I Have No Mouth, and I Must Rhyme: Uncovering Internal Phonetic Representations in LLaMA 3.2

Oliver McLaughlin [1] **Arjun Khurana** [1] **Jack Merullo** [1]

Abstract

Large language models demonstrate proficiency on phonetic tasks, such as rhyming, without explicit phonetic or auditory grounding. In this work, we investigate how `Llama-3.2-1B-Instruct` represents token-level phonetic information. Our results suggest that Llama uses a rich internal model of phonemes to complete phonetic tasks. We provide evidence for high-level organization of phoneme representations in its latent space. In doing so, we also identify a "phoneme mover head" which promotes phonetic information during rhyming tasks. We visualize the output space of this head and find that, while notable differences exist, Llama learns a model of vowels similar to the standard IPA vowel chart for humans, despite receiving no direct supervision to do so.

1. Introduction

Language models ("LMs") cannot hear speech. In spite of this, many large LMs can consistently produce rhymes, poetry, alliteration, and other phenomena which seem to require a fundamental understanding of phonetic properties. We propose that these models complete these tasks using underlying token-level phonetic representations as well as circuits to retrieve phonetic information.

Up to this point, study of LMs' phonetic behavior has been largely limited to training models on grounded phonetic information (English et al., 2023; Popescu-Belis et al., 2023). Anthropic's recent work *On the Biology of a Large Language Model* (Lindsey et al., 2025) uses attribution graphs to provide evidence of rhyme planning circuits in Claude 3.5 Haiku, but they do not rigorously study the internal phonetic representations that allow these circuits to function.

It is reasonable to ask the extent to which an LM is able to infer phonetic properties directly from tokens. Following results that LMs and even simple text models can extract information about the world without direct supervision (Louwerse and Benesh, 2012; Mikolov et al., 2013; Gurnee and Tegmark), we hypothesize that LMs also learn a robust, structured model of phonetic information that supports interventions. We investigate how Llama 3.2 (Grattafiori et al., 2024) represents phonetic information through methods commonly used in interpretability analysis and find evidence to support this hypothesis: rather than simply memorizing phonetic information for various tokens, Llama uses structure in its latent space across tokens to represent the phonetics of given input tokens.

Our experiments follow a straightforward methodology. To explore linear token-level phonetic information, we employ linear probes to identify subspaces of the residual stream and embedding and inspect the subspaces. To explore the mechanisms that leverage this information, we isolate components of the model which prove to be impactful in expressing the LM's phonetic beliefs. Using this simple approach, we find evidence of rich phonetic representations within `Llama-3.2-1B-Instruct`.

In Section 2, we find vectors in the embedding space corresponding to common English phonemes. We perform causal interventions in the embedding space using these vectors to alter the model's performance on a rhyming task, demonstrating their role in rhyming processes. In Section 3 we identify a single "phoneme mover head" using activation patching. We decode this head's result vectors using logit lens (nostalgebraist, 2020) and demonstrate that the phonetic information in this head is, to some extent, cross-lingual. In Section 4, we use the result vectors of this phoneme mover head to perform a phonetically-informed dimensionality reduction on the embedding space in order

[1]Brown University, Providence, Rhode Island. Correspondence to: Oliver McLaughlin <oliver_mclaughlin@brown.edu>, Arjun Khurana <arjun_khurana@alumni.brown.edu>, Jack Merullo <jack_merullo@brown.edu>.

ICML 2025 Workshop on Assessing World Models.

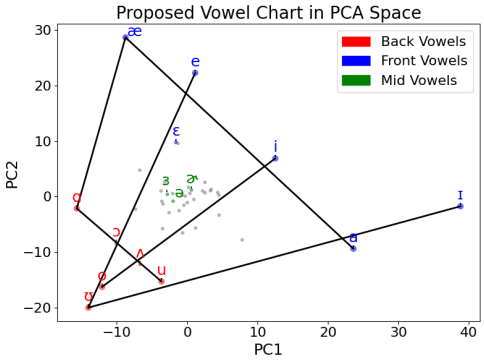

Figure 1: Patterns emerge from vowel representations under our methodology, revealing a world model partially inconsistent with human anatomy.

to analyze the geometry of the phoneme vectors from Section 2. We see consistent linear patterns across internal representations of vowel phonemes (Figure 1).

