ON LINEAR REPRESENTATIONS AND PRETRAINING DATA FREQUENCY IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

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's effect on downstream task behavior, we investigate its relationship to LM representations. Previous work has discovered that, in language models, some concepts are encoded as "linear representations", but what factors cause these representations to form (or not)? 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 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 (s-r-o) fact triplets 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 discover that a linear representation consistently (but not exclusively) forms when the subjects and objects within a relation co-occur at least 1-2k times, regardless of when these occurrences happen during pretraining. In the OLMo 1B model, consistent linearity only occurs after 4.4k occurrences, suggesting a connection to scale. Finally, we train a regression model on measurements of linear representation quality that can predict how often a term was seen in pretraining. We show such model achieves low error even for a different model and pretraining dataset, providing a new unsupervised method for exploring possible data sources of closed-source models. We conclude that the presence or absence of linear representations in LMs contains signal about their pretraining corpora that may provide new avenues for controlling and improving model behavior. We release our code to support future work[1]

1 INTRODUCTION

Understanding how the content of pretraining data affects language model (LM) behaviors and performance is an active area of research (Ma et al., 2024; Xie et al., 2024; Aryabumi et al., 2024; Longpre et al., 2024; Antoniades et al., 2024; Seshadri et al., 2024; Razeghi et al., 2023; Wang et al., 2024). For instance, it has been shown that for specific tasks, models perform better on instances containing higher frequency terms than lower frequency ones (Razeghi et al., 2022; Mallen et al., 2023a). The ways in which frequency affects the internal representations of LMs to cause this difference in performance remain unclear. We connect dataset statistics to recent work in interpretability, which focuses on the emergence of simple linear representations of factual relations in LMs. Our findings demonstrate a strong correlation between these features and the frequency of terms in the pretraining corpus.

Linear representations in LMs have become central to interpretability research in recent years (Ravfogel et al., 2020; Elazar et al., 2021; Elhage et al., 2021; Slobodkin et al., 2023; Olah et al., 2020; Park et al., 2024; Jiang et al., 2024; Black et al., 2022; Chanin et al., 2024). Linear representations are essentially linear approximations (linear transforms, directions in space) that are simple to understand, and strongly approximate the complex non-linear transformations that networks are

[1]Anonymized

implementing. These representations are crucial because they allow us to localize much of the behavior and capabilities of LMs to specific directions in activation space. This means that certain behaviors can be activated or modulated by intervening on these directions with linear projections at inference time, a process also known as steering (Todd et al., 2024; Subramani et al., 2022; Hendel et al., 2023; Rimsky et al., 2023).

Recent work by Hernandez et al. (2024) and Chanin et al. (2024) highlight how the linearity of different types of relations varies greatly depending on the specific relationships being depicted. For example, over 80% of "country largest city" relations can be approximated by a single linear transformation on the contextual embedding of the country, but less than 30% of "star in constellation" can be. Their methods for identifying representations with linear structure do not offer an explanation for this. Such findings complicate the understanding of the Linear Representation Hypothesis, which proposes that LMs will represent features linearly (Park et al., 2024). While Jiang et al. (2024) provide both theoretical and empirical evidence that the training objectives of LMs implicitly encourage linear representations, it remains unclear why some features are represented this way while others are not. This open question is a central focus of our investigation.

Whether linear representations for "common" concepts are actually more prevalent in models or simply easier to identify (using current methods) than those for less common concepts remains unclear. We hypothesize that factual relations exhibiting linear representations are correlated with higher mention frequencies in the pretraining data (as has been shown with static embeddings, see Ethayarajh et al., 2019), which we confirm in Section 4. Our results also indicate that this can occur at any point in pretraining, as long as a certain average frequency is reached across subject-object pairs in a relation. In order to count the appearance of terms in data corpora throughout training, we develop an efficient tool for counting tokens in tokenized batches of text, which we release to support future work in this area. We also explore whether the presence of linear representations can provide insights into relation frequency. In Section 5, we fit a regression model to predict the frequency of individual terms (such as "The Beatles") in pretraining data, based on metrics measuring the presence of a linear feature for some relation. For example, how well a linear transformation approximates the internal computation of the "lead singer of" relation mapping "John Lennon" to "The Beatles" can tell us about the frequency of those terms in the pretraining corpus.

Our findings indicate that the predictive signal, although approximate, is much stronger than that encoded in log probabilities and task accuracies alone, allowing us to estimate the frequencies of held-out relations and terms within approximate ranges. Importantly, this regression model generalizes beyond the specific LM it was trained on without additional supervision. This provides a valuable foundation for analyzing the pretraining corpora of closed-data models with open weights.

To summarize, in this paper we show that:

1. The development of linear representations for factual recall relations in LMs is related to frequency as well as model size.
2. Linear representations form at predictable frequency thresholds during training, regardless of when this frequency threshold is met for the nouns in the relation. The formation of these also correlates strongly with recall accuracy.
3. Measuring the extent to which a relation is represented linearly in a model allows us to predict the approximate frequencies of individual terms in the pretraining corpus of that model, even when we do not have access to the model's training data.
4. We release a tool for accurately and efficiently searching through tokenized text to support future research on training data.

2 Background

2.1 Linear Representations

Representing information in distributed vector spaces has a long history in language processing, where geometric properties of these spaces were used to encode semantic information (Salton et al., 1975; Paccanaro & Hinton, 2001). When and why linear structure emerges without explicit bias to do so has been of considerable interest since the era of static word embeddings. Work on skipgram

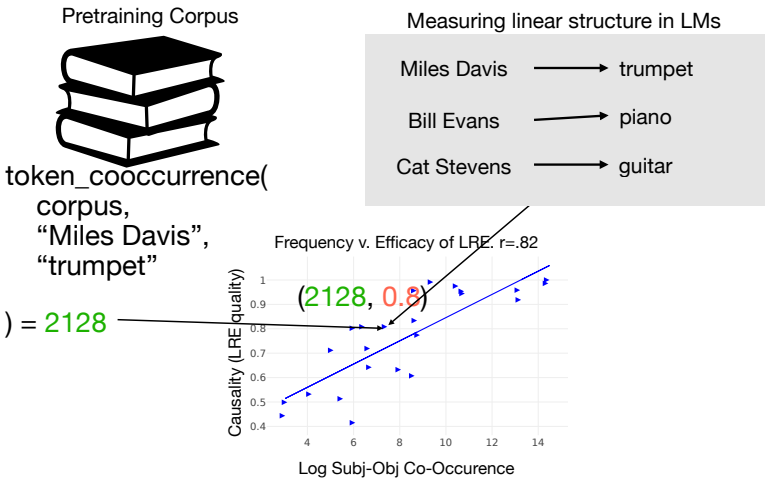


Figure 1: Overview of this work. Given a dataset of subject-relation-object factual relation triplets, we count subject-object co-occurrences throughout pretraining batches. We then measure how well the corresponding relations are represented within an LM across pretraining steps, using the Linear Relational Embeddings (LRE) method from Hernandez et al. (2024). We establish a strong relationship between average co-occurrence frequency and a model's tendency to form linear representations for relations. From this, we show that we can predict frequencies in the pretraining corpus

models (Mikolov, 2013) found that vector space models of language learn regularities which allow performing vector arithmetic between word embeddings to calculate semantic relationships (e.g., France-Paris+Spain=Madrid) (Mikolov et al., 2013; Pennington et al., 2014). This property was subject to much debate, as it was not clear why word analogies would appear for some relations and not others (Köper et al., 2015; Karpinska et al., 2018; Gladkova et al., 2016). Followup work showed that linguistic regularities form in static embeddings for relations under specific dataset frequency constraints for relevant terms (Ethayarajh et al., 2019), but does not clearly relate to how modern LMs learn. More recently, there has been renewed interest in the presence of similar linear structure in models with contextual embeddings like transformer language models (Park et al., 2024; Jiang et al., 2024; Merullo et al., 2024). As a result, there are many ways to find and test for linear representations in modern LMs, though the relationship to pretraing data is not addressed (Huben et al., 2023; Gao et al., 2024; Templeton et al., 2024; Rimsky et al., 2023; Todd et al., 2024; Hendel et al., 2023; Hernandez et al., 2024; Chanin et al., 2024). Many of these share similarities in how they compute and test for the linear representations, typically through counterfactuals. We focus on a particular class of linear representations called Linear Relational Embeddings (LREs) (Paccanaro & Hinton, 2001).

