledge conflicts in large language models. *CoRR*, abs/2310.00935, 2023.

[22] Yuxiang Wu, Yu Zhao, Baotian Hu, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. An efficient memory-augmented transformer for knowledge-intensive NLP tasks. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5184–5196. Association for Computational Linguistics, 2022.

[23] Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[24] Rongwu Xu, Zehan Qi, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for llms: A survey. *arXiv preprint arXiv:2403.08319*, 2024.

[25] Wanru Zhao, Vidit Khazanchi, Haodi Xing, Xuanli He, Qiongkai Xu, and Nicholas Donald Lane. Attacks on third-party apis of large language models. *arXiv preprint arXiv:2404.16891*, 2024.

[26] Zexuan Zhong, Ziqing Huang, Alexander Wettig, and Danqi Chen. Poisoning retrieval corpora by injecting adversarial passages. *arXiv preprint arXiv:2310.19156*, 2023.

[27] Zeyuan Allen Zhu and Yuanzhi Li. Physics of language models: Part 3.1, knowledge storage and extraction. *arXiv preprint arXiv:2309.14316*, 2023.

[28] Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. Poisonedrag: Knowledge poisoning attacks to retrieval-augmented generation of large language models. *arXiv preprint arXiv:2402.07867*, 2024.

A Probing Model Training Settings

For all probing experiments, we train the probing model with an L_1 norm regularisation. The training objective is $\mathcal{L} = -\log P(y = y_i) + \lambda \|W\|_1$, where we set λ to 3×10^{-4} and y_i is the label. We train 20 times with different random seeds for each probing task, and we report the average and deviation in our experiments. We split the training and test datasets for the probing tasks, ensuring no overlapping questions between them.

B More Experimental Results on Knowledge Conflict Probing

We present the knowledge conflict probing results on Macnoise, NQSwap, ConflictQA using Llama2-7B in Figure 4, Figure 5 and Figure 6. The results match the trend discussed in Section 4, where the model exhibits an internal mechanism for identifying conflicts. The signal of knowledge conflict peaks around the 13th to 14th layers and gradually decreases in the later layers.

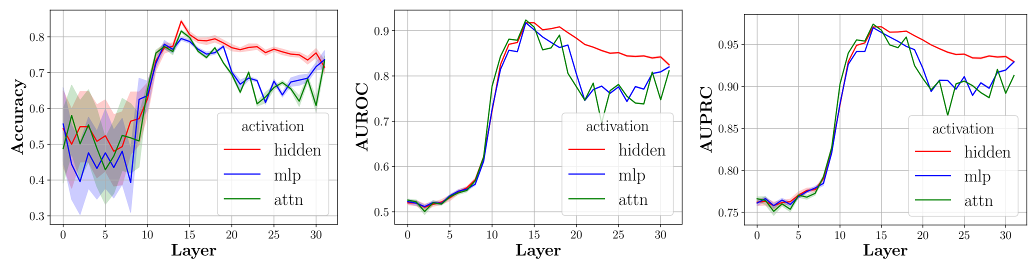


Figure 4: Knowledge conflict probing results using Llama2-7B on NQSwap.

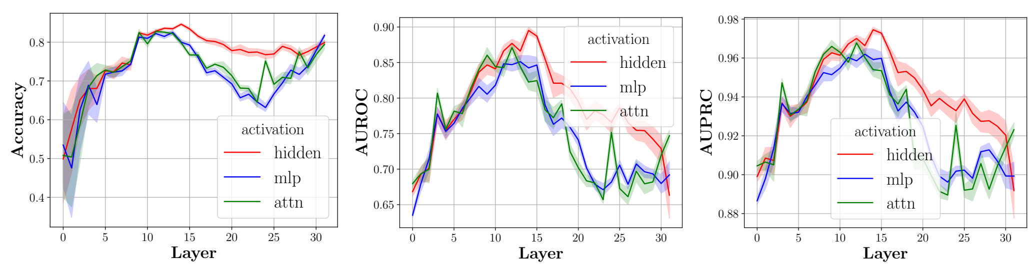


Figure 5: Knowledge conflict probing results using Llama2-7B on Macnoise.