These patterns differ from anatomical properties of vowel phonemes, suggesting the presence of a robust internal vowel model which is distinct from human vowel models. We construct an organized representation of these patterns and contrast it with existing anatomically-grounded vowel organization systems.

2. Phonetic Information in Token Embeddings

In Llama 3.2, like in most open-source LMs, the tokenization process does not explicitly encode any phonetic information about input tokens. We hypothesize that some phonetic information is encoded through token embeddings. To investigate this hypothesis, we train a multi-hot linear probe to predict which phonemes are present in a word from its embedding. We differentiate phonemes based on the International Phonetic Alphabet, using WikiPron (Lee et al., 2020) as a pronunciation reference.

Probing the embedding space Our probe predicts the correct phonemes for approximately 96% of single-token words, compared with 42% for the same probe architecture trained on embeddings from a randomly-generated embedding matrix. This indicates that the model's embedding matrix encodes some amount of recoverable phonetic information.

Since this probe is linear, it essentially constitutes a linear map from Llama 3.2's 2048-dimensional embedding space to what we call an "IPA phoneme space": a 44-dimensional space with one axis for each common English-language phoneme. Each row of our probe matrix, therefore, could be considered a representation of its corresponding phoneme in the embedding space.

Causal interventions on embeddings To test the degree to which the rows of our probe represent phonemes in latent space, we run a causal intervention experiment to change the expected rhyming output for a given target word by intervening on embedding vectors. We test the model on a simple rhyming task:

```
prompt = """Here are a few examples of words
that rhyme with <word>:"""
```

where `<word>` is a single-token word with one unique vowel sound. We then select two "phoneme vectors" (rows of our probe matrix) in the embedding space: ξ, the phoneme vector corresponding to the vowel in `<word>`; and μ, the phoneme vector corresponding to a different vowel. We perform a forward pass of Llama with the following intervention at the embedding step:

$$E = E + c(\mu - \xi)$$

where E is the embedding vector corresponding to our rhyming word `<word>`, and c is a scalar. As we increase c (the weight of our intervention), the model's prediction tends to switch from words with the ξ vowel to words with the μ vowel. Figure 2 shows example model output results where `<word>` is `leet` pronounced /liːt/, ξ corresponds to /i/, and μ corresponds to ɛ.

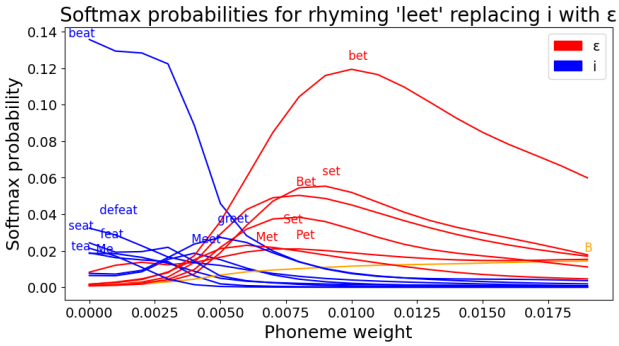

Figure 2: Example intervention on rhymes with `leet`. Blue indicates that the ξ phoneme /i/ is present, red indicates that the μ phoneme /ɛ/ is present.

3. Phoneme Mover Head

Given that embeddings seem to contain phonetic information, we seek to identify model components which use this information to complete rhyming tasks. We run activation patching experiments over all attention heads and MLP components across all layers.