Linear Relational Embeddings (LREs) Hernandez et al. (2024) use a particular class of linear representation called a Linear Relational Embedding (Paccanaro & Hinton, 2001) to approximate the computation performed by a model to predict the objects that complete common `subject-relation-object` triplets as an affine transformation. This transform is calculated from hidden state $\mathbf{s}$, the subject token representation at some middle layer of the model, to $\mathbf{o}$, the hidden state at the last token position and layer of the model (i.e., the final hidden state that decodes a token in an autoregressive transformer). For example, given the input sequence "Miles Davis (`subject`) plays the (`relation`)", the goal is to approximate the computation of the object "trumpet", assuming the model predicts the object correctly. It was found that this transformation holds for nearly every subject and object in the relation set (such as "Cat Stevens plays the guitar") for some relations. This is surprising because, despite the non-linearities within the many layers and token positions separating $\mathbf{s}$ and $\mathbf{o}$, a simple structure within the representation space well approximates the model's prediction process for a number of factual relations. In this work we study LREs under the same definition and experimental setup, because it allows us to predefine the concepts we want to search for (e.g., factual relations), as well as use a handful of representations to relate thousands of terms in the dataset by learning linear representations on a per-relation level.

Hernandez et al. calculate LREs to approximate an LM's computation as a first-order Taylor Series approximation. Let $F(\mathbf{s}, c) = \mathbf{o}$ be the forward pass through a model that produces object representation $\mathbf{o}$ given subject representation $\mathbf{s}$ and a few-shot context c, this computation is approximated 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 a specific subject $\mathbf{s}_i$. Hernandez et al. propose to compute W and b using the average of n examples from the relation (n=8 here) with $\frac{\partial F}{\partial \mathbf{s}}$ representing the Jacobian Matrix of F:

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

In practice, LREs are estimated using hidden states from LMs during the processing of the test example in a few-shot setup. For a relation like "instrument-played-by–musician", the model may see four examples (in the form "X plays the Y") and on the fifth example, when predicting e.g., "trumpet" from "Miles Davis plays the", the subject representation $\mathbf{s}$ and object representation $\mathbf{o}$ are extracted.

2.2 Inferring Training Data from Models

There has been significant interest recently in understanding the extent to which it is possible to infer the training data of a fully trained neural network, including LMs, predominantly by performing 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 inferring the distribution of data sources (Hayase et al., 2024b; Ateniese et al., 2015; Suri & Evans, 2022). Our work is related in that we find hints of the pretraining data distribution in model itself, but focus on how linear structure in the representations relates. Carlini et al. (2024); Finlayson et al. (2024) do not focus on extracting dataset information, but on inferring information architectural information about a black-box model behind an API.

3 Methods

Our analysis is twofold: 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 model v1.7 (0424 7B and 0724 1B) (Groeneveld et al., 2024) and GPT-J (6B) (Wang & Komatsuzaki, 2021) and their corresponding datasets: Dolma (Soldaini et al., 2024) and the Pile (Gao et al., 2020), respectively. To understand how these features form over training time, we test 8 model checkpoints throughout training in the OLMo family of models (Groeneveld et al., 2024).

3.1 Linear Relational Embeddings (LREs)

The original Relations dataset includes factual, commonsense, gender bias, and linguistic relations, but we reduce this set to the 25 factual relations used by Hernandez et al. (2024)[2]. These are relations such as capital-city and person-mother (full list in Appendix A). The reason for this is due to the way we count occurrences of a relation in training data not being accurate for non-factual relations (see §3.2). Across these relations there are 10,488 unique subjects and objects. Following Hernandez et al. (2024), we fit an LRE for each relation on 8 examples from that relation, each with a 5 shot prompt. We use the approach from this work as described in Section 2.1.

Fitting LREs Hernandez et al. (2024) find that Equation 1 underestimates the optimal slope of the linear transformation, so they scale each relation's W by a scalar hyperparameter β. Unlike the original work, which finds one β per model, we use one β per relation, as this avoids disadvantaging specific relations. 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 Matrix. Interestingly, using examples that models predict incorrectly to fit Equation 1 works as well as using only correct examples. We opt to use this variant as it allows us to compare different checkpoints and models (§4) with linear transformations trained on the same 8

[2]For the analysis, we drop "Landmark on Continent" because 74% of the answers are Antarctica, making it uninteresting for studying true relational knowledge.

examples, despite the fact that the models make different predictions on these instances. We explore the effect of example choice in Appendix A.

Metrics To evaluate the quality of LREs, (Hernandez et al., 2024) introduce two metrics that measure the quality of the learned transformations. **Faithfulness** measures whether the transformation learned by the LRE produces the same object token prediction as the original LM. **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*, instead of the *trumpet*). For specifics on implementation we refer the reader to Hernandez et al. (2024). We consider an LRE to be high 'quality' when it scores highly on these metrics, as this measures when an LRE works across subject-object pairs within the relation. In general, we prefer to use causality in our analysis, as faithfulness can be high when LMs predict the same token very often (like in early checkpoints).

3.2 Counting Frequencies Throughout Training

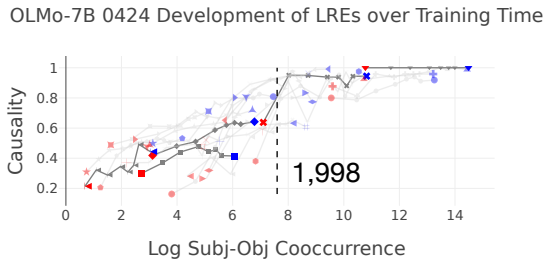
A key question we explore is how term frequencies affect the formation of linear representations. We hypothesize that more commonly occurring relations will lead to higher quality LREs for those relations. Following Elsahar et al. (2018); Elazar et al. (2022), we count an occurrence of a relation when a subject and object co-occur together. While term co-occurrence is used as a proxy for the frequency of the entire triplet mentioned in text, Elsahar et al. (2018) show that this approximation is quite accurate. We now discuss how to compute these co-occurrence 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 and co-occurrences within documents. However, this only gives us counts for the full dataset. Since we are interested in counting term frequencies throughout pretraining, we count these within training batches of OLMo instead. 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. We compare WIMBD co-occurrence counts to the Batch Search method (described below) for the final checkpoint of OLMo in Appendix C, and find that the counts are extremely close.

Batch Search 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. This also gives us the advantage of counting in a way that is highly accurate to how LMs are trained; since LMs are trained on batches of fixed lengths which often split documents into multiple sequences, miscounts may occur unless using tokenized sequences. Using this method, we note every time one of our 10k terms appears throughout a dataset used to pretrain an LM. 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 the approximately 2T tokens of the Dolma dataset (Soldaini et al., 2024) this way. Using our implementation we are able to complete this on 900 CPUs in about a day. To support future work, we release our code as Cython bindings that integrate out of the box with existing libraries.

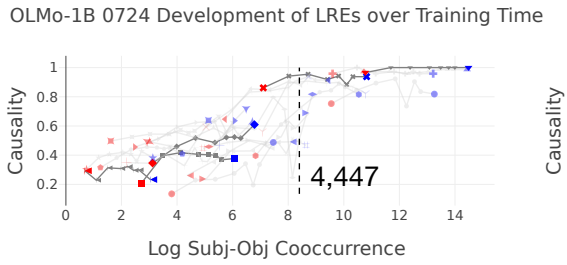
4 Frequency of Subject-Object Co-Occurrences Aligns with Emergence of Linear Representations

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 within a relation strongly correlates with whether an LRE will form; 2) the frequency effect is independent of the pretraining stage; if the average subject-object co-occurrence within a relation surpasses some threshold it is very likely to have a high-quality LRE, even for early pretraining steps. This finding is exclusive to *co-occurrences* rather than individual subject or object occurrences. In addition to confirming dataset frequencies strongly align with LREs forming, we aim to confirm that this relationship is strongest with subject-object **co-occurrences** rather than just mentions of relevant subjects or objects.