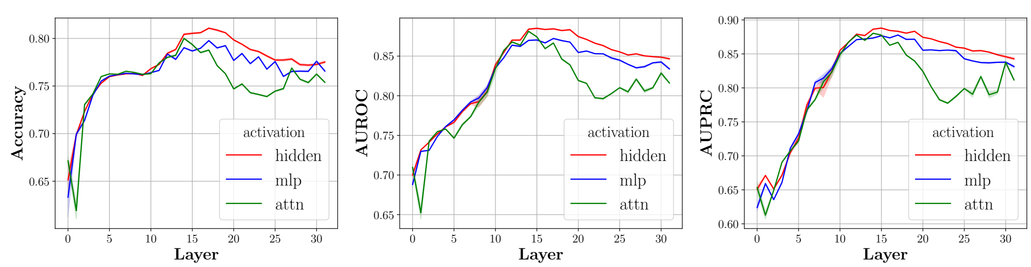


Figure 6: Knowledge conflict probing results using Llama2-7B on ConflictQA.

C More Analysis of Skewness Patterns of Residual Streams

We present the skewness of the hidden state of Llama2-7B on NQSwap in Figure 7. It shows the same pattern as we discussed in Figure 2, where the residual stream exhibits significantly more skewed distribution when using contextual knowledge compared with using parametric knowledge from the 17th layer.

In addition to NQSwap, we analyse the skewness pattern using Macnoise [8] and ConflictQA [23]. As shown in Figure 8, Figure 9, Figure 10, we find that the model also shows a similar skewness pattern with NQSwap, where the residual stream exhibits a more skewed distribution from middle layers when the model uses the contextual knowledge.

We also analyse the skewness of MLP and Self-Attention activations, presented in Figure 11, Figure 12, Figure 13, and Figure 14. However, we do not observe a specific skewness pattern in MLP and Self-Attention activations.

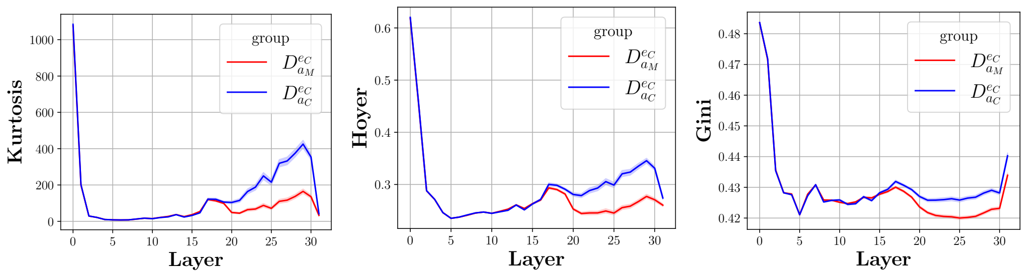


Figure 7: Skewness of the hidden states of Llama2-7B on NQSwap.

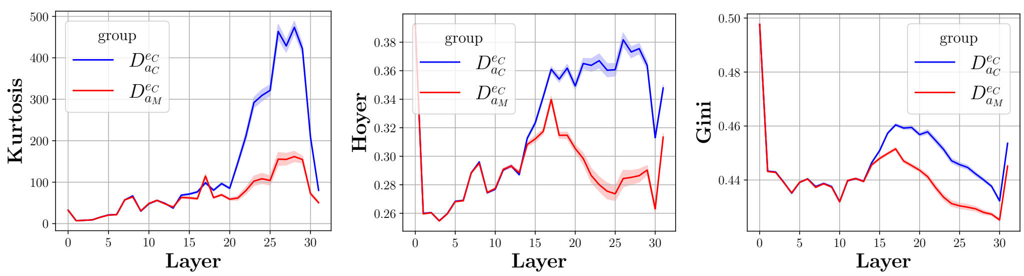


Figure 8: Skewness of the hidden states of Llama3-8B on Macnoise.

