On Linear Representations and Pretraining Data Frequency in Language Models

Jack Merullo[♠1], **Noah A. Smith**[♣], **Sarah Wiegreffe**[♣], **Yanai Elazar**[♣]

Brown University[♠], Allen Institute for Artificial Intelligence[♣]

jack_merullo@brown.edu

Abstract

Pretraining data has a direct impact on the behaviors and quality of language models (LMs), but we only understand the most basic principles of this relationship. While most work focuses on pretraining data and downstream task behavior, we look at the effect on LM representations. Previous work has discovered that, in language models, some concepts are encoded as "linear representations" argued to be highly interpretable and useful for controllable generation. We study the connection between differences in pretraining data frequency and differences in trained models' linear representations of factual recall relations. We find evidence that the two are directly linked, with the formation of linear representations strongly connected to pretraining term frequencies. First, we establish that the presence of linear representations for subject-relation-object-formatted facts is highly correlated with both subject-object co-occurrence frequency and in-context learning accuracy. This is the case across all phases of pretraining, i.e., it is not affected by the model's underlying capability. In OLMo 7B and GPT-J (6B), we find that a linear representation forms predictably when the subjects and objects within a relation co-occur at least 1–2k times. Thus, it appears linear representations form as a result of consistent repeated occurrences, not due to lengthy pretraining time. In the OLMo 1B model, formation of these features only occurs after 4.4k occurrences. Finally, we train a regression model on measurements of linear representation robustness that can predict how often a term was seen in pretraining with low error, which generalizes to GPT-J without additional training, providing a new unsupervised method for exploring how possible data sources of closed-source models. We conclude that the presence/absence of linear representations contain a weak but significant signal that reflects an imprint of the pretraining corpus across LMs.

1 Introduction

Understanding how the content of pretraining data affects language model (LM) behaviors and performance is a very active area of research [Ma et al., 2024, Xie et al., 2024, Aryabumi et al., 2024, Longpre et al., 2024]. In such work, a common goal is to understand how to understand how to encourage certain behaviors as cheaply as possible. It has been shown, for example, that the frequency of a concept in pretraining encourages the model to use it [Razeghi et al., 2022, Mallen et al., 2023]. The ways in which frequency is affecting the representations of LMs to cause this are unclear, but this problem has been investigated by a separate line of work determining what about the closed-source training data we can infer from an open model [Hayase et al., 2024a, Carlini et al., 2021]. We bridge connect this to recent work in interpretability which focuses on the emergence of

[*]Work done at the Allen Institute for Artificial Intelligence

ATTRIB: Workshop on Attributing Model Behavior at Scale at NeurIPS 2024.

simple *linear features* of concepts in LMs, which strongly correlate with the frequency of terms in the pretraining corpus.

Linear representations/features in LMs have been central to much interpretability work in the past few years [Elhage et al., 2021, Olah et al., 2020, Park et al., 2024, Jiang et al., 2024, Black et al., 2022, Chanin et al., 2024] . The significance of this idea is that much of the behaviors and capabilities of LMs can be localized to directions in activation space, allowing certain behaviors to be activated or modulated by intervening on these directions at inference time (also known as steering) [Todd et al., 2024, Subramani et al., 2022, Hendel et al., 2023, Merullo et al., 2024, Rimsky et al., 2023], or for changing weights for fact editing [Meng et al., 2022]. Testbeds for steering generation often focuses on common concepts like country-capital relations or sentiment. Hernandez et al. [2024], Chanin et al. [2024] both examine how linearity of different types of relations varies greatly depending on the exact relationship being depicted. Such findings complicate the picture of the Linear Representation Hypothesis [Elhage et al., 2021, Park et al., 2024] which proposes that LMs will represent features linearly, without mandating that *all* features be represented this way. Jiang et al. [2024] provide theoretical and empirical evidence that the training objective of LMs encourages linear representations implicitly, however, it is not well understood why some features are represented this way and not others.

