Linear Decoding of Morphology Relations in Language Models

Anonymous ACL submission

Abstract

The recent success of transformer language models owes much to their conversational fluency and productivity in linguistic and morphological aspects. An affine Taylor approximation has been found to be a good approximation for transformer computations over certain factual and encyclopedic relations. We show that the truly linear approximation $W\mathbf{s}$, where $\mathbf{s}$ is a middle layer representation of the base form and W is a local model derivative, is necessary and sufficient to approximate *morphological derivations*. This approach achieves above 80% faithfulness across most morphological tasks in the Bigger Analogy Test Set. We argue that morphological forms in transformer models are likely to be encoded by linear transformations, with implications for how entities are represented.

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 the contexts in which approximations perform well is an important area of study, with applications in interpretability and model editing.

Work to date around relational representation in LMs have primarily focused on relations in the context of factual subjects and objects (Meng et al., 2022), (Hernandez et al., 2023), (Chanin et al., 2023). However, relations in natural language encompass a much broader range of subject and object relations. Much of the mainstream success of LLMs has been due to the conversational nature of chat-oriented language models. The impressive conversational ability of LLMs depends on their linguistic competency, including lexical and morphological productivity, and uncovering how models are able to achieve this is an important aspect of model interpretability.

We demonstrate that for morphological relations, a transformation from the Jacobian alone is able to approximate object computations from an enriched subject exceptionally well. This suggests that transformers encode morphology linearly, in an even simpler fashion than the affine LRE discovered by Hernandez (2023). Linearly approximating the computation from an early hidden state of the base form to the final state of the derived form, we find that approximable relations include pluralization, nominalization, changes in tense, and resultative forms. These derivations range over different parts of speech, including noun to adjective **[noun+less]**, adjective to noun **[adj+ness]**, verb to noun **[verb+er]**, and verb to adjective **[verb+able]**, and involve diverse subjects and objects.

By linearly decoding linguistic relations in this manner, we offer an interpretation which reveals how internal ontological atomic representations, such as those espoused in the Linear Representational Hypothesis and concept theory (Park et al., 2023), (Wang et al., 2023), (Park et al., 2024), could be unfolded by a LM to encompass a range of morphological variations. We show that relational approximation in LMs can be applied to a broad range of linguistic phenomena, opening avenues for further research in model representations. To the best of our knowledge,

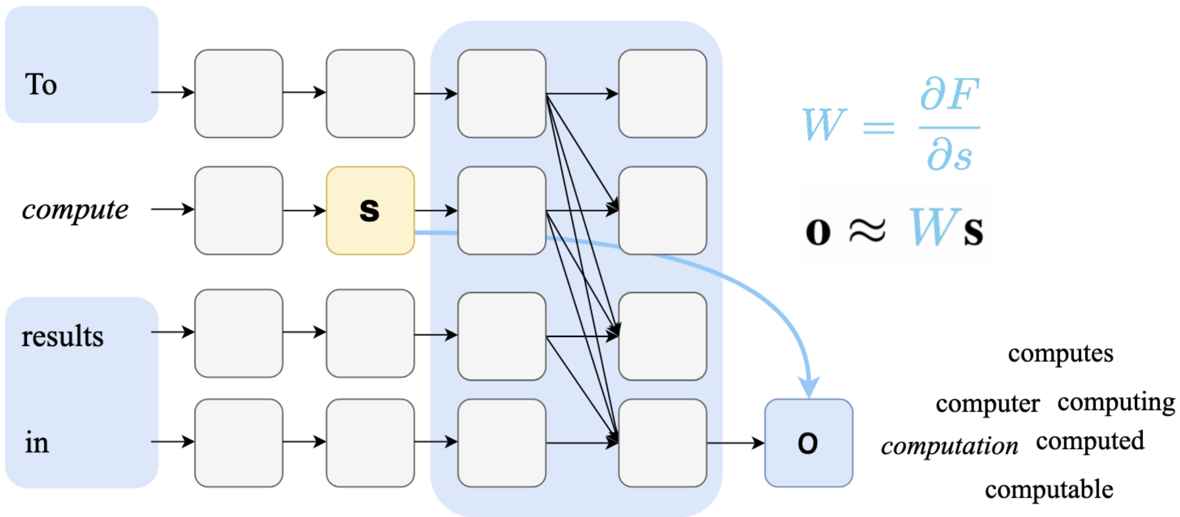


Figure 1: Adapting morphological analogies from the Bigger Analogy Test Set to relational contexts reveals that many are genuinely linearly approximable, such as **[verb+tion]**, **[verb+able]**, **[noun - plural]**, **[verb_inf]**, and **[verb+er].**

we are the first to provide empirical evidence of a linear transformation acting as an effective LM approximator over a broad range of inputs.

To understand affine approximation better, we also analyze linear projections of the LRE. We find that the Jacobian increases the geometric similarity of the subject and object spaces, while the bias contributes the majority of the movement. We confirm the previous hypothesis that a beta parameter is necessary to reproduce the output space. We find that approximation accuracy can be measured by embedding variance, and that linear representation of final layer embeddings is an effective way to diagnose relations which are not affinely approximable.

2 Background & Related Work

2.1 Transformer Computation

In autoregressive transformer language models, input text is converted to a sequence of tokens $t_1 \dots t_n$, which are subsequently embedded as $x_1 \dots x_n \in \mathbb{R}^d$ by an embedding matrix. The hidden states $x_1 \dots x_n$ are then passed through L transformer layers, each composed of a self-attention layer a^l and an multi-layer perceptron (MLP) layer m^l, and then decoded by the decoder head D to a probability distribution over tokens. The representation state x_i^l of the i^{th} token at layer l is obtained as:

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

Where a_i^l and m_i^l are

$$a_i^l = \text{attn}^l \left(x_1^{l-1}, x_2^{l-1}, \dots, x_i^{l-1} \right)$$
$$m_i^l = W_{out}^l \, \sigma \left(W_{in}^l (a_i^l + x_i^{l-1}) \right)$$

Here, attn^l is multiheaded Key-Value Query attention as described in Vaswani et al. (2017), W_{in}^l and W_{out}^l are projection matrices, and σ is a nonlinear activation function. In GPT-J, the output of the l^{th} MLP sublayer for the i^{th} representation depends only on x_i^{l-1} rather than $a_i^l + x_i^{l-1}$, so the attention and MLP modules operate in parallel (Wang and Komatsuzaki, 2021).