41B Tokens (10k steps)
Final Model

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



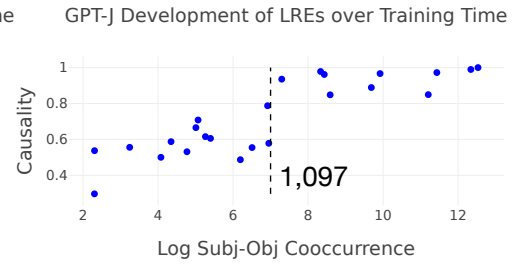


Figure 2: We find that LREs have consistently high causality scores across relations after some average frequency threshold is reached (table, top right). In OLMo models, red dots show the model's LRE performance at 41B tokens, and blue dots show the final checkpoint performance (550k steps in 7B). Gray 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.[5]

4.1 Setup

Using the factual recall relations from the Hernandez et al. (2024) dataset, we use the Batch Search method (§3.2) to count subject and object co-occurrences withing sequences in Dolma (Soldaini et al., 2024) used to train the OLMo 1B (v. 0724) and 7B (v. 0424) models (Groeneveld et al., 2024). The OLMo family of models provide tools for accurately recreating the batches from Dolma, which allow us to reconstruct the data the way the model was trained. We also use GPT-J (Wang & Komatsuzaki, 2021) and the Pile (Gao et al., 2020) as its training data, but since we do not have access to accurate batches used to train it, we use WIMBD (Elazar et al., 2024) to count s-o counts in the entire data. We fit LREs on each relation and model separately. Hyperparameter sweeps are in Appendix B. OLMo also releases intermediate checkpoints, which we use to track development over pretraining time. We use checkpoints that have seen {41B, 104B, 209B, 419B, 628B, 838B, 1T, and 2T} tokens[3]. We use the Pearson coefficient for measuring correlation unless other specified.

4.2 Results

Our results are summarized in Figure 2. We report training tokens because the step count differs between 7B and 1B. Co-occurrence frequencies highly correlate with causality (r=.82). This is notably higher than the correlations with subject frequencies: r=.66, and object frequencies: .59 for both OLMo 7B and OLMo 1B, respectively.

We consider a causality score above .9 to be nearly perfectly linear. The table in Figure 2 shows the co-occurrence counts above which 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 draw conclusions from only three models, it is possible that scale also affects this threshold: OLMo 7B and GPT-J (6B params) require far less exposure than OLMo 1B.

[3]In OLMo 7B 0424, this corresponds to 10k, 25k, 50k, 100k, 150k, 200k, 250k, 409k pretraining steps

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

4.3 Relationship to Accuracy

Increased frequency (or a proxy for it) is shown to lead to better factual recall in LMs (Chang et al., 2024; Mallen et al., 2023b). However, it remains unknown whether high accuracy entails the existence of a linear relationship. Such a finding would inform when we expect an LM to achieve high accuracy on a task. We find that the correlation between causality and subject-object frequency is higher than with 5-shot accuracy (.82 v.s. .74 in OLMo 7B), though both are clearly high. In addition, there are a few examples of high accuracy relations that do not form single consistent LREs. These relations are typically low frequency, such as star constellation name, which has 84% 5-shot accuracy but only 44% causality (OLMo 7B), with subjects and objects only co-occurring about 21 times on average across the full dataset. In general, few-shot accuracy closely tracks causality, consistent with arguments that in-context learning allows models to identify linear mappings between input-output pairs (Hendel et al., 2023; Garg et al., 2022). We find that causality increases first in some cases, like "food from country" having a causality of 65% but a 5-shot accuracy of only 42%. This gap is consistently closed through training. In the final model, causality and 5-shot accuracy is within 11% on average. We report the relationship between every relation, zero-shot, and few-shot accuracy for OLMo models across training in Appendix E.

A fundamental question in the interpretability community is why linear structures form. While previous work has claimed that the training objective encourages this type of representation (Jiang et al., 2024), our results suggest that the reason why some concepts form a linear representation while others do not, is strongly related to the pretraining frequency.

5 Linear Representations 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's training data from the representations alone. Such prediction model can be useful, if it generalizes, when applied to other models whose weights are open, but the data is closed. For instance, such predictive model could tell us whether a model was trained on specific domains (e.g., Java code) by measuring the presence of relevant LREs. First, we show that LRE features encode information about frequency that is not present using probabilities alone. Then, we show how a regression fit on one model generalizes to the features extracted from another without any information about the new model's counts.

5.1 Experimental Setup

We train a random forest regression model with 100 decision tree 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") from one of two sets of features. Our baseline set of features is based on likelihood of recalling a fact. Given some few-shot context from the relations dataset ("John Lennon is a lead singer of") we extract the log probability of the correct answer, as well as the average accuracy on this prompt across 5 trials. The intuition is that models will be more confident about highly frequent terms. The other set of features include the first, as well as faithfulness and causality measurement.

We use Faithfulness and Causality as defined in Hernandez et al. (2024) as well as too other metrics: **Faith Prob.**, which is the log probability of the correct answer as produced by an LRE, and **Hard Causality**, which is the same as the "soft" variant, but only counts the proportion of times the causality edit produces the target answer as the number one prediction. We use every example from the Relations for which there are more than 1 object occurrence or subject-object co-occurrence. We drop the "Landmark in Continent" relation because it is too imbalanced.[6] We do not provide an explicit signal for which relation an example comes from, but due to the bias of subjects/objects having similar frequencies within a relation, we train multiple models and evaluate on held out

[6]Most answers are "Antarctica" which was artificially inflating our results.

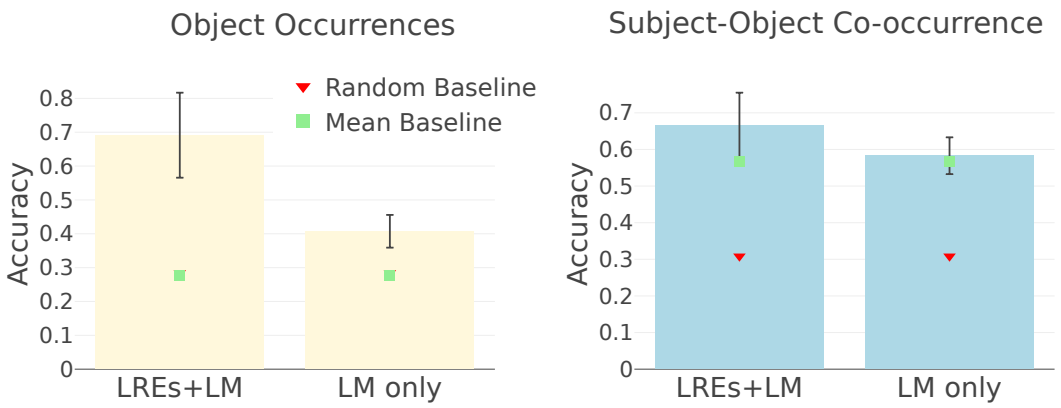


Figure 3: Within-Magnitude accuracy (aka the proportion of predictions within one order of magnitude of ground truth) for models predicting object and subject-object co-occurrences in heldout relations. Using LRE features outperforms LM only features by about 30%. We find that it is much easier to predict object frequencies; the subj-obj. prediction models with LRE features only marginally outperform baseline performance.

relations and average performance. In all settings, the held out set objects are guaranteed to not have been in the training set.

5.2 LRE Metrics encode Fine-grained Frequency Information

We fit a regression to the Relations dataset (Hernandez et al., 2024) using OLMo-7B LRE features and log probabilities. We fit 24 models such that each relation is held out once per random seed across 4 seeds. Because of the difficulty of predicting the exact number of occurrences, we report accuracy within one order of magnitude of the ground truth. This measures whether the predicted value is within a reasonable range of the actual value. Results are shown in Figure 3. We find that language modeling features do not provide any meaningful signal towards predicting object or subject-object frequencies, and are only marginally above the baseline of predicting the average or random frequencies from the training data. On object frequency predictions, we find that LRE features encode a strong signal allowing for accurate predictions about 70% of the time. Mean absolute error of the predictions (in natural log space) for LRE features (LM-only features) are 2.1, (4.2) and 1.9, (2.3) on object prediction and subject-object predictions tasks, respectively. We find that subject-object co-occurrence frequency is likely too difficult to predict given the signals that we have here, as our predictions are higher than, but within one standard deviation of

Feature Importance: How important are LRE features for predicting the frequency of an item? We perform feature permutation tests to see how much each feature (LRE features and log probs) contributes to the final answer. First, we check to see which features used to fit the regression are correlated, as if they are, then perturbing one will leave the signal present in another. In Appendix D, we show that only faithfulness and faith probability are strongly correlated, so for this test only, we train models with a single PCA component representing 89% of the variance of those two features. We find that hard causality is by far the most important feature for generalization performance, causing a difference of about 15% accuracy, followed by faithfulness measures with 5% accuracy, providing evidence that the LRE features are encoding an important signal.

