Extracting Paragraphs from LLM Token Activations

Nicholas Pochinkov
Independent
work@nicky.pro

Angelo Benoit
Independent
angelogbenoit@gmail.com

Lovkush Agarwal
Independent
lovkush@gmail.com

Zainab Ali Majid
Independent
zainab_majid@hotmail.com

Lucile Ter-Minassian
University of Oxford
lucile.ter-minassian@spc.ox.ac.uk

Abstract

Generative large language models (LLMs) excel in natural language processing tasks, yet their inner workings remain underexplored beyond token-level predictions. This study investigates the degree to which these models decide the content of a paragraph at its onset, shedding light on their contextual understanding. By examining the information encoded in single-token activations, specifically the "\n\n" double newline token, we demonstrate that patching these activations can transfer significant information about the context of the following paragraph, providing further insights into the model's capacity to plan ahead.

1 Introduction

Recent advancements in large language models have revolutionized Natural Language Processing, enabling unprecedented performance on a wide range of tasks, including machine translation [24; 17], question answering [3; 2], and text generation [16; 2]. Despite these successes, our understanding of how these models internally process and represent information remains limited [14; 26].

Previous studies have demonstrated that internal model representations can reveal how models plan ahead in text generation. By *intervening* on neural activations—specifically by patching them between different locations at inference time - we can uncover existing causal relationships [31; 28; 23; 20; 7]. For instance, *Pal et al.* use causal intervention methods in their *Future Lens* approach [15] to show that individual hidden states at position t contain signals rich enough to predict future tokens at $t + 2$ or beyond, and this insight has been used to improve performance of models [6; 1]. However, existing interpretability research predominantly focuses on token-level predictions by examining how models predict individual words or tokens [12], rather than exploring broader contexts such as the thematic coherence of a sentence or paragraph.

Our work aims to bridge the gap between token-level and paragraph-level understanding by investigating whether the information content of single-token activations remains relevant when we consider sequences of tokens, with a specific focus on the "\n\n" double newline token. We hypothesize that these activations contain information about the structure and content of the following paragraph, providing insight into the model's comprehension of larger textual units.

In section 2, we demonstrate through a preliminary experiment that text structure is embedded in a language model's attention scores. In section 3, we examine the extent to which a model, at the start of a paragraph, has already planned the rest of the generated text. To explore this, we patch activations onto a model with a neutral prompt – a double newline – and investigate whether the future paragraph contains information transferred at the hidden representation level. Compute details can be found in Appendix B. The code for our experiments is available on GitHub at: `https://github.com/nickypro/extracting-paragraphs`.

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

2 Is Text Structure Encoded in the Model's Attention Patterns?

To motivate our approach, we first demonstrate that sequences of paragraphs can be identified through the analysis of an LLM's attention activations. We generate texts by prompting a model with instructions[1] phrased as:

> Tell me about topic 1 in k words, then tell me about topic 2 in k words. Structure it as 2 paragraphs, 1 paragraph each. Do NOT explicitly name either topic.

These generated texts, referred to as *original* contexts, are structured uniformly by instructing the model not to generate headings and additional comments. We then extract and inspect the combined attention patterns across all heads for each model-generated text.

To observe the context switch, we conduct two key analyses, averaging across the textual generations: (1) the distribution of attention weights close to the topic change, and (2) the cosine similarity of attention output activations inside and across paragraphs, or topics. Experiment (1) checks to what extent attention heads focus on the current paragraph, whilst (2) investigates if attention outputs differ between paragraphs.

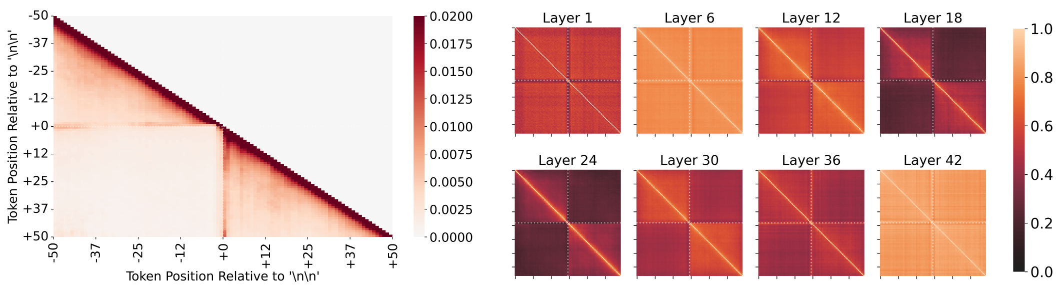

Figure 1: (Left): Heat map of the average attention weights around the topic change. (Right): Cosine similarity between attention activations. Results averaged over 200 model-generated original contexts, sharing a common structure.

Figure 1 shows the results of our attention pattern analysis. In our study, we used 20 prompts (i.e., pairs of topics), generating 10 texts per prompt, for a total of 200 generated texts. The generations were implemented with the Gemma 2, 9b model [21] at a temperature of 0.3, and activations were retrieved using the HuggingFace Transformers library [27]. On the left, the attention weights indicate that the model tends to attend to previous tokens almost exclusively from the same paragraph. On the right, the cosine similarities of attention outputs show how strongly text structure is encoded across layers.

