Linear Relational Decoding of Morphology in Language Models

Eric Xia
Brown University
eric_xia@brown.edu

Jugal Kalita
University of Colorado Colorado Springs
jkalita@uccs.edu

Abstract

A two-part affine approximation has been found to be a good approximation for transformer computations over certain subject-object relations. Adapting the Bigger Analogy Test Set, we show that the linear transformation $W\mathbf{s}$, where $\mathbf{s}$ is a middle layer representation of a subject token and W is derived from model derivatives, is also able to accurately reproduce final object states for many relations. This linear technique is able to achieve 90% faithfulness on morphological relations, and we show similar findings multi-lingually and across models. Our findings indicate that some conceptual relationships in language models, such as morphology, are readily interpretable from latent space, and are sparsely encoded by cross-layer linear transformations.

1 Introduction

Large language models display impressive capabilities for factual recall, which commonly involve relations between entities (Brown et al. 2020). Recent work has shown that affine transformations on subject representations can faithfully approximate model outputs for certain subject-object relations (Hernandez et al. 2023). Identifying transformer approximators is an important area of study, with applications in model training and editing.

The contributions of this paper are twofold. We reproduce and extend existing research. Specifically, we apply affine Linear Relational Embedding (LRE) method to novel diverse relational categories, including derivational and inflectional morphology, encyclopedic knowledge, and lexical semantics. By doing so, we confirm the efficacy of the affine technique. We show that relational approximation can be applied to adapted analogical datasets, and demonstrate relational approximation for a broad range of linguistic phenomena.

At the same time, this work makes a key contribution to research on relational representation in model latents. We show that for different relations, additive and multiplicative mechanisms play complementary roles in affine approximation. We find that an analogue to the original linear relational embedding developed by Paccanaro and Hinton (2001), using a single multiplicative operator, is effective within specific relations. In particular, linear approximation within contexts relating morphological forms reaches near-equivalent level of faithfulness to the affine LRE. We test faithfulness over eight different languages and find that this equivalence holds cross-typologically.

2 Related Work

Much work in machine learning has focused on learning concept representations with hierarchical structure. Relations between representations in concept spaces have been modeled successfully by both linear multiplicative and additive operations.

Multiplicative. Paccanaro and Hinton (2001) introduced the concept of the linear relational embedding for learning relational knowledge from triples (a, R, b). Concepts such as a and b are represented as n-length vectors, while relations such as R are represented as $n \times n$ matrices, akin to distributional models of compositional semantics proposed by Coecke et al. (2010).

Additive. Mikolov et al. (2013) used linear operations in word vector space derived from context-predictive neural nets, demonstrating a correspondence between directional binary relations (e.g. male-female, country-capital, verb tense) and the addition of embedding vectors. Later work found inflectional relations were better captured than derivational ones, and encyclopedic relations better than lexicographic ones. (Gladkova et al. 2016; Vylomova et al. 2016).
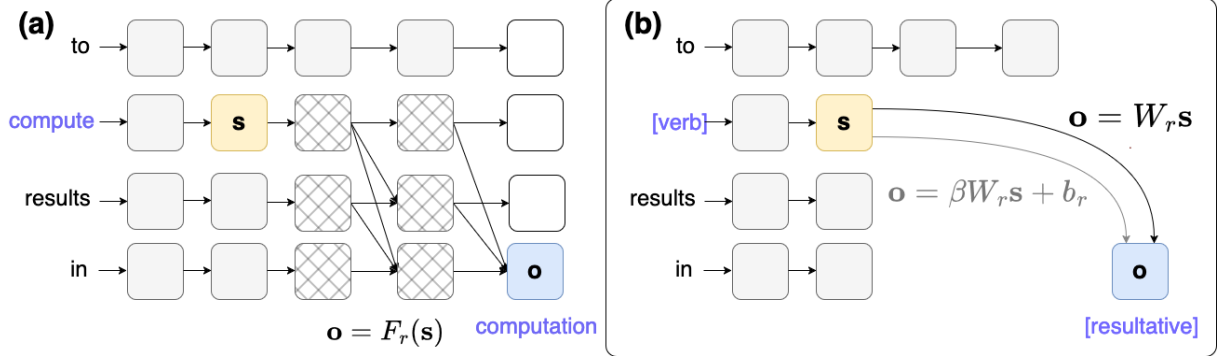
Figure 1: As seen in **(a)**, transformers resolve subject-object relations in a highly nonlinear fashion. As seen in **(b)**, both affine and linear approximators of the subject-object map $F_r(\mathbf{s})$ are demonstrated to be highly effective over relations such as morphology.