Following the insights of Meng (2022) and Geva (2023) that the last subject token state in middle layers are strongly casual, we are primarily interested in utilizing the Jacobian (derivative) between the hidden state at the last subject token x_s^i, and the last token position overall, the object prediction x_o^L. The LM computation between these two states is clearly highly nonlinear, but within certain relational contexts this derivative can yield faithful approximations (Hernandez et al., 2023).

2.2 Internal Relational Representation

Relations can be encoded as $n \times n$ matrices, or linear transformations. In transformer models, positional encodings are designed to linearly encode relative positions through a range of methods (Vaswani et al., 2017; Su et al., 2024).

Through linear probing and embedding analysis, transformer models have been found to encode high-level linguistic features in internal embedding representations, such as syntactic dependencies and thematic categories (Kann et al., 2018; Tenney et al., 2019; Wilson et al., 2023).

Meng et. al. (2022) found that factual statement predictions exhibit strongly causal states in middle layers at last subject token, supporting the idea that an enriched subject representation exists prior to output. Geva (2023) demonstrated that attribute extraction is often performed by specific attention heads in later layers, and takes the form of a query on the enriched representation. The AttributeLens directly applies this notion to extract encoded attributes from hidden states (Hernandez et al., 2023).

We directly build off of work by Hernandez et al. (2023), who present a an affine linear approximator, known by the corresponding internal hypothesis of the Linear Relational Encoding. With $\mathbf{s}$ denoting the hidden subject state and $\mathbf{o}$ denoting the final object state, they treat object-retrieval within a relational context as linearly approximable: $\mathbf{o} = F(\mathbf{s})$. They model o with a first-order Taylor approximation

$$o = F(\mathbf{s}) \approx W\mathbf{s} + b$$

using the transformer Jacobian $\frac{\partial F}{\partial s}$ from relation examples to approximate W, and utilizing a subject representation $\mathbf{s}$ from an intermediate layer. By doing so, they achieve over 60% faithfulness for LM predictions across certain factual, commonsense, linguistic, and bias relations.

2.3 Linear Embedding Spaces

Paccanaro and Hinton (2001) introduced the concept of the linear relational embedding for learning relational knowledge from triples (a, R, b). Along with prior work (Hinton, 1986), they were able to solve a family tree problem where data is given in relational triples (Colin, *child*, Victoria), where vector components captured implicit semantics such as generation. 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 Coecke's vector semantics (2010) and similar to the hidden state representation described above.

Mikolov et al. (2013) used linear operations in word vector space derived from context-predictive neural nets, demonstrating a correspondence between directional binary relations (male-female, country-capital, verb tense) and the addition of certain embedding vectors. Subsequent work found inflection relations (*comparative*, strong:stronger) are better captured than derivation relations (*lacking*, life:lifeless), and that encyclopedic relations (*capital-of*, Greece:Athens) are better captured than lexicographic relations (*member-of*, player:team) (Gladkova et al., 2016; Vylomova et al., 2016).

Park et al. (2023) formalize the compositional representation of concepts in embedding spaces. Extending prior work (Wang et al., 2023), they define a set of counterfactual outputs Y for a directional binary concept W. Letting W = male⇒female, the space of outputs comprise:

$$(Y(W{=}0), Y(W{=}1)) \in \{(''man'', ''woman''), \\ (''king'', ''queen''), \ldots\}$$

They formalize concept intervention as adding an embedding representation $\bar{\lambda}_W$ to change the probability of an output reflecting a concept W. For any concept Z linearly separable from W, an output word $Y(W, \ldots, Z)$, and concept embedding λ, an intervention is effective if it changes the probability of W but not Z.

Subsequent work (Park et al., 2024) illustrated concepts were linearly encoded in the final embedding layer, through noun projections with particular binary characteristics against estimated feature vectors.

2.4 In-context learning

Our work utilizes in-context learning (ICL) for both training and testing purposes, where input-label pairs are provided as demonstration for a novel task. Min (2022)'s in-depth empirical study of ICL finds that ground truth demonstrations are not necessary, and suggests that more important to ICL is the identification of a label and output space, while Wei et al. (2024) proposes that learning input-label mappings is an emergent ability of large LLMs. Yan (2023) also performs an in-depth study on what they term the token reinforcement loop, providing empirical evidence of $n = 8$ as an optimal number of examples for the LRE. During the testing process, we reproduce Hernandez et. al. (2023)'s findings that approximations without relation-specific contexts generally perform significantly worse than relation-specific contexts. Further work in this area is important for understanding how concepts are represented in transformers.

3 Problem Statement

The LRE is well motivated mathematically, under the assumption the transformer computation is linearly approximable for a specific contextual relation. The object retrieval function from a subject with a fixed relational context, $o = F(s)$, is hypothesized by Hernandez et al. (2023) to be modeled by a first-order Taylor approximation of F about a number of examples $s_1 \dots s_n$. For $i = 1 \dots n$:

$$\begin{aligned} F(s) &\approx F(s_i) + W(s - s_i) \\ &= F(s_i) + Ws - Ws_i \\ &= Ws + b, \\ \text{where } b &= F(s_i) - Ws_i \end{aligned}$$

Note that we get a W and b for each s_i. This motivates the following definition for W and b over a relation. They can be calculated as the mean Jacobian and bias between n enriched subjects $s_1 \dots s_n$ and outputs $F(s_1) \dots F(s_n)$ for a fixed relation:

$$W = \mathbb{E}_{s_i}\left[\left.\frac{\partial F}{\partial s}\right|_{s_i}\right]$$
$$b = \mathbb{E}_{s_i}\left[F(s) - \left.\frac{\partial F}{\partial s}\, s\right|_{s_i}\right]$$

The LRE diverges from its namesake, the linear relational embedding introduced by Hinton (1986), by introducing the bias b and scaling β terms:

$$\mathbf{o} \approx \beta W\mathbf{s} + b$$

They claim the 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.

However, while linearity is assumed by Hernandez by calculating W and b from $\mathbb{E}_{s_i}$ over $i = 1 \dots n$, defining the approximation as a Taylor series implicitly makes a weaker assumption. Under the assumption that the relation is not only linearly approximable, but truly linear, we would expect the following approximation to be valid:

$$\mathbf{o} \approx F'(s_i)\mathbf{s}$$

This motivates the definition for a linear approximation over $s_1 \dots s_n$ within the same relation to be simply the mean Jacobian [1]:

$$W = \mathbb{E}_{s_i}\left[\left.\frac{\partial F}{\partial s}\right|_{s_i}\right]$$
$$F(s) = Ws$$

This is the form given in the original linear relational embedding (Paccanaro and Hinton, 2001). We will test this approximation against the LRE; for truly linear relations, we anticipate equivalent performance.