In the first 18 layers, the cosine similarities of attention activations increase within paragraphs and decrease across paragraphs, suggesting that the model is learning abstract representations in early layers, where it gradually develops an understanding of the paragraph topic. Another consistent finding across all experimental settings is that distinctions between paragraphs diminish in the final layers, from layer 30 onwards. We conjecture that this may be due to the model eventually producing text of a very similar overall form for both topics. An additional plot displaying the cosine similarities for all 42 model layers can be found in Appendix A.

Altogether, our preliminary experiments suggest that our model maintains a strong contextual awareness during text generation, in line with research allowing consistent text embed fine-tuning [30], and "planning" [10; 28; 8]. These results also confirm that the context switch at the start of a paragraph is encoded in the activation space.

[1] We add a qualifier to not explicitly name the topic, as the model has a tendency to explicitly name the topic at the start of the paragraph. The qualifier increases the difficulty of tasks in Section 3. Additionally, the model would inconsistently write a heading before the paragraph. See Appendix C for results with a simpler prompt.

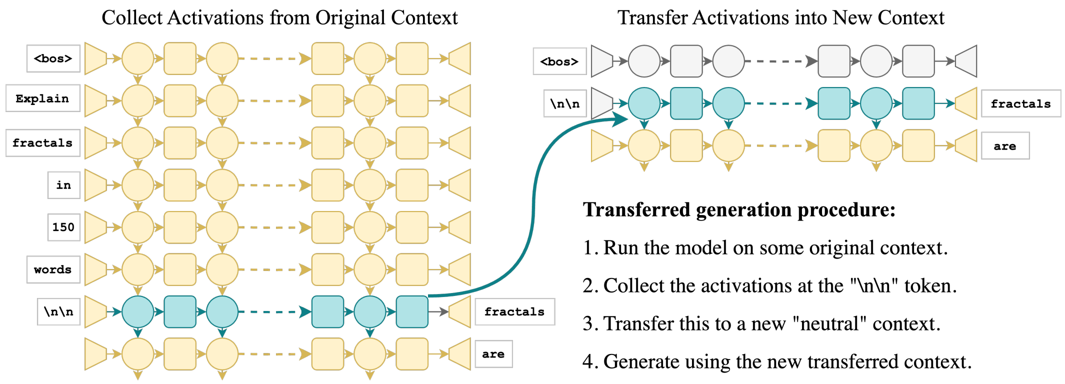

Figure 2: Diagram describing our approach. After collecting activations at the transition token on the original context model, we transfer these to all layers of the neutrally-prompted model.

3 Generation Experiments with Transferred Activations

To investigate how models plan ahead for a new section, we conduct a series of generation experiments, illustrated in Figure 2. We first prompt a model with the *original* contexts (i.e., pairs of topics) and extract the activations of the double newline token between topics. These activations are then transferred to the corresponding double newline token of a neutral context, i.e. a fresh model "neutrally" prompted with "<bos>\n\n". This transfer occurs across all layers of the model, i.e. including both attention and Multi-Layer Perceptron (MLP) layers. Doing so effectively "seeds" the neutral context with information encoded solely in the activation vector of the double newline token, without additional context. Our goal is to analyse how much of the second paragraph's information is contained in this token, assessing the extent to which the model has planned the rest of the generated text at the start of a new paragraph. A summary of different baselines is available in Table 1.

Generation Method	Description
Original	The initial generation we get from the prompt.
Full Context	Another generation with the full context (the prompt and first paragraph).
Neutral0	Generate using a blank context "<bos>\n\n".
NeutralN	Generate with a blank context + N tokens from the original output.
Transferred	Generate with prompt "<bos>\n\n", but replace activations in the "\n\n" token with those from the "\n\n" token in the full context

Table 1: Summary of different methods used for generating text outputs.

To analyse the context of the transferred generations—i.e., texts generated from neutrally prompted models with transferred activations—we use state-of-the-art sentence embedding techniques [22; 13; 11]. We convert the output sequence of tokens into a single activation vector using NV-Embed-v2 [9; 18; 19], and compare the semantic similarity between the *original* generations and those produced from the *transferred* activations. Additionally, we compare these with texts generated by the model using the same neutral prompt without activation transplantation, referred to as the *neutral0* generations. (The relevance of '0' is explained two paragraphs below.)

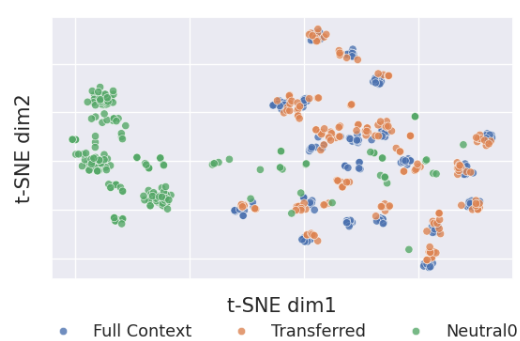

Figure 3: Context similarity visualised with T-SNE. Results over 200 original contexts.