3 Background

3.1 Transformer Computation

In auto-regressive transformer language models, input text is converted to a sequence of tokens $t_1 \ldots t_n$, which are subsequently embedded as $x_1 \ldots x_n \in \mathbb{R}^d$ by an embedding matrix. They are then passed through L transformer layers, each composed of a self-attention layer and an multi-layer perceptron (MLP) layer. In GPT-J, the representation x_i^l of the ith token at layer l is obtained as:

$$x_i^l = x_i^{l-1} + a_i^l + m_i^l$$

where a_i^l is multi-headed Key-Value Query attention over x^{l-1}(Vaswani et al. 2017) and m_i^l is the ith output of the lth MLP sublayer. In this case, the output of the l-th MLP sublayer for the i-th representation depends on x_i^{l-1}, rather than $a_i^l + x_i^{l-1}$ (Wang and Komatsuzaki 2021). The final prediction t_{n+1} is then determined by the final hidden state x_n passed through a decoder head D, which consists of a linear layer and softmax to a token vocabulary: $t_{n+1} = \underset{t}{\operatorname{argmax}} D(x_n^L)_t$.

3.2 Relational Representation

Throughout this paper, we will focus on the subject-object relationship as expressed through a single fixed context. Following prior work (Meng et al. 2022b; Geva et al. 2023) that the last token state of a subject in middle layers are strongly casual on predictions (e.g. "Needle" in "The Space Needle"), we are interested in utilizing the gradient between the last token position of the subject s at an intermediate layer, and the object prediction state o.

4 Approach

4.1 Problem Statement

We first consider what it means for a context to express a relation. Many statements can be expressed in terms of a subject, relation, and object (*s,r,o*). For instance, the statement *Miles Davis plays the trumpet* expresses a relation F_r, connecting the subject s (*Miles Davis*) to the object o (*trumpet*): $F_r(s) = o$. We can then relate new subjects to objects: $F_r(Jimi\ Hendrix) = guitar$ and $F_r(Elton\ John) = piano$. F_r is an inductive mechanism, from which statements relating subject and object pairs can be obtained. We are interested in how a language model implements this abstraction. **Affine LRE.** As a starting point, we look at the affine linear relational embedding (LRE) method developed by Hernandez et al. (2023). The authors are able to approximate the transformer's relational function $F_r(s)$ with the affine approximator LRE(s), such that when applied to novel subjects, they reproduce LM object predictions.

The object retrieval function from a subject with a fixed relational context, $o = F_r(s)$, is modeled to be a first-order Taylor approximation of F_r about a number of subjects $s_1 \ldots s_n$. For $i = 1 \ldots n$:

$$\begin{aligned} F_r(s) &\approx F_r(s_i) + W_r(s - s_i) \\ &= F(s_i) + W_r s - W_r s_i \\ &= W_r s + b_r, \end{aligned}$$
$$\text{where } b_r = F_r(s_i) - W_r s_i$$

In a relational context, a model may rely heavily on a singular subject state to produce the object state. Accordingly, the Jacobian matrix of derivatives between vector representations of the subject and object is hypothesized to serve as W_r. For a

226

(a)

```
[animal - youth]

The offspring of a dog is referred to as a puppy
The offspring of a sheep is referred to as a lamb
...
The offspring of a bear is referred to as a cub
```

(b)
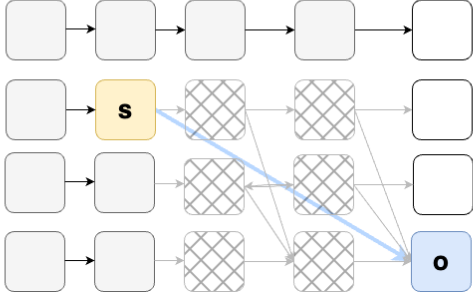
```
The offspring of a deer is
referred to as a fawn
```
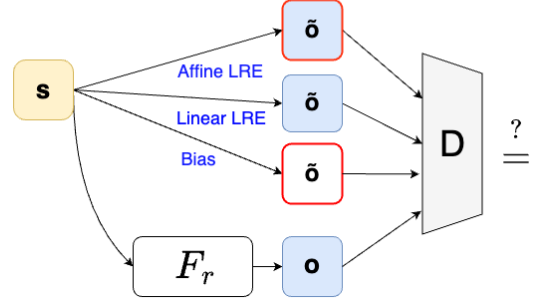
Figure 2: In **(a)**, we first assemble approximators from trained model Jacobians between middle-layer subject states and the final-layer object state. Then, in **(b)**, we evaluate approximated tokens against transformer computations.

fixed relation, they calculate the mean Jacobian and bias between n enriched subject states $\mathbf{s}_1 \dots \mathbf{s}_n$ and outputs $F_r(\mathbf{s}_1) \dots F_r(\mathbf{s}_n)$:

$$W_r = \mathbb{E}_{\mathbf{s}_i} \left[\left. \frac{\partial F_r}{\partial \mathbf{s}} \right|_{\mathbf{s}_i} \right] \qquad (d \times d \text{ matrix})$$

$$b_r = \mathbb{E}_{\mathbf{s}_i} \left[F_r(\mathbf{s}) - \left. \frac{\partial F_r}{\partial \mathbf{s}} \, \mathbf{s} \right|_{\mathbf{s}_i} \right] \qquad (d \text{ vector})$$

This yields a relational approximator capable of transforming a jth layer subject state $x_s^j = \mathbf{s}$ [1] into the final object hidden state $x_o^L = \mathbf{o}$ [2]:

$$\mathbf{o} \approx \text{LRE}(\mathbf{s}) = \beta W_r \mathbf{s} + b_r$$

For instance, $\mathbf{s}$ may be the hidden state of the 7th layer at the subject token, and $\mathbf{o}$ the hidden state of the 26th (last) layer at the object token, e.g. the next-token prediction state.