BOS Here are a few examples of words that rhyme with grab : *
→ **Promoted by head**: Gab, Bob, ブ, bara, Barbara, bo, gado, boot, Bob, b
→ **Top predictions**: Cab, cab, dab, Tab, Crab, Grab, Br, tab, Sab, b
BOS Here are a few examples of words that rhyme with plush : *
→ **Promoted by head**: ush, uss, ブ, uch, us, usk, usses, usc, ش, USH
→ **Top predictions**: crush, lush, h, Bush, flush, splash, plush, rush, crunch, Crush
BOS Here are a few examples of words that rhyme with clean : *
→ **Promoted by head**: enic, lean, enal, enas, ean, Wein, िन, Ben, LEAN, Neil
→ **Top predictions**: keen, clean, Screen, green, Seen, serene, seen, screen, Green, Clean

Figure 3: Across three examples: Attention patterns of H13L12 at the final token shows the head attends to the rhyme target token. Running logit lens on the corresponding result vector (shown after →) produces phonetically similar tokens.

Patching setup To patch, we perform two full passes of the model using parallel prompts containing two sufficiently different[1] rhyme target words. For example, setting `<word>="clean"` in our above template results in the top predicted token `"keen"`. Similarly, setting `<word>="track"` results in `"back"`. These then form our "clean" and "corrupted" runs (in line with Meng et al. (2023)) allowing us to inspect the mean normalized logit difference between `"keen"` and `"back"`. Head 13 of Layer 12 ("H13L12") emerged as the most critical for our rhyming task with a mean normalized logit difference of 0.48 (out of 1). This was significantly higher than both the mean (0.002) and the second highest value (0.19). See Appendix B for a visual representation.

Result vectors and cross-lingual features To understand the effect of this head on task completion, we explore its contributions to the residual stream by inspecting its result vectors[2]. We decode these vectors into the vocab space using logit lens (nostalgebraist, 2020) in order to study the tokens this head promotes. Figure 3 demonstrates these results along with last-token attention patterns. We detail further study of the head dimension in Appendix C.

There is a clear, if approximate, phonetic similarity between the target rhyme word and the decoded result vectors of H13L12. These results and the associated attention patterns suggest that this head moves phonetic information from the target word to the final token's residual stream. Because of this apparent phonetic promotion behavior, we call this head a "phoneme mover head". We investigate the connection between the result vectors produced by the head and completion of our rhyming task in Appendix D.

[1]Two words are "*sufficiently different*" iff there is no third word that rhymes with both simultaneously.

[2]i.e. the projection of a single head back to the residual stream dimension through its output matrix, see Elhage et al. (2021)

We also see clear evidence of the LM modeling phonetic information across languages, as decoding result vectors produces language-agnostic sets of similar phonemes. In Figure 3, for instance, we see that one of the top tokens promoted into the residual stream for `plush` is "ش" /ʃ/, for `clean` we see "ीन" (/iːn/), and for `grab` we see "ブ" (/bɯ/), among others.

Further Phonetic Heads Upon further inspection we discovered a set of heads (H13L12, H21L14, H22L14) with nearly identical attention patterns. After zero ablating all three of these heads, we noticed that the model could no longer correctly produce a *single* rhyming token. Instead, it produced a first token (typically a single letter) and then a corresponding second token which typically completed the response word (For example if the target word was "plush" we might get "l" "ush"). Results were generally mixed as to if the response produced was truly a correct rhyme or not. Omitting any one of these three heads from ablation reinstated the model's original rhyming ability. Of further interest is that the composition scores (Elhage et al., 2021) for all three of these heads and all prior heads were also essentially identical, suggesting a common channel.