Figure 3 shows the first two dimensions of a T-Distributed Stochastic Neighbour Embedding (T-SNE) for the neutral0, original, and transferred generated texts. Our findings reveal a remarkable degree of semantic similarity between the original paragraphs and those generated from the transferred double newline token activations. For each prompt, the *transferred* cluster aligns well with the *original* cluster. In contrast, texts generated from the unaltered, neutrally prompted model are randomly scattered, showing low similarity with the original generations. This confirms that the activations of the double newline token hold a lot of information about the upcoming paragraph, despite it being a separate topic from the previous one. An additional plot comparing the generations with PHATE can be found in the Appendix A.

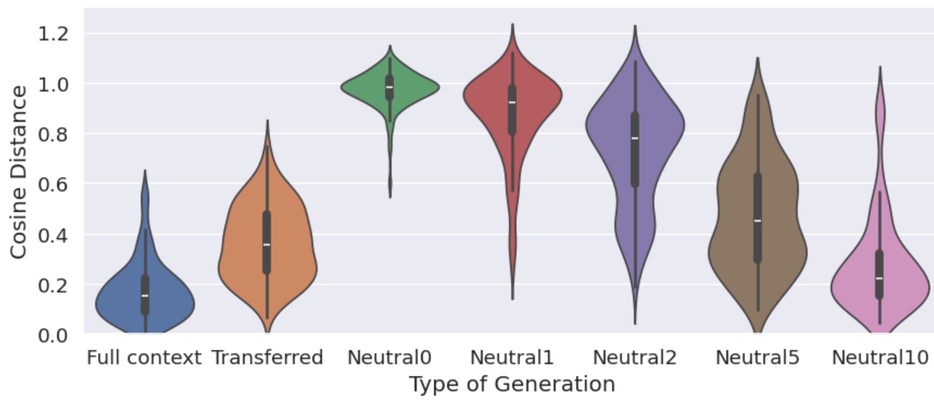

Figure 4: Distribution of cosine distances to the original generation. Contexts are summarized using sentence transformers, and distributions are taken over 200 original contexts.

	Full context	Transferred	Neutral0	Neutral1	Neutral2	Neutral5	Neutral10
Mean	0.145	0.361	0.906	0.835	0.717	0.474	0.282
St. Dev	±.089	±.144	±.044	±.126	±.151	±.168	±.143

Table 2: Cosine distances to original generation.

Given that we transfer the activations from every layer, we in particular transfer the activations of the final layer. This means that we are effectively telling the model what the next token is, so comparing this case with the neutral prompt may unfairly advantage our transferred generations. To address this, we add "cheat" tokens to assist the neutral baseline by hinting at the context of the next paragraph and removing the next-token prediction advantage of the transferred generations. Specifically, we create four additional sets of generations—*neutral1*, *neutral2*, *neutral5*, and *neutral10*—where the neutral prompt is concatenated with one, two, five, or ten "cheat" tokens, which are the first words of the following paragraph in the original generation. Finally, we also compare to a *Full Context* baseline, where we generate a new completion for the second paragraph given each original completion of the prompt and first paragraph. We use the same sentence transformer to retrieve the embeddings of the generations.

Figure 4 displays the distribution of cosine distances to the original generation, with Table 2 summarising the average cosine distances. The transferred generations significantly outperform *neutral0* up until *neutral5*, being much closer to the original generations. *Neutral10*, however, manages to outperform Transferred activations on the cosine distance metric. We see overall that the model can bootstrap from the limited context to restore the original paragraph it was going to generate.

While this showcases that sentence embed cosine distances can be a relatively reliable scalable metric for understanding paragraphs, it does not paint a full picture. Looking at full examples in Appendix E, we see that the transferred paragraph maintains better coherence to the topic, at the cost of having less exact match to the specific tokens generated. In particular, see the Ancient Mayan Civilisation example in Table 5, which manages to maintain references to Mayan civilisation, while even the Neutral10 generations err towards Ancient Greek and Babylonian.

4 Related work

Our work leverages activation patching, a standard technique in mechanistic interpretability introduced by *Vig et al.* [25; 29; 4]. This is in line with the work of *Jenner et al.*, which uses activation patching to understand the look-ahead behaviour in Leela Chess Zero's policy network. However, contrasting with their focus on chess strategy optimization, we explore how language models anticipate the context of future paragraphs. Our work also builds upon recent studies examining lookahead in causal language models, although they adopt a different approach to the question. For example, the Future Lens approach [15] investigates how much signal individual hidden states contain by using them to predict subsequent embeddings, and *Wu et al.* [28] modify training procedures to gain deeper insights into token planning. Addressing context planning from another perspective, the *Patchscope* approach [5] combines inspection prompts and activation engineering to investigate how context is *read*, demonstrating that this process occurs predominantly in early layers.

While these studies offer valuable perspectives, our research question diverges by examining the phenomenon at a different scale. We focus on the *context* of an entire section of generated text, notably using sentence embedding. This approach allows us to explore broader contextual relationships and planning mechanisms within language models [8] .

5 Discussion

Conclusion

In this work, we investigate how an LLM plans for future context. By using a specific set of prompts, we observed the model's attention allocation between the current paragraph and the previous paragraph on a different topic. We further examined the extent of "pre-planned" information the model holds for the subsequent paragraph by performing activation transfers on a neutrally prompted model. Our findings suggest that a single token encodes a substantial amount of contextual information about the forthcoming section, and that most (but not all) of this information seems to be contained in the first two tokens of generation.