4 Approach

4.1 Introducing New Relations

The Bigger Analogy Test Set, was originally introduced to explore linguistic regularities in word embeddings (Gladkova et al., 2016). It provides forty different categories, ten each in inflectional morphology, derivational morphology, encyclopedic knowledge, and lexical semantics. Each category consists of 50 unique word pairs; the dataset contains 2000 samples total. The data is compiled from various sources, including WordNet, SemEval2012-Task2, Wikipedia, the Google Analogy Test Set, and a color dataset built for evaluating multimodal models (Fellbaum, 1998; Jurgens et al., 2012; Mikolov et al., 2013; Bruni et al., 2012).

[1] Note that because the scale of the hidden state does not contribute to an output prediction, β is irrelevant:

$$\operatorname*{argmax}_t D(Ws) = \operatorname*{argmax}_t D(\beta Ws)$$

Inflections	Nouns	I01: regular plurals *(student:students)*
		I02: plurals - orthographic changes *(wife:wives)*
	Adjectives	I03: comparative degree *(strong:stronger)*
		I04: superlative degree *(strong:strongest)*
	Verbs	I05: infinitive: 3Ps.Sg *(follow:follows)*
		I06: infinitive: participle *(follow:following)*
		I07: infinitive: past *(follow:followed)*
		I08: participle: 3Ps.Sg *(following:follows)*
		I09: participle: past *(following:followed)*
		I10: 3Ps.Sg : past *(follows:followed)*
Derivation	No stem change	D01: noun+less *(life:lifeless)*
		D02: un+adj. *(able:unable)*
		D03: adj.+ly *(usual:usually)*
		D04: over+adj./Ved *(used:overused)*
		D05: adj.+ness *(same:sameness)*
		D06: re+verb *(create:recreate)*
		D07: verb+able *(allow:allowable)*
	Stem change	D08: verb+er *(provide:provider)*
		D09: verb+ation *(continue:continuation)*
		D10: verb+ment *(argue:argument)*

Lexicography	Hypernyms	L01: animals *(cat:feline)*
		L02: miscellaneous *(plum:fruit, shirt:clothes)*
	Hyponyms	L03: miscellaneous *(bag:pouch, color:white)*
	Meronyms	L04: substance *(sea:water)*
		L05: member *(player:team)*
		L06: part-whole *(car:engine)*
	Synonyms	L07: intensity *(cry:scream)*
		L08: exact *(sofa:couch)*
	Antonyms	L09: gradable *(clean:dirty)*
		L10: binary *(up:down)*
Encyclopedia	Geography	E01: capitals *(Athens:Greece)*
		E02: country:language *(Bolivia:Spanish)*
		E03: UK city:county *York:Yorkshire*
	People	E04: nationalities *(Lincoln:American)*
		E05: occupation *(Lincoln:president)*
	Animals	E06: the young *(cat:kitten)*
		E07: sounds *(dog:bark)*
		E08: shelter *(fox:den)*
	Other	E09: thing:color *(blood:red)*
		E10: male:female *(actor:actress)*

Figure 2: The BATS dataset structure from Gladkova et al. (2016)

We adapt the Bigger Analogy Test Set to a relational dataset by introducing relation-specific prompts for each analogy dataset. The derivational morphology dataset **[verb+ment]** uses the clozed prompt *"To* $\{\}$ *results in a"*, which is filled in with subjects to obtain the Jacobians used. For example, one corresponding prompt would be *"To fulfill results in a"*, eliciting the object *"fulfillment"*. We use the first item as the subject and the second as the object, except in **[verb_inf - Ving]**, where the reverse was used.

4.2 Utilizing ICL

Following the testing standards established by Hernandez (2023), we use 8 ICL examples for 8 different subject-object pairs to create an approximator for each relation. For instance, we might approximate **E06 [animal - youth]** with the pairs $\{(\textit{dog}, \textit{puppy}), (\textit{sheep}, \textit{lamb}), \ldots\}$, prepending the 7 other examples before each pair.

We restrict evaluation to the pairs for which the LM computation is successful in reproducing the actual object in question: for both GPT-J and Llama-7b, this is all or nearly all of the examples provided in BATS. See the Appendix for statistics on successful completion.

4.3 Evaluating the Jacobian

We primarily work with the six-billion parameter model GPT-J. Following Hernandez, we measure approximator faithfulness over a relation by the top-one token match rate for the approximation and the LM. Let the enriched subject state be $\mathbf{s}$ and the relation-expressing context be r. Let the transformer computation be $o = F(\mathbf{s})$ and the relational approximator be $\tilde{F}$. Then for token t and decoder head D, we say an approximator is faithful if the top token approximation matches the LM:

$$\operatorname*{argmax}_t D(F(\mathbf{s}))_t \stackrel{?}{=} \operatorname*{argmax}_t D(\tilde{F}(\mathbf{s}))_t$$

In the original LRE,

$$\tilde{F} = \beta W\mathbf{s} + b$$

We primarily test two variants of the LRE. First, the Jacobian approximator:

$$\tilde{F} = W\mathbf{s}$$

Second, the Bias approximator:

$$\tilde{F} = \mathbf{s} + b$$

We would like our approximations to generalize over new subjects. In order to do so, we omit the subject-object pairs used to build the approximator from the testing pool.

5 Results

5.1 The Jacobian Faithfully Approximates Morphological Relations

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, with randomized test prompts each iteration, and average the best performing approximation from each. For the LRE, we use $\beta = 7$, which was found to be optimal for BATS. We find that the Jacobian reproduces derivational and inflectional morphology particularly well [2]. In most other morphology categories, the LRE performance does not improve significantly past the Jacobian, suggesting that the object representation is well captured by the Jacobian alone.

The high faithfulness of the Jacobian shows that it is sufficient to approximate most morphological relations, but not that it is necessary. To show that the Jacobian is also necessary, we compare against the Bias approximation $\mathbf{s} + b$, (equivalent to $\mathbf{s} + \mathbb{E}(o - W\mathbf{s})$). We also compare against the TRANSLATION approximator, where the bias is formulated as $b = \mathbb{E}(o - \mathbf{s})$.[3] We find that without the Jacobian, bias approximations fail to approximate nearly all morphological relations, while successfully capturing some semantic and encyclopedic relations: the bias approximator achieves 67% faithfulness on **[things - color]**, while the TRANSLATION estimator attains 50% and 52% faithfulness on **[animal - shelter]** and **[hypernyms - misc]** respectively.