True Linear Encoding. The affine LRE diverges from the linear relational embedding introduced by Hinton (1986), in introducing a bias b_r and scaling term β. While linearity is assumed in Hernandez et al. (2023) by calculating W_r and b_r from

$\mathbb{E}_{s_i}$ over $i = 1 \dots n$, using a Taylor approximation makes a weaker assumption, simply that the subject-object relation F_r is differentiable. With linearity, we would expect the following:

$$\mathbf{o} \approx F_r'(s_i)\mathbf{s}$$
$$= W_r \mathbf{s}$$

In this case, the linear approximation over $\mathbf{s}_1 \dots \mathbf{s}_n$ within the same relation would be the mean Jacobian. If this approximation generalizes to unseen objects, it would indicate the presence of a linear subject-object map.

4.2 Introducing New Relations

Analogy is traditionally seen as a special case of role-based relational reasoning (Sternberg and Rifkin 1979, Gentner 1983, Holyoak 2012), motivating the adaptation of analogical pairs to a relational setting. We choose to adapt the Bigger Analogy Test Set (BATS), originally introduced to explore linguistic regularities in word embeddings by Gladkova et al. (2016). The dataset comprises forty different categories, each with fifty pairs of words sharing a common relation. The categories span inflectional morphology, derivational morphology, encyclopedic knowledge, and lexical semantics.

4.3 Utilizing ICL

As seen in Figure 2, we adapt the relational pairs in BATS by introducing prompts which are compatible with each instance of the analogy.

---

[1]Following Meng et al. (2022a), both this paper and the affine LRE focus primarily on middle-layer states.

[2]Note the introduction of a β scaling parameter. The authors claim the affine LRE is limited by layer normalization: the $\mathbf{s}$ representation is normalized before contributing to $\mathbf{o}$, and $\mathbf{o}$ is normalized before token prediction by the LM head, resulting in a mismatch in the scale of the output approximation. We find that this conclusion is supported by empirical evidence from linear projections.
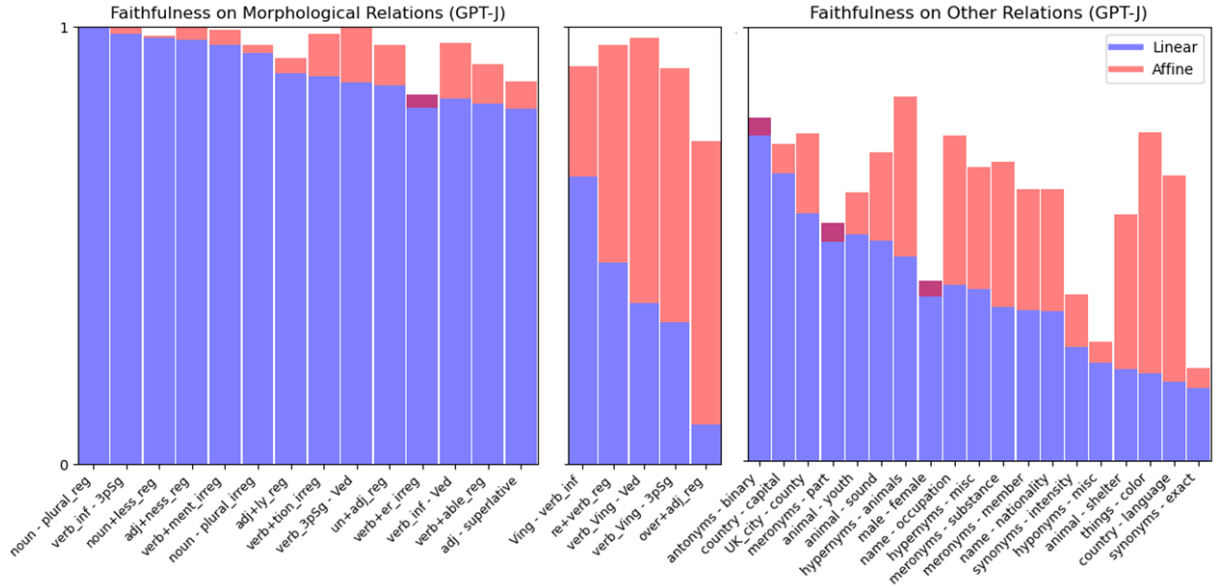
Figure 3: Comparing affine & linear LREs on GPT-J reveals many morphological relations are linearly approximable. With the exception of prefix and active form subjects, semantic and encyclopedic relations benefit more from the affine LRE than morphology. For subject layers 3-9, the best performing approximation is averaged ($n = 4$).