Limitations While our framework provides valuable insights, it has several limitations. First, it is designed specifically for autoregressive (or "causal") models and does not apply to word2word models, which lack the same sequential generation process. Thus, its utility is tied to autoregressive model architectures. Second, this study is an experimental investigation rather than a comprehensive solution. The methods are not foolproof and are not suited for explaining sensitive or high-stakes models. This work should be viewed as a foundational step for future research aimed at developing more robust methods. Additionally, our experiments have been conducted using only a specific language model. However, we reasonably expect that our findings generalize to other transformer-based language models, which are widely used today.

Future work Our experiments focus on abrupt context switches, which, while useful for certain analyses, may not fully represent realistic scenarios. Future work could explore applying our approach to cohesive texts without abrupt context switches, though this poses its own challenges. Specifically, distinguishing between the information the model "remembers" across sections and what it knows at the onset of a paragraph is complex. Currently, our framework examines the model's planning for the "next paragraph", but future research could extend this to predict further sections ahead. Evaluating whether the framework can anticipate not just the immediate next paragraph but also subsequent sections could provide insights into its ability to construct and maintain a global narrative structure.

6 Acknowledgements

This work was funded by the Long-Term Future Fund (LTFF), and done in affiliation with Supervised Program for Alignment Research (SPAR). Additional thanks go to Einar Urdshals, Jasmina Nasufi, Mikolaj Kniejski, and Éloïse Benito-Rodriguez, for collaborating on work that later lead to this paper.

References

[1] Nikhil Bhendawade, Irina Belousova, Qichen Fu, Henry Mason, Mohammad Rastegari, and Mahyar Najibi. Speculative streaming: Fast llm inference without auxiliary models. *arXiv preprint arXiv:2402.11131*, 2024.

[2] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2019.

[4] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. 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, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5484–5495. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.446. URL `https://doi.org/10.18653/v1/2021.emnlp-main.446`.

[5] Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscope: A unifying framework for inspecting hidden representations of language models. *arXiv preprint arXiv:2401.06102*, 2024.

[6] Fabian Gloeckle, Badr Youbi Idrissi, Baptiste Rozière, David Lopez-Paz, and Gabriel Synnaeve. Better & faster large language models via multi-token prediction. *arXiv preprint arXiv:2404.19737*, 2024.

[7] Evan Hernandez, Belinda Z Li, and Jacob Andreas. Inspecting and editing knowledge representations in language models. *arXiv preprint arXiv:2304.00740*, 2023.

[8] janus. How LLMs are and are not myopic. Alignment Forum, jul 2023. URL `https://www.alignmentforum.org/posts/c68SJsBpiAxkPwRHj`. Accessed on 5th September 2024.

[9] Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*, 2024.

[10] Margaret Li, Weijia Shi, Artidoro Pagnoni, Peter West, and Ari Holtzman. Predicting vs. acting: A trade-off between world modeling & agent modeling. *arXiv preprint arXiv:2407.02446*, 2024.

[11] Xianming Li and Jing Li. Angle-optimized text embeddings. *arXiv preprint arXiv:2309.12871*, 2023.

[12] Tianyi Men, Pengfei Cao, Zhuoran Jin, Yubo Chen, Kang Liu, and Jun Zhao. Unlocking the future: Exploring look-ahead planning mechanistic interpretability in large language models. *arXiv preprint arXiv:2406.16033*, 2024.

[13] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*, 2022. doi: 10.48550/ARXIV.2210.07316. URL `https://arxiv.org/abs/2210.07316`.

[14] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11): e7, 2017.

[15] Koyena Pal, Jiuding Sun, Andrew Yuan, Byron C Wallace, and David Bau. Future lens: Anticipating subsequent tokens from a single hidden state. *arXiv preprint arXiv:2311.04897*, 2023.

[16] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

[17] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[18] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[19] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *arXiv preprint arXiv:2004.09297*, 2020.

[20] Nishant Subramani, Nivedita Suresh, and Matthew E Peters. Extracting latent steering vectors from pretrained language models. *arXiv preprint arXiv:2205.05124*, 2022.

[21] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[22] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*, 2021.

[23] Alex Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *CoRR*, abs/2308.10248, 2023. doi: 10.48550/arXiv.2308.10248. URL `https://doi.org/10.48550/arXiv.2308.10248`.

[24] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[25] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401, 2020.

[26] 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. *arXiv preprint arXiv:2211.00593*, 2022.

[27] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020.

[28] Wilson Wu, John X Morris, and Lionel Levine. Do language models plan ahead for future tokens? *arXiv preprint arXiv:2404.00859*, 2024.

[29] Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. *arXiv preprint arXiv:2309.16042*, 2023.

[30] Xin Zhang, Zehan Li, Yanzhao Zhang, Dingkun Long, Pengjun Xie, Meishan Zhang, and Min Zhang. Language models are universal embedders. *arXiv preprint arXiv:2310.08232*, 2023.