5.2 Llama-7b Results

While these results show that the Jacobian is a faithful approximator for morphological relations in GPT-J, it is possible that the unique architecture decisions have contributed to the observed linearity. We repeat the process for Llama-7b, which like most popular LLMs, utilizes sequential attention and feedforward layers (Touvron et al., 2023). As seen in Figure 4, we obtain very similar results. [4]

We can conclude that morphological relations can be decoded linearly in LMs, and that they are likely to be linearly encoded.

In general, the Jacobian does poorly on semantic and encyclopedic relations, highlighting the complementary role of the bias term.

5.3 Underlying Mechanisms

Due to the layer normalization within the decoder head, the scale of the hidden state does not contribute to an output prediction. We have

$$\underset{t}{\operatorname{argmax}}\, D(\beta W\mathbf{s} + b)_t = \underset{t}{\operatorname{argmax}}\, D(W\mathbf{s} + \frac{b}{\beta})_t$$

In other words, the scale factor β on $W\mathbf{s}$ is equivalent to $\frac{1}{\beta}$ on b, and that the LRE approaches $W\mathbf{s}$ asymptotically for high β. In practice, we observed that for $\beta > 100$, the performance of the LRE becomes negligibly different from $W\mathbf{s}$.

We would like to interpret W and b. The approximation results in Figure 3 and 4 give credence to the idea that W and b play complementary roles in the approximation. One potential interpretation is that the weight provides variation in the embedding space, while the bias is a concept shift. With this interpretation, b can be compared to the vectors used by Mikolov and many others, and the concept vector subsequently formalized by Park. However, it's important to note the bias vector and the concept vector are not exactly analogous. The bias term describes an offset from the transformed subject to the object: $b = \mathbb{E}(o - W\mathbf{s})$, not $b = \mathbb{E}(o - s)$. We observe the bias vector does not generally lie in the same direction as $\mathbb{E}(o - s)$, suggesting it may play a different role in transforming the subject.

We find that it is possible to get interpretable subject representations through linear projection onto the span $\{b, \perp\}$, where $\perp$ is a vector orthogonalized through Gram–Schmidt to b [5]. They suggest W is primarily responsible for transforming the underlying distribution to be geometrically similar to the output, while b contributes the majority of movement in vector space. Figures 5 and 6 both display embeddings projected to $\{b, \perp\}$.

Note that the shapes of the transformed subject spaces **Ws** and **Ws + b** are both similar to the object space. Note that the b scale is much larger than the $\perp$ scale.

[2] The exceptions are the prefix relations **[re+verb_reg]** and **[over+adj_reg]**, and present participle relations. We will address these further below.

[3] This approximator, from Hernandez et. al. (2023), calculates the direct offset of the subject and object hidden states, and is inspired by Merullo et al. (2023) and Word2Vec arithmetic. See the appendix for the results.

[4] This diagram will be updated in the final version to look like the GPT-J one: with the best approximations after layer sweeping, and with β optimized for Llama-7b.

[5] There are many options for this orthogonal vector; we chose the first weight column vector W_0.

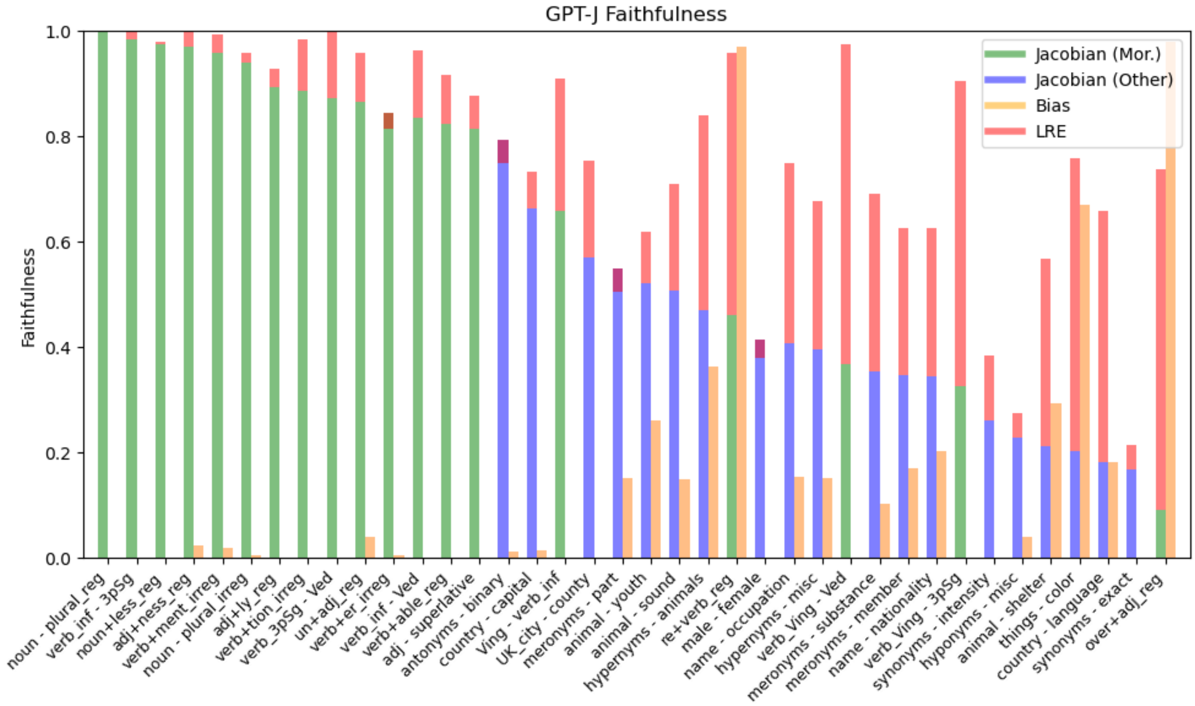


Figure 3: Breaking down the affine LRE into Jacobian ($W\mathbf{s}$) and Bias ($\mathbf{s}+b$) approximators suggests that W and b play complementary roles: the Jacobian is responsible for approximating morphology, while the bias is responsible for conceptual shifts.