Whether linear representation form for more 'common' concepts, or some concepts are just easier to identify is unclear. We hypothesize that concepts that form linear features are inherently linked to their higher frequency in their pretraining data, which we provide positive evidence for in Section 4. We also investigate whether the converse holds: that the measurable presence of linear features tells us something about concept frequency. In Section 5, we fit a regression model to predict the frequency of individual terms (like "The Beatles") in pretraining data, using the measurements of the presence/absence linearity of features (§2). We find that the signal is noisy but allows us to predict the frequencies of heldout relations/terms in approximate ranges; something that is not possible to predict using log probabilities and/or task performance as features alone. Importantly, the regression generalizes beyond the model it was trained on without additional supervision, thus providing the groundwork for analyzing the pretraining corpora of closed-data models with open weights.

2 Background

2.1 Linear Features

Despite how much effort goes in to understanding what features LMs learn, little is understood about why some features form linearly and not others, which we explore here. In his work we study Linear Relational Embeddings (LREs) in LMs from Hernandez et al. [2024], Paccanaro and Hinton [2001] because it allows us to predefine the concepts we want to search for, as well as use a handful of linear features to relate thousands of terms. We choose this method because alternatives, like Sparse Autoencoders (SAEs) [Huben et al., 2023, Gao et al., 2024, Templeton et al., 2024] can be be unwieldy. Finding interpretable latents is not always straightforward, training costs are high, and it is not clear whether we could extract the same features across checkpoints/models. Hernandez et al. [2024] approximate the computation performed by a model to predict the output of common subject-relation-object triplets (e.g., Miles Davis [subject] plays the [relation] Trumpet [object]) as an affine transformation from s, the subject token representation at some middle layer of the model to the hidden state o in the last layer when the model is about to predict the object (e.g., the final hidden state that decodes as trumpet in the above example). This transformation holds for approximately every subject and object in the relation set. This is surprising because, despite the non-linearities within the many layers and token positions separating the utterance of the subject and object, a simple structure within the representation space underlies the model.

2.2 Inferring Training Data from Models

There has been significant interest in understanding the extent to which it is possible to infer the training data from a fully trained network, predominantly in the form of membership inference attacks [Shokri et al., 2017, Carlini et al., 2022], judging memorization of text [Carlini et al., 2023, Oren et al., 2024, Shi et al., 2024], or distribution of data sources [Hayase et al., 2024b, Ateniese et al., 2015, Suri and Evans, 2022]. Our work is related in that we find that a significant imprint of the pretraining data appears in the structure of the LM activation space.