4. Geometry of Phoneme Vectors

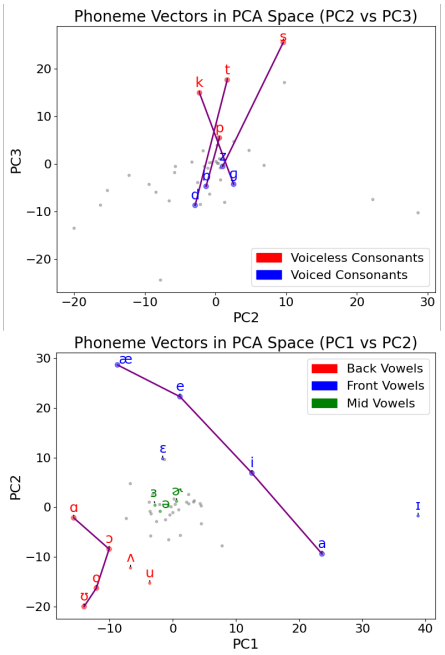

Figure 4: Phoneme vectors reduced to 2 dimensions using PCA trained on H13L12 result vectors. Phonetic patterns emerge among consonant voicedness (top) and vowel backness (bottom).

We visualize the result vectors of Head H13L12 for 5742 different target words on our template with principal component analysis (PCA) to understand how the model organizes phonetic information. We expect that if the model has an organized representation of phonetic information, we should see similar phonemes cluster together regardless of their source word.

For consonants, Principal Components 2 and 3 contain the least noisy phonetic information, in line with findings from Engels et al. (2025). Figure 4 shows that phoneme vectors corresponding to voiced consonants versus their voiceless equivalents follow a consistent linear pattern across PC3.

Vowels also follow consistent patterns in this space. These patterns partially align with anatomically-based IPA distinctions, but there are some key differences. Figure 4 shows that PC1 and PC2 seem to distinguish vowel backness: front vowels tend to have positive PCs, mid vowels center around the origin, and back vowels tend to have negative PCs. Likewise, vowel openness order is consistent within backness classes: more closed vowels have lesser PC2 components than more open vowels in the same backness class.

The notable exceptions are the extreme vowels. For instance, /a/, an open front vowel, has a large PC1 and small PC2, placing it at the "close" end of the front vowels in PCA space. /ɪ/, a near-close near-front vowel, has an unusually large PC1 value for its backness.

Llama's vowel chart It is important to remember that these phonetic representations are not explicitly grounded in model training. Rather, these patterns emerge from its architecture and training data. The patterns partially align with IPA representations based on the anatomy of a human mouth, which Llama does not have. We suspect that instances which break these patterns, such as /u/, /a/, and /ɪ/, emerge from the usage of these particular phonemes in context[3].

As an attempt to show this portion of Llama's internal vowel model in a human-readable manner, we construct a variant of the IPA vowel chart which is consistent with our findings, populated with common English vowel sounds (Figure 5). This chart maintains the surprising emergent colinearities we find in our PCA of the vowel vectors, as shown by the overlay of this chart in Figure 1. H13L12 result vectors also align with these geometries (see Appendix E).

[3]These particular phonemes are often present in English language diphthongs, which could explain the divergence from anatomical characteristics.

We note that the representations in Figure 5 do not necessarily constitute a comprehensive account of Llama's internal representations of phonetic information. However, the fact that we see some degree of linearity — and the fact that this linearity aligns, to some degree, with existing anatomical representations — indicates the presence of some kind of relativistic world model of vowel phonemes between the embedding space and the result vector of H13L12.

5. Future Work

While our experiments provide evidence of robust phonetic representation in `Llama-3.2-1B-Instruct`'s embedding and residual stream, several important questions remain. Our findings open up multiple avenues for further research, especially regarding the mechanisms and structures through which the model propagates and processes phonetic information. In particular, we are interested in further characterizing H13L12's behaviors, better understanding the cross-lingual features in the model's phonetic system, and isolating the propagation of phonetic information.

