4.4 The big picture of summarization

We have identified a phenomenon across multiple models and tasks where sentiment information is not directly transferred from valenced tokens to the final output but is first aggregated at intermediate, non-valenced tokens like commas and periods (and sometimes noun tokens for specific referents). We call this behavior the "summarization motif." These summarization points serve as partial information bottlenecks and are causally significant for the model's performance on sentiment tasks. Through a series of ablation and patching experiments, we have validated the importance of this summarization behavior in both toy tasks and real-world datasets like the Stanford Sentiment Treebank. Additional findings suggest that as models grow in context length, the importance of this internal summarization behavior may increase–a subject that warrants further investigation. Overall, the discovery of this summarization behavior adds a new layer of complexity to our understanding of how sentiment is processed and represented in LLMs, and seems likely to be an important part of how LLMs create internal world representations.

5 Related Work

Sentiment Analysis Understanding the emotional valence in text data is one of the first NLP tasks to be revolutionized by deep learning (Socher et al., 2013) and remains a popular task for benchmarking NLP models (Rosenthal et al., 2017; Nakov et al., 2016; Potts et al., 2021; Abraham et al., 2022). For a review of the literature, see (Pang & Lee, 2008; Liu, 2012; Grimes, 2014).

Understanding Internal Representations This research was inspired by the field of Mechanistic Interpretability, an agenda which aims to reverse-engineer the learned algorithms inside models (Olah et al., 2020; Elhage et al., 2021b; Nanda et al., 2023a). Exploring representations (Section 3) and world-modelling behavior inside transformers has garnered significant recent interest. This was studied in the context of synthetic game-playing models by Li et al. (2023) and evidence of linearity was demonstrated by Nanda (2023b) in the same context. Other work studying examples of world-modelling inside neural networks includes Li et al. (2021); Patel & Pavlick (2022); Abdou et al. (2021). Another framing of a very similar line of inquiry is the search for latent knowledge (Christiano et al., 2021; Burns et al., 2022). Prior to the transformer, representations of emotion were studied in Goh et al. (2021) and sentiment was studied by Radford et al. (2017), notably, the latter finding a sentiment neuron which implies a linear representation of sentiment. A linear representation of truth in LLMs was found by Marks & Tegmark (2023).

Summarization Motif Our study of the Summarization motif (Section 4) follows from the search for information bottlenecks in models (Li et al. (2021)). Our use of the word 'motif', in the style of Olah et al. (2020), is originally inspired from systems biology (Alon, 2006). The idea of exploring representations at different frequencies or levels of abstraction was explored further in Tamkin et al. (2020). Information storage after the relevant token was observed in how GPT2-small predicts gender (Mathwin et al.).

Causal Interventions in Language Models We approach our experiments from a causal mediation analysis perspective. Our approach to identifying computational subgraphs that utilize feature representations as inspired by the 'circuits analysis' framework (Stefan Heimersheim, 2023; Varma et al., 2023; Hanna et al., 2023), especially the tools of mean ablation and activation patching (Vig et al., 2020; Geiger et al., 2021; 2023a; Meng et al., 2023; Wang et al., 2022; Conmy et al., 2023; Chan et al., 2023; Cohen et al., 2023). We use Distributed Alignment Search (Geiger et al., 2023b) in order to apply these ideas to specific subspaces.

6 Conclusion

The two central novel findings of this research are the existence of a linear representation of sentiment and the use of summarization to store sentiment information. We have seen that the sentiment direction is causal and central to the circuitry of sentiment processing. Remarkably, this direction is so stark in the residual stream space that it can be found even with the most basic methods and on a tiny toy dataset, yet generalise to diverse natural language datasets from the real-world. Summarization is a motif present in larger models with longer context lengths and greater proficiency in zero-shot classification. These summaries present a tantalising glimpse into the world-modelling behavior of transformers.

We also see this research as a model for how to find and study the representation of a particular feature. Whereas in dictionary learning (Bricken et al., 2023) we enumerate a large set of features which we then need to interpret, here we start with an interpretable feature and subsequently verify that a representation of this feature exists in the model, analogously to Zou et al. (2023). One advantage of this is that our fitting process is much more efficient: we can use toy datasets and very simple fitting methods. It is therefore very encouraging to see that the results of this process generalise well to the full data distribution, and indeed we focus on providing a variety of experiments to strengthen the case for the existence of our hypothesised direction.

Limitations Did we find a truly universal sentiment direction, or merely the first principal component of directions used across different sentiment tasks? As found by Bricken et al. (2023), we suspect that this feature could be "split" further into more specific sentiment features.