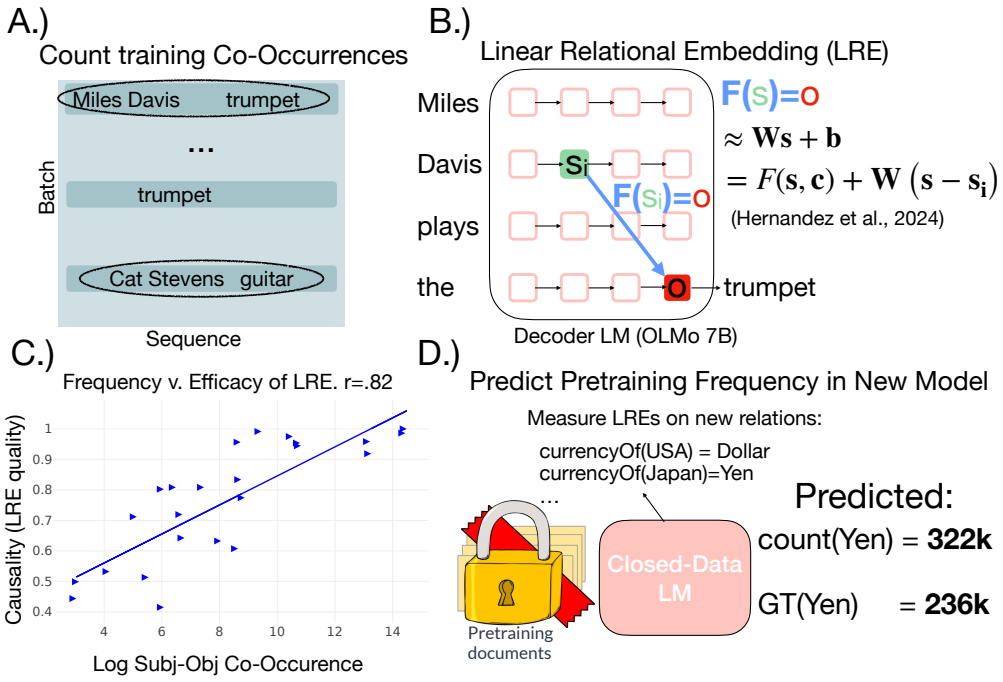


Figure 1: Overview of this work. A.) We count co-occurrences of subjects and objects in s-r-o factual relation triplets throughout pretraining batches. B.) We measure how well these relations are represented within an LM using the Linear Relational Embeddings (LRE) method from Hernandez et al. [2024] across training steps. C.) We establish a strong relationship between average co-occurrence freqeuncy and the tendency to form linear features for relations. D.) We introduce a method using LREs to roughly approximate the frequencies of individual terms in models that for which we do not have access to the training data in order to help infer domains in which a model may have been trained.

3 Methods

At a high level, we explore the connection between pretraining data frequency and the formation of linear features in LMs. Our analysis is based on two main aspects: Counts of terms in the pretraining corpus of LMs, and measurements of how well factual relations are approximated by affine transformations. We use the OLMo (7B and 1B) [Groeneveld et al., 2024] and GPT-J (6B) [Wang and Komatsuzaki, 2021] and their corresponding datasets: Dolma [Soldaini et al., 2024] and the Pile [Gao et al., 2020], respectively.

3.1 Linear Relational Embeddings (LREs)

Hernandez et al. [2024] approximate LREs in an LM as a first-order Taylor Series approximation. Let $F(\mathbf{s}, c)$ be the computation the model does to predict the object representation $\mathbf{o}$ given a subject $\mathbf{s}$ and a context c, this approximation is denoted as $F(\mathbf{s}, c) \approx W\mathbf{s} + b = F(\mathbf{s}_i, c) + W(\mathbf{s} - \mathbf{s}_i)$ where we approximate the relation about some specific subject point $\mathbf{s}_i$. Following the original work, W and b are approximated using the average of n examples from the relation (n=8 here) with $\frac{\partial F}{\partial \mathbf{s}}$ representing the Jacobian of F:

$$W = \mathbb{E}_{\mathbf{s}_i, c_i}\left[\left.\frac{\partial F}{\partial \mathbf{s}}\right|_{(\mathbf{s}_i, c_i)}\right] \quad \text{and} \quad b = \mathbb{E}_{\mathbf{s}_i, c_i}\left[\left. F(\mathbf{s}, c) - \frac{\partial F}{\partial \mathbf{s}}\,\mathbf{s}\right|_{(\mathbf{s}_i, c_i)}\right] \tag{1}$$

Empirically, Hernandez et al. [2024] find that this approximation underestimates the true slope of the transformation, so they scale by scalar hyperparameter β. Unlike the original work, which finds one β per model, we use one beta per relation, as it simplifies analysis. We also find that this does not make a very substantial difference in the results. Another difference in our calculation of LREs is that we do not impose the constraint that the model has to predict the answer correctly to be used as one of the 8 examples used to approximate the Jacobian. We were surprised that using incorrect examples could work as well as it did, but it is very convenient for analyzing early checkpoints (§4), as they can have low performance on some relations. We explore the effect of example choice in Appendix A.

We use the same 25 factual relations from the Relations dataset used in the Hernandez et al. [2024] work (such as capital-city, person-mother). Across these relations there are 10,488 unique subjects and objects. To evaluate the quality of LREs, the authors introduce the **Faithfulness** and **Causality** metrics. The former measures whether the LRE produces the same next token as the original LM, and causality measures the proportion of the time a prediction of an object can be changed to the output of a different example from the relation (e.g., editing the Miles Davis subject representation so that the LM predicts he plays the guitar). For specifics on implementation we refer the readers Hernandez et al. [2024].