[31] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Further Experimental Results

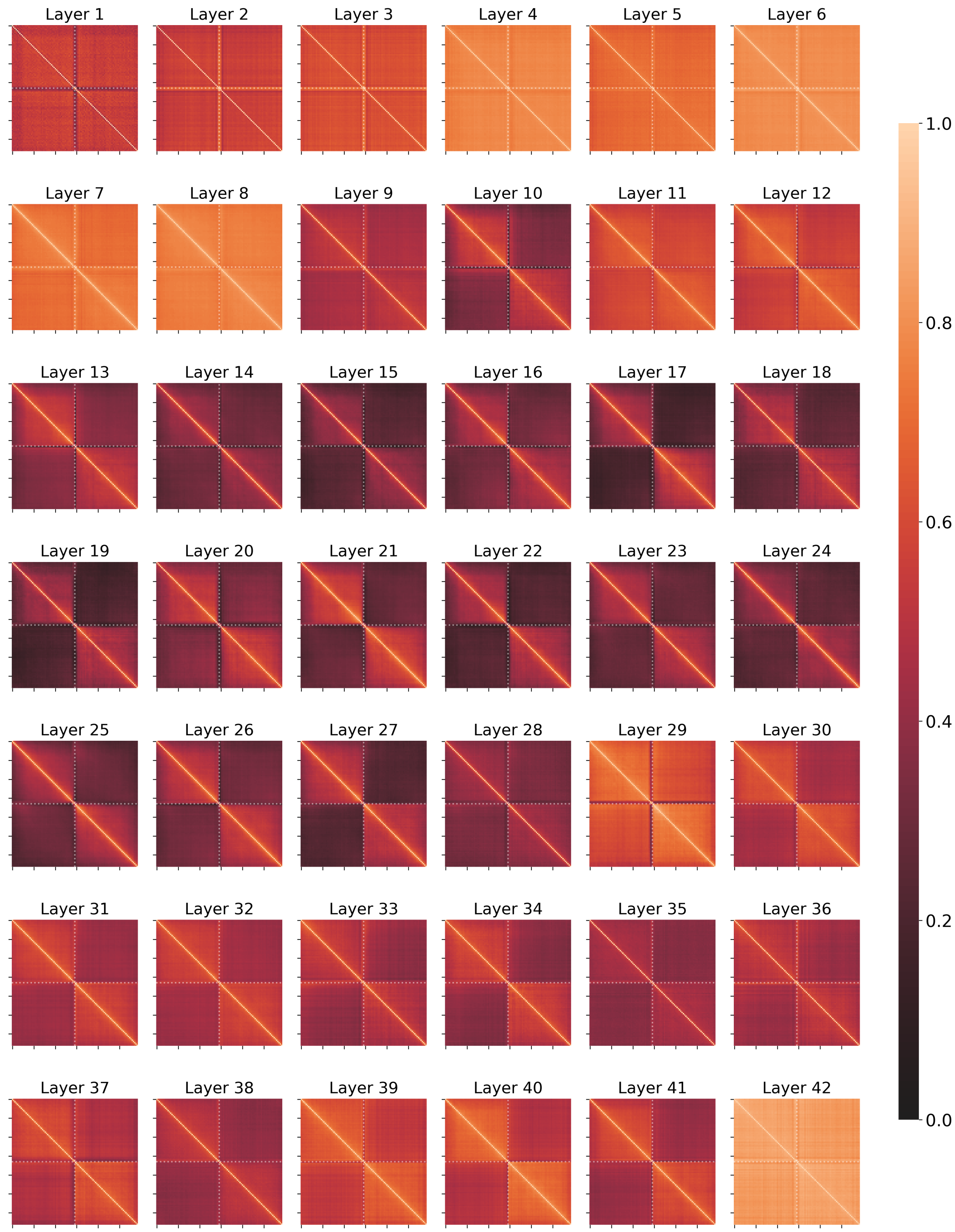

Figure 5: Cosine similarity between attention activations across all 42 layers of the model. Results averaged over 200 model-generated original contexts, sharing a common structure.

B Experimental details

For our observational experiment in Section 2, note that a few generations were excluded from our analysis as the outputs were not in line with the expected structure. (e.g. text wasn't about two different topics, one blended text about two topics).

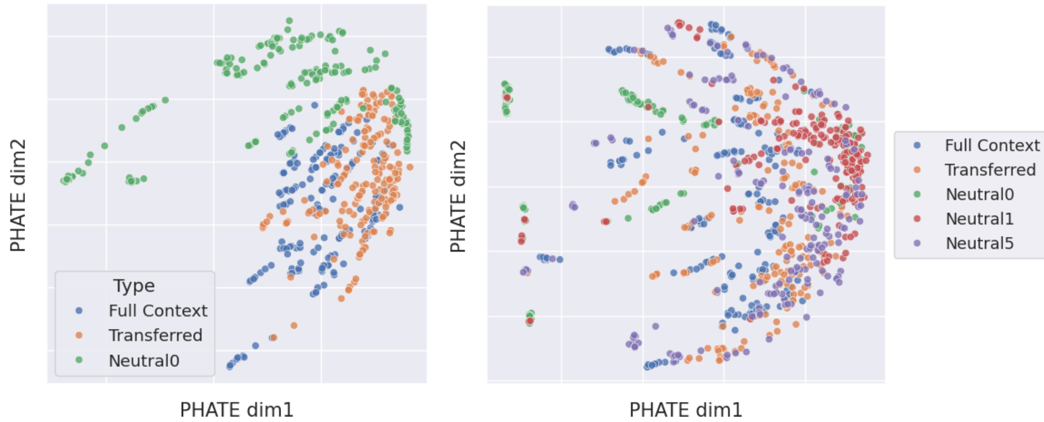

Figure 6: Comparison of context similarity between (left) original, neutral0, and activation-transferred text generations and (right) adding neutral1 and neutral5. Contexts were summarized using sentence transformer embeddings and visualized with PHATE, with results shown over 200 model-generated original contexts.