6. Conclusion

Our investigation provides strong evidence that `Llama-3.2-1B-Instruct` contains a rich internal phonetic model which partially diverges from human anatomical phonetic models. We demonstrate the recoverability of phonetic information from token embeddings, which produces directions for each phoneme in latent space, and we identify a "phoneme mover head" (H13L12) that moves and promotes phonetic information during rhyming tasks. We use the result vectors of this head to fit a phonetically-informed PCA. Applying this PCA to our latent space phoneme vectors yields a geometric arrangement of phonemes. The pattern of vowel phoneme vectors in particular suggests the use of an internal vowel model which partly aligns with anatomically-informed vowel taxonomies.

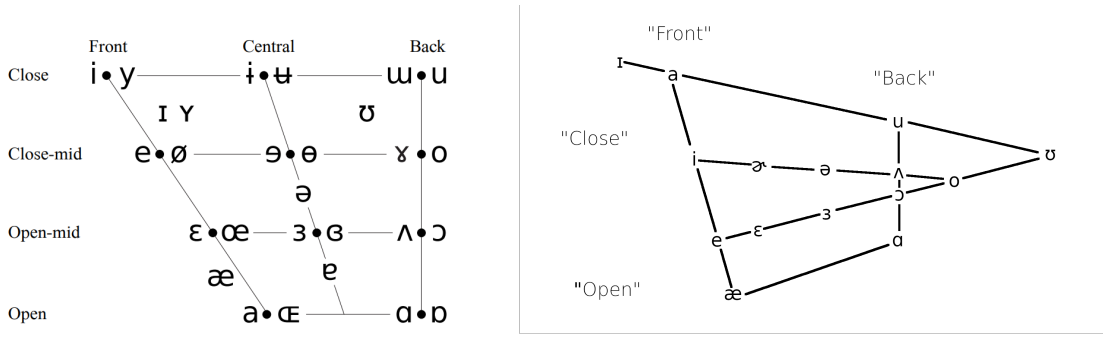

Figure 5: Standard IPA representation of vowels in humans (left) and a proposed IPA-style representation of Llama's internal vowel model (right). The positions of some phonemes, such as /a/ and /ɪ/, notably differ.

References

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear, 2025. URL `https://arxiv.org/abs/2405.14860`.

Patrick Cormac English, John D. Kelleher, and Julie Carson-Berndsen. Discovering phonetic feature event patterns in transformer embeddings. In *Interspeech 2023*, pages 4733–4737, 2023. doi: 10.21437/Interspeech.2023-1985.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross

Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim

Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Wes Gurnee and Max Tegmark. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*.

Tayyip Güzel. tyypgzl/oxford-5000-words, May 2025. URL https://github.com/tyypgzl/Oxford-5000-words.

Connor Kissane, Robert Krzyzanowski, Joseph Isaac Bloom, Arthur Conmy, and Neel Nanda. Interpreting attention layer outputs with sparse autoencoders. In *ICML 2024 Workshop on Mechanistic Interpretability*.

Jackson L. Lee, Lucas F.E. Ashby, M. Elizabeth Garza, Yeonju Lee-Sikka, Sean Miller, Alan Wong, Arya D. McCarthy, and Kyle Gorman. Massively multilingual pronunciation modeling with WikiPron. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4223–4228, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.521/.

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. URL https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

Max M Louwerse and Nick Benesh. Representing spatial structure through maps and language: Lord of the rings encodes the spatial structure of middle earth. *Cognitive science*, 36(8):1556–1569, 2012.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt, 2023. URL https://arxiv.org/abs/2202.05262.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Talking heads: Understanding inter-layer communication in transformer language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=LUsx0chTsL.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

nostalgebraist. Interpreting GPT: The logit lens. *LessWrong*, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Andrei Popescu-Belis, Àlex R. Atrio, Bastien Bernath, Etienne Boisson, Teo Ferrari, Xavier Theimer-Lienhard, and Giorgos Vernikos. GPoeT: a language model trained for rhyme generation on synthetic data. In Stefania Degaetano-Ortlieb, Anna Kazantseva, Nils Reiter, and Stan Szpakowicz, editors, *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 10–20, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.latechclfl-1.2. URL https://aclanthology.org/2023.latechclfl-1.2/.