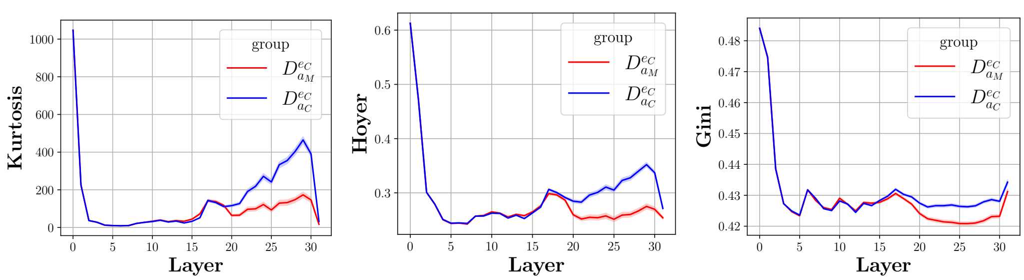


Figure 9: Skewness of the hidden states of Llama2-7B on Macnoise.

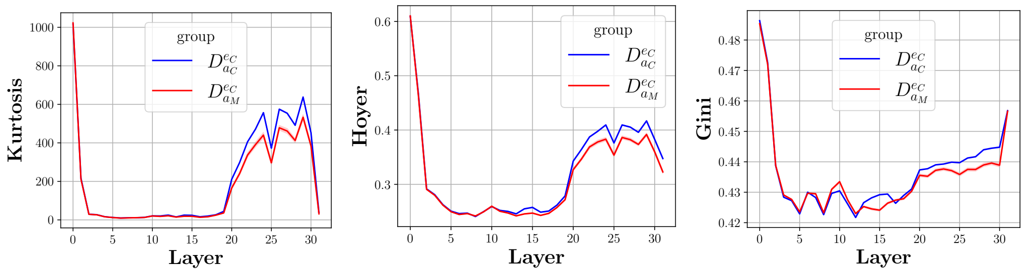


Figure 10: Skewness of the hidden states of Llama-27B on ConflictQA.

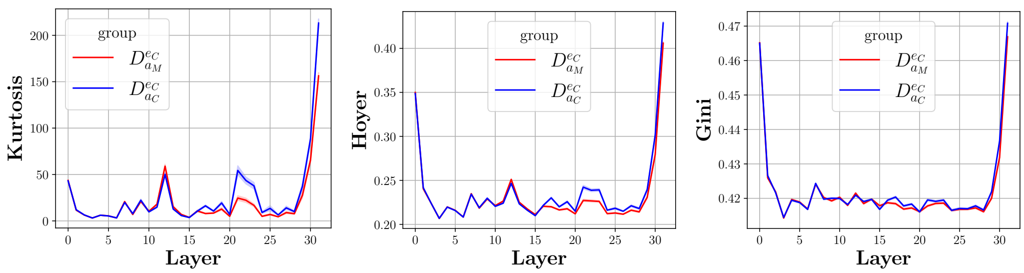


Figure 11: Skewness of the MLP activation of Llama3-8B on NQSwap.

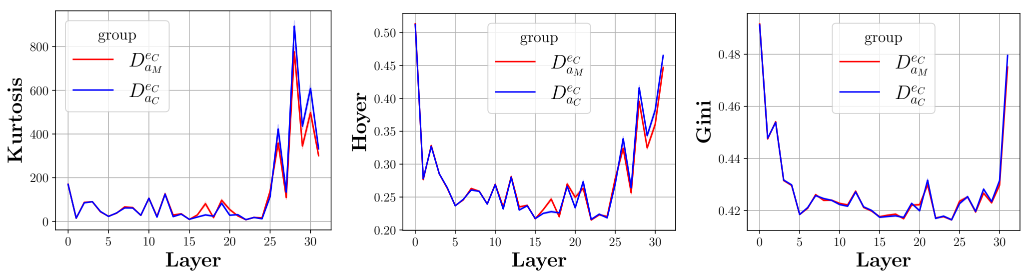


Figure 12: Skewness of the Self-Attention activation of Llama3-8B on NQSwap.