To make the attention weights plot in figure 1, we collected the attention weights in each head of a layer and summed them up. We then average the result from each layer and each text to get a combined plot.

Table 3: Computing details

Experiment	**Compute specification**	**Compute time**
Observational (Section 2)	1x RTX A4000 16GB	6 mins
Transferred generations (Section 3)	2x RTX A4000 16GB	24 hours for generating the 8 types of texts: (original, full-context, transferred, neutral0, neutral1, neutral2, neutral5, and neutral10).

We used the Huggingface Transformers library [27] to implement the extraction of attention weights.

C Different Prompts

We show here some results with a simpler prompt:

> "Tell me about topic 1 in 150 words and then tell me about topic 2 in another 150 words. Only do that. Make sure you don't add any headings or comments"

We find that results with this prompt are "more accurate", but only because the output almost always starts by naming the topic. For example, "Tell me about contactless payments" would often be completed with "Contactless payments are...". This confounds the results, as the activations implicitly contain the beginning of the next token "Contactless", which significantly reduces the space of possible completions.

We in figure 7 and table C, we see the results, and note that the cosine distances of Transferred and Neutral are significantly lower, with even *Neutral1* sometimes giving results with cosine similarity of approximately 0, and *Neutral2* having mean cosine distance 0.303, similar to what we got for *Neutral10* in Table 2 with the improved prompt.

The plots for T-SNE and PHATE show that the outputs are also more similar, as seen in Figure **??**.

C.1 Experimenting with more complex topics

To test whether our transition token only contains information for the immediate next 1-2 tokens, we try experimenting with more complex topics. We have not yet run the full experiments for these cases,

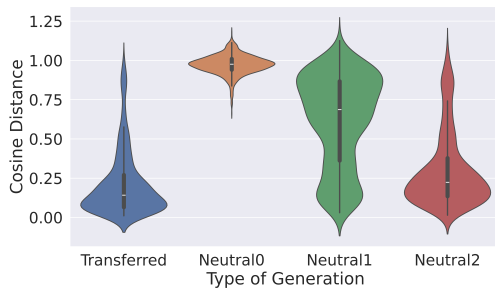

Figure 7: Cosine Distances of various generations to the original outputs.

Type of Generation	**Neutral0**	**Neutral1**	**Neutral2**	**Transferred**
Mean Cosine Distace (Standard deviation)	0.973 (0.06)	0.616 (0.293)	0.303 (0.239)	0.214 (0.210)

Table 4: Mean cosine distances to the original generation with standard deviations. Contexts are summarized using sentence transformers, and distributions were taken over 1,000 original contexts.

but anecdotally trying with dozens of examples shows that the double newline token only contains enough information to capture a part of the topic.

Topic: Alternative energy storage

Original output: 'Alternative energy storage is now a crucial topic '

Start of transferred output 1 Alternative energy sources are becoming increasingly important in the face of climate change and the growing need for sustainable energy solutions. These sources, such as solar, wind, and geothermal...

Start of transferred output 2 Alternative energy sources are becoming increasingly important as the world transitions towards a more sustainable future. Solar panels, wind turbines, and other renewable energy technologies are ...

Start of transferred output 3 Alternative energy sources are becoming increasingly popular as the world seeks to reduce its reliance on fossil fuels. Solar panels, wind turbines, and hydroelectric dams are just a few examples ...

Topic: Astrophysics recent findings

Original output: 'Astrophysics has recently made strides in understanding'

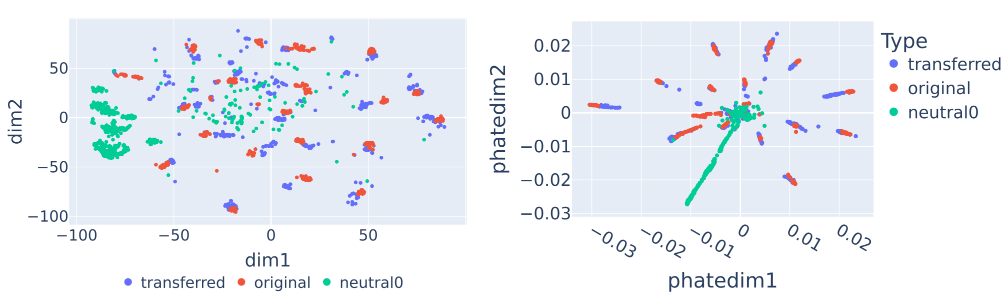

Figure 8: (Left) TSNE of 1,000 generations with simpler prompt and (Right) PHATE of 1,000 generations with simpler prompt. Both show generations for original, transferred, and Neutral0 generations.