A. Emergence of third-party vowels in interventions

In the Figure 2 example, the vowel sounds remain relatively isolated. Sometimes, however, unexpected third-party vowel sounds appear on the fringes, such as small /o/-to-/ɛ/ interventions manifesting /i/ sounds, as shown in Figure 6.

These "intermediate" vowel sounds often appear in consistent patterns for consistent ξ and μ vowels, regardless of the consonants involved in the starting

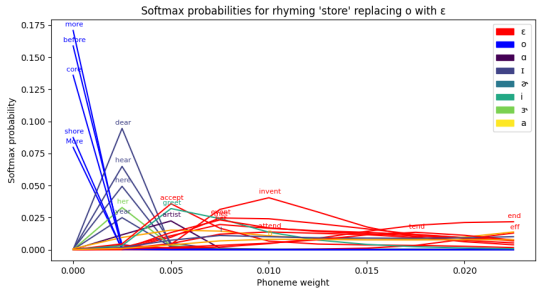

Figure 6: Example intervention on rhymes with `store`. When replacing the /o/ sound with /ɛ/ at small phoneme weight values, the model predicts words containing /i/. This is consistent across multiple starting words containing /o/.

word, which suggests that the model internally represents vowel sounds in the embedding space with meaningful relationships to each other. This motivates our attention to colinearities of phoneme vectors in PCA space in Section 4. When applying this methodology to non-phonetic interpretability tasks, we suggest carefully studying the presence of third-party characteristics (i.e. attributes of model results which are not strictly related to the specific intervention) in order to inform both the choice of dimensionality reduction technique and the expected emergent geometric patterns in PCA space.

B. Activation patching results

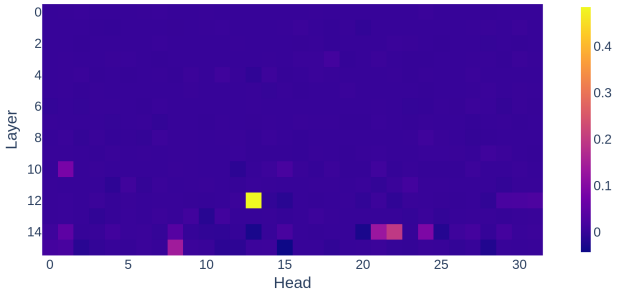

Figure 7: The mean normalized logit differences for our activation patching experiment described in Section 3. Head 13 Layer 12 clearly has the highest value at 0.48.

C. Exploring the head dimension

After discovering H13L12, we were interested in understanding both how the internal head dimension is organized (e.g. are there phoneme specific neurons in the output matrix?) and how phonetic information is pulled from the target word into the head.

The model seems to use complex interference patterns to approximate different phonetic directions. This is evidenced by the relative uninterpretability of the head dimension and the result matrix combined with the mutual consistency of similar phonemes within the head dimension with respect to cosine similarity. Notably, we were able to recover much of the result vector (measuring with cosine similarity) with the top and bottom (In terms of magnitude) eight values from the head vector, suggesting a kind of sparsity in the head dimension (See Figure 8). Interestingly, we found that all sixty four head dimensions were, for at least one word, present within the top or bottom eight z values. Future work could investigate the subspaces read from and written to by this head using more involved methods, like sparse coding (Kissane et al.) or communication channels (Elhage et al., 2021; Merullo et al., 2024).

