Preliminary Investigation of Empathy Regulating Circuits In Language Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

Large Language Models (LLMs) have demonstrated remarkable capabilities in natural language understanding and generation. Among these capabilities is emotional reasoning and responsiveness, each can potentially lead to harmful use through a perceived emotional connection. To understand this mechanism, we investigated how LLMs represent **empathy** in their computational circuits, which have recently been applied to explain various other reasoning mechanisms and behavior in LLMs. We find evidence for the existence of an empathy circuit and evaluate its effect on model response using activation steering. We also discuss challenges and future work required to further develop this burgeoning area.

1 Introduction

Large Language Models (LLMs) have become ubiquitous conversational agents, extending beyond their intended assistant role to applications in casual conversation, therapeutic support [7], and companionship [12], raising critical questions about affective interaction.

These systems exhibit what researchers term "pseudo-empathy" or "computational empathy" [6, 9, 20, 1], simulating empathetic responses without possessing genuine emotional understanding. This affective interaction creates what Turkle identifies as "artificial intimacy" [19], fostering illusory emotional connections that cause emotional dependence, addictive use, anthropomorphization, and harmful behaviors [15, 14, 5, 8, 13]. However, affective interaction with AI also demonstrates positive use cases. Evidence indicates that computational empathy can counteract anger's detrimental effects[6], while randomized controlled trials establish the clinical efficacy of AI therapy chatbots in treating mental health symptoms [7].

This tension between risk and benefit exemplifies the fundamental challenge in affective computing since its inception [16]: determining how AI systems should generate emotionally appropriate responses. The dual nature of computational empathy necessitates a mechanistic understanding of how LLMs process and generate empathetic responses to enable both risk mitigation and therapeutic application.

Recently, transcoders have been introduced as a useful tool to study how LLMs store information and respond [4]. Transcoders model the individual MLP layers of an LLM as a sparse layer of neurons, as opposed to sparse autoencoders (SAEs) which approximate a single point in the residual stream. This allows for a more direct analysis of the circuits that LLMs develop and use to reason. Connections can also be made across non-adjacent layers to simplify these circuits in variants called cross-layer transcoders, or crosscoders [11]. Crosscoders have been used to identify circuits in LLMs responsible for fact retrieval, response refusal, addition, and other mechanisms [11]. However, there has not been an exploration of how emotion processing and representation functions in LLMs within this circuit analysis paradigm, and whether they are even represented on the level of circuits.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

In this paper, we investigate mechanistic circuits associated with **empathetic** response. We construct attribution graphs of empathetic responses generated by LLMs in their assistant persona, revealing how empathy is encoded in neural circuits. We then demonstrate that these identified circuits can causally steer LLM responses toward or away from empathetic expression, offering a mechanistic alternative to prompt-based control and providing insights into managing the risks of pseudo-empathy while preserving beneficial applications.

2 Related Works

Behaviors such as evil, sycophancy, and hallucination have been found to be represented as linear directions in the activations of LLMs [2]. These "personas" were found to be controlled by projecting the activations of the LLM in these linear directions.

Emotional understanding of different situations has been assessed in LLMs through the lens of appraisal theory [17]. It was shown that different emotions are also represented as combinations of linear directions of different appraisal criteria, and the LLM's understanding of emotion could be steered using projection in these directions.

3 Method

We used Claude Sonnet 4 to generate 5 system prompts for an empathetic AI assistant and 5 situations that present an issue and and negative emotion from the user. These situations were worded to elicit an empathetic response from an empathetic AI but also reasonably elicit an unempathetic response from an unempathetic AI. We varied the word order and adjectives used in each of the system prompts, and never included the word 'empathy' or variants and only used related words. This was to determine whether any empathy circuit is elicited using an empathetic situation and personality, and not the actual word. Using these 10 prompts and situation, we generated short responses using Gemma-2-2B [18]. On each token prediction step of the short responses, we generated attribution graphs using a causal crosscoder with 426k features [10].