Following the procedure outlined in Hernandez (2023), we employ 8 in-context learning (ICL) examples for 8 different subject-object prompts for each relation. This allows us to obtain a Jacobian from the model computation which is most likely to exhibit the desired linear encoding.

We omit the subject-object pairs used in construction from the testing pool. We further restrict evaluation to the pairs for which the LM computation is successful in reproducing the object. [3]

4.4 Evaluating Operators

After passing through the activation function in the decoder, the approximated object tokens should faithfully replicate the true LM output.

Affine LRE. The original affine LRE is a two-step approximation involving both a weight term W_r and bias term b_r, which are applied to the subject state $\mathbf{s}$ to produce an approximated output state: $\tilde{\mathbf{o}} = \text{LRE}(\mathbf{s}) = \beta W_r \mathbf{s} + b_r$

Linear LRE. Our variants isolate the components of the LRE in order to inspect their contribution to the approximation. First, we define the linear LRE, a multiplicative operation. This is the subject hidden state $\mathbf{s}$ multiplied by the mean Jacobian for *other subject-object pairs* to derive a

final object state: $\tilde{\mathbf{o}} = \text{Linear}(\mathbf{s}) = W_r \mathbf{s}$

Bias. Second, we define the Bias approximator, an additive operation. This approximator calculates $\tilde{\mathbf{o}}$ by adding b_r, the mean difference between $W_r \mathbf{s}$ and $\mathbf{o}$ for *other subject-object pairs*, to $\mathbf{s}$: $\tilde{\mathbf{o}} = \text{Bias}(\mathbf{s}) = \mathbf{s} + b_r$

Following Hernandez et al. (2023), we define *faithfulness* of an approximator by the top-one token match rate. For token t and decoder head D, we say an approximator is faithful if the top token approximation matches that of the LM:

$$\underset{t}{\arg\max}\, D(\mathbf{o})_t \overset{?}{=} \underset{t}{\arg\max}\, D(\tilde{\mathbf{o}})_t$$

5 Results

5.1 The Linear LRE Faithfully Approximates Relations across Morphology

We first evaluate relational approximators for the GPT-J model (Wang and Komatsuzaki 2021). We build approximators for likely subject hidden states (layers 3-9) and the final object state (layer 27) through the process outlined above. We then evaluate the approximators four times for each relation, and average the best cross-layer approximation.[4] [5]

---

[3] For both GPT-J and Llama-7b, nearly all examples fit this criteria.

[4] There were two relations which were not tested on, **[adj+comparative]** and **[antonyms-gradable]**. This was due to preprocessing issues.

[5] For the LRE, we use $\beta = 7$, which was found to be optimal for BATS.
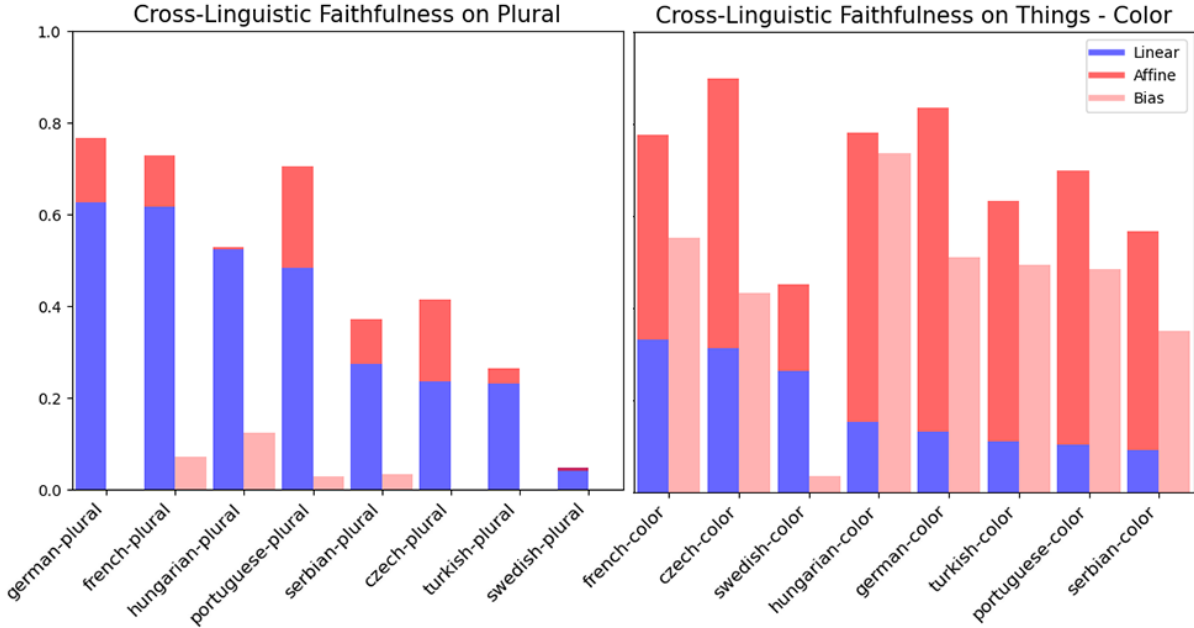
Figure 4: Evaluating languages present in Llama-7b reveal cross-typological linear encoding of morphology. Linear and affine LREs respectively score 56% and 68% on **[plural]** across German, French, Hungarian, and Portuguese. In contrast, on **[things - color]** relation the linear and affine techniques respectively score 19% and 70%. The Bias approximator scores 45%, suggesting the affine approximation for **[things - color]** is primarily additive.