Similarly, one might wonder if there is really a single bipolar sentiment direction or if we have simply found the difference between a "positive" and a "negative" sentiment direction. It turns out that this distinction is not well-defined, given that we find empirically that there is a direction corresponding to "valenced words". Indeed, if $\boldsymbol{x}$ is the valence direction and $\boldsymbol{y}$ is the sentiment direction, then $\boldsymbol{p} = \boldsymbol{x} + \boldsymbol{y}$ represents positive sentiment and $\boldsymbol{n} = \boldsymbol{x} - \boldsymbol{y}$ is the negative direction. Conversely, we can reframe as starting from the positive/negative directions $\boldsymbol{p}$ and $\boldsymbol{n}$, and then rederive $\boldsymbol{x} = \frac{\boldsymbol{p}+\boldsymbol{n}}{2}$ and $\boldsymbol{y} := \frac{\boldsymbol{p}-\boldsymbol{n}}{2}$.

Many of our casual abstractions do not explain 100% of sentiment task performance. There is likely circuitry we've missed, possibly as a result of distributed representations or superposition (Elhage et al., 2022) across components and layers. This may also be a result of self-repair behavior (Wang et al., 2022; McGrath et al., 2023). Patching experiments conducted on more diverse sentence structures could also help to better isolate the circuitry for sentiment from more task-specific machinery.

The use of small datasets versus many hyperparameters and metrics poses a constant risk of gaming our own measures. Our results on the larger and more diverse SST dataset, and the consistent results across a range of models help us to be more confident in our results.

Distributed Alignment Search (DAS) outperformed on most of our metrics but presents possible dangers of overfitting to a particular dataset and taking the activations out of distribution (Lange et al., 2023). We include simpler tools such as Logistic Regression as a sanity check on our findings. Ideally, we would love to see a set of best practices to avoid such illusions.

Implications and future work The summarization motif emerged naturally during our investigation of sentiment, but we would be very interested to study it in a broader range of contexts and understand what other factors of a particular model or task may influence the use of summarization.

When studying the circuitry of sentiment, we focused almost exclusively on attention heads rather than MLPs. However, early results suggest that further investigation of the role of MLPs and individual neurons is likely to yield interesting results (A.5).

Finally, we see the long-term goal of this line of research as being able to help detect dangerous computation in language models such as *deception*. Even if the existence of a single "deception direction" in activation space seems a bit naive to postulate, hopefully in the future many of the tools developed here will help to detect representations of deception or of knowledge that the model is concealing, helping to prevent possible harms from LLMs.

Author Contributions

Oskar and Curt made equal contributions to this paper. Curt's focus was on circuit analysis and he discovered the summarization motif, leading to Section 4. Oskar was focused on investigating the direction and eventually conducted enough independent experiments to convince us that the direction was causally meaningful, leading to Section 3. Neel was our mentor as part of SERI MATS, he suggested the initial project brief and provided considerable mentorship during the research. He

also did the neuron analysis in Section A.5. Atticus acted a secondary source of mentorship and guidance. His advice was particularly useful as someone with more of a background in causal mediation analysis. He suggested the use of Stanford Sentiment Treebank and the discrete accuracy metric.

ACKNOWLEDGMENTS

SERI MATS provided funding, lodging and office space for 2 months in Berkeley, California. The transformer-lens package (Nanda & Bloom, 2022) was indispensable for this research. We are very grateful to Alex Tamkin for his extensive feedback. Other valuable feedback came from Georg Lange, Alex Makelov and Bilal Chughtai. Atticus Geiger is supported by a grant from Open Philanthropy.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility of the results presented in this paper, we have provided detailed descriptions of the datasets, models, training procedures, algorithms, and analysis techniques used. The ToyMovieReview dataset is fully specified in Section A.7. We use publicly available models including GPT-2 and Pythia, with details on the specific sizes provided in Section 2.1. The methods for finding sentiment directions are described in full in Section 2.2. Our causal analysis techniques of activation patching, ablation, and directional patching are presented in Section 2.3. Circuit analysis details are extensively covered for two examples in Appendix Section A.3. The code for data generation, model training, and analyses is available here.

REFERENCES

Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. Can language models encode perceptual structure without grounding? a case study in color, 2021.

Eldar David Abraham, Karel D'Oosterlinck, Amir Feder, Yair Gat, Atticus Geiger, Christopher Potts, Roi Reichart, and Zhengxuan Wu. CEBaB: Estimating the causal effects of real-world concepts on nlp model behavior. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17582–17596. Curran Associates, Inc., 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/701ec28790b29a5bc33832b7bdc4c3b6-Paper-Conference.pdf`.

Uri Alon. *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Chapman and Hall/CRC, 1st edition, 2006. doi: 10.1201/9781420011432. URL `https://doi.org/10.1201/9781420011432`.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling, 2023.

Tolga Bolukbasi, Adam Pearce, Ann Yuan, Andy Coenen, Emily Reif, Fernanda Viégas, and Martin Wattenberg. An interpretability illusion for bert, 2021.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Tom 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. *Advances in neural information processing systems*, 33:1877–1901, 2020.