3.2 Counting Frequencies Throughout Training

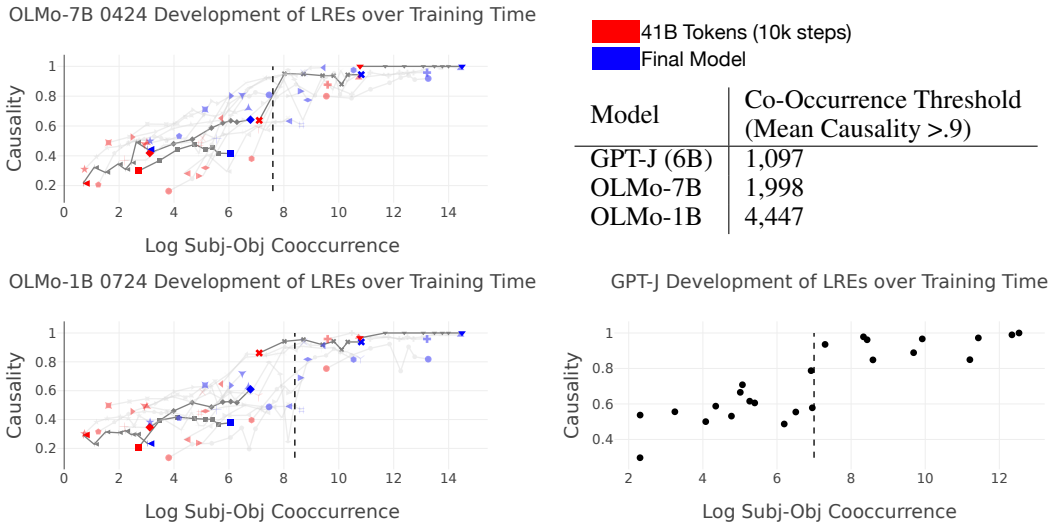
A major question we explore in this work is how term frequencies affect the formation of linear features. We hypothesize that more commonly occurring relations will lead to more robust LREs for those relations. Following Elsahar et al. [2018], we count an occurrence of a relation when a subject and object co-occur together. They show that when a subject and object occur together, the relation triplet is also likely to have occurred in the same span. To rule out the confound that linear features simply form due to extended pretraining time, we test model checkpoints in the OLMo family of models [Groeneveld et al., 2024] to test this. In this section we discuss how we compute these counts.

What's in My Big Data? (WIMBD) Elazar et al. [2024] index many popular pretraining datasets, including Dolma and the Pile, and provide search tools that allows for counting individual terms/co-occurrences within a pretraining document. However, we are interested in counting term frequencies throughout pretraining, and WIMBD would only give us accurate counts for the full training set. Additionally, counting co-occurrences in a document is not exactly accurate to how LMs are pretrained. Since LMs receive gradient updates from batches of fixed lengths which often split documents into multiple sequences, miscounts are likely to occur. Additionally, case sensitivity and tokenization can differ from the LMs we want to study. Because of these reasons we count tokens in batched tokenized text. That being said, we compare WIMBD co-occurrence counts to the Batch Search method (§3.2) for counts in the final checkpoint in Appendix B and find that the counts are extremely close. When per-batch counts are not available, WIMBD offers a good approximation for final checkpoints, which is what we do in the case of GPT-J.

Batch Search In order to accurately count subject-object co-occurrences throughout pretraining, we count how often s-o pairs in the Relations dataset appear in the same relation in pretraining batches. Data counting tools can not typically provide accurate counts for model checkpoints at arbitrary training steps. Thus, we design a tool to efficiently count exact co-occurrences within sequences of tokenized batches. The OLMo family of models [Groeneveld et al., 2024] provide tools for accurately recreating the batches from the Dolma dataset [Soldaini et al., 2024]. We use these to reconstruct the training data batches as it was presented to the model, noting every time one of our 10k terms appears throughout the dataset. We count a co-occurrence as any time two terms appear in the same sequence within a batch (a $(\text{batch-size}, \text{sequence-length})$ array). We search 10k terms in 2T tokens of the Dolma dataset this way. Using our implementation we are able to complete this on a standard compute node of CPUs in about a day. To support future work, we release our code as well as the occurrence positions for every term.

4 Frequency of Subject-Object Co-Occurrences Aligns with Emergence of Linear Features

In this section we explore when LREs begin to appear in training time, and how these are related to pretraining term frequencies. Our main findings are that 1.) average co-occurrence frequency



Model	Co-Occurrence Threshold (Mean Causality >.9)
GPT-J (6B)	1,097
OLMo-7B	1,998
OLMo-1B	4,447