As seen in Figure 3, the linear LRE achieves 90% faithfulness across 14 morphology relations, while the affine LRE achieves a faithfulness of 95%. In contrast, the linear LRE achieves 40% faithfulness over non-morphological relations, while the affine LRE achieves 61% faithfulness. This confirms the efficacy of the affine LRE found by Hernandez et al. (2023), while suggesting that some relations, e.g. morphology, may be encoded as truly linear.

To show that the Jacobian is not only sufficient but also necessary, in Appendix Figure 5 and Appendix Figure 6 we compare the LREs against two additive approximations, Bias and TRANSLA-TION. TRANSLATION adds the mean difference between the subject and object states to each subject state. In both cases, we find that an additive operator is unable to reproduce morphology.

5.2 Llama-7b Results

GPT-J utilizes parallel MLP and attention layers, unlike many other language models. Consequently, it is possible the observed linearity does not generalize to different architectures. We repeat the procedure for Llama-7b, which like most LLMs utilizes sequential attention and feedforward layers (Touvron et al. 2023). In the Appendix Figure 7, we display similar results to Figure 3; suggesting similar encoding mechanisms exist across models.

5.3 Cross-Linguistic Evidence

We have shown that morphological relations in English are largely linearly decodable. However, these results may be limited to fusional-analytic languages with fewer unique affixes. For Llama-7b, we test Czech, French, German, Hungarian, Portuguese, Serbian, Swedish, and Turkish, each comprising significant portions of the training dataset. Hungarian and Turkish are both highly agglutinative. We create templates for one morphological (**[plural]**) and non-morphological relation (**[things - color]**). We evaluate approximators as above.

As seen in Figure 4, affine and linear approximators achieve similar results on **[plural]**, while the additive operation performs well on **[things - color]**. These results indicate a multiplicative linear relational embedding for certain morphological relations, independent of linguistic typology. The high performance of the additive Bias operator on **[things - color]** provides evidence for complementary additive and multiplicative mechanisms.

6 Conclusion

In this work, we have adapted a large relational dataset for testing transformer approximation. We formulate the transformer version of the linear relational embedding found in Paccanaro and Hinton (2001) more precisely to be equivalent to a

matrix-vector multiplication with the mean Jacobian. Surprisingly, we find this linear operation is able to model certain relations such as morphology nearly as well as the affine LRE. This suggests that certain conceptual relations surface linearly in the residual space of language models, and are sparsely encoded multiplicatively as opposed to additively.

7 Limitations

Our experiments were conducted exclusively on GPT-J and Llama-7b due to hardware constraints, which limited the scope of our evaluations. However, smaller models serve as a likely proxy for studying the interpretability of transformer-based language models due to identical architectures.

Throughout the work, we assume linear transformations observed are employed in token prediction through the same mechanism as in explicit relational contexts. Existing literature in activation patching and editing indicates that subject enrichment occurs independently from surrounding contexts (Geva et al. 2021), indicating that the relational embedding outlined here is consistent.

Unlike previous investigations of linear approximation, we did not investigate whether the faithfulness of the Jacobian approximation is associated with causality. Based on prior work which finds a consistent relationship between these variables (Hernandez et al. 2023), these two measures appear correlated.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. 2010. Mathematical foundations for a compositional distributional model of meaning. *arXiv preprint arXiv:1003.4394*.

Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2):155–170.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting Recall of Factual Associations in Auto-Regressive Language Models. *arXiv preprint*. ArXiv:2304.14767 [cs].

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. 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*, pages 8–15, San Diego, California. Association for Computational Linguistics.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. 2023. Linearity of relation decoding in transformer language models. *arXiv preprint arXiv:2308.09124*.

Geoffrey E Hinton. 1986. Learning distributed representations of concepts. In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*, volume 1, page 12. Amherst, MA.

Keith J Holyoak. 2012. Analogy and relational reasoning. *The Oxford handbook of thinking and reasoning*, pages 234–259.

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

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. 2022b. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. 2023. Language models implement simple word2vec-style vector arithmetic. *arXiv preprint arXiv:2305.16130*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Alberto Paccanaro and Geoffrey E. Hinton. 2001. Learning distributed representations of concepts using linear relational embedding. *IEEE Transactions on Knowledge and Data Engineering*, 13(2):232–244.