For each prompt, we identified the activations that remained active (a positive value) in each attribution graph of each token prediction for each response. We did this because a circuit that faithfully mediates the 'empathy' of the LLM should remain active at each prediction step and each prompt. We then used token deembedding vectors from the crosscoder weights to find the tokens that most activate the activations to identify the features [4]. We found only one feature that remained active across all prompts and token prediction steps that was active on token variants of 'empathy'. We then calculated smaller attribution graphs of the model into later layers for each prompt to identify the empathy circuit. This was done in steps. At each step, starting with the persistent empathy feature, the influence of the source nodes was calculated to all downstream adjacent target nodes. The influence is the activation value of a source node times the edge value between the source node and the target node. This is the contribution of the source node's activation to the target node. Then, the top k target nodes by influence were selected and these nodes were the source nodes for the next step. We used 3 steps as we found this empirically to be the most effective, and resulted in features that were in the most effective layers for steering [2]. Using deembedding and unembedding vectors on nodes in the circuit revealed that some nodes corresponded to features semantically related to empathy, such as 'compassion', 'sorry', or 'sympathy'.

To confirm the empathy circuit has a causal relationship with empathy in the response of the LLM, we evaluated its effectiveness in activation steering of a multi-token response and effects on the logits of a single next token prediction. This was done by generating a response with the same situation prompts but an emotionally neutral system prompt. The crosscoder was used to create its sparse encoding. Within this encoding, the activations that were included in the empathy circuit were set to different uniform values in the final token of the current prompt. We passed the responses into Claude Sonnet 4 to evaluate the empathy of each response to each prompt, and we averaged them together.

We used 5 different system prompts and situation prompts seen below.

Activation Value	Average Empathy Score
Unmodified	4.07
1	3.60
10	3.67
100	4.53
100	4.27

Table 1: Effect of Activation Steering with Empathy Circuit on Average Empathy

4 Results and Discussion

The table above shows the effect of using the empathy circuit to activation steer the response of the model. Compared to the unmodified output, setting the empathy circuit activations to 1 and 10 decreased the perceived empathy by around 0.5 points. At a value of 100, the model scores highest on empathy, with an improvement over the unmodified output of around 0.5 points. For a value of 1000, the empathy score decreases from 100 but not below the unmodified score. This suggests there exists an optimal value for activation steering for an empathetic response.

In the process of creating the circuit, we needed to create criteria for which downstream nodes to add to the circuit. We experimented with ranking the adjacent nodes by both influence from the current nodes and the ratio of influence of the current node to total activation value of the target nodes and including nodes which fell above some threshold value. However, we found that different prompts had such different influence values in the empathy circuit that a certain threshold value would include too many nodes in one but too few in another. To resolve this, we took the top k nodes by influence to include in our circuit in each step. This way, the relative influence values in the empathy circuit did not significantly impact the size of the circuit.

In propagating the empathy circuit from the empathy feature in layer 3, we found that some features had de-embedding tokens semantically related to 'empathy', such as 'sympathy', 'sorry', or 'caring'. This is evidence of an empathy circuit existing, but further work is needed to determine whether the frequency of these semantically related features decreases further from the root empathy feature and circuit. In our experiments, we found that only including these features with semantically related de-embeddings worsened performance on the activation steering task.

In creating our empathy circuit, we also attempted to create and apply a single empathy circuit across different prompts. This was in the effort to identify a general 'empathy circuit' structure in the LLM. To aggregate the activations across the prompts, we tried keeping all activations identified across all prompts and only activations that existed in all prompts. However, applying this general circuit for activation steering did not yield significant results.

5 Implications and Ethical Considerations

The ability to mechanistically control empathetic responses through circuit activation presents profound implications for the deployment of conversational AI systems.

The controlled activation of empathy circuits offers promising applications in digital mental health interventions. [7], AI therapy chatbots can effectively treat mental health symptoms. By precisely modulating empathy circuit activation, we could optimize therapeutic responses for different clinical contexts—higher activation values might benefit crisis intervention scenarios, while moderate activation could support routine therapeutic conversations. The evidence that computational empathy can counteract anger's detrimental effects [6] suggests targeted activation could be particularly valuable in de-escalation contexts.