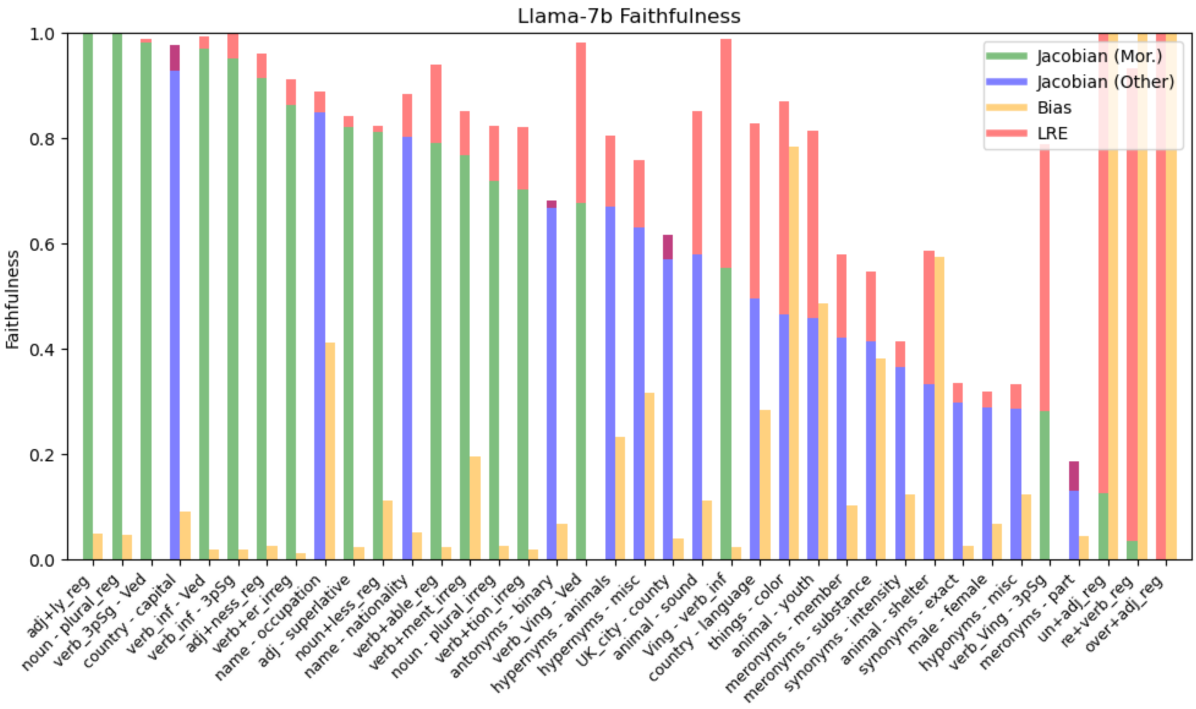


Figure 4: Comparing the LRE and Jacobian for Llama-7b reproduces the results seen above for GPT-J, suggesting the high faithfulness of the Jacobian for morphological relations is widespread among LMs.

Projection also aids in interpreting β in the LRE, which scales the output approximation $\beta W\mathbf{s} + b$. When this approximation approaches the output embeddings $\mathbf{o}$, the performance of the