Robert J Sternberg and Bathsheva Rifkin. 1979. The development of analogical reasoning processes. *Journal of experimental child psychology*, 27(2):195–232.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ekaterina Vylomova, Laura Rimell, Trevor Cohn, and Timothy Baldwin. 2016. Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1671–1682.

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

A Reproducibility Statement

The approximation code is based on the LRE repository (Hernandez et al. 2023), and loads GPT-J and Llama-7b in half-precision. The code and dataset are available at {link}. Experiments were run remotely on a workstation with 24GB NVIDIA RTX 3090 GPUs using HuggingFace Transformers.

B Evidence of non-stemmed forms

As seen in Table 1, the linear LRE successfully replicates full forms for many derived object states. In Table 3, we can see consistent preferences for correct forms over stemmed forms on morphological relations. All examples shown are for GPT-J.

C Bias Results demonstrate W necessity

A comparison of linear and affine approximators against the bias approximator demonstrates that the bias term b_r alone cannot explain the relational encoding but contributes alongside the Jacobian W_r. This suggests that these operations play complementary roles in semantic and encyclopedic relations.

The TRANSLATION operator, inspired by Merullo et al. (2023) and vector arithmetic, is also additive and performs similarly to the Bias operator. Figure 6 demonstrates the additive TRANSLATION approximator against both the affine and linear LRE. Like the bias approximator, the TRANSLATION approximator succeeds when the gap between the Jacobian and LRE is large. This suggests that semantic information plays a crucial role in bridging some subject-object relations.

D Linear Projection

We find that linear projection to $\mathbb{R}^2$ can yield interpretable geometric representations. Specifically, we use a basis of the bias vector b and a random normalized vector, which has been orthogonalized with Gram-Schmidt to b, and compare approximated transformations against true object states. As seen in Figure 8, we find subspace distance corresponding heavily to faithfulness. Additionally, we validate that the β hyperparameter is necessary for recovering scale lost in layer normalization, as conjectured by Hernandez et al. (2023).

We project approximations $\mathbf{s}$, $\beta W \mathbf{s}$, $\beta W \mathbf{s} + b$, as well as a calculated hidden state for the correct object output $\mathbf{o}$. These projections suggest W is

Subject	Jacobian Top-3
society	societies, Soc, soc
child	children, children, Children
success	successes, success, Success
series	series, Series, Series
woman	women, women, Women
righteous	righteousness, righteous, . . .
conscious	consciousness, conscious, . . .
serious	seriousness, serious, serious
happy	happiness, happy, happy
mad	madness, mad, being
invest	investment, invest, investing
amuse	amusement, amuse, amusing
accomplish	accomplishment, accomplish, . . .
displace	displacement, displ, dis
reimburse	reimbursement, reimburse, reimb
globalize	globalization, global, international
install	installation, install, Installation
continue	continuation, continu, contin
authorize	authorization, Authorization, . . .
restore	restoration, restitution, re
manage	manager, managers, manager
teach	teacher, teachers, teach
compose	compos, composer, composing
borrow	borrower, lender, debtor
announce	announcer, announ, ann

Table 1: **[noun_plural], [verb+er], [verb+ment], [adj+ness], [verb+tion]** Selected examples of full subject tokens demonstrate that the linear Jacobian approximation captures irregular morphology effectively, reproducing both stemmed and full subject forms.

Relation	# Unique
un+adj	**7**
over+adj	**4**
re+verb	**15**
name - nationality	13
animal - shelter	18
synonyms - intensity	35
verb+able	47
noun - plural	47

Table 2: The number of unique start tokens for correct objects across selected BATS relations. Start tokens which occur frequently among objects indicate a non-injective subject-object map, making linear approximation a less suitable choice as an approximator.