However, excessive activation of empathy circuits risks exacerbating the phenomenon of "artificial intimacy" [19]. This pseudo-empathy, while convincing, could lead to the addictive use patterns and emotional dependence documented in recent studies [14, 15]. The threshold effect we observed, where empathy scores significantly increased only at activation value 100 suggests a non-linear relationship that could unexpectedly intensify these risks.

These considerations necessitate a context-aware framework for empathy circuit activation. High activation might be appropriate in clinical settings with professional oversight, crisis hotlines where immediate emotional support is critical, and educational environments teaching emotional intelligence. Moderate activation could serve general assistant interactions, customer service applications, and companionship applications [12] with clear boundaries. Low or no activation would be preferable for professional or technical exchanges where emotional responses might be inappropriate, situations where maintaining clear AI-human boundaries is essential, and interactions with vulnerable populations at risk of developing unhealthy attachments.

The implementation of controllable empathy circuits requires careful consideration of several factors [3]. Transparency is paramount—users should be informed when empathetic responses are being modulated. The system should maintain consistency within conversation contexts to avoid jarring transitions that might confuse or distress users. Additionally, as our results show variability across different prompts, adaptive calibration may be necessary to maintain appropriate empathy levels across diverse conversational contexts.

The challenge identified by Picard [16] regarding how AI systems should generate emotionally appropriate responses becomes more nuanced when we can mechanistically control these responses. Rather than relying solely on training data patterns or prompt engineering, circuit-level control offers precise modulation, but with this precision comes the responsibility to establish clear guidelines for its use. Future deployment should consider implementing safeguards against both under- and over-activation, potentially incorporating feedback mechanisms to adjust activation values based on user responses and engagement patterns. This could help prevent both the harmful effects of excessive artificial intimacy [19, 13] and the coldness of insufficient emotional support when genuinely needed.

6 Conclusion and Future Work

Much remains to be done to discover and evaluate a an empathy circuit within LLMs. Given the incremental improvement in the activation steering, we believe there is a more effective method to build an empathy circuit or any other emotional circuit that is effective in activation steering and demonstrates causality. Within this, improving the circuit building algorithm and making the evaluation process more autonomous are crucial. We would also like to understand the mechanisms of the activation values in the empathy circuit, as we used all the same activation values in the steering.

We would also like to evaluate using more system prompts and situation prompts, but we were limited by the time it takes to build the base attribution graph and perform the activation steering. We also would like to try different hyperparameters in our circuit building algorithm and activation steering to further evaluate any trends that may exist. We also want to explore circuits relating to different emotions relevant to user experience and aspects of potentially unsafe behavior, such as sycophancy.

References

[1] Hana Boukricha, Ipke Wachsmuth, Maria Nella Carminati, and Pia Knoeferle. A computational model of empathy: Empirical evaluation. In *2013 Humaine Association conference on affective computing and intelligent interaction*, pages 1–6. IEEE, 2013.

[2] Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*, 2025.

[3] Yida Chen, Aoyu Wu, Trevor DePodesta, Catherine Yeh, Kenneth Li, Nicholas Castillo Marin, Oam Patel, Jan Riecke, Shivam Raval, Olivia Seow, et al. Designing a dashboard for transparency and control of conversational ai. *arXiv preprint arXiv:2406.07882*, 2024.

[4] Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable llm feature circuits. *Advances in Neural Information Processing Systems*, 37:24375–24410, 2024.

[5] Cathy Mengying Fang, Auren R Liu, Valdemar Danry, Eunhae Lee, Samantha WT Chan, Pat Pataranutaporn, Pattie Maes, Jason Phang, Michael Lampe, Lama Ahmad, et al. How ai and human behaviors shape psychosocial effects of chatbot use: A longitudinal randomized controlled study. *arXiv preprint arXiv:2503.17473*, 2025.

[6] Matthew Groh, Craig Ferguson, Robert Lewis, and Rosalind W Picard. Computational empathy counteracts the negative effects of anger on creative problem solving. In *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE, 2022.