Figure 2: We find that LREs have consistently high causality scores across relations after some average frequency threshold is reached (table top right). In OLMo models, red dots show the model's LRE performance at 41B steps, which is at 10k training steps for OLMo 7B, blue dots show the final checkpoint performance (550k steps in 7B), and gray dots show intermediate checkpoints. Even at very early training steps, if the average subject-object cooc. count is high enough, the models are very likely to already have robust LREs formed in the representation space. We are able to derive loose threshold bounds above which average causality for a relation exceeds .9. For clarity, we highlight five relations to demonstrate the pattern of frequency determining performance.[3]

within a relation strongly correlates with whether a LRE will form. In two similarly sized models, we can draw a similar threshold where LREs are consistently robust. And 2.) that this frequency effect is independent of pretraining time; if the average subject-object co-occurrence within a relation surpasses some threshold it is very likely to have a robust LRE, and this is true even for extremely early pretraining steps. This is also exclusive to *co-occurrences* rather than individual subject or object occurrences. Our results are summarized in Figure 2. In the OLMo models we use checkpoints at [41B, 104B, 209B, 419B, 628B, 838B, 1T, and 2T] pretraining tokens[4]. Co-occurrence frequencies highly correlate with causality (r=.82), the metric used by Hernandez et al. [2024] to judge the linear structure of the intermediate representations for items in a relation. This is notably higher than the correlations with subject frequencies and object frequencies: r=.66, .59 (.66, .59) for OLMo 7B (OLMo 1B), respectively.

We consider a causality score above .9 to be nearly perfectly linear. The top left table shows the frequency threshold where the average causality is above .9 and is shown by dashed black lines on the scatterplots. Regardless of pretraining step, models that surpass this threshold have very high causality scores. Although we can not define strict rules for what this threshold is, it appears to be scale dependent, as both OLMo 7B and GPT-J. Lastly, in all models, ICL accuracy correlates strongly with causality. An interesting direction for future work would be testing whether it is this linear structure that facilitates ICL accuracy; the finding of task vectors produced during ICL [Hendel et al., 2023] may be suggestive of this.

A fundamental question in the interpretability community is why linear structures form. While we do know that the training objective encourages this type of representation [Jiang et al., 2024], our results suggest that why some concepts form this way and not others is strongly controlled by the pretraining frequency.

[3]These are: 'country largest city', 'country currency', 'company hq', 'company CEO', and 'star constellation name' in order from best to worst performing final checkpoints.

[4]In OLMo 7B 0424, this corresponds to 10k, 25k, 50k, 100k, 150k, 200k, 250k, 409k pretraining steps. We report training tokens because the step count differs between 7B and 1B.

	Predicting Object Occs.		Predicting Subject-Object Co-Occs.	
	Train OLMo	Train GPT-J	Train OLMo	Train GPT-J
LRE Features	0.65±0.12	0.49±0.12	0.76±0.12	0.68±0.08
LogProb Features	0.42±0.10	0.41±0.09	0.66±0.09	0.60±0.07
Mean Freq. Baseline	0.31±0.15	0.41±0.17	0.57±0.15	0.67±0.16

Table 1: Overall, we find that fitting a regression on one model's LREs and evaluating on the other provides a meaningful signal compared to fitting using only log probability and task performance, or predicting the average training data frequency. The metric here is proportion of predictions within one order of 10x the ground truth.

5 Linear Features Help Predict Pretraining Corpus Frequencies

In this section, we aim to understand this relationship further by exploring what we can understand about pretraining term frequency from linearity of LM representations. We target the challenging problem of predicting how often a term, or co-occurrence of terms appears in an LM in which we do not have access to the training data. In this setting, we train a regression model to predict frequencies given either LRE features or the log probabilities of the correct answer. The intuition here is that more common answers will have higher probabilities of being the next token for completing the relation, and if the linearity contains an especially meaningful signal about ferquency, using those features will outperform log probs alone.

5.1 Experimental Setup

We train a random forest regression model with 100 estimators to predict the frequency of terms (either the subject-object frequency, or the object frequency alone; e.g., predicting "John Lennon" and "The Beatles" or just "The Beatles") on features from one of two models: either OLMo 7B (final checkpoint) or GPT-J, treating the other as the 'closed' model. We test the hypothesis that LRE features (**faithfulness, causality**) are useful in predicting term frequencies compared to baseline features such as the LM log probabilities or example accuracy. We remove any examples with 0 co-occurrences or less than 10 object occurrences. We evaluate on held out objects and relations to remove any bias of certain relations having tight frequency distributions. To give a better idea of whether the model's predictions are in the right *range* of values, we report the accuracy of predictions within one order of magnitude of the ground truth.