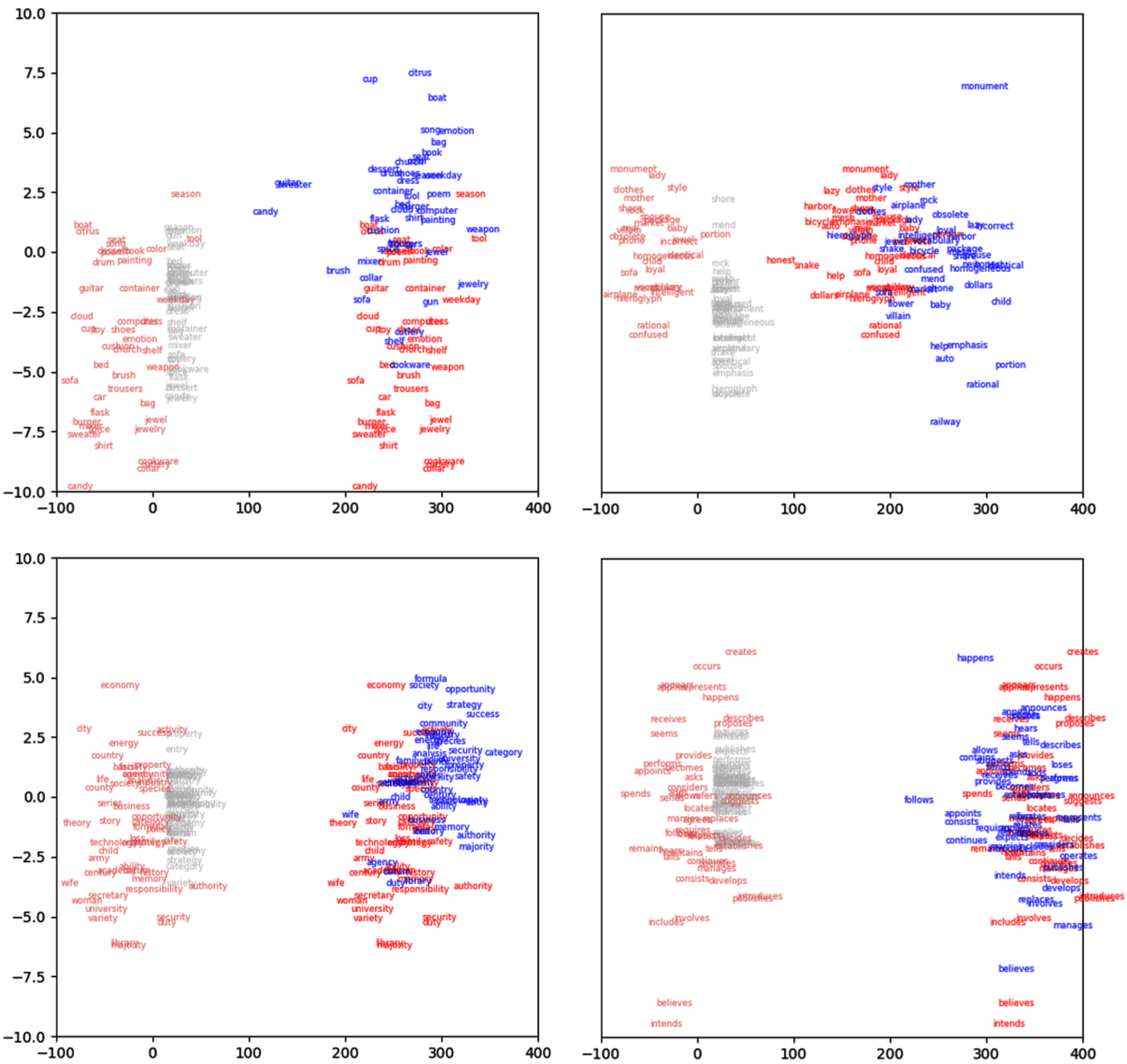


Figure 5: Output space projections of s, $\beta W\mathbf{s}$, $\beta W\mathbf{s}+b$ and $\mathbf{o}$ can be used to diagnose nonapproximable relations. Above, the ineffective **[hyponyms - misc]** and **[synonyms - exact]** approximations do not resemble their corresponding outputs, despite high cosine similarity scores. Below, the effective **[noun - plural_irreg]** and **[verb_3pSg - Ved]** approximations closely resemble their outputs.

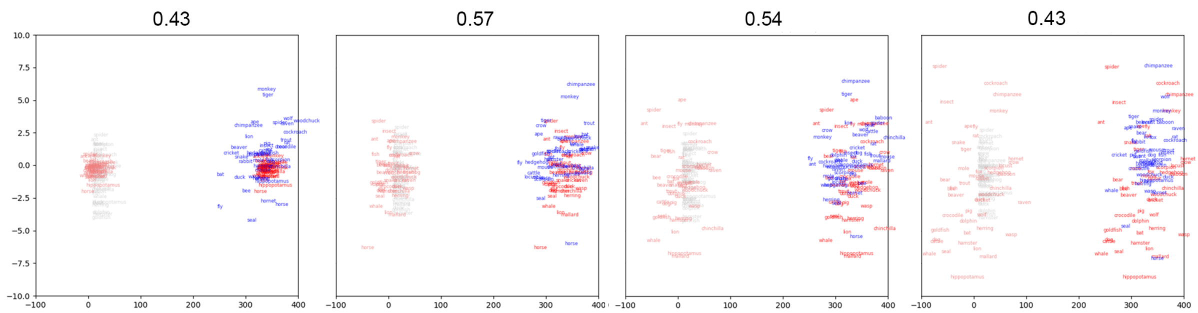


Figure 6: A projection of s, $\beta W\mathbf{s}$, $\beta W\mathbf{s}+b$ and $\mathbf{o}$ for **[animal - shelter]** for $\beta = 1, 3, 5, 7$ with faithfulness scores demonstrates that embedding distances corresponds to the accuracy of the approximation.

approximator improves. In Figure 6, we see empirical evidence that β restores the magnitude of change that was lost through layer normalization, as conjectured by Hernandez et. al. (2023).

Discussion

5.4 Counterarguments

We have shown empirically that it is possible to linearly decode morphological relations in LMs with a high degree of faithfulness. However, several potential issues must be addressed prior to considering the theoretical implications.

Subject	Jacobian Top-3
society	societies, Soc, soc
child	children, children, Children
success	successes, success, Success
series	series, Series, Series
woman	women, women, Women
manage	manager, managers, manager
teach	teacher, teachers, teach
compose	compos, composer, composing
borrow	borrower, lender, debtor
announce	announcer, announ, ann
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

Table 1: **[noun_plural], [verb+er], [verb+ment], [adj+ness], [verb+tion]** Selected examples from GPT-J show that relational Jacobian approximation captures irregular morphology effectively, and does not only reproduce stemmed subject forms.

What if the Jacobian is just modeling syntax? One argument against linear encoding is that the Jacobian is not learning morphology, but instead some regular syntax in the relation. Then, the high faithfulness reported merely reflects some orthographic change from the base, and not a true morphological relation.

What if the Jacobian is replicating the subject? Another argument against linear encoding is that the faithfulness metric is a bad choice for measuring morphological faithfulness. High faithfulness scores on many reported inflectional tasks can be achieved simply by reproducing a substring of the subject token.

To provide a counterargument against the two possibilities above, we provide specific approximation token predictions in Table 1. While generally the fact that approximation outputs tend to be stemmed forms is not a cause for concern, we observe there exist many tokens such as #25303 ' sadness' and #24659 ' continuation' which faithfully replicate morphology.

There are two inflectional relationships the Jacobian failed to approximate as well over the tests performed, **[Ving - 3psg]** and **[Ving - Ved]**.It's possible that transformations from the verb active form make the LM computation non-linear. For the majority of the relations on which the Jacobian achieves high faithfulness, the subject is the unmarked form, such as the verb infinitive or third person singular.

There are two derivational prefix tasks for which the LRE, but not the Jacobian, faithfully approximates, **[re+verb]** and **[over+adj]**. The Jacobian does achieves a high faithfulness on the prefix relation **[un+adj]**, so the notion that prefix relations are distinctive from other morphology is not supported. A partial explanation for this phenomenon may be that the object tokens "over" and "re" are idiosyncratically related by an offset to the subjects, unlike other relations. With fewer correct object tokens than average, a linear subject transformation without any bias may fail to model the relation effectively.

5.5 Implications for Concept Theory

Geometrically, the findings suggests that morphological relations between words do not involve additional concept vectors. Above, we have demonstrated that the bias term is not necessary for morphological terms, and even results in incorrect approximation for low values of β. This is compatible with the Linear Relational Hypothesis, which

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 starting object tokens for selected BATS relations.

posits that subspace distances in LMs are fundamentally about semantics. If morphology is encoded as linear transformation, vectors can retain their semantic interpretations.

Conclusion

In this work, we have adapted the Bigger Analogy Test Set to create a large novel testing dataset for relations, covering forty relations over morphological, factual, and semantic relations. We find Jacobian approximation models morphological relations well. We hypothesise that the Jacobian serves the role of *extending* a subject entity to alternative forms (including morphological derivations), and the bias term serves the role of *shifting* underlying concepts. We validate this hypothesis through embedding projections of model transformations.

Through linear approximation of a language model, we arrive at a better understanding of its internal structure, which is crucial for controlling its outputs effectively. This ultimately has implications for many downstream applications of transformer language models, including as knowledge bases, dialogue agents, and as robust tools for inference and reasoning.