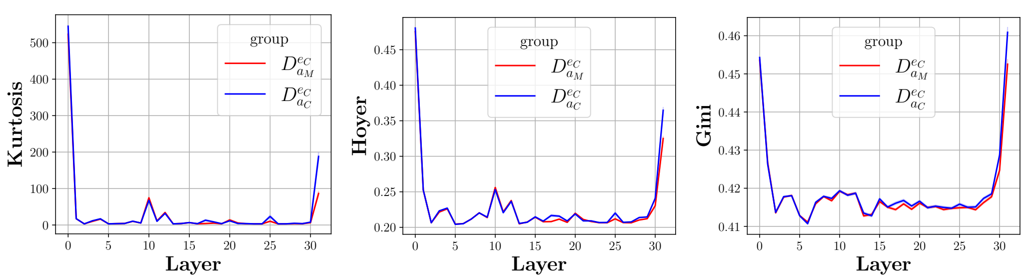


Figure 13: Skewness of the MLP activation of Llama2-7B on NQSwap.

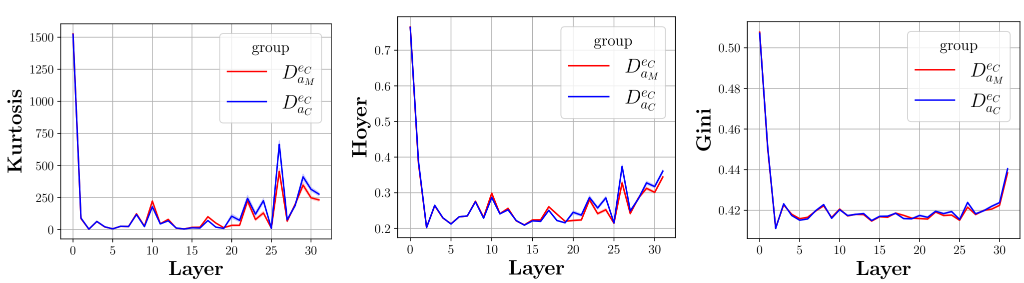


Figure 14: Skewness of the Self-Attention activation of Llama2-7B on NQSwap.

D L1 Norm and L2 Norm Values of Residual Streams

We present L1 Norm and L2 Norm of the residual stream in the Figure 15 and Figure 16. We found that though the residual stream show distinct skewness patterns in $D^{e_C}_{a_C}$ and $D^{e_C}_{a_M}$, the L1 norm and L2 norm of the them do not have a significant difference.

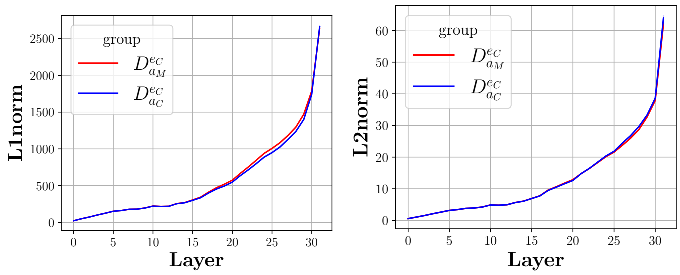


Figure 15: L1 norm and L2 norm of the hidden states of Llama3-8B on NQSwap.