5.2 Results

Our results are presented in Table 1. First, we find that there is a signal in the LRE features that does not exist in the LM features: We are able to fit a much better generalizable model when using LRE features as opposed to the LM probabilities alone. Second, evaluating on the LRE features of a heldout model (scaled by the ratio of total tokens trained between the two models) maintains around the same accuracy, allowing us to predict whether. We find that predicting either the subj-obj. co-occurrences or object frequencies using LREs alone is extremely noisy, and closing these bounds likely require specialized solutions integrating related approaches on dataset inference. Nevertheless, we show that linearity of features within LM representations encode a rich signal representing dataset frequency.

6 Conclusion

We find a connection between linear representations of subject-relation-object factual triplets in LMs and the pretraining frequencies of the subjects and objects in those relations. This finding can guide future interpretabilty work in deciphering whether a linear feature for a given concept will exist in a model, since it seems reasonably certain that frequencies below a certain threshold will not yield LREs (a particular class of linear feature). From there we show that we can use the presence of linear features to predict with some accuracy, the frequency of terms in the pretraining corpus of a closed-data model without supervision. Future work could aim to improve on our bounds of predicted frequencies. Overall, our work presents a meaningful step towards understanding the interactions between pretraining data and internal LM representations.

References

Viraat Aryabumi, Yixuan Su, Raymond Ma, Adrien Morisot, Ivan Zhang, Acyr Locatelli, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. To code, or not to code? exploring impact of code in pre-training. *arXiv preprint arXiv:2408.10914*, 2024.

Giuseppe Ateniese, Luigi V Mancini, Angelo Spognardi, Antonio Villani, Domenico Vitali, and Giovanni Felici. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. *International Journal of Security and Networks*, 10(3):137–150, 2015.

Sid Black, Lee Sharkey, Leo Grinsztajn, Eric Winsor, Dan Braun, Jacob Merizian, Kip Parker, Carlos Ramón Guevara, Beren Millidge, Gabriel Alfour, and Connor Leahy. Interpreting neural networks through the polytope lens, 2022. URL `https://arxiv.org/abs/2211.12312`.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914, 2022. doi: 10.1109/SP46214.2022.9833649.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=TatRHT_1cK`.

David Chanin, Anthony Hunter, and Oana-Maria Camburu. Identifying Linear Relational Concepts in Large Language Models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1524–1535. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.85. URL `https://aclanthology.org/2024.naacl-long.85`.

Yanai Elazar, Akshita Bhagia, Ian Helgi Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Evan Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hannaneh Hajishirzi, Noah A. Smith, and Jesse Dodge. What's in my big data? In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=RvfPnOkPV4`.

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.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. T-REx: A large scale alignment of natural language with knowledge base triples. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL `https://aclanthology.org/L18-1544`.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*, 2024.

Jonathan Hayase, Alisa Liu, Yejin Choi, Sewoong Oh, and Noah A. Smith. Data mixture inference: What do bpe tokenizers reveal about their training data?, 2024a. URL https://arxiv.org/abs/2407.16607.

Jonathan Hayase, Alisa Liu, Yejin Choi, Sewoong Oh, and Noah A. Smith. Data mixture inference: What do bpe tokenizers reveal about their training data?, 2024b. URL https://arxiv.org/abs/2407.16607.

Roee Hendel, Mor Geva, and Amir Globerson. In-Context Learning Creates Task Vectors. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9318–9333. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.624. URL https://aclanthology.org/2023.findings-emnlp.624.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. Linearity of Relation Decoding in Transformer Language Models. 2024. URL https://openreview.net/forum?id=w7LU2s14kE.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse Autoencoders Find Highly Interpretable Features in Language Models. 2023. URL https://openreview.net/forum?id=F76bwRSLeK.

Yibo Jiang, Goutham Rajendran, Pradeep Kumar Ravikumar, Bryon Aragam, and Victor Veitch. On the Origins of Linear Representations in Large Language Models. 2024. URL https://openreview.net/forum?id=otuTw4Mghk&referrer=%5Bthe%20profile%20of%20Goutham%20Rajendran%5D(%2Fprofile%3Fid%3D~Goutham_Rajendran1).