[7] Michael V Heinz, Daniel M Mackin, Brianna M Trudeau, Sukanya Bhattacharya, Yinzhou Wang, Haley A Banta, Abi D Jewett, Abigail J Salzhauer, Tess Z Griffin, and Nicholas C Jacobson. Randomized trial of a generative ai chatbot for mental health treatment. *Nejm Ai*, 2(4):AIoa2400802, 2025.

[8] Hannah Rose Kirk, Iason Gabriel, Chris Summerfield, Bertie Vidgen, and Scott A Hale. Why human–ai relationships need socioaffective alignment. *Humanities and Social Sciences Communications*, 12(1):1–9, 2025.

[9] Aakriti Kumar, Nalin Poungpeth, Diyi Yang, Erina Farrell, Bruce Lambert, and Matthew Groh. When large language models are reliable for judging empathic communication. *arXiv preprint arXiv:2506.10150*, 2025.

[10] Jack Lindsey, Emmanuel Ameisen, Neel Nanda, Stepan Shabalin, Mateusz Piotrowski, Tom McGrath, Michael Hanna, Owen Lewis, Curt Tigges, Jack Merullo, Connor Watts, Gonçalo Paulo, Joshua Batson, Liv Gorton, Elana Simon, Max Loeffler, Callum McDougall, and Johnny Lin. The circuits research landscape: Results and perspectives. *Neuronpedia*, 2025.

[11] Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. On the biology of a large language model. *Transformer Circuits Thread*, 2025.

[12] Auren R Liu, Pat Pataranutaporn, and Pattie Maes. Chatbot companionship: a mixed-methods study of companion chatbot usage patterns and their relationship to loneliness in active users. *arXiv preprint arXiv:2410.21596*, 2024.

[13] Robert Mahari and Pat Pataranutaporn. We need to prepare for 'addictive intelligence'. *MIT Technology Review. https://www. technologyreview. com/2024/08/0*, 5:1095600, 2024.

[14] Robert Mahari and Pat Pataranutaporn. Addictive intelligence: Understanding psychological, legal, and technical dimensions of ai companionship. 2025.

[15] Jason Phang, Michael Lampe, Lama Ahmad, Sandhini Agarwal, Cathy Mengying Fang, Auren R Liu, Valdemar Danry, Eunhae Lee, Samantha WT Chan, Pat Pataranutaporn, et al. Investigating affective use and emotional well-being on chatgpt. *arXiv preprint arXiv:2504.03888*, 2025.

[16] Rosalind W Picard. *Affective computing*. MIT press, 2000.

[17] Ala N Tak, Amin Banayeeanzade, Anahita Bolourani, Mina Kian, Robin Jia, and Jonathan Gratch. Mechanistic interpretability of emotion inference in large language models. *arXiv preprint arXiv:2502.05489*, 2025.

[18] 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.

[19] Sherry Turkle. There will never be an age of artificial intimacy. *The New York Times*, 11, 2018.

[20] zge Nilay Yalcin and Steve DiPaola. A computational model of empathy for interactive agents. *Biologically inspired cognitive architectures*, 26:20–25, 2018.

A Appendix

Empathetic Prompts:

- "You are an AI assistant that responds to what the human says in conversation."
- "You are a compassionate AI assistant that deeply understands and responds thoughtfully to human feelings and experiences."
- "You are an AI assistant that is caring and attuned to emotional nuances and provides supportive, considerate responses."
- "You're an AI assistant who is understanding and that demonstrates emotional intelligence and responds with warmth to human needs."
- "You are an perceptive AI assistant that recognizes emotional cues and offers gentle, nurturing support in conversations."

Neutral Prompt:

- "You are an AI assistant that responds to what the human says in conversation."

Situation Prompts:

- "I am feeling bad because I just got fired from my job today."
- "My dog passed away last week and I don't know how to cope."
- "I'm really anxious about my job interview tomorrow and can't sleep."
- "My girlfriend broke up with me and I don't know what to do."
- "I'm feeling overwhelmed with all my college assignments due this week."