Reproducibility statement

The code is based on the LRE repository, and loads GPT-J in half-precision. The code and the dataset are available at [link to be released after review]. Experiments were run remotely on a workstation with 24GB NVIDIA RTX 3090 GPUs using HuggingFace Transformers.

Acknowledgments

All work herein reported is supported by the Nation Science Foundation under Grant No. 2349452. Any opinion, finding, or conclusion in this study is that of the authors and does not necessarily reflect the views of the National Science Foundation.

References

[Brown et al.2020] 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.

[Bruni et al.2012] Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 136–145.

[Chanin et al.2023] David Chanin, Anthony Hunter, and Oana-Maria Camburu. 2023. Identifying linear relational concepts in large language models. *arXiv preprint arXiv:2311.08968.*

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

[Fellbaum1998] Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

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

[Gladkova et al.2016] 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 Jacob Andreas, Eunsol Choi, and Angeliki Lazaridou, editors, *Proceedings of the NAACL Student Research Workshop*, pages 8–15, San Diego, California, June. Association for Computational Linguistics.

[Hernandez et al.2023] 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.*

[Hinton1986] 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.

[Jurgens et al.2012] David Jurgens, Saif Mohammad, Peter Turney, and Keith Holyoak. 2012. SemEval-2012 task 2: Measuring degrees of relational similarity. In Eneko Agirre, Johan Bos, Mona Diab, Suresh Manandhar, Yuval Marton, and Deniz Yuret, editors, **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 356–364, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

[Kann et al.2018] Katharina Kann, Alex Warstadt, Adina Williams, and Samuel R Bowman. 2018. Verb argument structure alternations in word and sentence embeddings. *arXiv preprint arXiv:1811.10773*.

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

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

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

[Min et al.2022] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?, October. arXiv:2202.12837 [cs].

[Paccanaro and Hinton2001] 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.

[Park et al.2023] Kiho Park, Yo Joong Choe, and Victor Veitch. 2023. The linear representation hypothesis and the geometry of large language models.

[Park et al.2024] Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. 2024. The geometry of categorical and hierarchical concepts in large language models.

[Su et al.2024] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing*, 568:127063.

[Tenney et al.2019] Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601.

[Touvron et al.2023] 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*.

[Vaswani et al.2017] 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.

[Vylomova et al.2016] 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.

[Wang and Komatsuzaki2021] Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. `https://github.com/kingoflolz/mesh-transformer-jax`, May.

[Wang et al.2023] Zihao Wang, Lin Gui, Jeffrey Negrea, and Victor Veitch. 2023. Concept algebra for score-based conditional model. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*.

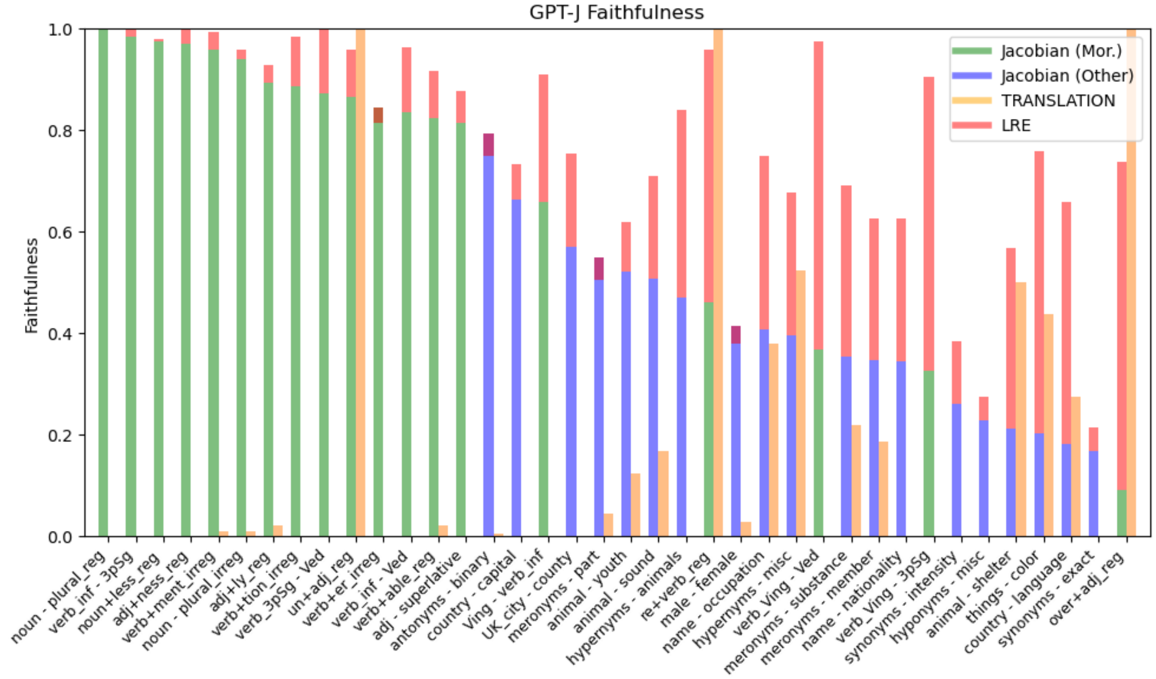
[Wei et al.2024] Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2024. Larger language models do in-context learning differently.

[Wilson et al.2023] Michael Wilson, Jackson Petty, and Robert Frank. 2023. How abstract is linguistic generalization in large language models? experiments with argument structure. *Transactions of the Association for Computational Linguistics*, 11:1377–1395.

[Yan et al.2023] Jianhao Yan, Jin Xu, Chiyu Song, Chenming Wu, Yafu Li, and Yue Zhang. 2023. Understanding in-context learning from repetitions. In *The Twelfth International Conference on Learning Representations*.