5.3 Generalization to a new LM

In this section, we test the ability to generalize the regression fit on one LM to another for which we do not have access to pretraining term counts, without requiring further supervision. We keep the objective the same and apply the regression model, fit for example on OLMo ("Train OLMo" setting), to features extracted from GPT-J, using ground truth counts from The Pile (or vice versa, i.e., the "Train GPT-J" setting).

	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: Within-Magnitude accuracy for different settings of train and test models. 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.

Predicting Object Frequency in GPT-J, Regression fit on OLMo					
Relation	Subject	Object	Prediction	Ground Truth	Error
landmark-in-country	Menangle Park	Australia	2,986,989	3,582,602	1.2x
country-language	Brazil	Portuguese	845,406	561,005	1x
star-constellation name	Arcturus	Boötes	974,550	2,817	346x
person-mother	Prince William	Princess Diana	5,826	27,094	4.6x
person-mother	Prince Harry	Princess Diana	131	27,094	207x

Table 2: Examples of a regression fit on OLMo LRE metrics and evaluated on GPT-J on held-out relations, demonstrating common error patterns: 1. Predictions are better for relations that are closer to those found in fitting the relation (country related relations), 2. Some relations, like star-constellation perform very poorly, possibly due to low frequency, 3. the regression model can sensitive to the choice of subject (e.g., William vs. Harry), telling us the choice of data to measure LREs for is important for predictions.

We again train a random forest regression model 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 across different models, with the hope that this could be applied to dataset inference methods in the future, where access to the ground truth pretraining data counts is limited or unavailable.

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 log probability 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 when fit on exact counts from OLMo, allowing us to predict occurrences without access to the GPT-J pretraining data. We find that predicting either the subject-object co-occurrences or object frequencies using LREs alone is barely better than the baseline. This task is much more difficult than predicting the frequency of the object alone, but our model may just also be unable to account for outliers in the data, which is tightly clusterd around the mean (thus giving the high mean baseline performance of between approx. 60-70%). Nevertheless, we show that linearity of features within LM representations encode a rich signal representing dataset frequency.

5.4 Error Analysis

In Table 2 we show example predictions from a regression model fit on OLMo evaluated on heldout relations with LREs measured on GPT-J. We find that some relations transfer more easily than others, with the star constellation name transferring especially poorly. In general, the regression transfers well, without performance deteriorating much (about 5% within-magnitude accuracy), suggesting LREs are encoding information in a consistent way across models. We also find that the regression makes use of the full prediction range, producing values in the millions (see Table 2) or in the tens: The same regression shown in the table also predicts 59 occurrences for "Caroline Bright" (Will Smith's mother) where the ground truth is 48.

6 DISCUSSION

Connection to Factual Recall Work in interpretability has focused largely around linear representations in recent years, and our work aims to address the open question of the conditions in which they form. We find that coherent linear representations form when the relevant terms (in this case subject-object co-occurrences) appear in pretraining at a consistent enough rate. Analogously, Chang et al. (2024) show that repeated exposure encourages higher retention of facts. It isn't clear whether accuracy on factual recall entails that a linear representation exists (at least for some cases) from our work, however future research could study this connection more closely.

Linear Representations in LMs The difficulty of disentangling the formation of linear representations from increases in relation accuracy, especially in the few-shot case, is interesting. Across 24 relations, only the "star constellation name" and "product by company" relations have few shot accuracies that far exceed their causality scores (and both are low frequency). Thus, it is still an open question how LMs are able to recall these tasks. While there is not a single LRE that can solve the relation, it is not necessarily true that the model is preferring a non-linear solution, as multiple incomplete LREs could account for different parts of the data. The fact that few-shot accuracy and causality seem so closely linked is consistent with findings that ICL involves locating the right task (Min et al., 2022) and applying a 'function' to map input examples to outputs (Hendel et al., 2023; Todd et al., 2024). That frequency controls this ability is perhaps unsurprising, as frequency also controls this linear structure emerging in static embeddings (Ethayarajh et al., 2019). Jiang et al. (2024) prove a strong frequency-based condition (based on matched log-odds between subjects and objects) and an implicit bias of gradient descent (when the frequency condition is not met) encourage linearity in LLMs; our work empirically shows conditions where linear representations tend to form in more realistic settings. If LMs are 'only' solving factual recall or performing ICL through linear structures, it is surprising how well this works at scale, but the simplicity also provides a promising way to understand LMs and ICL in general. An interesting avenue for future work would be to understand if and when LMs use a method that is not well approximated linearly to solve these types of tasks, as non-linear representations, as recent work has shown non-linearity can be preferred for some tasks in recurrent networks (Csordás et al., 2024).

Future Work in Predicting Dataset Frequency The ability to predict the contents of pretraining data is an important area for investigating memorization, contamination, and privacy of information used to train models. In our approach, we show it's possible to extract signal without supervision by first fitting on an opens source model. The fact that there is some transferable signal between models is indicative that this relationship between pretraining frequency and linearity is consistent between models. Mosbach et al. (2024) discuss the role of interpretability on the broader field of NLP. Without interpretability work on the nature of representations in LMs, we would not know of this implicit dataset signal, and we argue that interpretability can generate useful insights more broadly as well. Extensions on this work could include more information to tighten the prediction bounds on frequency, such as extracting additional features from the tokenizer (Hayase et al., 2024a). A likely candidate task that could integrate our method is for predicting whether a certain domain of data (e.g., code) was included in pretraining, since extensive exposure would lead to LREs forming. Regardless, we hope this work encourages future research in other ways properties of pretraining data affect LM representations for both improving and better understanding these models.

7 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 representation for a given concept will exist in a model, since we observe that frequencies below a certain threshold for a given model will not yield LREs (a particular class of linear representation). From there we show that we can use the presence of linear representations 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.

8 LIMITATIONS

While our approach thoroughly tracks exposure to individual terms and formation of LRE features across pretraining, we can not draw causal claims about how exposure affects individual representations, due to the cost of counterfactual pretraining. We try to address this by showing the frequency of individual terms can be predicted with some accuracy from measurements of LRE presence. We motivate this approach as a possible way to detect the training data of closed-data LMs, however, we are not able to make any guarantees on its efficacy in settings not shown here, and would caution drawing strong conclusions without additional information. Furthermore, we find that our method is relatively worse at predicting subject-object co-occurrences than object occurrences, and our method fails to account for the harder task. Future work could expand on this tool by incorporating it with other data inference methods for greater confidence. We also do not discuss the role of the presentation of facts on the formation of LRE features, but following Elsahar et al. (2018) and the strength of the relationship we find, we speculate this has minimal impact. Note that the BatchSearch tool we release tracks the exact position index of the searched terms, thus facilitating future work on questions about templates/presentation of information.

REFERENCES

Antonis Antoniades, Xinyi Wang, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, and William Yang Wang. Generalization v.s. memorization: Tracing language models' capabilities back to pretraining data. *ArXiv*, abs/2407.14985, 2024.

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, 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)*, pp. 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`.

Nicholas Carlini, Daniel Paleka, Krishnamurthy Dj Dvijotham, Thomas Steinke, Jonathan Hayase, A Feder Cooper, Katherine Lee, Matthew Jagielski, Milad Nasr, Arthur Conmy, et al. Stealing part of a production language model. In *Forty-first International Conference on Machine Learning*, 2024.

Hoyeon Chang, Jinho Park, Seonghyeon Ye, Sohee Yang, Youngkyung Seo, Du-Seong Chang, and Minjoon Seo. How Do Large Language Models Acquire Factual Knowledge During Pretraining? 2024. URL `http://arxiv.org/abs/2406.11813`.