Correct	Stemmed	Incorrect
42	0	0
23	11	9
7	35	6

Table 3: Correct, stemmed, and incorrect suffix counts for **[noun_plural]**, **[verb+tion]** and **[adj+ness]** from the top prediction of a fixed layer Jacobian approximation further suggests consistent linear encoding beyond stemmed forms.
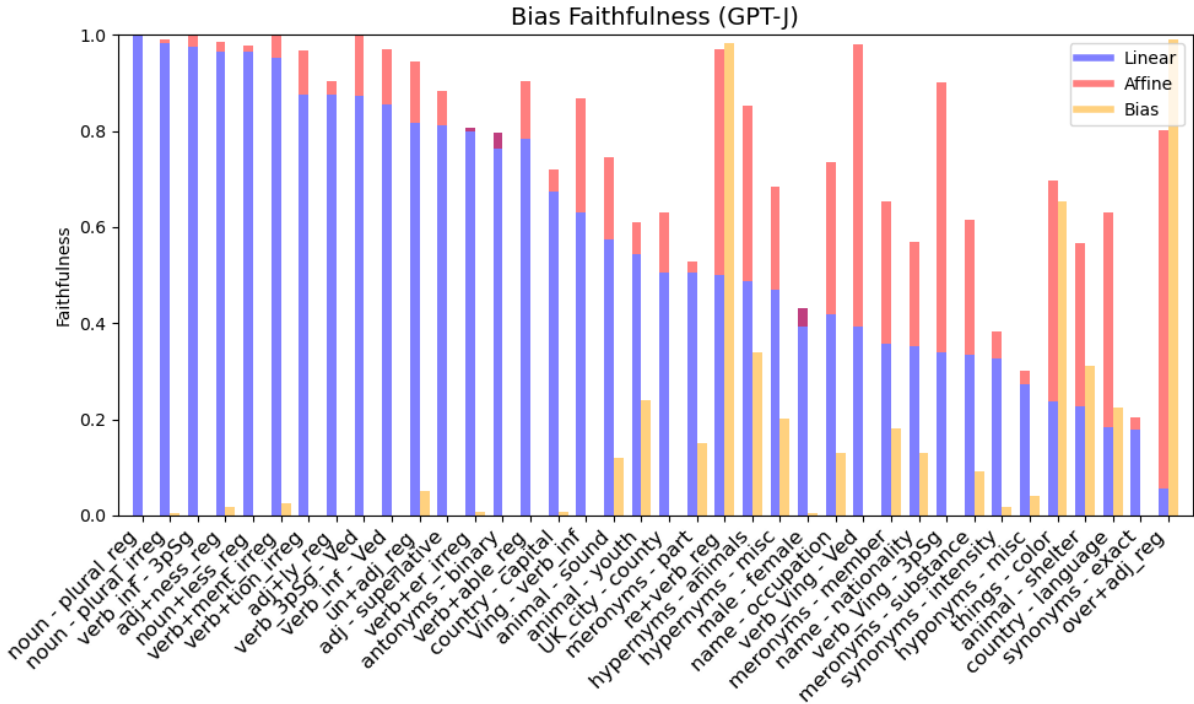
Figure 5: A comparison of the affine LRE against the Bias approximator demonstrates the necessity of the multiplicative (Jacobian) operator. Across semantic and encyclopedic relations, the additive Bias operator exhibits far better performance on morphology, providing evidence for complementary additive and multiplicative mechanisms.
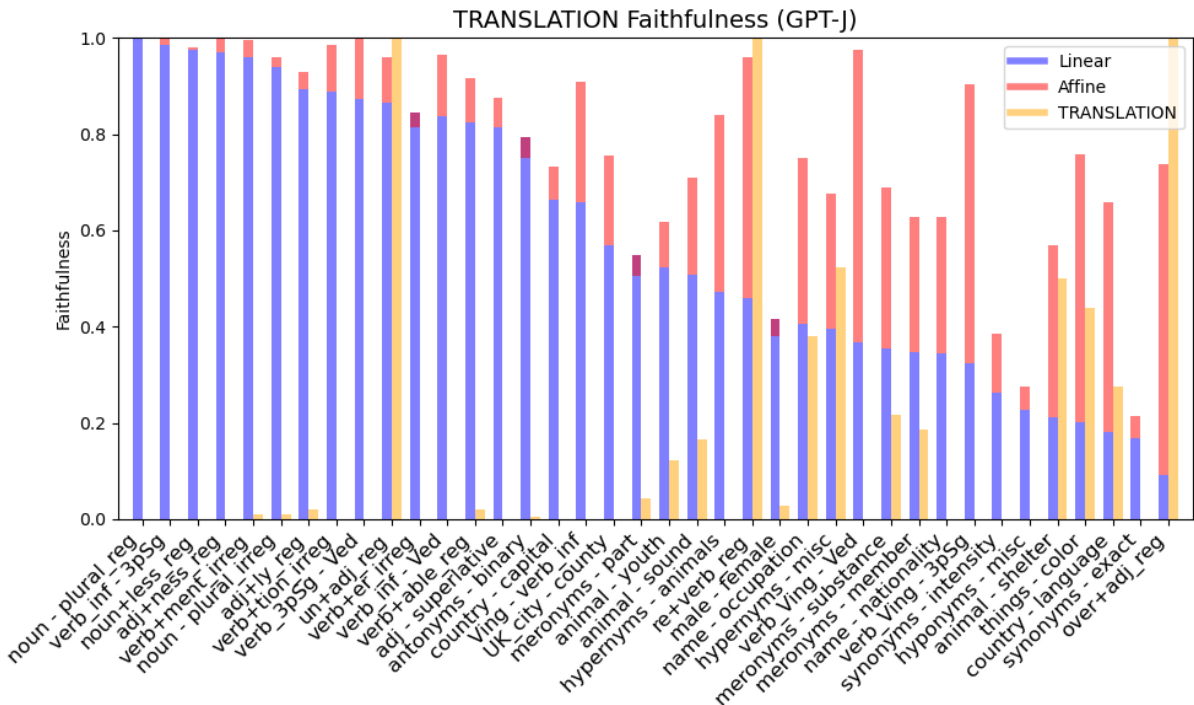

Figure 6: The TRANSLATION approximator $\tilde{\mathbf{o}} = \text{Bias}(\mathbf{s}) = \mathbf{s} + b_r$, with $b_r = \mathbb{E}(\mathbf{o} - \mathbf{s})$, performs well on semantic and encyclopedic relations, similar to the Bias approximator.

primarily responsible for transforming the underlying distribution to be geometrically similar to the output, while b contributes the majority of movement in vector space.