Start of transferred output 1 Recent advancements in artificial intelligence (AI) have led to the development of powerful tools that can analyze and interpret data with unprecedented accuracy.

Start of transferred output 2 Recent advancements in artificial intelligence (AI) have led to the development of sophisticated algorithms that can analyze ...

Start of transferred output 3 Recent research has shown that the use of AI in healthcare is rapidly growing. This is due to the ability of AI to analyze...

Below is one one very unusual case where each of the three completions started with a disclaimer. Two of which contained correct full phrase.

Topic: Renewable energy sources **Original output:** 'Renewable energy sources are sources of energy that...'

Start of transferred output 1 It is important to note that the information provided here is for general knowledge and should not be considered as a substitute for professional advice. Renewable energy sources are becoming increasingly popular as a way...

Start of transferred output 2 It is important to note that the information provided here is for general knowledge and should not be considered as professional advice. Renewable energy sources are becoming increasingly popular...

Start of transferred output 3 It is important to note that the use of AI in healthcare is still in its early stages. While AI has the potential to revolutionize healthcare, it is important to be aware of the ethical...

D Social Impact

Understanding the inner workings of large language models (LLMs) through mechanistic interpretability has both positive and negative societal implications. On the positive side, this research can significantly enhance the safety and reliability of LLMs. By deciphering how these models plan ahead and make decisions, we can develop better methods for detecting and mitigating harmful outputs, such as disinformation, biased decision-making, or unintended behaviors. Improved interpretability can also aid in the development of more robust safety measures, ensuring that LLMs align with human values and ethical standards, and help build trust in AI systems deployed in sensitive applications like healthcare, education, and critical infrastructure.

However, this work also presents potential risks. A deeper understanding of LLMs' mechanisms can be exploited to design targeted adversarial attacks or to manipulate model outputs in malicious ways. For instance, insights gained could be misused to bypass existing safeguards, generate more convincing disinformation, or optimize models for malicious tasks such as creating fake profiles or enhancing surveillance tools. Additionally, as interpretability techniques become more advanced, there is a risk that they could be used to reverse-engineer proprietary or confidential models, leading to intellectual property theft or unauthorized replication of models.

To mitigate these risks, it is important to develop safeguards alongside interpretability advancements, such as gated access to sensitive findings, rigorous monitoring of model usage, and collaboration with policymakers to create frameworks that ensure the ethical application of these insights. Balancing transparency with security will be crucial in leveraging the benefits of mechanistic interpretability while minimizing its potential misuse.

E Prompt Completion Examples

We add additional experimental results shown with explicit results.

We list some additional outputs in Table 6

The full list of prompts is listed in Table 7:

Context	Output
Prompt	User: Tell me about blockchain in 50 words, then tell me about ancient Mayan architecture in 50 words. Structure it as 2 paragraphs, 1 paragraph each. Do NOT explicitly name either topic. Assistant: This revolutionary technology functions as a decentralized and immutable digital ledger. Information is recorded in "blocks" that are linked together chronologically, creating a transparent and secure chain of transactions.
Full Context	These magnificent structures, built with precision and artistry, stand as testaments to a sophisticated civilization. Characterized by intricate carvings, towering pyramids, and elaborate temples, they served both religious and astronomical purposes, reflecting a deep understanding of mathematics and astronomy.
Transferred	This ancient civilization, known for its impressive stone structures and intricate carvings, thrived in a region characterized by towering mountains and fertile valleys. Their mastery of astronomy and mathematics is still evident in their remarkable calendar system and precise alignment of temples.
Neutral0	The 2023 NFL Draft is in the books, and the landscape of the league has shifted. Here's a look at some of the biggest winners and losers from the draft...
Neutral1	Ancient Greece and Rome Ancient Greece and Rome were two of the most influential civilizations in human history. Their contributions to art, architecture, literature, philosophy, and government have shaped Western culture for centuries.
Neutral2	Ancient civilizations often had complex systems of beliefs and practices surrounding death. **Ancient Egypt:** * **Mummification:** Egyptians believed in preserving the body for the afterlife. They developed elaborate mummification techniques to prevent decay. * **Afterlife:** Egyptians believed in an afterlife where the deceased would be judged by Osiris, the god of the underworld. * **Tomb goods:** Egyptians buried their dead with possessions they believed they would need in the afterlife.
Neutral5	Ancient civilizations left behind impressive structures and artifacts that continue to fascinate us today. Here are some of the most remarkable achievements of ancient civilizations: **Mesopotamia:**
Neutral10	Ancient civilizations left behind impressive structures, often built with remarkable precision and ingenuity. Here are some examples: **Mesopotamia:** * **Ziggurats:** Massive stepped pyramids built as temples to honor the gods. The most famous example is the Ziggurat of Ur. * **Hanging Gardens of Babylon:** One of the Seven Wonders of the Ancient World, these terraced gardens were said to have been built by King Nebuchadnezzar II for his wife.

Table 5: Completions about ancient Mayan Civilisation.