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

Róbert Csordás, Christopher Potts, Christopher D. Manning, and Atticus Geiger. Recurrent neural networks learn to store and generate sequences using non-linear representations, 2024. URL https://arxiv.org/abs/2408.10920.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 03 2021. URL https://doi.org/10.1162/tacl_a_00359.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Amir Feder, Abhilasha Ravichander, Marius Mosbach, Yonatan Belinkov, Hinrich Schütze, and Yoav Goldberg. Measuring causal effects of data statistics on language model'sfactual'predictions. *arXiv preprint arXiv:2207.14251*, 2022.

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 (eds.), *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.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. Towards Understanding Linear Word Analogies. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3253–3262. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1315. URL https://aclanthology.org/P19-1315.

Matthew Finlayson, Swabha Swayamdipta, and Xiang Ren. Logits of api-protected llms leak proprietary information. *arXiv preprint arXiv:2403.09539*, 2024.

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.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pp. 8–15, 2016.

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 (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 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).

Marzena Karpinska, Bofang Li, Anna Rogers, and Aleksandr Drozd. Subcharacter information in japanese embeddings: When is it worth it? In *Proceedings of the Workshop on the Relevance of Linguistic Structure in Neural Architectures for NLP*, pp. 28–37, 2018.

Maximilian Köper, Christian Scheible, and Sabine Schulte im Walde. Multilingual reliability and "semantic" structure of continuous word spaces. In *Proceedings of the 11th international conference on computational semantics*, pp. 40–45, 2015.

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)*, pp. 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 (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546.

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 (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822. Association for Computational Linguistics, 2023b. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Language models implement simple Word2Vec-style vector arithmetic. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 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.

Tomas Mikolov. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064, 2022.

Marius Mosbach, Vagrant Gautam, Tomás Vergara-Browne, Dietrich Klakow, and Mor Geva. From insights to actions: The impact of interpretability and analysis research on nlp. *arXiv preprint arXiv:2406.12618*, 2024.

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.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162. URL https://aclanthology.org/D14-1162.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.647.

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*, pp. 840–854, 2022.

Yasaman Razeghi, Hamish Ivison, Sameer Singh, and Yanai Elazar. Backtracking mathematical reasoning of language models to the pretraining data. In *NeurIPS Workshop on Attributing Model Behavior at Scale*, 2023. URL https://openreview.net/forum?id=EKvqw9k3lC.

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.

G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, November 1975. ISSN 0001-0782. doi: 10.1145/361219.361220. URL https://doi.org/10.1145/361219.361220.

Preethi Seshadri, Sameer Singh, and Yanai Elazar. The bias amplification paradox in text-to-image generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6367–6384, 2024.

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)*, pp. 3–18. IEEE, 2017.

Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. The curious case of hallucinatory (un) answerability: Finding truths in the hidden states of over-confident large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3607–3625, 2023.

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 (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 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.

Xinyi Wang, Alfonso Amayuelas, Kexun Zhang, Liangming Pan, Wenhu Chen, and William Yang Wang. Understanding reasoning ability of language models from the perspective of reasoning paths aggregation. In *Forty-first International Conference on Machine Learning*, 2024.

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.

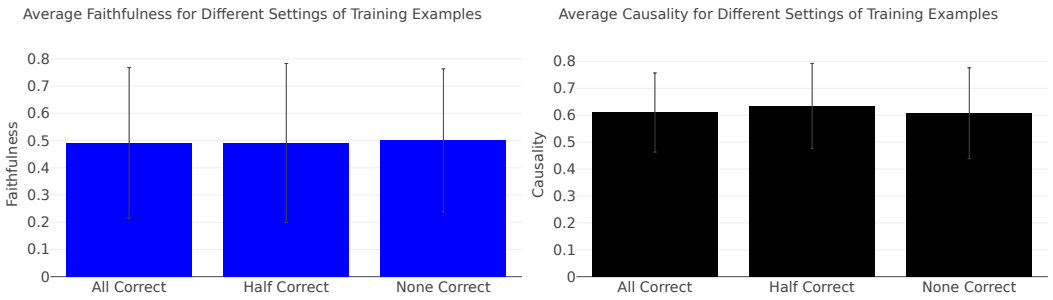


Figure 4: Average Causality and Faithfulness results across relations depending on if the LRE was fit with correct or incorrect samples. We find no notable difference in the choice of examples.

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 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 5, 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. Averages across relations for which we have enough data are shown in Figure 4, which shows that there is not a considerable difference in the choice of LRE samples to train with.

B LRE Hyperparameter Tuning

There are three hyperparameters for fitting LREs: **layer** at which to edit the subject, the **beta** term used to scale the LRE weight matrix, and the **rank** of the pseuoinverse matrix used to make edits for measuring causality. Beta is exclusive to measuring faithfulness and rank is exclusive to causality. We test the same ranges for each as in Hernandez et al. (2024): [0, 5] beta and [0, full_rank] in for causality at varying intervals. Those intervals are every 2 from [0,100], every 5 from [100,200], every 25 from [200, 500], every 50 from [500, 1000], every 250 from [1000, hidden_size]. We perform the hyperparameter sweeps across faithfulness and causality, but we choose the layer to edit based on the causality score. In cases where this is not the same layer as what faithfulness would decide, we use the layer causality chooses, as it would not make sense to train one LRE for each metric. We refer the reader to Hernandez et al. (2024) for more details on the interactions between hyperparameters and the choice of layer. The results of our sweeps on OLMo 7B across layers in Figures 6 and 7 and across beta and rank choices in Figures 8 and 9.

C Batch Search Counts Compared to WIMBD

In Figure 10, 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.

D Feature Correlations and Importances

Our feature importance test is shown in Figure 12. This permutation test was done on the heldout data to show which features contribute the most to generalization performance. We use PCA to reduce the faithfulness features to one feature for the purposes of this test. Correlations are shown in Figure 11

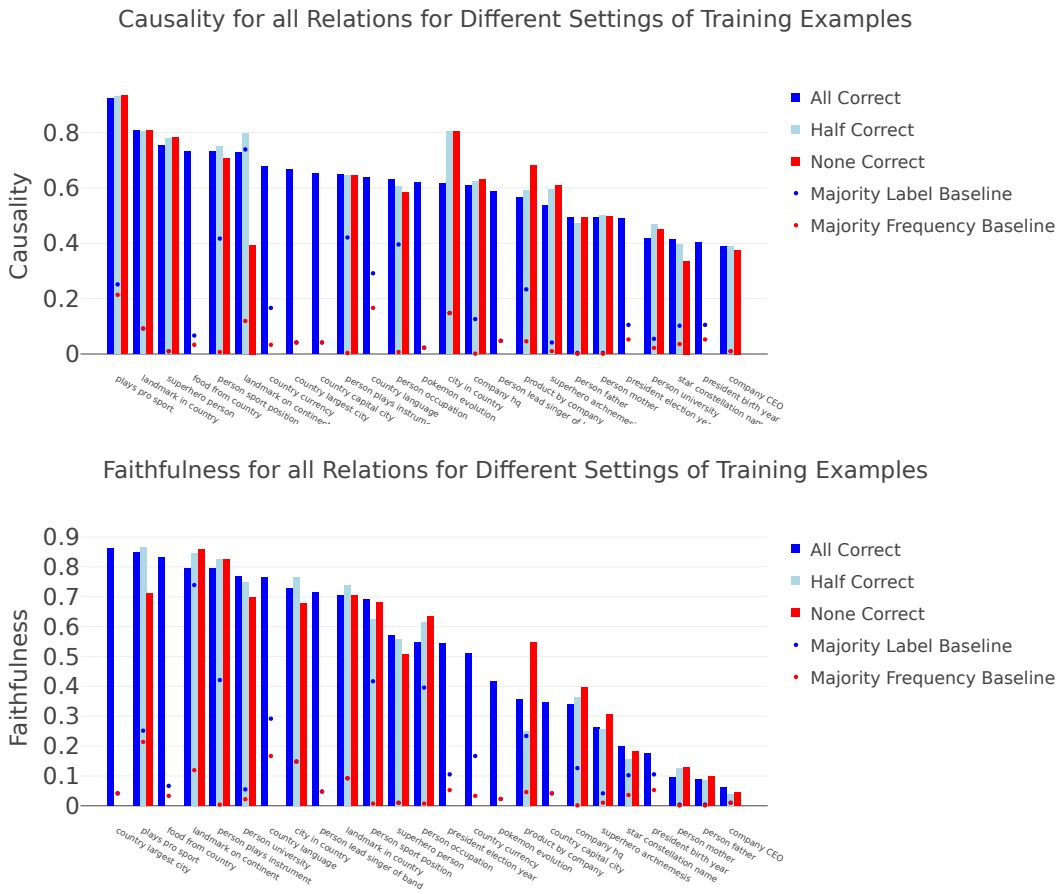


Figure 5: Causality and Faithfulness results for each relation depending on if the LRE was fit with correct or incorrect samples. Note that relations with only one bar do not have zeros in the other categories. It means that there was not enough data that the model (OLMo 7B) got wrong to have enough examples to fit.