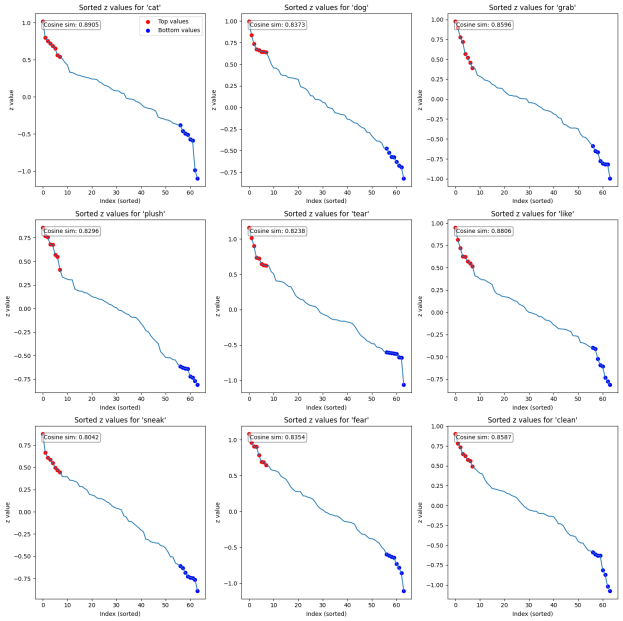

Figure 8: Sorted Z values with the top and bottom eight highlighted. Cosine similarity is between the full result vector and the vector generated by only taking the top and bottom eight values (setting every other index to zero) and passing it through the output matrix.

D. Manual inspection of decoded result vectors

To better understand the role the phoneme mover head plays in downstream rhyme production, we perform a manual survey of one hundred different runs of our

rhyme task (described in Section 2), chosen randomly after filtering for tokenization from the Oxford 5000 dataset (Güzel, 2025).

We judge result vectors according to the following criteria: we call a result vector (R.V.) "*coherent*" if and only if five of the top ten tokens it promotes contain similar phonemes to the target rhyme. For example, if the target word was `plush` then the target rhyme would be /ʌʃ/. We consider the task to be passed if a correct rhyme is within the top ten tokens the model predicts as the next token. Although our sample size

	Coherent R.V.	Incoherent R.V.
Pass	55%	0%
Fail	25%	18%

is quite small, the effect is clear. Rarely occur, if ever, is the task completed successfully when an incoherent result vector is produced. In every example we analyzed, having a coherent R.V. was a prerequisite for completing the task.

E. Result vectors cluster around the proposed vowel chart

The vowel chart we propose in Figure 5 aligns with phoneme vector geometries in PCA space as shown in Figure 1. Importantly, this geometry is also present among result vectors. We use the same PCA to transform the H13L12 result vectors from 2000 single-token single-vowel words. Figure 9 shows that up to linear transformation[4], result vectors cluster around the phoneme vectors corresponding to their vowels.

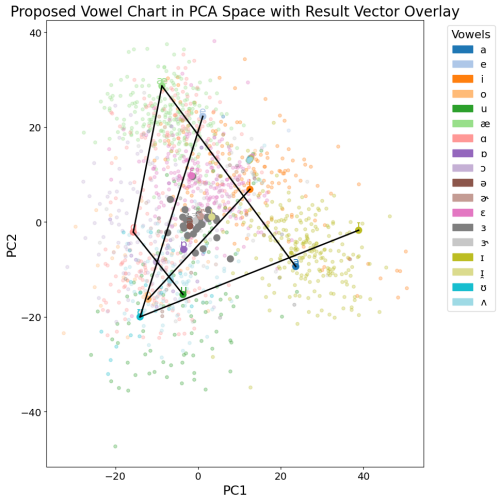

Figure 9: (High opacity) phoneme vectors corresponding to common English vowels in PCA space.
(Low opacity) H13L12 result vectors of single-token single-vowel words in the context of our rhyming task, scaled and shifted. The relative geometry of these result vector clusters matches the geometry of the proposed vowel chart in Figure 5.

[4]The result vectors tend to have much smaller magnitude than our phoneme vectors due to differences in normalization, so we scale up the PCs of the result vectors by a factor of 25 and shift each component by +8.