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, et al. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3245–3276, 2024.

Yingwei Ma, Yue Liu, Yue Yu, Yuanliang Zhang, Yu Jiang, Changjian Wang, and Shanshan Li. At which training stage does code data help LLMs reasoning? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KIPJKST4gw.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Language models implement simple Word2Vec-style vector arithmetic. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5030–5047, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.281. URL https://aclanthology.org/2024.naacl-long.281.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

Yonatan Oren, Nicole Meister, Niladri S. Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. Proving test set contamination in black-box language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KS8mIvetg2.

Alberto Paccanaro and Geoffrey E Hinton. Learning Hierarchical Structures with Linear Relational Embedding. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001. URL https://papers.nips.cc/paper_files/paper/2001/hash/814a9c18f5abff398787c9cfcbf3d80c-Abstract.html.

Kiho Park, Yo Joong Choe, and Victor Veitch. The Linear Representation Hypothesis and the Geometry of Large Language Models. 2024. URL https://openreview.net/forum?id=UGpGkLzwpP.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854, 2022.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2023.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zWqr3MQuNs.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. Dolma: An open corpus of three trillion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159*, 2024.

Nishant Subramani, Nivedita Suresh, and Matthew Peters. Extracting Latent Steering Vectors from Pretrained Language Models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 566–581. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-acl.48. URL https://aclanthology.org/2022.findings-acl.48.

Anshuman Suri and David Evans. Formalizing and estimating distribution inference risks. *Proceedings on Privacy Enhancing Technologies*, 2022, 2022.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function Vectors in Large Language Models. 2024. URL https://openreview.net/forum?id=AwyxtyMwaG&noteId=6Qv7kx00La.

Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax, May 2021.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy S Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. Doremi: Optimizing data mixtures speeds up language model pretraining. *Advances in Neural Information Processing Systems*, 36, 2024.

A Effect of Training on Incorrect Examples

In Hernandez et al. [2024], examples are filtered to ones in which the LM gets correct, assuming that an LRE will only exist once a model has attained the knowledge to answer the relation accuracy (e.g., knowing many country capitals). We find that the choice of examples for fitting LREs is not

Figure 3: Comparison in performance of LREs when fitting with examples the LM gets correct or incorrect. We found this did not make a significant difference in LRE quality.

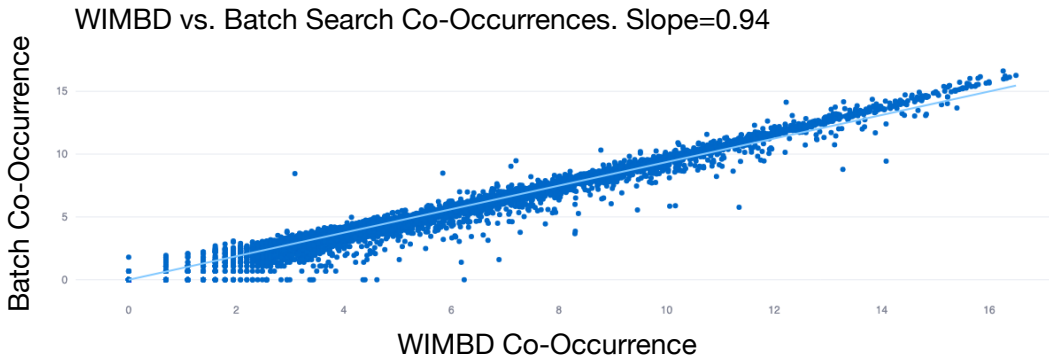


Figure 4: Comparison between WIMBD and Batch Search subject-object co-occurrences

entirely dependent on the model 'knowing' that relation perfectly (i.e., attains high accuracy). This is convenient for our study, where we test early checkpoint models, that do not necessarily have all of the information that they will have seen later in training. In Figure 3, we show faithfulness on relations where the LRE was fit with all, half, or zero correct examples. We omit data for which the model did not get enough incorrect examples.

B Batch Search Counts Compared to WIMBD

In Figure 4, we find that What's in My Big Data [Elazar et al., 2024] match very well to batch search co-occurrences, however, WIMBD tends to overpredict co-occurrences (slope less than 1), due to the sequence length being shorter than many documents, as discussed in the main paper.