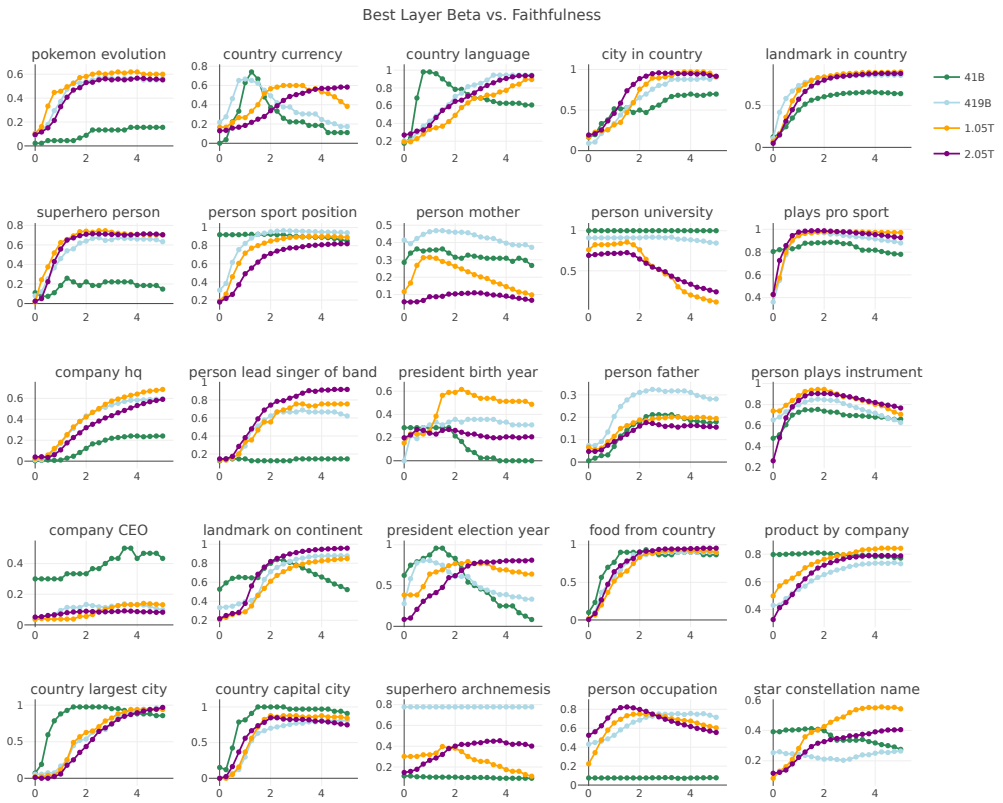


Figure 6: OLMo 0424 7B per layer faithfulness scores as a function of the choice of layer at which to fit the LRE. Note we do not use these results to choose the layer for the LRE, instead preferring the results from the causality sweep.

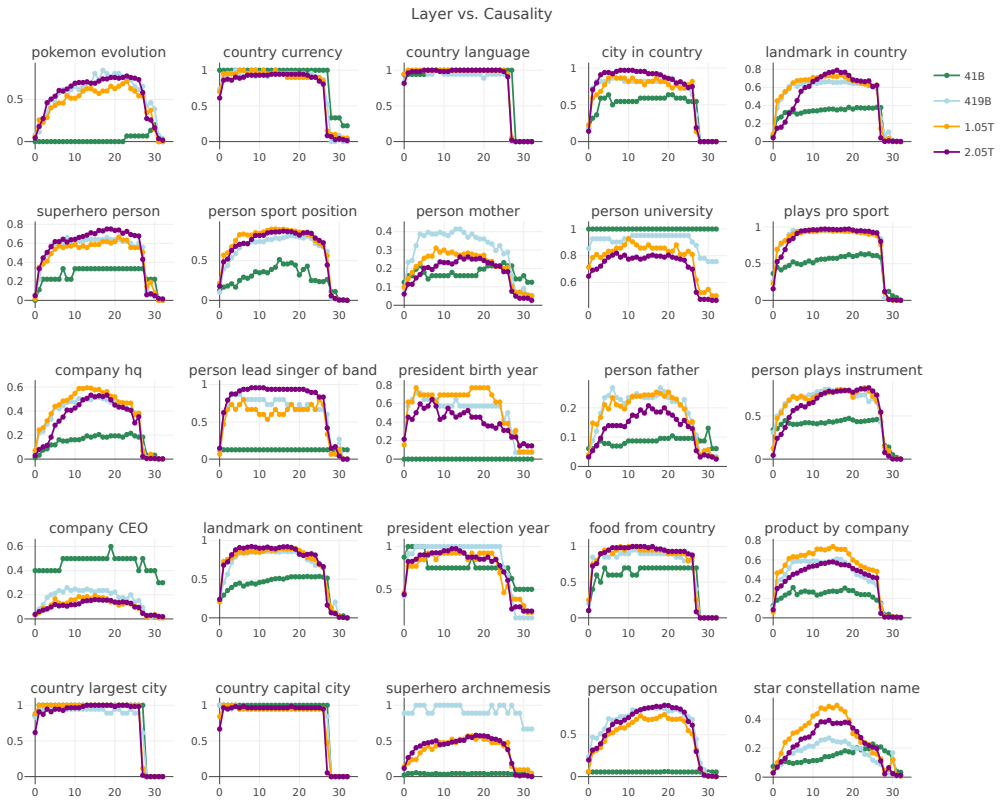


Figure 7: OLMo 0424 7B per layer causality scores as a function of the choice of layer at which to fit the LRE.

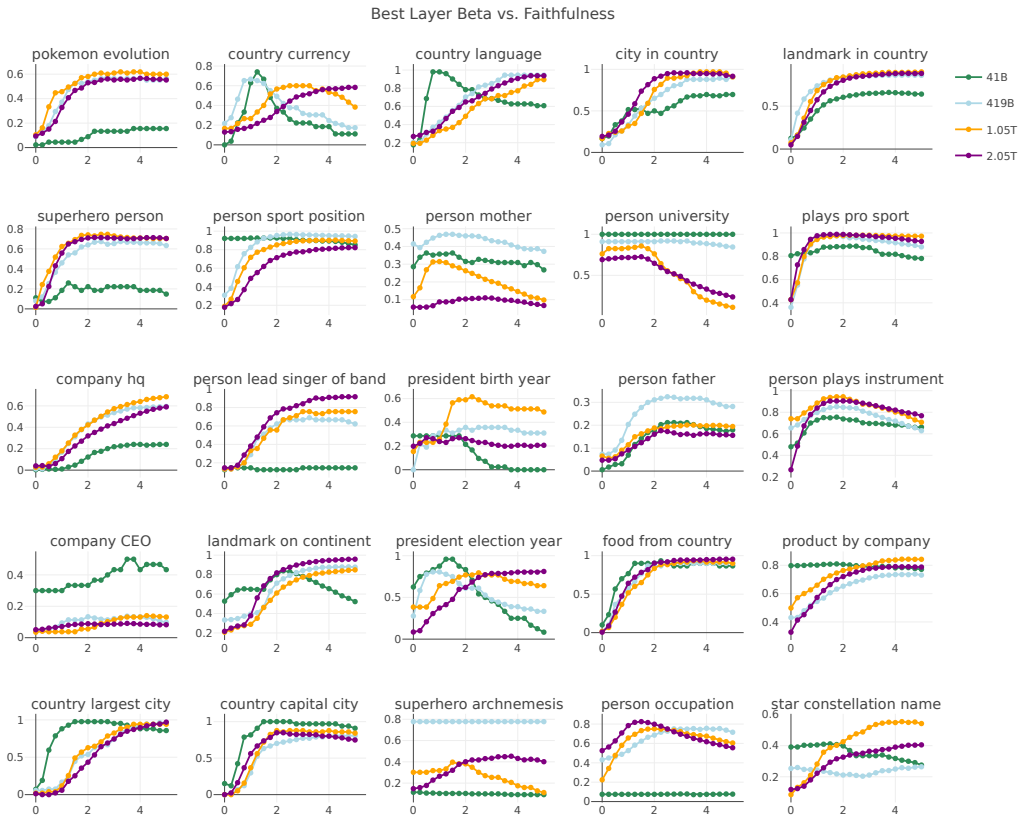


Figure 8: OLMo 0424 7B LRE Beta hyperparameter sweep at highest performing layer.