The term b_r could be compared to the vectors used by Mikolov and many others, and the concept vector subsequently formalized by Park. However, the bias vector and the concept vector are not truly analogous. The bias term describes an offset from the transformed subject to the object: $b_r = \mathbb{E}(o - W_r \mathbf{s})$, not $b_r = \mathbb{E}(o - \mathbf{s})$. In practice, we find that bias and concept vectors are close in cosine similarity, and likely serve similar roles.
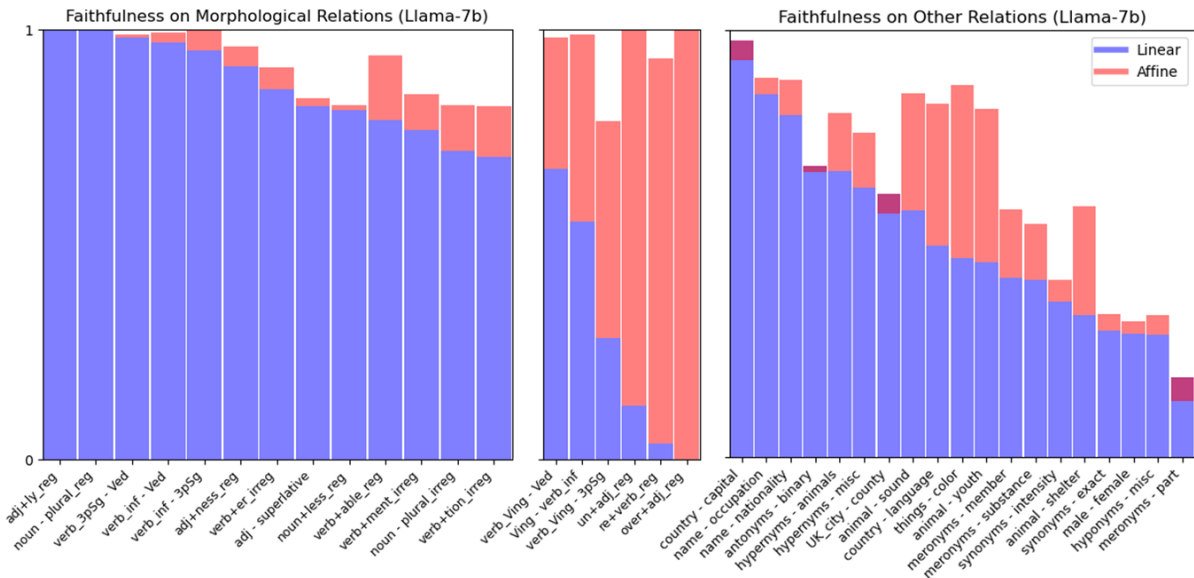
Figure 7: Llama-7b results support a generalization across models: many morphological relations are linearly approximable, while semantic and encyclopedic relations benefit greatly from the affine method. Out of a range of subject layers 4-16, the best performing approximation is averaged ($n = 4$).
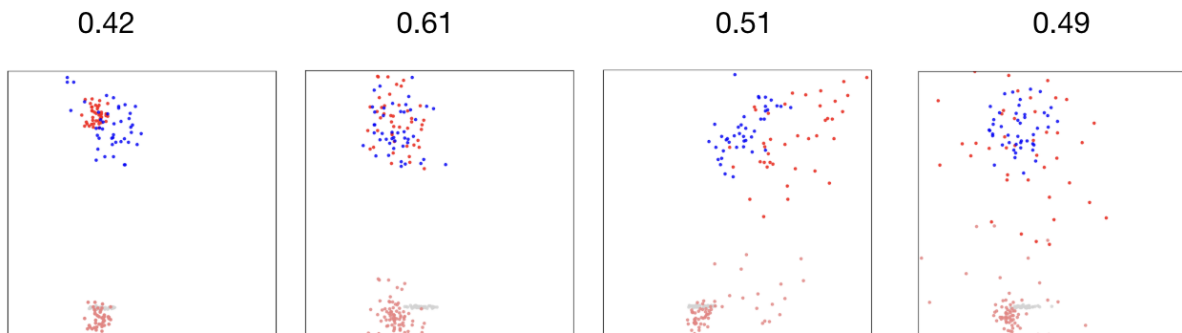

Figure 8: Projected subspace distances for fifty approximated object states $\beta W\mathbf{s} + b_r$ and true object states $\mathbf{o}$ for [animal - youth]. The subspace used is $\{\perp, b_r\}$, where $\perp$ is a randomly chosen orthogonal vector to b_r. The faithfulness scores of each relation are displayed above. With β values of 1, 3, 5, and 7, the hyperparameter β is shown to be crucial for faithful approximation in the affine LRE.