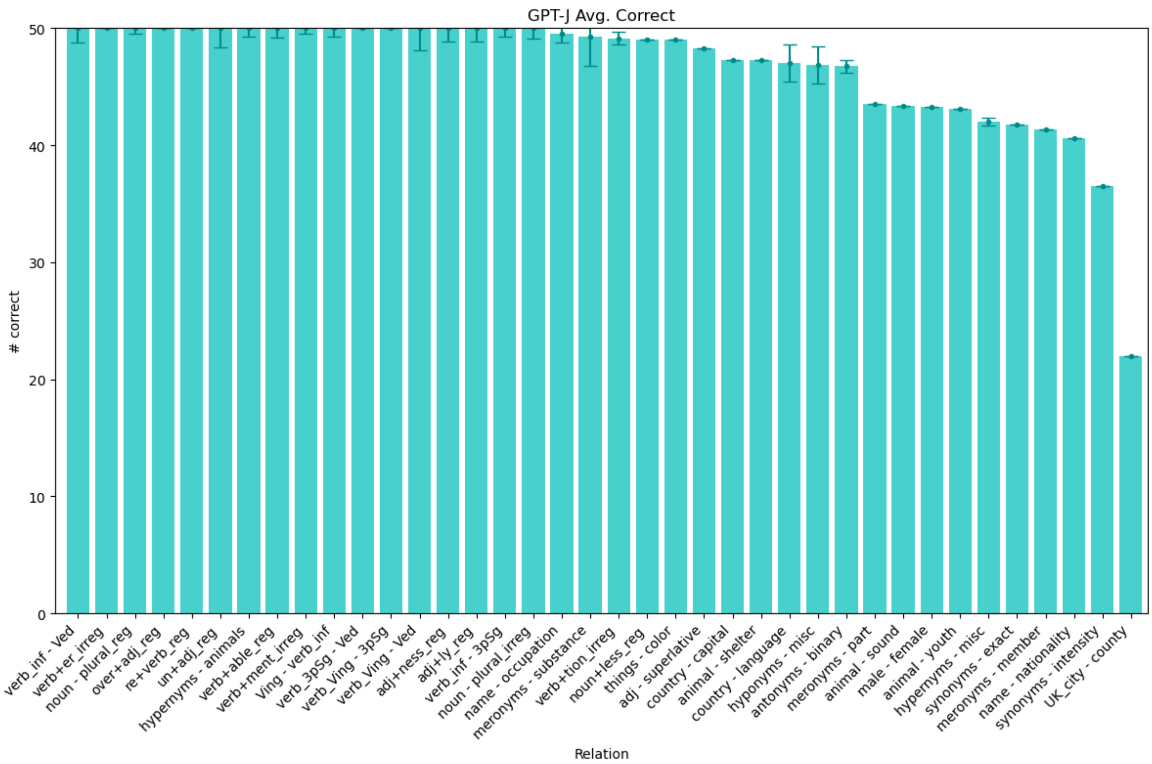
A The Jacobian with TRANSLATION

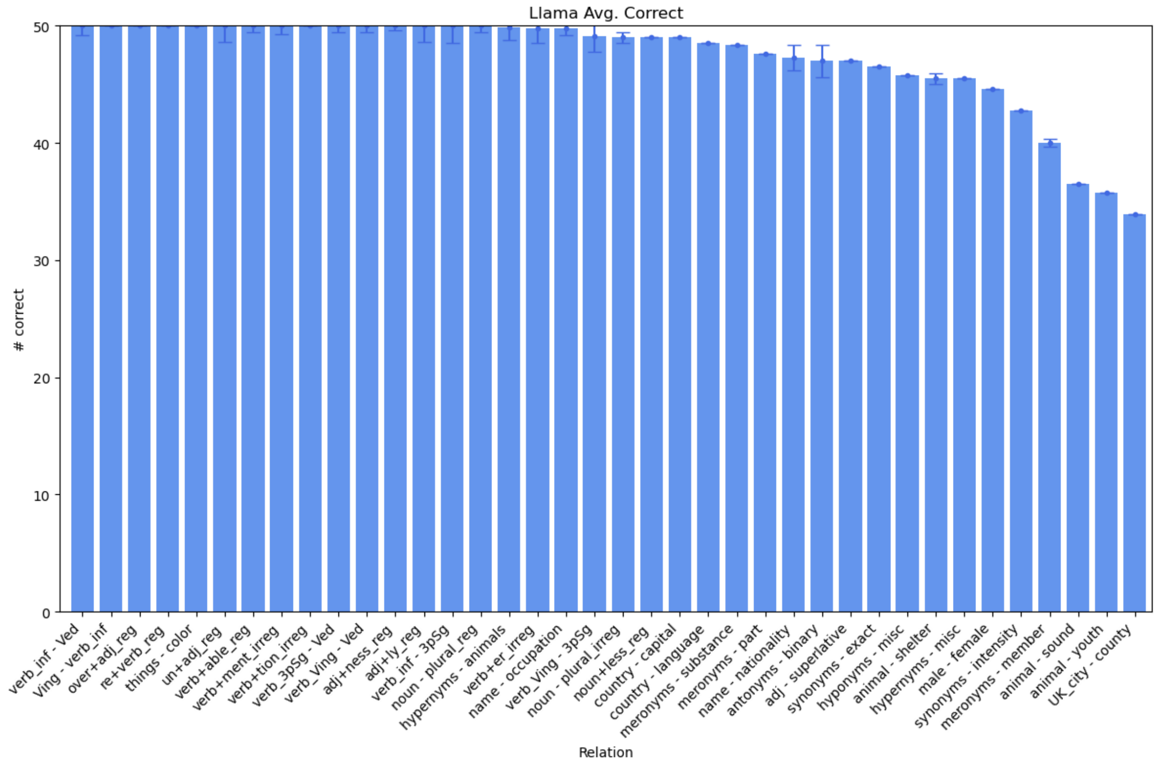
Figure 7: Comparing the affine LRE with the Jacobian ($W\mathbf{s}$) and TRANSLATION ($\mathbb{E}(\mathbf{o} - \mathbf{s})$) approximators yields similar results to above, suggesting that W and b play complementary roles.



B GPT-J and Llama-7b Ability

Figure 8: Both GPT-J and Llama-7b demonstrate mastery of most relations.




Llama Avg. Correct
correct
0
10
20
30
40
50
verb_inf - Ved
Ving - verb_inf
over+adj_reg
re+verb_reg
things - color
un+adj_reg
verb+able_reg
verb+ment_irreg
verb+tion_irreg
verb_3pSg - Ved
verb_Ving - Ved
adj+ness_reg
adj+ly_reg
verb_inf - 3pSg
noun - plural_reg
hypernyms - animals
verb+er_irreg
name - occupation
verb_Ving - 3pSg
noun - plural_irreg
noun+less_reg
country - capital
country - language
meronyms - substance
meronyms - part
name - nationality
antonyms - binary
adj - superlative
synonyms - exact
hyponyms - misc
animal - shelter
hypernyms - misc
male - female
synonyms - intensity
meronyms - member
animal - sound
animal - youth
UK_city - county
Relation