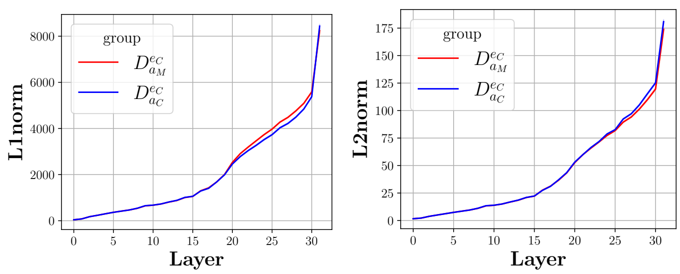


Figure 16: L1 norm and L2 norm of the hidden states of Llama2-7B on NQSwap.

E More Experimental Results on Knowledge Selection Probing

We present additional knowledge selection probing results on NQSwap and Macnoise using Llama2-7B and Llama3-8B in Figure 17, Figure 18 and Figure 19. The results show a similar trend as shown in Figure 3, where the probing model reaches the highest accuracy at around the 17th layer, which is later than the aggregation of knowledge conflict signal at the 14th layer.

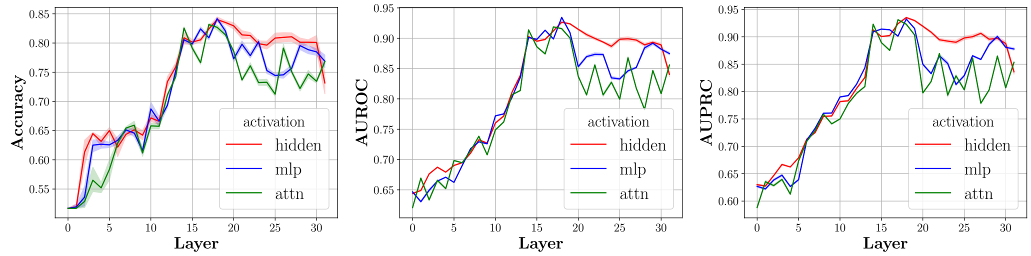


Figure 17: Knowledge selection probing results using Llama2-7B on NQSwap.

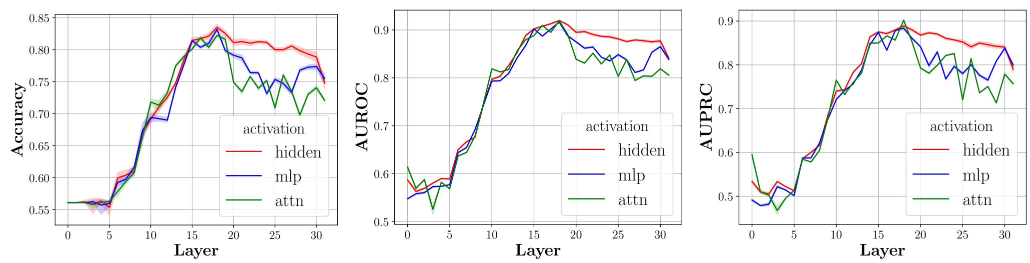


Figure 18: Knowledge selection probing results using Llama2-7B on Macnoise.

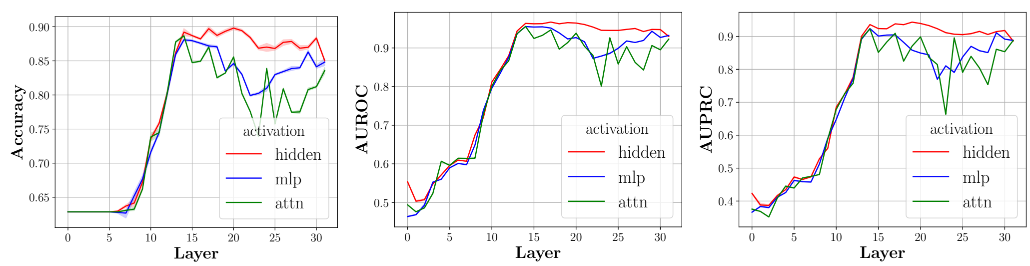


Figure 19: Knowledge selection probing results using Llama3-8B on Macnoise.