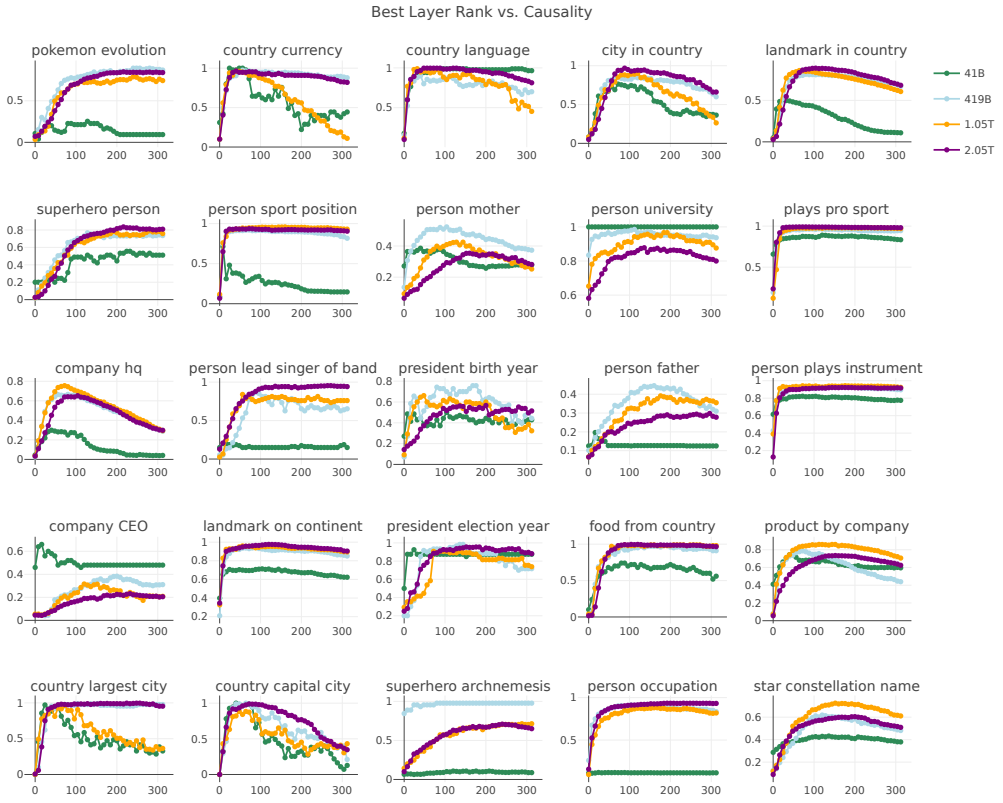


Figure 9: OLMo 0424 7B LRE Rank hyperparameter sweep at highest performing layer.

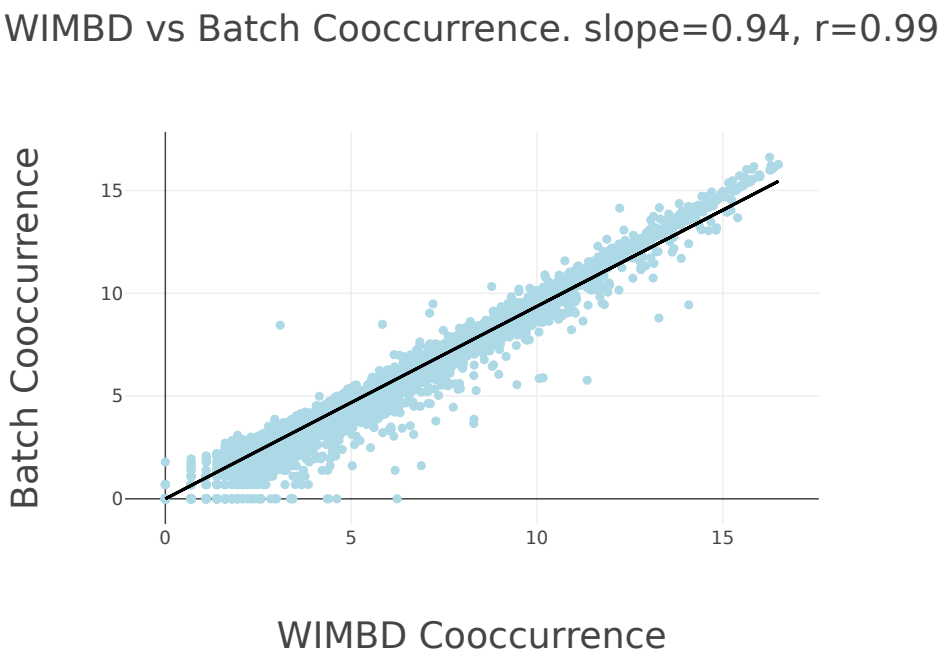


Figure 10: Comparison between WIMBD and Batch Search subject-object co-occurrences

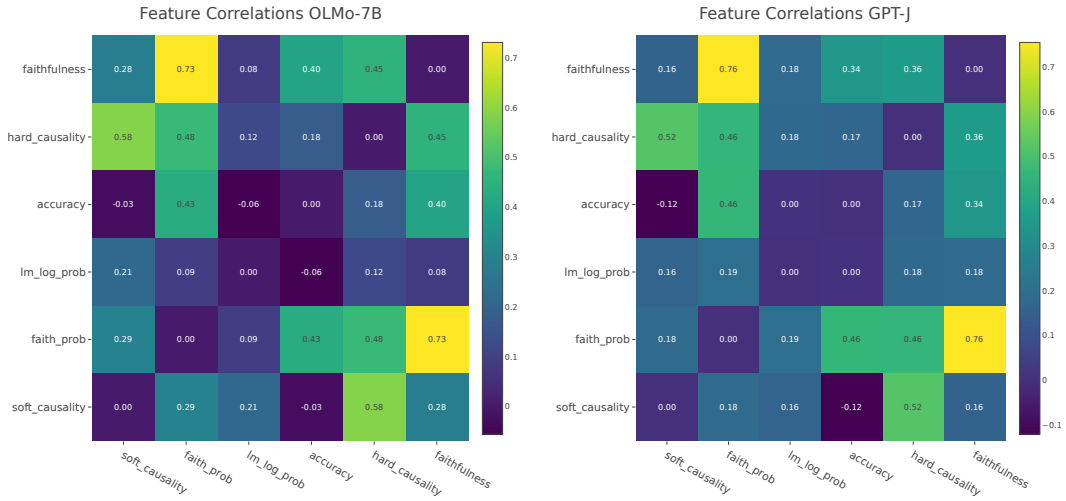


Figure 11: Correlations between each feature in our regression analysis. Because of the high correlation between faithfulness metrics, we use a single dimensional PCA to attain one feature that captures 89% of the variance of both for the purposes of doing feature importance tests. Note that we zero out the diagonal (which has values of 1) for readability.