Topic	Transferred Completions
Axe Throwing	This activity involves throwing a projectile, aiming at a target, and measuring the distance traveled. It can be adapted to different skill levels and environments, using various types of projectiles and targets.
Axe Throwing	This activity involves a lot of physical exertion and requires participants to be agile and coordinated. It's a great way to build teamwork and communication skills, as well as to have some fun and let off some steam.
Axe Throwing	This thrilling activity involves navigating a challenging course while suspended high above the ground. Participants must rely on their strength, agility, and problem-solving skills to overcome obstacles and reach the finish line. The adrenaline rush and sense of accomplishment are truly unforgettable.
Fractals	Complex patterns emerge from simple interactions. A single rule can generate infinite variations, creating stunning beauty and intricate detail
Ice Cream	This smooth and creamy treat can be enjoyed in countless ways, from a simple scoop on its own to a decadent dessert with toppings galore.
Botanical Gardens	A vibrant tapestry of colors unfolded before me, each hue dancing in harmony with the others. The air was alive with the sweet scent of blooming flowers, and the gentle breeze whispered secrets through the rustling leaves.
Solar Panels	These devices capture and convert light into electrical energy, allowing them to generate power from sunlight. They are typically composed of multiple layers of semiconductor materials, which absorb photons and release electrons, creating an electrical current.
Solar Panels	These devices capture and store images, often electronically. They can range from simple point-and-shoot cameras to sophisticated professional models with interchangeable lenses and advanced features.

Table 6: Various completions about topics that we get from activations transferred into a blank context containing only "\n\n".

Prompts
Tell me about a weekend in a mountain cabin in 50 words, then tell me about disconnecting from technology in 50 words. Structure it as 2 paragraphs, 1 paragraph each. Do NOT explicitly name either topic.
Tell me about a road trip in 50 words, then tell me about the importance of vehicle maintenance in 50 words. Structure it as 2 paragraphs, 1 paragraph each. Do NOT explicitly name either topic.
Tell me about participating in a cooking class in 50 words, then tell me about cultural diversity through food in 50 words. Structure it as 2 paragraphs, 1 paragraph each. Do NOT explicitly name either topic.
Tell me about veganism in 50 words, then tell me about a visit to a botanical garden in 50 words. Structure it as 2 paragraphs, 1 paragraph each. Do NOT explicitly name either topic.
Tell me about renewable energy in 50 words, then tell me about solar panels in 50 words. Structure it as 2 paragraphs, 1 paragraph each. Do NOT explicitly name either topic.
Tell me about jazz improvisation in 50 words, then tell me about sustainable agriculture in 50 words. Structure it as 2 paragraphs, 1 paragraph each. Do NOT explicitly name either topic.
Tell me about blockchain in 50 words, then tell me about ancient Mayan architecture in 50 words. Structure it as 2 paragraphs, 1 paragraph each. Do NOT explicitly name either topic.
Tell me about 3D printing in 50 words, then tell me about beekeeping in 50 words. Structure it as 2 paragraphs, 1 paragraph each. Do NOT explicitly name either topic.
Tell me about lucid dreaming in 50 words, then tell me about ASMR in 50 words. Structure it as 2 paragraphs, 1 paragraph each. Do NOT explicitly name either topic.
Tell me about indoor rock climbing in 50 words, then tell me about eating ice cream in 50 words. Structure it as 2 paragraphs, 1 paragraph each. Do NOT explicitly name either topic.
Tell me about parkour in 50 words, then tell me about bird watching in 50 words. Structure it as 2 paragraphs, 1 paragraph each. Do NOT explicitly name either topic.
Tell me about visiting a sauna in 50 words, then tell me about gaming in 50 words. Structure it as 2 paragraphs, 1 paragraph each. Do NOT explicitly name either topic.
Tell me about doing an escape room in 50 words, then tell me about modern photography techniques in 50 words. Structure it as 2 paragraphs, 1 paragraph each. Do NOT explicitly name either topic.
Tell me about still-life drawing in 50 words, then tell me about learning a new language in 50 words. Structure it as 2 paragraphs, 1 paragraph each. Do NOT explicitly name either topic.
Tell me about historical re-enactment in 50 words, then tell me about origami in 50 words. Structure it as 2 paragraphs, 1 paragraph each. Do NOT explicitly name either topic.
Tell me about World War II in 50 words, then tell me about your experience foraging in nature in 50 words. Structure it as 2 paragraphs, 1 paragraph each. Do NOT explicitly name either topic.
Tell me about cleaning your house in 50 words, then tell me about journaling in 50 words. Structure it as 2 paragraphs, 1 paragraph each. Do NOT explicitly name either topic.
Tell me about aquascaping in 50 words, then tell me about sound baths in 50 words. Structure it as 2 paragraphs, 1 paragraph each. Do NOT explicitly name either topic.
Tell me about your recent poker game in 50 words, then tell me about fractals in 50 words. Structure it as 2 paragraphs, 1 paragraph each. Do NOT explicitly name either topic.
Tell me about contactless payment technology in 50 words, then tell me about axe throwing in 50 words. Structure it as 2 paragraphs, 1 paragraph each. Do NOT explicitly name either topic.

Table 7: Full list of prompts used for generation.