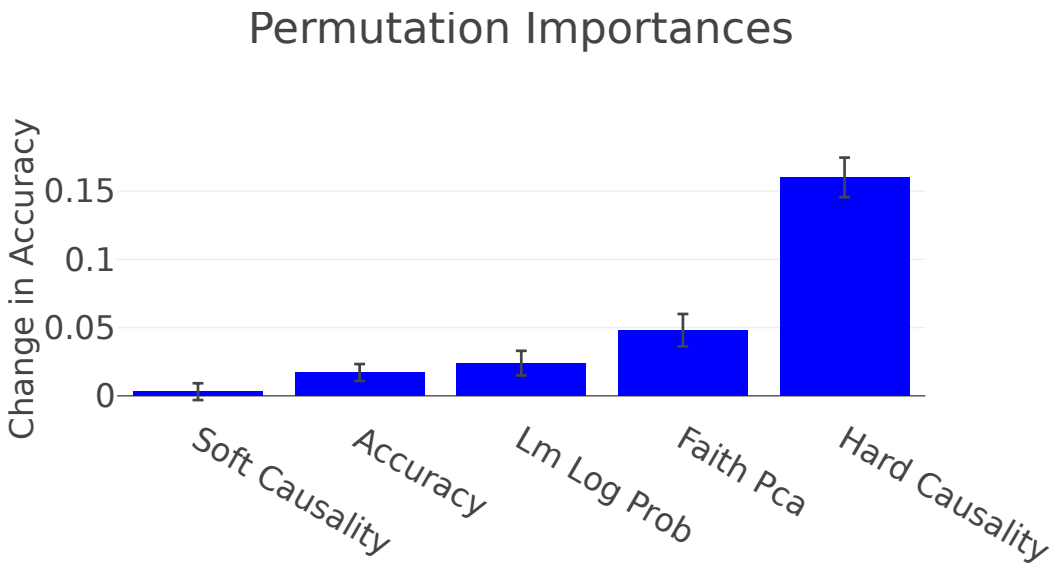


Figure 12: Hard causality is by far the most important feature for generalizing to new relations when predicting Object frequencies, causing a change in about 15% accuracy.

E Relationship Between Causality and Accuracy

In this section we provide more detail on the relationship between the formation of linear representations and accuracy on in-context learning tasks. Although the two are very highly correlated, we argue that accuracy and LRE formation are somewhat independent.

We show this relationship across training For OLMo 1B in Figure 13 and 7B in Figure 14.

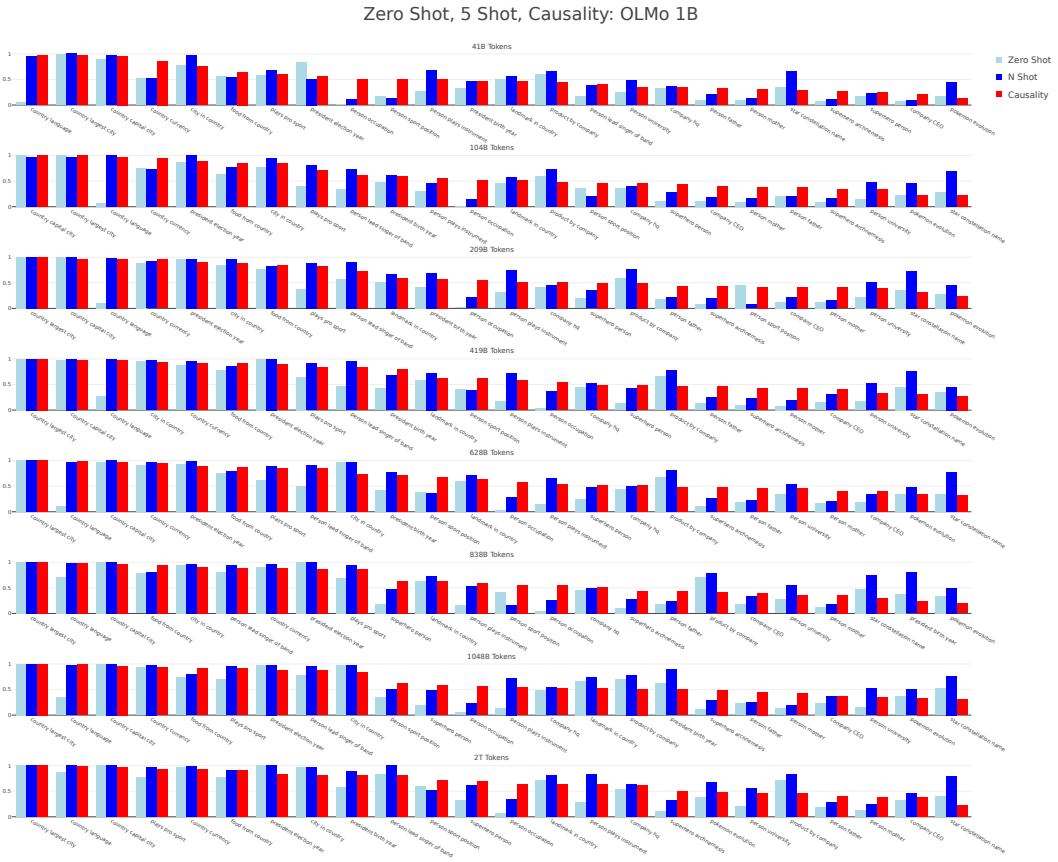


Figure 13: Zero shot, 5-shot accuracies against causality for each relation across training time in OLMo-1B

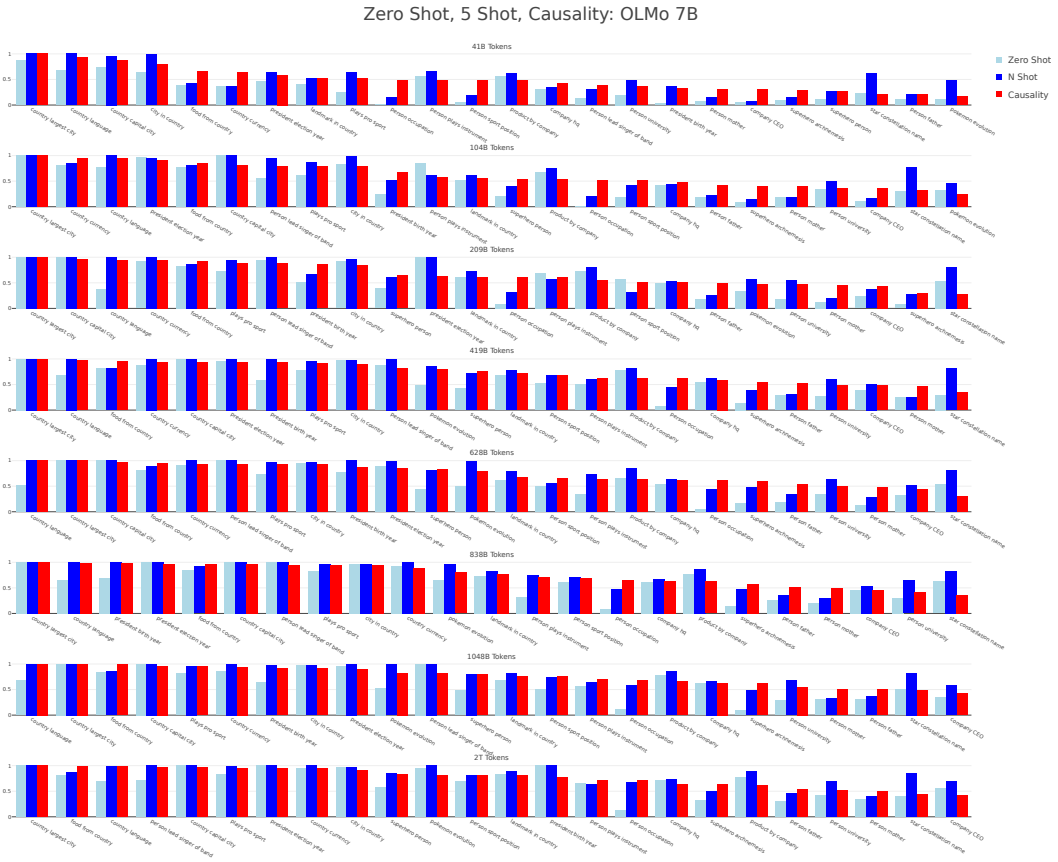


Figure 14: Zero shot, 5-shot accuracies against causality for each relation across training time in OLMo-7B